Uniform attachment with freezing

Etienne Bellin, Arthur Blanc-Renaudie, Emmanuel Kammerer, Igor Kortchemski

Uniform attachment with freezing

Étienne Bellin ♠ Arthur Blanc-Renaudie ♥ Emmanuel Kammerer ♦ Igor Kortchemski ♣

Abstract

In the classical model of random recursive trees, trees are recursively built by attaching new vertices to old ones. What happens if vertices are allowed to freeze, in the sense that new vertices cannot be attached to already frozen ones? We are interested in the impact of freezing on the height of such trees.

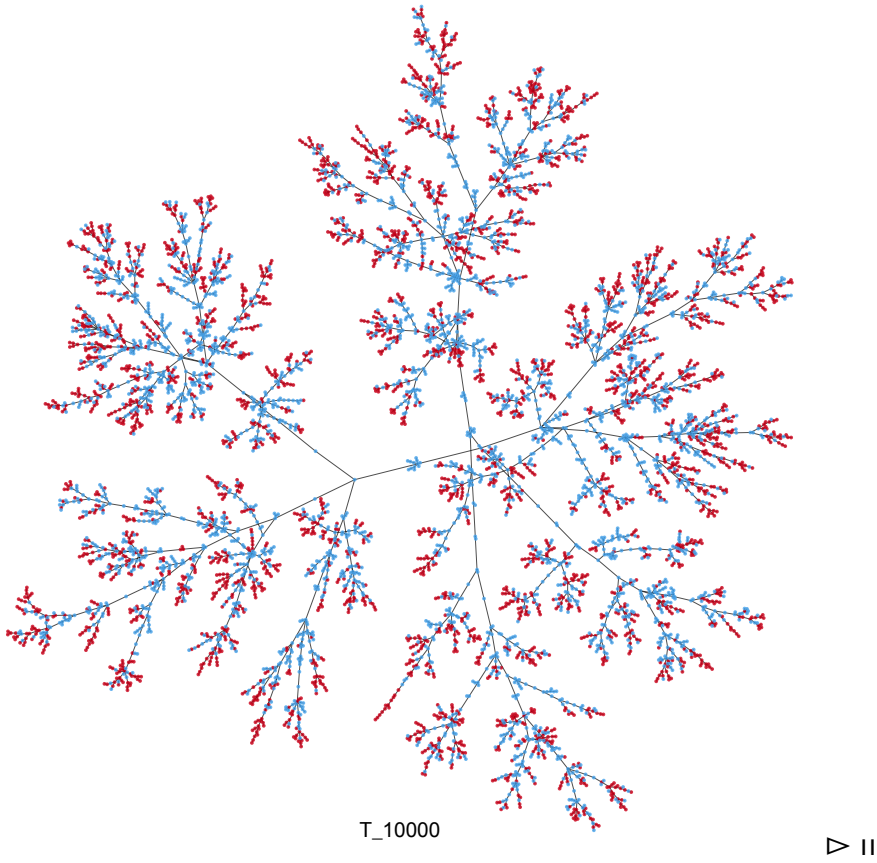

Figure 1: Simulation of a tree of size 10000 built by uniform attachment with freezing, when the number of active (i. e. non-frozen) vertices roughly evolves as a positive fraction of the total number of vertices. Frozen vertices are blue; active vertices are red. The animation (played with Acrobat Reader) shows the resulting process as the number of steps increases.

♠CMAP, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France, etienne.bellin@polytechnique.edu

♥Tel Aviv University, Israel. Supported by the ERC consolidator grant 101001124 (UniversalMap). ablancrenaudiepro@gmail.com

♦CMAP, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France, emmanuel.kammerer@polytechnique.edu

♣CNRS & CMAP, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France, igor.kortchemski@math.cnrs.fr

Contents

1 Introduction

Random graphs are instrumental in the study of real-world networks. Uniform recursive trees (sometimes also called uniform attachment trees) are one of such models. They are constructed recursively by starting with one single vertex, and successively attaching new vertices to a previous existing vertex, chosen uniformly at random. This model has been introduced in [31] in the context of system generation, and has received considerable interest since, starting with graph-theoretical properties such as the height, number of leaves, etc. (see e.g. [35] and references therein, as well as the survey [37]). Recursive trees have been proposed as models for the spread of epidemics [30], the family trees of preserved copies of ancient or medieval texts [32], pyramid schemes [19], internet interface maps [23] and appear in the study of Hopf algebras [21]. Very interesting connections have been made with other probabilistic objects such as the Bolthausen-Sznitman coalescent [20] and elephant random walks [27]. Uniform recursive trees have also been extended in several directions, for instance by introducing deterministic weights [26] or random weights with a random environment [10].

In the present work, we introduce and study a modification of this model by introducing *freezing*, in that existing vertices can freeze and new vertices cannot be attached to frozen vertices.

Our motivation is twofold. First, in the context of real-world networks such mechanisms are natural: for instance, on the social network Twitter a user can choose to set their account to "private" which prevents strangers from "following" them; also performing an infection-tracing of an SIR epidemics falls within this framework (see Section 6). Second, it is natural to investigate from a mathematical point of view the impact of freezing in dynamically-built random graph models. In particular, in a companion paper [7], we investigate the regime where the number of active vertices roughly evolves as the total number of vertices to the power α, for fixed $\alpha \in (0, 1]$, we describe a phase transition where the macroscopic geometry of our model drastically changes according to the value of α. Let us also mention the work [5], which considers a growth-fragmentation-isolation process on random recursive trees in the context of contact tracing, see also [8].

Uniform attachment with freezing. Let us first define our model. In order to incorporate freezing in the model of uniform attachment trees and to obtain results in a rather general setup, we shall fix beforehand the steps where an attachment takes place and the steps where freezing takes place. Specifically, our input is a deterministic sequence $\mathbf{x} = (\mathrm{x}_i)_{i\geq 1}$ of elements of $\{-1, +1\}$. Starting from a sole active vertex, we recursively build random trees by reading the elements of the sequence one after the other, by applying a "freezing" step when reading -1 (which amounts to freezing an active vertex chosen uniformly at random) and a "uniform attachment" step when reading $+1$ (which amounts to attaching a new vertex to an active vertex chosen uniformly at random).

More precisely, given a sequence $\mathbf{x} = (\mathrm{x}_i)_{i\geq 1} \in \{-1, +1\}^{\mathbb{N}}$, we set $S_0(\mathbf{x}) := 1$ and for every $n \geq 1$

$$S_n(\mathbf{x}) := 1 + \sum_{i=1}^{n} \mathrm{x}_i \qquad \text{then} \qquad \tau(\mathbf{x}) := \inf\{n \geq 1 : S_n(\mathbf{x}) = 0\}. \tag{1}$$

Observe that $\tau(\mathbf{x})$, if it is finite, is the first time when all the vertices are frozen, so that afterwards the tree does not evolve any more. For every $0 \leq n \leq \tau(\mathbf{x})$, let $\mathcal{T}_n(\mathbf{x})$ be the random tree recursively built in this fashion after reading the first n elements of $\mathbf{x}$ (see Algorithm 1 in Sec. 2.1 for a precise definition and Fig. 3 for an example).

Let us comment on our choice of parametrization. It would have been possible to define the model starting with a random sequence $\mathbf{x}$, but the choice of a deterministic sequence defines a more general model, for which our results can then be applied.

Local limits. When $\tau(\mathbf{x}) = \infty$, the next result allows to give a meaning to $\mathcal{T}_\infty(\mathbf{x})$ as a local limit of finite trees (see Sec. 3 for background on the local topology).

Theorem 1. *Let $(\mathbf{x}^n)_{n\geq 0}$ be a sequence of elements of $\{-1, 1\}^{\mathbb{N}}$. Suppose that there exists $\mathbf{x}$ such that $\tau(\mathbf{x}) = \infty$ and such that for all $i \geq 1$, $\mathbf{x}^n_i = \mathbf{x}_i$ for all n large enough. Then the following assertions are equivalent:*

(1) The sequence of trees $(\mathcal{T}_n(\mathbf{x}^n))_{n\geq 0}$ converges locally, in distribution.

(2) The sequence of trees $(\mathcal{T}_n(\mathbf{x}))_{n\geq 0}$ converges locally, almost surely, towards a tree $\mathcal{T}_\infty(\mathbf{x})$.

(3) The sum $\sum_{i\geq 1} \frac{1}{S_i(\mathbf{x})} \mathbb{1}_{\{\mathrm{x}_i=-1\}}$ diverges.

In this case $\mathcal{T}_\infty(\mathbf{x})$ is the local limit of $(\mathcal{T}_n(\mathbf{x}^n))_{n\geq 0}$.

Examples. It is interesting to note that this model encompasses the two classical models of random recursive trees and random uniform plane trees:

– when $x_i = 1$ for every $i \geq 1$, then $\mathcal{T}_n(\mathbf{x})$ is a random recursive tree on n vertices built by uniform attachment;
– when $\mathbf{X} = (X_i)_{i\geq 1}$ is a sequence of non constant i.i.d. uniform random variables on $\{-1, +1\}$, then for every $n \geq 1$, conditionally given $\tau(\mathbf{X}) = n$, $\mathcal{T}_n(\mathbf{X})$ is a uniform plane tree when the plane order among the vertices of $\mathcal{T}_n(\mathbf{X})$ is chosen uniformly at random.

More generally we have the following result, in which we consider the case of random sequences. In this case we will write $\mathbf{X}$ in upper case rather than $\mathbf{x}$ to emphasize the fact that the sequence is random. And, the law of $(\mathcal{T}_n(\mathbf{X}))_{n\geq 0}$ given the random sequence $\mathbf{X}$ follows the description above where $\mathbf{X}$ is considered to be fixed. In other words, we first choose $\mathbf{X}$ randomly, then, conditionally on $\mathbf{X}$, we construct $\mathcal{T}_n(\mathbf{X})$ following Algorithm 1 where we consider the sequence $\mathbf{X}$ to be deterministic.

Theorem 2. *Let $p \in [0, 1)$ and let $\mathbf{X} = (X_i)_{i\geq 1}$ be a sequence of i.i.d random variables such that $\mathbb{P}(X_1 = +1) = p$ and $\mathbb{P}(X_1 = -1) = 1 - p$. Then $\mathcal{T}_\infty(\mathbf{X})$ has the law of a Bienaymé tree with offspring distribution μ given by $\mu(k) = (1 - p)p^k$ for $k \geq 0$.*

Roughly speaking, this comes from the fact that the subtrees grafted on the initial vertex evolve in an i.i.d. fashion.

Height of uniform attachment trees with freezing. Since we consider large trees, let us consider for $n \geq 1$, a sequence $\mathbf{x}^n \in \{-1, 1\}^{\mathbb{N}}$ such that $\tau(\mathbf{x}^n) > n$, and let $\mathcal{T}_n := \mathcal{T}_n(\mathbf{x}^n)$. Also, to simplify notation, let $S_k^n := S_k(\mathbf{x}^n)$ for $0 \leq k \leq n$.

Our next main theorem is that the height of $\mathcal{T}_n$ is of order

$$\mathsf{h}_n^+ = \sum_{i=1}^{n} \frac{1}{S_i^n} \mathbb{1}_{\{x_i^n = 1\}}.$$

Theorem 3. *The following results hold.*

(1) Let U^n be an active vertex of $\mathcal{T}_n$ chosen uniformly at random, and denote by $\mathsf{H}(U^n)$ its height. Then for all $p \geq 1$:

$$\frac{\mathsf{H}(U^n)}{\mathsf{h}_n^+} \xrightarrow[n\to\infty]{\mathbb{L}^p} 1.$$

(2) For all $\varepsilon \in (0, 1)$:

$$\mathbb{P}\left(1 - \varepsilon \leq \frac{\mathsf{Height}(\mathcal{T}_n)}{\mathsf{h}_n^+} \leq e + \varepsilon\right) \xrightarrow[n\to\infty]{} 1.$$

(3) Assume that $\ln(n)/\mathsf{h}_n^+ \to 0$ as $n \to \infty$. Then for all $p \geq 1$:

$$\frac{\mathsf{Height}(\mathcal{T}_n)}{\mathsf{h}_n^+} \xrightarrow[n\to\infty]{\mathbb{L}^p} 1.$$

The proof of Theorem 3 is based on an alternative construction of $\mathcal{T}_n$, based on time-reversal, through a growth-coalescent process of rooted forests. This construction can be roughly described as follows (see Algorithm 2 for a precise definition): start with a forest made of S_n^n rooted one-vertex trees. Read successively $x_n^n, x_{n-1}^n, \ldots, x_1^n$; when reading +1 add a new one-vertex tree to the forest; when reading -1, choose successively two different forests uniformly at random, connect their roots by an edge, and root this new tree at the root of the first forest. It turns out that, in the end, one gets a tree having the same distribution as $\mathcal{T}^n$ (see Theorem 8 for a precise statement). This is a generalization of the connection between random recursive trees and Kingman's coalescent which first appeared in [17] (see [16, Sec. 6], [35, Sec. 3], [1, Sec. 2.2], [3] for applications), in the context of union-find data structures (which are data structures that

store a collection of disjoint sets where merging sets and finding a representative member of a set), see e.g. [25]. The main difference is that a growth feature must be added when freezing is introduced.

We are next interested in the specific regime where roughly speaking the number of active vertices increases linearly compared to the size of the tree.

Theorem 4. *Let $c \in (0,1]$. Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of positive numbers such that $A_n = o(\log n)$ as $n \to \infty$. Assume that*

$$\lim_{\varepsilon\to 0} \limsup_{n\to\infty} \max_{A_n \le i \le \varepsilon n} \left| \frac{S_i^n}{i} - c \right| = 0 \qquad \text{and} \qquad \forall \varepsilon \in (0,1),\ \liminf_{n\to\infty} \min_{\varepsilon n \le i \le n} \frac{S_i^n}{n} > 0. \tag{2}$$

The following assertions hold.

(1) We have $\dfrac{\mathsf{h}_n^+}{\ln n} \xrightarrow[n\to\infty]{} \dfrac{c+1}{2c}$.

(2) For all $n \ge 1$, conditionally given $\mathcal{T}_n$, let V_1^n and V_2^n be independent uniform vertices of $\mathcal{T}_n$. Then for every $p \ge 1$

$$\frac{\mathsf{H}(V_1^n)}{\ln n} \xrightarrow[n\to\infty]{\mathbb{L}^p} \frac{c+1}{2c} \qquad \text{and} \qquad \frac{d^n(V_1^n, V_2^n)}{\ln n} \xrightarrow[n\to\infty]{\mathbb{L}^p} \frac{c+1}{c}, \tag{3}$$

where $\mathsf{H}(V_1^n)$ is the height of V_1^n in $\mathcal{T}_n$ and d^n denotes the graph distance in $\mathcal{T}_n$.

(3) We have

$$\frac{\mathsf{Height}(\mathcal{T}_n)}{\ln(n)} \xrightarrow[n\to\infty]{\mathbb{P}} \frac{1+c}{2c} f(c),$$

where $f(c)$ is the unique solution with $f(c) > 1$ to $f(c)(\ln f(c) - 1) = (c-1)/(c+1)$.

Let us make some comments on these results. Roughly speaking, $\mathcal{T}_n$ looks like a "tentacular bush" in the sense that two typical vertices are always at the same distance of order $2\mathsf{h}_n^+ \sim \frac{c+1}{c} \cdot \ln(n)$, but the total height is of order $f(c) \cdot \mathsf{h}_n^+$. A typical example where the assumptions (2) are satisfied is e.g. when $\max_{A_n \le i \le n} |S_i^n/i - c| \to 0$. This is for instance the case when $p \in (1/2, 1)$, $\mathbf{X} = (\mathrm{X}_i)_{i\ge 1}$ is a sequence of i.i.d random variables such that $\mathbb{P}(X_1 = +1) = p$ and $\mathbb{P}(\mathrm{X}_1 = -1) = 1 - p$, setting $S_i^n = S_i(\mathbf{X})$, conditionally given $\tau(\mathbf{X}) > n$, Theorem 4 applies almost surely with $c = 2p - 1$ thanks to the laws of large numbers. The reason we rather use (2) is for applications to contact-tracing in the SIR model (see Sec. 6).

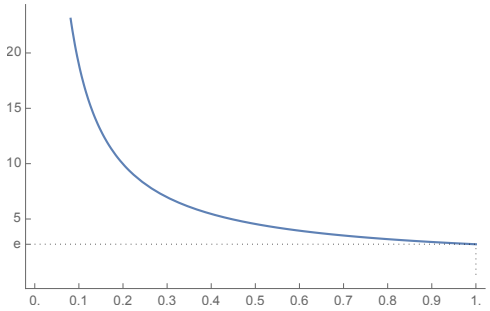

Figure 2: Plot of the function $(\frac{1+c}{2c} f(c) : 0 < c \le 1)$.

Theorem 4 (3) gives a more precise result than Theorem 3 (2) under an additional assumption. It generalizes the well known result that when $\mathcal{T}_n$ is a random recursive tree (which is obtained by taking $\mathrm{x}_n = 1$ for every $n \ge 1$), we have [16, Theorem 10]: $\mathsf{Height}(\mathcal{T}_n)/\ln(n) \to e$ in probability (see also [35, Theorem 1] for a more robust approach based on branching processes).

The proof of Theorem 4 (3) is rather delicate: the main difficulty is that the presence of freezing impedes the direct use of branching process techniques. The alternative growth-coalescent process construction previously mentioned lies at the heart of our proof.

Let us mention that it is possible to check that $(\mathcal{T}_n/\ln(n))_{n\geq 1}$ is not tight for the so-called Gromov–Hausdorff–Prokhorov topology. However, for the topology considered in [18, 24], $\mathcal{T}_n/\ln(n)$ converges to the so-called long dendron Υ_ν (we use the notation of [24, Example 3.12]) with ν being the Dirac mass $\nu = \delta_{(c+1)/2c}$.

Application to contact-tracing in a stochastic SIR dynamics. Our results may be applied to analyze the geometry of the so-called "infection tree" of a stochastic SIR dynamics, in which the vertices are individuals and where edges connect two individuals if one has infected the other. To keep this introduction at reasonable length, we refer to Section 6 for details and to Theorem 24 for a precise statement.

Plan of the paper. The rest of the paper is organized as follows: In Section 2, we present our alternative construction, and describe a few of its properties. In Section 3, we show Theorem 1 and 2 about local limits. In Sections 4 and 5, we prove respectively Theorems 3 and 4 about the height of our trees. We then apply our results in Section 6 to a stochastic SIR dynamics. Finally, in Section 7 we give a few open problems.

Acknowledgments. We thank Christina Goldschmidt for stimulating discussions at early stages of this work. We also thank the wonderful organization committee (Serte Donderwinkel, Christina Goldschmidt, Remco van der Hofstad, and Joost Jorritsma) of the RandNET Summer School and Workshop on Random Graphs, where this work was initiated.

2 Trees constructed by uniform attachment with freezing

We start by defining our model. We also provide for future use a table of notation below (Table 1).

Table 1: Table of the main notation and symbols introduced in Section 2 and used later.

$\mathbb{N}$	$= \{1, 2, 3, \dots\}$ positive integers
$[\![i, j]\!]$	$= \{i, i+1, \dots, j-1, j\}$ all integers between i and j
$\#A$	cardinality of a finite set A
$\mathbf{x} = (\mathrm{x}_n)_{n\in\mathbb{N}}$	a sequence of elements of $\{-1, 1\}$
$S_n(\mathbf{x})$	$= 1 + \sum_{i=1}^n \mathbf{x}_i$
$\tau(\mathbf{x})$	$= \inf\{n \geq 1 : S_n(\mathbf{x}) = 0\}$
$\mathcal{T}_n(\mathbf{x})$	tree built at time n by Algorithm 1; $S_n(\mathbf{x})$ is its number of active vertices
$N_n(\mathbf{x})$	total number of vertices in $\mathcal{T}_n(\mathbf{x})$; $N_n(\mathbf{x}) = (S_n(\mathbf{x}) + n + 1)/2$ when $n \leq \tau(\mathbf{x})$
$\mathcal{F}_n^n(\mathbf{x}), \mathcal{F}_{n-1}^n(\mathbf{x}), \dots, \mathcal{F}_0^n(\mathbf{x})$	the forest of trees built by Algorithm 2
$\mathcal{T}^n(\mathbf{x})$	$= \mathcal{F}_0^n(\mathbf{x})$ the -output of Algorithm 2
$\mathbb{F}_n(\mathbf{x})$	$= \{i \in [\![1, n]\!] : \mathrm{x}_i = -1\}$ the labels of frozen vertices of $\mathcal{T}^n(\mathbf{x})$
$\mathbb{A}_n(\mathbf{x})$	$= \{a_1, \dots, a_{S_n(\mathbf{x})}\}$ the labels of active vertices of $\mathcal{T}^n(\mathbf{x})$
$\mathbb{V}_n(\mathbf{x})$	$= \mathbb{F}_n(\mathbf{x}) \cup \mathbb{A}_n(\mathbf{x})$ the labels of all vertices of $\mathcal{T}^n(\mathbf{x})$
$\mathrm{b}_n(u)$	the birth time of $u \in \mathbb{V}_n(\mathbf{x})$ in the construction of $\mathcal{T}^n(\mathbf{x})$ by Algorithm 2
$\mathrm{c}_n(u, v)$	the coalescence time between $u, v \in \mathbb{V}_n(\mathbf{x})$ in the construction of $\mathcal{T}^n(\mathbf{x})$ by Algorithm 2
$\mathrm{H}_i^n(u)$	the height of vertex u in $\mathcal{F}_i^n(\mathbf{x})$

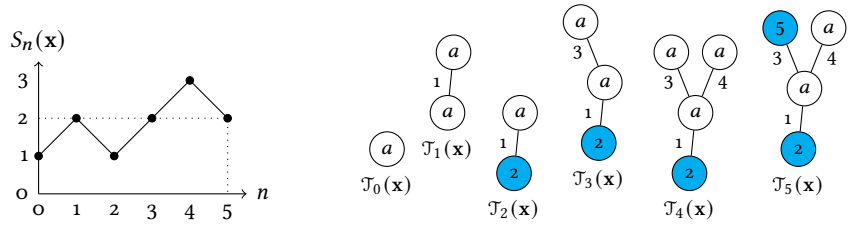

Figure 3: On the left is represented the walk $(S_n(\mathbf{x}))_{n\geq 0}$ up to time $n = 5$ associated with the sequence $\mathbf{x} = +1, -1, +1, +1, -1, \ldots$. On the right, a possible realisation of the trees $\mathcal{T}_0(\mathbf{x})$ to $\mathcal{T}_5(\mathbf{x})$ given this sequence. Frozen vertices have been colored in blue.

2.1 Uniform attachment with freezing: recursive construction

Let $\mathbf{x} = (\mathrm{x}_i)_{i\geq 1} \in \{-1, +1\}^{\mathbb{N}}$. In what follows, we formally construct the random trees $(\mathcal{T}_n(\mathbf{x}))_{n\geq 0}$. These trees will be rooted, edge-labelled, and vertex-labelled. The label of an edge is the time it appears. And the label of a vertex is either the time it freezes, or the label "a" if it is still active at time n.

Algorithm 1.

- Start with the tree $\mathcal{T}_0(\mathbf{x})$ made of a single root vertex labelled a.
- For every $n \geq 1$, if $\mathcal{T}_{n-1}(\mathbf{x})$ has no vertices labelled a, then set $\mathcal{T}_n(\mathbf{x}) = \mathcal{T}_{n-1}(\mathbf{x})$. Otherwise let V_n be a random uniform active vertex of $\mathcal{T}_{n-1}(\mathbf{x})$, chosen independently from the previous ones. Then:
  - if $\mathrm{x}_n = -1$, build $\mathcal{T}_n(\mathbf{x})$ from $\mathcal{T}_{n-1}(\mathbf{x})$ simply by replacing the label a of V_n with the label n;
  - if $\mathrm{x}_n = 1$, build $\mathcal{T}_n(\mathbf{x})$ from $\mathcal{T}_{n-1}(\mathbf{x})$ by adding an edge labelled n between V_n and a new vertex labelled a.

For $n \geq 0$, we view $\mathcal{T}_n(\mathbf{x})$ as a rooted, double-labelled tree (that is edge-labelled and vertex-labelled). If $N_n(\mathbf{x})$ represents the total number of vertices of $\mathcal{T}_n(\mathbf{x})$, observe that by construction $N_n(\mathbf{x}) = (S_n(\mathbf{x}) + n + 1)/2$ for $0 \leq n \leq \tau(\mathbf{x})$.

2.2 Uniform attachment with freezing: growth-coalescent construction

As before, let $\mathbf{x} = (\mathrm{x}_n)_{n\geq 1} \in \{-1, 1\}^{\mathbb{N}}$. We introduce in this section an alternative time-reversed construction of uniform attachment trees with freezing. It may be seen as a growth coalescence process of forests, and as previously mentioned, is a generalization of the connection between random recursive trees and Kingman's coalescent introduced in [17]. Most of our proofs are based on this construction.

Algorithm 2. Fix $0 \leq n \leq \tau(\mathbf{x})$. We construct a sequence $(\mathcal{F}_n^n(\mathbf{x}), \mathcal{F}_{n-1}^n(\mathbf{x}), \ldots, \mathcal{F}_0^n(\mathbf{x}))$ of forests of rooted, edge-labelled, vertex-labelled, unoriented trees by induction as follows.

- Let $\mathcal{F}_n^n(\mathbf{x})$ be a forest made of $S_n(\mathbf{x})$ one-vertex trees labelled $a_1, \ldots, a_{S_n(\mathbf{x})}$.
- For every $1 \leq i \leq n$, if $\mathcal{F}_i^n(\mathbf{x})$ has been constructed, define $\mathcal{F}_{i-1}^n(\mathbf{x})$ as follows:
  - if $\mathrm{x}_i = -1$, $\mathcal{F}_{i-1}^n(\mathbf{x})$ is obtained by adding to $\mathcal{F}_i^n(\mathbf{x})$ a new one-vertex tree labelled i;
  - if $\mathrm{x}_i = 1$, let (T_1, T_2) be a pair of different random trees in $\mathcal{F}_i^n(\mathbf{x})$ chosen uniformly at random, independently of the previous choices; then $\mathcal{F}_{i-1}^n(\mathbf{x})$ is obtained from $\mathcal{F}_i^n(\mathbf{x})$ by adding an edge labelled i between the roots $r(T_1)$ and $r(T_2)$ of respectively T_1 and T_2, and rooting the tree thus obtained at $r(T_1)$;
- Let $\mathcal{T}^n(\mathbf{x})$ be the only tree of $\mathcal{F}_0^n(\mathbf{x})$.

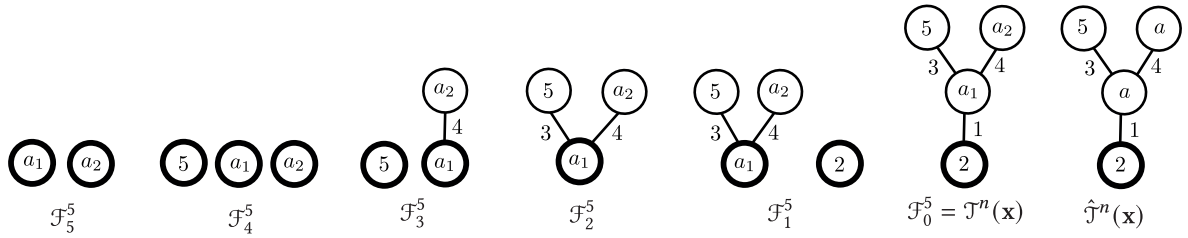

Figure 4: An illustration of Algorithm 2 with $n = 5$ and $(x_5, x_4, x_3, x_2, x_1) = (-1, 1, 1, -1, 1)$ (this is the same sequence as in Fig. 3). For example, since $x_2 = -1$, $\mathcal{F}_1^5$ is obtained from $\mathcal{F}_2^5$ by adding a new tree made of a vertex labeled 2. Since $x_4 = 1$, to build $\mathcal{F}_3^5$ from $\mathcal{F}_4^5$ we have chosen in $\mathcal{F}_3^5$ two trees (T_1, T_2) with T_1 being the vertex a_1 and T_2 being the vertex a_2, and we have added an edge labelled 4 between the roots $r(T_1) = a_1$ and $r(T_2) = a_2$ of respectively T_1 and T_2, and rooting the tree thus obtained at $r(T_1) = a_1$.

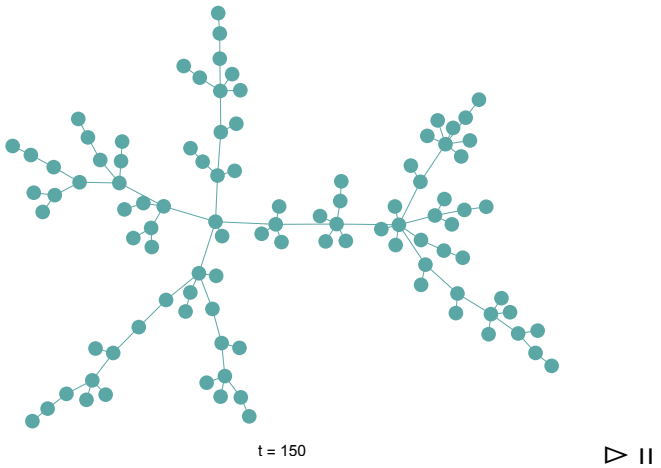

Figure 5: Simulation of Algorithm 2: the animation (played with Acrobat Reader) shows the resulting process as the number of steps increases. When a new vertex appears, its color is chosen at random. When two trees merge, the resulting tree keeps the color of the largest of the two trees.

We shall soon see in Sec. 2.4 the equivalence between Algorithm 1 and Algorithm 2, but before that, we introduce some important notation and consequences which will be useful in the analysis of the height of trees built by uniform attachment with freezing.

Important convention Throughout this article, when the context is clear, we will sometimes drop the parameter $\mathbf{x}$ to lighten notation: in particular, we shall write τ, S_n, $\mathcal{T}_n$, $\mathcal{F}_i^n$ instead of $\tau(\mathbf{x})$, $S_n(\mathbf{x})$, $\mathcal{T}_n(\mathbf{x})$, $\mathcal{F}_i^n(\mathbf{x})$, etc.

2.3 Laws of the birth and coalescence times

Let $\mathbf{x} = (x_n)_{n\geq 1} \in \{-1, 1\}^{\mathbb{N}}$. First, define

$$\mathbb{F}_n(\mathbf{x}) = \{i \in [\![1, n]\!] : x_i = -1\}, \qquad \mathbb{A}_n(\mathbf{x}) = \{a_1, \ldots, a_{S_n(\mathbf{x})}\}, \qquad \mathbb{V}_n(\mathbf{x}) = \mathbb{F}_n(\mathbf{x}) \cup \mathbb{A}_n(\mathbf{x}). \tag{4}$$

It is crucial to observe that while $\mathcal{T}^n$ is a random tree, the labels of its vertices are deterministic quantities that only depend on $\mathbf{x}$: $\mathbb{A}_n$ are the labels of the active vertices of $\mathcal{T}^n$ and $\mathbb{F}_n$ are the labels of the frozen

vertices of $\mathcal{T}^n$. In particular, the elements of $\mathbb{V}_n$ will be called vertices of $\mathcal{T}^n$.

Next, for every $u \in \mathbb{V}_n$, let $\mathrm{b}_n(u)$ be the largest $i \in \{0, 1, \dots, n\}$ such that u belongs to the forest $\mathcal{F}_i^n$. Explicitly, if $u \in \mathbb{A}_n$ is an active vertex then $\mathrm{b}_n(v) = n$, and if $u \in \mathbb{F}_n$ (note that u is then an integer) then $\mathrm{b}_n(u) = u - 1$ (see Fig. 4 for an example). We say that $\mathrm{b}_n(u)$ is the *birth time* of u, since it encodes the first time when vertex u appears in Algorithm 2. For $0 \le i \le \mathrm{b}_n(u)$, let $\mathrm{H}_i^n(u)$ be the height of vertex u in $\mathcal{F}_i^n$ (that is, the graph distance between u and the root of $\mathcal{F}_i^n$).

Finally, for $u, v \in \mathbb{V}_n$, let $\mathrm{c}_n(u, v)$ be the largest $i \in \{0, 1, \dots, n\}$ such that u and v belong to the same tree in the forest $\mathcal{F}_i^n$ obtained when building $\mathcal{T}^n$ in Algorithm 2. We say that $\mathrm{c}_n(u, v)$ is the *coalescence time* between u and v, since it encodes the first time u and v belong to the same tree in Algorithm 2 (observe that while $\mathrm{b}_n(u)$ is deterministic, $\mathrm{c}_n(u, v)$ is random).

We now state several simple consequences, which will be useful to study the geometry of $\mathcal{T}^n$.

Lemma 5. *Fix $1 \le n \le \tau$ and $u \in \mathbb{V}_n$. For every $1 \le i \le \mathrm{b}_n(v)$:*

$$\mathbb{P}\big(\mathrm{H}_{i-1}^n(v) - \mathrm{H}_i^n(v) = 1 | \mathcal{F}_n^n, \dots, \mathcal{F}_i^n\big) = \frac{1}{S_i} \mathbb{1}_{\{\mathrm{x}_i = 1\}}.$$

This is clear from the definition of Algorithm 2. In particular, if $(Y_i)_{1 \le i \le n}$ are independent Bernoulli random variables of respective parameters $(1/S_i)_{1 \le i \le n}$, we have for every $u \in \mathbb{V}_n$

$$\mathrm{H}_0^n(u) \overset{(d)}{=} \sum_{i=1}^{\mathrm{b}_n(u)} Y_i \mathbb{1}_{\{\mathrm{x}_i = 1\}}. \tag{5}$$

The next result identifies the law of the birth time of a vertex of $\mathcal{T}^n$ chosen uniformly at random.

Lemma 6. *Fix $1 \le n \le \tau$. Let V be an element of $\mathbb{V}_n$ chosen uniformly at random. For every $1 \le m \le n$,*

$$\mathbb{P}(\mathrm{b}_n(V) < m) = \frac{m + 1 - S_m}{n + 1 + S_n}.$$

Proof. Observe that $\mathrm{b}_n(u) = n$ if $u \in \mathbb{A}_n$ and $\mathrm{b}_n(u) = u - 1$ if $u \in \mathbb{F}_n$. Also, conditionally given the fact that $V \in \mathbb{F}_n$, $\mathrm{b}_n(V)$ is uniform on $\mathbb{F}_n$. Since $\#\mathbb{A}_n = S_n$ and $\#\mathbb{F}_n = \sum_{i=1}^n \mathbb{1}_{\{\mathrm{x}_i = -1\}}$, it follows by definition of $\mathbb{F}_n$ that

$$\mathbb{P}(\mathrm{b}_n(V) < m) = \frac{\sum_{i=1}^m \mathbb{1}_{\{\mathrm{x}_i = -1\}}}{S_n + \sum_{i=1}^n \mathbb{1}_{\{\mathrm{x}_i = -1\}}} = \frac{m + 1 - S_m}{n + 1 + S_n},$$

where for the last equality we have used the fact that $\sum_{i=1}^k \mathbb{1}_{\{\mathrm{x}_i = -1\}} = (k - S_k + 1)/2$ for every $1 \le k \le n$. □

The last useful result identifies the law of the coalescence times between two vertices.

Lemma 7. *Fix $1 \le n \le \tau$ and consider $u, v \in \mathbb{V}_n$. Then for every $0 \le c < \mathrm{b}_n(u) \wedge \mathrm{b}_n(v)$ with $\mathrm{x}_{c+1} = 1$:*

$$\mathbb{P}(\mathrm{c}_n(u, v) = c) = \frac{1}{\binom{S_{c+1}}{2}} \prod_{\substack{i = c+2 \\ \text{s.t. } \mathrm{x}_i = 1}}^{\mathrm{b}_n(u) \wedge \mathrm{b}_n(v)} \left(1 - \frac{1}{\binom{S_i}{2}}\right).$$

Proof. When running Algorithm 2, the factors in the product correspond to the probability that the two vertices born at times $\mathrm{b}_n(u)$ and $\mathrm{b}_n(v)$ do not coalesce until time c and the factor $1/\binom{S_{c+1}}{2}$ corresponds to the probability that the two vertices coalesce at time c. □

Table 2: Table of the main notation and symbols introduced in Section 2.4.

$\hat{\mathcal{T}}^n(\mathbf{x})$	the tree obtained from $\mathcal{T}^n(\mathbf{x})$ by relabelling all its active vertices by "a"
$\mathfrak{T}_{k,n}(A)$, $n \geq 1$ $0 \leq k \leq n$, $A \subset \mathbb{N}$ $\#A = 2n-k-1$	set of rooted trees with n vertices, such that all the $n-1$ edges and $n-k$ of the vertices are labelled in a one-to-one manner with the integers of A, while the other k vertices are labelled with the letter "a"
$\mathfrak{T}^+_{k,n}(A)$	set of all trees in $\mathfrak{T}_{k,n}(A)$ such that edge-labels increase along paths directed from the root, and all vertices labelled with an integer (i.e. not "a") have a label which is larger than the labels of their adjacent edges
$\mathfrak{T}^+_{k,n}$	$\mathfrak{T}^+_{k,n}(A)$ with $A = [\![1, 2n-k-1]\!]$

2.4 Equivalence between the two constructions

Observe that the active vertices of $\mathcal{T}_n$ are all labelled a while the active vertices of $\mathcal{T}^n$ are labelled $a_1, \dots, a_{S_n}$. It turns out that $\mathcal{T}_n$ is equal in law to $\mathcal{T}^n$ when its active vertices are all relabelled a. More precisely, let $\hat{\mathcal{T}}^n$ be the tree obtained from $\mathcal{T}^n$ by relabeling its S_n active vertices by a (see the right-most part of Fig. 4 for an illustration).

Theorem 8. *The two trees $\mathcal{T}_n$ and $\hat{\mathcal{T}}^n$ have the same distribution.*

To prove this result, we first identify the law of $\mathcal{T}_n$. We first look at the range of $\mathcal{T}_n$. For $n \in \mathbb{N}$, $0 \leq k \leq n$ and $A \subset \mathbb{N}$ with $\#A = 2n-k-1$, let $\mathfrak{T}_{k,n}(A)$ be the set of rooted trees with n vertices, such that all the $n-1$ edges and $n-k$ of the vertices are labelled in a one-to-one manner with the integers of A, while the other k vertices are labelled with the letter a. We then define $\mathfrak{T}^+_{k,n}(A)$ to be the set of all trees in $\mathfrak{T}_{k,n}(A)$ such that edge-labels increase along paths directed from the root, and all vertices labelled with an integer (i.e. not a) have a label which is larger than the labels of their adjacent edges. We set $\mathfrak{T}^+_{k,n} = \mathfrak{T}^+_{k,n}([\![1, 2n-k-1]\!])$. Note that $\mathcal{T}_n$ is an element of $\mathfrak{T}^+_{S_n,N_n}$ for $0 \leq n \leq \tau$ (see the left-most tree in Fig. 7 for an example) where $N_n = (S_n + n + 1)/2$ is the total number of vertices in $\mathcal{T}_n$. In particular, $\mathfrak{T}^+_{k,n}$ is the set of all possible trees with k active vertices and $n-k$ frozen vertices which can be obtained by Algorithm 1.

Lemma 9. *Fix $0 \leq n \leq \tau$ and $T_n \in \mathfrak{T}^+_{S_n,N_n}$. Then*

$$\mathbb{P}(\mathcal{T}_n = T_n) = \prod_{1 \leq i \leq n} \frac{1}{S_{i-1}}.$$

Proof. For two trees T, T' with $T \in \mathfrak{T}^+_{k,n}$, we write $T \rightsquigarrow T'$ if either $T' \in \mathfrak{T}^+_{k+1,n}$ and T' is obtained by replacing in T the label of a vertex labeled a by $2n-k$, or $T' \in \mathfrak{T}^+_{k,n+1}$ and T' is obtained from T by adding an edge labelled by $2n-k$ between a new vertex labeled a and an existing a-labeled vertex.

Since $T_n \in \mathfrak{T}^+_{S_n,N_n}$, observe that there is a unique sequence $(T_0, \dots, T_{n-1})$ such that $T_i \in \mathfrak{T}^+_{S_i,N_i}$ for every $0 \leq i \leq n-1$ and such that $T_i \rightsquigarrow T_{i+1}$ for every $0 \leq i \leq n-1$. In particular,

$$\mathbb{P}(\mathcal{T}_n = T_n) = \mathbb{P}((\mathcal{T}_0, \mathcal{T}_1 \dots, \mathcal{T}_n) = (T_0, T_1, \dots, T_n)) = \prod_{i=1}^{n} \mathbb{P}(\mathcal{T}_i = T_i | \mathcal{T}_{i-1} = T_{i-1}).$$

Then observe that $\mathbb{P}(\mathcal{T}_i = T_i | \mathcal{T}_{i-1} = T_{i-1}) = \frac{1}{S_{i-1}}$. Indeed, T_{i-1} has S_{i-1} active vertices. Thus, if $\mathrm{x}_i = 1$, then the probability of attaching a new vertex to a given active vertex is $1/S_{i-1}$; if $\mathrm{x}_i = -1$, then the probability of freezing a given active vertex is $1/S_{i-1}$. □

Proof of Theorem 8. First of all, recall that by Lemma 9, for every $T_n \in \mathfrak{T}^+_{S_n,N_n}$ we have

$$\mathbb{P}(\mathcal{T}_n = T_n) = \prod_{1\le i\le n} \frac{1}{S_{i-1}}.$$

Now let $\widetilde{\mathfrak{T}}^+_{k,n}$ be the set of all trees obtained from $\mathfrak{T}^+_{k,n}$ by labelling $a_1,\dots,a_k$ their k active vertices. It is enough to show that for every $\widetilde{T}_n \in \widetilde{\mathfrak{T}}^+_{k,n}$ we have

$$\mathbb{P}\Big(\mathcal{T}^n = \widetilde{T}_n\Big) = \frac{1}{S_n!}\prod_{1\le i\le n} \frac{1}{S_{i-1}}. \tag{6}$$

Indeed, since there are $S_n!$ ways to relabel by $a_1,\dots,a_k$ the S_n active vertices of any given tree $T_n \in \mathfrak{T}^+_{k,n}$, this will imply that $\mathbb{P}(\mathcal{T}_n = T_n) = \mathbb{P}(\hat{\mathcal{T}}^n = T_n)$.

To establish (6) we need to introduce some notation. Let $\widetilde{\mathfrak{F}}^n_k$ be the set of all forests of rooted, edge-labelled and vertex-labelled trees, such that:

- S_n vertices are labelled $a_1,\dots,a_{S_n}$;
- all the other vertices are labelled in a one-to-one fashion by $[\![k+1,n]\!] \cap \{1 \le i \le n : \mathrm{x}_i = -1\}$;
- the edges are labelled in a one-to-one fashion by $[\![k+1,n]\!] \cap \{1 \le i \le n : \mathrm{x}_i = 1\}$.

In addition, the labelling satisfies the following condition: edge-labels increase along paths directed from the roots, and all vertices labelled with an integer (i.e. not a) have a label which is larger than the labels of their adjacent edges. Notice that forests in $\widetilde{\mathfrak{F}}^n_k$ are made of S_k trees.

Observe that the random forest $\mathcal{F}^n_i$ satisfies $\mathcal{F}^n_i \in \widetilde{\mathfrak{F}}^n_i$ for every $0 \le i \le n$. If $F_i \in \widetilde{\mathfrak{F}}^n_i$ and $F_{i-1} \in \widetilde{\mathfrak{F}}^n_{i-1}$, we write $F_i \rightsquigarrow F_{i-1}$ if F_{i-1} is obtained from F_i by either adding a new tree made of a sole vertex labeled i, or choosing two different trees T_1, T_2 in F_i and adding an edge labeled i between the roots $r(T_1)$ and $r(T_2)$ of respectively T_1 and T_2, and rooting the tree thus obtained at $r(T_1)$.

Now, let F^n_0 be the forest made of the tree $\widetilde{T}_n$. Observe that there is a unique sequence $(F^n_n,\dots,F^n_1)$ such that $F^n_i \in \widetilde{\mathfrak{F}}^n_i$ for every $0 \le i \le n-1$ and such that $F^n_i \rightsquigarrow F^n_{i-1}$ for every $1 \le i \le n$. Notice that F^n_i is made of S_i trees. In particular,

$$\mathbb{P}\Big(\mathcal{T}^n = \widetilde{T}_n\Big) = \mathbb{P}\big((\mathcal{F}^n_n,\mathcal{F}^n_{n-1},\dots,\mathcal{F}^n_0) = (F^n_n,F^n_{n-1},\dots,F^n_0)\big) = \prod_{i=1}^{n} \mathbb{P}\big(\mathcal{F}^n_{i-1} = F^n_{i-1} | \mathcal{F}^n_i = F^n_i\big).$$

Then observe that if $\mathrm{x}_i = -1$ we have $\mathbb{P}\big(\mathcal{F}^n_{i-1} = F^n_{i-1} | \mathcal{F}^n_i = F^n_i\big) = 1$ since to build $\mathcal{F}^n_{i-1}$ from $\mathcal{F}^n_i$ with probability one we add a new one-vertex tree labelled i to $\mathcal{F}^n_{i-1}$, and since $\mathrm{x}_i = -1$, the forest F^n_{i-1} is also obtained from F^n_i by adding a new one-vertex tree labelled i. If $\mathrm{x}_i = 1$ we have the equality $\mathbb{P}\big(\mathcal{F}^n_{i-1} = \mathcal{F}^n_{i-1} | \mathcal{F}^n_i = \mathcal{F}^n_i\big) = \frac{1}{S_i(S_i-1)}$; indeed, there are $S_i(S_i-1)$ ways of choosing a pair (T_1,T_2) of different trees in F^n_i, and one of them gives F^n_{i-1} when adding an edge labeled i between the roots $r(T_1)$ and $r(T_2)$ of respectively T_1 and T_2, and rooting the tree thus obtained at $r(T_1)$.

As a consequence,

$$\mathbb{P}\Big(\mathcal{T}^n = \widetilde{T}_n\Big) = \prod_{\substack{1\le i\le n\\ \mathrm{x}_i=1}} \frac{1}{S_i(S_i-1)} = \frac{1}{S_n!}\prod_{1\le i\le n} \frac{1}{S_{i-1}},$$

where the last equality is readily checked by induction. This establishes (6) and completes the proof. □

2.5 A connection with increasing binary trees

It turns out that $\mathfrak{T}_{k,n}(A)$ is in bijection with increasing plane binary trees. Such a bijection is well-known in the case $k = 0$ (see e.g. [25] and [16, Sec. 6]). To make this connection explicit, we need some further notation. For $n \ge 1$, $0 \le k \le n$ and $A \subset \mathbb{N}$ with cardinality $\#A = 2n-k-1$ we define $\mathfrak{B}_{k,n}(A)$ to be the

Table 3: Table of the main notation and symbols introduced in Section 2.5.

$\mathfrak{B}_{k,n}(A)$	set of rooted binary plane trees with $2n-1$ vertices, such that k of the leaves are labelled with the letter a and the $2n-k-1$ other vertices are labelled in a one-to-one manner with the integers of A
$\mathfrak{B}^+_{k,n}(A)$	set of trees in $\mathfrak{B}_{k,n}(A)$ such that the vertices have increasing labels paths directed from the root.
$\mathfrak{B}^+_{k,n}$	$\mathfrak{B}^+_{k,n}(A)$ with $A = [\![1, 2n-k-1]\!]$

set of rooted binary (i.e. every vertex has either zero or two children) plane (i.e. there is an order among the children of internal vertices) trees with $2n-1$ vertices, such that k of the leaves are labelled with the letter a and the $2n-k-1$ other vertices are labelled in a one-to-one manner with the integers of A. We also define $\mathfrak{B}^+_{k,n}(A)$ to be the set of trees in $\mathfrak{B}_{k,n}(A)$ such that the vertices have increasing labels paths directed from the root. We set $\mathfrak{B}^+_{k,n} = \mathfrak{B}^+_{k,n}([\![1, 2n-k-1]\!])$ (see the right of Figure 7).

We are now ready to define a function Φ from $\bigcup \mathfrak{T}_{k,n}(A)$ to $\bigcup \mathfrak{B}_{k,n}(A)$ where the unions are over all possible triplets (k, n, A). The definition is done by induction. If τ is the trivial tree with one vertex then we simply set $\Phi(\tau) = \tau$. If τ has at least 2 vertices, then let u be the child of the root ρ such that the edge between ρ and u has the smallest label, denoted by ℓ, among all the edges adjacent to ρ. Denote by τ_1 the sub-tree rooted in u composed of all the descendants of u and τ_2 the sub-tree rooted in ρ composed of all the vertices that are not in τ_1. Then $\Phi(\tau)$ is the binary tree starting with a root, labelled ℓ, with two children: the first child on the left gives birth to $\Phi(\tau_1)$ and the second one on the right gives birth to $\Phi(\tau_2)$ (see Figure 6 and 7).

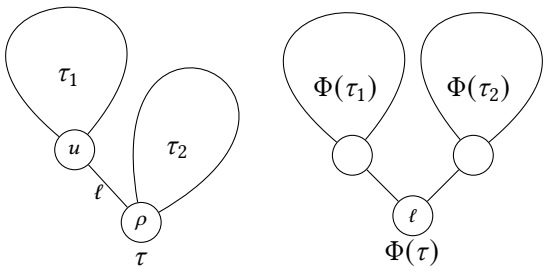

Figure 6: Illustration of the construction of the bijection Φ by induction.

Proposition 10. *The function Φ defined above is a bijection between $\mathfrak{T}^+_{k,n}$ and $\mathfrak{B}^+_{k,n}$.*

Proof. It suffices to construct explicitly the inverse bijection Ψ of Φ. We proceed by induction. For a tree τ with a single vertex let $\Psi(\tau) = \tau$. If τ is a binary tree with at least 2 children and a root labelled ℓ, let τ_1 be the tree on the left and let τ_2 be the tree on the right of the root. Then $\Psi(\tau)$ is obtained by attaching the root of $\Psi(\tau_1)$ to the root of $\Psi(\tau_2)$ with an edge labelled ℓ. And the root of $\Psi(\tau_2)$ becomes the root of $\Psi(\tau)$. It is a simple matter to check that Ψ is indeed the inverse bijection of Φ. □

As a consequence, we can identify the cardinality of $\mathfrak{T}^+_{0,n}$, which is the set of all possible trees with 0 active vertices and n frozen vertices which can be obtained by Algorithm 1.

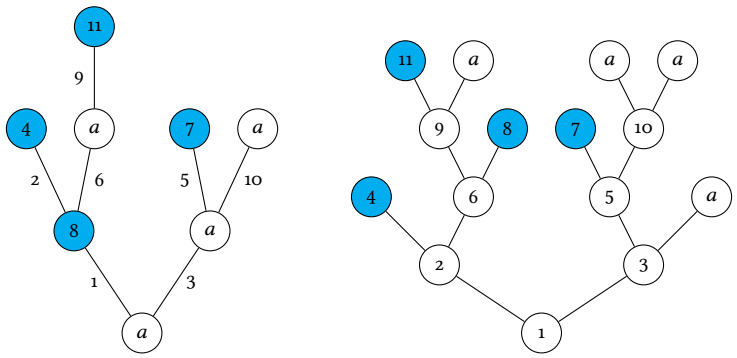

Figure 7: An example of a tree $\tau \in \mathfrak{T}^+_{4,8}$ on the left and its image by the bijection $\Phi(\tau) \in \mathfrak{B}^+_{4,15}$ on the right.

Corollary 11. *The cardinality of* $\mathfrak{T}^+_{0,n}$ *is given by*

$$\#\mathfrak{T}^+_{0,n} = \frac{4^n(4^n-1)|B_{2n}|}{2n} = \tan^{(2n-1)}(0)$$

where B_n *is the n-th Bernoulli number and* $\tan^{(n)}$ *stands for the n-th derivative of the tangent function. This sequence of numbers is given by OEIS A000182.*

3 Local limits

Let $\mathfrak{T}$ be the set of all rooted, vertex-labelled, edge-labelled, and locally finite (meaning that every vertex has finite degree) trees. For every $\tau \in \mathfrak{T}, h \in \mathbb{N}$ let $[\![\tau]\!]_h$ be the finite rooted vertex-labelled edge-labelled tree obtained from τ by keeping only the vertices at distance at most h from the root vertex together with the edges between them, and their labels. It is standard to construct a metric $\mathrm{d_{loc}}$ on $\mathfrak{T}$ such that the space $(\mathfrak{T}, \mathrm{d_{loc}})$ is Polish (i.e. separable and complete) and $\tau_n \to \tau$ for $\mathrm{d_{loc}}$ if and only if for every $h \geq 1$, we have $[\![\tau_n]\!]_h = [\![\tau]\!]_h$ for n large enough. Similar metrics can be defined for other families of trees (plane and unlabelled, plane and vertex-labelled and edge-labelled) and will be also denoted by $\mathrm{d_{loc}}$.

3.1 Proof of Theorem 1

We begin by showing that conditions 2 and 3 are equivalent. Since the sequence of trees $(\mathcal{T}_n(\mathbf{x}))_{n\geq 0}$ is increasing, it converges almost surely for $\mathrm{d_{loc}}$, if and only if, almost surely, the degree of each vertex converges (to a finite value). First, we show that every vertex degree converges a.s. if and only if every vertex freezes a.s. at some finite time. The converse implication is obvious (once a vertex is frozen its degree does not change), so we only need to prove the direct implication. Set $\mathsf{P} = \{n \geq 1 : \mathrm{x}_n = +1\}$ and $\mathsf{M} = \{n \geq 1 : \mathrm{x}_n = -1\}$. Observe that

$$\sum_{n\in\mathsf{P}} \frac{1}{S_n(\mathbf{x})} = +\infty. \tag{7}$$

Indeed, since the walk $(S_n(\mathbf{x}))_{n\geq 0}$ never touches 0 it is easy to see that $n^{-1}\#(\mathsf{P} \cap \{1,\dots,n\})$ does not converge to 0 (since it is at least equal to $1/2$). Thus $\sum_{n\in\mathsf{P}} 1/S_n(\mathbf{x}) \geq \sum_{n\in\mathsf{P}} 1/(n+1)$ diverges. Now fix $n_0 \in \mathsf{P}$ and denote by v the vertex that appears at time n_0. Let A be the event "v never freezes". We show that $\mathbb{P}(A) > 0$ implies that with positive probability the degree of v does not converge. To this end, for $n \in \mathsf{P}$, $n \geq n_0$, let E_n denote the event "the degree of v increases by one at time n". Then, since the event E_n involves times in P while the event A involves times in M, conditionally given A the events $(E_n)_{n\geq 1}$ are independent

and $\mathbb{P}(E_n|A) = 1/S_n(\mathbf{x})$. Thus, by (7) and Borel-Cantelli, $\mathbb{P}(\text{the degree of } v \text{ diverges as } n \to \infty|A) = 1$. Thus with positive probability the degree of v does not converge.

Now, to show that (2) and (3) are equivalent, it remains to show that, almost surely, the fact that all the vertices freeze at some point is equivalent to the divergence of the sum $\sum_{n\in\mathsf{M}} 1/S_n(\mathbf{x})$. Fix a vertex v appearing at time $n_0 \geq 0$. The probability that it never freezes is $\prod_{n\in\mathsf{M},n>n_0}(1-1/S_n(\mathbf{x}))$ which converges towards 0 if and only if $\sum_{n\in\mathsf{M}} 1/S_n(\mathbf{x}) = +\infty$.

Finally we prove that 1 is equivalent to 2 and 3. Let t be a finite tree with no vertex labelled a (so they are all frozen). If m is the biggest label of t then for all h and n the event "$[\![\mathcal{T}_n(\mathbf{x})]\!]_h = t$" depends only on the first m steps of $\mathbf{x}$. Therefore for all n large enough $\mathbb{P}([\![\mathcal{T}_n(\mathbf{x})]\!]_h = t) = \mathbb{P}([\![\mathcal{T}_n(\mathbf{x}^n)]\!]_h = t)$. If $\mathcal{T}_n(\mathbf{x})$ converges locally to $\mathcal{T}_\infty(\mathbf{x})$, then, by 3, all the vertices of $\mathcal{T}_\infty(\mathbf{x})$ are frozen so we deduce that $\mathcal{T}_n(\mathbf{x}^n)$ also converges locally to $\mathcal{T}_\infty(\mathbf{x})$. Now we suppose that $\mathcal{T}_n(\mathbf{x})$ does not converge locally. Then there exists i such that $\mathrm{x}_i = +1$ and the vertex v_i, added at time i, in $\mathcal{T}_n(\mathbf{x})$ has a degree converging towards $+\infty$ with positive probability. In other words

$$\mathbb{P}(\deg(v_i, \mathcal{T}_n(\mathbf{x})) \to +\infty) = \lim_{m\to\infty} \mathbb{P}\left(\bigcup_{n\geq i}\{\deg(v_i, \mathcal{T}_n(\mathbf{x})) \geq m\}\right) > 0.$$

Let φ be an increasing function such that for all $k \leq n$, $\mathbf{x}_k^{\varphi(n)} = \mathbf{x}_k$. There is an obvious coupling such that $\mathcal{T}_n(\mathbf{x}) = \mathcal{T}_n(\mathbf{x}^{\varphi(n)})$ for all n. Therefore, for all m,

$$\bigcup_{n\geq i}\{\deg(v_i, \mathcal{T}_n(\mathbf{x})) \geq m\} = \bigcup_{n\geq i}\{\deg(v_i, \mathcal{T}_n(\mathbf{x}^{\varphi(n)})) \geq m\},$$

and

$$\lim_{m\to\infty} \mathbb{P}\left(\bigcup_{n\geq i}\{\deg(v_i, \mathcal{T}_{\varphi(n)}(\mathbf{x}^{\varphi(n)})) \geq m\}\right) \geq \lim_{m\to\infty} \mathbb{P}\left(\bigcup_{n\geq i}\{\deg(v_i, \mathcal{T}_n(\mathbf{x}^{\varphi(n)})) \geq m\}\right) > 0.$$

We deduce that the degree of v_i in $\mathcal{T}_{\varphi(n)}(\mathbf{x}^{\varphi(n)})$ tends to ∞ with positive probability so $\mathcal{T}_n(\mathbf{x}^n)$ can't converge locally. □

3.2 A connection with Geometric Bienaymé trees: Proof of Theorem 2

To set up the proof of Theorem 2, we need to introduce the formalism of plane trees. Let $\mathcal{U}$ be the set of labels defined by

$$\mathcal{U} = \bigcup_{n=0}^{\infty} (\mathbb{N})^n,$$

where, by convention, $(\mathbb{N})^0 = \{\varnothing\}$. In other words, an element of $\mathcal{U}$ is a (possibly empty) sequence $u = u_1 \cdots u_j$ of positive integers. When $u = u_1 \cdots u_j$ and $v = v_1 \cdots v_k$ are elements of $\mathcal{U}$, we let $uv = u_1 \cdots u_j v_1 \cdots v_k$ be the concatenation of u and v. In particular, $u\varnothing = \varnothing u = u$. Finally, a *plane tree* is a subset of $\mathcal{U}$ satisfying the following three conditions:

(i) $\varnothing \in \tau$,

(ii) if $v \in \tau$ and $v = uj$ for a certain $j \in \mathbb{N}$, then $u \in \tau$,

(iii) for every $u \in \tau$, there exists an integer $k_u(\tau) \geq 0$ such that for every $j \in \mathbb{N}$, $uj \in \tau$ if and only if $1 \leq j \leq k_u(\tau)$.

If τ is a plane tree and $u \in \tau$ let $\theta_u\tau = \{v \in \mathcal{U} : uv \in \tau\}$ be the subtree of descendants of u. We finally let $\mathbb{T}$ be the set of all plane trees (observe that such trees may be infinite, but are always locally finite).

Given a probability distribution μ on $\mathbb{Z}_+$, the law $\mathbb{P}_\mu$ of a Bienaymé tree with offspring distribution μ (sometimes known as a Galton–Watson tree, but we prefer to use the terminology suggested in [2]) is a probability measure on $\mathbb{T}$ that can be characterized by the following two properties (see e.g. [34] for more general statements):

(i) $\mathbb{P}_\mu(k_\varnothing = j) = \mu(j)$ for $j \in \mathbb{Z}_+$,

(ii) for every $j \geq 1$ with $\mu(j) > 0$, the shifted trees $\theta_1\tau, \ldots, \theta_j\tau$ are independent under the conditional probability $\mathbb{P}_\mu(\mathrm{d}\tau | k_\varnothing = j)$ and their conditional distribution is $\mathbb{P}_\mu$.

The following simple result gives a characterization of Bienaymé trees with geometric offspring distribution in terms of an invariance property involving subtrees (since we have not managed to find a reference in the literature we include the short proof for completeness).

Lemma 12. *Let $p \in [0,1)$. Let $\mathcal{T}$ be a random plane tree such that $\mathbb{P}(k_\varnothing(\mathcal{T}) = 0) = 1 - p$ and under the conditional probability $\mathbb{P}(\cdot | k_\varnothing(\mathcal{T}) > 0)$ the two trees $\theta_1\mathcal{T}$ and $\mathcal{T}\backslash\theta_1\mathcal{T}$ are independent with conditional distribution equal to the law of $\mathcal{T}$. Then $\mathcal{T}$ is a Bienaymé tree with offspring distribution μ given by $\mu(k) = (1-p)p^k$ for $k \geq 0$.*

Proof. First, for $j \geq 1$ we have $\mathbb{P}(k_\varnothing(\mathcal{T}) \geq j) = \mathbb{P}(k_\varnothing(\mathcal{T}) > 0)\mathbb{P}(k_\varnothing(\mathcal{T}\backslash\theta_1\mathcal{T}) \geq j-1 | k_\varnothing(\mathcal{T}) > 0)$, which implies $\mathbb{P}(k_\varnothing(\mathcal{T}) \geq j) = p\mathbb{P}(k_\varnothing(\mathcal{T}) \geq j-1)$. Thus $\mathbb{P}(k_\varnothing(\mathcal{T}) = k) = (1-p)p^k$ for $k \geq 0$.

Second, we argue by induction. Take $n \geq 1$ and assume that the shifted trees $\theta_1\mathcal{T}, \ldots, \theta_n\mathcal{T}$ are independent under the conditional probability $\mathbb{P}(\,\cdot\,| k_\varnothing(\mathcal{T}) = n)$ with conditional distribution equal to the law of $\mathcal{T}$. Then for $f_1, \ldots, f_{n+1} \geq 0$ measurable

$$\mathbb{E}[f_1(\theta_1\mathcal{T}) \cdots f_{n+1}(\theta_{n+1}\mathcal{T}) | k_\varnothing(\mathcal{T}) = n+1] = \mathbb{E}[f_1(\theta_1\mathcal{T}) g_n(\mathcal{T}\backslash\theta_1\mathcal{T}) | k_\varnothing(\mathcal{T}) > 0]$$

with $g_n(\tau) = \mathbb{1}_{k_\varnothing(\tau)=n} f_2(\theta_1\tau) \cdots f_{n+1}(\theta_n\tau)\mathbb{P}(k_\varnothing(\mathcal{T}) > 0)/\mathbb{P}(k_\varnothing(\mathcal{T}) = n+1)$. Thus, by assumption, we have $\mathbb{E}[f_1(\theta_1\mathcal{T}) \cdots f_{n+1}(\theta_{n+1}\mathcal{T}) | k_\varnothing(\mathcal{T}) = n+1] = \mathbb{E}[f_1(\mathcal{T})]\mathbb{E}[g_n(\mathcal{T})]$. But

$$\mathbb{E}[g_n(\mathcal{T})] = \frac{\mathbb{P}(k_\varnothing(\mathcal{T}) = n)\mathbb{P}(k_\varnothing(\mathcal{T}) > 0)}{\mathbb{P}(k_\varnothing(\mathcal{T}) = n+1)} \mathbb{E}[f_2(\theta_1\mathcal{T}) \cdots f_{n+1}(\theta_n\mathcal{T}) | k_\varnothing(\mathcal{T}) = n],$$

which is equal to $\mathbb{E}[f_2(\theta_1\mathcal{T}) \cdots f_{n+1}(\theta_n\mathcal{T}) | k_\varnothing(\mathcal{T}) = n]$ thanks to the first step. The desired result follows from the induction hypothesis. □

For every tree $\tau \in \mathfrak{T}^+_{k,n}(A)$, we define $\tilde{\tau}$ to be the unique element of $\mathfrak{T}^+_{k,n}$ obtained by shifting the labels of τ; specifically if $f : \{1, \ldots, 2n-k-1\} \to A$ denotes the unique increasing bijection, an edge/vertex is labeled $f(i)$ in τ if and only if its corresponding edge/vertex is labeled i in $\tilde{\tau}$ (see Figure 8).

We are now ready to establish Theorem 2.

Proof of Theorem 2. First, we claim that almost surely the sequence $(\mathcal{T}_n(\mathbf{X}))_{n\geq 0}$ converges locally towards a tree denoted by $\mathcal{T}_\infty$. If $p \leq 1/2$, $(S_n)_{n\geq 0}$ touches 0 almost surely at some point, so the sequence of trees $(\mathcal{T}_n(\mathbf{X}))_{n\geq 0}$ is almost surely stationary. If $p > 1/2$, conditionally on the fact that $(S_n)_{n\geq 0}$ never touches 0, this walk satisfies the condition of Theorem 1, so $(\mathcal{T}_n(\mathbf{X}))_{n\geq 0}$ has a local limit. We conclude that, for every value of $p \in [0,1)$, almost surely the sequence $(\mathcal{T}_n(\mathbf{X}))_{n\geq 0}$ converges locally towards a tree $\mathcal{T}_\infty$.

We shall first consider $\mathcal{T}_\infty$ conditioned on having at least 2 vertices (which amounts to condition on the event $\{X_1 = +1\}$). Conditionally given $\{X_1 = +1\}$, $\mathcal{T}_n(\mathbf{X})$ is composed of two trees: the tree $\mathcal{T}^1_n(\mathbf{X})$ representing all the descendants of the first child of the root of $\mathcal{T}_n(\mathbf{X})$, and the tree $\mathcal{T}^2_n(\mathbf{X})$ representing all the other vertices, including the root of $\mathcal{T}_n(\mathbf{X})$ (see Figure 8).

The main idea is then to check that if one of the trees $\mathcal{T}_n^1(\mathbf{X})$ or $\mathcal{T}_n^2(\mathbf{X})$ has at least one active vertex, then it will eventually almost surely evolve. This will show that the two trees $\tilde{\mathcal{T}}_n^1(\mathbf{X})$ or $\tilde{\mathcal{T}}_n^2(\mathbf{X})$ evolve independently with the same transition probabilities, which in turn will entail that conditionally given $\{\mathrm{X}_1 = 1\}$, $(\tilde{\mathcal{T}}_n^1(\mathbf{X}))_{n\geq 1}$ and $(\tilde{\mathcal{T}}_n^2(\mathbf{X}))_{n\geq 1}$ converge locally, jointly, towards two independent trees distributed like $\mathcal{T}_\infty$.

Let us now make these considerations more formal. Notice that $(\mathcal{T}_n(\mathbf{X}))_{n\geq 0}$ is a Markov chain in the state space $E := \bigcup_{n\geq 1}\bigcup_{0\leq k\leq n} \mathfrak{T}_{n,k}^+$ starting at the trivial tree, denoted by $t_\circ$, composed of a single active vertex. Denote by α its transition matrix. The absorbing states are exactly the trees with no active vertices. For $n \geq 2$, let I_n be the random variable which is equal to 1 when $S_k = 0$ for some $k \leq n-1$ or when the n-th action (corresponding to the step X_n) occurs in the tree $\mathcal{T}_{n-1}^1(\mathbf{X})$ and 2 if it occurs in the tree $\mathcal{T}_{n-1}^2(\mathbf{X})$. More precisely, saying that $\mathrm{X}_n = -1$ and $S_k > 0$ for all $k \leq n-1$, is equivalent to saying that an active vertex of $\mathcal{T}_{n-1}(\mathbf{X})$, chosen uniformly at random, will freeze. If this vertex is chosen in $\mathcal{T}_{n-1}^1(\mathbf{X})$, then $I_n = 1$, otherwise, it is chosen in $\mathcal{T}_{n-1}^2(\mathbf{X})$ and $I_n = 2$. By convention we also set $I_1 = 1$. Denote by $a(t)$ the number of active vertices of a tree $t \in E$.

The previous discussion shows that, conditionally given the event $\{\mathrm{X}_1 = 1\}$, $\left(\tilde{\mathcal{T}}_n^1(\mathbf{X}), \tilde{\mathcal{T}}_n^2(\mathbf{X}), I_n\right)_{n\geq 1}$ is a Markov chain in the state space $E^2 \times \{1,2\}$, starting at $(t_\circ, t_\circ, 1)$, with transitions given by

$$\begin{aligned}
(x,y,i) &\to (x',y,1) \qquad \text{with probability} \quad k(x,y)\alpha(x,x')\\
(x,y,i) &\to (x,y',2) \qquad \text{with probability} \quad (1-k(x,y))\alpha(y,y').
\end{aligned}$$

for every $x, x', y, y' \in E$ and $i \in \{1,2\}$ where

$$k(x,y) := \frac{a(x)}{a(x)+a(y)} \quad \text{if } a(x)+a(y) > 0 \qquad \text{and} \qquad k(x,y) := 1 \ \text{ otherwise.}$$

Let $x \in E$ be a non-absorbing state (i.e. $a(x) > 0$). Let $y \in E$ and $n \geq 1$. We show that conditionally given $\{\tilde{\mathcal{T}}_n^1(\mathbf{X}) = x \text{ and } \tilde{\mathcal{T}}_n^2(\mathbf{X}) = y\}$ the probability that $I_k = 1$ for some $k > n$ is 1. The forthcoming Lemma 13 then entails that conditionally given $\{\mathrm{X}_1 = 1\}$, the trees $(\tilde{\mathcal{T}}_n^1(\mathbf{X}))_{n\geq 1}$ and $(\tilde{\mathcal{T}}_n^2(\mathbf{X}))_{n\geq 1}$ converge locally, jointly, towards two independent trees distributed like $\mathcal{T}_\infty$.

Now, to show that conditionally given $\{\tilde{\mathcal{T}}_n^1(\mathbf{X}) = x \text{ and } \tilde{\mathcal{T}}_n^2(\mathbf{X}) = y\}$ the probability that $I_k = 1$ for some $k > n$ is 1, observe that since $a(x) > 0$, we have $S_k > 0$ for every $1 \leq k \leq n$. If $S_k = 0$ for some $k > n$, this implies that both $\mathcal{T}_k^1(\mathbf{X})$ and $\mathcal{T}_k^2(\mathbf{X})$ have no active vertices left. So in particular, there must be an integer $n < \ell \leq k$ such that $I_\ell = 1$. Conditionally given $(S_k)_{k\geq n}$, if $S_k > 0$ for all $k > n$, then the probability that $I_k = 2$ for all $k > n$ is

$$\prod_{k\geq n} \frac{S_k - a(x)}{S_k} \leq \prod_{k\geq n}\left(1 - \frac{1}{S_k}\right) \leq \prod_{k\geq n}\left(1 - \frac{1}{k+1}\right) = 0.$$

Therefore we can apply the forthcoming Lemma 13 and deduce that, conditionally given $\{\mathrm{X}_1 = 1\}$, the trees $(\tilde{\mathcal{T}}_n^1(\mathbf{X}))_{n\geq 1}$ and $(\tilde{\mathcal{T}}_n^2(\mathbf{X}))_{n\geq 1}$ converge locally, jointly, towards two independent trees distributed like $\mathcal{T}_\infty$.

To conclude, for a labelled locally finite tree τ denote by $\overline{\tau}$ to be the plane tree obtained from τ by first ordering the vertices in such a way that the labels of edges connecting vertices to their children are increasing from left to right, and second erasing all the labels (for example the vertices of the tree on the left of Fig. 7 are ordered in such a way). By continuity of the operation $\tau \mapsto \overline{\tau}$ on the set of labelled locally finite trees with respect to the local topology, we conclude that conditionally on the fact that $\overline{\mathcal{T}}_\infty$ is not a single vertex, the two trees $\overline{\mathcal{T}}_\infty^1$ and $\overline{\mathcal{T}}_\infty^2$ are independent copies of $\overline{\mathcal{T}}_\infty$. Since $\overline{\mathcal{T}}_\infty$ has only one vertex with probability $1-p$ (this happens if and only if $\mathrm{X}_1 = -1$), the desired result follows from Lemma 12. □

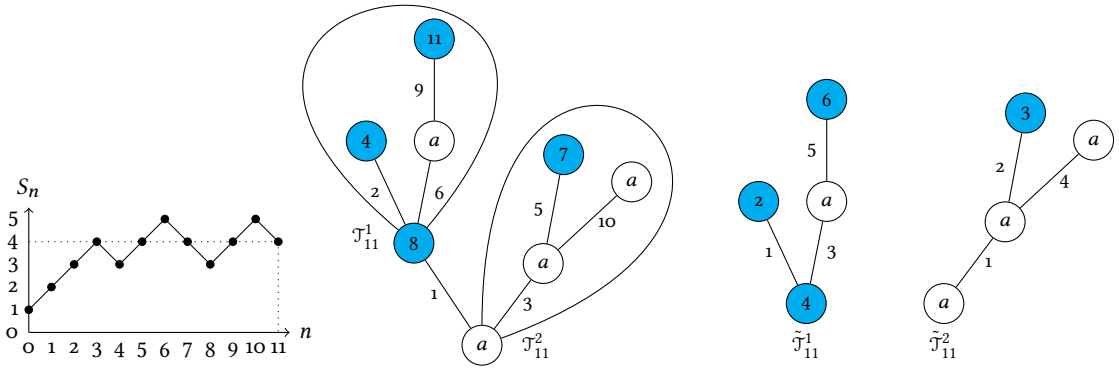

Figure 8: Illustration of the proof of Theorem 2

Lemma 13. *Let $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ be two Markov chains on the countable state space E with the same transition matrix α. Denote by $A := \{x \in E : \alpha(x,x) = 1\}$ their set of absorbing states. Let $k : E \times F \to [0,1]$. We define a Markov chain $(U_n, V_n, I_n)_{n\geq 0}$ on the state space $E^2 \times \{1,2\}$ with transition matrix given by, for all $x, x', y, y' \in E$ and $i \in \{1,2\}$:*

$$(x,y,i) \to (x',y,1) \qquad \text{with probability} \quad k(x,y)\alpha(x,x')$$
$$(x,y,i) \to (x,y',2) \qquad \text{with probability} \quad (1-k(x,y))\alpha(y,y').$$

Suppose that, for all $(x,y,i) \in (E\setminus A)\times E\times\{1,2\}$, $\mathbb{P}_{(x,y,i)}(\exists n \geq 1,\, I_n = 1) = 1$. Similarly, suppose that for all $(x,y,i) \in E\times(E\setminus A)\times\{1,2\}$, $\mathbb{P}_{(x,y,i)}(\exists n \geq 1,\, I_n = 2) = 1$. Let $n_0 = m_0 = 0$. For all $k \geq 1$, if the set $\{n : n > n_{k-1} \text{ and } I_n = 1\}$ is non-empty, we define n_k as its minimum, otherwise we set $n_k = n_{k-1}$. Similarly, if the set $\{m : m > m_{k-1} \text{ and } I_m = 2\}$ is non-empty, we define m_k as its minimum, otherwise we set $m_k = m_{k-1}$. Then, for every $(x,y) \in E^2$, $\big((U_{n_k})_{k\geq 0}, (V_{m_k})_{k\geq 0}\big)$ under $\mathbb{P}_{(x,y,1)}$ has the same law as $((X_k)_{k\geq 0}, (Y_k)_{k\geq 0})$ under $\mathbb{P}_x \otimes \mathbb{P}_y$.

Proof of Lemma 13. We show that $\big((U_{n_k})_{0\leq k\leq r}, (V_{m_k})_{0\leq k\leq s}\big)$ under $\mathbb{P}_{(x,y,1)}$ has the same law as $((X_k)_{0\leq k\leq r}, (Y_k)_{0\leq k\leq s})$ under $\mathbb{P}_x \otimes \mathbb{P}_y$ by double induction on (r,s). If $r = s = 0$, the result is obvious. Suppose it is true for some pair (r,s). We show it for $(r,s+1)$ (the case $(r+1,s)$ is similar). Let $\mathbf{x} = (x_1,\dots,x_r) \in E^r$, $\mathbf{y} = (y_1,\dots,y_s) \in E^s$ and $y_{s+1} \in E$. Write $\mathbf{U} := (U_{n_k})_{1\leq k\leq r}$, $\mathbf{V} := (V_{m_k})_{1\leq k\leq s}$, $\mathbf{X} := (X_k)_{1\leq k\leq s}$, $\mathbf{Y} := (Y_k)_{1\leq k\leq s}$ and $\mathbb{P} := \mathbb{P}_{(x,y,1)}$. Suppose that $y_s \notin A$, then $m_{s+1} > m_s$ and

$$\mathbb{P}\big(\mathbf{U}=\mathbf{x}, \mathbf{V}=\mathbf{y}, V_{m_{s+1}} = y_{s+1}\big) = \mathbb{P}(\mathbf{U}=\mathbf{x}, \mathbf{V}=\mathbf{y})\mathbb{P}_{(x_r,y_s,1)}(\exists n \geq 1,\, I_1 = \cdots = I_{n-1} = 1,\, I_n = 2,\, V_n = y_{s+1}).$$

By induction hypothesis, $\mathbb{P}(\mathbf{U}=\mathbf{x}, \mathbf{V}=\mathbf{y}) = \mathbb{P}_x(\mathbf{X}=\mathbf{x})\mathbb{P}_y(\mathbf{Y}=\mathbf{y})$. Moreover

$$\begin{aligned}\mathbb{P}_{(x_r,y_s,1)}(\exists n \geq 1,\, I_1 = \cdots = I_{n-1} = 1,\, I_n = 2,\, V_n = y_{s+1}) &= \mathbb{P}_{(x_r,y_s,1)}(\exists n \geq 1, I_n = 2)\alpha(y_s, y_{s+1}) \\ &= \alpha(y_s, y_{s+1}).\end{aligned}$$

This shows the formula for $(r,s+1)$ in the case $y_s \notin A$. Now, if $y_s \in A$ then

$$\begin{aligned}\mathbb{P}\big(\mathbf{U}=\mathbf{x}, \mathbf{V}=\mathbf{y}, V_{m_{s+1}} = y_{s+1}\big) &= \mathbb{P}(\mathbf{U}=\mathbf{x}, \mathbf{V}=\mathbf{y})\mathbb{P}_{(x_r,y_s,1)}(\forall m \geq 1, I_m = 1)\mathbb{1}_{y_s = y_{s+1}} \\ &\quad + \mathbb{P}(\mathbf{U}=\mathbf{x}, \mathbf{V}=\mathbf{y})\mathbb{P}_{(x_r,y_s,1)}(\exists m \geq 1, I_m = 2)\alpha(y_s, y_{s+1}) \\ &= \mathbb{P}_x(\mathbf{X}=\mathbf{x})\mathbb{P}_y(\mathbf{Y}=\mathbf{y}, Y_{s+1} = y_{s+1}).\end{aligned}$$

This completes the proof. □

4 The height of uniform attachment trees with freezing

We shall now study the geometry of uniform attachment trees with freezing. It will be convenient to work with $\mathcal{T}^n(\mathbf{x})$ as built using Algorithm 2. Recall Theorem 8: the only difference between $\mathcal{T}_n(\mathbf{x})$ and $\mathcal{T}^n(\mathbf{x})$ is that all the active vertices of $\mathcal{T}^n(\mathbf{x})$ are labeled $a_1, \dots, a_{S_n(\mathbf{x})}$, while all the active vertices of $\mathcal{T}_n(\mathbf{x})$ are labelled a. In particular the graph structure of both trees is the same in law, so it is equivalent to establish our main results with $\mathcal{T}_n(\mathbf{x})$ replaced with $\mathcal{T}^n(\mathbf{x})$.

Also recall from (4) the definition of $\mathbb{V}_n(\mathbf{x})$, which is a deterministic set representing the labels of the vertices of $\mathcal{T}^n(\mathbf{x})$, and that by a slight abuse of notation we view elements of $\mathbb{V}_n(\mathbf{x})$ as vertices of $\mathcal{T}^n(\mathbf{x})$.

An important ingredient is Bennett's concentration inequality. Since it will be used multiple times, we state it here, taylored for our purpose:

Proposition 14 (Bennett's inequality). *Let $(Y_i)_{1\le i\le n}$ be independent Bernoulli random variables of respective parameters $(p_i)_{1\le i\le n}$. Set $m_n = \sum_{i=1}^n p_i$ and assume that $m_n > 0$. Then for every $t > 0$:*

$$\mathbb{P}\left(\sum_{i=1}^n Y_i > m_n + t\right) \le \exp\left(-m_n g\left(\frac{t}{m_n}\right)\right) \qquad \text{and} \qquad \mathbb{P}\left(\sum_{i=1}^n Y_i < m_n - t\right) \le \exp\left(-m_n g\left(\frac{t}{m_n}\right)\right),$$

where $g(u) = (1+u)\ln(1+u) - u$ for $u > 0$.

4.1 Uniform bounds on the height

We keep the notation introduced before (see in particular Table 1). In particular, $\mathbf{x} = (\mathrm{x}_n)_{n\ge1}$ is a fixed sequence of elements of $\{-1, 1\}$, $S_0(\mathbf{x}) := 1$ and for every $n \ge 1$

$$S_n(\mathbf{x}) := 1 + \sum_{i=1}^n \mathrm{x}_i; \qquad \tau(\mathbf{x}) := \inf\{n \ge 1 : S_n(\mathbf{x}) = 0\}. \tag{8}$$

To lighten notation we will drop the $\mathbf{x}$ in parenthesis (e.g. we will write S_n instead of $S_n(\mathbf{x})$, τ instead of $\tau(\mathbf{x})$, $\mathcal{T}^n$ instead of $\mathcal{T}^n(\mathbf{x})$, $\mathbb{V}_n(\mathbf{x})$ instead of $\mathbb{V}_n$, etc.) but we keep in mind that all the quantities depend on the sequence $\mathbf{x}$. We first start with an estimate concerning the height of $\mathcal{T}^n$ (recall its construction using Algorithm 2 in Section 2.2).

Proposition 15. *Fix $n \ge 1$ and assume that $\tau > n$. Let $\mathsf{Height}(\mathcal{T}^n)$ denote the height of $\mathcal{T}^n$. Then, for every $t > 0$,*

$$\mathbb{P}\big(\mathsf{Height}(\mathcal{T}^n) - h_n^+ > t\big) \le \exp\left(-h_n^+ g\left(\frac{t}{h_n^+}\right) + \ln(n)\right)$$

and

$$\mathbb{P}\big(\mathsf{Height}(\mathcal{T}^n) - h_n^+ < -t\big) \le \exp\left(-h_n^+ g\left(\frac{t}{h_n^+}\right)\right)$$

where $g(u) = (1+u)\ln(1+u) - u$ for $u > 0$.

Proof. Fix $u \in \mathbb{V}_n$. From (5) if $(Y_i)_{1\le i\le n}$ are independent Bernoulli random variables of respective parameters $(1/S_i)_{1\le i\le n}$, we have

$$\mathsf{H}_0^n(u) \overset{(d)}{=} \sum_{i=1}^{\mathrm{b}_n(u)} Y_i \mathbb{1}_{\{\mathrm{x}_i=1\}}.$$

Thus, by a union bound over all vertices of $\mathcal{T}^n$ we get

$$\mathbb{P}\big(\mathsf{Height}(\mathcal{T}^n) > h_n^+ + t\big) \le n\mathbb{P}\left(\sum_{i=1}^n \left(Y_i - \frac{1}{S_i}\right)\mathbb{1}_{\{\mathrm{x}_i=1\}} > t\right).$$

In addition, when $u \in \mathbb{V}_n$ is an active vertex (so that $\mathrm{b}_n(u) = n$), using the fact that $\mathrm{H}_0^n(u) \le \mathsf{Height}(\mathcal{T}_n)$, we get

$$\mathbb{P}\big(\mathsf{Height}(\mathcal{T}^n) < h_n^+ - t\big) \le \mathbb{P}\left(\sum_{i=1}^{n}\left(Y_i - \frac{1}{S_i}\right)\mathbb{1}_{\{\mathrm{x}_i=1\}} < -t\right).$$

The desired result then follows from Bennett's inequality. □

4.2 A simple lower bound for h_n^+

In view of Theorem 3, which tells us that h_n^+ is the order of magnitude of $\mathsf{Height}(\mathcal{T}^n)$, it is useful to have estimates on h_n^+. We give here a simple lower bound on h_n^+.

Lemma 16. *Let $n \ge 1$. If $\tau > n$ then $\mathrm{h}_n^+ \ge \ln(n/4)$.*

Proof. Set $p_n := \#\{1 \le i \le n : \mathrm{x}_i = +1\}$ and $m_n := \#\{1 \le i \le n : \mathrm{x}_i = -1\}$. If there exists $1 \le i \le n-1$ such that $\mathrm{x}_i = -1$ and $\mathrm{x}_{i+1} = +1$, set $\mathbf{x}' = (\mathrm{x}_1, \ldots, \mathrm{x}_{i-1}, +1, -1, \mathrm{x}_{i+2}, \mathrm{x}_{i+3}, \ldots)$. Then it is a simple matter to check that $\mathrm{h}_n^+(\mathbf{x}) \ge \mathrm{h}_n^+(\mathbf{x}')$. By iteration, it readily follows that $\mathrm{h}_n^+(\mathbf{x}) \ge \mathrm{h}_n^+((+1, \ldots, +1, -1, \ldots, -1, \mathbf{x}_{n+1}, \mathbf{x}_{n+2}, \ldots))$ where $+1$ appears p_n times and -1 appears m_n times. Since $\tau(\mathbf{x}) > n$ entails $p_n \ge n/2$, it follows that $\mathrm{h}_n^+(\mathbf{x}) \ge \sum_{i=1}^{p_n} 1/(i+1) \ge \ln((p_n+1)/2) \ge \ln(n/4)$. □

4.3 Limit theorems for the height

We are now in position to establish Theorem 3.

Proof of Theorem 3. We start with (1). Recall that for an active vertex $u \in \mathbb{A}_n$ we always have $\mathrm{b}_n(u) = n$. Therefore, by (5)

$$\mathrm{H}(U^n) \overset{(d)}{=} \sum_{i=1}^{n} Y_i^n \mathbb{1}_{\{\mathrm{x}_i=1\}}.$$

Using Bennett's inequality (Proposition 14), for all $\varepsilon > 0$

$$\mathbb{P}\left(\left|\frac{\mathrm{H}(U^n)}{\mathrm{h}_n^+} - 1\right| > \varepsilon\right) \le 2\exp\big(-\mathrm{h}_n^+ g(\varepsilon)\big).$$

This upper bound goes to 0 as n tends to infinity since h_n^+ goes to infinity by Lemma 16. We deduce that the sequence $\big(\mathrm{H}(U^n)/\mathrm{h}_n^+ - 1\big)_n$ converges to 0 in probability. To conclude it suffices to show that it is bounded in $\mathbb{L}^p$ for all $p \ge 1$. Let $p \ge 1$, we have

$$\mathbb{E}\left[\left|\frac{\mathrm{H}(U^n)}{\mathrm{h}_n^+} - 1\right|^p\right] = p\int_0^\infty \mathbb{P}\left(\left|\frac{\mathrm{H}(U^n)}{\mathrm{h}_n^+} - 1\right| \ge u\right) u^{p-1}\,\mathrm{d}u.$$

Using Bennett's inequality again and by splitting the integral in half at $u = 7$, the above quantity is bounded by

$$7^p + 2p\int_7^\infty \exp\big(-\mathrm{h}_n^+ g(u)\big) u^{p-1}\,\mathrm{d}u.$$

Notice that for $u \ge 7 \ge e^2 - 1$, we have $g(u) \ge u$ thus

$$\int_7^\infty \exp\big(-\mathrm{h}_n^+ g(u)\big) u^{p-1} du \le \int_7^\infty \exp\big(-\mathrm{h}_n^+ u\big) u^{p-1}\,\mathrm{d}u = \frac{1}{(\mathrm{h}_n^+)^p}\int_{7\mathrm{h}_n^+}^\infty \exp(-u) u^{p-1}\,\mathrm{d}u.$$

By an iterated integration by parts, the last quantity is $\mathcal{O}\big(\exp(-7\mathrm{h}_n^+)/\mathrm{h}_n^+\big)$. Therefore the boundedness in $\mathbb{L}^p$ follows.

We now establish (2) and (3). Fix $C > 1$. By Proposition 15, for all $n \geq 1$:

$$\mathbb{P}\left(\frac{\mathsf{Height}(\mathcal{T}^n)}{\mathsf{h}_n^+} > C\right) \leq \exp\big(-\mathsf{h}_n^+ g(C-1) + \ln(n)\big).$$

This upper bound goes to 0 when $\ln(n) = o(\mathsf{h}_n^+)$ since $g(C-1) > 0$. It also goes to 0 when $C = e + \varepsilon$ since $g(e - 1 + \varepsilon) > 1$ and by Lemma 16, $\mathsf{h}_n^+ \geq \ln(n/4)$. Similarly,

$$\mathbb{P}\left(\frac{\mathsf{Height}(\mathcal{T}^n)}{\mathsf{h}_n^+} < 1 - \varepsilon\right) \leq \exp\big(-\mathsf{h}_n^+(\mathbf{x}) g(\varepsilon)\big) \quad \underset{n\to\infty}{\longrightarrow} \quad 0.$$

This completes the proof of (2). It also shows that the convergence

$$\frac{\mathsf{Height}(\mathcal{T}^n)}{\mathsf{h}_n^+} \quad \underset{n\to\infty}{\overset{\mathbb{P}}{\longrightarrow}} \quad 1$$

holds in probability when $\ln(n) = o(\mathsf{h}_n^+)$. To fully show (3) we just need to prove that the sequence $\big(\mathsf{Height}(\mathcal{T}^n)/\mathsf{h}_n^+ - 1\big)_n$ is bounded in $\mathbb{L}^p$ for all $p \geq 1$ when $\ln(n) = o(\mathsf{h}_n^+)$. It is done using the same method as in the proof of (1) (the extra n in the upper bound won't change the fact that it is indeed bounded in $\mathbb{L}^p$). □

5 Regime with a linear amount of active vertices

We shall now investigate the regime where the number of active vertices grows roughly linearly in the size of the trees: the main goal of this section is to establish Theorem 4. Here we assume that (2) is in force, namely there exist $c \in (0, 1]$ and a sequence $(A_n)_{n\in\mathbb{N}}$ of positive numbers such that $A_n = o(\log n)$ as $n \to \infty$ and it is assumed that

$$\lim_{\varepsilon\to 0} \limsup_{n\to\infty} \max_{A_n \leqslant i \leqslant \varepsilon n} \left|\frac{S_i^n}{i} - c\right| = 0 \qquad \text{and} \qquad \forall \varepsilon > 0, \ \liminf_{n\to\infty} \min_{\varepsilon n \leqslant i \leqslant n} \frac{S_i^n}{n} > 0.$$

5.1 An asymptotic equivalent for h_n^+

Here we show the convergence of Theorem 4 (1), namely:

$$\frac{\mathsf{h}_n^+}{\ln n} \quad \underset{n\to\infty}{\longrightarrow} \quad \frac{c+1}{2c}. \tag{9}$$

To simplify notation, for all $n \in \mathbb{N}$ and for all $\varepsilon > 0$, set

$$\eta_n(\varepsilon) \coloneqq \max_{A_n \leqslant i \leqslant \varepsilon n} \left|\frac{S_i^n}{i} - c\right|.$$

By (2), we know that $\limsup \eta_n(\varepsilon) \to 0$ when $\varepsilon \to 0$.

Lemma 17. *Suppose that the assumptions (2) hold with $A_n \to \infty$ when $n \to \infty$. Then, for all $C > 1$*

$$\lim_{\varepsilon\to 0} \limsup_{n\to\infty} \max_{\substack{A_n \leqslant a < b \leqslant \varepsilon n \\ aC < b}} \left|\frac{1}{\ln(b/a)} \sum_{i=a+1}^{b} \frac{1}{S_i^n} \mathbb{1}_{\{\mathrm{x}_i^n = 1\}} - \frac{1+c}{2c}\right| = 0. \tag{10}$$

Proof of Lemma 17. We prove (10) with a double bound. First, let us start with the upper bound. Let $\varepsilon \in (0, 1)$ and $\beta > 1$. For all $A_n \leq a < b \leq \varepsilon n$:

$$\sum_{i=a+1}^{b} \frac{1}{S_i^n} \mathbb{1}_{\{\mathrm{x}_i^n = 1\}} = \sum_{k=1+\ell}^{m} \sum_{\beta^{k-1} < i \leqslant \beta^k} \frac{1}{S_i^n} \mathbb{1}_{\{\mathrm{x}_i^n = 1\}} + \sum_{a < i \leqslant \beta^\ell} \frac{1}{S_i^n} \mathbb{1}_{\{\mathrm{x}_i^n = 1\}} + \sum_{\beta^m < i \leqslant b} \frac{1}{S_i^n} \mathbb{1}_{\{\mathrm{x}_i^n = 1\}}$$

where $\ell := \lceil \ln a/\ln\beta\rceil$, $m := \lfloor \ln b/\ln\beta\rfloor$ and where, in all the sums over i, it is implicit that i is an integer. Let $P_n := \{1 \le i \le n : \mathrm{x}_i = 1\}$ and $M_n := \{1 \le i \le n : \mathrm{x}_i = -1\}$. Note that for all $1 \le i < j \le n$, we have

$$\#(P_n \cap [i+1, j]) - \#(M_n \cap [i+1, j]) = S_j^n - S_i^n, \quad \#(P_n \cap [i+1, j]) + \#(M_n \cap [i+1, j]) = j - i.$$

Moreover, when $A_n \le i < j \le \varepsilon n$, we have $(j-i)c - 2j\eta_n(\varepsilon) \le S_j^n - S_i^n \le (j-i)c + 2j\eta_n(\varepsilon)$. As a result,

$$\frac{c+1}{2}(j-i) - j\eta_n(\varepsilon) \le \#(P_n \cap [i+1, j]) = \frac{S_j^n - S_i^n + j - i}{2} \le \frac{c+1}{2}(j-i) + j\eta_n(\varepsilon). \tag{11}$$

Therefore, for all integer $1 + \ell \le k \le m$

$$\begin{aligned}\sum_{\beta^{k-1}<i\leqslant\beta^k} \frac{1}{S_i^n}\mathbb{1}_{\{\mathrm{x}_i^n=1\}} &\le \sum_{\beta^{k-1}<i\leq\beta^k} \frac{1}{i(c-\eta_n(\varepsilon))}\mathbb{1}_{\{\mathrm{x}_i^n=1\}} \le \#\{\beta^{k-1} < i \le \beta^k : \mathrm{x}_i^n = 1\}\frac{1}{\beta^{k-1}(c-\eta_n(\varepsilon))} \\ &\le \left(\frac{c+1}{2}(\beta^k - \beta^{k-1}{+}1) + \beta^k\eta_n(\varepsilon)\right)\frac{1}{\beta^{k-1}(c-\eta_n(\varepsilon))} \\ &= \left(\frac{c+1}{2}(\beta - 1 + 1/A_n) + \beta\eta_n(\varepsilon)\right)\frac{1}{(c-\eta_n(\varepsilon))} =: U_n(\varepsilon)\end{aligned}$$

where the third inequality comes from (11) and where we used that $\beta^{k-1} \ge \beta^\ell \ge a \ge A_n$. Thus

$$\sum_{i=a+1}^{b} \frac{1}{S_i^n}\mathbb{1}_{\{\mathrm{x}_i^n=1\}} \le (m-\ell)U_n(\varepsilon) + \frac{\beta^\ell - \beta^{\ell-1}}{\beta^{\ell-1}(c-\eta_n(\varepsilon))} + \frac{\beta^{m+1} - \beta^m}{\beta^m(c-\eta_n(\varepsilon))} \le \frac{\ln(b/a)}{\ln\beta}U_n(\varepsilon) + \frac{2(\beta-1)}{c-\eta_n(\varepsilon)}.$$

For the lower bound, the same ideas apply. Indeed, for every integer $1 + \ell \le k \le m$ we have

$$\begin{aligned}\sum_{\beta^{k-1}<i\leqslant\beta^k} \frac{1}{S_i^n}\mathbb{1}_{\{\mathrm{x}_i^n=1\}} &\ge \#\{\beta^{k-1} < i \le \beta^k : \mathrm{x}_i^n = 1\}\frac{1}{\beta^k(c+\eta_n(\varepsilon))} \\ &\ge \left(\frac{c+1}{2}\left(\beta^k - \beta^{k-1} - 2\right) - \beta^k\eta_n(\varepsilon)\right)\frac{1}{\beta^k(c+\eta_n(\varepsilon))} \\ &\ge \left(\frac{c+1}{2}\left(1 - \frac{1}{\beta} - \frac{2}{\beta A_n}\right) - \eta_n(\varepsilon)\right)\frac{1}{(c+\eta_n(\varepsilon))} =: L_n(\varepsilon)\end{aligned}$$

where the second inequality comes from (11) and the third one comes from $\beta^k \ge \beta^{\ell+1} \ge \beta a \ge \beta A_n$. Thus

$$\sum_{i=a+1}^{b} \frac{1}{S_i^n}\mathbb{1}_{\{\mathrm{x}_i^n=1\}} \ge (m-\ell)L_n(\varepsilon) \ge \frac{\ln(b/a)}{\ln\beta}L_n(\varepsilon) - 2L_n(\varepsilon) \ge \frac{\ln(b/a)}{\ln\beta}L_n(\varepsilon) - \frac{c+1}{c}\left(1 - \frac{1}{\beta} - \frac{2}{\beta A_n}\right).$$

We end up with the following inequalities

$$\frac{L_n(\varepsilon)}{\ln\beta} - \frac{c+1}{c\ln C}\left(1 - \frac{1}{\beta} - \frac{2}{\beta A_n}\right) \le \frac{1}{\ln(b/a)}\sum_{i=a+1}^{b} \frac{1}{S_i^n}\mathbb{1}_{\{\mathrm{x}_i^n=1\}} \le \frac{U_n(\varepsilon)}{\ln\beta} + \frac{2(\beta-1)}{(c-\eta_n(\varepsilon))\ln C}.$$

Notice that $\limsup U_n(\varepsilon) \to (\beta-1)(c+1)/(2c)$ and $\liminf L_n(\varepsilon) \to (1-1/\beta)(c+1)/(2c)$ when $\varepsilon \to 0$. Therefore

$$\limsup_{\varepsilon\to 0}\limsup_{n\to\infty} \max_{\substack{A_n\leqslant a<b\leqslant\varepsilon n\\ aC<b}} \left|\frac{1}{\ln(b/a)}\sum_{i=a+1}^{b} \frac{1}{S_i^n}\mathbb{1}_{\{\mathrm{x}_i^n=1\}} - \frac{1+c}{2c}\right| \le \frac{1+c}{2c}\left|\frac{\beta-1}{\ln\beta} - 1\right| + \frac{c+1}{c\ln C}\left(1 - \frac{1}{\beta}\right) + \frac{2(\beta-1)}{c\ln C}.$$

Making β tend to 1 gives the result. □

Proof of Theorem 4 (1). Let $(A'_n)_{n\in\mathbb{N}}$ be a sequence of positive integers such that $A_n \le A'_n$, $A'_n = o(\ln n)$ and $A'_n \to \infty$ as $n \to \infty$. Applying Lemma 17 we see that

$$\lim_{\varepsilon\to 0}\limsup_{n\to\infty}\left|\frac{1}{\ln n}\sum_{A'_n<i\leqslant\varepsilon n} \frac{1}{S_i^n}\mathbb{1}_{\{\mathrm{x}_i^n=1\}} - \frac{1+c}{2c}\right| = 0$$

Since $S_i^n \geq 1$ for all $1 \leq i \leq n$ and $A'_n = o(\ln n)$, we have

$$\sum_{i=1}^{A'_n} \frac{1}{S_i^n} \mathbb{1}_{\{x_i^n=1\}} = o(\ln n).$$

Finally, for all $0 < \varepsilon < 1$, by the second assumption of (2):

$$\sum_{\varepsilon n \leqslant i \leqslant n} \frac{1}{S_i^n} \mathbb{1}_{\{x_i^n=1\}} = \mathcal{O}(1).$$

Combining everything gives the desired convergence (9). □

5.2 Distances between uniform vertices

Here we prove Theorem 4 (2). We keep the notation introduced in Section 2.3: for a vertex $v \in \mathbb{V}_n$, let $b_n(v)$ be the largest $i \in \{0, 1, \dots, n\}$ such that v belongs to the forest $\mathcal{F}_i^n$ obtained when building $\mathcal{T}^n$ using Algorithm 2. We first focus on $H(V_1^n)$ and introduce some notation. Set

$$Z_n := \sum_{i=1}^{b_n(V_1^n)} Y_i^n 1_{\{x_i^n=1\}} \qquad \text{and} \qquad M_n := \sum_{i=1}^{b_n(V_1^n)} \frac{1}{S_i^n} \mathbb{1}_{\{x_i^n=1\}}$$

where $(Y_i^n)_{1\leq i\leq n}$ are independent Bernoulli random variables of respective parameters $(1/S_i^n)_{1\leq i\leq n}$ (independent from $b_n(V_1^n)$). Observe that $M_n = \mathbb{E}\big[Z_n | b_n(V_1^n)\big]$. By (5), we have

$$H(V_1^n) \overset{(d)}{=} Z_n. \tag{12}$$

Fix $p \geq 1$. First we establish that M_n/h_n^+ converges to 1 in $\mathbb{L}^p$. Actually, since $0 \leq M_n \leq h_n^+$ almost surely it is enough to show the convergence in probability.

Let $\varepsilon > 0$ small enough so that $\eta_n(\varepsilon) \leq c/2$ for all n large enough. Let $\gamma_\varepsilon > 0$ such that for all n large enough we have

$$\gamma_\varepsilon \leq \min_{\varepsilon n \leq i \leq n} \frac{S_i^n}{n}.$$

Such a γ_ε exists thanks to (2). Besides, since $A_n = o(\ln n) = o(n)$, by Lemma 6 we have $b_n(V_1^n) \geq A_n$ with high probability. Using this, as well as the fact that $h_n^+ \geq \ln(n/4)$ (see Lemma 16), we obtain, with high probability as $n \to \infty$,

$$\begin{aligned} 0 \leq 1 - \frac{M_n}{h_n^+} = \frac{1}{h_n^+} \sum_{i=b_n(V_1^n)}^{n} \frac{1}{S_i^n} \mathbb{1}_{\{x_i^n=1\}} &\leq \frac{1}{h_n^+} \sum_{i=b_n(V_1^n)}^{n} \left(\frac{1}{i(c-\eta_n(\varepsilon))} + \frac{1}{\gamma_\varepsilon n} \right) \mathbb{1}_{\{x_i^n=1\}} \\ &\leq \frac{1}{\ln(n/4)} \frac{2}{c} \ln\left(\frac{n}{b_n(V_1^n)} \right) + \frac{1}{\gamma_\varepsilon h_n^+}. \end{aligned}$$

Now, fix $0 < \delta < 1$ and take $m_n = n^{1-\delta} 2^\delta$. Then by Lemma 6

$$\mathbb{P}\left(\ln\left(\frac{n}{b_n(V_1^n)} \right) > \delta \ln\left(\frac{n}{2} \right) \right) = \frac{m_n + 1 - S_{m_n}^n}{n + 1 + S_n^n} \xrightarrow[n\to\infty]{} 0.$$

We conclude that M_n/h_n^+ converges towards 1 in probability (and therefore in $\mathbb{L}^p$ for all $p \geq 1$). Now we show that Z_n/M_n converges to 1 in $\mathbb{L}^p$ following the same spirit as in the proof of Theorem 3. We have

$$\mathbb{E}\left[\left| \frac{Z_n}{M_n} - 1 \right|^p \,\middle|\, b_n(V_1^n) \right] = p \int_0^\infty \mathbb{P}\left(\left| \frac{Z_n}{M_n} - 1 \right| \geq u \,\middle|\, b_n(V_1^n) \right) u^{p-1} du.$$

By Bennett's inequality (Proposition 14)

$$\mathbb{P}\left(\left|\frac{Z_n}{M_n}-1\right|\geq u\;\middle|\;\mathsf{b}_n(V_1^n)\right)\leq 2\exp(-M_n g(u))$$

where $g(u)=(u+1)\ln(u+1)-u$. Since M_n is of order $\ln n$ it is a simple matter to check that

$$\mathbb{E}\left|\frac{Z_n}{M_n}-1\right|^p\leq 2p\mathbb{E}\left[\int_0^\infty \exp(-M_n g(u))u^{p-1}du\right]\xrightarrow[n\to\infty]{}0.$$

Since $M_n\leq \mathsf{h}_n^+$ we deduce that

$$\mathbb{E}\left|\frac{Z_n-M_n}{\mathsf{h}_n^+}\right|^p\leq \mathbb{E}\left|\frac{Z_n}{M_n}-1\right|^p\xrightarrow[n\to\infty]{}0.$$

So $(Z_n-M_n)/\mathsf{h}_n^+$ converges to 0 in $\mathbb{L}^p$. Since M_n/h_n^+ converges to 1 in $\mathbb{L}^p$, Z_n/h_n^+ goes to 1 as well. Finally, using Theorem 4 (1), we conclude that $Z_n/\ln n$ converges to $(c+1)/(2c)$ in $\mathbb{L}^p$. This proves the first convergence of Theorem 4 (2).

So as to deduce the asymptotic behaviour of the distance between two uniform points, we next show that two uniform random vertices coalesce near the root.

Lemma 18. *Assume that the assumptions of (2) are in force. For all $n\geq 1$, let V_1^n, V_2^n be two independent uniform vertices of $\mathbb{V}_n$. Then*

$$\frac{\mathsf{c}_n(V_1^n,V_2^n)}{n}\xrightarrow[n\to\infty]{\mathbb{P}}0.$$

Proof. Let $0<\varepsilon<1$. Let $i,j\leq n-1$ distinct with $\mathsf{x}_{i+1}^n=\mathsf{x}_{j+1}^n=-1$ and $k\leq i\wedge j$ with $\mathsf{x}_{k+1}^n=1$. Denote by B the event "$\mathsf{b}_n(V_1^n)=i$ and $\mathsf{b}_n(V_2^n)=j$" and by C the event "$\mathsf{c}_n(V_1^n,V_2^n)=k$". By Lemma 7, it is clear that

$$\mathbb{P}(C\,|\,B)=\mathbb{P}\big(C\text{ and }\mathsf{c}_n(V_1^n,V_2^n)\leq k\,|\,B\big)\leq\mathbb{P}\big(C\,|\,\mathsf{c}_n(V_1^n,V_2^n)\leq k,B\big)=\binom{S_{k+1}^n}{2}^{-1}.$$

Therefore

$$\mathbb{P}\big(\mathsf{c}_n(V_1^n,V_2^n)\geq\varepsilon n\,|\,B\big)\leq\sum_{k=\lceil\varepsilon n\rceil}^{i\wedge j}\binom{S_{k+1}^n}{2}^{-1}\mathbb{1}_{\{\mathsf{x}_{k+1}^n=1\}}\leq\sum_{k=\lceil\varepsilon n\rceil}^{n-1}\binom{S_{k+1}^n}{2}^{-1}\mathbb{1}_{\{\mathsf{x}_{k+1}^n=1\}}.$$

And

$$\mathbb{P}\big(\mathsf{c}_n(V_1^n,V_2^n)\geq\varepsilon n\big)\leq\sum_{k=\lceil\varepsilon n\rceil}^{n-1}\binom{S_{k+1}^n}{2}^{-1}\mathbb{1}_{\{\mathsf{x}_{k+1}^n=1\}}.$$

Besides, by the second hypothesis of (2), we know that

$$\sum_{k=\lceil\varepsilon n\rceil}^{n-1}\binom{S_{k+1}^n}{2}^{-1}\mathbb{1}_{\{\mathsf{x}_{k+1}^n=1\}}=\mathcal{O}\left(\frac{1}{n}\right)\qquad\text{as}\qquad n\to\infty,$$

hence the desired result. □

The second convergence of the point (2) of Theorem 4 is then an immediate consequence of the first convergence of the same point using a dominated convergence and Lemma 18.

5.3 Limit theorem for the height

Here we show Theorem 4 (3). The proof of the upper bound is readily obtained by a union bound, while the lower bound is much more delicate.

Proof of the upper bound in Theorem 4 (3). Fix $\eta > 0$. First observe that by definition of $f(c)$ we have

$$g(f(c)-1) = f(c)\ln f(c) - f(c) + 1 = 1 + \frac{c-1}{1+c} = \frac{2c}{1+c},$$

so that $g(f(c)-1+\eta) > 2c/(1+c)$. By Theorem 4 (1) we have $\mathsf{h}_n^+ \sim (1+c)/(2c)\ln(n)$ as $n \to \infty$. As a consequence, there exists $\delta \in (0,1)$ such that $h_n^+ g(f(c)-1+\eta) \geq (1+\delta)\ln(n)$ for n sufficiently large. Hence, by Proposition 15, for n sufficiently large,

$$\mathbb{P}\big(\mathsf{Height}(\mathcal{T}^n) > \mathsf{h}_n^+(f(c)+\eta)\big) \leq \exp\big(-\mathsf{h}_n^+ g(f(c)-1+\eta) + \ln n\big) \leq \exp(-\delta\ln(n)) \underset{n\to\infty}{\longrightarrow} 0.$$

This completes the proof of the upper bound. □

It therefore remains to show the lower bound on $\mathsf{Height}(\mathcal{T}_n)$, which is the delicate part of the proof. Let us explain the main idea of our approach. Since the upper-bound has been obtained by using a union-bound over all vertices, one would hope to obtain the lower-bound from the fact that the height of the vertices are "almost independent". However, their heights are highly correlated. To overcome this issue, we use the so-called chaining technique and estimate by induction on k the height of the subtrees $\mathsf{Height}(\mathcal{T}_k(\mathbf{x}^n))$. More precisely, for every increasing sequence of integers (chain) $0 = R_0 < R_1 < \cdots < R_m = n$, we have

$$\mathsf{Height}(\mathcal{T}_n) = \sum_{0\leq i<m} \mathsf{Height}(\mathcal{T}_{R_{i+1}}(\mathbf{x}^n)) - \mathsf{Height}(\mathcal{T}_{R_i}(\mathbf{x}^n)).$$

We shall choose an appropriate chain of integers $(R_i)_{0\leq i\leq m}$ and then bound the differences appearing in the last display. To this end, the main idea is that if R_{i+1}/R_i is close to some large constant $C > 0$, then $\mathcal{T}_{R_{i+1}}(\mathbf{x}^n)$ can be approximated by a forest of independent trees attached on $\mathcal{T}_{R_i}(\mathbf{x}^n)$. We then show by induction on i that many vertices have a large height in this forest, and among them many are attached on vertices of $\mathcal{T}_{R_i}(\mathbf{x}^n)$ with large height.

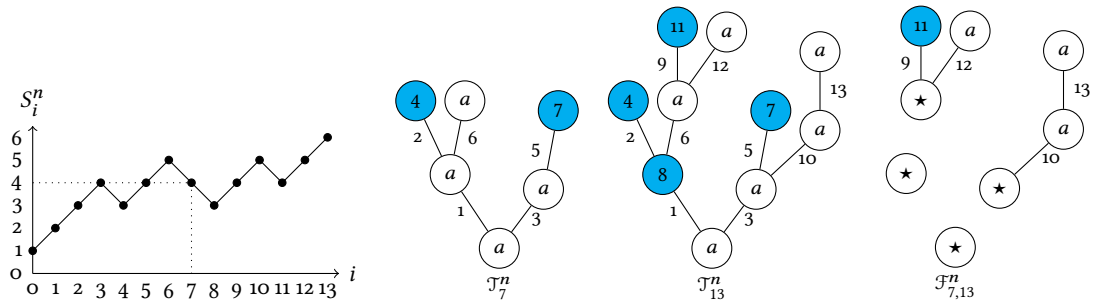

Figure 9: Illustration of the construction of $\mathcal{F}_{k,k'}^n$ with $k = 7$ and $k' = 13$. From left to right: the sequence $\mathbf{x}^n$, a realization of $\mathcal{T}_k^n$, a realization of $\mathcal{T}_{k'}^n$ and the forest $\mathcal{F}_{k,k'}^n$ composed of four trees all rooted at their unique vertex labelled $\star$.

To make this intuition rigorous we first need to define some objects. For $n \in \mathbb{N}$ and $0 \leq k \leq n$, recall the definition of $\mathcal{T}_k^n := \mathcal{T}_k(\mathbf{x}^n)$ from Algorithm 1. We denote by $\mathcal{A}_k^n$ the set of all vertices in $\mathcal{T}_n^n$ that were active at time k when running Algorithm 1 (these are vertices of $\mathcal{T}_k^n$ that are either labelled with a and have an adjacent edge labelled by $j \leq k$, or labelled with $i > k$ and have an adjacent edge labelled by $j \leq k$). For $0 \leq k \leq k' \leq n$, let $\mathcal{F}_{k,k'}^n$ be the forest obtained as follows:

(i) Consider $\mathcal{T}^n_{k'}$ and relabel by $\star$ all the vertices of $\mathcal{T}^n_{k'}$ belonging to $\mathcal{A}^n_k$ (see Fig. 9 for an example);
(ii) then remove all the edges of $\mathcal{T}^n_{k'}$ that belongs to $\mathcal{T}^n_k$ (i.e. all the edges with labels in $[\![1,k]\!]$);
(iii) finally remove all the vertices with labels in $[\![1,k]\!]$.

Note that all the trees in $\mathcal{F}^n_{k,k'}$ have exactly one vertex labelled $\star$. We will thus consider the vertices labelled $\star$ as the roots of the trees in $\mathcal{F}^n_{k,k'}$. Observe that one cannot deterministically reconstruct $\mathcal{T}^n_{k'}$ from $\{\mathcal{F}^n_{k,k'}, \mathcal{T}^n_k\}$, but in distribution it is possible: if one merges uniformly at random every active vertex of $\mathcal{T}^n_{k'}$ with one vertex labelled $\star$ from $\mathcal{F}^n_{k,k'}$, then one obtains a tree having the same distribution as $\mathcal{T}^n_{k'}$ (see Fig. 9). Furthermore, it will be crucial to keep in mind that that $\mathcal{T}^n_k$ and $\mathcal{F}^n_{k,k'}$ are independent. Indeed $\mathcal{F}^n_{k,k'}$ may be constructed following Algorithm 1, independently of the steps $1 \le i \le k$.

For every $0 \le k \le k' \le n$, for every vertex $v \in \mathcal{F}^n_{k,n}$ we write $H^n_{k,k'}(v)$ for the height of v in the forest $\mathcal{F}^n_{k,n}$ minus its height in $\mathcal{F}^n_{k',n}$ (where, by convention, the height of v in $\mathcal{F}^n_{k',n}$ is 0 if v is not in this forest). In words, if v belongs to a certain tree τ of $\mathcal{F}^n_{k',n}$, the quantity $H^n_{k,k'}(v)$ represents the height in $\mathcal{F}^n_{k,k'}$ of the vertex associated with the root τ.

For $\eta > 0$, for every integers $n, k, C \in \mathbb{N}$ such that $C^k \le n/C$, set

$$N^n_k(C,\eta) = \#\left\{v \in \mathcal{A}^n_{C^{k+1}} : H^n_{C^k,C^{k+1}}(v) > \frac{1+c}{2c}\ln(C)(f(c)-\eta)\right\}.$$

The set of vertices involved in the definition of N^n_k will play an important role in our approach. Roughly speaking, they correspond to vertices active at time C^{k+1} which are "quite far" from a vertex active at time C^k. The first main input in the proof of the lower bound of the height is the following result, which shows that $C > 0$ can be chosen such that if R_{i+1}/R_i is close to $C > 0$, then there are many active vertices with a "large" height in the forest $\mathcal{F}^n_{R_i,R_{i+1}}$. Its proof is deferred to Section 5.4.

Lemma 19. *For every $\eta \in (0,1)$, there exists an integer $C > 1$, there exist $\lambda > 0$ and $\varepsilon \in (0,1)$ such that for every n large enough, for every $k \in \mathbb{N}$, with $A_n \le C^k \le \varepsilon n/C$, we have*

$$\mathbb{P}\left(N^n_k(C,\eta) > 4C^k\right) \ge 1 - \frac{\lambda}{C^k}$$

For every $n, k \in \mathbb{N}$, with $C^k \le n/C$, let M^n_k be the maximal number of active vertices of a tree in the forest $\mathcal{F}^n_{C^k,C^{k+1}}$. The second main input in the proof of the lower bound of the height is the following result, which shows that the this quantity cannot be too large. Its proof is deferred to Section 5.4.

Lemma 20. *For every n large enough, for every $k \in \mathbb{N}$, with $A_n \le C^k \le n/C$,*

$$\mathbb{P}\left(M^n_k > k^3\right) \le \frac{1}{k^2}.$$

Let us now explain how the lower bound for the height follows from Lemmas 19 and 20.

Proof of the lower-bound for Theorem 4 (3). Fix $\eta \in (0,1)$ and take $C > 1, \lambda > 0, \varepsilon \in (0,1)$ such that the conclusion of Lemma 19 holds and write N^n_k instead of $N^n_k(C,\eta)$ to simplify notation. For all integers $n, k \in \mathbb{N}$ such that $A_n \le C^k \le n/C$, set

$$\overline{N}^n_k = \#\left\{v \in \mathcal{A}^n_{C^k}, \quad \forall \ell \in [\![1,k-1]\!] \text{ s.t. } A_n \le C^\ell, \quad H^n_{C^\ell,C^{\ell+1}}(v) > \frac{1+c}{2c}\ln(C)(f(c)-\eta)\right\}. \tag{13}$$

We shall show that with probability tending to 1 as $n \to \infty$, if k is the largest integer such that $C^k \le \varepsilon n/C$, then $\overline{N}^n_k \ge 1$. Indeed, the fact that $\overline{N}^n_k \ge 1$ implies that there exists a vertex $v_k \in A^n_{C^k}$ and vertices $v_i \in A^n_{C^i}$

for every $i \in [\![1, k-1]\!]$ such that $A_n \le C^i$, v_i is an ancestor of v_{i+1} and $d^n(v_i, v_{i+1}) > \frac{1+c}{2c}\ln(C)(f(c)-\eta)$. Thus the height of v_k in $\mathcal{T}^n$ is at least

$$\left(\left\lfloor \frac{\ln(n)}{\ln(C)} + \frac{\ln(\varepsilon/C)}{\ln(C)} \right\rfloor - \left\lceil \frac{\ln(A_n)}{\ln(C)} \right\rceil\right) \cdot \frac{1+c}{2c}\ln(C)\cdot(f(c)-\eta) \quad \underset{n\to\infty}{\sim} \quad \ln(n)\frac{1+c}{2c}(f(c)-\eta).$$

Since $\eta \in (0,1)$ was chosen arbitrary, this will imply the desired result.

The main step of the proof is to establish that for every $k \in \mathbb{N}$ with $A_n \le C^k \le \varepsilon n/C$ we have

$$\mathbb{P}\left(\overline{N}^n_{k+1} < 2^{k+1} \middle| \overline{N}^n_k \ge 2^k\right) = \mathcal{O}\left(\frac{1}{k^2}\right), \tag{14}$$

where here, and in the rest of the proof, the $\mathcal{O}$ is uniform in $A_n \le C^k \le \varepsilon n/C$. This means that there exists a constant $\Delta > 0$ such that the term $\mathcal{O}(1/k^2)$ can be bounded from above by Δ/k^2 for every $A_n \le C^k \le \varepsilon n/C$ and n large enough. Indeed, we will then have

$$\mathbb{P}\left(\exists k \text{ with } A_n \le C^k \le \varepsilon n/C \text{ and } \overline{N}^n_k < 2^k\right) \le \mathbb{P}\left(N^n_{k_n} < 2^{k_n}\right) + \mathcal{O}\left(\sum_{k=k_n}^{\infty}\frac{1}{k^2}\right)$$

with $k_n = \lceil \ln(A_n)/\ln(C)\rceil$. By Lemma 19 the first term of the sum tends to 0. This implies that with probability tending to 1 as $n \to \infty$, for every integer k such that $A_n \le C^k \le \varepsilon n/C$, we have $\overline{N}^n_k \ge 2^k$. A fortiori, with probability tending to 1 as $n \to \infty$, k is the largest integer such that $C^k \le \varepsilon n/C$, when $\overline{N}^n_k \ge 1$, which will complete the proof.

It thus remains to establish (14). Note that the random variable $\overline{N}^n_k$ is $\mathcal{T}^n_{C^k}$ measurable, and so is independent of the random variables N^n_k and M^n_k which are $\mathcal{F}^n_{C^k,C^{k+1}}$ measurable. As a consequence, we get for every $k \in \mathbb{N}$ with $A_n \le C^k \le \varepsilon n/C$:

$$\begin{aligned}\mathbb{P}\left(\overline{N}^n_{k+1} < 2^{k+1} \middle| \overline{N}^n_k \ge 2^k\right) &\le \mathbb{P}\left(N^n_k < 4C^k\right) + \mathbb{P}\left(M^n_k > k^3\right) + \mathbb{P}\left(\overline{N}^n_{k+1} < 2^{k+1}, N^n_k \ge 4C^k, M^n_k \le k^3 \middle| \overline{N}^n_k \ge 2^k\right)\\ &\le \mathcal{O}\left(\tfrac{1}{C^k}\right) + \mathcal{O}\left(\tfrac{1}{k^2}\right) + \mathbb{P}\left(\overline{N}^n_{k+1} < 2^{k+1}, N^n_k \ge 4C^k, M^n_k \le k^3 | \overline{N}^n_k \ge 2^k\right),\end{aligned} \tag{15}$$

where we have used Lemmas 19 and 20.

To bound the third term, we shall estimate $\overline{N}^n_k$ by induction on k by using a second moment technique. To this end, we recall that one may reconstruct a tree with the same distribution as $\mathcal{T}^n_{C^{k+1}}$, by merging uniformly at random each active vertex from $\mathcal{T}^n_{C^k}$ with one vertex labelled $\star$ from $\mathcal{F}^n_{C^k,C^{k+1}}$. With this in mind, it follows that

$$\mathbb{E}\left[\overline{N}^n_{k+1} \middle| \mathcal{T}^n_{C^k}, \mathcal{F}^n_{C^k,C^{k+1}}\right] = \frac{\overline{N}^n_k N^n_k}{S^n_{C^k}}, \tag{16}$$

where we recall that $S^n_{C^k} = S_{C^k}(\mathbf{x}^n)$ represents the number of active vertices of $\mathcal{T}^n_{C^k}$.

Next, denote by $\mathcal{A}^n_{C^k,C^{k+1}}$ the set of all active vertices of the forest $\mathcal{F}^n_{C^k,C^{k+1}}$. For $v \in \mathcal{A}^n_{C^k,C^{k+1}}$, let $E^n_k(v)$ be the event defined by

$$E^n_k(v) = \left\{\forall \ell \in [\![1, k-1]\!] \text{ s.t. } A_n \le C^\ell, \quad H^n_{C^\ell,C^{\ell+1}}(v) > \frac{1+c}{2c}\ln(C)(f(c)-\eta)\right\}.$$

Note that for all $v, v' \in \mathcal{A}^n_{C^k,C^{k+1}}$, conditionally on $\mathcal{T}^n_{C^k}$ and $\mathcal{F}^n_{C^k,C^{k+1}}$, the events $E^n_{k+1}(v)$ and $E^n_{k+1}(v')$ are negatively correlated unless v, v' are in the same tree in $\mathcal{F}^n_{C^k,C^{k+1}}$. Indeed, on the one hand, if the trees of $\mathcal{F}^n_{C^k,C^{k+1}}$ containing v and v' are different, then when we attach the tree containing v to an active vertex $w \in \mathcal{A}^n_{C^k}$ such that the event $E^n_k(w)$ happens then the tree containing v' will have less chance to be attached to a vertex w' in $\mathcal{A}^n_{C^k}$ such that the event $E^n_k(w')$ occurs, and vice versa. On the other hand, if v and v' are in the same tree and if we have the inequalities $H^n_{C^k,C^{k+1}}(v) > \frac{1+c}{2c}\ln(C)(f(c)-\eta)$ and

$H^n_{C^k,C^{k+1}}(v') > \frac{1+c}{2c}\ln(C)(f(c)-\eta)$, if we denote by w the (random) active vertex of $\mathcal{T}^n_{C^k}$ on which this tree is attached, then the events $E^n_{k+1}(v)$ and $E^n_{k+1}(v')$ occur if and only if the event $E^n_k(w)$ occurs. As a result, writing $\widetilde{\mathbb{P}}$ and $\widetilde{\mathbb{E}}$ for the conditional probability and expectation conditionally given $\mathcal{T}^n_{C^k}$ and $\mathcal{F}^n_{C^k,C^{k+1}}$, we get

$$\widetilde{\mathcal{V}\mathrm{ar}}\left[\overline{N}^n_{k+1}\right] = \widetilde{\mathcal{V}\mathrm{ar}}\left[\sum_{v\in\mathcal{A}^n_{C^k,C^{k+1}}} \mathbb{1}_{E^n_{k+1}(v)}\right] = \sum_{v,v'\in\mathcal{A}^n_{C^k,C^{k+1}}} \widetilde{\mathcal{C}\mathrm{ov}}(\mathbb{1}_{E^n_{k+1}(v)}, \mathbb{1}_{E^n_{k+1}(v')}).$$

$$\leq M^n_k \sum_{v\in\mathcal{A}^n_{C^k,C^{k+1}}} \widetilde{\mathbb{P}}\big(E^n_{k+1}(v)\big) = M^n_k \cdot \frac{\overline{N}^n_k N^n_k}{S^n_{C^k}}. \tag{17}$$

Then, by (16) under the event $\left\{\overline{N}^n_k \geq 2^k, N^n_k \geq 4C^k, M^n_k \leq k^3\right\}$ (which is measurable with respect to $\mathcal{T}^n_{C^k}, \mathcal{F}^n_{C^k,C^{k+1}}$) we have:

$$\widetilde{\mathbb{E}}\left[\overline{N}^n_{k+1}\right] \geq \frac{\overline{N}^n_k N^n_k}{S^n_{C^k}} \geq 2^{k+2},$$

and by (17) we have

$$\widetilde{\mathbb{P}}\left(\overline{N}^n_{k+1} < 2^{k+1}\right) \leq \widetilde{\mathbb{P}}\left(\overline{N}^n_{k+1} < \frac{1}{2}\widetilde{\mathbb{E}}\left[\overline{N}^n_{k+1}\right]\right) \leq 4\frac{\widetilde{\mathcal{V}\mathrm{ar}}\left[\overline{N}^n_{k+1}\right]}{\widetilde{\mathbb{E}}\left[\overline{N}^n_{k+1}\right]^2} \leq 4\frac{M^n_k \overline{N}^n_k N^n_k / S^n_{C^k}}{\left(\overline{N}^n_k N^n_k / S^n_{C^k}\right)^2} \leq \frac{4M^n_k S^n_{C^k}}{N^n_k N^n_k} \leq \frac{4k^3 C^k}{2^k C^k} = \mathcal{O}\left(\frac{1}{k^2}\right).$$

As a consequence,

$$\mathbb{P}\left(\overline{N}^n_{k+1} < 2^{k+1}, N^n_k \geq 4C^k, M^n_k \leq k^3 \middle| \overline{N}^n_k \geq 2^k\right) = \mathbb{E}\left[\widetilde{\mathbb{P}}\left(\overline{N}^n_{k+1} < 2^{k+1}\right)\mathbb{1}_{N^n_k \geq 4C^k, M^n_k \leq k^3} \middle| \overline{N}^n_k \geq 2^k\right] = \mathcal{O}\left(\frac{1}{k^2}\right),$$

and (14) follows from (15). This completes the proof. □

5.4 Proof of Lemmas 19 and 20

We keep the notation introduced in Section 5.3. The following estimate is the first step for the proof of Lemma 19.

Lemma 21. *For every $\eta \in (0,1)$, there exists an integer $C > 1$ and $\varepsilon \in (0,1)$ such that for every n large enough for every $k \in \mathbb{N}$, with $A_n \leq C^k \leq \varepsilon n/C$, we have*

$$\mathbb{E}\left[N^n_k(C,\eta)\right] \geq 5C^k.$$

Proof of Lemma 21. Recall that $g(x) = (1+x)\ln(1+x) - x$. Set $\delta = 1 - \frac{1+c}{2c}\cdot g(f(c)-1-\eta)$, which belongs to $(0,1)$ since $g(f(c)-1-\eta) < (2c)/(1+c)$, and then choose an integer $C > 1$ large enough such that

$$\frac{1}{\ln(C)^4}\exp(-(1-\delta)\ln(C)) \geq \frac{6}{cC} \qquad \text{and} \qquad \ln(C) \geq \frac{1+c}{2c}(f(c)-\eta). \tag{18}$$

It remains to check that this value of C satisfies the desired conclusion. By Lemma 5 for every $v \in \mathcal{A}^n_{C^{k+1}}$,

$$H^n_{C^k,C^{k+1}}(v) \overset{(d)}{=} \sum_{C^k < i \leq C^{k+1}} Y^n_i,$$

where $(Y^n_i)_{1\leq i\leq n}$ are independent Bernoulli random variables of respective parameters $(1/S^n_i \mathbb{1}_{\mathrm{x}^n_i=1})_{1\leq i\leq n}$. Lemma 17 yields

$$\lim_{\varepsilon\to 0}\limsup_{n\to\infty}\max_{A_n\leq C^k\leq \varepsilon n}\left|\sum_{C^k<i\leq C^{k+1}}\frac{1}{S^n_i}\mathbb{1}_{\mathrm{x}^n_i=1} - \frac{1+c}{2c}\ln(C)\right| \underset{n\to\infty}{\longrightarrow} 0.$$

Hence, by [6, Theorem 1], we have

$$\lim_{\varepsilon\to 0}\limsup_{n\to\infty}\max_{A_n\le C^k\le \varepsilon n} d_{\mathrm{TV}}\left(H^n_{C^k,C^{k+1}}(v), \mathcal{P}\mathrm{oi}\left(\frac{1+c}{2c}\ln(C)\right)\right)=0,$$

where $\mathcal{P}\mathrm{oi}(\lambda)$ denotes a Poisson random variable with parameter λ and d_{TV} stands for the total variation distance. As a consequence, setting

$$P(C) := \mathbb{P}\left(\mathcal{P}\mathrm{oi}\left(\frac{1+c}{2c}\ln(C)\right) > \frac{1+c}{2c}\ln(C)(f(c)-\eta)\right),$$

we have

$$\lim_{\varepsilon\to 0}\limsup_{n\to\infty}\max_{A_n\le C^k\le \varepsilon n}\left|\mathbb{P}\left(H^n_{C^k,C^{k+1}}(v) > \frac{1+c}{2c}\ln(C)(f(c)-\eta)\right) - P(C)\right| = 0. \tag{19}$$

Then, for $\lambda > 2$ and $t > 1$, using the bound $\lceil u\rceil! \le u^3(u/e)^u$ for $u \ge 2$, observe that

$$\mathbb{P}(\mathcal{P}\mathrm{oi}(\lambda) > \lambda t) \ge e^{-\lambda}\frac{\lambda^{\lceil \lambda t\rceil}}{\lceil \lambda t\rceil!} \ge e^{-\lambda}\lambda^{\lambda t}\left(\frac{e}{\lambda t}\right)^{\lambda t}\frac{1}{(\lambda t)^3} = e^{-\lambda g(t-1)}\frac{1}{(\lambda t)^3},$$

Taking into account our choice of C (see (18)), it follows that

$$P(C) \ge \frac{1}{\ln(C)^4}\exp\left(-\frac{1+c}{2c}\ln(C)g(f(c)-1-\eta)\right) \ge \frac{6}{cC}. \tag{20}$$

Then observe that by linearity we have

$$\mathbb{E}\left[N^n_k\right] = \#\mathcal{A}^n_{C^{k+1}}\mathbb{P}\left(H^n_{C^k,C^{k+1}}(v) > \frac{1+c}{2c}\ln(C)(f(c)-\eta)\right).$$

But $|\#\mathcal{A}^n_{C^{k+1}}/(cC^{k+1}) - 1| = |S^n_{C^{k+1}}/(cC^{k+1}) - 1| \le \eta_n(\varepsilon)/c$ for $C^{k+1} \le \varepsilon n$, where we recall that

$$\eta_n(\varepsilon) = \max_{A_n\le i\leqslant \varepsilon n}\left|\frac{S^n_i}{i} - c\right|.$$

The desired result follows by using (19) and (20). □

Observe that by (2) we may choose ε such that $\epsilon \in (0, \varepsilon_0]$ with ε_0 such that

$$\limsup_{n\to\infty}\max_{A_n\le i\le \varepsilon_0 n}\left|\frac{S^n_i}{i} - c\right| < \frac{c}{2} \qquad \text{and} \qquad \liminf_{n\to\infty}\min_{\varepsilon_0 n\le i\le n}\frac{S^n_i}{n} > 0,$$

which implies the existence of a constant $c_0 > 0$ such that:

$$\text{for every } n \text{ sufficiently large}, \quad \text{for every } 1 \le i \le n, \qquad c_0 i \le S^n_i \le i+1, \tag{21}$$

where the upper bound comes from the fact that $\mathrm{x}^n_i \le 1$.

We are now ready to prove Lemma 19.

Proof of Lemma 19. Fix $\eta \in (0,1)$. Consider $C > 1$ given by Lemma 21. Take $\varepsilon \in (0,1)$ small enough such that the conclusion of Lemma 21 holds. The main idea of the proof is to show that for some fixed $\lambda > 0$ we have uniformly for every $k \in \mathbb{N}$ with $A_n \le C^k \le \varepsilon n/C$

$$\mathcal{V}\mathrm{ar}(N^n_k(C,\eta)) \le \lambda C^k. \tag{22}$$

Indeed, it will then directly follow from Lemma 21 and Bienaymé-Chebyshev's inequality that

$$\mathbb{P}\left(N^n_k(C,\eta) < 4C^k\right) \le \mathbb{P}\left(\left|N^n_k(C,\eta) - \mathbb{E}[N^n_k(C,\eta)]\right| > C^k\right) \le \mathcal{V}\mathrm{ar}(N^n_k(C,\eta))/C^{2k} \le \lambda/C^k.$$

We shall show that

$$\mathbb{E}\big[(N_k^n(C,\eta))^2\big] \le \big(\mathbb{E}\big[N_k^n(C,\eta)\big]\big)^2 + \mathcal{O}(C^k). \tag{23}$$

This indeed implies (22). Observe that by definition of $N_k^n(C,\eta)$, setting $t = \frac{1+c}{2c}\ln(C)(f(c)-\eta)$, we have

$$N_k^n(C,\eta) = \sum_{v\in\mathcal{A}^n_{C^{k+1}}} \mathbb{1}_{H^n_{C^k,C^{k+1}}(v)>t}.$$

Note from Algorithm 2, that for every $v, w \in \mathcal{A}^n_{C^{k+1}}$ with $v \neq w$

$$\left(H^n_{C^k,C^{k+1}}(v) \quad , \quad H^n_{C^k,C^{k+1}}(w)\right) \overset{(d)}{=} \left(\sum_{C^k<i\le C^{k+1}} Y_i^n \quad , \quad \sum_{C^k<i\le C^{k+1}} Z_i^n\right)$$

where $(Z_i^n)_{1\le i\le n}$ are independent Bernoulli random variables of respective parameters $(1/S_i^n \mathbb{1}_{\mathrm{x}_i^n=1})_{1\le i\le n}$. The problem is that $(Z_i^n)_{1\le i\le n}$ and $(Y_i^n)_{1\le i\le n}$ are not independent. The idea is to say that, conditionally on the fact that v and w do not coalesce (i.e they do not belong to the same tree) before time C^{k+1}, then, when the height of v goes up by 1 in the coalescence process, the height of w doesn't change and vice versa. Therefore their heights are negatively correlated. One can see that for all $v, w \in \mathcal{A}^n_{C^{k+1}}$,

$$\begin{aligned}&\mathbb{P}\Big(H^n_{C^k,C^{k+1}}(v) > t \text{ and } H^n_{C^k,C^{k+1}}(w) > t\Big)\\ &\quad\le \mathbb{P}\Big(C^k \le \mathrm{c}_n(v,w) \le C^{k+1}\Big) + \mathbb{P}\Big(\mathrm{c}_n(v,w) < C^k \text{ and } H^n_{C^k,C^{k+1}}(v) > t \text{ and } H^n_{C^k,C^{k+1}}(w) > t\Big).\end{aligned} \tag{24}$$

The first probability of (24) is bounded from above as follows:

$$\mathbb{P}\Big(C^k \le \mathrm{c}_n(v,w) \le C^{k+1}\Big) \le 1 - \prod_{C^k\le i\le C^{k+1}}\left(1 - \mathbb{1}_{\mathrm{x}_i^n=1}\frac{1}{\binom{S_i^n}{2}}\right) \le 1 - \exp\left(\sum_{C^k\le i\le C^{k+1}} \ln\left(1 - \frac{1}{\binom{S_i^n}{2}}\right)\right) = \mathcal{O}\left(\frac{1}{C^k}\right), \tag{25}$$

where the last equality comes from (2) and where the $\mathcal{O}(1/C^k)$ is uniform in v, w and in $n \ge C^{k+1}$.

We then focus on the event involved in the second probability of (24). One can write

$$H^n_{C^k,C^{k+1}}(v) = \sum_{C^k<i\le C^{k+1}} Y_i^n,$$

where $(Y_i^n)_{1\le i\le n}$ are independent Bernoulli random variables of respective parameters $(1/S_i^n \mathbb{1}_{x_i^n=1})_{1\le i\le n}$. Moreover, we note that if we know that $Y_i^n = 1$ and that $\mathrm{c}_n(v,w) \le i-1$, then it implies that there is no coalescence with the tree containing w. More precisely, one can write

$$H^n_{C^k,C^{k+1}}(w) = \sum_{C^k<i\le C^{k+1}} Z_i^n.$$

with

$$Z_i^n = Y_i^n \mathbb{1}_{\mathrm{c}_n(v,w)\ge i} + \xi_i^n \mathbb{1}_{\mathrm{c}_n(v,w)<i \text{ and } Y_i^n=0},$$

where the ξ_i^n's are independent Bernoulli r.v. of parameter $(1/(S_i^n-1))\mathbb{1}_{\mathrm{x}_i^n=1}$ respectively and are taken

independently from the Y_i^n's. As a result, the second probability in (24) is bounded from above by

$$\begin{aligned}
&\mathbb{P}\left(\mathrm{c}_n(v,w) < C^k, \sum_{C^k<i\le C^{k+1}} Y_i^n > t \quad \text{and} \quad \sum_{C^k<i\le C^{k+1}} Z_i^n > t\right)\\
&= \mathbb{P}\left(\mathrm{c}_n(v,w) < C^k, \sum_{C^k<i\le C^{k+1}} Y_i^n > t \quad \text{and} \quad \sum_{C^k<i\le C^{k+1}} \xi_i^n \mathbb{1}_{Y_i^n=0} > t\right)\\
&\le \mathbb{P}\left(\sum_{C^k<i\le C^{k+1}} Y_i^n > t \quad \text{and} \quad \sum_{C^k<i\le C^{k+1}} \xi_i^n > t\right)\\
&= \mathbb{P}\left(\sum_{C^k<i\le C^{k+1}} Y_i^n > t\right)\mathbb{P}\left(\sum_{C^k<i\le C^{k+1}} \xi_i^n > t\right) \le \mathbb{P}\left(\sum_{C^k<i\le C^{k+1}} \xi_i^n > t\right)^2. \qquad (26)
\end{aligned}$$

Now, thanks to the fact that $1/S_i^n \le 1/(S_i^n-1)$, one can define some random variables $\widetilde{Y}_i^n$'s for $i \ge 1$ such that for all $i \ge 1$, we have $\widetilde{Y}_i^n \le \xi_i^n$ and $(\widetilde{Y}_i^n)_{i\ge1}$ is a family of independent Bernoulli random variables of parameter $(1/S_i^n)\mathbb{1}_{\mathrm{x}_i^n=1}$, having thus the same law as $(Y_i^n)_{i\ge1}$. It follows that

$$\mathbb{P}\left(\sum_{C^k<i\le C^{k+1}} \xi_i^n > t\right) - \mathbb{P}\left(\sum_{C^k<i\le C^{k+1}} Y_i^n > t\right) = \mathbb{P}\left(\sum_{C^k<i\le C^{k+1}} \xi_i^n > t \ge \sum_{C^k<i\le C^{k+1}} \widetilde{Y}_i^n\right) \le \mathbb{P}\left(\sum_{C^k<i\le C^{k+1}} (\xi_i^n - \widetilde{Y}_i^n) > 0\right)$$

Thus

$$\mathbb{P}\left(\sum_{C^k<i\le C^{k+1}} \xi_i^n > t\right) - \mathbb{P}\left(\sum_{C^k<i\le C^{k+1}} Y_i^n > t\right) \le \mathbb{E}\left[\sum_{C^k<i\le C^{k+1}} (\xi_i^n - \widetilde{Y}_i^n)\right] = \sum_{C^k<i\le C^{k+1}} \frac{1}{S_i^n(S_i^n-1)}\mathbb{1}_{\mathrm{x}_i^n=1} = \mathcal{O}\left(\frac{1}{C^k}\right),$$

where the $\mathcal{O}(1/C^k)$ is uniform in $n \ge C^{k+1}$. By combining this estimate with (24) and (26), we conclude that

$$\mathbb{P}\left(H^n_{C^k,C^{k+1}}(v) > t \text{ and } H^n_{C^k,C^{k+1}}(w) > t\right) \le \mathbb{P}\left(H^n_{C^k,C^{k+1}}(v) > t\right)^2 + \mathcal{O}\left(\frac{1}{C^k}\right)$$

Therefore

$$\begin{aligned}
\mathbb{E}\left[(N_k^n(C,\eta))^2\right] &= \sum_{v\in\mathcal{A}^n_{C^{k+1}}} \mathbb{P}\left(H^n_{C^k,C^{k+1}}(v) > t\right) + \sum_{u\ne v\in\mathcal{A}^n_{C^{k+1}}} \mathbb{P}\left(H^n_{C^k,C^{k+1}}(v) > t \text{ and } H^n_{C^k,C^{k+1}}(w) > t\right)\\
&= \left(\#\mathcal{A}^n_{C^{k+1}}\mathbb{P}\left(\sum_{C^k<i\le C^{k+1}} \xi_i^n > t\right)\right)^2 + \mathcal{O}\left(\frac{1}{C^k}\right). \qquad (27)
\end{aligned}$$

where we have used the inequalities $N_k^n(C,\eta) \le \#\mathcal{A}^n_{C^{k+1}} \le C^{k+1}+1$. This implies (22) and completes the proof. □

To establish Lemma 20 and bound M_k^n we use the following estimate on Pólya urns, which may be shown by following verbatim the proof of Lemma A.1 from [9].

Lemma 22. *Fix an integer $z_0 > 0$ and set $U_0 = 1$. Let $(z_n)_{n\ge1} \in \{-1,1\}^{\mathbb{N}}$, set $Z_n = \sum_{i=0}^n z_i$. Assume that $Z_n > 0$ for every $n \ge 1$ and $\sum_{n=0}^\infty 1/Z_n^2 < \infty$. Let $(U_n)_{n\ge1}$ be a sequence of random non-negative integers such that for every $n \ge 0$,*

$$\mathbb{P}(U_{n+1} = U_n + z_{n+1}|U_n) = \frac{U_n}{Z_n} \quad ; \quad \mathbb{P}(U_{n+1} = U_n|U_n) = \frac{Z_n - U_n}{Z_n}.$$

Almost surely for every $t \ge 0$

$$\mathbb{P}\left(\sup_{i\ge1}\left|\frac{U_i}{Z_i} - \frac{1}{z_0}\right| > t\frac{1}{z_0}\right) \le 2\exp\left(-\frac{(t^2/4)(1/z_0)}{\sum_{n\ge1} 1/Z_n^2 + t\max\left(\sum_{n\ge1} 1/Z_n^2, \max_{n\ge1} 1/Z_n\right)}\right).$$

Proof of Lemma 20. We apply Lemma 22 with $t = k^2$, $z_0 = S^n_{C^k}$, $z_i = x^n_{C^k+i}$ for $1 \le i \le C^{k+1} - C^k$ and $z_i = 1$ for all $i > C^{k+1} - C^k$, and U_i being the number of active vertices in a given tree of $\mathcal{F}^n_{C^k,C^k+i}$. Thus, writing X_k for the number of active vertices of a given tree of the forest $\mathcal{F}^n_{C^k,C^{k+1}}$, writing $\delta^n_k := \sum_{C_k<n\le C^{k+1}} 1/(S^n_i)^2 + \sum_{j\ge 1} 1/(S^n_{C^{k+1}} + j)^2$, and $\varepsilon^n_k := \max_{C^k<i\le C^{k+1}} 1/S^n_i$,

$$\mathbb{P}\Big(X_k/S^n_{C^{k+1}} > (1+k^2)/S^n_{C^k}\Big) \le 2\exp\left(-\frac{(k^4/4)/S^n_{C^k}}{\delta^n_k + k^2\max(\delta^n_k,\varepsilon^n_k)}\right).$$

And by (21), as $k, n \to \infty$ with $A_n \le C^k \le n/C$, we have $S^n_{C^k} \le C^k + 1$, $1/S^n_{C^k} = \mathcal{O}(1/C^k)$, $\varepsilon^n_k = \mathcal{O}(1/C^k)$ and $\delta^n_k = \mathcal{O}(1/C^k)$, thus for every k large enough,

$$\mathbb{P}\big(X_k > k^3\big) \le 2e^{-k^2/\mathcal{O}(1)}.$$

Then, since there are at most $S^n_{C^k} = \mathcal{O}(C^k)$ trees in $\mathcal{F}^n_{C^k,C^{k+1}}$, for n large enough, for every $k \in \mathbb{N}$, with $A_n \le C^k \le n/C$,

$$\mathbb{P}\big(M^n_k > k^3\big) \le S^n_{C^k} 2e^{-k^2/\mathcal{O}(1)} \le 1/k^2. \qquad \square$$

6 Application: contact-tracing in a stochastic SIR dynamics

Our results can be applied to study the so-called "infection tree" of a stochastic SIR dynamics, which is a classical model for the evolution of epidemics (for background on stochastic epidemic models, see [4, 12]). We assume that initially there is 1 infectious individual (that has just become infected) and n susceptible individuals. The infectious periods of different infectives are i.i.d distributed according to an exponential random variable of parameter 1. During her infectious period an infective makes contacts with a given individual at the time points of a time homogeneous Poisson process with intensity λ_n. If a contacted individual is still susceptible, then she becomes infectious and is immediately able to infect other individuals. An individual is considered "removed" once her infectious period has terminated, and is then immune to new infections, playing no further part in the epidemic spread. The epidemic ceases as soon as there are no more infectious individuals present in the population. All Poisson processes are assumed to be independent of each other; they are also independent of the infectious periods.

We call a "step" an event where either a susceptible individuals becomes infective, or where an infectious period terminates. Denote by τ_n the number of steps made when the epidemic ceases. For $0 \le k \le \tau_n$, let $\mathcal{T}^n_k$ be the infection tree after k steps, in which the vertices are individuals and where edges correspond between two individuals if one has infected the other. We are interested in the evolution of the associated "infection tree" $(\mathcal{T}^n_k)_{0\le k\le\tau_n}$, as well as in the shape of the full infection tree $\mathcal{T}^n_{\tau_n}$ when the epidemic ceases.

Let $(H^n_k, I^n_k)_{k\ge 0}$ be a Markov chain with initial state $(H^n_0, I^n_0) = (n, 1)$ and transition probabilities given by

$$(H^n_{k+1}, I^n_{k+1}) = \begin{cases} (H^n_k - 1, I^n_k + 1) & \text{with probability } \frac{\lambda_n H^n_k}{1+\lambda_n H^n_k} \\ (H^n_k, I^n_k - 1) & \text{with probability } \frac{1}{1+\lambda_n H^n_k} \end{cases}$$

with $\{(k,0) : 0 \le k \le n\}$ as absorbing states. Observe that evolution of the number of susceptible individuals and the number of infectious individuals in the infection tree evolves according to this Markov chain. Then define the random sequence $\mathbf{X}^n = (\mathrm{X}^n_i)_{1\le i\le\tau'_n}$ of ± 1 as follows: let τ'_n be the absorption time of the Markov chain, and for $1 \le i \le \tau'_n$ set $\mathrm{X}^n_i = I^n_i - I^n_{i-1}$.

Then by construction, it is clear that:

$$(\mathcal{T}^n_k)_{0\le k\le\tau_n} \overset{(d)}{=} (\mathcal{T}_k(\mathbf{X}^n))_{0\le k\le\tau'_n}.$$

To simplify notation, let $\mathcal{T}^n = \mathcal{T}^n_{\tau_n}$ be the full infection tree when the epidemic ceases. We also identify the random variables in the above equality in distribution.

Theorem 23. *The following assertions hold.*

(1) Assume that $\lambda_n \sim \lambda/n$ for some $\lambda > 0$. Then $\mathcal{T}^n$ converges in distribution locally towards a Bienaymé tree with offspring distribution $\mathcal{G}(1/(1+\lambda))$.

(2) Assume that $\lambda_n \gg 1/n$. Then $\mathcal{T}^n$ is not tight for the topology of the local convergence.

Proof. Let us first assume $\lambda_n \sim \lambda/n$ as $n \to \infty$ for some $\lambda > 0$. To prove the local convergence in distribution of $\mathcal{T}^n$ towards a Bienaymé tree, it is enough by Theorem 1 and Theorem 2 to prove that $(I^n_k)_{k\geq 0}$ converges in distribution for the product topology towards the random walk $(S_k)_{k\geq 0}$ such that $S_{k+1}-S_k = 1$ with probability $\lambda/(1+\lambda)$ and $S_{k+1} - S_k = -1$ with probability $1/(1+\lambda)$. Indeed, the local convergence of $\mathcal{T}_n(\mathbf{X}^n)$ implies the local convergence of $\mathcal{T}^n$. So as to achieve this, one can first see that since $H^n_k \leq n$ for all $k \geq 0$, we have

$$\frac{\lambda_n H^n_k}{1+\lambda_n H^n_k} \leq \frac{n\lambda_n}{1+n\lambda_n} \xrightarrow[n\to\infty]{} \frac{\lambda}{1+\lambda}.$$

But we also have

$$\inf_{k\leq\sqrt{n}} \frac{\lambda_n H^n_k}{1+\lambda_n H^n_k} \geq \frac{\lambda_n(n-\sqrt{n})}{1+n\lambda_n} \xrightarrow[n\to\infty]{} \frac{\lambda}{1+\lambda}.$$

For all $k_0 \geq 0$, one can thus build a coupling between $(I^n_k)_{0\leq k\leq k_0}$ and $(S_k)_{0\leq k\leq k_0}$ such that the two walks coincide w.h.p. until time k_0 as $n \to \infty$. This concludes the proof of the local limit result.

For (2), the absence of local limit for $\mathcal{T}^n$ stems from Theorem 1 and from the fact that $(I^n_k)_{k\geq 0}$ converges in distribution with respect to the product topology towards $(k+1)_{k\geq 0}$ since for all fixed $k \geq 0$, the transition probability $\lambda_n H^n_k/(1+\lambda_n H^n_k)$ converges to 1 as $n \to \infty$. □

Theorem 24. *The following assertions hold.*

(1) Assume that $\lambda_n \sim \lambda/n$ for some $\lambda > 1$. Let A be the event of survival of a Bienaymé process with offspring distribution $\mathcal{G}(1/(1+\lambda))$.

(a) The convergence

$$\left(\frac{I^n_{\lfloor nt\rfloor}}{n}\right)_{t\geq 0} \xrightarrow[n\to\infty]{(d)} (\max(2-2g(t)-t, 0)\mathbb{1}_A)_{t\geq 0}, \tag{28}$$

holds in distribution for the topology of uniform convergence on compact sets, where g is the solution of the ordinary differential equation $g'(t) = -\lambda g(t)/(1+\lambda g(t))$ with $g(0) = 1$.

(b) For all $n \geq 1$, conditionally given $\mathcal{T}^n$, let V^n_1 and V^n_2 be independent uniform vertices of $\mathcal{T}^n$. Then

$$\frac{\mathsf{H}(V^n_1)}{\ln n} \xrightarrow[n\to\infty]{(d)} \frac{\lambda}{\lambda-1}\mathbb{1}_A \qquad \text{and} \qquad \frac{d^n(V^n_1, V^n_2)}{\ln n} \xrightarrow[n\to\infty]{(d)} \frac{2\lambda}{\lambda-1}\mathbb{1}_A, \tag{29}$$

where $\mathsf{H}(V^n_1)$ is the height of V^n_1 in $\mathcal{T}^n$ and d^n denotes the graph distance in $\mathcal{T}^n$.

(2) Assume that $\lambda_n \gg 1/n$.

(a) The convergence

$$\left(\frac{I^n_{\lfloor nt\rfloor}}{n}\right)_{0\leq t\leq 2} \xrightarrow[n\to\infty]{\mathbb{P}} (\min(t, 2-t))_{0\leq t\leq 2}, \tag{30}$$

holds in probability for the topology of uniform convergence.

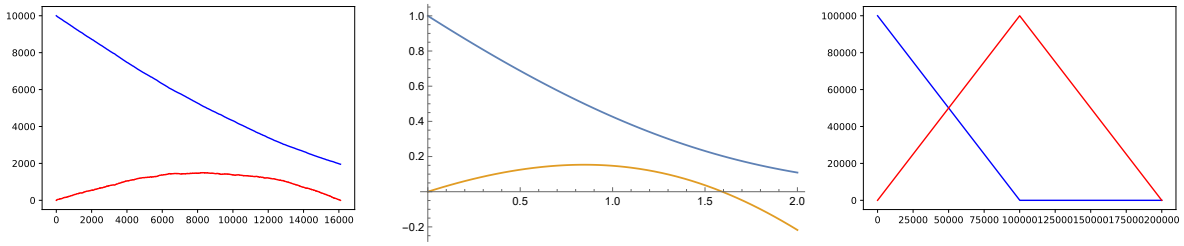

Figure 10: Left: simulation of (H^n, I^n) for $\lambda_n = 2/n$ and $n = 10000$ (H^n in blue and I^n in red). Centre: the corresponding fluid limit. Right: simulation of (H^n, I^n) for $\lambda_n = 2$ and $n = 100000$.

(b) We have

$$\frac{\mathsf{H}(V_1^n)}{\ln n} \xrightarrow[n\to\infty]{\mathbb{L}^p} 1, \qquad \frac{d^n(V_1^n, V_2^n)}{\ln n} \xrightarrow[n\to\infty]{\mathbb{L}^p} 2 \qquad \textit{and} \qquad \frac{\mathsf{Height}(\mathcal{T}^n)}{\ln(n)} \xrightarrow[n\to\infty]{\mathbb{P}} e. \tag{31}$$

We observe that the solution g of the differential equation can be expressed using the Lambert W function.

Remark 25. It would be very interesting to obtain a limit theorem for $\mathsf{Height}(\mathcal{T}^n)$ in the setting of Theorem 24 (1) (b). This requires more work, because the behavior of the number of infected/active vertices near the extinction time τ_n' should give a non-negligible contribution to the height.

Proof of Theorem 24. For (1) (a), we assume $\lambda > 1$ and we determine the fluid limit of the chain (H^n, I^n). Let $(\widetilde{H}^n, \widetilde{I}^n)$ be a Markov chain which has the same initial conditions as (H^n, I^n) and the same transition probabilities but which is not stopped when $\widetilde{I}^n$ reaches zero. We check that (28) holds using the classical theory of fluid limits of Markov chains (see e.g. [28] or [15]), and more precisely a combination of Grönwall's inequality with Doob's maximal inequality.

We claim that it is enough to show that

$$\left(\frac{\widetilde{H}^n_{\lfloor nt \rfloor}}{n}\right)_{t\geq 0} \xrightarrow[n\to\infty]{\mathbb{P}} (g(t))_{t\geq 0}, \tag{32}$$

uniformly on compacts. Indeed, it is easy to see that $\widetilde{I}^n_k = 2(n - \widetilde{H}^n_k) - k + 1$ so that (32) entails the convergence

$$\left(\frac{\widetilde{I}^n_{\lfloor nt \rfloor}}{n}\right)_{t\geq 0} \xrightarrow[n\to\infty]{\mathbb{P}} (2 - 2g(t) - t)_{t\geq 0}, \tag{33}$$

uniformly on compacts. Let us check that this implies (28). Since $t \mapsto 2 - 2g(t) - t$ is concave, starts from zero at time zero, stays positive and then goes below zero, it is thus enough to identify the limit of the probability that the chain $\widetilde{I}^n$ reaches zero before time εn as $n \to \infty$ and $\varepsilon \to 0$ as the probability of extinction of a $\mathcal{G}(1/(1+\lambda))$-Bienaymé process. To this end, for all $\varepsilon > 0$, for n large enough, for all $k \leq \varepsilon n$, since $n \geq \widetilde{H}^n_k \geq (1-\varepsilon)n$ we have

$$\frac{\lambda}{1+\lambda} - \varepsilon \leq (1-\varepsilon)\frac{n\lambda_n}{1+n\lambda_n} \leq \frac{\lambda_n \widetilde{H}^n_k}{1+\lambda_n \widetilde{H}^n_k} \leq \frac{n\lambda_n}{1+n\lambda_n} \leq \frac{\lambda}{1+\lambda} + \varepsilon.$$

Thus, one can couple (I^n, H^n) until time εn with two simple random walks $(\overline{Y}^\varepsilon_k)_{k\geq 0}$ and $(\underline{Y}^\varepsilon_k)_{k\geq 0}$ starting from one such that

$$\mathbb{P}(\overline{Y}^\varepsilon_1 = 1) = \frac{\lambda}{1+\lambda} + \varepsilon \qquad \text{and} \qquad \mathbb{P}(\underline{Y}^\varepsilon_1 = 1) = \frac{\lambda}{1+\lambda} - \varepsilon$$

so that for all n large enough, for all $k \le \varepsilon n$,

$$\underline{Y}^{\varepsilon}_k \le \widetilde{I}^n_k \le \overline{Y}^{\varepsilon}_k. \tag{34}$$

Since the probabilities that the random walks $\underline{Y}^{\varepsilon}$ and $\overline{Y}^{\varepsilon}$ starting from 1 reach the negative integers converge both as $\varepsilon \to 0$ to the probability of extinction of a $\mathcal{G}(1/(1+\lambda))$-Bienaymé process, we get the desired result.

It remains to check (32). This is a rather direct application of Lemma 6.5 in [14] (which remains clearly true if we add a superscript n to all the objects). More precisely, let $t > 0$, for all $k \ge 0$ let $X^n_k = \widetilde{H}^n_k - ng(k/n)$ and let $(\mathcal{F}^n_k)_{k\ge 0}$ be the natural filtration associated with X^n. This lemma states that if there exists a constant $C > 0$ such that for all $n \ge 1$, for all $0 \le k \le nt$,

$$\left|\mathbb{E}\left(X^n_{k+1} - X^n_k \middle| \mathcal{F}^n_k\right)\right| \le \frac{C}{n}\left(1 + \sup_{0\le j\le nt} |X^n_j|\right) \qquad \text{a.s. and} \qquad \mathbb{E}\left(\left(X^n_{k+1} - X^n_k\right)^2\right) \le C, \tag{35}$$

then $\sup_{0\le k\le nt} |X^n_k|/n \to 0$ in probability as $n \to \infty$, hence (32).

To show (35), we first compute

$$\mathbb{E}\left(X^n_{k+1} - X^n_k \middle| \mathcal{F}^n_k\right) = -\frac{\lambda \widetilde{H}^n_k/n}{1 + \lambda \widetilde{H}^n_k/n} - n\left(g\left(\frac{k+1}{n}\right) - g\left(\frac{k}{n}\right)\right).$$

Thus, using Taylor at order 2 and using the definition of g one gets that

$$\left|\mathbb{E}\left(X^n_{k+1} - X^n_k + \frac{\lambda \widetilde{H}^n_k/n}{1 + \lambda \widetilde{H}^n_k/n} - \frac{\lambda g(k/n)}{1 + \lambda g(k/n)} \middle| \mathcal{F}^n_k\right)\right| \le \frac{C}{n},$$

for some constant $C > 0$. Moreover, the function $x \mapsto x/(1+x)$ is 1-Lipschitz on $[0, \infty)$. Thus, the first inequality of (35) is satisfied. The second one is clearly satisfied since $\widetilde{H}^n_{k+1} - \widetilde{H}^n_k \in \{0, -1\}$ and since g is differentiable, hence (32).

We now turn to (1) (b). We essentially apply Theorem 4 with $c = (\lambda - 1)/(\lambda + 1)$ and $A_n = \log\log n$. However, the rub is that the fluid limit reaches zero so that we can not apply it directly. To overcome this issue, let $t_0 > 0$ be the unique positive time at which $2 - 2g(t_0) - t_0 = 0$. We shall check that we are in position to apply Theorem 4 to the subtree $\mathcal{T}_{\lfloor (t_0-\varepsilon)n\rfloor}(\mathbf{X}^n)$. This will indeed enable us to conclude since if V^n_1 and V^n_2 are two independent uniform random vertices of $\mathcal{T}^n$, then by Lemma 6,

$$\mathbb{P}(\mathrm{b}_{\tau'_n}(V^n_1) \le (t_0 - \varepsilon)n) = \mathbb{P}(\mathrm{b}_{\tau'_n}(V^n_2) \le (t_0 - \varepsilon)n) \ge 1 - \varepsilon + o(1) \quad \text{as } n \to \infty.$$

To check that Theorem 4 applies to the subtree $\mathcal{T}_{\lfloor (1-\varepsilon)\tau'_n\rfloor}(\mathbf{X}^n)$ with $c = (\lambda - 1)/(\lambda + 1)$ and $A_n = \log\log n$, it remains to show that for all $\varepsilon > 0$, there exists $\delta > 0$ such that on the event "$I^n_k > 0$ for all $k \le \varepsilon n$",

$$\max_{\log\log n\le k\le \delta n} \left|\frac{I^n_k}{k} - \frac{\lambda - 1}{\lambda + 1}\right| \le \varepsilon + o(1)$$

in probability as $n \to \infty$. In turn, it is enough to prove that for every $\varepsilon > 0$ we have

$$\max_{\log\log n\le k\le \varepsilon n} \left|\frac{\widetilde{I}^n_k}{k} - \frac{\lambda - 1}{\lambda + 1}\right| \le \varepsilon + o(1)$$

in probability as $n \to \infty$. This follows by combining the coupling (34) with the law of large numbers.

We now turn to (2) (a) and assume $\lambda_n \gg 1/n$. Let us first establish that if $\delta > 0$, then

$$\left(\frac{I^n_{\lfloor nt\rfloor}}{n}\right)_{0\le t\le 1} \xrightarrow[n\to\infty]{\mathbb{P}} (t)_{0\le t\le 1} \qquad \text{and} \qquad \sup_{\log\log n\le k\le (1-\delta)n} \left|\frac{I^n_k}{k} - 1\right| \xrightarrow[n\to\infty]{\mathbb{P}} 0, \tag{36}$$

where the first convergence holds for the topology of the uniform convergence. Let $\delta > 0$. Note that for all $k \leq (1-\delta)n$ we have $H_k^n \geq \delta n$. As a result, for all $\varepsilon > 0$, for n large enough, for all $k \leq (1-\delta)n$.

$$1 - \varepsilon \leq \frac{\lambda_n H_k^n}{1 + \lambda_n H_k^n} \leq 1.$$

One can thus obtain (36) by coupling I_k^n with a simple random walk (whose step is +1 with probability $1-\varepsilon$).

We now explain why (36) implies (30). By (36), $H_n^n = o(n)$ with high probability, so that after time n there are $o(n)$ remaining steps of +1 for I^n and all the other steps are -1. Hence, for the topology of the uniform convergence,

$$\left(\frac{I^n_{\lfloor nt \rfloor}}{n}\right)_{1\leq t\leq 2} \xrightarrow[n\to\infty]{\mathbb{P}} (2-t)_{1\leq t\leq 2},$$

and this concludes the proof of (30).

We finally turn to (2) (b). As in the proof of (1) (a), the first two convergences of (31) are proved by applying Theorem 4 to the tree $\mathcal{T}_{\lfloor (2-\varepsilon)n \rfloor}(\mathbf{X}^n)$ for some $\varepsilon > 0$ small enough. The last convergence of (2) (b) is more delicate. The idea is to apply Theorem 4 to the tree $\mathcal{T}_n(\mathbf{X}^n)$ and then to show that the vertices added between times n and τ'_n, in a first approximation, do not affect the height.

More precisely, by Theorem 4 the height of $\mathcal{T}_n(\mathbf{X}^n)$ divided by $\ln n$ converges in probability towards e. To understand how the height of $\mathcal{T}^n$ behaves compared to $\mathcal{T}_n(\mathbf{X}^n)$ we need to estimate the quantity

$$\mathsf{h}^n := \sum_{i=n}^{\tau'_n - 1} \frac{1}{I_i^n} \mathbb{1}_{\{\mathrm{X}_i^n = 1\}}.$$

Indeed, as we have already seen multiple times, roughly speaking this quantity represents "the height" added between times n and τ'_n. Define $\hat{\mathrm{X}}_i^n = -\mathrm{X}^n_{\tau'_n - i}$ and $\hat{S}_i^n = I^n_{\tau'_n - i} = \hat{\mathrm{X}}_1^n + \cdots + \hat{\mathrm{X}}_i^n$ for all $0 \leq i \leq \tau'_n$. Then

$$\mathsf{h}^n = \sum_{i=1}^{\tau'_n - n} \frac{1}{\hat{S}_i^n} \mathbb{1}_{\{\hat{\mathrm{X}}_i^n = -1\}}.$$

In a similar way that we have proved that $(I_i^n)_{0\leqslant i\leqslant n}$ satisfies the assumptions (2) with high probability, one shows that $(\hat{S}_i^n)_{0\leqslant i\leqslant \tau'_n - n}$ also satisfies those assumptions with high probability. Therefore, one could easily adapt the convergence of Theorem 4 (1) to show that that h^n divided by $\ln n$ converges to 0 in probability (indeed, if in Theorem 4 we look at h_n^- instead of h_n^+, where h_n^- consists of the sum over the -1 steps instead of the +1 steps, then, we have that h_n^- divided by $\ln n$ converges to $(1-c)/(2c)$). It remains to show how it implies that $\mathsf{Height}(\mathcal{T}^n) = \mathsf{Height}(\mathcal{T}_n(\mathbf{X}^n)) + o(\ln n)$. To do so we will use, once again, the coalescence construction and Bennett's inequality. By Algorithm 2, conditionally on τ'_n and $(I_i^n)_{0\leqslant i\leqslant \tau'_n}$, the height of any new vertex added between time n and τ'_n is stochastically dominated by $\mathsf{Height}(\mathcal{T}_n(\mathbf{X}^n)) + \sum_{i=n}^{\tau'_n - 1} Y_i$ where $(Y_i^n)_{n\leqslant i\leqslant \tau'_n - 1}$ are i.i.d Bernoulli random variables of respective parameters $(1/I_i^n \mathbb{1}_{\{\mathrm{X}_i^n = 1\}})_{n\leqslant i\leqslant \tau'_n - 1}$. Using Bennett's inequality and a union bound we obtain that, for all $\varepsilon > 0$

$$\mathbb{P}(\mathsf{Height}(\mathcal{T}^n) - \mathsf{Height}(\mathcal{T}_n(\mathbf{X}^n)) \geq \mathsf{h}^n + \varepsilon \ln n) \leq n \exp\left(-\mathsf{h}^n g\left(\frac{\varepsilon \ln n}{\mathsf{h}^n}\right)\right)$$

where $g(u) = (u+1)\ln(u+1) - u \sim u \ln u$ as $u \to +\infty$. Since $\mathsf{h}^n = o(\ln n)$ it is a simple matter to check that the above upper bound goes to 0 when n goes to ∞ and thus $\mathsf{Height}(\mathcal{T}^n)$ is of order $e \ln n$. □

7 Perspectives and extensions

We mention here some perspectives for future research in the direction of studying the impact of freezing in random graph models.

(1) What is the number of ends of $\mathcal{T}_\infty$?

(2) It would be interesting to study the evolution of the sizes of the trees in the forests obtained by Algorithm 2 (in the particular case where $\mathbf{x}_n = 1$ for every $n \geq 1$ this corresponds to Kingman's coalescent), as well as the evolution of degree vertices (in the particular case where $\mathbf{x}_n = 1$ for every $n \geq 1$ there are nice connections with Pólya urns).

(3) Is the height of $\mathcal{T}_n$ stochastically maximized when $\mathrm{x}_k = 1$ for every $1 \leq k \leq n$.

(4) A natural question is to obtain a limit theorem for the height of the epidemic tree $\mathcal{T}^n$ in the setting of Theorem 24 (1) (b) (see Remark 25).

(5) It would be very interesting to extend our results to other attachment mechanisms, such as preferential attachment, where new vertices attach to existing vertices proportional to their degree.

(6) Similarly one may wonder what happens when the frozen vertices are not chosen uniformly, notably when vertices with high degrees are more or less likely to freeze.

(7) An important question arising in the context of growing real-world networks or in the context of infection tracing is how to find the root, or patient zero [36]. This is a very active area of research, sometimes called *network archeology* [33]. Such questions have been for instance considered for uniform attachment and preferential attachment trees [22, 13, 29], as well as Bienaymé trees [11].

References

[1] L. Addario-Berry. Partition functions of discrete coalescents: from Cayley's formula to Frieze's ζ (3) limit theorem. In *XI Symposium on Probability and Stochastic Processes*, pages 1–45. Springer, 2015.

[2] L. Addario-Berry, A. Brandenberger, J. Hamdan, and C. Kerriou. Universal height and width bounds for random trees. *Electronic Journal of Probability*, 27(none):1 – 24, 2022.

[3] L. Addario-Berry and L. Eslava. High degrees in random recursive trees. *Random Structures & Algorithms*, 52(4):560–575, 2018.

[4] H. Andersson and T. Britton. *Stochastic epidemic models and their statistical analysis*, volume 151. Springer Science & Business Media, 2012.

[5] V. Bansaye, C. Gu, and L. Yuan. A growth-fragmentation-isolation process on random recursive trees and contact tracing. *To appear in* Ann. Appl. Probab.

[6] A. D. Barbour and P. Hall. On the rate of Poisson convergence. *Mathematical Proceedings of the Cambridge Philosophical Society*, 95(3):473–480, 1984.

[7] E. Bellin, A. Blanc-Renaudie, E. Kammerer, and I. Kortchemski. Uniform attachment with freezing: scaling limits. *Preprint available on arxiv*, 2023.

[8] J. Bertoin. A model for an epidemic with contact tracing and cluster isolation, and a detection paradox. *Journal of Applied Probability*, pages 1–17, 2023.

[9] A. Blanc-Renaudie. Compactness and fractal dimensions of inhomogeneous continuum random trees. *Probability Theory and Related Fields*, pages 1–31, 2022.

[10] K. A. Borovkov and V. A. Vatutin. On the Asymptotic Behaviour of Random Recursive Trees in Random Environments. *Advances in Applied Probability*, 38(4):1047–1070, 2006.

[11] A. M. Brandenberger, L. Devroye, and M. K. Goh. Root estimation in Galton–Watson trees. *Random Structures & Algorithms*, 61(3):520–542, 2022.

[12] T. Britton, E. Pardoux, F. Ball, C. Laredo, D. Sirl, and V. C. Tran. *Stochastic epidemic models with inference*. Springer, 2019.

[13] S. Bubeck, L. Devroye, and G. Lugosi. Finding Adam in random growing trees. *Random Structures & Algorithms*, 50(2):158–172, 2017.

[14] N. Curien. A random Walk among random Graphs, 2023. Lecture notes available online https://www.dropbox.com/s/bf6a61x3mpad4um/cours-GA-online.pdf.

[15] R. W. R. Darling and J. R. Norris. Differential equation approximations for Markov chains. *Probab. Surv.*, 5:37–79, 2008.

[16] L. Devroye. Branching processes in the analysis of the heights of trees. *Acta Informatica*, 24(3):277–298, 1987.

[17] J. Doyle and R. L. Rivest. Linear expected time of a simple union-find algorithm. *Inf. Process. Lett.*, 5(5):146–148, 1976.

[18] G. Elek and G. Tardos. Convergence and Limits of Finite Trees. *Combinatorica*, 42(6):821–852, 2022.

[19] J. L. Gastwirth. A Probability Model of a Pyramid Scheme. *The American Statistician*, 31(2):79–82, 1977.

[20] C. Goldschmidt and J. Martin. Random Recursive Trees and the Bolthausen-Sznitman Coalesent. *Electronic Journal of Probability*, 10(none):718 – 745, 2005.

[21] R. Grossman and R. G. Larson. Hopf-algebraic structure of families of trees. *Journal of Algebra*, 126(1):184–210, 1989.

[22] J. Haigh. The recovery of the root of a tree. *Journal of Applied Probability*, 7(1):79–88, 1970.

[23] M. Janic, F. Kuipers, X. Zhou, and P. Van Mieghem. Implications for qos provisioning based on traceroute measurements. In B. Stiller, M. Smirnow, M. Karsten, and P. Reichl, editors, *From QoS Provisioning to QoS Charging*, pages 3–14, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[24] S. Janson. Tree limits and limits of random trees. *Combinatorics, Probability and Computing*, 30(6):849–893, 2021.

[25] D. E. Knuth and A. Schönhage. The expected linearity of a simple equivalence algorithm. *Theoretical Computer Science*, 6(3):281–315, 1978.

[26] P. L. Krapivsky and S. Redner. Organization of growing random networks. *Physical Review E*, 63(6):066123, 2001.

[27] R. Kürsten. Random recursive trees and the elephant random walk. *Phys. Rev. E*, 93:032111, Mar 2016.

[28] T. G. Kurtz. *Approximation of population processes*, volume 36 of *CBMS-NSF Reg. Conf. Ser. Appl. Math.* Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 1981.

[29] G. Lugosi and A. S. Pereira. Finding the seed of uniform attachment trees. *Electronic Journal of Probability*, 24(none):1 – 15, 2019.

[30] J. W. Moon. *The distance between nodes in recursive trees*, pages 125–132. London Mathematical Society Lecture Note Series. Cambridge University Press, 1974.

[31] H. S. Na and A. Rapoport. Distribution of nodes of a tree by degree. *Mathematical Biosciences*, 6:313–329, 1970.

[32] D. Najock and C. C. Heyde. On the number of terminal vertices in certain random trees with an application to stemma construction in philology. *Journal of Applied Probability*, 19(3):675–680, 1982.

[33] S. Navlakha and C. Kingsford. Network archaeology: uncovering ancient networks from present-day interactions. *PLoS computational biology*, 7(4):e1001119, 2011.

[34] J. Neveu. Arbres et processus de Galton-Watson. In *Annales de l'IHP Probabilités et statistiques*, volume 22, pages 199–207, 1986.

[35] B. Pittel. Note on the heights of random recursive trees and random m-ary search trees. *Random Structures & Algorithms*, 5(2):337–347, 1994.

[36] C. Shah, N. Dehmamy, N. Perra, M. Chinazzi, A.-L. Barabási, A. Vespignani, and R. Yu. Finding patient zero: Learning contagion source with graph neural networks. *arXiv preprint arXiv:2006.11913*, 2020.

[37] R. T. Smythe and H. M. Mahmoud. A survey of recursive trees. *Theory of Probability and Mathematical Statistics*, (51):1–28, 1995.