The Moran forest

François Bienvenu, Jean-Jil Duchamps, Félix Foutel-Rodier

The Moran forest

François Bienvenu[1,2], Jean-Jil Duchamps[1,2], and Félix Foutel-Rodier[1,2]

[1] *Center for Interdisciplinary Research in Biology (CIRB), CNRS UMR 7241, Collège de France, PSL Research University, Paris, France*
[2] *Laboratoire de Probabilités, Statistique et Modélisation (LPSM), CNRS UMR 8001, Sorbonne Université, Paris, France*

June 25, 2019

Abstract

Starting from any graph on $\{1, \ldots, n\}$, consider the Markov chain where at each time-step a uniformly chosen vertex is disconnected from all of its neighbors and reconnected to another uniformly chosen vertex. This Markov chain has a stationary distribution whose support is the set of non-empty forests on $\{1, \ldots, n\}$. The random forest corresponding to this stationary distribution has interesting connections with the uniform rooted labeled tree and the uniform attachment tree. We fully characterize its degree distribution, the distribution of its number of trees, and the limit distribution of the size of a tree sampled uniformly. We also show that the size of the largest tree is asymptotically $\alpha \log n$, where $\alpha = (1 - \log(e-1))^{-1} \approx 2.18$, and that the degree of the most connected vertex is asymptotically $\log n / \log \log n$.

Contents

1 Introduction

1.1 The model

Consider a Markov chain on the space of graphs on $\{1, \dots, n\}$ whose transition probabilities are defined as follows: at each time-step,

1. Choose an ordered pair of distinct vertices (u, v) uniformly at random.
2. Disconnect v from all of its neighbors, then connect it to u.

Note that if u is the only neighbor of v at time t, then the graph is unchanged at time $t + 1$. A simple example illustrating the dynamics of this Markov chain is depicted in Figure 1.

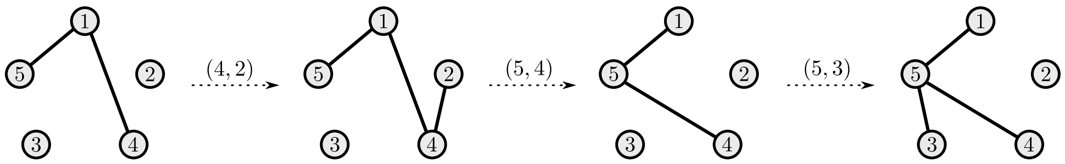


Figure 1: Example of four successive transitions of the Markov chain. Starting from the left-most graph, transitions are represented by dashed arrows decorated with the pair (u, v) that is chosen uniformly at each step.

This Markov chain has a stationary distribution whose support is the set of non-empty forests on $\{1, \dots, n\}$. Indeed,

- The graph cannot be empty because there is always an edge between the two vertices involved in the last transition.
- Starting from any graph, the chain will eventually reach a forest (for instance, the sequence of transitions $(1, 2), (1, 3), \dots, (1, n)$ will at some point turn the graph into the star graph centered on vertex 1).
- The chain cannot leave the set of non-empty forests because its transitions cannot create cycles.
- Any non-empty forest is accessible from any other graph (if not clear, this will become apparent in Section 2).
- The chain is aperiodic because it can stay in the same state.

The stationary distribution of this chain is the random forest model that we study in this paper. We denote it by $\mathscr{F}_n$ and call it the *Moran forest*, in reference to the Moran model of population genetics, where at each time step two distinct individuals are sampled uniformly at random, and the second one is replaced by a copy of the first [9, 11, 17].

1.2 Main results

Our first result, which we detail in Section 2, is that there is a simple way to sample $\mathscr{F}_n$. This construction enables us to study several of its statistics, such as its number of trees (Section 3.1), its degree distribution (Section 4.1), and the typical size of its trees (Section 5.2). Some of these results are presented in Table 1.

Notation	Variable	Distribution
N_n	Number of trees	$\sum_{\ell=1}^{n} I_\ell$, where $I_\ell \sim \mathrm{Ber}\left(\frac{\ell-1}{n-1}\right)$
D	Asymptotic degree distribution	$\mathrm{Ber}(1-U) + \mathrm{Poisson}(U)$, where $U \sim \mathrm{Unif}([0,1])$
T^U	Asymptotic size of a uniform tree	$\mathrm{Geometric}(e^{-X})$, where $X \sim 2xdx$ on $[0,1]$

Table 1: Some statistics of the Moran forest, for fixed n in the case of the number of trees, and as $n \to \infty$ for the degree and the size of a uniform tree. Note that the degree also has a simple, explicit distribution for fixed n (see Proposition 4.1). The Bernoulli variables I_ℓ used to describe the distribution of N_n are independent and, conditional on U, so are the Bernoulli and Poisson variables used for the distribution of D.

In Section 3.2, we show that the Moran forest is closely linked to uniform rooted labeled trees. Specifically, we prove the following theorem.

Theorem 3.4. *Let $\mathcal{T}$ be a uniform rooted tree on $\{1, \dots, n-1\}$. From this tree, build a forest $\mathcal{F}$ on $\{1, \dots, n\}$ according to the following procedure:*

1. *Remove all decreasing edges from $\mathcal{T}$ (that is, edges $\vec{uv}$ pointing away from the root such that $u > v$).*
2. *Add a vertex labeled n and connect it to a uniformly chosen vertex of $\mathcal{T}$*
3. *Relabel vertices according to a uniform permutation of $\{1, \dots, n\}$.*

Then, the resulting forest $\mathcal{F}$ has the law of the Moran forest $\mathscr{F}_n$.

Finally, we study the asymptotic concentration of the largest degree and of the size of the largest tree of $\mathscr{F}_n$. The following theorems are proved in Sections 4.2 and 5.3, respectively.

Theorem 4.5. *Let $D_n^{\max}$ denote the largest degree of $\mathscr{F}_n$. Then,*

$$D_n^{\max} = \frac{\log n}{\log\log n} + \Big(1 + o_{\mathrm{p}}(1)\Big)\frac{\log n \, \log\log\log n}{(\log\log n)^2},$$

where $o_{\mathrm{p}}(1)$ denotes a sequence of random variables that goes to 0 in probability.

Theorem 5.8. *Let $T_n^{\max}$ denote the size of the largest tree of $\mathscr{F}_n$. Then,*

$$T_n^{\max} = \alpha\Big(\log n - (1 + o_{\mathrm{p}}(1)) \log\log n\Big),$$

where $\alpha = (1 - \log(e-1))^{-1} \approx 2.18019$.

2 Sampling of the stationary distribution

2.1 Backward construction

Consider an i.i.d. sequence $((V_t, W_t),\, t \in \mathbb{Z})$, where (V_t, W_t) is uniformly distributed on the set of ordered pairs of distinct elements of $\{1, \dots, n\}$. These variables are meant to encode the transitions of the chain: W_t represents the vertex that is disconnected at step t, and V_t the vertex to which W_t is then connected. We now explain how to construct a chain $(\mathscr{F}_n(t),\, t \in \mathbb{Z})$ of forests by looking at the sequence $((V_t, W_t),\, t \in \mathbb{Z})$ backwards in time.

Fix a focal time $t \in \mathbb{Z}$. For each vertex v, let us denote by

$$\tau_t(v) \coloneqq \max\{s \leqslant t : W_s = v\}$$

the last time before t that v was chosen to be disconnected, and define

$$m_t(v) \coloneqq V_{\tau_t(v)}$$

to be the vertex to which it was then reconnected. We refer to the time $\tau_t(v)$ as the *birth time* of v, and to the vertex $m_t(v)$ as its *mother*. Note that the variables $(\tau(v),\, 1 \leqslant v \leqslant n)$ are independent of $(m(v),\, 1 \leqslant v \leqslant n)$.

Now, for each $s \leqslant t$, let the vertices be in one of two states, *active* or *inactive*, as follows: vertex v is active at times s such that $\tau_t(v) \leqslant s \leqslant t$, and inactive at times $s < \tau_t(v)$. Finally, let $\mathscr{F}_n(t)$ be the forest obtained by connecting each vertex v to its mother if the mother is active at the time of birth of v, that is,

$$v \text{ is connected to } m_t(v) \iff \tau_t(m_t(v)) < \tau_t(v).$$

This procedure is illustrated in Figure 2.

Let us show that the chain $(\mathscr{F}_n(t),\, t \in \mathbb{Z})$ has the same transitions as the chain described in the introduction. First, note that for $v \neq W_t$ we have $\tau_t(v) = \tau_{t-1}(v)$, and thus $m_t(v) = m_{t-1}(v)$. As a result, edges that do not involve W_t are the same in $\mathscr{F}_n(t)$ and in $\mathscr{F}_n(t-1)$. Now, $\tau_t(W_t) = t$, so that W_t is always inactive as a mother in the construction of $\mathscr{F}_n(t)$, and $m_t(W_t) = V_t$ with $\tau_t(V_t) < t$, so that W_t is linked to V_t in $\mathscr{F}_n(t)$. In other words, $\mathscr{F}_n(t)$ is obtained from $\mathscr{F}_n(t-1)$ by disconnecting W_t from its neighbors, and then connecting it to V_t. This corresponds to the transitions of the chain described in the introduction.

Finally, $(\mathscr{F}_n(t),\, t \in \mathbb{Z})$ is stationary by construction, and thus $\mathscr{F}_n(t)$ is distributed as the Moran forest for all time $t \in \mathbb{Z}$.

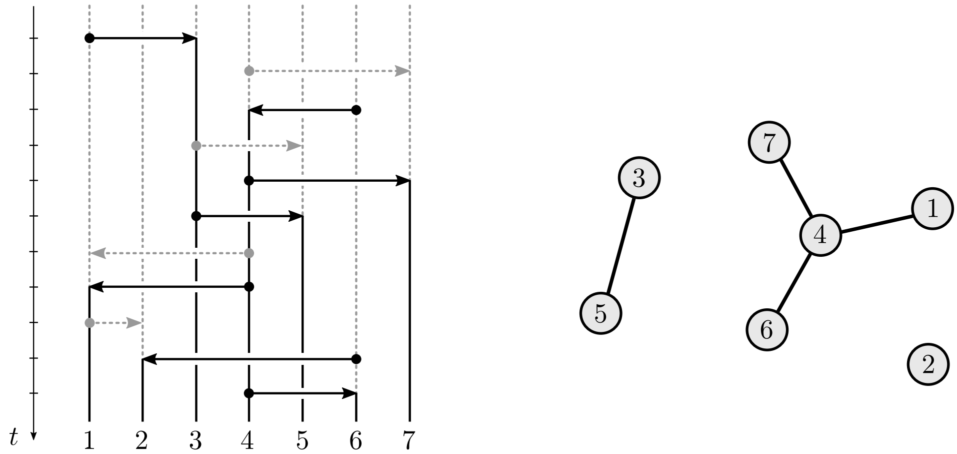


Figure 2: Illustration of the backward construction. Each vertex corresponds to a vertical line. A pair (V_t, W_t) is represented by an arrow $V_t \to W_t$. The line representing a vertex is solid black when that vertex is active, and dashed grey when it is inactive. Arrows pointing to inactive vertices are represented in dashed grey because they have no impact on the state of the graph at the focal time: their effect has been erased by subsequent arrows.

2.2 Uniform attachment construction

We now give a forward-in-time variant of the construction described in the previous section. This forward-in-time procedure, which we call the *uniform attachment construction* (UA construction for short), is our main tool to study $\mathscr{F}_n$ and will be used throughout the rest of the paper.

Let $(U_n(\ell),\ 1 \leqslant \ell \leqslant n)$ be a vector of independent variables such that $U_n(\ell)$ is uniformly distributed on $\{1,\dots,n\} \setminus \{\ell\}$. Consider the forest $\mathscr{F}_n^*$ on $\{1,\dots,n\}$ obtained by setting

$$k \text{ is connected to } \ell, \text{ with } k < \ell \iff U_n(\ell) = k.$$

We will show that, after relabeling the vertices of $\mathscr{F}_n^*$ according to a uniform permutation of $\{1,\dots,n\}$, we obtain the Moran forest. Before this let us make a few remarks.

First, it will be helpful to think of the construction of $\mathscr{F}_n^*$ as a sequential process where, starting from a single vertex labeled 1, for $\ell = 2,\dots,n$ we add a new vertex labeled ℓ and connect it to $U_n(\ell)$ if $U_n(\ell) < \ell$. See Figure 3. This will make the link with some well-known stochastic processes more intuitive. This also explains that we speak of the *ℓ-th vertex in the UA construction* to refer to vertex ℓ in $\mathscr{F}_n^*$.

Second, the edges of $\mathscr{F}_n^*$ are by construction increasing, in the sense that if we root every tree of $\mathscr{F}_n^*$ by letting the root of a tree be its smallest vertex, then each edge $\vec{uv}$ pointing away from the root of its tree is such that $u < v$.

Rooted trees that have only increasing edges are known as *recursive trees* [8], and forests of recursive trees have been called *recursive forests* [4]. Recursive trees have been studied extensively. In particular, the uniform attachment tree, which corresponds to the uniform distribution over the set of recursive trees, has received much attention [5, 15, 16]. However, the random forest $\mathscr{F}_n^*$ does not seem to correspond to any previously studied model of random recursive forest (in particular, it is not uniformly distributed over the set of recursive forests).

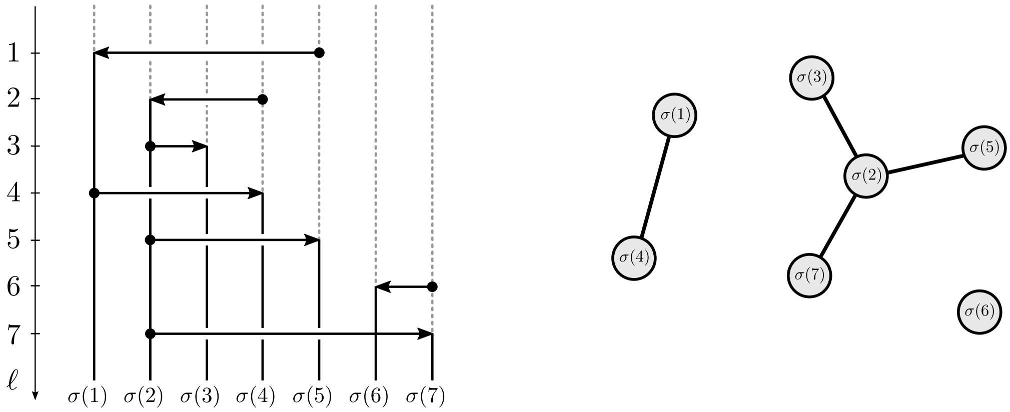


Figure 3: Illustration of the uniform attachment construction for $n = 7$ and the vector $(U_n(1), \dots, U_n(n)) = (5, 4, 2, 1, 2, 7, 2)$. The ℓ-th vertical line from the left corresponds to vertex $\sigma(\ell)$ (i.e, in the sequential vision, to the ℓ-th vertex that is added). $U_n(\ell)$ is represented by the arrow pointing from the $U_n(\ell)$-th line to the ℓ-th one at time ℓ. Compare this with Figure 2: the vertical lines corresponding to the vertices have been reordered in increasing order of their birth time, and the grey arrows that left no trace on the graph at the focal time have been removed.

Proposition 2.1. *The random forest obtained by relabeling the vertices of* $\mathscr{F}_n^*$ *according to a uniform permutation of* $\{1, \dots, n\}$ *is distributed as the Moran forest.*

Proof. Consider the forest $\mathscr{F}_n(0)$ built from the variables $((V_t, W_t),\ t \in \mathbb{Z})$ in the previous section. To ease notation, we will omit the subscript in τ_0 and m_0.

Let us relabel the vertices in increasing order of their birth time: since the variables $(\tau(v),\ 1 \leqslant v \leqslant n)$ are all distinct, there exists a unique permutation σ of $\{1, \dots, n\}$ such that

$$\tau(\sigma(1)) < \cdots < \tau(\sigma(n)).$$

In words, $\sigma(\ell)$ is the ℓ-th vertex that was born in the construction of $\mathscr{F}_n(0)$. Using the new labeling, let us denote its birth time by $\tau^*(\ell) = \tau(\sigma(\ell))$ and its mother by $m^*(\ell) = \sigma^{-1}(m(\sigma(\ell)))$.

Now, for every vertex $v = \sigma(\ell)$,

$$\begin{aligned} v \text{ is connected to } m(v) \text{ in } \mathscr{F}_n(0) \iff & \tau(m(v)) < \tau(v) \\ \iff & \tau^*(m^*(\ell)) < \tau^*(\ell) \\ \iff & m^*(\ell) < \ell\,. \end{aligned}$$

Thus, if we set $U_n(\ell) = m^*(\ell)$ in the construction of $\mathscr{F}_n^*$ then ℓ is connected to $m^*(\ell)$ if and only if $v = \sigma(\ell)$ is connected to $m(v) = \sigma(m^*(\ell))$ in $\mathscr{F}_n(0)$. Therefore, to finish the proof we have to show that:

(i) The variables $(m^*(\ell),\ 1 \leqslant \ell \leqslant n)$ are independent and such that $m^*(\ell)$ is uniformly distributed on $\{1, \dots, n\} \setminus \{\ell\}$.

(ii) The permutation σ is uniform and independent of $(m^*(\ell),\ 1 \leqslant \ell \leqslant n)$.

First, note that by construction the variables $(m(v),\ 1 \leqslant v \leqslant n)$ are independent and that for each v, $m(v)$ is uniformly distributed on $\{1, \dots, n\} \setminus \{v\}$. Since the permutation σ depends only on the variables $(\tau(v),\ 1 \leqslant v \leqslant n)$, which are independent of $(m(v),\ 1 \leqslant v \leqslant n)$, we see that σ is independent of $(m(v),\ 1 \leqslant v \leqslant n)$.

Moreover, the variables $(\tau(v),\ 1 \leqslant v \leqslant n)$ are exchangeable so the permutation σ is uniform. Now, for any fixed permutation π of $\{1, \dots n\}$ and any fixed map $f : \{1, \dots, n\} \to \{1, \dots, n\}$ such that $f(\ell) \neq \ell$ for all ℓ, we have

$$\begin{aligned} \mathbb{P}(\sigma = \pi,\ m^* = f) \;&=\; \mathbb{P}\Big(\sigma = \pi,\ m = \pi \circ f \circ \pi^{-1}\Big) \\ &=\; \frac{1}{n!}\frac{1}{(n-1)^n}\,, \end{aligned}$$

concluding the proof. □

3 Number of trees

3.1 Law of the number of trees

In the UA construction, let $I_\ell = \mathbb{1}_{\{U_n(\ell) < \ell\}}$ be the indicator variable of the event "the ℓ-th vertex was linked to a previously added vertex". The variables $(I_1, \dots, I_n)$ are thus independent Bernoulli variables such that

$$I_\ell \;\sim\; \mathrm{Bernoulli}\Big(\tfrac{\ell-1}{n-1}\Big)\,.$$

With this notation, the number of edges $|E_n|$ and the number of trees N_n are

$$|E_n| = \sum_{\ell=1}^{n} I_\ell \quad \text{and} \quad N_n = \sum_{\ell=1}^{n} (1 - I_\ell)\,.$$

Moreover, since $I_\ell \overset{d}{=} 1 - I_{n-\ell+1}$, we see that

$$\mathbb{P}(N_n = k) \;=\; \mathbb{P}(N_n = n-k) \;=\; \mathbb{P}(|E_n| = k)\,,$$

that is, the number of trees and the number of edges have the same, symmetric distribution. In consequence, from now on we only use the notation N_n and refer to it as the number of trees of $\mathscr{F}_n$ when stating our results—even though we sometimes work with the number of edges in the proofs.

From the representation of N_n as a sum of independent Bernoulli variables, we immediately get the following result.

Proposition 3.1. *Let N_n denote the number of trees of $\mathscr{F}_n$.*

(i) $\mathbb{E}(N_n) = \dfrac{n}{2}$

(ii) $\mathrm{Var}(N_n) = \dfrac{n(n-2)}{6(n-1)}$.

(iii) $G_{N_n}(z) := \mathbb{E}\Big(z^{N_n}\Big) = \displaystyle\prod_{k=1}^{n-1}\Big(1 + \tfrac{k}{n-1}(z-1)\Big)$.

The representation of N_n as a sum of independent Bernoulli variables also makes it straightforward to get the following central limit theorem.

Proposition 3.2. *Let N_n denote the number of trees of $\mathscr{F}_n$. Then,*

$$\frac{N_n - n/2}{\sqrt{n/6}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)\,.$$

Proof. This is an immediate consequence of the Lyapunov CLT for triangular arrays of independent random variables. Indeed, $\mathbb{E}\Big(|I_\ell - \mathbb{E}(I_\ell)|^3\Big) \leqslant 1$. Therefore,

$$\frac{1}{n^{3/2}} \sum_{\ell=1}^{n} \mathbb{E}\Big(|I_\ell - \mathbb{E}(I_\ell)|^3\Big) \leqslant \frac{1}{\sqrt{n}} \xrightarrow[n\to\infty]{} 0\,,$$

and the result follows, e.g., from Corollary 11.1.4 in [3]. □

3.2 Link with uniform labeled trees

As announced in the introduction, there is a strong connection between the Moran forest and uniform labeled trees. Our starting point is the following observation about the probability generating function of N_n. First,

$$\begin{aligned}
\sum_{k\geqslant 0} \mathbb{P}(N_n = k)\, z^k \;&=\; \prod_{k=1}^{n-1}\Big(1 + \tfrac{k}{n-1}(z-1)\Big) \\
&=\; \frac{z}{(n-1)^{n-2}} \prod_{k=1}^{n-2} \Big(n-1-k+kz\Big) \\
&=\; \sum_{k=0}^{n-2} \frac{a(n-1,k)}{(n-1)^{n-2}}\, z^{k+1},
\end{aligned}$$

where

$$\sum_{k=0}^{n-2} a(n-1,k)\, z^k \;=\; \prod_{k=1}^{n-2} \Big(n-1-k+kz\Big).$$

Second, the coefficients of this polynomial have a simple combinatorial interpretation: $a(n-1,k)$ is the number of rooted trees on $\{1,\dots,n-1\}$ with k increasing edges, where an edge $\vec{uv}$ pointing away from the root is said to be increasing if $u < v$. This fact is known in the literature as a consequence of the more general Theorem 1.1 of [10] (see also Example 1.7.2 in [7] and Theorem 9.1 in [12]).

This simple observation already gives us the following proposition.

Proposition 3.3. *The probability mass function of the number of trees of $\mathscr{F}_n$ is*

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

where $a(n,k)$ is the number of rooted trees on $\{1,\dots n\}$ with k increasing edges (sequence A067948 of the On-Line Encyclopedia of Integer Sequences [1]).

Looking for a bijective proof of Proposition 3.3 naturally leads to the following more general result about the link between the Moran forest and uniform rooted labeled trees.

Theorem 3.4. *Let $\mathcal{T}$ be a uniform rooted tree on $\{1, \dots, n-1\}$. From this tree, build a forest $\mathcal{F}$ on $\{1, \dots, n\}$ according to the following procedure:*

1. *Remove all decreasing edges from $\mathcal{T}$ (that is, edges $\vec{uv}$ pointing away from the root such that $u > v$).*
2. *Add a vertex labeled n and connect it to a uniformly chosen vertex of $\mathcal{T}$*
3. *Relabel vertices according to a uniform permutation of $\{1, \dots, n\}$.*

Then, the resulting forest $\mathcal{F}$ has the law of the Moran forest $\mathscr{F}_n$.

Proof. In the UA construction, let $F_{|n-1}$ denote the forest obtained after the addition of $n-1$ vertices, before their relabeling. After this, the n-th vertex will be linked to a uniformly chosen vertex of $F_{|n-1}$. As a result, to prove the theorem it suffices to show that $F_{|n-1}$ has the same law as the forest obtained from $\mathcal{T}$ by removing its decreasing edges.

To do so, we couple $F_{|n-1}$ and $\mathcal{T}$ in such a way that the edges of $F_{|n-1}$ are exactly the increasing edges of $\mathcal{T}$. Formally, $F_{|n-1}$ is a deterministic function of the random vector $\mathbf{U} = (U_n(2), \dots, U_n(n-1))$. Moreover, $\mathbf{U}$ is uniform on the set

$$\mathscr{S}^\star_{n-1} \;=\; \Big\{\mathbf{u} \in \{1, \dots, n\}^{\{2,\dots,n-1\}} \,:\, u_\ell \neq \ell\Big\}\,.$$

Thus, to end the proof it is sufficient to find a bijection Φ from $\mathscr{S}^\star_{n-1}$ to the set of rooted trees on $\{1, \dots, n-1\}$ and such that

$$k\ell \in F_{|n-1}(\mathbf{u}) \iff k\ell \text{ is an increasing edge of } \Phi(\mathbf{u})\,.$$

First, let

$$\mathscr{S}_{n-1} \;=\; \{1, \dots, n-1\}^{\{2,\dots,n-1\}}$$

and consider the bijection $\Theta : \mathscr{S}^\star_{n-1} \to \mathscr{S}_{n-1}$ defined by

$$\Theta\mathbf{u} : \ell \mapsto u_\ell - \mathbb{1}_{\{u_\ell > \ell\}}\,.$$

Importantly, note that Θ does not modify the entries of $\mathbf{u}$ that correspond to edges of $F_{|n-1}(\mathbf{u})$, that is, for all $k < \ell$,

$$k\ell \in F_{|n-1}(\mathbf{u}) \iff u_\ell = k \iff (\Theta\mathbf{u})(\ell) = k\,.$$

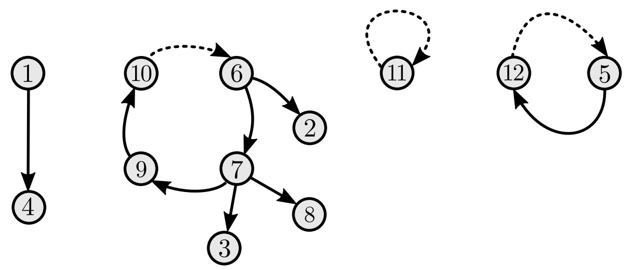
As a result, it remains to find a bijection Ψ from $\mathscr{S}_{n-1}$ to the set of rooted trees on $\{1, \dots, n-1\}$ such that

$$u_\ell < \ell \iff u_\ell \text{ and } \ell \text{ are linked by an increasing edge in } \Psi(\mathbf{u})\,.$$

This bijection will essentially be that used in [10], which can itself be seen as a variant of Joyal's bijection [2, 14].

Let $\mathscr{G}_\mathbf{u}$ be the directed graph on $\{1, \dots, n-1\}$ obtained by putting a directed edge going from u_ℓ to ℓ for all $\ell \geqslant 2$.

If $\mathscr{G}_\mathbf{u}$ has no cycle or self-loop, then it is a tree. Moreover, the orientation of its edges uniquely identify vertex 1 as its root. Thus we set $\Psi(\mathbf{u}) = \mathscr{G}_\mathbf{u}$.



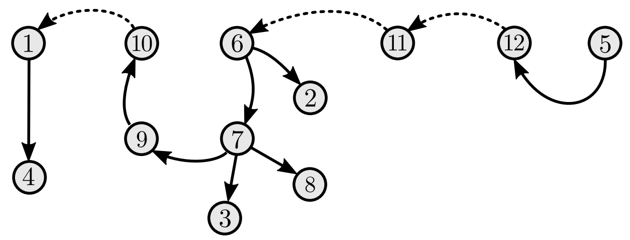


Figure 4: Example of construction of $\Phi(\mathbf{u})$, for $\mathbf{u} = (7, 8, 1, 13, 11, 6, 7, 7, 9, 12, 5)$. Applying Θ yields $\mathbf{u'} = \Theta\mathbf{u} = (6, 7, 1, 12, 10, 6, 7, 7, 9, 11, 5)$. The directed graph $\mathscr{G}_{\mathbf{u'}}$ encoding $\mathbf{u'}$ is represented on top. Its cycles are $\mathscr{C}_1 = (10, 6, 7, 9)$, $\mathscr{C}_2 = (11)$ and $\mathscr{C}_3 = (12, 5)$, and we set $\mathscr{C}_0 = (1)$. The edges $\vec{m_i s_i}$ are dashed. Rewiring them as described in the main text turns $\mathscr{G}_{\mathbf{u'}}$ into the rooted tree $\Psi(\mathbf{u'})$ represented on bottom. No information is lost when turning the cycles $(1)(10, 6, 7, 9)(11)(12, 5)$ into the path going from 5 to 1 encoded by the word $(1, 10, 6, 7, 9, 11, 12, 5)$, because the left-to-right maxima of that word—here 1, 10, 11 and 12—each mark the start of a new cycle.

If $\mathscr{G}_{\mathbf{u}}$ is not a tree, set $\mathscr{C}_0 = \{1\}$ and let $\mathscr{C}_1, \dots, \mathscr{C}_k$ denote the cycles of $\mathscr{G}_{\mathbf{u}}$, taken in increasing order of their largest element and treating self-loops as cycles of length 1. Note that because each vertex has exactly one incoming edge, except for vertex 1 which has none, these cycles are vertex-disjoint and directed.

To turn $\mathscr{G}_{\mathbf{u}}$ into a tree, set $s_0 = 1$ and for $i \geqslant 1$ let m_i denote the largest element of $\mathscr{C}_i$ and $\vec{m_i s_i}$ its out-going edge in $\mathscr{C}_i$. With this notation, for $i = 1, \dots, k$ remove the edge $\vec{m_i s_i}$ from $\mathscr{G}_{\mathbf{u}}$ and replace it by $\vec{m_i s_{i-1}}$. Note that

- This turns $\mathscr{C}_0 \sqcup \dots \sqcup \mathscr{C}_k$ into a directed path $\mathscr{P}$ going from s_k to 1.
- Because $m_i = \max \mathscr{C}_i$ and that $1 < m_1 < \dots < m_k$, every edge $\vec{m_i s_i}$ was non-increasing and has been replaced by the decreasing edge $\vec{m_i s_{i-1}}$.

Therefore, this procedure turns $\mathscr{G}_{\mathbf{u}}$ into a tree $\Psi(\mathbf{u})$ rooted in s_k, without modifying its increasing edges. Consequently, the increasing edges of $\Psi(\mathbf{u})$ are exactly the pairs $k\ell$ for which $k = u_\ell < \ell$.

To see that Ψ is a bijection, it suffices to note that the cycles $\mathscr{C}_0, \dots, \mathscr{C}_k$ can be recovered unambiguously from the path $\mathscr{P}$ going from the root to vertex 1. Indeed, writing this path as the word $1 m_1 \cdots s_1 m_2 \cdots s_k$, the m_i are exactly the left-to-right maxima of that word.

Setting $\Phi = \Psi \circ \Theta$ thus gives us the bijection that we were looking for, concluding the proof. $\square$

4 Degrees

4.1 Degree of a fixed vertex

Using the UA construction and the notation from Section 2.2, let us denote by

- $I_\ell = \mathbb{1}_{\{U_n(\ell)<\ell\}}$ the indicator variable of the event "the ℓ-th vertex is linked to a previously added vertex".
- $X_\ell^{(v)} = \mathbb{1}_{\{U_n(\ell)=\sigma^{-1}(v)\}}$ the indicator variable of the event "the ℓ-th vertex is linked to vertex v".
- $B_v = \sigma^{-1}(v)$ the step of the construction at which vertex v is added.

With this notation, the degree of vertex v is

$$D_n^{(v)} \;=\; I_{B_v} \;+ \sum_{\ell=B_v+1}^{n} X_\ell^{(v)}.$$

Moreover, conditional on $\{B_v = b\}$, $(X_{b+1}^{(v)}, \ldots, X_n^{(v)})$ are i.i.d. Bernoulli variables with parameter $1/(n-1)$ and I_b is a Bernoulli variable with parameter $\frac{b-1}{n-1}$ that is independent of $(X_{b+1}^{(v)}, \ldots, X_n^{(v)})$. As a result, conditional on B_v and writing L_v for $n - B_v$,

$$D_n^{(v)} \;\overset{d}{=}\; \mathrm{Ber}\Big(1 - \tfrac{L_v}{n-1}\Big) \;+\; \mathrm{Bin}\Big(L_v,\, \tfrac{1}{n-1}\Big)\,,$$

where the Bernoulli and the binomial variables are independent conditional on L_v. Using that L_v is uniformly distributed on $\{0, \ldots, n-1\}$, the mean, variance and probability generating function of $D_n^{(v)}$ are obtained by routine calculations.

Proposition 4.1. *Let D_n be the degree of a fixed vertex of $\mathscr{F}_n$. Then,*

(i) $\mathbb{E}(D_n) = 1$.

(ii) $\mathrm{Var}(D_n) = \dfrac{2(n-2)}{3(n-1)}$.

(iii) $G_{D_n}(z) := \mathbb{E}(z^{D_n}) = \dfrac{1}{n}\displaystyle\sum_{\ell=0}^{n-1}\Big(1 + (1 - \tfrac{\ell}{n-1})(z-1)\Big)\Big(1 + \tfrac{1}{n-1}(z-1)\Big)^{\ell}$.

(iii') $G_{D_n}(z) = 2\Big(1 - \dfrac{1}{n}\Big)\dfrac{\Big(1 + \frac{z-1}{n-1}\Big)^{n} - 1}{z \,-\, 1} - 1$.

Remark 4.2. Note that we also have $\mathbb{E}(D_n^{(v)} \mid L_v) = 1$, that is, the average degree of a vertex is independent of the step at which it was added in the UA construction. ◇

Proposition 4.3. *The degree D_n of a fixed vertex of $\mathscr{F}_n$ converges in distribution to the variable D satisfying:*

(i) $D \sim \mathrm{Ber}(1-U) + \mathrm{Poisson}(U)$, *where U is uniform on $[0,1]$ and the Bernoulli and Poisson variables are independent conditional on U.*

(ii) $G_D(z) := \mathbb{E}\Big(z^D\Big) = \displaystyle\int_0^1 \Big(1 + (1-x)(z-1)\Big)e^{x(z-1)}dx = 2\,\frac{e^{z-1} - 1}{z-1} - 1$.

(iii) *For all* $p \geqslant 1$, $\mathbb{E}\Big(D(D-1)\cdots(D-p+1)\Big) = \dfrac{2}{p+1}$.

(iv) $\mathbb{P}(D=0) = 1 - 2/e$ *and, for* $k \geqslant 1$,

$$\mathbb{P}(D=k) \;=\; \frac{2}{e}\sum_{j>k}\frac{1}{j!}\,.$$

Proof. First, for all $z \in \setminus\{1\}$,

$$G_{D_n}(z) \;=\; 2\Big(1-\frac{1}{n}\Big)\frac{\Big(1+\frac{z-1}{n-1}\Big)^n - 1}{z\;-\;1} - 1 \;\xrightarrow[n\to\infty]{}\; 2\,\frac{e^{z-1}-1}{z-1} - 1.$$

This pointwise convergence of the probability generating function of D_n proves the convergence in distribution of D_n to a random variable D satisfying (ii). Point (i) then follows immediately from the integral expression of G_D.

To compute the factorial moments of D, note that

$$G_D(z) \;=\; 2\sum_{k\geqslant 0}\frac{(z-1)^k}{(k+1)!} \;-\; 1.$$

As a result, for $p \geqslant 1$ the p-th derivative of G_D is

$$G_D^{(p)}(z) \;=\; 2\sum_{k\geqslant 0}\frac{(z-1)^k}{(k+1+p)k!}\,,$$

and, in particular, $\mathbb{E}(D(D-1)\cdots(D-p+1)) = G_D^{(p)}(1) = \frac{2}{p+1}$, proving (iii).

Finally, to prove (iv), using (i) we see that

$$\mathbb{P}(D=0) \;=\; \int_0^1 xe^{-x}\,dx \;=\; 1-\frac{2}{e}$$

and that, for $k \geqslant 1$,

$$\mathbb{P}(D=k) \;=\; \frac{1}{k!}\int_0^1\Big(kx^{k-1}-kx^k+x^{k+1}\Big)\,e^{-x}\,dx.$$

Noting that $\Big(kx^{k-1}-kx^k+x^{k+1}\Big)e^{-x} = 2x^ke^{-x} + \frac{d}{dx}\Big((x^k-x^{k+1})e^{-x}\Big)$, we get

$$\mathbb{P}(D=k) = \frac{2}{k!}\int_0^1 x^ke^{-x}\,dx,$$

and an easy integration by parts yields

$$\mathbb{P}(D=k+1) \;=\; \mathbb{P}(D=k) \;-\; \frac{2}{e(k+1)!}\,,$$

from which (iv) follows by induction. □

Before closing this section, let us give an asymptotic equivalent of the tail of D_n. We will need it in the proof of Theorem 4.5 on the largest degree.

Proposition 4.4. *Let D_n be the degree of a fixed vertex of $\mathscr{F}_n$ and let D have the asymptotic distribution of D_n.*

(i) *For all $k \geqslant 1$,*

$$\frac{2/e}{(k+1)!} \;\leqslant\; \mathbb{P}(D \geqslant k) \;\leqslant\; \left(1+\frac{1}{k}\right)^2 \frac{2/e}{(k+1)!}\,.$$

(ii) *For all $K_n = o(\sqrt{n})$, there exists $\varepsilon_n = o(1)$ such that, for all $k \leqslant K_n$,*

$$|\mathbb{P}(D_n \geqslant k) - \mathbb{P}(D \geqslant k)| \;\leqslant\; \varepsilon_n \, \mathbb{P}(D \geqslant k)\,.$$

(iii) *For all $k_n \to +\infty$ and $K_n \geqslant k_n$ such that $K_n = o(\sqrt{n})$,*

$$\mathbb{P}(D_n \geqslant k) \;\sim\; \frac{2/e}{(k+1)!}\,,$$

uniformly in k such that $k_n \leqslant k \leqslant K_n$.

Proof. First, observe that

$$\frac{1}{(\ell+1)!} \;\leqslant\; \frac{1}{\ell \cdot \ell!} - \frac{1}{(\ell+1)\cdot(\ell+1)!}\,,$$

so that

$$\sum_{\ell > i} \frac{1}{\ell!} \;\leqslant\; \frac{1}{i \cdot i!} \;=\; \left(1+\frac{1}{i}\right)\frac{1}{(i+1)!}\,.$$

Recalling from Proposition 4.3 that

$$\mathbb{P}(D \geqslant k) \;=\; \frac{2}{e}\sum_{i \geqslant k}\sum_{\ell > i}\frac{1}{\ell!}\,,$$

point (i) follows readily.

The proof of (ii) is somewhat technical so we only outline it here and refer the reader to Section A.1 of the Appendix for the detailed calculations.

Consider the function

$$\Delta_n(z) \;=\; \sum_{i \geqslant 0}\Big(\mathbb{P}(D \geqslant i) - \mathbb{P}(D_n \geqslant i)\Big) z^i.$$

With this function, (ii) can be re-expressed as

$$\Delta_n^{(k)}(0) \;=\; \frac{\varepsilon_n}{k+1} \quad \text{for all } k \leqslant K_n = o(\sqrt{n}),$$

where $\Delta_n^{(k)}$ denotes the k-th derivative of Δ_n. But Δ_n can be expressed in terms of the generating functions of D and D_n, namely as

$$\Delta_n(z) \;=\; \left(1 + \frac{1}{z-1}\right)\Big(G_D(z) - G_{D_n}(z)\Big).$$

The expressions of G_D and G_{D_n} obtained in Propositions 4.1 and 4.3 thus make it straightforward to obtain a power series expansion of Δ_n at $z = 1$, and this expansion can be used to bound $\Delta_n^{(k)}(0)$ and conclude the proof.

Finally, (iii) is a direct consequence of (i) and (ii). $\square$

4.2 Largest degree

The aim of this section is to prove the following result.

Theorem 4.5. *Let $D_n^{\max} = \max_v D_n^{(v)}$ denote the largest degree of $\mathscr{F}_n$. Then,*

$$D_n^{\max} = \frac{\log n}{\log\log n} + \Big(1 + o_{\mathrm{p}}(1)\Big)\frac{\log n \, \log\log\log n}{(\log\log n)^2},$$

where $o_{\mathrm{p}}(1)$ denotes a sequence of random variables that goes to 0 in probability.

Our proof uses a first and second moment method that will also be used in the proof of Theorem 5.8 concerning the size of the largest tree. In order to avoid repeating ourselves, we isolate this classic part of our reasoning as a lemma, whose proof we recall for the sake of completeness.

Lemma 4.6. *For all integers n, let $(X_n^{(1)}, \dots, X_n^{(n)})$ be a vector of exchangeable random variables and*

$$X_n^{\max} = \max\Big\{X_n^{(i)} : i = 1, \dots, n\Big\}.$$

Write $p_n(k)$ for $\mathbb{P}(X_n^{(i)} \geqslant k)$, and suppose that there exists a sequence (m_n) and a constant β such that, for all $\varepsilon > 0$, as $n \to \infty$,

(i) $np_n((\beta + \varepsilon)m_n) \to 0$.

(ii) $np_n((\beta - \varepsilon)m_n) \to +\infty$.

(iii) $\mathbb{P}\Big(X_n^{(1)} \geqslant (\beta - \varepsilon)m_n,\, X_n^{(2)} \geqslant (\beta - \varepsilon)m_n\Big) \sim p_n((\beta - \varepsilon)m_n)^2$.

Then for all $\varepsilon > 0$,

$$\mathbb{P}\Big(X_n^{\max} \geqslant (\beta + \varepsilon)\, m_n\Big) \to 0 \quad \text{and} \quad \mathbb{P}\Big(X_n^{\max} \geqslant (\beta - \varepsilon) m_n\Big) \to 1,$$

which can also be written

$$X_n^{\max} = \Big(\beta + o_{\mathrm{p}}(1)\Big)\, m_n,$$

where $o_{\mathrm{p}}(1)$ denotes a sequence of random variables that goes to 0 in probability.

Proof. First,

$$\begin{aligned}\mathbb{P}\Big(X_n^{\max} \geqslant (\beta + \varepsilon)m_n\Big) &= \mathbb{P}\left(\bigcup_{i=1}^{n}\Big\{X_n^{(i)} \geqslant (\beta + \varepsilon)m_n\Big\}\right)\\ &\leqslant np_n\Big((\beta + \varepsilon)m_n\Big),\end{aligned}$$

which goes to zero by (i). Now, denote by

$$Z_n = \sum_{i=1}^{n} \mathbb{1}_{\{X_n^{(i)} \geqslant (\beta - \varepsilon)m_n\}}$$

the number of variables $X_n^{(i)}$ that are greater than or equal to $(\beta - \varepsilon)m_n$. By the Cauchy–Schwartz inequality,

$$\mathbb{P}\Big(X_n^{\max} \geqslant (\beta - \varepsilon)m_n\Big) \;=\; \mathbb{P}(Z_n > 0) \;\geqslant\; \frac{\mathbb{E}(Z_n)^2}{\mathbb{E}(Z_n^2)}.$$

Moreover,

$$\begin{aligned}\mathbb{E}\Big(Z_n^2\Big) &= np_n\Big((\beta-\varepsilon)m_n\Big)\\ &\quad+\ n(n-1)\mathbb{P}\Big(X_n^{(1)}\geqslant(\beta-\varepsilon)m_n,\,X_n^{(2)}\geqslant(\beta-\varepsilon)m_n\Big)\,,\end{aligned}$$

and so, by (ii) and (iii), $\mathbb{E}(Z_n)^2/\mathbb{E}(Z_n^2)\to 1$ as $n\to\infty$. □

Remark 4.7. Note that under assumption (ii) of this lemma, for any $\varepsilon>0$, letting $n\to\infty$ in $\mathbb{E}(Z_n)^2/\mathbb{E}(Z_n^2)\leqslant 1$ shows that

$$p_n((\beta-\varepsilon)m_n)^2\ \leqslant\ \mathbb{P}\Big(X_n^{(1)}\geqslant(\beta-\varepsilon)m_n,\,X_n^{(2)}\geqslant(\beta-\varepsilon)m_n\Big)(1+o(1)).$$

Therefore, to prove (iii) it suffices to show

(iii') $\mathbb{P}\Big(X_n^{(1)}\geqslant(\beta-\varepsilon)m_n,\,X_n^{(2)}\geqslant(\beta-\varepsilon)m_n\Big)\ \leqslant\ p_n((\beta-\varepsilon)m_n)^2(1+o(1)).$ ◇

We now turn to the proof of Theorem 4.5.

Proof of Theorem 4.5. Instead of proving the theorem directly for the variables $(D_n^{(1)},\dots,D_n^{(n)})$, we prove it for some auxiliary variables $(\tilde{D}_n^{(1)},\dots,\tilde{D}_n^{(n)})$ whose maximum has the same asymptotic behavior as $D_n^{\max}$. The point in doing this is that the tails of the variables $\tilde{D}_n^{(v)}$ are less correlated than those of the variables $D_n^{(v)}$, making it easier to study their maximum by the first and second moment method.

Remember from Section 4.1 that, in the UA construction,

$$D_n^{(v)}\ =\ I_{B_v}\ +\sum_{\ell=B_v+1}^{n}X_\ell^{(v)},$$

where B_v is the step at which vertex v was added, $X_\ell^{(v)}$ is the indicator of "the ℓ-th vertex is linked to vertex v", and I_ℓ is the indicator of "the ℓ-th vertex is linked to a previously added vertex". With this notation, set

$$\tilde{D}_n^{(v)}\ =\sum_{\ell=B_v+1}^{n}X_\ell^{(v)}$$

and $\tilde{D}_n^{\max}=\max\{\tilde{D}_n^{(v)}:v=1,\dots,n\}$. Since $\tilde{D}_n^{\max}$ and $D_n^{\max}$ differ by at most 1, for any $m_n\to+\infty$,

$$D_n^{\max}-\tilde{D}_n^{\max}\ =\ o_{\mathrm{p}}(m_n)\,,$$

i.e. $(D_n^{\max}-\tilde{D}_n^{\max})/m_n$ goes to 0 in probability. Thus, to prove the theorem we apply Lemma 4.6 to the variables

$$\left(\tilde{D}_n^{(1)}-\frac{\log n}{\log\log n},\dots,\ \tilde{D}_n^{(n)}-\frac{\log n}{\log\log n}\right),$$

with $m_n=(\log n)(\log\log\log n)/(\log\log n)^2$ and $\beta=1$.

Using Proposition 4.4 and Stirling's formula, we see that for any $k_n=o(\sqrt{n})$,

$$\log\Big(\mathbb{P}(D_n\geqslant k_n)\Big)\ =\ -k_n\log k_n\ +\ k_n\ +\ O(\log k_n)\,.$$

Writing $\tilde{D}_n$ to refer to the common distribution of the variables $\tilde{D}_n^{(v)}$, since

$$\mathbb{P}\Big(D_n \geqslant k_n + 1\Big) \;\leqslant\; \mathbb{P}\Big(\tilde{D}_n \geqslant k_n\Big) \;\leqslant\; \mathbb{P}\Big(D_n \geqslant k_n\Big)\,,$$

we also have

$$\log\Big(\mathbb{P}\Big(\tilde{D}_n \geqslant k_n\Big)\Big) \;=\; -k_n \log k_n \;+\; k_n \;+\; O(\log k_n)\,.$$

In particular, for $k_n = (\log n)/(\log\log n) + \gamma\, m_n$ with

$$m_n = \frac{\log n\, \log\log\log n}{(\log\log n)^2}\,,$$

this gives

$$\log\Big(\mathbb{P}\Big(\tilde{D}_n \geqslant k_n\Big)\Big) \;=\; -\log n \;-\; \Big(\gamma - 1\Big)\,\frac{\log n\;\log\log\log n}{\log\log n} \;+\; O\bigg(\frac{\log n}{\log\log n}\bigg) \quad (1)$$

As a result, for all $\varepsilon > 0$,

(i) $\quad n\,\mathbb{P}\bigg(\tilde{D}_n - \dfrac{\log n}{\log\log n} \geqslant (1+\varepsilon)m_n\bigg) \to 0.$

(ii) $\quad n\,\mathbb{P}\bigg(\tilde{D}_n - \dfrac{\log n}{\log\log n} \geqslant (1-\varepsilon)m_n\bigg) \to +\infty.$

Thus, to apply Lemma 4.6 and finish the proof it suffices to show that

$$\mathbb{P}\Big(\tilde{D}_n^{(1)} \geqslant k_n,\, \tilde{D}_n^{(2)} \geqslant k_n\Big) \;\sim\; \mathbb{P}\Big(\tilde{D}_n \geqslant k_n\Big)^2$$

whenever $k_n = (\log n)/(\log\log n) + (1-\varepsilon)m_n$. More precisely, using Remark 4.7 it is sufficient to show that

$$\mathbb{P}\Big(\tilde{D}_n^{(1)} \geqslant k_n,\, \tilde{D}_n^{(2)} \geqslant k_n\Big) \;\leqslant\; \mathbb{P}\Big(\tilde{D}_n \geqslant k_n\Big)^2 \;+\; o\Big(\mathbb{P}\Big(\tilde{D}_n \geqslant k_n\Big)^2\Big).$$

First let us fix $b_1 \neq b_2 \in \{1,\dots,n\}$. Conditional on $\{B_1 = b_1, B_2 = b_2\}$, recall that the variables $(X_\ell^{(2)},\, b_2 + 1 \leqslant \ell \leqslant n)$ are independent Bernoulli variables with parameter $1/(n-1)$. By further conditioning on the variables $X_\ell^{(1)}$, the independence of $(X_\ell^{(2)},\, b_2+1 \leqslant \ell \leqslant n)$ still holds but their distribution is changed. Indeed, choose $(x_\ell,\, \ell \neq b_1) \in \{0,1\}^{n-1}$ and consider the event

$$A \coloneqq \Big\{B_1 = b_1, B_2 = b_2,\; \forall \ell \neq b_1, X_\ell^{(1)} = x_\ell\Big\}.$$

Then by construction, for all $\ell \notin \{b_1, b_2\}$, we have

$$\mathbb{P}\Big(X_\ell^{(2)} = 1 \,\Big|\, A\Big) = \begin{cases} 0 & \text{if } x_\ell = 1 \\ \frac{1}{n-2} & \text{if } x_\ell = 0. \end{cases}$$

Consequently $X_\ell^{(2)}$ is always stochastically dominated by a Bernoulli($\frac{1}{n-2}$) random variable, and so we bound the distribution of $\tilde{D}_n^{(2)} = \sum_{\ell > b_2} X_\ell^{(2)}$ conditional on A by

$$\Big(\tilde{D}_n^{(2)} \,\Big|\, A\Big) \;\overset{d}{\leqslant}\; \text{Binomial}\Big(n - b_2, \frac{1}{n-2}\Big)\,.$$

To get a bound on the distribution of $\tilde{D}_n^{(2)}$ conditional on $\tilde{D}_n^{(1)} = i$ for some i, first note that summing over all configurations $b_1, b_2, (x_\ell, \ell \neq b_1)$ such that $\sum_{\ell > b_1} x_\ell = i$ gives

$$\left(\tilde{D}_n^{(2)} \,\middle|\, B_1 = b_1, B_2 = b_2,\, \tilde{D}_n^{(1)} = i\right) \;\overset{d}{\leqslant}\; \text{Binomial}\left(n - b_2, \frac{1}{n-2}\right).$$

Let us now write for conciseness $L_1 = n - B_1$ and $L_2 = n - B_2$. Note that L_2 is not independent of $\{\tilde{D}_n^{(1)} = i\}$ because they are linked by L_1. Indeed, L_1 is positively correlated to $\tilde{D}_n^{(1)}$ and we always have $L_2 \neq L_1$. Nevertheless, since conditional on L_1, L_2 is independent of $\tilde{D}_n^{(1)}$ and uniform on $\{0, \dots, n-1\} \setminus L_1$, we have the following stochastic ordering:

$$\left(L_2 \,\middle|\, B_1 = b_1,\, \tilde{D}_n^{(1)} = i\right) \;\overset{d}{\leqslant}\; \bar{L}_2\,,$$

where $\bar{L}_2$ is uniformly distributed on $\{1, \dots, n-1\}$. Summing over b_1 and b_2, we thus get

$$\left(\tilde{D}_n^{(2)} \,\middle|\, \tilde{D}_n^{(1)} = i\right) \;\overset{d}{\leqslant}\; \text{Binomial}\left(\bar{L}_2, \frac{1}{n-2}\right).$$

Let us define a random variable $M_n \sim \text{Bin}\left(\bar{L}_2, \frac{1}{n-2}\right)$. As the previous bound is uniform in i, we have

$$\mathbb{P}\left(\tilde{D}_n^{(2)} \geqslant k_n \,\middle|\, \tilde{D}_n^{(1)} \geqslant k_n\right) \;\leqslant\; \mathbb{P}(M_n \geqslant k_n).$$

To conclude, it is sufficient to show that $\mathbb{P}(M_n \geqslant k_n) \sim \mathbb{P}\left(\tilde{D}_n \geqslant k_n\right)$ since this implies

$$\mathbb{P}\left(\tilde{D}_n^{(1)} \geqslant k_n,\, \tilde{D}_n^{(2)} \geqslant k_n\right) \;\leqslant\; \mathbb{P}\left(\tilde{D}_n \geqslant k_n\right)\mathbb{P}(M_n \geqslant k_n) \sim \mathbb{P}\left(\tilde{D}_n \geqslant k_n\right)^2.$$

For this, define on the same probability space as the variables $\bar{L}_2$ and M_n the variable

$$\underline{L}_2 \coloneqq \bar{L}_2 \mathbb{1}_{\{\bar{L}_2 \leqslant n-2\}}.$$

$\underline{L}_2$ is then uniformly distributed on $\{0, \dots, n-2\}$, and we have the equality in distribution

$$M_n \mathbb{1}_{\{\bar{L}_2 \leqslant n-2\}} \;\overset{d}{=}\; \tilde{D}_{n-1} \;\sim\; \text{Binomial}\left(\underline{L}_2, \frac{1}{n-2}\right).$$

As the two variables M_n and $M_n \mathbb{1}_{\{\bar{L}_2 \leqslant n-2\}}$ differ on an event of probability no greater than $1/(n-1)$, we have

$$\mathbb{P}(M_n \geqslant k_n) = \mathbb{P}\left(\tilde{D}_{n-1} \geqslant k_n\right) + O\left(\tfrac{1}{n}\right),$$

and finally (1) with $\gamma = (1 - \varepsilon)$ allows us to conclude that this expression is indeed equivalent to $\mathbb{P}\left(\tilde{D}_n \geqslant k_n\right)$. □

5 Tree sizes

In this section, we study the size of the trees composing the Moran forest. Section 5.2 is concerned with the typical size of these trees, while Section 5.3 focuses on the asymptotics of the size of the largest tree. But before going any further we need to introduce a process that will play a central role throughout the rest of this paper.

5.1 A discrete-time Yule process

Let $\Upsilon_n = (\Upsilon_n(\ell),\, \ell \geqslant 0)$ be the Markov chain defined by $\Upsilon_n(0) = 1$ and the following transition probabilities:

$$\mathbb{P}\Big(\Upsilon_n(\ell+1) = j \,\Big|\, \Upsilon_n(\ell) = i\Big) = \begin{cases} \frac{i}{n-1} & \text{if } j = i+1 \\ 1 - \frac{i}{n-1} & \text{if } j = i, \end{cases}$$

and stopped when reaching n.

The reason why this process will play an important role when studying the trees of $\mathscr{F}_n$ is the following: let $\mathscr{T}_n^{(v)}$ denote the tree containing v, and $\tilde{\mathscr{T}}_n^{(v)}$ the subtree descending from v in the UA construction—that is, letting $m(v)$ denote the mother of v and $\mathscr{T}_n^{(v)} \setminus \{vm(v)\}$ the forest obtained by removing the edge between v and $m(v)$ from $\mathscr{T}_n^{(v)}$ (if that edge existed), $\tilde{\mathscr{T}}_n^{(v)}$ is the tree of $\mathscr{T}_n^{(v)} \setminus \{vm(v)\}$ containing v. Recalling that L_v denotes the number of steps after vertex v was added in the UA construction and letting $\tilde{T}_n^{(v)} = |\tilde{\mathscr{T}}_n^{(v)}|$ be the size of $\tilde{\mathscr{T}}_n^{(v)}$, we have

$$\tilde{T}_n^{(v)} \;\overset{d}{=}\; \Upsilon_n(L_v)\,,$$

where Υ_n is independent of L_v. In particular, the size of a tree created at step $n-h$ of the UA construction is distributed as $\Upsilon_n(h)$.

In the rest of this section, we list a few basic properties of Υ_n that will be used in subsequent proofs.

Lemma 5.1. *For all $0 \leqslant \ell \leqslant n-1$,*

$$\mathbb{E}(\Upsilon_n(\ell)) \;=\; \left(1 + \frac{1}{n-1}\right)^{\ell}.$$

Proof. For $0 \leqslant \ell < n-1$, we have $\Upsilon_n(\ell) < n$ almost surely, therefore we can write

$$\begin{aligned} \mathbb{E}\Big(\Upsilon_n(\ell+1) \,\Big|\, \Upsilon_n(\ell)\Big) \;&=\; \frac{\Upsilon_n(\ell)}{n-1}\Big(\Upsilon_n(\ell)+1\Big) \;+\; \left(1 - \frac{\Upsilon_n(\ell)}{n-1}\right)\Upsilon_n(\ell) \\ &=\; \Upsilon_n(\ell)\left(1 + \frac{1}{n-1}\right), \end{aligned}$$

and the result follows by induction. □

We now compare the discrete-time process Υ_n to the Yule process. By Yule process, we refer to the continuous-time Markov chain $(Y(t),\, t \geqslant 0)$ that jumps from i to $i+1$ at rate i (see e.g. [18], Section 5.3).

Lemma 5.2. *As $n \to \infty$,*

$$\Big(\Upsilon_n(\lfloor tn \rfloor),\, t \geqslant 0\Big) \;\Longrightarrow\; \Big(Y(t),\, t \geqslant 0\Big),$$

where "$\Longrightarrow$" denotes convergence in distribution in the Skorokhod space [6], and $(Y(t),\, t \geqslant 0)$ is a Yule process.

Proof. Since both processes only have increments of $+1$, it suffices to prove that the sequence of jump times of $(\Upsilon_n(\lfloor tn \rfloor),\, t \geqslant 0)$ converges in distribution to that of the Yule process. For $1 \leqslant i \leqslant n$, let

$$t_n(i) = \inf\Big\{\ell \geqslant 0 : \Upsilon_n(\ell) = i\Big\}$$

be the jump times of the chain Υ_n. By the strong Markov property, the variables $(t_n(i+1) - t_n(i),\, 1 \leqslant i \leqslant n-1)$ are independent, and $t_n(i+1) - t_n(i) \sim \mathrm{Geometric}(\frac{i}{n-1})$. Therefore,

$$\Big(\tfrac{1}{n}\Big(t_n(i+1) - t_n(i)\Big),\, 1 \leqslant i \leqslant n-1\Big) \xrightarrow[n\to\infty]{d} (\mathscr{E}(i),\, i \geqslant 1),$$

where the variables $(\mathscr{E}(i),\, i \geqslant 1)$ are independent and $\mathscr{E}(i) \sim \mathrm{Exponential}(i)$. This concludes the proof. $\square$

Lemma 5.3. *For all integers* $0 \leqslant k \leqslant \ell \leqslant n-1$,

$$\mathbb{P}\Big(Y\Big(\tfrac{\ell-k+1}{n-1}\Big) > k\Big) \;\leqslant\; \mathbb{P}(\Upsilon_n(\ell) > k) \;\leqslant\; \mathbb{P}\Big(Y\Big(\lambda_n(k)\tfrac{\ell}{n-1}\Big) > k\Big),$$

where

$$\lambda_n(k) = -\frac{n-1}{k}\log\bigg(1 - \frac{k}{n-1}\bigg).$$

Proof. Let us start with the upper bound, and write $\lambda \coloneqq \lambda_n(k)$ for simplicity. Note that, for all $t \geqslant 0$ and $i \geqslant 1$,

$$\mathbb{P}\Big(Y\Big(t + \tfrac{\lambda}{n-1}\Big) = i \,\Big|\, Y(t) = i\Big) = e^{-\frac{i\lambda}{n-1}},$$

and that we have chosen λ such that if $i \leqslant k$ then

$$e^{-\frac{i\lambda}{n-1}} \;\leqslant\; 1 - \frac{i}{n-1} \;=\; \mathbb{P}\Big(\Upsilon_n(\ell+1) = i \,\Big|\, \Upsilon_n(\ell) = i\Big).$$

Thus, until it reaches $k+1$ individuals, the process Υ_n is dominated by the Markov chain $(Y(\frac{\lambda\ell}{n-1}),\, 0 \leqslant \ell \leqslant n-1)$. This shows that

$$\mathbb{P}(\Upsilon_n(\ell) > k) \;\leqslant\; \mathbb{P}\Big(Y\Big(\tfrac{\lambda\ell}{n-1}\Big) > k\Big),$$

proving the second inequality of the lemma.

To prove the first inequality, we couple Υ_n with a "censored" Yule process Y_{c}. Intuitively, this censoring consists in ignoring births that occur less than $1/(n-1)$ unit of time after another birth.

Formally, we define Y_{c} by specifying the sequence $t_0 = 0 < t_1 < t_2 < \ldots$ of times corresponding to births in the population. Let $(\mathscr{E}_i,\, i \geqslant 1)$ be an independent sequence of exponential random variables where $\mathscr{E}_i \sim \mathrm{Exponential}(i)$. Set $t_0 = 0$ and, for each $i \geqslant 1$,

$$t_i \;\coloneqq\; \mathscr{E}_1 + \sum_{j=2}^{i}\bigg(\frac{1}{n-1} + \mathscr{E}_i\bigg) \;=\; \frac{i-1}{n-1} + \sum_{j=1}^{i}\mathscr{E}_i. \tag{2}$$

We now define, for all $t \geqslant 0$,

$$Y_{\mathrm{c}}(t) \;\coloneqq\; 1 + \sum_{i\geqslant 1}\mathbb{1}_{\{t_i \leqslant t\}} \;=\; \sum_{i \geqslant 1} i\, \mathbb{1}_{\{t_{i-1} \leqslant t < t_i\}}.$$

The censoring of the Yule process after birth events implies that for any time $t \geqslant 0$, the random variable $Y_{\mathrm{c}}(t+\frac{1}{n-1}) - Y_{\mathrm{c}}(t)$ takes values in $\{0,1\}$. Furthermore, for any $i \in \mathbb{N}$,

$$\mathbb{P}\Big(Y_{\mathrm{c}}(t+\tfrac{1}{n-1}) = i+1 \,\Big|\, Y_{\mathrm{c}}(t)=i\Big) \;\leqslant\; 1-e^{-\frac{i}{n-1}} \;\leqslant\; \frac{i}{n-1}.$$

Therefore, we can couple $(\Upsilon_n(\ell),\, 0 \leqslant \ell \leqslant n-1)$ and $(Y_{\mathrm{c}}(t),\, t \geqslant 0)$ in such a way that, for all $0 \leqslant \ell \leqslant n-1$,

$$Y_{\mathrm{c}}\Big(\tfrac{\ell}{n-1}\Big) \;\leqslant\; \Upsilon_n(\ell).$$

Now, by construction, the sequence $(t_i - \frac{i-1}{n-1},\, i \geqslant 1)$ has the distribution of the sequence of jump times of a Yule process. Therefore,

$$\begin{aligned}
\mathbb{P}(\Upsilon_n(\ell) > k) \;&\geqslant\; \mathbb{P}\Big(Y_{\mathrm{c}}\Big(\tfrac{\ell}{n-1}\Big) > k\Big) \\
&=\; \mathbb{P}\Big(t_k \leqslant \tfrac{\ell}{n-1}\Big) \\
&=\; \mathbb{P}\Big(t_k - \tfrac{k-1}{n-1} \leqslant \tfrac{\ell-k+1}{n-1}\Big) \\
&=\; \mathbb{P}\Big(Y\Big(\tfrac{\ell-k+1}{n-1}\Big) > k\Big),
\end{aligned}$$

which yields the lower bound of the lemma. $\square$

5.2 Size of some random trees

In this section, we study the size of some typical trees of $\mathscr{F}_n$. In particular, we study the asymptotics of the size $T_n^{(1)}$ of the tree containing vertex 1 and of the size T_n^U of a tree sampled uniformly at random among the trees composing $\mathscr{F}_n$. Our main result is the following theorem.

Theorem 5.4.

(i) *Let T_n^U be the size of a uniform tree of $\mathscr{F}_n$. Then,*

$$\mathbb{P}\Big(T_n^U = k\Big) \xrightarrow[n\to\infty]{} 2\int_0^1 x e^{-x}(1-e^{-x})^{k-1}dx,$$

that is, $T_n^U \xrightarrow{d} T^U$ where $T^U \sim \mathrm{Geometric}(e^{-X})$, and $X \sim 2x dx$ on $[0,1]$.

(ii) *Let $T_n^{(1)}$ be the size of the tree containing vertex 1 in $\mathscr{F}_n$. Then,*

$$\mathbb{P}\Big(T_n^{(1)} = k\Big) \xrightarrow[n\to\infty]{} k\int_0^1 x e^{-x}(1-e^{-x})^{k-1}dx,$$

that is, $T_n^{(1)}$ converges in distribution to the size-biasing of T^U.

Remark 5.5. Note that even though the limit distribution of $T_n^{(1)}$ is the size-biased limit distribution of T_n^U, for finite n the distribution of $T_n^{(1)}$ is *not* the size-biased distribution of T_n^U. $\diamond$

We start by giving the distribution of $T_n^{(1)}$ in terms of the process Υ_n defined in Section 5.1. For this, we first need to introduce some notation. Let $\mathcal{T}_n^{(v)}$ be the tree containing vertex v in $\mathscr{F}_n$. We denote by $H_n^{(v)}$ the number of steps after the root of $\mathcal{T}_n^{(v)}$ was added in the UA construction. Recalling the notation from Section 2.2, where $\sigma^{-1}(v) \in \{1, \dots, n\}$ denotes the step of the UA construction at which vertex v was added, we thus have

$$H_n^{(v)} \;=\; n - \min\Big\{\sigma^{-1}(u) : u \in \mathscr{T}_n^{(v)}\Big\}\,.$$

Proposition 5.6. *Let $T_n^{(1)}$ be the size of the tree containing vertex 1 in $\mathscr{F}_n$, and denote by $H_n^{(1)}$ the number of steps after the root of that tree was added in the UA construction. Then,*

(i) *For $0 \leqslant h \leqslant n-1$, $\mathbb{P}(H_n^{(1)} = h) = \dfrac{h}{n(n-1)}\Big(1 + \dfrac{1}{n-1}\Big)^h$.*

(ii) *Conditional on $\{H_n^{(1)} = h\}$, $T_n^{(1)}$ is distributed as the size-biasing of $\Upsilon_n(h)$.*

Remark 5.7. The size-biasing of $\Upsilon_n(h)$ can be easily represented as follows. Consider the Markov chain $\Upsilon_n^* = (\Upsilon_n^*(\ell),\, 0 \leqslant \ell \leqslant n-1)$ defined by $\Upsilon_n^*(0) = 1$ and the following transition probabilities:

$$\mathbb{P}\Big(\Upsilon_n^*(\ell+1) = j \;\Big|\; \Upsilon_n^*(\ell) = i\Big) = \begin{cases} \frac{i+1}{n} & \text{if } j = i+1 \\ 1 - \frac{i+1}{n} & \text{if } j = i. \end{cases}$$

A straightforward induction on ℓ shows that $\Upsilon_n^*(\ell)$ is distributed as the size-biasing of $\Upsilon_n(\ell)$. $\diamond$

Proof. First, note that $H_n^{(1)} = h$ if and only if a new tree is created at step $n-h$, and vertex 1 belongs to this tree. Now, the probability that a new tree is created at step $n-h$ is $\frac{h}{n-1}$, and the size of this tree is then distributed as $\Upsilon_n(h)$. Moreover, at the end of the UA construction the labels are assigned to the vertices uniformly. As a result, conditional on a tree having size i, the probability that it contains vertex 1 is i/n. We thus have

$$\mathbb{P}\Big(H_n^{(1)} = h,\, T_n^{(1)} = i\Big) \;=\; \frac{h}{n-1} \cdot \frac{i}{n}\, \mathbb{P}(\Upsilon_n(h) = i).$$

Summing over i and using Lemma 5.1 yields

$$\mathbb{P}\Big(H_n^{(1)} = h\Big) \;=\; \frac{h}{n(n-1)}\Big(1 + \frac{1}{n-1}\Big)^h.$$

Finally,

$$\mathbb{P}\Big(T_n^{(1)} = i \;\Big|\; H_n^{(1)} = h\Big) = i\,\mathbb{P}\Big(\Upsilon_n(h) = i\Big)\Big(1 + \frac{1}{n-1}\Big)^{-h},$$

which concludes the proof. $\square$

We can now turn to the proof of our main result.

Proof of Theorem 5.4. (i) First recall the notation of the UA construction and Section 4, and note that conditional on the event

$$\{I_{n-h} = 0\} = \{\text{a new tree is created at step } n-h \text{ of the UA construction}\},$$

the total number of trees has distribution

$$(N_n \mid I_{n-h} = 0) \;\overset{d}{=}\; 1 + \sum_{\substack{\ell=1 \\ \ell \neq n-h}}^{n} (1 - I_\ell),$$

where $I_\ell \sim \mathrm{Ber}(\frac{\ell-1}{n-1})$ are independent random variables. From this, it is clear that uniformly in h,

$$\mathbb{E}(N_n \mid I_{n-h} = 0) \sim \frac{n}{2}\,. \tag{3}$$

On the event $\{I_{n-h} = 0\}$, let us denote by $\mathcal{T}_{n,h}$ the size of the tree created at step $n-h$. Note that the marginal distribution of $\mathcal{T}_{n,h}$ is simply $\Upsilon_n(h)$. Let us now compute

$$\begin{aligned}\mathbb{P}\Big(T_n^U = k,\, H_n^U = h\Big) &= \frac{h}{n-1}\mathbb{P}\Big(\mathcal{T}_{n,h} = k,\, H_n^U = h \,\Big|\, I_{n-h} = 0\Big)\\ &= \frac{h}{n-1}\mathbb{E}\Big(\frac{1}{N_n}\mathbb{1}_{\{\mathcal{T}_{n,h}=k\}} \,\Big|\, I_{n-h} = 0\Big),\end{aligned}$$

and note that

$$\begin{aligned}\frac{n}{2}\,\mathbb{E}\Big(\frac{1}{N_n}\mathbb{1}_{\{\mathcal{T}_{n,h}=k\}} \,\Big|\, I_{n-h} = 0\Big) \;&=\; \mathbb{P}(\Upsilon_n(h) = k)\\ &\quad+\; \mathbb{E}\Big(\Big(\frac{n/2}{N_n} - 1\Big)\mathbb{1}_{\{\mathcal{T}_{n,h}=k\}} \,\Big|\, I_{n-h} = 0\Big).\end{aligned} \tag{4}$$

The last term in this display goes to zero as $n \to \infty$, uniformly in h. Indeed, using (3) and applying Hoeffding's inequality [13] to N_n, which is a sum of n independent Bernoulli random variables, we get, for $\varepsilon > 0$ and uniformly in h,

$$\mathbb{P}\Big(\Big|\tfrac{N_n}{n/2} - 1\Big| > \varepsilon \,\Big|\, I_{n-h} = 0\Big) \;\leqslant\; 2e^{-Cn},$$

where C is a positive constant that depends only on ε. Using that for any ε such that $0 < \varepsilon < 1/2$ and any $x > 0$, $|\frac{1}{x} - 1| > 2\varepsilon \implies |x-1| > \varepsilon$, we may bound

$$\begin{aligned}\mathbb{E}\Big(\Big|\tfrac{n/2}{N_n} - 1\Big| \,\Big|\, I_{n-h} = 0\Big) \;&\leqslant\; 2\varepsilon \,+\, \frac{n}{2}\mathbb{P}\Big(\Big|\tfrac{n/2}{N_n} - 1\Big| > 2\varepsilon \,\Big|\, I_{n-h} = 0\Big)\\ &\leqslant\; 2\varepsilon \,+\, \frac{n}{2}\mathbb{P}\Big(\Big|\tfrac{N_n}{n/2} - 1\Big| > \varepsilon \,\Big|\, I_{n-h} = 0\Big)\\ &\leqslant\; 2\varepsilon \,+\, ne^{-Cn}.\end{aligned}$$

This shows that the last term in (4) goes to zero uniformly in h. We thus get

$$\mathbb{P}\Big(T_n^U = k,\, H_n^U = h\Big) \;=\; \frac{2h}{n^2}\Big(\mathbb{P}(\Upsilon_n(h) = k) + o(1)\Big),$$

and so summing over h and using Lemma 5.2 yields

$$\begin{aligned}\mathbb{P}\Big(T_n^U = k\Big) \;&=\; \frac{2}{n}\sum_{h=0}^{n-1}\frac{h}{n}\mathbb{P}(\Upsilon_n(h) = k) + o(1)\\ &\xrightarrow[n\to\infty]{}\; 2\int_0^1 x\mathbb{P}(Y(x) = k)\,dx.\end{aligned}$$

Recalling the well-known fact that $Y(x)$ has a Geometric(e^{-x}) distribution (see for instance Section 5.3 in [18]) proves the first point.

(ii) We know from Proposition 5.6 that

$$\begin{aligned}\mathbb{P}\Big(T_n^{(1)} = k\Big) &= \frac{1}{n}\sum_{h=0}^{n-1}\frac{h}{n-1}\mathbb{E}\Big(\Upsilon_n(h)\mathbb{1}_{\{\Upsilon_n(h)=k\}}\Big)\\ &= \frac{k}{n}\sum_{h=0}^{n-1}\frac{h}{n-1}\mathbb{P}(\Upsilon_n(h) = k).\end{aligned}$$

Again, using Lemma 5.2 and dominated convergence, we have

$$\frac{k}{n}\sum_{h=0}^{n-1}\frac{h}{n-1}\mathbb{P}(\Upsilon_n(h)=k) \xrightarrow[n\to\infty]{} k\int_0^1 x\mathbb{P}(Y(x)=k)dx,$$

which yields the result. □

5.3 Size of the largest tree

The goal of this section is to derive asymptotics for $T_n^{\max} := \max_v T_n^{(v)}$, the size of the largest tree in the Moran forest on n vertices, when $n \to \infty$.

Theorem 5.8. *Let $T_n^{\max}$ denote the size of the largest tree in $\mathscr{F}_n$. Then*

$$T_n^{\max} = \alpha\Big(\log n - (1+o_{\mathrm{p}}(1))\log\log n\Big),$$

where $\alpha = (1-\log(e-1))^{-1} \approx 2.18019$ and $o_{\mathrm{p}}(1)$ denotes a sequence of random variables that goes to 0 in probability.

As in Section 5.1, for any vertex v let us define $\tilde{\mathscr{T}}_n^{(v)} \subset \mathscr{T}_n^{(v)}$ as the subtree descending from v in the UA construction. For our purpose, it will be sufficient to study the size $\tilde{T}_n^{(v)} := |\tilde{\mathscr{T}}_n^{(v)}|$ of those subtrees instead of that of the trees $\mathscr{T}_n^{(v)}$. Indeed, observe that

$$T_n^{\max} = \max_v \tilde{T}_n^{(v)},$$

so that applying Lemma 4.6 with $m_n = \alpha\log\log n$ and $\beta = -1$ to the exchangeable variables $(\tilde{T}_n^{(1)} - \alpha\log n, \ldots, \tilde{T}_n^{(n)} - \alpha\log n)$ will prove the theorem. Again, we omit the superscript and denote by $\tilde{T}_n$ a random variable with distribution equal to that of $\tilde{T}_n^{(1)}$.

For the rest of the section, we thus study the tail probabilities of the variable $\tilde{T}_n$. Recall from the UA construction that the number L of steps after a fixed vertex was added is uniformly distributed on $\{0,\ldots,n-1\}$, and from Section 5.1 that, conditional on $\{L=\ell\}$,

$$\tilde{T}_n \overset{d}{=} \Upsilon_n(\ell)\,.$$

Proposition 5.9. *For any sequence of integers $k_n \to \infty$ with $k_n = o(\sqrt{n})$,*

$$\mathbb{P}\Big(\tilde{T}_n > k_n\Big) \sim \frac{e}{k_n}(1-e^{-1})^{k_n+1}.$$

Proof. Using the upper bound in Lemma 5.3 and the fact that L is uniform on $\{0,\ldots,n-1\}$, we have

$$\begin{aligned}
\mathbb{P}\Big(\tilde{T}_n > k_n\Big) &\leqslant \frac{1}{n}\sum_{\ell=0}^{n-1}\mathbb{P}\Big(Y\Big(\lambda_n(k)\tfrac{\ell}{n-1}\Big) > k_n\Big) \\
&= \frac{n-1}{n}\int_0^1 \mathbb{P}\Big(Y\Big(\lambda_n(k_n)\tfrac{\lfloor x(n-1)\rfloor}{n-1}\Big) > k_n\Big)\,dx \;+\; \frac{1}{n}\mathbb{P}(Y(\lambda_n(k_n)) > k_n) \\
&\leqslant \int_0^1 \mathbb{P}(Y(\lambda_n(k_n)x) > k_n)\,dx \;+\; \frac{1}{n}(1-e^{-\lambda_n(k_n)})^{k_n} \\
&= \int_0^1 \Big(1-e^{-\lambda_n(k_n)x}\Big)^{k_n}dx \;+\; \frac{1}{n}(1-e^{-\lambda_n(k_n)})^{k_n}.
\end{aligned}$$

Now recall that $\lambda_n(k_n) = -\frac{n-1}{k_n}\log\Big(1-\frac{k_n}{n-1}\Big) = 1+O(\frac{k_n}{n})$, so uniformly in $x \in [0,1]$,

$$e^{-\lambda_n(k_n)x} = e^{-x} + O\Big(\frac{k_n}{n}\Big).$$

Since $k_n = o(\sqrt{n})$, we have $k_n/n = o(1/k_n)$ and thus Lemma A.1 from the Appendix gives

$$\int_0^1 \Big(1 - e^{-\lambda_n(k_n)x}\Big)^{k_n} dx \sim \frac{e}{k_n}(1-e^{-1})^{k_n+1}.$$

Elementary calculations show that when $k_n = o(\sqrt{n})$, we also have

$$\frac{1}{n}(1-e^{-\lambda_n(k_n)})^{k_n} \sim \frac{1}{n}(1-e^{-1})^{k_n} = o\Big(\frac{(1-e^{-1})^{k_n}}{k_n}\Big).$$

It remains to examine the lower bound in Lemma 5.3. As above, we get an integral

$$\begin{aligned}
\mathbb{P}\Big(\tilde{T}_n > k_n\Big) &\geqslant \frac{1}{n}\sum_{\ell=0}^{n-1} \mathbb{P}\Big(Y\Big(\tfrac{\ell-k_n+1}{n-1}\Big) > k_n\Big)\\
&\geqslant \frac{n-1}{n}\int_0^1 \mathbb{P}\Big(Y\Big(\tfrac{\lceil x(n-1)\rceil - k_n}{n-1}\Big) > k_n\Big)\, dx\\
&\geqslant \frac{n-1}{n}\int_0^1 \mathbb{P}\Big(Y\Big(x - \tfrac{k_n}{n-1}\Big) > k_n\Big)\, dx.
\end{aligned}$$

Since

$$\mathbb{P}\Big(Y\Big(x-\tfrac{k_n}{n-1}\Big) > k_n\Big) = \Big(1-\exp(-x+\tfrac{k_n}{n-1})\Big)^{k_n} = \Big(1-e^{-x}+O(k_n/n)\Big)^{k_n},$$

using Lemma A.1 again, we get

$$\mathbb{P}\Big(\tilde{T}_n > k_n\Big) \geqslant \frac{n-1}{n}\int_0^1 \mathbb{P}\Big(Y\Big(x-\tfrac{k_n}{n-1}\Big) > k_n\Big)\, dx \sim \frac{e}{k_n}(1-e^{-1})^{k_n+1},$$

which completes the proof. □

Note that if k_n is not integer-valued, then

$$\mathbb{P}\Big(\tilde{T}_n > k_n\Big) = \mathbb{P}\Big(\tilde{T}_n > \lfloor k_n \rfloor\Big) \sim \frac{e}{k_n}(1-e^{-1})^{\lfloor k_n\rfloor+1},$$

which is not necessarily equivalent to $\frac{e}{k_n}(1-e^{-1})^{k_n+1}$ since $k_n - \lfloor k_n \rfloor$ may oscillate between 0 and 1. However, we do have $\mathbb{P}(\tilde{T}_n > k_n) = \Theta((1-e^{-1})^{k_n}/k_n)$, where the Bachmann–Landau notation $u_n = \Theta(v_n)$ indicates that there exist two positive constants c and C such that $c\,v_n \leqslant u_n \leqslant C v_n$ for n large enough. This approximation is sufficient for our purpose.

We may now prove Theorem 5.8 using the first and second moment method that we already used for the largest degree.

Proof of Theorem 5.8. We apply Lemma 4.6 to the exchangeable variables

$$(X_n^{(1)}, \ldots, X_n^{(n)}) = (\tilde{T}_n^{(1)} - \alpha\log n, \ldots, \tilde{T}_n^{(n)} - \alpha\log n)\,,$$

with $m_n = \alpha\log\log n$ and $\beta = -1$. The first two points of the lemma are readily checked, since Proposition 5.9 tells us that for

$$\alpha = (1-\log(e-1))^{-1} = -(\log(1-e^{-1}))^{-1}$$

and any $\gamma > 0$, we have for $k_n := \alpha(\log n - \gamma \log\log n)$

$$\mathbb{P}\Big(\tilde{T}_n - \alpha \log n \geqslant -\gamma\alpha \log\log n\Big) = \mathbb{P}\Big(\tilde{T}_n \geqslant k_n\Big) = \Theta\Big(\frac{(\log n)^{\gamma-1}}{n}\Big). \qquad (5)$$

Thus, for all $\varepsilon > 0$,

(i) $n\mathbb{P}\Big(\tilde{T}_n - \alpha \log n \geqslant (-1+\varepsilon)\alpha \log\log n\Big) \to 0.$

(ii) $n\mathbb{P}\Big(\tilde{T}_n - \alpha \log n \geqslant (-1-\varepsilon)\alpha \log\log n\Big) \to +\infty.$

All that remains to check is the third point of the lemma. From now on, we fix $k_n = \alpha(\log n - (1+\varepsilon)\log\log n)$ for some $\varepsilon > 0$, and for the sake of readability, we set $R_n := \mathbb{P}(\tilde{T}_n \geqslant k_n)$. With this notation, given Remark 4.7 we need to show

$$\mathbb{P}\Big(\tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\Big) \;\leqslant\; R_n^2 + o(R_n^2). \qquad (6)$$

Since this is rather technical, we defer the complete proof to Lemma A.2 in Appendix A.2, and only outline the main ideas of the proof here. As in the study of the largest degree, we prove this by showing that the law of $\tilde{T}_n^{(2)}$ conditional on $\{\tilde{T}_n^{(1)} \geqslant k_n\}$ is close to its unconditional law. We first prove that

$$\mathbb{P}\Big(A_n,\, \tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\Big) \;=\; o(R_n^2),$$

where $A_n := \{\tilde{\mathscr{T}}_n^{(2)} \subset \tilde{\mathscr{T}}_n^{(1)}\} \sqcup \{\tilde{\mathscr{T}}_n^{(1)} \subset \tilde{\mathscr{T}}_n^{(2)}\}$ is the event that one of the two vertices 1 and 2 is an ancestor of the other in the UA construction. We then show

$$\mathbb{P}\Big(A_n^{\mathrm{c}},\, \tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\Big) \;\leqslant\; R_n^2 + o(R_n^2)\,,$$

where A_n^{c} denotes the complement of A_n. This is done by showing that, conditional on $\{\tilde{T}_n^{(1)} = i\}$, on the event A_n^{c} the process counting the number of vertices of the tree $\mathscr{T}_n^{(2)}$ in the UA construction behaves as a modified Υ_n process, which we essentially bound from above by Υ_{n-i}. Therefore, $\tilde{T}_n^{(2)}$ can be compared with an independent variable with distribution $\tilde{T}_{n-i}$. Finally, we show that

$$\sum_{i \geqslant k_n} \mathbb{P}\Big(\tilde{T}_n^{(1)} = i\Big)\mathbb{P}\Big(\tilde{T}_{n-i} \geqslant k_n\Big) \;\leqslant\; R_n^2 + o(R_n^2),$$

thereby proving (6) and concluding the proof of Theorem 5.8. □

Acknowledgments

We thank Amaury Lambert for initiating the discussion that led us to consider this model and for comments on the first version of this manuscript.

References

[1] The On-Line Encyclopedia of Integer Sequences, published electronically at http://oeis.org, 2019.

[2] M. Aigner and G. M. Ziegler. *Proofs from* THE BOOK. Springer-Verlag Berlin, 6th edition, 2018. doi:10.1007/978-3-642-00856-6.

[3] K. B. Athreya and S. N. Lahiri. *Measure Theory and Probability Theory*. Springer Science+Business Media, 2006. doi:10.1007/978-0-387-35434-7.

[4] K. T. Balińska, L. V. Quintas, and J. Szymański. Random recursive forests. *Random Structures & Algorithms*, 5(1):3–12, 1994. doi:10.1002/rsa.3240050103.

[5] F. Bergeron, P. Flajolet, and B. Salvy. Varieties of increasing trees. In *CAAP '92*, pages 24–48, Berlin, Heidelberg, 1992. Springer Berlin Heidelberg. doi:10.1007/3-540-55251-0_2.

[6] P. Billingsley. *Convergence of Probability Measures*. Wiley, 2nd edition, 1999.

[7] B. Drake. *An Inversion Theorem for Labeled Trees and Some Limits of Areas Under Lattice Paths*. PhD thesis, Brandeis University, 2008.

[8] M. Drmota. *Random Trees: An Interplay Between Combinatorics and Probability*. Springer-Verlag Vienna, 1st edition, 2009. doi:10.1007/978-3-211-75357-6.

[9] R. Durrett. *Probability Models for DNA Sequence Evolution*. Springer-Verlag New York, 2nd edition, 2008. doi:10.1007/978-1-4757-6285-3.

[10] Ö. Eğecioğlu and J. B. Remmel. Bijections for Cayley trees, spanning trees, and their q-analogues. *Journal of Combinatorial Theory, Series A*, 42(1): 15–30, 1986. doi:10.1016/0097-3165(86)90004-X.

[11] A. Etheridge. *Some Mathematical Models From Population Genetics. École d'Été de Probabilités de Saint-Flour XXXIX-2009*, volume 2012. Springer Science+Business Media, 2011. doi:10.1007/978-3-642-16632-7.

[12] I. M. Gessel and S. Seo. A refinement of Cayley's formula for trees. *The Electronic Journal of Combinatorics*, 11(2):R27, 2006.

[13] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, Mar. 1963. doi:10.1080/01621459.1963.10500830.

[14] A. Joyal. Une théorie combinatoire des séries formelles. *Advances in Mathematics*, 42(1):1–82, 1981. doi:10.1016/0001-8708(81)90052-9.

[15] H. M. Mahmoud and R. T. Smythe. On the distribution of leaves in rooted subtrees of recursive trees. *The Annals of Applied Probability*, 1(3):406–418, 1991. doi:10.1214/aoap/1177005874.

[16] A. Meir and J. W. Moon. Cutting down recursive trees. *Mathematical Biosciences*, 21(3):173–181, 1974. doi:10.1016/0025-5564(74)90013-3.

[17] P. A. P. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(1):60–71, 1958. doi:10.1017/S0305004100033193.

[18] S. M. Ross. *Stochastic Processes*. Wiley, 2nd edition, 1995.

A Appendix

A.1 Proof of point (ii) of Proposition 4.4

We want to prove that, for all $K_n = o(\sqrt{n})$, there exists $\varepsilon_n = o(1)$ such that, for all $k \leqslant K_n$,

$$|\mathbb{P}(D_n \geqslant k) - \mathbb{P}(D \geqslant k)| \;\leqslant\; \varepsilon_n \, \mathbb{P}(D \geqslant k)\,.$$

Doing this directly from the expressions of D_n and D involves unappealing calculations. To somewhat circumvent this, we make use of the simple expressions of the probability generating functions G_{D_n} and G_D. For this, let

$$\Delta_n(z) \;\coloneqq\; \sum_{i \geqslant 0} \Big(\mathbb{P}(D \geqslant i) - \mathbb{P}(D_n \geqslant i)\Big) z^i,$$

so that the k-th derivative of Δ_n evaluated at $z = 0$ is

$$\Delta_n^{(k)}(0) \;=\; k! \Big(\mathbb{P}(D \geqslant k) - \mathbb{P}(D_n \geqslant k)\Big).$$

Since $\mathbb{P}(D \geqslant k) \geqslant \frac{2/e}{(k+1)!}$, we have to show that for any given sequence $K_n = o(\sqrt{n})$,

$$\Delta_n^{(k)}(0) \;=\; \frac{\varepsilon_n}{k+1}$$

for some $\varepsilon_n \to 0$ and all $k \leqslant K_n$. Now, since for any non-negative integer-valued random variable X,

$$\sum_{i \geqslant 0} \mathbb{P}(X \geqslant i)\, z^i \;=\; \frac{z\, \mathbb{E}\big(z^X\big) - 1}{z - 1}\,,$$

we can express Δ_n in terms of the generating functions of D and D_n, that is,

$$\Delta_n(z) \;=\; \left(1 + \frac{1}{z-1}\right)\Big(G_D(z) - G_{D_n}(z)\Big)\,.$$

Moreover, we know from Proposition 4.3 that

$$G_D(z) = \; 2\,\frac{e^{z-1} - 1}{z - 1} - 1 \;=\; 2 \sum_{i \geqslant 0} \frac{(z-1)^i}{(i+1)!} - 1$$

and from Proposition 4.1 that

$$\begin{aligned} G_{D_n}(z) &= \; 2\left(1 - \frac{1}{n}\right) \frac{\left(1 + \frac{z-1}{n-1}\right)^n - 1}{z - 1} \;-\; 1 \\ &= \; 2\left(1 - \frac{1}{n}\right) \sum_{i=0}^{n-1} \binom{n}{i+1} \left(\frac{1}{n-1}\right)^{i+1} (z-1)^i \;-\; 1 \\ &= \; 2 \sum_{i=0}^{n-1} \left(\prod_{\ell=1}^{i} \frac{n - \ell}{n - 1}\right) \frac{(z-1)^i}{(i+1)!} \;-\; 1\,, \end{aligned}$$

where the empty product is 1. Therefore,

$$G_D(z) - G_{D_n}(z) = \sum_{i \geqslant 0} A(n,i) \frac{(z-1)^i}{(i+1)!}\,,$$

where

$$A(n,i) = 2\left[1 - \left(\prod_{\ell=1}^{i} \frac{n-\ell}{n-1}\right)\mathbb{1}_{\{i \leqslant n-1\}}\right].$$

Using that $A(n,0) = A(n,1) = 0$ and rearranging a bit, we obtain the following expansion of Δ_n at $z = 1$:

$$\Delta_n(z) \;=\; \sum_{i \geqslant 1}\left(A(n,i) + \frac{A(n,i+1)}{i+2}\right)\frac{(z-1)^i}{(i+1)!}\,,$$

from which we get

$$\Delta_n^{(k)}(0) \;=\; \sum_{i \geqslant k}\left(A(n,i) + \frac{A(n,i+1)}{i+2}\right)\frac{(-1)^{i-k}}{(i-k)!\,(i+1)}\,.$$

Now, pick any $J_n = o(\sqrt{n})$ such that $K_n = o(J_n)$. For all $i < J_n$,

$$\left|A(n,i) + \frac{A(n,i+1)}{i+2}\right| \;\leqslant\; 4\left(1 - \prod_{\ell=1}^{J_n} \frac{n-\ell}{n-1}\right) \;=\; \varepsilon_n\,,$$

with $\varepsilon_n \to 0$, since

$$\prod_{\ell=1}^{J_n} \frac{n-\ell}{n-1} \;\geqslant\; \left(\frac{n-J_n}{n-1}\right)^{J_n} \;=\; \exp\!\left(-\tfrac{J_n^2}{n} + o\!\left(\tfrac{J_n^2}{n}\right)\right).$$

For $i \geqslant J_n$, we have

$$\left|A(n,i) + \frac{A(n,i+1)}{i+2}\right| \;\leqslant\; 4\,.$$

Combining these two upper bounds, we get

$$\begin{aligned}\left|\Delta_n^{(k)}(0)\right| \;&\leqslant\; \sum_{i=k}^{J_n-1} \frac{\varepsilon_n}{(i-k)!\,(i+1)} \;+\; \sum_{i \geqslant J_n} \frac{4}{(i-k)!\,(i+1)}\\ &\leqslant\; \frac{\varepsilon_n\,C_1}{(k+1)} \;+\; \frac{C_2}{(J_n+1)}\,.\end{aligned}$$

Finally, since $K_n = o(J_n)$, we have for all $k \leqslant K_n$,

$$\frac{1}{J_n+1} \;\leqslant\; \frac{1}{k+1}\cdot\frac{K_n+1}{J_n+1}\,,$$

with $(K_n+1)/(J_n+1) = o(1)$. This concludes the proof.

Note that although we have been quite crude in that we have used the triangle inequality on an alternating series, a more careful analysis would show that the $o(\sqrt{n})$ requirement on K_n is in fact optimal.

A.2 Technical lemmas used in the proof of Theorem 5.8

Lemma A.1. *For any sequence $k_n \to \infty$ and any sequence of measurable maps $f_n : [0,1] \to \mathbb{R}$ such that for all $x \in [0,1]$, $(1 - e^{-x} + f_n(x)) \geqslant 0$ and $\sup_x |f_n(x)| = o(1/k_n)$, we have*

$$\int_0^1 \left(1 - e^{-x} + f_n(x)\right)^{k_n} dx \;\sim\; \frac{e}{k_n}(1 - e^{-1})^{k_n+1}.$$

Proof. Let us compute

$$\int_0^1 \frac{(1-e^{-x}+f_n(x))^{k_n}}{(1-e^{-1})^{k_n}} k_n\,dx = \int_0^1 \left(1-\frac{e^{1-x}-1}{e-1}+\frac{e}{e-1}f_n(x)\right)^{k_n} k_n\,dx$$
$$= \int_0^{k_n} \left(1-\frac{y}{k_n}+g_n(y)\right)^{k_n} \frac{e-1}{1+(e-1)\frac{y}{k_n}}\,dy,$$

where we used the change of variable $y = k_n(e^{1-x}-1)(e-1)^{-1}$, and defined the map g_n as

$$g_n(y) = \frac{e}{e-1} f_n\left(1-\log\left(1+\frac{y}{k_n}(e-1)\right)\right).$$

Since $(1-\frac{y}{k_n}+g_n(y))^{k_n} \leqslant \exp(-y+\frac{e}{e-1}k_n \sup_x f_n(x))$, it follows from dominated convergence that

$$\int_0^1 \frac{(1-e^{-x}+f_n(x))^{k_n}}{(1-e^{-1})^{k_n}} k_n\,dx \xrightarrow[n\to\infty]{} \int_0^\infty e^{-y}(e-1)\,dy = e-1,$$

concluding the proof. $\square$

Lemma A.2. *Let $\tilde{T}_n^{(v)}$ denote the size of the subtree descending from v in the UA construction of $\mathscr{F}_n$. Then, for $\alpha = -1/\log(1-e^{-1})$ and any $\varepsilon > 0$, letting $k_n = \alpha(\log n - (1+\varepsilon)\log\log n)$ and $R_n = \mathbb{P}\left(\tilde{T}_n \geqslant k_n\right)$,*

$$\mathbb{P}\left(\tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\right) \leqslant R_n^2 + o(R_n^2).$$

Proof. Let us denote by $A_n := \{\tilde{\mathscr{T}}_n^{(2)} \subset \tilde{\mathscr{T}}_n^{(1)}\} \sqcup \{\tilde{\mathscr{T}}_n^{(1)} \subset \tilde{\mathscr{T}}_n^{(2)}\}$ the event that one of the vertices 1 and 2 is an ancestor of the other. We start by showing that

$$\mathbb{P}\left(A_n,\, \tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\right) = o(R_n^2). \tag{7}$$

By exchangeability, we have

$$\begin{aligned}
&\mathbb{P}\left(A_n,\, \tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\right) \\
&\quad = 2\,\mathbb{P}\left(\tilde{\mathscr{T}}_n^{(2)} \subset \tilde{\mathscr{T}}_n^{(1)},\, \tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\right) \\
&\quad = \sum_{i \geqslant k_n} \mathbb{P}\left(\tilde{\mathscr{T}}_n^{(2)} \subset \tilde{\mathscr{T}}_n^{(1)},\, \tilde{T}_n^{(2)} \geqslant k_n \,\middle|\, \tilde{T}_n^{(1)} = i\right) \mathbb{P}\left(\tilde{T}_n = i\right).
\end{aligned}$$

Let us call the *height* of a vertex the number of steps after it was added in the UA construction. Conditional on $\{\tilde{T}_n^{(1)} = i\}$ and on the heights of the vertices of $\tilde{\mathscr{T}}_n^{(1)}$ being $\ell_1 > \ldots > \ell_i$, the height L_2 of vertex 2 is uniformly distributed on $\{0, \ldots n-1\} \setminus \{\ell_1\}$. Moreover, in order to have

$$\{\tilde{\mathscr{T}}_n^{(2)} \subset \tilde{\mathscr{T}}_n^{(1)},\, \tilde{T}_n^{(2)} \geqslant k_n\},$$

the height of vertex 2 must belong to $\{\ell_2, \ldots, \ell_{i-(k_n-1)}\}$, which happens with probability $\frac{i-k_n}{n-1}$. Therefore,

$$\begin{aligned}
&\mathbb{P}\left(A_n,\, \tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\right) \\
&\quad \leqslant \sum_{i \geqslant k_n} \mathbb{P}\left(\tilde{T}_n = i\right) \frac{i-k_n}{n-1} \\
&\quad = \frac{1}{n-1} \sum_{i > k_n} \mathbb{P}\left(\tilde{T}_n \geqslant i\right).
\end{aligned}$$

To show that this is small enough, we let $K_n := k_n + \alpha(\log n)^\delta$ with $0 < \delta < \min(1, \varepsilon)$ and $K'_n := 2\alpha \log n$, and crudely bound

$$\sum_{i>k_n} \mathbb{P}\Big(\tilde{T}_n \geqslant i\Big) \leqslant (K_n - k_n)\mathbb{P}\Big(\tilde{T}_n \geqslant k_n\Big) + K'_n\,\mathbb{P}\Big(\tilde{T}_n \geqslant K_n\Big) + n\,\mathbb{P}\Big(\tilde{T}_n \geqslant K'_n\Big).$$

Now let us show that these three terms are negligible compared to nR_n^2. Recalling from (5) that $R_n = \Theta\Big(\frac{(\log n)^\varepsilon}{n}\Big)$, we have $nR_n^2 = \Theta((\log n)^{2\varepsilon}/n)$ and therefore

- $(K_n - k_n)\mathbb{P}\Big(\tilde{T}_n \geqslant k_n\Big) \sim \alpha(\log n)^\delta R_n = \Theta\Big(\dfrac{(\log n)^{\delta+\varepsilon}}{n}\Big) = o(nR_n^2)$.
- $K'_n\,\mathbb{P}\Big(\tilde{T}_n \geqslant K_n\Big) = \Theta\Big(\log n\, R_n e^{-(\log n)^\delta}\Big) = o(R_n) = o(nR_n^2)$.
- $n\,\mathbb{P}\Big(\tilde{T}_n \geqslant K'_n\Big) = \Theta\Big(n\dfrac{n^{-2}}{\log n}\Big) = o(1/n) = o(nR_n^2)$.

As a result, (7) is proven and it remains to show that

$$\mathbb{P}\Big(A_n^{\mathrm{c}},\, \tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\Big) \leqslant R_n^2 + o(R_n^2),$$

where A_n^{c} denotes the complement of A_n. We now fix $n \geqslant 1$, $i \geqslant k_n$, and a finite sequence $n-1 \geqslant \ell_1 > \ldots > \ell_i \geqslant 0$. Let us write B for the event that $\tilde{\mathscr{T}}_n^{(1)}$ contains exactly the vertices with heights $\ell_1 > \ldots > \ell_i$. Conditional on B, let us examine the distribution of $\tilde{\mathscr{T}}_n^{(2)}$. Recall that the height L_2 of vertex 2 is uniformly distributed on $\{0, \ldots n-1\} \setminus \{\ell_1\}$. In the UA construction, define $\mathbb{T}$ as the tree obtained by starting from a root arrived at height L_2 and allowing the attachment of a vertex with height ℓ to $\mathbb{T}$ only if $\ell \notin \{\ell_1, \ldots \ell_i\}$. Then, on the event A_n^{c}, this tree must coincide with $\mathscr{T}_n^{(2)}$, and so

$$\mathbb{P}\Big(A_n^{\mathrm{c}},\, \tilde{T}_n^{(2)} \geqslant k_n \,\Big|\, B\Big) \;=\; \mathbb{P}\Big(A_n^{\mathrm{c}},\, |\mathbb{T}| \geqslant k_n \,\Big|\, B\Big).$$

From the UA construction, for any $\ell \notin \{\ell_1, \ldots, \ell_i\}$, conditional on $B \cap \{L_2 = \ell\}$, we can describe $|\mathbb{T}|$ using the process $(\tilde{\Upsilon}_\ell(m),\, 0 \leqslant m \leqslant \ell)$ defined by

- $\tilde{\Upsilon}_\ell(0) = 1$.
- For all $0 < m \leqslant \ell$, $\tilde{\Upsilon}_\ell(m) - \tilde{\Upsilon}_\ell(m-1) \in \{0, 1\}$ and, conditional on $\{\tilde{\Upsilon}_\ell(m-1) = j\}$, $\tilde{\Upsilon}_\ell(m) = j+1$ with probability

$$\begin{cases} \dfrac{j}{n-1-J_m} & \text{if } \ell - m \notin \{\ell_1, \ldots, \ell_i\} \\ 0 & \text{if } \ell - m \in \{\ell_1, \ldots, \ell_i\}, \end{cases}$$

  where $J_m = |\{\ell_1, \ldots, \ell_i\} \cap \{\ell - m, \ldots, n\}|$ is the number of vertices of $\tilde{T}_n^{(1)}$ with height greater than $\ell - m$ in the UA construction.

With this definition, for any $\ell \notin \{\ell_1, \ldots, \ell_i\}$, conditional on $B \cap \{L_2 = \ell\}$, we have by construction $|\mathbb{T}| \overset{d}{=} \tilde{\Upsilon}_\ell(\ell)$. Now, note that the probability of increasing is always bounded by $j/(n-1-i)$. Therefore, $\tilde{\Upsilon}_\ell$ can be coupled with Υ_{n-i} in such a way that, for all $0 \leqslant m \leqslant \ell < n-i$,

$$\tilde{\Upsilon}_\ell(m) \leqslant \Upsilon_{n-i}(m).$$

For $\ell \geqslant n-i$, we use the crude bound $\mathbb{P}(\tilde{\Upsilon}_\ell(\ell) \geqslant k_n) \leqslant \mathbb{E}(\tilde{\Upsilon}_\ell(\ell))/k_n$. Using the same reasoning as in Lemma 5.1, we get

$$\mathbb{E}(\tilde{\Upsilon}_\ell(\ell)) \leqslant (1+\frac{1}{n-i-1})^{n-i-1} \leqslant e\,.$$

We thus have

$$\begin{aligned}
\mathbb{P}\Big(A_n^{\mathrm{c}},\, |\mathbb{T}| \geqslant k_n \,\Big|\, B\Big) &\leqslant \mathbb{P}\Big(L_2 \notin \{\ell_1,\dots,\ell_i\},\, |\mathbb{T}| \geqslant k_n \,\Big|\, B\Big) \\
&= \frac{1}{n-1} \sum_{\substack{\ell=0 \\ \ell\notin\{\ell_1,\dots\ell_i\}}}^{n-1} \mathbb{P}\Big(\tilde{\Upsilon}_\ell(\ell) \geqslant k_n\Big), \\
&\leqslant \frac{ei}{k_n(n-1)} + \frac{1}{n-1}\sum_{\ell=0}^{n-i-1} \mathbb{P}(\Upsilon_{n-i}(\ell) \geqslant k_n) \qquad (8) \\
&= \frac{ei}{k_n(n-1)} + \frac{n-i}{n-1}\mathbb{P}\Big(\tilde{T}_{n-i} \geqslant k_n\Big). \qquad (9)
\end{aligned}$$

Since this bound depends on the set $\{\ell_1,\dots,\ell_i\}$ only via its cardinality i, one can integrate with respect to the distribution of $\mathcal{T}_n^{(1)}$ to get

$$\mathbb{P}\Big(A_n^{\mathrm{c}},\, |\mathbb{T}| \geqslant k_n \,\Big|\, \tilde{T}_n^{(1)} = i\Big) \;\leqslant\; \frac{ei}{k_n(n-1)} + \frac{n-i}{n-1}\mathbb{P}\Big(\tilde{T}_{n-i} \geqslant k_n\Big).$$

Finally, because $\Upsilon_{n-(i+1)}(\ell) \overset{d}{\geqslant} \Upsilon_{n-i}(\ell)$, the expression (8)—and therefore (9)—is nondecreasing in i, and we have

$$\begin{aligned}
&\mathbb{P}\Big(A_n^{\mathrm{c}},\, \tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\Big) \\
&\quad\leqslant \sum_{i\geqslant k_n} \mathbb{P}\Big(\tilde{T}_n = i\Big)\left(\frac{ei}{k_n(n-1)} + \frac{n-i}{n-1}\mathbb{P}\Big(\tilde{T}_{n-i} \geqslant k_n\Big)\right) \\
&\quad\leqslant \sum_{i=k_n}^{K_n} \mathbb{P}\Big(\tilde{T}_n = i\Big)\left(\frac{e\,K_n}{k_n(n-1)} + \frac{n-K_n}{n-1}\mathbb{P}\Big(\tilde{T}_{n-K_n} \geqslant k_n\Big)\right) \qquad (10) \\
&\qquad + \sum_{i\geqslant K_n} \mathbb{P}\Big(\tilde{T}_n = i\Big)\left(\frac{ei}{k_n(n-1)} + \frac{n-i}{n-1}\mathbb{P}\Big(\tilde{T}_{n-i} \geqslant k_n\Big)\right), \qquad (11)
\end{aligned}$$

for any sequence $K_n \geqslant k_n$. Letting $K_n \coloneqq \alpha(\log n)^{1+\varepsilon/2}$, we then show that (10) is asymptotically no greater than R_n^2, and that (11) is negligible compared to R_n^2. Indeed, (10) is bounded from above by

$$R_n\left(\frac{e\,K_n}{k_n(n-1)} + \frac{n-K_n}{n-1}\mathbb{P}\Big(\tilde{T}_{n-K_n} \geqslant k_n\Big)\right).$$

Now note that $\frac{e\,K_n}{k_n(n-1)} = O(\frac{(\log n)^{\varepsilon/2}}{n}) = o(R_n)$, and that since $n-K_n \sim n$, we have $k_n = o(\sqrt{n-K_n})$. Therefore, by Proposition 5.9, $\mathbb{P}\Big(\tilde{T}_{n-K_n} \geqslant k_n\Big) \sim R_n$. Finally, up to a multiplicative constant, (11) is bounded from above by

$$\mathbb{P}\Big(\tilde{T}_n \geqslant K_n\Big) = \Theta\Big(\frac{n^{-(\log n)^{\varepsilon/2}}}{K_n}\Big) = o(n^{-2}) = o(R_n^2).$$

Putting everything together, we have proved that

$$\mathbb{P}\Big(\tilde{T}_n^{(1)} \geqslant k_n, \tilde{T}_n^{(2)} \geqslant k_n\Big) \;\leqslant\; R_n^2 + o(R_n^2),$$

which concludes the proof. $\square$