

# A branching process with coalescence to model random phylogenetic networks

François Bienvenu\*      Jean-Jil Duchamps†

## Abstract

We introduce a biologically natural, mathematically tractable model of random phylogenetic network to describe evolution in the presence of hybridization. One of the features of this model is that the hybridization rate of the lineages correlates negatively with their phylogenetic distance. We give formulas/characterizations for quantities of biological interest that make them straightforward to compute in practice. We show that the appropriately rescaled network, seen as a metric space, converges to the Brownian continuum random tree, and that the uniformly rooted network has a local weak limit, which we describe explicitly.

**Keywords:** reticulate evolution; explicit phylogenetic network; random DAG; tree-like structure; logistic branching process; local limit; continuum random tree.

**MSC2020 subject classifications:** Primary 60J80; 05C80, Secondary 60F05; 60J27; 92B10.

Submitted to EJP on November 9, 2022, final version accepted on January 18, 2024.

Supersedes arXiv:2211.02407.

Supersedes HAL:hal-03841257.

## Contents

<b>1 Introduction</b>	<b>2</b>
1.1 Biological context . . . . .	2
1.2 Setting and main results . . . . .	3
1.3 Comments and perspectives . . . . .	4
<b>2 Probability of extinction and growth rate</b>	<b>5</b>
2.1 The logistic branching process . . . . .	5
2.2 Offspring distribution and extinction probability of $\mathcal{T}$ . . . . .	8
2.3 Growth rate of the number of lineages . . . . .	13

---

\*Institute for Theoretical Studies, ETH Zürich, 8092 Zürich, Switzerland.  
 E-mail: francois.bienvenu@normalesup.org; <http://www.normalesup.org/~bienvenu>  
 †Université de Franche-Comté, CNRS, LmB, F-25000 Besançon, France.  
 E-mail: jean-jil.duchamps@univ-fcomte.fr; <https://duchamps.perso.math.cnrs.fr>

<b>3</b>	<b>Convergence to the CRT</b>	<b>15</b>
3.1	The rooted Gromov–Hausdorff–Prokhorov distance . . . . .	16
3.2	Conditioning on the number of colors . . . . .	18
3.3	Technical lemmas . . . . .	19
3.4	Proof of the convergence to the CRT . . . . .	24
<b>4</b>	<b>Local weak limit</b>	<b>29</b>
4.1	Local topology . . . . .	29
4.2	Construction of the limit as a decorated tree . . . . .	30
4.3	Geometry of the focal and spinal networks . . . . .	33
	<b>Appendices</b>	<b>36</b>
A.1	Path decompositions of Markov chains . . . . .	36
A.2	Proofs for Section 3.1 . . . . .	40
A.3	An extension of Skorokhod’s representation theorem . . . . .	43
A.4	Tail of the size of critical Galton–Watson trees . . . . .	45
	<b>References</b>	<b>45</b>

## 1 Introduction

### 1.1 Biological context

Random trees play a central role in evolutionary biology: ultimately, much of what we know about evolution relies on a random tree being used as a null model. Meanwhile, the genomic revolution of the past decades has shown that phenomena once thought to play a minor role in large-scale evolution, such as hybrid speciation [36, 37] or horizontal gene transfers [8, 26, 40], are in fact widespread and crucial to our understanding of evolutionary processes. As a result, there have been growing calls by biologists to replace trees by networks when studying phylogenies [6, 17, 18], which lead to the emergence of the flourishing field of phylogenetic networks (see e.g. [32] for a recent review).

Despite this, there is still a notable lack of biologically relevant, mathematically tractable models of random phylogenetic networks. To the best of our knowledge, so far only two models of random phylogenetic networks have been studied extensively from a probabilistic standpoint: uniform ranked tree-child networks [9] and uniform level- $k$  networks [48]. Uniform ranked tree-child networks are generated by a biologically natural process where species split at constant rate and pairs of species hybridize at a constant rate. They turn out to be highly tractable [9, 12, 25]. However, they fail to take into account the fact that phylogenetically distant species are less likely to hybridize than closely related ones, which results in a very non-tree-like structure whose biological relevance is questionable. By contrast, uniform level- $k$  networks have a tree-like large-scale structure [48]; but they do not have a biological interpretation that would justify their relevance as a model of random phylogenetic network, and they are not as mathematically tractable (at least for generic values of  $k$ ).

In this work, we introduce a model of random phylogenetic network that has a natural biological interpretation while remaining mathematically tractable. The idea of this model is to consider species that (1) speciate and go extinct at constant rates and (2) hybridize, subject to some constraints: each species has a type, which can be thought of as a proxy for the genetic distance, and species of the same type hybridize at a constant rate. Types are created at a constant rate in an infinite-allele fashion, and inherited by descendants. The formal description of the model is given in the next section, along with an overview of our main results.

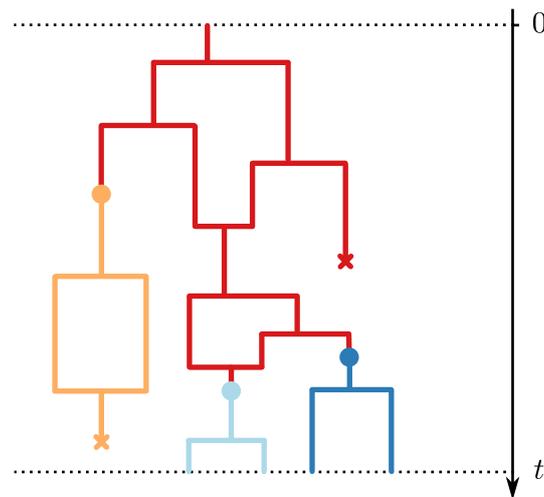


Figure 1: Graphical depiction of a realization of the process generating the network  $\mathcal{G}$ . The vertical axis is the time, flowing from top to bottom, and the vertical lines represent the lineages. Dots correspond to mutations (i.e. to a lineage changing color) and crosses correspond to deaths (i.e. to a lineage stopping). Horizontal lines correspond to either branching or coalescence, and serve to indicate the genealogical relationship between lineages; they should be treated as having length 0.

## 1.2 Setting and main results

Starting from one colored lineage at time  $t = 0$ , consider the continuous-time interacting particle system where:

- each lineage splits into two lineages at rate 1 (*branching*);
- each lineage dies at rate  $\alpha > 0$  (*death*);
- each pair of lineages of the same color merge at rate  $2\beta > 0$  (*coalescence*);
- each lineage takes a new, never-seen-before color at rate  $\mu > 0$  (*mutation*).

As illustrated in Figure 1, this process defines a time-embedded random network which can be seen as a random metric measure space  $(\mathcal{G}, d_{\mathcal{G}}, \lambda_{\mathcal{G}})$ . Formally, a point  $x \in \mathcal{G}$  corresponds to a lineage  $\ell$  and a time  $t$  at which that lineage is alive. Since the lineages can be seen as segments,  $\mathcal{G}$  can be seen as a collection of segments glued together at their endpoints, and  $\lambda_{\mathcal{G}}$  as the usual Lebesgue measure on this union of segments.

There is a natural metric  $d_{\mathcal{G}}$  on  $\mathcal{G}$ , obtained by defining  $d_{\mathcal{G}}(x, y)$  to be the length of a shortest path between two points  $x, y \in \mathcal{G}$ . Since genetic material cannot be transmitted back in time, a more biologically relevant notion of distance between two points  $x, y \in \mathcal{G}$  would consist in considering only the paths that lie in the past of the focal points. Letting  $h(x) = t$  denote the *height* of a point  $x = (\ell, t) \in \mathcal{G}$  and  $x \wedge y$  the most recent common ancestor of  $x$  and  $y$ , this notion of distance can be expressed as

$$h(x) + h(y) - 2h(x \wedge y).$$

However, this does not define a distance in the mathematical sense, as the triangle inequality is not satisfied when the network contains coalescence points.

In this document, we are mostly interested in the structure of  $\mathcal{G}$  conditioned on being large. More specifically, we consider  $(\mathcal{G}_n, d_{\mathcal{G}_n}, \lambda_{\mathcal{G}_n})$ , the metric measure space having the

law of  $(\mathcal{G}, d_{\mathcal{G}}, \lambda_{\mathcal{G}})$  conditioned on having  $n$  colors, and we study various limits of  $\mathcal{G}_n$  as  $n$  goes to infinity.

For this, it will be convenient to see  $\mathcal{G}$  as a decorated Galton–Watson tree. For each color  $k$ , let  $\mathcal{X}_k$  denote the subnetwork of  $\mathcal{G}$  formed by the lineages of color  $k$ , endowed with the information of which endpoint corresponds to the creation of the color  $k$  (henceforth referred to as the *root* of  $\mathcal{X}_k$ ) and of which of the other endpoints correspond to mutations as opposed to deaths. Let  $\mathcal{T}$  denote the genealogical tree of the colors – that is, the ordered tree whose vertices are the colors of the lineages and where  $k'$  is a child of  $k$  if and only if  $k'$  was created by the mutation of a lineage of color  $k$ , the children of a color being ordered according to the order of apparition of the corresponding mutations. Finally, let  $\mathcal{T}^*$  denote the tree  $\mathcal{T}$  where each vertex  $k$  is decorated by the corresponding network  $\mathcal{X}_k$ . Note that  $\mathcal{G}$  and  $\mathcal{T}^*$  contain the same information, since to reobtain  $\mathcal{G}$  from  $\mathcal{T}^*$  it suffices to glue, for each color  $k$  and each child of that color, the root of the decoration the  $i$ -th child of  $k$  to the endpoint of  $\mathcal{X}_k$  corresponding to its  $i$ -th mutation.

In Section 2, we list miscellaneous results that are used throughout the document, starting with properties of the process describing the dynamics of the number of lineages of a given color (the so-called *logistic branching process*). We then study the random variable  $M$  giving the number of new colors that a color produces over its lifetime, i.e. the offspring distribution of the Galton–Watson tree  $\mathcal{T}$ . We show that its expected value is given by

$$\mathbb{E}(M) = \mu \sum_{j \geq 1} \prod_{k=1}^j \frac{1}{\rho_k},$$

where  $\rho_k = \alpha + \mu + (k - 1)\beta$ . Since every color has an almost surely finite lifetime, the process generating  $\mathcal{G}$  goes extinct with probability 1 if and only if  $\mathbb{E}(M) \leq 1$ . We also show that the probability generating function of  $M$  can be expressed as the continued fraction

$$g(z) = \frac{\alpha + \mu z}{1 + \rho_1 - \frac{\alpha + \beta + \mu z}{1 + \rho_2 - \frac{\alpha + 2\beta + \mu z}{1 + \rho_3 - \ddots}}}$$

This expression makes it straightforward to numerically compute the probability of extinction of the process – that is, the smallest fixed-point of  $g$  in  $[0, 1]$ . Similarly, we give a characterization of the asymptotic growth rate of the total number of lineages that makes it possible to compute it in practical applications.

In Sections 3 and 4, we study the geometry of the network  $\mathcal{G}_n$ . Section 3 deals with the global, large-scale structure of  $\mathcal{G}_n$  as  $n$  goes to infinity. This structure is tree-like: in Theorem 3.12 we show that, letting  $|\mathcal{G}_n| = \lambda_{\mathcal{G}_n}(\mathcal{G}_n)$ , for some well-characterized constant  $C$  the rescaled space

$$\left( \mathcal{G}_n, \frac{C}{\sqrt{n}} d_{\mathcal{G}_n}, \frac{1}{|\mathcal{G}_n|} \lambda_{\mathcal{G}_n} \right)$$

converges to Aldous’ Brownian continuum random tree in distribution for the Gromov–Hausdorff–Prokhorov topology. Finally, Section 4 focuses on the local structure of  $\mathcal{G}_n$ : we show that  $\mathcal{G}_n$  rooted at a uniform point has a local weak limit, which we describe explicitly.

### 1.3 Comments and perspectives

Our proof of the convergence to the CRT is based on [48], where most of the ideas that we use in Section 3 can already be found. Nevertheless, some specificities of

our model – in particular the fact that the number of new colors produced by a color during its lifetime and the total length of the corresponding subnetwork are not bounded random variables – require a different treatment and have necessitated a fine-grained study of the logistic branching process with mutation. The existence of various local weak limits was also obtained in [48]. The ideas are similar, insofar as we are also dealing with local weak limits of blow-ups of Galton–Watson trees. However, there are some notable differences – such as the fact that our focal point is chosen uniformly with respect to the length measure of our time-embedded network (as opposed to uniformly on the vertices of the underlying graph) and that we are able to give a more explicit description of the local geometry of the limit.

In an effort to make this paper accessible to mathematical biologists who do not have specific knowledge about Galton–Watson trees or Gromov–Hausdorff–Prokhorov convergence, we have strived to make it as self-contained as possible by (1) providing detailed reminders about most of the notions and results that are used and (2) whenever possible, expressing our results as general statements that are not tied to our particular setting. In particular, Proposition 3.4 provides a general recipe for proving Gromov–Hausdorff–Prokhorov convergence to a random  $\mathbb{R}$ -tree, and Lemma 3.10 makes it straightforward to apply this proposition to decorated Galton–Watson trees.

We close this introduction by mentioning an interesting line of research: our study hinges on the fact that our model can be seen as a decorated Galton–Watson tree. This crucial connection stems from the fact that the hybridization rate is a 0-1 function of the phylogenetic distance, which has the simple form  $d(\ell, \ell') = 1$  if  $\ell$  and  $\ell'$  are the same color, and 0 otherwise. However, from a modelling point of view it would be more natural to use a more nuanced notion of phylogenetic distance, and to let the phylogenetic distance be a gradually decreasing function of that distance.

For instance – as a first step and in keeping with the idea of colors representing incompatibility alleles – one could let lineages carry several colors, and make the hybridization rate between two lineages a decreasing function of the number of colors that differ between these lineages. Based on the biological interpretation, one might expect such a model to have properties that are very similar to our model. However, because the link with Galton–Watson trees is lost, it is not clear whether this is the case, and how to study this. Therefore, studying such models of phylogenetic networks – whose large-scale geometry is expected to be tree-like, even though there is no immediate, rigorous connection with branching processes – seems like an interesting and challenging problem that will likely require developing new tools and methods.

## 2 Probability of extinction and growth rate

### 2.1 The logistic branching process

Throughout this document, we denote by  $X = (X_t : t \geq 0)$  the process counting the number of lineages of the first color. It is a birth-death process started from  $X_0 = 1$ , killed in 0, and with transition rates:

- $k \rightarrow k + 1$  at rate  $k$ ;
- $k \rightarrow k - 1$  at rate  $k\rho_k = (\alpha + (k - 1)\beta + \mu)k$ .

This process has been called the *branching process with logistic growth* (or, more succinctly, the *logistic branching process*) and has been studied, e.g, in [34, 43]. It is also a special case of a branching process with interactions, see [13, 29, 42].

The qualitative behaviour of  $X$  can be described as transient fluctuations in a potential well. Indeed, letting  $K = 1 + (1 - \alpha - \mu)/\beta$ , when  $X$  is smaller than  $K$  it tends to increase whereas when it is greater than  $K$  it tends to decrease. Thus, in particular when  $K$  is

large, typical trajectories of  $X$  quickly relax towards a quasi-stationary distribution and then fluctuate until they eventually hit 0, which happens in finite time with probability 1.

Although this qualitative behaviour is well-understood, the quadratic term in the death rate makes the obtention of exact quantitative results difficult – and, to some extent, impossible. For instance, a classic approach to study birth-death processes consists in using the Kolmogorov forward equations to obtain a characterization of the probability generating function  $f(z, t) = \mathbb{E}(z^{X_t})$  as the solution of a partial differential equation. Here, standard calculations show that  $f$  is the unique analytic solution on  $[0, 1] \times \mathbb{R}_+$  of

$$\partial_t f = (z - \alpha - \mu)(z - 1) \partial_z f + \beta z(1 - z) \partial_{zz} f$$

with  $f(z, 0) = z$ . However, this partial differential equation is known not to have a closed-form solution – see Proposition 1.2 in [1]. Another powerful approach to study birth-death processes is the integral representation of the transition probabilities using orthogonal polynomials introduced by Karlin and McGregor [30, 31], but to our knowledge in the case of the logistic branching process this does not yield useful explicit expressions.

One of the important properties of the logistic branching process is that it comes down from infinity (meaning that there is a unique way to start it from  $X_0 = \infty$  and yet have  $X_t < \infty$  for any  $t > 0$ ), as shown by Lambert in [34]. We denote by  $\mathbb{E}_\infty$  the expectation under the initial condition  $X_0 = \infty$ . A recurring quantity throughout this paper is the extinction time  $T = \inf\{t \geq 0 : X_t = 0\}$ . In his Theorem 2.3, Lambert gives Laplace transform of  $T$  under  $\mathbb{E}_\infty$  as a function of the solution of a Riccati equation, and shows that its expected value is finite. In fact,  $T$  also has finite exponential moments under  $\mathbb{E}_\infty$ . This can be deduced, e.g. from [24, Proposition 2.4] or [5, Proposition 2.2], and will play an important role in our study – even though, for reasons that will become clear, we actually need a variant of this result (namely Lemma 3.6 in Section 3.3).

Because in our setting the mutations associated to the logistic branching process play a crucial role, the following change of measure will be useful. In what follows, we fix the parameters  $\alpha$  and  $\beta$ , and we denote by  $\mathbb{E}_\mu$  the expectation under a logistic branching process with mutation rate  $\mu$ .

**Proposition 2.1.** *Let  $M$  be the number of mutations associated to the logistic branching process  $X$ , and let  $L = \int_0^\infty X(t) dt$ . Then, for any positive number  $s$  and any nonnegative measurable functional  $f$  of the trajectory of  $X$ ,*

$$\mathbb{E}_\mu(f(X) s^M) = \mathbb{E}_{s\mu}(f(X) e^{(s-1)\mu L}).$$

*Proof.* For this proof, it will be convenient to use the “extended” chain  $\bar{X}$  which, in addition to the trajectory of  $X$ , contains the information about which transitions correspond to mutations. In other words,  $\bar{X}$  is a continuous-time Markov chain on  $\bar{\mathbb{N}} := \mathbb{N} \times \{\circ, \bullet\}$ , where the second coordinate  $\bullet$  means: “the last transition was a mutation”. The chain is started from  $(1, \circ)$  and its transition rates are the following:

$$\begin{aligned} (i, \circ) &\rightarrow (j, \circ) && \text{at rate } q_{ij}^\circ := i \mathbb{1}_{\{j=i+1\}} + (\rho_i - \mu) i \mathbb{1}_{\{j=i-1\}}, \\ (i, \bullet) &\rightarrow (j, \circ) && \text{at rate } q_{ij}^\circ, \\ (i, \circ) &\rightarrow (j, \bullet) && \text{at rate } q_{ij}^\bullet := \mu i \mathbb{1}_{\{j=i-1\}}, \\ (i, \bullet) &\rightarrow (j, \bullet) && \text{at rate } q_{ij}^\bullet. \end{aligned}$$

For convenience, we also use the notation  $q_i = \sum_j q_{ij}$ , i.e. in our case  $q_i = (1 + \rho_i) i$ .

Note that  $\bar{X}$  almost surely has a finite number of jumps before hitting 0 – i.e. hitting either  $(0, \circ)$  or  $(0, \bullet)$ , both of which are absorbing states. Let  $\Gamma$  be the embedded chain

of  $\bar{X}$ , whose distribution is given by:

$$\mathbb{P}(\Gamma = \gamma) = \left( \prod_{i \overset{\circ}{\rightarrow} j} \frac{q_{ij}^{\circ}}{q_i} \right) \left( \prod_{i \overset{\bullet}{\rightarrow} j} \frac{q_{ij}^{\bullet}}{q_i} \right),$$

where  $\gamma$  is a finite sequence of states, starting at  $(1, \circ)$  and ending at  $0$ ; the first product is on the transitions from either  $(i, \circ)$  or  $(i, \bullet)$  to  $(j, \circ)$  in  $\gamma$ , and the second product is defined similarly.

Also, given  $\{\Gamma = \gamma\}$ , where  $\gamma$  is of length  $n + 1$ , with  $n \geq 1$ , the distribution of the holding times  $\Xi = (\xi_1, \dots, \xi_n)$  in each of the  $n$  first states is given by:

$$\mathbb{P}(\Xi \in dx \mid \Gamma = \gamma) = \prod_{k=1}^n q_{\gamma_k} e^{-q_{\gamma_k} x_k} dx_1 \dots dx_n,$$

where, by a slight abuse of notation,  $\gamma_k \in \mathbb{N}$  denotes the first coordinate  $i$  of the  $k$ -th state  $(i, \circ)$  or  $(i, \bullet)$  of  $\gamma$ .

Noting that  $M = M(\Gamma) = \sum_{i \overset{\bullet}{\rightarrow} j} 1$  and that  $L = L(\Gamma, \Xi) = \sum_{k=1}^n \Gamma_k \xi_k$ , and recalling that  $q_{ij}^{\bullet} = \mu_i \mathbb{1}_{\{j=i-1\}}$ , we get

$$\begin{aligned} & \mathbb{E}_{\mu}(f(\Gamma, \Xi) s^M) \\ &= \sum_{\gamma} s^{M(\gamma)} \int f(\gamma, x) \left( \prod_{i \overset{\circ}{\rightarrow} j} q_{ij}^{\circ} \right) \left( \prod_{i \overset{\bullet}{\rightarrow} j} \mu_i \right) \left( \prod_{k=1}^n e^{-q_{\gamma_k} x_k} \right) dx \\ &= \sum_{\gamma} \int f(\gamma, x) \left( \prod_{i \overset{\circ}{\rightarrow} j} q_{ij}^{\circ} \right) \left( \prod_{i \overset{\bullet}{\rightarrow} j} s \mu_i \right) \left( \prod_{k=1}^n e^{-(q_{\gamma_k} + (s-1)\mu_{\gamma_k}) x_k} \right) \left( \prod_{k=1}^n e^{(s-1)\mu_{\gamma_k} x_k} \right) dx \\ &= \mathbb{E}_{s\mu} \left( f(\Gamma, \Xi) e^{(s-1)\mu L} \right), \end{aligned}$$

where the sums run on all sequences  $\gamma$  that end in  $0$  after a finite number of steps, and the integrals are over  $\mathbb{R}_+^n$ , with  $n$  is the number of jumps of  $\gamma$ . This concludes the proof since  $X$ , seen as a random variable living in the Skorokhod space  $\mathbb{D}(\mathbb{R}_+, \mathbb{N})$ , is clearly a measurable functional of the pair  $(\Gamma, \Xi)$ .  $\square$

Finally, it will also be useful to describe the trajectory of  $X$  as seen from a uniform mutation time. For this, we first need to introduce some notation for yet another type of changes of measures that will appear several times in the paper.

**Notation 2.2.** Let  $A$  and  $B$  be random variables defined on the same probability space such that  $B$  is almost surely nonnegative and  $0 < \mathbb{E}(B) < \infty$ . We write  $\mathcal{L}(A \dagger B)$  for the distribution of  $A$  biased by  $B$ , that is, under the probability measure defined by  $\mathbb{P}(\cdot \dagger B) = \mathbb{E}(\mathbb{1}_{\{\cdot\}} B) / \mathbb{E}(B)$ .

With this notation, by “the process  $X$  as seen from a uniform mutation time” we rigorously mean

$$X^m \sim \mathcal{L}((X_{U+t})_{-U \leq t < T-U} \dagger M),$$

where  $U$  is chosen uniformly at random among the atoms of the point process  $\mathcal{M}$  giving the times of the mutations associated to the trajectory of  $X$ ; note that  $U$  need not be defined when  $\mathcal{M}$  is empty because  $\mathbb{P}(M = 0 \dagger M) = 0$ . Equivalently, the distribution of  $X^m$  is characterized by

$$\mathbb{E}(F(X^m)) = \frac{1}{\mathbb{E}(M)} \mathbb{E} \left[ \sum_{u \in \mathcal{M}} F((X_{u+t})_{-u \leq t < T-u}) \right]$$

for any measurable bounded functional  $F$ .

It turns out that it is also possible to obtain  $X^m$  by a simple construction. For this, we need to introduce one last definition.

**Definition 2.3.** Let  $f: [0, T_f[ \rightarrow \mathbb{R}$  and  $g: [0, T_g[ \rightarrow \mathbb{R}$  be two càdlàg functions. The back-to-back pasting of  $f$  to  $g$  is the càdlàg function  $f \wr g: [-T_f, T_g[ \rightarrow \mathbb{R}$  defined by

$$f \wr g: t \mapsto \begin{cases} \lim_{s \downarrow t} f(-s) & \text{if } t < 0 \\ g(t) & \text{if } t \geq 0. \end{cases}$$

**Proposition 2.4.** Let  $\nu_\circ$  be the probability distribution on the positive integers defined by

$$\nu_\circ(n) = C \prod_{k=1}^n \frac{1}{\rho_k}, \tag{2.1}$$

with  $C$  the corresponding normalizing constant. Let  $K \sim \nu_\circ$  and, conditional on  $K$ , let  $X'$  and  $X''$  be two independent realizations of the logistic branching process  $X$  started from  $X'_0 = K$  and  $X''_0 = K - 1$ . Then,

$$X^m \stackrel{d}{=} X' \wr X''.$$

The proof uses general results about the decomposition of trajectories of Markov chains that are recalled in Appendix A.1, and therefore is deferred to the end of that appendix.

## 2.2 Offspring distribution and extinction probability of $\mathcal{T}$

In this section, we focus on the law of the random variable  $M$  giving the number of mutations of a color (that is, the number of new colors that it produces; also the offspring distribution of the Galton–Watson tree  $\mathcal{T}$ ). Our main result is the following theorem, on which much of our study relies.

**Theorem 2.5.** Let  $M$  be the offspring distribution of  $\mathcal{T}$ , and let  $g$  be its probability generating function. Then, letting  $\rho_k = \alpha + \mu + (k - 1)\beta$ ,

$$\mathbb{E}(M) = \mu \sum_{j \geq 1} \prod_{k=1}^j \frac{1}{\rho_k}$$

and

$$g(z) = \frac{\alpha + \mu z}{1 + \rho_1 - \frac{\alpha + \beta + \mu z}{1 + \rho_2 - \frac{\alpha + 2\beta + \mu z}{1 + \rho_3 - \ddots}}},$$

which, using Gauss’s notation for continued fractions, can be written

$$g(z) = -\mathbf{K}_{k=1}^{\infty} \frac{\mu - \rho_k - \mu z}{1 + \rho_k}.$$

Moreover,  $g$  is meromorphic on  $\mathbb{C}$ . The radius of convergence of its power series expansion around 0 is  $R > 1$ , and  $g$  has a pole in  $R$ .

*Proof.* Let  $\tilde{X}$  be the embedded chain of  $X$ , that is,  $\tilde{X}_i = X_{\tau_i}$  where  $\tau_0 = 0$  and  $\tau_{i+1} = \inf\{t > \tau_i : X_t \neq X_{\tau_i}\}$ . Note that, conditional on the trajectory of  $\tilde{X}$ , each step from  $k$  to  $k - 1$  corresponds to a mutation with probability  $p_k := \mu/\rho_k$ , independently of everything else. Let us refer to a trajectory of  $\tilde{X}$  started from  $k$  and killed when it

first hits  $k - 1$  as a  $k$ -excursion of  $\tilde{X}$ . Every  $k$ -excursion of  $\tilde{X}$  can be decomposed into  $N_k$  independent  $(k + 1)$ -excursions, followed by a single step from  $k$  to  $k - 1$ . By the strong Markov property,  $N_k$  follows a geometric distribution on  $\{0, 1, \dots\}$  with parameter  $\theta_k := \rho_k / (1 + \rho_k)$ . Therefore, letting  $M_k$  have the distribution of the number of mutations along a  $k$ -excursion of  $\tilde{X}$ , we have

$$M_k \stackrel{d}{=} \sum_{i=1}^{N_k} M_{k+1}^{(i)} + \text{Ber}(p_k), \tag{2.2}$$

where  $M_{k+1}^{(i)}$  are independent copies of  $M_{k+1}$  that are also independent of  $N_k$ , and  $\text{Ber}(p_k)$  is a Bernoulli variable that is independent of everything else.

Applying Wald’s formula to Equation (2.2) gives

$$\mathbb{E}(M_k) = \frac{1}{\rho_k} \mathbb{E}(M_{k+1}) + \frac{\mu}{\rho_k},$$

and solving this first-order linear recurrence yields the formula for the expected value of  $M \stackrel{d}{=} M_1$ .

Let us now turn to the generating function of  $M$  and let  $g_k(z) := \mathbb{E}(z^{M_k})$ , defined for  $|z| < R_k$ , where  $R_k = \sup\{r \geq 0 : \mathbb{E}(r^{M_k}) < \infty\}$ . Note that since  $M_k$  stochastically dominates  $M_{k+1}$ , the  $R_k$  are nondecreasing. Then, for all  $z$  such that  $|(1 - \theta_k) g_{k+1}(z)| < 1$  – note that this is true for all  $|z| < 1$  – we have

$$\begin{aligned} g_k(z) &= (1 - p_k + p_k z) \sum_{i \geq 0} g_{k+1}(z)^i \theta_k (1 - \theta_k)^i \\ &= \frac{(1 - p_k + p_k z) \theta_k}{1 - (1 - \theta_k) g_{k+1}(z)} \\ &= \frac{\alpha + (k - 1)\beta + \mu z}{1 + \rho_k - g_{k+1}(z)}. \end{aligned}$$

This gives the representation of  $g = g_1$  as the continued fraction of the theorem.

We now show that  $g$  extends to a meromorphic function on all of  $\mathbb{C}$ . Note that  $M_k$  is stochastically dominated by  $H_k$ , the hitting time of 0 by the simple random walk started from 1 that goes up with probability  $1 - \theta_k$  and down with probability  $\theta_k$ , independently of its current position. A standard calculation (see e.g. [11, Section 6.4]) shows that the probability generating function of  $H_k$  is

$$h_k(z) = \frac{1 - \sqrt{1 - 4\theta_k(1 - \theta_k)z^2}}{2(1 - \theta_k)z},$$

whose power series expansion around zero has a radius of convergence equal to  $(4\theta_k(1 - \theta_k))^{-1/2}$ . Moreover,  $\mathbb{P}(H_k < \infty) = h_k(1^-) = \frac{1 - |1 - 2\theta_k|}{2(1 - \theta_k)}$  is equal to 1 for all  $k$  large enough, and  $\mathbb{P}(H_k < \infty) = 1$  implies that  $\mathbb{E}(z^{H_k}) = h_k(z)$  inside the disk of convergence of  $\mathbb{E}(z^{H_k})$ . Since  $\mathbb{E}(z^{M_k}) \leq \mathbb{E}(z^{H_k})$  for  $z \geq 1$  and since  $(4\theta_k(1 - \theta_k))^{-1/2} \rightarrow +\infty$  as  $k$  goes to infinity, this shows that for any  $r > 0$  there exists  $k_r$  such that  $g_{k_r}$  is analytic on  $\mathbb{D}_r = \{z : |z| < r\}$ . It then follows by induction that  $g_{k_r-1}, \dots, g_1$  are meromorphic on  $\mathbb{D}_r$ .

Finally, recall that the *dominant* singularities of a function that is analytic at 0 are those singularities that are closest to the origin. To see that the dominant pole of  $g$  is in  $R > 1$ , note that since the power series representation of  $g$  around the origin has nonnegative coefficients, Pringsheim’s theorem (see e.g. [23, Theorem IV.6]) ensures that it has a dominant singularity in  $]0, +\infty[$ . Since  $g(1) = 1$  is finite and since all singularities of  $g$  are poles, this means that  $g$  has no singularity in 1. Hence,  $g$  has a dominant pole in  $R$  for some  $R > 1$ . □

## A branching process with coalescence for phylogenetic networks

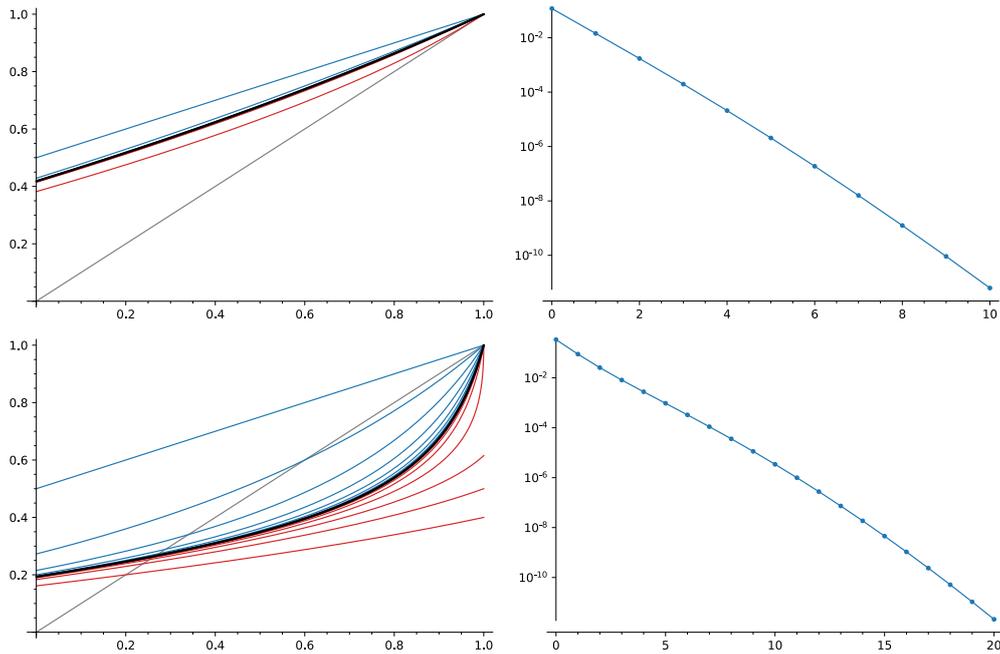


Figure 2: Left, the generating function of  $M$ , in black, along with the modified convergents of Proposition 2.6 giving upper (blue) and lower (red) bounds; right, supremum on  $[0, 1]$  of the difference between the upper and the lower bound, as a function of  $n$ . Top, subcritical regime, with  $(\alpha, \beta, \mu) = (1, 1, 1)$ ; bottom, supercritical regime, with  $(\alpha, \beta, \mu) = (0.2, 0.2, 0.2)$ .

One of the advantages of the expression of  $g$  as a generalized continued fraction is that this makes its numerical evaluation straightforward and very efficient. Indeed, modified convergents of this continued fraction provide us with upper and lower bounds on  $g$ , as the next proposition shows. The rapid convergence of these bounds is illustrated in Figure 2.

**Proposition 2.6.** *For all  $z \in [0, 1]$  and all  $n \geq 1$ ,*

$$\frac{1 + \rho_1 - \frac{\alpha + \mu z}{\alpha + \beta + \mu z}}{1 + \rho_2 - \frac{\alpha + 2\beta + \mu z}{\alpha + (n-1)\beta + \mu z}} \leq g(z) \leq \frac{1 + \rho_1 - \frac{\alpha + \mu z}{\alpha + \beta + \mu z}}{1 + \rho_2 - \frac{\alpha + 2\beta + \mu z}{\alpha + (n-1)\beta + \mu z}}$$

where

$$\bar{g}_n(z) = \frac{1}{2} \left( 1 + \rho_n - \sqrt{(1 - \rho_n)^2 - 4\mu(z - 1)} \right).$$

Letting  $\bar{R}_n \geq 1$  denote the radius of convergence of the power series expansion around 0 of the left-hand side, for  $n$  large enough the reversed inequalities hold for  $z \in [1, \bar{R}_n[$ . Moreover, the difference between the right-hand side and the left-hand side is  $O(n^c \beta^{-n} / n!)$ , where  $c = 1 - \frac{\alpha + \mu}{\beta}$ , uniformly in  $z \in [0, 1]$ .

*Proof.* We give a probabilistic proof. Although it is possible to give a shorter analytic proof, we think that the probabilistic one is more instructive.

Let  $\bar{M}^{(n)}$  denote the number of mutations associated to a modified version  $\bar{X}^{(n)}$  of the process  $X$ , where coalescences happen at rate  $n(n - 1)\beta$  instead of  $k(k - 1)\beta$  whenever

$\bar{X}^{(n)} = k \geq n$ . Let  $\underline{M}^{(n)}$  denote the number of mutations that correspond to transitions from  $k$  to  $k - 1$  with  $k \leq n$  in the original process  $X$ . For all  $n \geq 1$ , we have the stochastic dominations

$$\underline{M}^{(n)} \stackrel{d}{\leq} M \stackrel{d}{\leq} \bar{M}^{(n)}. \tag{2.3}$$

Let  $\bar{G}_n(z)$  and  $\underline{G}_n(z)$  denote the left- and right-hand sides of the inequality of the proposition, respectively. The same reasoning as in the proof of Theorem 2.5 shows that  $\bar{G}_n$  and  $\underline{G}_n$  are, respectively, the generating functions of  $\bar{M}^{(n)}$  and of  $\underline{M}^{(n)}$ . Note however that, in the case of  $\bar{M}^{(n)}$ , for small values of  $n$  there can be a positive probability that  $\bar{X}^{(n)}$  never hits 0 – in which case it is not possible to decompose its trajectory into finite excursions. Nevertheless, letting  $A_k$  be the event that  $\bar{X}^{(n)}$  started from  $k$  never hits  $k - 1$  and  $\bar{N}_k^{(n)}$  a geometric variable with parameter  $\rho_{k \wedge n} / (1 + \rho_{k \wedge n})$ , we have

$$\mathbb{P}(A_k) = \mathbb{P}\left(\bigcup_{i=1}^{\bar{N}_k^{(n)}} A_{k+1}^{(i)}\right),$$

where  $A_{k+1}^{(i)}$  are independent realizations of  $A_{k+1}$  that are also independent of  $\bar{N}_k^{(n)}$ . Since, up to a negligible event,  $A_k$  and  $\{\bar{M}_k^{(n)} = \infty\}$  are equal, this means that Equation (2.2) holds for  $\bar{M}_k^{(n)}$ , mutatis mutandis, even when  $\mathbb{P}(\bar{M}_k^{(n)} = \infty) > 0$ . Finally, the expression of  $\bar{g}_n$  is obtained by solving

$$\bar{g}_n(z) = \frac{\alpha + (k - 1)\beta + \mu z}{1 + \rho_k - \bar{g}_n(z)},$$

since, by construction,  $\bar{M}_{n+1}^{(n)} \stackrel{d}{=} \bar{M}_n^{(n)}$ .

Being generating functions,  $\bar{G}_n$  and  $\underline{G}_n$  are analytic at 0 with radius of convergence at least 1, and we have  $\mathbb{E}(z^{\bar{M}^{(n)}}) = \bar{G}_n(z)$  and  $\mathbb{E}(z^{\underline{M}^{(n)}}) = \underline{G}_n(z)$  for all  $z \in [0, 1[$ . Combining this with (2.3) and taking the limit  $z \rightarrow 1^-$  proves the inequality of the proposition for all  $z \in [0, 1]$ .

For  $z > 1$ , since  $\underline{M}^{(n)}$  is stochastically dominated by  $M$  and since  $M$  is almost surely finite, for all  $n$  we have  $\underline{G}_n(z) = \mathbb{E}(z^{\underline{M}^{(n)}}) \leq \mathbb{E}(z^M) = g(z)$  for all  $z \in [1, R[$ , where  $R$  is the radius of convergence of  $g$  around 0. Similarly, for  $n$  large enough  $\mathbb{P}(\bar{M}^{(n)} < +\infty) = 1$  and thus  $g(z) \leq \bar{G}_n(z)$  for all  $z \in [1, \bar{R}_n[$ , where  $\bar{R}_n$  is the radius of convergence of  $\bar{G}_n$  around 0. Note however that for small  $n$  we can have  $\bar{G}_n(1) = \mathbb{P}(\bar{M}^{(n)} < +\infty) < 1$ , and thus  $\bar{G}_n(z) < g(z)$  for  $z \in [1, \bar{R}_n[$ .

Finally, to see that  $\sup_{z \in [0, 1]} |\underline{G}_n(z) - \bar{G}_n(z)| = O(n^c \beta^{-n} / n!)$ , note that

$$\frac{\alpha + (k - 1)\beta + \mu z}{1 + \rho_k - A} - \frac{\alpha + (k - 1)\beta + \mu z}{1 + \rho_k - B} = \frac{(\alpha + (k - 1)\beta + \mu z)(A - B)}{(1 + \rho_k - A)(1 + \rho_k - B)},$$

so that for all  $z, A, B \in [0, 1]$ ,

$$\left| \frac{\alpha + (k - 1)\beta + \mu z}{1 + \rho_k - A} - \frac{\alpha + (k - 1)\beta + \mu z}{1 + \rho_k - B} \right| \leq \frac{1}{\rho_k} |A - B|.$$

Since  $|\bar{g}_n(z) - 1| \leq 1$ , an immediate induction gives

$$\sup_{z \in [0, 1]} |\underline{G}_n(z) - \bar{G}_n(z)| \leq \prod_{k=1}^n \frac{1}{\rho_k} \sim \Gamma\left(\frac{\alpha + \mu}{\beta}\right) n^{1 - \frac{\alpha + \mu}{\beta}} \beta^{-n} / n!,$$

finishing the proof. □

Besides numerical evaluation, the bounds of Proposition 2.6 can be used to obtain rigorous bounds on the probability of extinction of the model. For instance, taking  $n = 2$  for the left-hand side,  $n = 3$  for the right-hand side, and finding the corresponding fixed points, we get the simple bounds

$$\frac{\alpha}{2\mu} \left( \beta + \mu - 1 + \sqrt{(\beta + \mu - 1)^2 + 4\mu} \right) \leq p_{\text{ext}} \leq \frac{\alpha((\alpha + \beta + \mu)(\alpha + 2\beta + \mu) + \mu)}{\mu(1 + 2\alpha + 2\beta + \mu)}.$$

In fact, it is possible to get one such upper bound up to  $n = 9$ . However, the resulting expression, although very sharp, is too complex to be of any practical use – so we do not reproduce it here.

Let us now point out two immediate consequences of Theorem 2.5 that will be useful in the rest of this document.

**Corollary 2.7.**

(i)  $M$  has finite exponential moments:

$$\exists \varepsilon > 0 \text{ s.t. } \mathbb{E}(e^{\varepsilon M}) < +\infty.$$

(ii) There is an exponential tilt of  $M$  with mean 1:

$$\exists \zeta > 0 \text{ s.t. } \frac{\mathbb{E}(M\zeta^M)}{\mathbb{E}(\zeta^M)} = 1.$$

*Proof.* (i) is merely saying that the radius of convergence of  $g$  is greater than 1; (ii) is a classic consequence of the fact that  $g(s) \rightarrow +\infty$  as  $s \uparrow R$ , see e.g. point (iv) of Lemma 3.1 in [28]. For the sake of completeness, we recall the proof here: for any  $a \geq 0$  and any  $s \in [0, R[$ ,

$$\frac{\mathbb{E}(Ms^M)}{\mathbb{E}(s^M)} - a \geq \frac{\mathbb{E}((M - a)s^M \mathbb{1}_{\{M < a\}})}{\mathbb{E}(s^M)}.$$

Since  $|\mathbb{E}((M - a)s^M \mathbb{1}_{\{M < a\}})| < aR^a$  and  $\mathbb{E}(s^M) \rightarrow +\infty$  as  $s \uparrow R$ , the right-hand side of this inequality goes to 0 as  $s \uparrow R$ . Therefore,  $\lim_{s \uparrow R} \mathbb{E}(Ms^M)/\mathbb{E}(s^M) \geq a$  for all  $a \geq 0$ , i.e.  $\mathbb{E}(Ms^M)/\mathbb{E}(s^M) \rightarrow +\infty$  as  $s \uparrow R$ . The existence of  $\zeta \in ]0, R[$  such that  $\mathbb{E}(M\zeta^M)/\mathbb{E}(\zeta^M) = 1$  follows by continuity.  $\square$

The main consequence of Corollary 2.7 is that, for all  $\alpha, \beta, \mu > 0$ , when conditioned to have  $n$  vertices  $\mathcal{T}$  is distributed as a critical Galton–Watson tree conditioned to have  $n$  vertices. We will come back to this in Section 3.

Finally, we close this section with a brief remark about  $\mathcal{M}$ , the point process of mutation times. We state it as a proposition for ease of reference, but it is not specific to our setting and follows readily from the infinitesimal definition of a continuous-time Markov chain – so we omit the proof.

**Proposition 2.8.** *Let  $\mathcal{M}$  be the point process on  $\mathbb{R}_+$  giving the birth times of the children of the first color (that is, every atom  $t \in \mathcal{M}$  corresponds to a mutation of a lineage of the first color). The intensity measure of  $\mathcal{M}$  is  $\mu \mathbb{E}(X_t) dt$  – that is, for any Borel set  $A \subset \mathbb{R}_+$ ,*

$$\mathbb{E}(\#\mathcal{M} \cap A) = \mu \int_A \mathbb{E}(X_t) dt.$$

In particular,  $\mathbb{E}(M) = \mu \mathbb{E}(\int_0^\infty X_t dt)$ .

### 2.3 Growth rate of the number of lineages

Let us start by focusing on the number the colors. We will turn to the number of lineages at the end of the section. Let  $Z_t$  denote the number of colors alive at time  $t$ . The process  $Z = (Z_t : t \geq 0)$  is a Crump–Mode–Jagers process, or CMJ for short, where individuals give birth according to a point process distributed as  $\mathcal{M}$ , the point process of mutations of the first color; and die after a time distributed as  $T$ , the extinction time of the logistic branching process started from 1.

The next proposition is an application of standard results from the theory of CMJ processes [14, 15, 27]. Essentially, CMJ processes grow/decrease exponentially with a growth rate known as their *Malthusian parameter*. Proposition 2.9 recalls the precise meaning of this “exponential growth” and gives the usual, generic characterization of the growth rate. Another characterization – one that is specific to our setting and makes it possible to compute the growth rate numerically – will be given in Proposition 2.10.

**Proposition 2.9.** *Let  $\lambda$  be the unique solution of*

$$\mathbb{E}\left(\sum_{t \in \mathcal{M}} e^{-\lambda t}\right) = 1 \quad \text{or, equivalently,} \quad \mu \mathbb{E}\left(\int_0^\infty X_t e^{-\lambda t} dt\right) = 1.$$

*Then,  $\lambda$  has the same sign as  $\mathbb{E}(M) - 1$ . Moreover,*

- (i) *If  $\mathbb{E}(M) > 1$ , then the process  $Z$  of the number of colors satisfies*

$$e^{-\lambda t} Z_t \xrightarrow[t \rightarrow \infty]{L^2, \text{ a.s.}} W,$$

*where  $W$  is a random variable with  $\mathbb{E}(W) = 1$  that is almost surely positive on non-extinction, i.e. on the event  $\{Z_t > 0 \text{ for all } t\}$ , and where the convergence holds both almost surely and in mean square.*

- (ii) *If  $\mathbb{E}(M) \leq 1$ , then  $\mathbb{E}(Z_t) \sim Ce^{\lambda t}$  as  $t \rightarrow \infty$  for some constant  $C > 0$ .*

*Proof.* The fact that the two characterizations of  $\lambda$  are equivalent follows from Proposition 2.8 and Campbell’s formula. The uniqueness of  $\lambda$  is standard, and so is the fact that  $\lambda$  is guaranteed to exist whenever  $\mathbb{E}(M) \geq 1$ , see [14, Section 6]. To see that  $\lambda$  also exists when  $\mathbb{E}(M) < 1$ , letting  $\tau$  denote the time of the first jump of  $X$  one can consider the random variable  $Y_\eta$  that takes the value  $e^{-\eta\tau}$  if the first jump of  $X$  is a mutation, and 0 otherwise. Thus,  $Y_\eta \leq \sum_{t \in \mathcal{M}} e^{-\eta t}$ . A straightforward calculation then shows that  $\mathbb{E}(Y_\eta) = \mu \int_0^\infty e^{-(1+\alpha+\mu+\eta)t} dt \rightarrow +\infty$  as  $\eta$  decreases to  $-(1 + \alpha + \mu)$  and therefore can be made greater than 1 by decreasing  $\eta$ . By the same reasoning as in the proof of the meromorphy of  $g$  in Theorem 2.5,  $\eta \mapsto \mathbb{E}(\sum_{t \in \mathcal{M}} e^{-\eta t})$  cannot jump to infinity. Since it is equal to  $\mathbb{E}(M) < 1$  when  $\eta = 0$ , the existence of  $\lambda$  follows by continuity.

Since Theorem 2.5 entails that  $\mathbb{E}(M^2) < \infty$ , the mean-square convergence in point (i) follows immediately from [15, Theorem 3.1]. Similarly, the almost sure convergence follows from [15, Theorem 3.2], provided that the intensity function of  $\mathcal{M}$ , namely  $m: t \mapsto \mu \mathbb{E}(X_t)$ , is differentiable and such that  $\int_0^\infty |m'(t)|^p dt < \infty$  for some  $p > 1$ . Now, since

$$\frac{d}{dt} \mathbb{E}(X_t) = (1 - \alpha - \mu) \mathbb{E}(X_t) - \beta \mathbb{E}(X_t(X_t - 1))$$

and that  $X_t$  is integer-valued, we have  $|m'(t)| < K \mathbb{E}(X_t^2)$  for some constant  $K$ . Thus, by Jensen’s inequality, to complete the proof of point (i) it suffices to show that  $\int_0^\infty \mathbb{E}(X_t^p) dt < \infty$  for some  $p > 2$ . Standard calculations, again using the decomposition of the trajectory of  $X$  into excursions, as in the proof of Theorem 2.5, show that

$$\int_0^\infty \mathbb{E}(X_t^p) dt = \sum_{j \geq 1} \left( \prod_{k=1}^j \frac{1}{\rho_k} \right) j^p.$$

Since  $\prod_{k=1}^j \rho_k^{-1} = O(n^c \beta^{-n}/n!)$  for some constant  $c$ , as already seen in the proof of Proposition 2.6, the integral is finite for all  $p$ , finishing the proof of point (i).

Finally, letting  $T$  denote the extinction time of  $X$ ,  $\mathbb{1}_{\{T>t\}} \leq X_t$  and therefore  $\mathbb{P}(T > t) \leq \mathbb{E}(X_t)$ . Since, by definition of  $\lambda$ ,  $\int_0^\infty \mathbb{E}(X_t) e^{-\lambda t} dt = 1/\mu$ , we have  $\int_0^\infty \mathbb{P}(T > t) e^{-\lambda t} dt < \infty$  and thus point (ii) follows from [14, Theorem 6.2].  $\square$

We now give another characterization of  $\lambda$ , which makes use of the measure  $\nu_\circ$  introduced in Proposition 2.4. Here, we let  $\mathbb{E}^\circ(\cdot)$  denote the expectation for the process  $X$  started from a random state with distribution  $\nu_\circ$ .

**Proposition 2.10.** *Let  $T$  denote the extinction time of  $X$ . Then, the growth rate  $\lambda$  is the unique solution of*

$$\mathbb{E}(M) \mathbb{E}^\circ(e^{-\lambda T}) = 1.$$

Furthermore,

$$\mathbb{E}(M) \mathbb{E}^\circ(e^{-\lambda T}) = \mu \sum_{j \geq 1} \prod_{k=1}^j \frac{f_k(\lambda)}{\rho_k},$$

where  $f_k(\lambda)$  is given by the continued fraction

$$f_k(\lambda) = \frac{\rho_k}{1 + \rho_k + \frac{\lambda}{k} - \frac{\rho_{k+1}}{1 + \rho_{k+1} + \frac{\lambda}{k+1} - \dots}} = -\prod_{i=k}^\infty \frac{-\rho_i}{1 + \rho_i + \frac{\lambda}{i}}.$$

The interest of this proposition is that, since the functions  $f_k(\lambda)$  can be evaluated efficiently, so can  $\mathbb{E}(M) \mathbb{E}^\circ(e^{-\lambda T})$ . This makes it straightforward to determine  $\lambda$  numerically, for instance using the bisection method.

*Proof.* The first part of the proposition is a consequence of the standard characterization of  $\lambda$ , which is recalled in Proposition 2.9, and of the construction of the process  $X^m$  given in Proposition 2.4. Indeed, first note that  $\mathbb{E}(\sum_{t \in \mathcal{M}} e^{-\lambda t}) = \mathbb{E}(M) \mathbb{E}(e^{-\lambda U} \dagger M)$ , where  $U$  is a uniform atom of  $\mathcal{M}$ , and also corresponds to minus the infimum of the times for which  $X^m$  is defined. Second, recall that  $X^m$  is distributed as  $X' \wr X''$ , where  $X'$  is distributed as  $X$  started from  $\nu_\circ$ , and that in this construction  $U$  corresponds to the extinction time of  $X'$ . As a result,  $\mathbb{E}(e^{-\lambda U} \dagger M) = \mathbb{E}^\circ(e^{-\lambda T})$ .

To express  $\mathbb{E}(M) \mathbb{E}^\circ(e^{-\lambda T})$  as a function of  $\lambda$ , for  $k \geq 1$  let

$$f_k(\lambda) = \mathbb{E}(e^{-\lambda T_{k-1}} \mid X_0 = k),$$

where  $T_{k-1}$  denotes the hitting time of  $k - 1$  by  $X$ . By the strong Markov property,

$$\mathbb{E}(M) \mathbb{E}^\circ(e^{-\lambda T}) = \mathbb{E}(M) \sum_{j \geq 1} \nu_\circ(j) \prod_{k=1}^j f_k(\lambda).$$

From the expression of  $\mathbb{E}(M)$  in Theorem 2.5, we see that the normalizing constant in Equation (2.1) where  $\nu_\circ$  is defined is  $C = \mu/\mathbb{E}(M)$ , from which deduce that

$$\mathbb{E}(M) \nu_\circ(j) = \mu \prod_{k=1}^j \frac{1}{\rho_k}.$$

Therefore, to finish the proof it only remains to show that

$$f_k(\lambda) = \frac{\rho_k}{1 + \rho_k + \frac{\lambda}{k} - f_{k+1}(\lambda)}.$$

The reasoning is exactly the same as for the expression of the generating function of  $M$  in Theorem 2.5, so we do not detail it.  $\square$

So far, we have been focusing on the growth rate of  $Z_t$ , the number of colors at time  $t$ . But from a biological point of view it is arguably more natural to consider  $\Upsilon_t$ , the number of lineages at time  $t$ . We therefore close this section with a proposition showing that the asymptotic growth rate of the number of lineages is the same as that of the number of colors. For simplicity we do not try to state the results in full generality.

**Proposition 2.11.** *Let  $\lambda$  be the growth rate of  $Z$ , as given in Proposition 2.9 and 2.10, and let  $\Upsilon_t$  be the number of lineages alive at time  $t$ . If  $\lambda > 1$ , then*

$$e^{-\lambda t} \Upsilon_t \xrightarrow[t \rightarrow \infty]{\text{a.s.}} \Xi,$$

where  $\Xi$  is a random variable that is almost surely positive on non-extinction.

*Proof.* Again, this is a standard application of general results for CMJ processes counted with a random characteristic, see e.g. [41]. More specifically, let the characteristic associated to each color be the number of lineages of that color. Note that the characteristic of a color is not independent of its lifespan and of its reproduction, but that the characteristics of different colors are independent. Since  $\mathbb{E}(M) < \infty$ , Condition 5.1 in [41] holds with  $g(t) = e^{-\lambda t}$ . Moreover, by using the same argument as for  $M$  it is straightforward to show that the total number of jumps of  $X$  has finite exponential moments. Since  $X$  has bounded jumps, this implies that  $\mathbb{E}(\sup_t X_t) < \infty$ , and so Condition 5.2 in [41] holds with  $h(t) = e^{-\lambda t}$ . As a result, the proposition follows from [41, Theorem 5.4].  $\square$

### 3 Convergence to the CRT

In this section, we study the large-scale geometry of  $\mathcal{G}$ . We will show that, after being conditioned to have  $n$  colors and appropriately rescaled, as  $n$  goes to infinity  $\mathcal{G}$  converges in distribution to the Brownian continuum random tree (CRT) for the rooted Gromov–Hausdorff–Prokhorov topology.

The Brownian CRT, introduced by Aldous in [4], is the universal scaling limit of critical Galton–Watson trees when the offspring distribution has finite variance. Since its first description as a random subset of  $\ell^1$  obtained by successively glueing segments of random lengths along orthogonal directions, it has become standard (see e.g. [35, Section 2] and [22, Section 2.4]) to view it as the random rooted compact metric probability space  $(\mathcal{C}, r, d, \lambda)$  defined in the following way:

- Take a standard Brownian excursion  $(e(t))_{t \in [0,1]}$ .
- Define a pseudo-metric  $d_e$  on  $[0, 1]$  by  $d_e(x, y) = e(x) + e(y) - 2 \inf_{z \in [x,y]} e(z)$ , where  $[x, y]$  is a slight abuse of notation for the segment  $[x \wedge y, x \vee y]$ .
- Let  $(\mathcal{C}, d)$  be the quotient metric space obtained by identifying the points of  $[0, 1]$  at distance zero for  $d_e$ , and let the root  $r \in \mathcal{C}$  be the equivalence class of 0.
- Let  $\lambda$  be the pushforward on  $\mathcal{C}$  of the Lebesgue measure on  $[0, 1]$ .

The rest of this section is organized as follows: first, we give a brief reminder about convergence in the rooted Gromov–Hausdorff–Prokhorov topology. Coming back to our model, we then detail how to condition  $\mathcal{G}$  on having  $n$  colors, and we introduce some notation. Finally, we prove a series of technical lemmas which, when put together, readily give us the desired convergence to the CRT.

### 3.1 The rooted Gromov–Hausdorff–Prokhorov distance

Here we recall, mostly without proof, the minimal set of notions about convergence of metric probability spaces that are needed to state and prove our results. More detailed treatments can be found, e.g. in [39, Section 6] or in [22, Section 4]. In particular, Proposition 3.4 below provides a general-purpose, simple way to establish convergence to the CRT by following the approach used in [48]. See also [44] for related results.

Since our network  $\mathcal{G}$  has a distinguished point, namely the point that corresponds to the first lineage at time 0, it is natural to work with a rooted version of the Gromov–Hausdorff–Prokhorov distance. We adapt the definition of [39, Section 6.2] to the rooted setting: let  $\mathbb{M}$  be the set<sup>1</sup> of isometry classes of rooted compact metric probability spaces  $\mathcal{X} = (\mathcal{X}, r, d, \lambda)$ , where  $r \in \mathcal{X}$  is called the root of  $\mathcal{X}$ ;  $d$  is a metric on  $\mathcal{X}$ ; and  $\lambda$  is a probability measure on  $\mathcal{X}$ . The rooted Gromov–Hausdorff–Prokhorov distance  $d_{\text{GHP}}(\mathcal{X}, \mathcal{X}')$  between two elements  $(\mathcal{X}, r, d, \lambda)$  and  $(\mathcal{X}', r', d', \lambda')$  of  $\mathbb{M}$  is defined as the infimum of the  $\varepsilon > 0$  such that there exists a well-defined metric  $\delta$  on the disjoint union  $\mathcal{Y} := \mathcal{X} \sqcup \mathcal{X}'$  satisfying:

- (i) For all  $x, y \in \mathcal{X}$  and  $x', y' \in \mathcal{X}'$ ,  $\delta(x, y) = d(x, y)$  and  $\delta(x', y') = d'(x', y')$ .
- (ii)  $\delta(r, r') \leq \varepsilon$ .
- (iii) The Hausdorff distance  $\delta_{\text{H}}(\mathcal{X}, \mathcal{X}')$  between  $\mathcal{X}$  and  $\mathcal{X}'$  is at most  $\varepsilon$ ; in other words,  $\mathcal{X}' \subset \mathcal{X}^\varepsilon$  and  $\mathcal{X} \subset (\mathcal{X}')^\varepsilon$ , where  $A^\varepsilon = \{y \in \mathcal{Y} : \exists x \in A, \delta(x, y) < \varepsilon\}$ .
- (iv) Extending  $\lambda$  and  $\lambda'$  to  $\mathcal{Y}$  via  $\lambda_{\mathcal{Y}}(A) = \lambda(A \cap \mathcal{X})$  and  $\lambda'_{\mathcal{Y}}(A) = \lambda'(A \cap \mathcal{X}')$ , the Prokhorov distance between  $\lambda_{\mathcal{Y}}$  and  $\lambda'_{\mathcal{Y}}$  is at most  $\varepsilon$ , i.e. for all Borel subset  $A \subset \mathcal{Y}$ , we have  $\lambda_{\mathcal{Y}}(A) \leq \lambda'_{\mathcal{Y}}(A^\varepsilon) + \varepsilon$ .

The space  $(\mathbb{M}, d_{\text{GHP}})$  is a complete separable metric space (see e.g. [39, Theorem 6 and Proposition 8] for a proof in the unrooted setting; we let the interested reader check that the proof carries over to the rooted setting, and refer them to [22, Section 4.3.3] where this is done for the Gromov–Hausdorff distance).

Because our metric spaces are tree-like, in our setting it will be more convenient to work with height processes than to manipulate  $d_{\text{GHP}}$  directly. Let us start by recalling how one can obtain a metric space from a càdlàg function, and introducing some notation. Note that this construction is simply a generalization of the construction of the Brownian CRT recalled at the beginning of this section, but with more general functions as contour processes.

**Definition 3.1.** Let  $h: [0, 1] \rightarrow \mathbb{R}$  be a nonnegative càdlàg function such that  $h(0) = 0$ . We denote by  $d_h$  the pseudometric on  $[0, 1]$  defined by

$$d_h(x, y) = h(x) + h(y) - 2 \inf_{z \in [x, y]} h(z),$$

where, as previously,  $[x, y]$  is shorthand for  $[x \wedge y, x \vee y]$ . We then denote by  $\mathcal{T}_h$  the rooted compact metric probability space obtained by: (1) identifying points  $x, y \in [0, 1]$  such that  $d_h(x, y) = 0$ ; (2) taking the completion of the space with respect to  $d_h$ ; (3) taking the equivalence class of 0 as the root; and (4) endowing the resulting rooted metric space with the pushforward of the Lebesgue measure on  $[0, 1]$ . This metric space is a subset of an  $\mathbb{R}$ -tree and consists of a countable number of connected components – see Figure 3 for an illustration, and e.g. [22] for an introduction to  $\mathbb{R}$ -trees.

<sup>1</sup> It is not obvious that  $\mathbb{M}$  can be defined as a set, because the class of compact metric spaces is not a set. However, since a compact metric space has cardinal at most  $\mathfrak{c} = \text{Card } \mathbb{R}$ , all isometry classes are obtained by considering subsets of  $\mathbb{R}$  endowed with a metric and a measure – and these do indeed form a well-defined set.

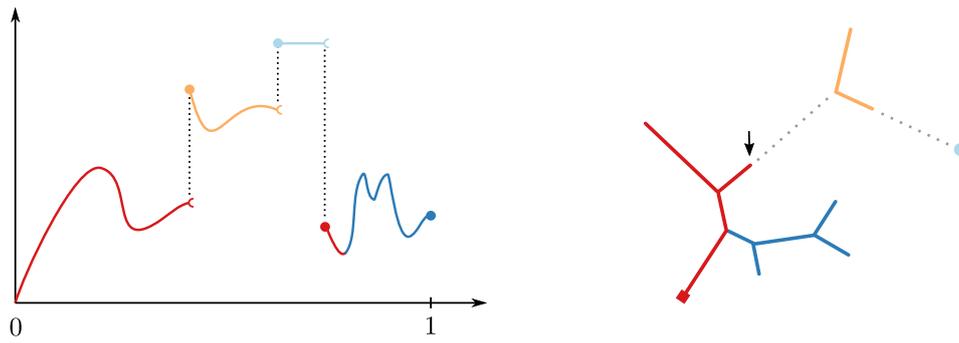


Figure 3: A nonnegative càdlàg function and the corresponding metric space, as given by Definition 3.1. The square indicates the root. The colors are irrelevant and are merely here to help show which part of the function corresponds to which part of the metric space. The dotted lines are not part of the metric space but are here to indicate how it can be embedded in an  $\mathbb{R}$ -tree. Note that we have added the tip of the red branch pointed at by the arrow in step (2) of the construction.

The interest of working with  $\mathbb{R}$ -trees and their height processes comes from the following lemma, which is a straightforward extension of [35, Lemma 2.4]. Let us denote by  $D$  the space of càdlàg functions from  $[0, 1]$  to  $\mathbb{R}$  that are also continuous at 1, endowed with the usual Skorokhod topology [10].

**Lemma 3.2.** *The map  $h \in D \mapsto \mathcal{T}_h \in \mathbb{M}$  is continuous. In other words, if  $h_1, h_2, \dots$  and  $h$  satisfy the hypotheses of Definition 3.1, then*

$$h_n \longrightarrow h \text{ in } D \implies \mathcal{T}_{h_n} \longrightarrow \mathcal{T}_h \text{ in } (\mathbb{M}, d_{\text{GHP}}).$$

A self-contained proof can be found in the appendices.

From Lemma 3.2, we get the upcoming Proposition 3.4, which provides a general recipe for proving convergence to the CRT in the rooted Gromov–Hausdorff–Prokhorov topology, and is going to be our main tool for the rest of this section.

**Definition 3.3.** *Let  $(\mathcal{X}, r, d, \lambda)$  be a random rooted compact metric probability space. A random càdlàg function  $\phi: [0, 1] \rightarrow \mathcal{X}$  is said to be a parametrization of  $\mathcal{X}$  if  $\phi(0) = r$  and  $\phi([0, 1])$  is almost surely dense in  $\mathcal{X}$ .*

*A parametrization  $\phi$  is said to be admissible if can write  $\phi(t) = \Phi(\mathcal{X}, \Theta, t)$ , where  $\Phi$  is a deterministic functional and  $\Theta$  is a random variable with values in  $[0, 1]$  that is independent from  $\mathcal{X}$ , in such a way that the functions  $t \mapsto \lambda(\phi([0, t]))$  and  $t \mapsto d(r, \phi(t))$  are well-defined random variables in the Skorokhod space  $D$ .*

We require our parametrizations to be admissible for measurability issues – namely, we need this assumption in order to use a variant of Skorokhod’s representation theorem in the proof of Proposition 3.4 below. In practice, admissibility should not be a restrictive requirement. In our case, we will define a parametrization of our network  $\mathcal{G}$  through a randomized traversal algorithm where, conditional on the network, the additional randomness that is needed amounts to a finite number of coin tosses; such a parametrization is readily checked to be admissible.

**Proposition 3.4.** *Let  $(\mathcal{X}_n, r_n, d_n, \lambda_n)_{n \geq 1}$  be a sequence of random rooted compact metric probability spaces such that, for each  $n \geq 1$ , there exists an admissible parametrization  $\phi_n: [0, 1] \rightarrow \mathcal{X}_n$ . Assume that, setting  $h_n(t) = d_n(r_n, \phi_n(t))$ :*

- (i)  $\sup_{s, t \in [0, 1]} |d_n(\phi_n(s), \phi_n(t)) - d_{h_n}(s, t)| \xrightarrow{d} 0$ .

(ii)  $\sup_{t \in [0,1]} |\lambda_n(\phi_n([0, t])) - t| \xrightarrow{d} 0.$

(iii)  $(h_n(t))_{t \in [0,1]} \xrightarrow{d} (h(t))_{t \in [0,1]}$  for the Skorokhod topology, where  $(h(t))_{t \in [0,1]}$  is a random càdlàg function.

Then,  $\mathcal{X}_n \xrightarrow{d} \mathcal{T}_h$  for the rooted Gromov–Hausdorff–Prokhorov topology.

Again, this proposition is proved in the appendices.

### 3.2 Conditioning on the number of colors

We now introduce some notation for conditioning  $\mathcal{G}$  on its number of colors. This notation will also be used in Section 4, where we study the local weak limit of  $\mathcal{G}$  conditioned to have  $n$  colors. First, recall that  $\mathcal{G}$  can be viewed as the decorated Galton–Watson tree  $\mathcal{T}^*$  obtained as follows:

1. Sample a Galton–Watson tree  $\mathcal{T}$  with offspring distribution  $M$ .
2. Conditional on  $\mathcal{T}$ , decorate each vertex  $k$  with the network  $\mathcal{X}_k$  associated to an independent realization of the process  $X$  conditioned on having  $M_k$  mutations (where  $M_k$  denotes the number of children of  $k$  in  $\mathcal{T}$ ).

Let  $A_n$  be the event  $\{\mathcal{G} \text{ has } n \text{ colors}\}$ , i.e.  $\{\mathcal{T} \text{ has } n \text{ vertices}\}$ . Since  $A_n$  is a deterministic function of  $\mathcal{T}$  and that the networks  $(\mathcal{X}_k)$  depend on  $\mathcal{T}$  only,  $\mathcal{G}_n \sim (\mathcal{G} | A_n)$  can be obtained by replacing  $\mathcal{T}$  with  $\mathcal{T}_n \sim (\mathcal{T} | A_n)$  in step 2 of the construction above – i.e. by decorating a Galton–Watson tree with offspring distribution  $M$  conditioned to have  $n$  vertices.

Note that we have not assumed that  $\mathbb{E}(M) = 1$ . Thus,  $\mathcal{T}$  is not necessarily critical. However, we know from Corollary 2.7 that there exists  $\zeta > 0$  such that

$$\frac{\mathbb{E}(M\zeta^M)}{\mathbb{E}(\zeta^M)} = 1.$$

Thus, letting  $\hat{M}$  be a  $\zeta$ -tilt of  $M$ , i.e. a random variable whose distribution is characterized by

$$\mathbb{E}(f(\hat{M})) = \frac{\mathbb{E}(f(M)\zeta^M)}{\mathbb{E}(\zeta^M)} \quad \text{for all bounded } f: \mathbb{N} \rightarrow \mathbb{R},$$

by considering a Galton–Watson tree with offspring distribution  $\hat{M}$  we get a critical Galton–Watson tree  $\hat{\mathcal{T}}$ . It is classic – and straightforward to check by writing down the probability distributions explicitly – that  $\mathcal{T}_n$  has the same distribution as  $\hat{\mathcal{T}}$  conditioned to have  $n$  vertices.

Since we will be conditioning on  $A_n$  and that  $(\mathcal{T} | A_n) \sim (\hat{\mathcal{T}} | \hat{A}_n)$ , it may not be clear at this point what the interest of working with  $\hat{\mathcal{T}}$  instead of  $\mathcal{T}$  is; this will become apparent later – see e.g. Remark 3.11 – but for now let us simply note that for any nonnegative function  $f$  we have  $\mathbb{E}(f(\hat{\mathcal{T}}) | \hat{A}_n) \leq \mathbb{E}(f(\hat{\mathcal{T}})) / \mathbb{P}(\hat{A}_n)$  and that, as the following classic proposition shows, it is straightforward to get an asymptotic equivalent of  $\mathbb{P}(\hat{A}_n)$  as  $n$  goes to infinity.

**Proposition 3.5.** *Let  $\hat{\mathcal{T}}$  be a critical Galton–Watson tree whose offspring distribution has a finite variance  $\sigma^2 > 0$  and is not supported on  $k\mathbb{N}$ , for any  $k \geq 2$ . Let  $\hat{A}_n$  denote the event  $\{\hat{\mathcal{T}} \text{ has } n \text{ vertices}\}$ . Then,*

$$\mathbb{P}(\hat{A}_n) \underset{n \rightarrow \infty}{\sim} \frac{1}{\sqrt{2\pi\sigma^2}} n^{-3/2}.$$

This result is well-known (see e.g. [33, Lemma 1.11] for a more general statement), but since it is central to our study we recall a short proof in the appendices.

### 3.3 Technical lemmas

To clarify the proof of the convergence to the CRT given in Section 3.4, we gather some of the more technical details here. Lemma 3.6 is a result about the logistic branching process with mutation. Once we have recognized that we are working with a decorated Galton–Watson tree, this lemma is the key specificity of our model for the convergence to the CRT. Lemma 3.10 is a streamlined, model-agnostic synthesis of the approach developed in [48, Section 4]. It provides generic concentration inequalities for sums of random variables associated to the vertices/edges of a size-conditioned Galton–Watson tree.

Before stating our first lemma, recall the notation for the quantities associated to a generic color:

- $X = (X_t)_{t \geq 0}$  denotes the trajectory of the number of individuals of that color, starting from a single individual at time  $t = 0$ .
- $T = \inf\{t \geq 0 : X_t = 0\}$  denotes the time of extinction of the color.
- $L = \int_0^\infty X(t) dt$  denotes the total length of the corresponding subnetwork.
- $M$  denotes the number of offspring of the color, i.e. the number of new colors that it produces by mutation.

Finally, recall that  $\zeta > 0$  is the unique real number – whose existence is guaranteed by Corollary 2.7 – such that  $\mathbb{E}(M\zeta^M) = \mathbb{E}(\zeta^M)$ .

**Lemma 3.6.** *There exists  $\eta > 0$  such that*

$$\mathbb{E}(M(\zeta \vee 1)^M e^{\eta L}) < +\infty.$$

**Remark 3.7.** Since  $T \leq L$ , we also have  $\mathbb{E}(M(\zeta \vee 1)^M e^{\eta T}) < +\infty$ .

*Proof.* Let us fix  $s > \zeta \vee 1$  such that  $\mathbb{E}(s^M) < \infty$  – such an  $s$  exists by Theorem 2.5. Note that to prove the lemma it is sufficient to show that there exists  $\eta > 0$  such that  $\mathbb{E}(s^M e^{\eta L}) < \infty$ ; and that since  $0 \in A_s := \{\eta \in \mathbb{R} : \mathbb{E}(s^M e^{\eta L}) < \infty\}$ , it in fact suffices to show that  $A_s$  is an open subset of  $\mathbb{R}$ .

Recall from Proposition 2.1 that for any numbers  $\mu$  and  $s$  and any nonnegative measurable function  $f$ ,

$$\mathbb{E}_\mu(f(X) s^M) = \mathbb{E}_{s\mu}(f(X) e^{(s-1)\mu L}). \tag{3.1}$$

Now, on the one hand by applying (3.1) to  $f(X) = e^{\eta L}$  we get that for any  $\eta$ ,

$$\mathbb{E}_\mu(s^M e^{\eta L}) = \mathbb{E}_{s\mu}(e^{((s-1)\mu + \eta)L}).$$

On the other hand, for  $\eta < \mu$ , by taking  $f(X) = 1$  and replacing  $(\mu, s)$  with  $(\mu - \eta, \frac{s\mu}{\mu - \eta})$  in (3.1) we get

$$\mathbb{E}_{\mu - \eta} \left[ \left( \frac{s\mu}{\mu - \eta} \right)^M \right] = \mathbb{E}_{s\mu}(e^{((s-1)\mu + \eta)L}).$$

Combining these two equalities, we see that for  $\eta < \mu$ ,

$$\mathbb{E}_\mu(s^M e^{\eta L}) = \mathbb{E}_{\mu - \eta} \left[ \left( \frac{s\mu}{\mu - \eta} \right)^M \right].$$

Now, from the explicit expression of the probability generating function of  $M$  given in Theorem 2.5, we see that if

$$\eta \mapsto \mathbb{E}_{\mu - \eta} \left[ \left( \frac{s\mu}{\mu - \eta} \right)^M \right]$$

is finite at  $\eta = 0$ , then it is also finite in a neighborhood of 0. This concludes the proof.  $\square$

**Remark 3.8.** The relation  $\mathbb{E}_\mu(s^M) = \mathbb{E}_{s\mu}(e^{(s-1)\mu L})$  gives an explicit expression for the distribution of  $M$  as a function of the family  $(\mathcal{L}_\mu(L))_\mu$  of distributions of  $L$ . In particular, if conditional on  $L$  we let  $Y$  be a Poisson variable with parameter  $\mu L$ , then  $\mathbb{E}_\mu(s^M) = \mathbb{E}_{s\mu}(s^Y)$ . However, because in  $\mathbb{E}_{s\mu}$  the distribution of  $Y$  depends on  $s$ , this does not give a simple construction of  $M$  as a random function of  $L$ .

We now give a simple Chernoff-type subpolynomial bound on the tail probabilities of a sum of independent random variables with finite exponential moments. The reasoning is classic – see e.g. [48], where it is used repeatedly – but we could not find a generic statement in the literature; so to streamline some of our proofs we state it as a lemma here.

**Lemma 3.9.** *Let  $Z_1, Z_2, \dots$  be i.i.d. copies of a random variable  $Z$  such that  $\mathbb{E}(Z) = 0$  and that there exists  $\eta > 0$  for which  $\mathbb{E}(e^{\eta|Z|}) < \infty$ . Then there exists  $C > 0$  such that, for all  $\varepsilon > 0$  and all  $n \geq 1$ ,*

$$\mathbb{P}\left(\left|\sum_{i=1}^n Z_i\right| \geq n^{1/2+\varepsilon}\right) \leq Ce^{-n^\varepsilon}.$$

*Proof.* We write  $u_n = n^{1/2+\varepsilon}$  to ease the notation. Let us start by focusing on positive deviations. For any  $\theta < \eta$ , by taking the exponential of the sum, applying Markov’s inequality and using the independence of the  $Z_i$ ’s, we get:

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^n Z_i \geq u_n\right) &= \mathbb{P}\left(\prod_{i=1}^n e^{\theta Z_i} \geq e^{\theta u_n}\right) \\ &\leq e^{-\theta u_n} \mathbb{E}\left(\prod_{i=1}^n e^{\theta Z_i}\right) \\ &\leq e^{-\theta u_n} (\mathbb{E}(e^{\theta Z}))^n. \end{aligned}$$

Now, since  $\mathbb{E}(Z) = 0$ , taking  $K > \mathbb{E}(Z^2)/2$  we have  $\mathbb{E}(e^{\theta Z}) \leq 1 + K\theta^2$  for all  $\theta$  small enough. Thus, with  $\theta_n = n^{-1/2}$  we get

$$\mathbb{P}\left(\sum_{i=1}^n Z_i \geq u_n\right) \leq \exp(K\theta_n^2 n - \theta_n u_n) = \exp(K - n^\varepsilon).$$

The negative deviations are treated similarly (or, more directly, by applying this bound to the variables  $-Z_1, -Z_2, \dots$ ), yielding

$$\mathbb{P}\left(\sum_{i=1}^n Z_i \leq -u_n\right) \leq \exp(K - n^\varepsilon).$$

Therefore,  $\mathbb{P}(|\sum_{i=1}^n Z_i| \geq n^{1/2+\varepsilon}) \leq 2e^{K-n^\varepsilon}$ , concluding the proof.  $\square$

Our next lemma is a general result about sums of decorations in critically tiltable size-conditioned Galton–Watson trees. In what follows, by *vertex decorations* of a tree  $\mathcal{T}$  we mean a family  $(F_k)_{k \geq 1}$  of real-valued random variables such that, letting  $k = 1, 2, \dots$  be the vertices of  $\mathcal{T}$ , in arbitrary order, and denoting by  $M_1, M_2, \dots$  their outdegrees, there exists i.i.d. random variables  $\Theta_1, \Theta_2, \dots$  that are independent of  $\mathcal{T}$  and such that  $F_k = F(M_k, \Theta_k)$  for some deterministic function  $F$ .

Similarly, letting  $E(\mathcal{T})$  denote the edge set of  $\mathcal{T}$  and writing  $e = k \rightarrow \ell$  for the edge from vertex  $k$  to vertex  $\ell$ , oriented away from the root, we say that  $(G_e)_{e \in E(\mathcal{T})}$  are *edge decorations* of  $\mathcal{T}$  if, for any vertex  $k$ , we have  $(G_{k \rightarrow \ell}, \ell \text{ child of } k) = G(M_k, \Theta_k)$  for some

deterministic function  $G$  such that, for all  $m \geq 1$ ,  $G(m, \Theta_k)$  is an exchangeable real vector of length  $m$ . In particular, note that for each edge  $e$ ,  $G_e$  is a real-valued random variable; and that for a given  $k$  the family  $(G_{k \rightarrow \ell}, \ell \text{ child of } k)$  is not assumed to be independent. Note also that because of exchangeability, for any edge  $e = k \rightarrow \ell$  the law of  $G_e$  depends only on  $M_k$ , and so with a slight abuse of notation we write  $(M_k, G_{k \rightarrow \ell})$  for a typical pair – for instance the pair of variables corresponding the edge from the root to a uniformly chosen child of the root.

**Lemma 3.10.** *Let  $\mathcal{T}$  be a Galton–Watson tree whose offspring distribution  $M$  is such that there exists  $\zeta > 0$  for which  $\mathbb{E}(M\zeta^M) = \mathbb{E}(\zeta^M) < \infty$ , and let  $k = 1, 2, \dots$  denote its vertices, in arbitrary order. Let  $(F_k)_{k \geq 1}$  be vertex decorations of  $\mathcal{T}$  such that there exists  $\eta > 0$  for which  $\mathbb{E}(e^{\eta|F_k|}\zeta^{M_k}) < \infty$ . Let  $A_n$  denote the event  $\{\mathcal{T} \text{ has } n \text{ vertices}\}$ . Then, along any sequence of integers  $n$  such that  $\mathbb{P}(A_n) > 0$ :*

(i) *There exists  $C > 0$  such that*

$$\mathbb{P}(\max\{F_1, \dots, F_n\} \geq C \log(n) \mid A_n) \xrightarrow[n \rightarrow \infty]{} 0.$$

(ii) *Letting  $v_1, v_2, \dots$  denote the vertices of  $\mathcal{T}$ , labeled in depth-first order: conditional on  $A_n$ , for all  $\varepsilon > 0$ ,*

$$\Delta := \max_{1 \leq k \leq n} \left| \sum_{i=1}^k (F_{v_i} - \hat{m}) \right| = o_p(n^{1/2+\varepsilon}), \quad \text{where } \hat{m} = \frac{\mathbb{E}(F_k \zeta^{M_k})}{\mathbb{E}(\zeta^M)}.$$

*More precisely,  $\mathbb{P}(\Delta \geq n^{1/2+\varepsilon} \mid A_n) \leq Ce^{-cn^\varepsilon}$  for some constants  $C, c > 0$  that may depend on  $\varepsilon$ , and all  $n \in \mathbb{N}$ .*

*Let  $(G_e)_{e \in E(\mathcal{T})}$  be edge decorations of  $\mathcal{T}$  such that there exists  $\eta > 0$  for which we have  $\mathbb{E}(e^{\eta|G_{k \rightarrow \ell}|} M_k \zeta^{M_k}) < \infty$ . Then, letting  $\Gamma(v)$  denote the path from the root of  $\mathcal{T}$  to its vertex  $v$ :*

(iii) *Conditional on  $A_n$ , for all  $\varepsilon > 0$ ,*

$$\Delta^* := \max_{v \in \mathcal{T}} \left| \sum_{e \in \Gamma(v)} (G_e - m^*) \right| = o_p(n^{1/4+\varepsilon}), \quad \text{where } m^* = \frac{\mathbb{E}(G_{k \rightarrow \ell} M_k \zeta^{M_k})}{\mathbb{E}(M \zeta^M)}.$$

*More precisely,  $\mathbb{P}(\Delta^* \geq n^{1/4+\varepsilon} \mid A_n) \leq Ce^{-cn^a}$  for some constants  $C, c, a > 0$  that may depend on  $\varepsilon$ , and all  $n \in \mathbb{N}$ .*

*Proof.* First, note that the main difficulty comes from the fact that, under  $\mathbb{P}(\cdot \mid A_n)$ , the decorations are not independent.

Let  $\hat{\mathcal{T}}$  be a Galton–Watson tree whose offspring distribution  $\hat{M}$  is the  $\zeta$ -tilt of  $M$ , and recall from Section 3.2 that  $\hat{\mathcal{T}}$  is critical and satisfies  $(\mathcal{T} \mid A_n) \sim (\hat{\mathcal{T}} \mid \hat{A}_n)$ , where  $\hat{A}_n = \{\hat{\mathcal{T}} \text{ has } n \text{ vertices}\}$ . Choose  $\hat{\mathcal{T}}$  to be independent of  $\Theta_1, \Theta_2, \dots$ , and let  $\hat{M}_1, \hat{M}_2, \dots$  be the number of children of its vertices. Finally, let  $\hat{F}_k = F(\hat{M}_k, \Theta_k)$ . Then,

$$\begin{aligned} \mathbb{P}(\max\{F_1, \dots, F_n\} \geq C \log(n) \mid A_n) &= \mathbb{P}(\max\{\hat{F}_1, \dots, \hat{F}_n\} \geq C \log(n) \mid \hat{A}_n) \\ &\leq \sum_{k=1}^n \mathbb{P}(\hat{F}_k \geq C \log(n) \mid \hat{A}_n) \\ &\leq \frac{n \mathbb{P}(\hat{F} \geq C \log(n))}{\mathbb{P}(\hat{A}_n)}, \end{aligned} \tag{3.2}$$

where we write  $\hat{F}$  – instead of, say,  $\hat{F}_1$  – for brevity. By assumption, there exists  $\eta > 0$  such that

$$\mathbb{E}\left(e^{\eta|\hat{F}|}\right) = \frac{\mathbb{E}\left(e^{\eta|F|}\zeta^M\right)}{\mathbb{E}\left(\zeta^M\right)} < \infty.$$

Therefore, for  $C > \frac{5}{2}\eta^{-1}$  Markov’s inequality yields  $\mathbb{P}(\hat{F} \geq C \log(n)) = o(n^{-5/2})$ . Moreover, by Proposition 3.5 we know that  $\mathbb{P}(\hat{A}_n) = \Theta(n^{-3/2})$ . Plugging these two estimates in (3.2) proves point (i).

Point (ii) is proved similarly: we fix  $\varepsilon > 0$  and use a union bound to get

$$\begin{aligned} \mathbb{P}\left(\Delta \geq n^{1/2+\varepsilon} \mid A_n\right) &= \mathbb{P}\left(\hat{\Delta} \geq n^{1/2+\varepsilon} \mid \hat{A}_n\right) \\ &\leq \frac{n \mathbb{P}\left(\left|\sum_{i=1}^n (\hat{F}_{v_i} - \hat{m})\right| \geq n^{1/2+\varepsilon}\right)}{\mathbb{P}(\hat{A}_n)}. \end{aligned}$$

Under the unconditional probability  $\mathbb{P}$ , the random variables  $\hat{F}_{v_i}$  are i.i.d. and their expected value is  $\mathbb{E}(\hat{F}) = \hat{m}$ . Therefore, by applying Lemma 3.9 we get

$$\mathbb{P}\left(\left|\sum_{i=1}^n (\hat{F}_{v_i} - \hat{m})\right| \geq n^{1/2+\varepsilon}\right) \leq C e^{-n^\varepsilon}$$

for some  $C > 0$ . Since  $n/\mathbb{P}(\hat{A}_n) = \Theta(n^{5/2})$ , this implies  $\mathbb{P}(\Delta \geq n^{1/2+\varepsilon} \mid A_n) = O(e^{-cn^\varepsilon})$  for some  $c > 0$ .

The proof of point (iii) requires a few extra ingredients. As for (i) and (ii), we start from  $\mathbb{P}(\Delta^* \geq u_n \mid A_n) = \mathbb{P}(\hat{\Delta}^* \geq u_n \mid \hat{A}_n)$ . Next, we recall that the maximum of the distance between a vertex and the root in a Galton–Watson tree  $\mathcal{T}_n$  conditioned to have  $n$  vertices is of order  $n^{1/2}$ . More specifically, if we denote by  $H(\mathbf{t})$  the maximal distance to the root in a tree  $\mathbf{t}$ , then  $n^{-1/2}H(\mathcal{T}_n)$  converges in distribution as  $n \rightarrow \infty$ , see e.g. [2]. Therefore, for every  $\varepsilon > 0$  there exists  $c > 0$  such that  $\mathbb{P}(H(\hat{\mathcal{T}}) > c\sqrt{n} \mid \hat{A}_n) < \varepsilon$  for all  $n$  large enough – which in turns entails

$$\mathbb{P}(\cdot \mid \hat{A}_n) \leq \mathbb{P}\left(\cdot \cap \{H(\hat{\mathcal{T}}) \leq c\sqrt{n}\} \mid \hat{A}_n\right) + \varepsilon.$$

As a result, we can pick an integer sequence  $(w_n)$  such that  $\sqrt{n} = o(w_n)$  and assume in what follows that, conditional on  $\hat{A}_n$ , we have  $H(\hat{\mathcal{T}}) \leq w_n$ .

Let us denote by  $\hat{\mathcal{T}}_h$  the set of vertices at distance  $h$  from the root in  $\hat{\mathcal{T}}$  and, to keep notation light, set  $S(v) := |\sum_{(k \rightarrow \ell) \in \Gamma(v)} (G_{k \rightarrow \ell} - m^*)|$ . For any sequence  $(u_n)$ ,

$$\begin{aligned} \mathbb{P}\left(\hat{\Delta}^* \geq u_n \mid \hat{A}_n\right) &= \mathbb{P}\left(\max_{1 \leq h \leq w_n} \left\{ \max_{v \in \hat{\mathcal{T}}_h} \{S(v)\} \right\} \geq u_n \mid \hat{A}_n\right) \\ &\leq \mathbb{E}\left(\sum_{h=1}^{w_n} \sum_{v \in \hat{\mathcal{T}}_h} \mathbb{1}_{\{S(v) \geq u_n\}} \mid \hat{A}_n\right) \\ &\leq w_n \max_{1 \leq h \leq w_n} \mathbb{E}\left(\sum_{v \in \hat{\mathcal{T}}_h} \mathbb{1}_{\{S(v) \geq u_n\}} \mid \hat{A}_n\right) \\ &\leq w_n \mathbb{P}(\hat{A}_n)^{-1} \max_{1 \leq h \leq w_n} \mathbb{E}\left(\sum_{v \in \hat{\mathcal{T}}_h} \mathbb{1}_{\{S(v) \geq u_n\}}\right). \end{aligned} \tag{3.3}$$

Now, let  $(\hat{\mathcal{T}}^{(h)}, v^*)$  be the random pointed tree obtained in the following way:

- Let  $v_1$  be the root, and start with a path  $v_1, v_2, \dots, v_{h+1} = v^*$  from  $v_1$  to  $v^*$ . This path will be referred to as the *spine* of the tree.

- For  $k = 1$  to  $h$ , add  $M_k^* - 1$  children to  $v_k$ , where  $M_k^*$  is an independent copy of the size-biasing of  $\hat{M}$ , i.e. a random variable whose distribution is  $\mathbb{P}(M_k^* = i) = i \zeta^i \mathbb{P}(M = i) / \mathbb{E}(M \zeta^M)$ .
- Let each of the vertices added at the previous step, as well as  $v^*$ , be the root of an independent Galton–Watson tree with offspring distribution  $\hat{M}$ .

It is classic (see e.g. [48, Section 4.2]) and readily checked that for any fixed tree  $\mathbf{t}$  and each vertex  $v \in \mathbf{t}|_h$ ,

$$\mathbb{P}\left[(\hat{\mathcal{F}}^{(h)}, v^*) = (\mathbf{t}, v)\right] = \mathbb{P}(\hat{\mathcal{F}} = \mathbf{t}).$$

As a result, for any function  $f$  on pointed trees,

$$\mathbb{E}\left(f(\hat{\mathcal{F}}^{(h)}, v^*)\right) = \sum_{\mathbf{t}} \sum_{v \in \mathbf{t}|_h} f(\mathbf{t}, v) \mathbb{P}\left[(\hat{\mathcal{F}}^{(h)}, v^*) = (\mathbf{t}, v)\right] = \mathbb{E}\left(\sum_{v \in \hat{\mathcal{F}}|_h} f(\hat{\mathcal{F}}, v)\right).$$

Applying this identity to  $f(\hat{\mathcal{F}}, v) = \mathbf{1}_{\{S(v) \geq u_n\}}$  in (3.3), we get

$$\mathbb{P}\left(\hat{\Delta} \geq u_n \mid \hat{A}_n\right) \leq w_n \mathbb{P}(\hat{A}_n)^{-1} \max_{1 \leq h \leq w_n} \mathbb{E}\left(f(\hat{\mathcal{F}}^{(h)}, v^*)\right). \tag{3.4}$$

By construction, on the spine of  $\hat{\mathcal{F}}^{(h)}$  the number of children of the vertices is distributed as the vector  $(M_1^*, \dots, M_h^*, \hat{M}_{h+1})$ , whose components are independent. As a result, letting  $G_k^*$  be the first component of the vector  $G(M_k^*, \Theta_k)$  and  $m^* = \mathbb{E}(G_{k \rightarrow \ell} M \zeta^M) / \mathbb{E}(M \zeta^M)$  its expected value, if  $u_n \rightarrow \infty$  as  $n \rightarrow \infty$ , then

$$\begin{aligned} \mathbb{E}\left(f(\hat{\mathcal{F}}^{(h)}, v^*)\right) &= \mathbb{P}\left(\left|\sum_{k=1}^{h-1} (G_k^* - m^*) + \hat{G}_h - m^*\right| \geq u_n\right) \\ &\leq \mathbb{P}\left(\left|\sum_{k=1}^{w_n-1} (G_k^* - m^*) + \hat{G}_{w_n} - m^*\right| \geq u_n\right) \text{ for } n \text{ large enough.} \end{aligned}$$

Moreover, we then also have, as  $n \rightarrow \infty$ ,

$$\mathbb{P}\left(\left|\sum_{k=1}^{w_n-1} (G_k^* - m^*) + \hat{G}_{w_n} - m^*\right| \geq u_n\right) \sim \mathbb{P}\left(\left|\sum_{k=1}^{w_n} (G_k^* - m^*)\right| \geq u_n\right).$$

Finally, taking  $u_n = n^{1/4+\varepsilon}$  for some  $\varepsilon > 0$  and  $w_n = \lfloor n^{1/2+\delta} \rfloor$  for some  $\delta > 0$  such that  $(1/2 + \delta)^2 < 1/4 + \varepsilon$ , Lemma 3.9 yields

$$\mathbb{P}\left(\left|\sum_{k=1}^{w_n} (G_k^* - m^*)\right| \geq n^{1/4+\varepsilon}\right) \leq C e^{-n^a}$$

for some positive constants  $a$  and  $C$ . Plugging this back in (3.4) and using that  $w_n \mathbb{P}(\hat{A}_n)^{-1} = \Theta(n^{2+\delta})$  concludes the proof.  $\square$

**Remark 3.11.** This proof illustrates the point of working with a critical Galton–Watson tree: for instance, even though we also have

$$\mathbb{P}\left(\max\{F_1, \dots, F_n\} \geq C \log(n) \mid A_n\right) \leq \frac{n \mathbb{P}(F \geq C \log(n))}{\mathbb{P}(A_n)},$$

because in the non-critical case  $\mathbb{P}(A_n)$  decays exponentially, the mere fact that  $F$  has finite exponential moments would not have been sufficient to get an adequate upper bound on the expression above.

### 3.4 Proof of the convergence to the CRT

We are now ready to prove the main theorem of this section. Recall that  $\zeta > 0$  is the unique real number such that  $\mathbb{E}(M\zeta^M) = \mathbb{E}(\zeta^M)$ .

**Theorem 3.12.** *Let  $(\mathcal{G}_n, r_n, d_{\mathcal{G}_n}, \lambda_{\mathcal{G}_n})$  denote the random rooted metric probability space  $(\mathcal{G}, r, d_{\mathcal{G}}, \lambda_{\mathcal{G}})$  conditioned to have  $n$  colors, and let  $\mathcal{C}$  be the Brownian CRT. Then, as  $n \rightarrow \infty$ ,*

$$(\mathcal{G}_n, r_n, \frac{C}{\sqrt{n}}d_{\mathcal{G}_n}, \frac{1}{|\mathcal{G}_n|}\lambda_{\mathcal{G}_n}) \xrightarrow{d} \mathcal{C}$$

for the rooted Gromov–Hausdorff–Prokhorov topology, with

$$C := \frac{\hat{\sigma}}{2\mathbb{E}(U^*)} = \frac{\sqrt{\mathbb{E}(\zeta^M)(\mathbb{E}(M^2\zeta^M) - \mathbb{E}(\zeta^M))}}{2\mathbb{E}(\sum_{t \in \mathcal{M}} t \zeta^M)},$$

where  $U^*$  is sampled uniformly at random among the mutation times of the biased process  $X^* \sim \mathcal{L}(X \dagger M\zeta^M)$ , and  $\hat{\sigma}^2$  is the variance of  $\hat{M} \sim \mathcal{L}(M \dagger \zeta^M)$ . Moreover,

$$|\mathcal{G}_n| := \lambda_{\mathcal{G}_n}(\mathcal{G}_n) = \frac{n\mathbb{E}(L\zeta^M)}{\mathbb{E}(\zeta^M)} + o_p(n^{1/2+\varepsilon}), \quad \forall \varepsilon > 0.$$

**Remark 3.13.** Using the probability measure  $\nu_{\circ}$  introduced in Proposition 2.4, the constant  $C$  can also be expressed in terms of  $T$ , the extinction time of the logistic branching process. Indeed, as explained after the proof of Theorem 3.12,

$$\mathbb{E}(U^*) = \frac{\zeta \mathbb{E}(M)}{\mathbb{E}(\zeta^M)} \sum_{k>0} \nu_{\circ}(k) \mathbb{E}_k(T \zeta^M) \mathbb{E}_{k-1}(\zeta^M).$$

**Remark 3.14.** Addario-Berry et al. [2] have shown that for critical Galton–Watson trees with finite-variance offspring distribution, the normalized height (and width) of the tree conditioned to have  $n$  vertices satisfy uniform sub-Gaussian tail bounds. More precisely, letting  $H(\mathcal{T}_n)$  denotes the height of the Galton–Watson tree  $\mathcal{T}_n$ , there exist  $K, k > 0$  such that for every  $n \in \mathbb{N}$  and every  $x \in \mathbb{R}_+$  we have

$$\mathbb{P}(H(\mathcal{T}_n) \geq h\sqrt{n}) \leq Ke^{-kh^2}. \tag{3.5}$$

Now consider, as in Lemma 3.10, any family of edge decorations  $(G_e)_{e \in E(\mathcal{T})}$  such that  $\mathbb{E}(e^{\eta|G_{k \rightarrow \ell}|} M_k \zeta^{M_k}) < \infty$  for some  $\eta > 0$ . Let us write

$$\tilde{\Delta}_n = \max_{v \in \mathcal{T}} \frac{1}{\sqrt{n}} \left| \sum_{e \in \Gamma(v)} G_e \right|.$$

With the notation of Lemma 3.10 (iii), we have  $\tilde{\Delta}_n \leq (\Delta^* + H(\mathcal{T})m^*)/\sqrt{n}$ . Thus, applying (3.5) and Lemma 3.10 (iii) with  $\varepsilon = 1/4$  yields

$$\begin{aligned} \mathbb{P}(\tilde{\Delta}_n \geq x \mid A_n) &\leq \mathbb{P}(\Delta^* \geq x\sqrt{n}/2 \mid A_n) + \mathbb{P}(H(\mathcal{T}_n) \geq x\sqrt{n}/2) \\ &\leq K(e^{-kx^a} + e^{-kx^2}) \end{aligned}$$

for some positive constants  $K, k$  and  $a$  that do not depend on  $n$ . This uniform tail bound implies that for any  $p \geq 1$ , the sequence  $(\tilde{\Delta}_n^p)_{n \geq 1}$  is tight. In our context, if we let  $G_{k \rightarrow \ell}$  be the distance between the root of  $\mathcal{T}_k$  and the mutation corresponding to the vertex  $\ell$ , then  $\tilde{\Delta}_n$  is the height of  $(\mathcal{G}_n, r_n, d_{\mathcal{G}_n}/\sqrt{n})$ . Therefore, we conclude that in addition to the convergence in distribution of Theorem 3.12, all moments of the height of the network, diameter and related quantities converge to the moments of the properly rescaled Brownian CRT.

*Proof of Theorem 3.12.* The proof is an application of Proposition 3.4. Set

$$d_n = C n^{-1/2} d_{\mathcal{G}_n}, \quad \lambda_n = |\mathcal{G}_n|^{-1} \lambda_{\mathcal{G}_n},$$

and let us define an admissible parametrization  $\phi_n: [0, 1] \rightarrow (\mathcal{G}_n, r_n, d_n, \lambda_n)$ . Recall that, in the forward-time process defining  $\mathcal{G}$ , a *branching point* is a point where a lineage splits and a *coalescence point* is a point where two lineages merge. Using some arbitrary procedure, distinguish one of the two outgoing lineages of each branching point of  $\mathcal{G}_n$  and one of the two incoming lineages of each coalescence point. Note that by (1) disconnecting the tip of each of the distinguished lineages that correspond to coalescence points from those coalescence points and (2) drawing distinguished lineages that correspond to branching points to the right of their undistinguished counterparts, we get a rooted plane R-tree  $\mathcal{G}_n^\#$  (not to be confused with  $\mathcal{T}_n$ , the combinatorial tree encoding the genealogy of the colors of  $\mathcal{G}_n$ ).

Now, pick a depth-first ordering of the vertices of  $\mathcal{T}_n$ , and visit the points of  $\mathcal{G}_n$  as follows:

- Visit the subnetworks corresponding to the vertices of  $\mathcal{T}_n$  in depth-first order.
- Within each subnetwork  $\mathcal{X}_k$ , do a depth-first traversal of the corresponding “unreticulated” R-tree  $\mathcal{X}_k^\#$ , that is: starting from the root, travel along the lineages at constant speed  $nL_k = n\lambda_{\mathcal{G}_n}(\mathcal{X}_k)$ , in depth-first order and visiting the “left” subtree first when encountering a branching point.

This construction is illustrated in Figure 4. Note that each jump of  $\phi_n$  corresponds to either the tip of a lineage or the second visit of a coalescence point, and that those jumps can be negative (typical case) or positive (which can only happen when finishing the exploration of a color and moving to a new one). Moreover, each point of  $\mathcal{G}_n$  is visited exactly once, except for:

- The tips of lineages, which – with the exception of  $\phi_n(1)$  – correspond to the left-limits of some of the jumps of  $\phi_n$ .
- Branching points, which are visited twice.

As a result,  $\phi_n([0, 1])$  is dense in  $\mathcal{G}_n$  and  $\phi_n$  is an admissible parametrization of  $\mathcal{G}_n$ .

Next, let us show that  $\phi_n$  satisfies assumptions (i–iii) of Proposition 3.4. Starting with (i), pick  $s, t \in [0, 1]$  with  $s < t$ , and let  $(x, y) = (\phi_n(s), \phi_n(t))$  be the corresponding points of  $\mathcal{G}_n$ . Let then  $x \wedge y$  be the most recent common ancestor of  $x$  and  $y$  in  $\mathcal{G}_n$ , i.e. the (unique) oldest point in a (non necessarily unique) shortest path between  $x$  and  $y$ , and let  $\mathcal{X}_c$  be the subnetwork containing  $x \wedge y$ . Let  $z_c$  be the root of  $\mathcal{X}_c$  and, for  $i \in \{x, y\}$ , let  $z_i$  be:  $z_c$  if  $i \in \mathcal{X}_c$ ; otherwise, the root of the subnetwork through which every path from  $x \wedge y$  to  $i$  exits  $\mathcal{X}_c$ . These definitions are illustrated in Figure 5.

With this notation, and recalling that  $h_n(t) = d_n(r_n, \phi_n(t))$ , observe that

$$\begin{aligned} & |d_n(x, y) - h_n(s) - h_n(t) + 2 \inf_{[s, t]} h_n| \\ &= |d_n(x, z_x) + d_n(z_x, z_y) + d_n(z_y, y) - d_n(r_n, x) - d_n(r_n, y) + 2 \inf_{[s, t]} h_n| \\ &\leq d_n(z_x, z_y) + |d_n(r_n, z_x) - \inf_{[s, t]} h_n| + |d_n(r_n, z_y) - \inf_{[s, t]} h_n|. \end{aligned} \tag{3.6}$$

Now,  $d_{\mathcal{G}_n}(z_x, z_y) < 2T_c$ , where  $T_c$  is the extinction time of the logistic process  $X_c$  associated to the color  $c$ . Therefore,

$$d_n(z_x, z_y) < 2Cn^{-1/2} T_c. \tag{3.7}$$

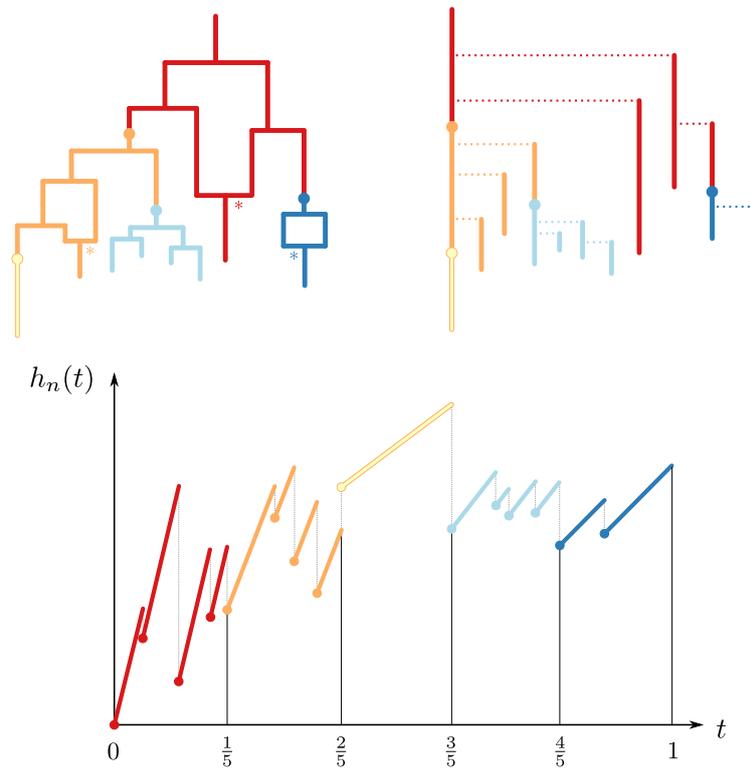


Figure 4: Illustration of the construction of the admissible parametrization  $\phi_n$  used in the proof. Left, a realization of  $\mathcal{G}_n$  for  $n = 5$ , with the same drawing conventions as in Figure 1. The distinguished lineages associated to coalescences are indicated by asterisks and, to avoid cluttering, the distinguished lineages associated to branchings are taken to be the lineages drawn to the right. Right, the rooted plane R-tree  $\mathcal{G}_n^\#$  obtained by “disconnecting” coalescence points of  $\mathcal{G}_n$ . This tree is to provide us with a natural order in which to visit the lineages of  $\mathcal{G}_n$ . Bottom, the height function  $h_n: t \mapsto d_n(r_n, \phi_n(t))$  associated to  $\phi_n$ . The speed of travel along the lineages of the subnetwork corresponding to a given color is proportional to the total length of that subnetwork, ensuring that each color is allotted the same amount of time by  $\phi_n$ .

Moreover, since by construction of  $\phi_n$  the vertices of  $\mathcal{T}_n$  are visited in depth-first order, for all  $u \in [s, t]$  the point  $\phi_n(u)$  belongs to either  $\mathcal{X}_c$  or one of its descendants, which implies that  $h_n(u) \geq d_n(r_n, z_c)$ ; and since there exists  $u \in [s, t]$  such that  $\phi_n(u)$  is at distance 0 from  $\mathcal{X}_c$  (indeed, if  $y \in \mathcal{X}_c$  one can take  $u = t$ , and if  $y \notin \mathcal{X}_c$  then one can take  $u$  such that  $\phi_n(u) = z_y$ ), which in turns implies  $h_n(u) \leq d_n(r_n, z_c) + Cn^{-1/2}T_c$ , we get

$$d_n(r_n, z_c) \leq \inf_{[s,t]} h_n \leq d_n(r_n, z_c) + Cn^{-1/2}T_c.$$

Similarly, for  $i \in \{x, y\}$  we have  $d_n(r_n, z_c) \leq d_n(r_n, z_i) \leq d_n(r_n, z_c) + Cn^{-1/2}T_c$ , so that

$$|d_n(r_n, z_i) - \inf_{[s,t]} h_n| \leq Cn^{-1/2}T_c, \quad i \in \{x, y\}. \tag{3.8}$$

Plugging (3.7) and (3.8) in (3.6), we get  $|d_n(x, y) - d_{h_n}(s, t)| < 4Cn^{-1/2}T_c$ . Therefore,

$$\sup_{s,t \in [0,1]} |d_n(\phi_n(s), \phi_n(t)) - d_{h_n}(s, t)| \leq 4Cn^{-1/2} \max(T_1, \dots, T_n).$$

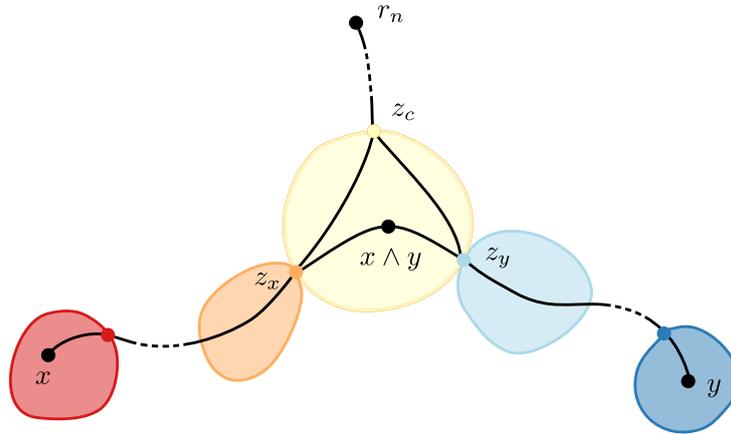


Figure 5: Graphical depiction of some of the notation used in the proof. The black lines correspond to shortest paths between various points of  $\mathcal{G}_n$ , and the colored blobs to the subnetworks associated to the colors. Although there can be several shortest paths between  $x$  and  $y$ , each of these paths goes through  $z_x$  and  $z_y$ , hence  $d_n(x, y) = d_n(x, z_x) + d_n(z_x, z_y) + d_n(z_y, y)$ . Similarly,  $d_n(r_n, x) = d_n(r_n, z_c) + d_n(z_c, z_x) + d_n(z_x, x)$ . Finally, note that within each subnetwork  $\mathcal{X}_k$  the distances are bounded above by  $2T_k$ , where  $T_k$  is the extinction time of the corresponding process  $X_k$ .

Applying point (i) of Lemma 3.10 to the extinction times  $T_1, \dots, T_n$ , which is made possible by the fact that we know from Lemma 3.6 that  $T$  has finite exponential moments under  $\mathcal{L}(\cdot \uparrow M\zeta^M)$ , we get  $\max(T_1, \dots, T_n) = o_p(n^\varepsilon)$  for any  $\varepsilon > 0$ , which in turns implies

$$\sup_{s, t \in [0, 1]} |d_n(\phi_n(s), \phi_n(t)) - d_{h_n}(s, t)| \xrightarrow{d} 0,$$

thereby proving that  $\phi_n$  satisfies assumption (i) of Proposition 3.4.

Let us now turn to assumption (ii). Let  $\mathcal{X}_1, \dots, \mathcal{X}_n$  be the subnetworks of  $\mathcal{G}_n$ , in order of their visit by  $\phi_n$ . By construction of  $\phi_n$ , for all  $t \in [0, 1]$  we have  $\phi_n(t) \in \mathcal{X}_{c_t}$ , where  $c_t := (\lfloor tn \rfloor + 1) \wedge n$ . Moreover,

$$\sum_{k=1}^{c_t-1} L_k \leq \lambda_{\mathcal{G}_n}(\phi_n([0, t])) \leq \sum_{k=1}^{c_t} L_k,$$

where  $L_k = \lambda_{\mathcal{G}_n}(\mathcal{X}_k)$ . Applying point (ii) of Lemma 3.10 to  $L_1, \dots, L_n$ , which again is made possible by Lemma 3.6, we get that for any  $\varepsilon > 0$ ,

$$\lambda_{\mathcal{G}_n}(\phi_n([0, t])) = tn\ell + o_p(n^{1/2+\varepsilon}),$$

uniformly in  $t \in [0, 1]$  and with  $\ell = \mathbb{E}(L\zeta^M)/\mathbb{E}(\zeta^M)$ . Taking  $t = 1$ , we see that  $|\mathcal{G}_n| = n\ell + o_p(n^{1/2+\varepsilon})$ , as claimed in the statement of the theorem. From there, we get

$$\sup_{t \in [0, 1]} |\lambda_n(\phi_n([0, t])) - t| = o_p(n^{-1/2+\varepsilon}),$$

which shows that  $\phi_n$  satisfies assumption (ii).

To show that  $\phi_n$  satisfies assumption (iii), let  $h_{\mathcal{T}_n}$  be the height process of  $\mathcal{T}_n$ , that is

$$h_{\mathcal{T}_n} : \begin{cases} \{1, \dots, n\} & \longrightarrow \mathbb{N} \\ k & \longmapsto d_{\mathcal{T}_n}(v_1, v_k), \end{cases}$$

where  $v_1, \dots, v_n$  are the vertices of  $\mathcal{T}_n$ , in order of their visit by  $\phi_n$  and  $d_{\mathcal{T}_n}(u, v)$  is the number of edges of the path joining  $u$  and  $v$ . Since  $\mathcal{T}_n \sim (\hat{\mathcal{T}} \mid \hat{A}_n)$  is a critical Galton–Watson tree conditioned to have  $n$  vertices, it is well-known – see e.g. Corollary 1 in [38], from which this readily follows – that, as  $n \rightarrow \infty$ ,

$$\left(\frac{1}{\sqrt{n}}h_{\mathcal{T}_n}(c_t)\right)_{t \in [0,1]} \xrightarrow{d} \left(\frac{2}{\hat{\sigma}}e(t)\right)_{t \in [0,1]} \text{ in the Skorokhod space } D,$$

where  $(e(t))_{t \in [0,1]}$  is a standard Brownian excursion and  $\hat{\sigma}^2 = \text{Var}(\hat{M})$  is the variance of the offspring distribution of  $\hat{\mathcal{T}}$ , and  $c_t = (\lfloor tn \rfloor + 1) \wedge n$ . Therefore, to conclude the proof of Theorem 3.12 it suffices to show that

$$\sup_{t \in [0,1]} \left| h_{\mathcal{G}_n}(t) - \mathbb{E}(U^*) h_{\mathcal{T}_n}(c_t) \right| = o_p(n^{1/2}), \tag{3.9}$$

where  $h_{\mathcal{G}_n}(t) = d_{\mathcal{G}_n}(r_n, \phi_n(t))$  and  $U^*$  is the time of a mutation sampled uniformly at random among the mutations of the process  $X^* \sim \mathcal{L}(X \dagger M\zeta^M)$ .

For this, for all  $k \in \{1, \dots, n\}$  let us denote by  $z_k$  the root of  $\mathcal{X}_k$ . As a result, letting  $\Gamma(k) = (i_1 \rightarrow \dots \rightarrow i_p)$  be such that  $(v_{i_1} = v_1, \dots, v_{i_p} = v_k)$  is the path from  $v_1$  to  $v_k$  in  $\mathcal{T}_n$ , and recalling that  $\phi_n(t) \in \mathcal{X}_{c_t}$ , we see that, for all  $t \in [0, 1]$ ,

$$h_{\mathcal{G}_n}(t) = \sum_{(i \rightarrow j) \in \Gamma(c_t)} d_{\mathcal{G}_n}(z_i, z_j) + d_{\mathcal{G}_n}(z_{c_t}, \phi_n(t)).$$

As we have already seen, for any  $\varepsilon > 0$ ,  $d_{\mathcal{G}_n}(z_{c_t}, \phi_n(t)) < T_{c_t} = o_p(n^\varepsilon)$ , uniformly in  $t$ . Since, by definition of  $h_{\mathcal{T}_n}$ , the number of edges of  $\Gamma(c_t)$  is  $h_{\mathcal{T}_n}(c_t)$ , we get that for any constant  $\kappa$ ,

$$h_{\mathcal{G}_n}(t) - \kappa h_{\mathcal{T}_n}(c_t) = \sum_{(i \rightarrow j) \in \Gamma(c_t)} (d_{\mathcal{G}_n}(z_i, z_j) - \kappa) + o_p(n^\varepsilon).$$

Moreover, along  $\Gamma(c_t)$  each  $d_{\mathcal{G}_n}(z_i, z_j)$  is the time elapsed between the creation of  $\mathcal{X}_i$  and that of  $\mathcal{X}_j$  – which, conditional on  $\mathcal{X}_i$ , is distributed as the random functional  $U_i = U(\mathcal{X}_i)$  giving the time of a mutation sampled uniformly at random among the mutations of  $\mathcal{X}_i$ . As a result, applying point (iii) of Lemma 3.10, we get

$$\sup_{t \in [0,1]} \left| \sum_{(i \rightarrow j) \in \Gamma(c_t)} (d_{\mathcal{G}_n}(z_i, z_j) - \mathbb{E}(U^*)) \right| = o_p(n^{1/2}),$$

where  $\mathbb{E}(U^*) = \mathbb{E}(UM\zeta^M)/\mathbb{E}(\zeta^M)$  does indeed correspond to the expected value of the time of a mutation sampled uniformly at random among the mutations of the biased process  $X^* \sim \mathcal{L}(X \dagger M\zeta^M)$ . Putting the pieces together, this proves (3.9), thereby concluding the proof of Theorem 3.12.  $\square$

Finally, before closing this section, let us justify the expression of  $\mathbb{E}(U^*)$  given in Remark 3.13. To make things simpler, we work with the “extended” process  $\bar{X}$ , which, in addition to the trajectory of  $X$ , contains the information of which jump corresponds to a mutation. Thus,  $M = M(\bar{X})$  is a deterministic function of  $\bar{X}$ . First, note that

$$\mathbb{E}(U^*) = \mathbb{E}(\zeta^M)^{-1} \mathbb{E}\left(\sum_{t \in \mathcal{M}} t \zeta^M\right),$$

and that, considering the shift operators  $(\Theta_t)_{t \in \mathbb{R}}$  defined by  $\Theta_t \bar{X} := (\bar{X}_{t+s})_{-t \leq s \leq T-t}$  and the function

$$F: \bar{X} \longmapsto \mathbb{E}(\zeta^M)^{-1} \zeta^{M(\bar{X})} \sup\{t : \bar{X}_{-t} > 0\},$$

we have

$$\mathbb{E}(\zeta^M)^{-1} \mathbb{E}\left(\sum_{t \in \mathcal{M}} t \zeta^M\right) = \mathbb{E}\left(\sum_{t \in \mathcal{M}} F(\Theta_t \bar{X})\right).$$

Recalling the definition of the process  $X^m$  introduced in Section 2.1, this is also

$$\mathbb{E}\left(\sum_{t \in \mathcal{M}} F(\Theta_t \bar{X})\right) = \mathbb{E}(M) \mathbb{E}(F(\bar{X}^m)).$$

Therefore, using the  $\bar{X}^m \stackrel{d}{=} \bar{X}' \wr \bar{X}''$  decomposition given in Proposition 2.4, together with the fact that

$$F(\bar{X}' \wr \bar{X}'') = \mathbb{E}(\zeta^M)^{-1} \zeta^{M(\bar{X}') + M(\bar{X}'') + 1} T(\bar{X}'),$$

where  $T(\bar{X}')$  denotes the extinction time of  $\bar{X}'$ , we get

$$\mathbb{E}(F(\bar{X}^m)) = \zeta \mathbb{E}(\zeta^M)^{-1} \sum_{k > 0} \nu_{\circ}(k) \mathbb{E}_k(T\zeta^M) \mathbb{E}_{k-1}(\zeta^M).$$

Putting the pieces together, this yields the expression given in Remark 3.13.

## 4 Local weak limit

In this section, we describe the structure of  $\mathcal{G}_n$  around a uniformly chosen point. More specifically, we give an algorithmic construction of the local weak limit of  $\mathcal{G}_n$  around a focal point distributed according to the normalized measure  $\lambda_{\mathcal{G}_n}/|\mathcal{G}_n|$ . The notion of local weak convergence after uniform rooting was introduced by Benjamini and Schramm in [7], and is therefore also known as *Benjamini–Schramm* convergence; see e.g. [47, Section 2.2] or [16, Section 1.2] for a general introduction. Throughout this section, unless specified otherwise the term *local weak limit* will always refer to the to the Benjamini–Schramm limit.

This section is organized as follows: first, we briefly lay out the topological notions that are used in our proof of the local convergence. We then describe the local weak limit  $(\mathcal{G}^\dagger, x^\dagger)$  as a decorated random tree. This random tree is a biased – that is, non-uniformly rooted – local weak limit of the size-conditioned Galton–Watson tree  $\mathcal{T}_n$  giving the genealogy of the colors of  $\mathcal{G}_n$  (see Section 3.2), and the decorations are modifications of the subnetwork  $\mathcal{X}$  corresponding to a generic color. We close the section by describing the geometry of these various modifications of  $\mathcal{X}$ .

### 4.1 Local topology

In order to define our local topology, we first need to specify a local space of decorated trees. A locally finite pointed rooted plane tree – henceforth simply referred to as a *pointed tree* for brevity – is a pair  $(\mathcal{T}, v^*)$  where  $\mathcal{T}$  is a rooted plane tree in which every vertex has a finite degree, and  $v^*$  is a vertex of  $\mathcal{T}$  known as the focal vertex. Note that in the case where  $\mathcal{T}$  is infinite, its root can be located at infinity: in that case, instead of corresponding to a vertex, the root corresponds to a topological end of  $\mathcal{T}$ . Another way to see this is that  $\mathcal{T}$  being rooted actually means for any pair of adjacent vertices  $(u, v)$  we know who is the parent and who is the child.

A *decorated pointed tree*  $(\mathcal{T}, v^*, (D_v)_{v \in \mathcal{T}})$  is a pointed tree where each vertex  $v \in \mathcal{T}$  is associated to a random variable  $D_v$  taking value in a Polish space  $\mathcal{D}$ . In our setting,  $\mathcal{D}$  will be a space in which the color networks  $(\mathcal{X}_v)$  used in the construction of  $\mathcal{G}$  as a decorated tree are well-defined Borel-measurable random variables; but for now let us view it simply as an abstract Polish space. We denote by  $\mathfrak{T}_{\text{loc}}$  the space of decorated pointed trees.

The *local topology* on  $\mathfrak{T}_{\text{loc}}$  is the topology generated by the following basis of open sets:

$$U(r, \mathbf{t}, (V_v)_{v \in \mathbf{t}}) = \{(\mathcal{T}, v^*, (D_v)_{v \in \mathcal{T}}) \in \mathfrak{T}_{\text{loc}} : B_{\mathcal{T}}(v^*, r) = \mathbf{t} \text{ and } \forall v \in \mathbf{t}, D_v \in V_v\},$$

where  $r$  runs over the positive integers,  $\mathbf{t}$  over the finite pointed trees, and  $(V_u)_{u \in \mathbf{t}}$  over the opens sets of  $\mathcal{D}$ . The notation  $B_{\mathcal{T}}(v^*, r)$  stands for the ball of radius  $r$  centered at  $v^*$  in  $\mathcal{T}$ .

To make our description of  $\mathfrak{T}_{\text{loc}}$  fully explicit, we would need to give a formal definition of the space  $\mathcal{D}$  of decorations. While this is relatively straightforward to do, this is not only tedious but also uninformative. Therefore, we leave it to the reader to convince themselves that this can be done while ensuring that the following properties hold:

- The decorations are pointed networks, i.e. pairs  $(\mathcal{X}, x^*)$  where  $x^* \in \mathcal{X}$  and, as previously, the network  $\mathcal{X}$  – which is meant to represent the subnetwork of  $\mathcal{G}$  that corresponds to a given color – can be seen as a collection of segments glued together at their endpoints (see Section 1.2). We denote by  $\lambda_{\mathcal{X}}$  the Lebesgue measure on  $\mathcal{X}$ .

To keep the notation light, the fact that the decorations are pointed will be considered implicit: we occasionally write  $\mathcal{X}$  instead of  $(\mathcal{X}, x^*)$  when the focal point  $x^*$  is irrelevant.

- The map  $\mathcal{X} \mapsto L_{\mathcal{X}} := \int d\lambda_{\mathcal{X}}$  is continuous.
- For all continuous bounded maps  $F: \mathcal{D} \rightarrow \mathbb{R}$ , the map

$$\mathcal{X} \mapsto \int F(\mathcal{X}, x) \lambda_{\mathcal{X}}(dx)$$

is continuous.

#### 4.2 Construction of the limit as a decorated tree

First, recall from Section 3.2 that if  $\hat{\mathcal{T}}$  is Galton–Watson tree whose offspring distribution  $\hat{M}$  is given by

$$\mathbb{P}(\hat{M} = k) = \frac{\zeta^k \mathbb{P}(M = k)}{\mathbb{E}(\zeta^M)}, \quad k \geq 0,$$

where  $\zeta$  is as in Corollary 2.7, then  $\hat{\mathcal{T}}$  conditioned to have  $n$  vertices has the same distribution as the tree  $\mathcal{T}_n$  used to construct  $\mathcal{G}_n$  as a decorated tree.

Next, let us describe  $(\mathcal{T}^*, v^*)$ , the local weak limit of  $\mathcal{T}_n$ . The local weak limit of size-conditioned critical Galton–Watson trees after random rooting is the invariant random sin-tree introduced by Aldous in [3] – see [46] for a detailed presentation. With our notation, this pointed tree  $(\mathcal{T}^*, v^*)$  can be constructed as follows:

- Let  $v^*$  be the focal vertex and let  $(v^*, v_1, v_2, \dots)$  be the *spine* of  $\mathcal{T}^*$ , i.e. an infinite path going towards the root (thus,  $v_1$  is the parent of  $v^*$ ,  $v_2$  is the parent of  $v_1$ , etc).
- For each  $k \geq 1$ , add  $M_k^* - 1$  children to  $v_k$ , where  $(M_k^*)_{k \geq 1}$  is an i.i.d. sequence with the size-biased distribution of  $\hat{M}$ .
- Let  $v^*$ , as well as each of the vertices added at the previous step be the root of a Galton–Watson tree with offspring distribution  $\hat{M}$ , and call  $\mathcal{T}^*$  the resulting infinite random tree.

Now, let us consider the infinite random network  $\mathcal{G}^*$  obtained by decorating  $\mathcal{T}^*$  using the same procedure as when decorating  $\mathcal{T}_n$  to obtain  $\mathcal{G}_n$ : conditional on  $\mathcal{T}^*$ , let us decorate each vertex  $v$ , independently of everything else, with a random network  $\mathcal{X}_v$  having the distribution of the generic color network  $\mathcal{X}$  conditioned to have a number of mutations equal to the number of children of  $v$ . Note that the subnetwork  $\mathcal{X}^*$  corresponding to the focal vertex  $v^*$  plays a special role in  $\mathcal{G}^*$ : we refer to that subnetwork as the *focal network*.

The pair  $(\mathcal{G}^*, \mathcal{X}^*)$  is not the local weak limit of  $\mathcal{G}_n$ : indeed, it describes the limit of neighborhoods of a *color network* picked uniformly at random in  $\mathcal{G}_n$ , rather than the limit of the neighborhoods of a *point* picked uniformly at random in  $\mathcal{G}_n$ . To see why the two differ, note in particular that picking the focal point  $x^*$  according to the normalized length measure  $\lambda_{\mathcal{G}_n}/|\mathcal{G}_n|$  biases the focal network by its total length  $L^* = \lambda_{\mathcal{X}^*}(\mathcal{X}^*)$ .

To construct the local weak limit of  $\mathcal{G}_n$ , conditional on  $(\mathcal{G}^*, \mathcal{X}^*)$  let  $x^* \sim \lambda_{\mathcal{X}^*}/L^*$  be a random point of  $\mathcal{X}^*$ . Note that – to fall back on the topological framework of Section 4.1 – the pointed network  $(\mathcal{G}^*, x^*)$  can be seen as an element of  $\mathfrak{T}_{\text{loc}}$  by identifying it with a copy of  $(\mathcal{T}^*, v^*)$  where the focal vertex  $v^*$  is decorated with the pointed network  $(\mathcal{X}^*, x^*)$ , and the decorations of the other vertices are arbitrarily pointed. Now recall from Notation 2.2 that  $\mathcal{L}(\cdot \dagger L^*)$  denotes the distribution under the  $L^*$ -biased probability measure  $\mathbb{E}(\mathbb{1}_{\{\cdot\}} L^*)/\mathbb{E}(L^*)$ , and let  $(\mathcal{G}^\dagger, x^\dagger)$  be the random pointed network rooted at infinity characterized by  $(\mathcal{G}^\dagger, x^\dagger) \sim \mathcal{L}((\mathcal{G}^*, x^*) \dagger L^*)$ .

**Theorem 4.1.** *The pointed network  $(\mathcal{G}^\dagger, x^\dagger)$  is the local weak limit of  $(\mathcal{G}_n)_{n \geq 1}$ .*

Before proving this theorem, let us point out that  $(\mathcal{G}^\dagger, x^\dagger)$  can also be constructed as follows; we leave it to the reader to convince themselves of the equivalence of the definitions:

- Let  $v^\dagger$  be the focal vertex, and let  $(v^\dagger, v_1, v_2, \dots)$  be an infinite spine going towards the root.
- For each  $k \geq 1$ , add  $M_k^* - 1$  children to  $v_k$ , where  $(M_k^*)_{k \geq 1}$  is an i.i.d. sequence with the size-biased distribution of  $\hat{M}$ .
- Let  $\mathcal{X}^\dagger \sim \mathcal{L}(\mathcal{X} \dagger L \zeta^M)$  where  $\mathcal{X}$  is a generic color network and  $M$  and  $L$  are respectively its number of mutations and total length. Write  $M^\dagger$  and  $L^\dagger$  for the corresponding quantities in  $\mathcal{X}^\dagger$ , and add  $M^\dagger$  children to  $v^\dagger$ .
- Let each children of  $v^\dagger$ , as well as each of the children that were added to the nodes  $(v_k)_{k \geq 1}$ , be the root of a Galton–Watson tree with offspring distribution  $\hat{M}$ . Let  $\mathcal{T}^\dagger$  be the resulting tree.
- Decorate each node  $v \in \mathcal{T}^\dagger$ ,  $v \neq v^*$ , with a network  $\mathcal{X}_v$  having the law of a generic color network  $\mathcal{X}$  conditioned to have as many mutations as the number of children of  $v$ . Let  $\mathcal{X}^\dagger$  be the decoration of  $v^\dagger$ .
- Conditional on  $\mathcal{X}^\dagger$ , let  $x^\dagger \sim \lambda_{\mathcal{X}^\dagger}/L^\dagger$  be the focal point of  $(\mathcal{G}^\dagger, x^\dagger)$ .

*Proof of Theorem 4.1.* Let  $x_n$  be a uniformly chosen point of  $\mathcal{G}_n$ , and let  $v_n^*$  be the vertex of  $\mathcal{T}_n$  such that  $x_n \in \mathcal{X}_{v_n^*}$ . Remember that, since here we view  $\mathcal{G}_n$  as the tree  $\mathcal{T}_n$  decorated with *pointed networks*, if we let  $x_n$  be the focal point of  $\mathcal{X}_{v_n^*}$  then  $(\mathcal{G}_n, x_n)$  and  $(\mathcal{T}_n, v_n^*)$  can be seen as the same object. Thus, we will use these two notations interchangeably.

To prove the theorem, it suffices to show that

$$\mathbb{E}(F(\mathcal{G}_n, x_n)) \xrightarrow{n \rightarrow \infty} \mathbb{E}(F(\mathcal{G}^\dagger, x^\dagger)) \tag{4.1}$$

for any function  $F: \mathfrak{T}_{\text{loc}} \rightarrow \mathbb{R}$  of the form

$$F(\mathcal{G}, x) = F(\mathcal{T}, v^*) = \mathbb{1}_{\{B_{\mathcal{D}}(v^*, r) = \mathbf{t}\}} \prod_{v \in \mathbf{t}} F_v(\mathcal{X}_v, x_v),$$

where  $r$  is a positive integer;  $\mathbf{t}$  is a finite pointed rooted plane tree; and  $(F_v)_{v \in \mathbf{t}}$  is a family of nonnegative continuous bounded maps  $\mathcal{D} \rightarrow \mathbb{R}$  such that for all  $v \neq v^*$ ,  $F_v(\mathcal{X}_v, x_v) =$

$\tilde{F}_v(\mathcal{X}_v)$  depends only on  $\mathcal{X}_v$ . Note that any such map  $F$  is continuous for the local topology and bounded, with  $\|F\|_\infty := \sup_{(\mathcal{G}, x)} |F(\mathcal{G}, x)| \leq \prod_{v \in \mathbf{t}} \sup_{(\mathcal{X}, x)} |F_v(\mathcal{X}, x)| < \infty$ . Moreover, since the map  $\mathcal{X} \mapsto L_{\mathcal{X}}$  giving the total length of a network is continuous, we can restrict ourselves to functions  $F$  for which there exists  $\ell > 0$  such that, for all  $v \in \mathbf{t}$ ,  $F_v(\mathcal{X}, x) = 0$  if  $L_{\mathcal{X}} > \ell$ .

Let us show that to finish the proof it suffices to show that

$$\mathbb{E} \left( \left| \left( \sum_{v \in \mathcal{T}_n} L_v \right)^{-1} - \frac{1}{n \mathbb{E}(\hat{L})} \right| \sum_{v \in \mathcal{T}_n} \int F(\mathcal{G}_n, x) \lambda_{\mathcal{X}_v}(dx) \right) \xrightarrow{n \rightarrow \infty} 0, \tag{4.2}$$

where  $\hat{L} \sim \mathcal{L}(L \dagger \zeta^M)$  is the total length of a generic decoration of the critical Galton-Watson tree  $\hat{\mathcal{T}}$  such that  $\mathcal{T}_n \sim (\hat{\mathcal{T}} \mid \hat{\mathcal{T}} \text{ has } n \text{ vertices})$ . By definition of  $x_n$ ,

$$\mathbb{E}(F(\mathcal{G}_n, x_n)) = \mathbb{E} \left( \left( \sum_{v \in \mathcal{T}_n} L_v \right)^{-1} \sum_{v \in \mathcal{T}_n} \int F(\mathcal{G}_n, x) \lambda_{\mathcal{X}_v}(dx) \right).$$

Thus, if (4.2) holds we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{E}(F(\mathcal{G}_n, x_n)) &= \lim_{n \rightarrow \infty} \frac{1}{n \mathbb{E}(\hat{L})} \mathbb{E} \left( \sum_{v \in \mathcal{T}_n} \int F(\mathcal{G}_n, x) \lambda_{\mathcal{X}_v}(dx) \right) \\ &= \frac{1}{\mathbb{E}(\hat{L})} \lim_{n \rightarrow \infty} \mathbb{E} \left( \frac{1}{n} \sum_{v \in \mathcal{T}_n} G(\mathcal{T}_n, v) \right), \end{aligned}$$

where the map

$$\begin{aligned} G(\mathcal{T}_n, v) &:= \int F(\mathcal{G}_n, x) \lambda_{\mathcal{X}_v}(dx) \\ &= \mathbf{1}_{\{B_{\mathcal{T}_n}(v^*, r) = \mathbf{t}\}} \left( \prod_{v \neq v^*} \tilde{F}_v(\mathcal{X}_v) \right) \int F_{v^*}(\mathcal{X}_{v^*}, x) \lambda_{\mathcal{X}_{v^*}}(dx) \end{aligned}$$

is continuous (and bounded by  $\|F\|_\infty \ell$ ) for the local topology, because the maps  $\mathcal{X} \mapsto \int F_v(\mathcal{X}, x) \lambda_{\mathcal{X}}(dx)$  are continuous on  $\mathcal{D}$ . Since the pointed tree  $(\mathcal{T}^*, v^*)$  used in the construction of  $(\mathcal{G}^\dagger, x^\dagger)$  is the local weak limit of  $\mathcal{T}_n$ , we get

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{E} \left( \frac{1}{n} \sum_{v \in \mathcal{T}_n} G(\mathcal{T}_n, v) \right) &= \mathbb{E}(G(\mathcal{T}^*, v^*)) \\ &= \mathbb{E} \left( \int F(\mathcal{G}^*, x) \lambda_{\mathcal{X}_{v^*}}(dx) \right) \\ &= \mathbb{E}(\hat{L}) \mathbb{E}(F(\mathcal{G}^\dagger, x^\dagger)), \end{aligned}$$

where the last equality holds because, by definition,  $(\mathcal{G}^\dagger, x^\dagger) \sim \mathcal{L}((\mathcal{G}^*, x^*) \dagger L^*)$  where  $L^* = \int d\lambda_{\mathcal{X}_{v^*}} \sim \hat{L}$ . Putting the pieces together, this proves (4.1).

Let us now prove (4.2), i.e. show that  $\mathbb{E}(Y_n) \rightarrow 0$ , where

$$Y_n = \left| \left( \sum_{v \in \mathcal{T}_n} L_v \right)^{-1} - \frac{1}{n \mathbb{E}(\hat{L})} \right| \sum_{v \in \mathcal{T}_n} \int F(\mathcal{G}_n, x) \lambda_{\mathcal{X}_v}(dx).$$

For this, on the one hand, note that  $F$  is bounded and

$$\sum_{v \in \mathcal{T}_n} \int F(\mathcal{G}_n, x) \lambda_{\mathcal{X}_v}(dx) \leq \|F\|_\infty \sum_{v \in \mathcal{T}_n} L_v, \tag{4.3}$$

and that on the other hand, since we have assumed that if the total length of the subnetwork containing  $x$  is greater than  $\ell$  then  $F(\mathcal{G}, x) = 0$ , we also have

$$\sum_{v \in \mathcal{T}_n} \int F(\mathcal{G}_n, x) \lambda_{\mathcal{X}_v}(dx) \leq \|F\|_\infty n \ell.$$

As a result,  $Y_n \leq \|F\|_\infty (1 + \ell/\mathbb{E}(\hat{L}))$ . Thus, by dominated convergence, to prove that  $\mathbb{E}(Y_n) \rightarrow 0$  it suffices to show that  $Y_n \rightarrow 0$  in probability. Using again (4.3),

$$0 \leq Y_n \leq \left| n \mathbb{E}(\hat{L}) - \sum_{v \in \mathcal{T}_n} L_v \right| \frac{\|F\|_\infty}{n \mathbb{E}(\hat{L})}.$$

Finally, by Lemma 3.6 we can apply point (ii) of Lemma 3.10 to the random variables  $(L_v)_{v \in \mathcal{T}_n}$  to get that for any  $\varepsilon > 0$ ,

$$\left| n \mathbb{E}(\hat{L}) - \sum_{v \in \mathcal{T}_n} L_v \right| = o_p(n^{1/2+\varepsilon}),$$

concluding the proof. □

### 4.3 Geometry of the focal and spinal networks

In order to complete the picture of the local weak limit of  $(\mathcal{G}_n)_{n \geq 1}$ , let us zoom in on the decorations composing  $\mathcal{G}^\dagger$  and describe their distributions more finely than in the previous section. Specifically, we are interested in

- $(\mathcal{X}^\dagger, x^\dagger)$ , the focal network and its distinguished point;
- $(\mathcal{X}^\circ, x^\circ)$ , which we call a *spinal network*. This network is distributed as the color network that is the parent of the focal network, and its distinguished point is the mutation point that corresponds to the root of the focal network.

Recall that, by the construction of  $\mathcal{G}^\dagger$  given in Section 4.2, these objects satisfy, for any positive measurable functional  $F$  on pointed color networks:

$$\mathbb{E}(F(\mathcal{X}^\dagger, x^\dagger)) = \frac{1}{\mathbb{E}(L\zeta^M)} \mathbb{E}\left(\zeta^M \int F(\mathcal{X}, x) \lambda_{\mathcal{X}}(dx)\right) \tag{4.4}$$

$$\mathbb{E}(F(\mathcal{X}^\circ, x^\circ)) = \frac{1}{\mathbb{E}(M\zeta^M)} \mathbb{E}\left(\zeta^M \sum_{x \in \mathcal{M}} F(\mathcal{X}, x)\right) \tag{4.5}$$

where, by a slight abuse of notation,  $\mathcal{M}$  denotes the point process of mutations on the space  $\mathcal{X}$  (previously,  $\mathcal{M}$  denoted the point process on  $\mathbb{R}$  corresponding to the mutation times).

Our next result shows that focal and spinal networks can be constructed by “glueing” two half-networks that are independent conditional on their number of tips. Moreover, there is an explicit procedure to build these networks from their profile, i.e. from the process giving their number of lineages as a function of time. Let us start by introducing some notation.

Let  $I = [t_0, t_1]$ , with  $t_0 < 0 \leq t_1$ , and let  $\gamma = (\gamma_t)_{t \in I}$  be a càdlàg, positive except at time  $t_1$ , integer-valued trajectory consisting of a finite number of  $\pm 1$  jumps, starting at 1 and ending with a jump to 0. As usual, let  $\mathcal{X}$  denote a generic color network, and let  $X$  be the corresponding logistic branching process. With a slight abuse of notation, we will write  $\{X = \gamma\}$  for the event on which the trajectory of the Markov chain  $X$ , started from 1 at time  $t_0$ , is exactly  $\gamma$ . Note that for any  $t_0 \in \mathbb{R}$ , it makes sense to consider a random network  $\mathcal{X}$  started from a single individual at time  $t_0$ , for which the logistic branching process  $X$  started from 1 at time  $t_0$  is the “number of lineages” process. The distribution of  $\mathcal{X}$  does not depend on  $t_0$ .



**Proposition 4.3.** For each  $k \geq 0$ , let  $X'_k$  and  $X''_k$  be independent realizations of  $X$  that are started from  $k$  and also independent of everything else, and let  $K \sim \nu_\circ$ .

- (i)  $(\mathcal{X}^\dagger, x^\dagger)$  is distributed as  $\mathcal{L}(\mathcal{X}[X'_K \wr X''_K] \uparrow \zeta^{M^\dagger})$ ;
- (ii)  $(\mathcal{X}^\diamond, x^\diamond)$  is distributed as  $\mathcal{L}(\mathcal{X}^m[X'_K \wr X''_{K-1}] \uparrow \zeta^{M^\diamond})$ ;

where  $M^\dagger$  denotes the number of mutations of  $\mathcal{X}[X'_K \wr X''_K]$  and  $M^\diamond$  the number of mutations of  $\mathcal{X}^m[X'_K \wr X''_{K-1}]$ .

**Remark 4.4.** This construction of the focal/spinal networks makes it possible to get expressions for some characteristics of the local weak limit  $(\mathcal{G}^\dagger, x^\dagger)$ . For instance, if we let  $N$  be the number of lineages of the same color as  $x^\dagger$  that are alive at the same time as  $x^\dagger$ , then by (i) we have, for a normalizing constant  $C \geq 0$ ,

$$\mathbb{P}(N = k) = C \nu_\circ(k) \mathbb{E}_k(\zeta^M)^2,$$

where  $\mathbb{E}_k$  denotes the expectation conditional on  $\{X_0 = k\}$ . The limitation comes from the fact that if  $\zeta \neq 1$ , then the expressions  $\mathbb{E}_k(\zeta^M)$  are not explicit. However, they can be expressed as continuous fractions, which would it possible compute them numerically (see Theorem 2.5).

Similarly, if  $T$  denotes the time since the last mutation in the ancestry of  $x^\dagger$ , then for any bounded measurable function  $F: \mathbb{R} \rightarrow \mathbb{R}$ ,

$$\mathbb{E}(F(T)\mathbb{1}_{\{N=k\}}) = C \nu_\circ(k) \mathbb{E}_k(\zeta^M) \mathbb{E}_k(F(T_0) \zeta^M),$$

where here  $T_0$  denotes the hitting time of 0 for the process  $X$ .

*Proof of Proposition 4.3.* The proof is very similar to that of Proposition 2.4, and also relies on the path decomposition Markov chains presented in Section A.1.

Let  $Y$  be the birth-death chain on  $\mathbb{N}$  that goes from  $k$  to  $k + 1$  at rate 1 and from  $k$  to  $k - 1$  at rate  $\rho_k$ . Note that  $Y$  is distributed as the chain  $X$  slowed-down by a factor  $k$  when in state  $k$ , and resurrected at rate 1 when it hits 0. Thus,  $Y$  is positive recurrent – and therefore, reversible (as any positive recurrent birth-death chain). Moreover, it is straightforward to check that the stationary distribution of  $Y$  is the probability distribution  $\pi$  defined by

$$\pi(n) = C \prod_{k=1}^n \frac{1}{\rho_k}, \quad n \geq 0,$$

where  $C$  is a normalizing constant. Note that by definition of  $\nu_\circ$ , if  $K_0$  is a random variable with distribution  $\pi$ , then its conditional distribution given  $\{K_0 \geq 1\}$  is  $\nu_\circ$ .

Let  $Y$  be started from 1, and denote  $T_0$  the hitting time of 0. Conditional on  $T_0$ , let  $U$  be uniform on  $[0, T_0]$ , and set  $K := Y_U$ . Now, from the trajectory of  $(Y_t)_{t \in [0, T_0]}$ , construct a path with the same distribution as  $X$  by speeding up time by a factor  $k$  when in state  $k$ . Let  $V$  be the point corresponding to  $U$  in the new timescale. Note that this shows that  $T_0$ , the hitting time of 0 by  $Y$ , has the same distribution as  $L = \int_0^\infty X_t dt$ , since in this construction the two quantities are equal.

Now, consider the biased probability measure

$$\mathbb{P}_0(\cdot) := \frac{\mathbb{E}(\mathbb{1}_{\{\cdot\}} T_0)}{\mathbb{E}(T_0)}.$$

By Proposition A.1, under  $\mathbb{P}_0$  we have  $K \sim \nu_\circ$  and, conditional on  $K$ , letting  $\Theta_u^t\{Y\} := (Y_{u+s})_{-u \leq s \leq t-u}$ , we have

$$\Theta_U^{T_0}\{Y\} \stackrel{d}{=} Y' \wr Y''$$

where  $Y'$  and  $Y''$  are independent copies of  $Y$  started from  $K$ . In other words, for any measurable positive functional  $F$  of trajectories:

$$\mathbb{E}[T_0 F(\Theta_U^{T_0}\{Y\})] = \mathbb{E}(T_0) \mathbb{E}[F(Y'_K \wr Y''_K)].$$

Using our coupling of  $X$  and  $Y$ , and recalling that  $T_0 = L$ , this yields:

$$\mathbb{E}[L F(\Theta_V^{\tau_0}\{X\})] = \mathbb{E}(L) \mathbb{E}[F(X'_K \wr X''_K)]. \tag{4.6}$$

where  $\tau_0$  is the extinction time of  $X$ , and  $X'_K$  and  $X''_K$  are independent copies of  $X$  started from  $K$ , as they are defined in the statement of the proposition.

Let us now define a pointed network  $(\tilde{\mathcal{X}}, x^*)$  as follows: conditional on the trajectory of  $X$  constructed above, using Definition 4.2 let  $(\tilde{\mathcal{X}}, x^*) = \mathcal{X}[\Theta_V^{\tau_0}\{X\}]$ . Thus,  $\tilde{\mathcal{X}}$  is distributed as a standard color network whose root is located at time  $-V$ . Recall that by definition, conditional on  $\tilde{\mathcal{X}}$ , the focal point  $x^*$  is chosen uniformly at random among the points that correspond to the  $K$  lineages alive at time 0.

By construction,  $x^*$  is uniform on  $\tilde{\mathcal{X}}$  with respect to its length measure. Therefore, for any measurable positive functional  $F$  on pointed networks:

$$\mathbb{E}\left(\zeta^M \int F(\mathcal{X}, x) \lambda_{\mathcal{X}}(dx)\right) = \mathbb{E}\left(L \zeta^M \int F(\mathcal{X}, x) \frac{1}{L} \lambda_{\mathcal{X}}(dx)\right) = \mathbb{E}(L \zeta^M F(\tilde{\mathcal{X}}, x^*)),$$

where  $M$  is the number of mutations of  $\tilde{\mathcal{X}}$ . Moreover, by applying Equation (4.6) to the functional  $\gamma \mapsto \mathbb{E}(\zeta^{M(\mathcal{X}[\gamma])} F(\mathcal{X}[\gamma]))$ , we get

$$\mathbb{E}(L \zeta^M F(\tilde{\mathcal{X}}, x^*)) = \mathbb{E}(L) \mathbb{E}(\zeta^{M^\dagger} F(\mathcal{X}[X'_K \wr X''_k])),$$

where  $M^\dagger$  is the total number of mutations of  $\mathcal{X}[X'_K \wr X''_k]$ . Therefore,

$$\mathbb{E}(\zeta^{M^\dagger} F(\mathcal{X}[X'_K \wr X''_k])) = \frac{1}{\mathbb{E}(L)} \mathbb{E}\left(\zeta^M \int F(\mathcal{X}, x) \lambda_{\mathcal{X}}(dx)\right).$$

Finally, taking  $F \equiv 1$ , we get  $\mathbb{E}(\zeta^{M^\dagger}) = \mathbb{E}(L \zeta^M) / \mathbb{E}(L)$ , and so comparing the previous display with Equation (4.4) characterizing the law of  $(\mathcal{X}^\dagger, x^\dagger)$ , we see that

$$\mathbb{E}(\zeta^{M^\dagger})^{-1} \mathbb{E}(\zeta^{M^\dagger} F(\mathcal{X}[X'_K \wr X''_k])) = \mathbb{E}(F(\mathcal{X}^\dagger, x^\dagger)),$$

finishing the proof of point (i).

Point (ii) is proved similarly, but using Proposition A.2 instead of Proposition A.1 to view the network from a uniform mutation point instead of from a uniform point.  $\square$

## Appendices

### A.1 Path decompositions of Markov chains

In this appendix, we give a description of the trajectory of a Markov chain as seen from a random point in time. The ideas are standard, but we could not find the two propositions below in the literature. Their proofs are elementary, but somewhat tedious; so since they are very similar we present only the most involved of the two and leave the other one to the reader. This appendix also contains the proof of Proposition 2.4.

Let  $E$  be a countable set and let  $Y$  be a continuous-time Markov chain on  $E$  with transition rate matrix  $Q = (q_{ij})_{i,j \in E}$ , started from the initial state  $0 \in E$ . Let us write  $\tau$  for the first jump time of the chain and  $T_0$  for the return time to 0, i.e.  $T_0 = \inf\{t \geq \tau : Y_t = 0\}$ . Assume that  $Y$  is positive recurrent, with stationary distribution  $\pi$ , and define the

reversed chain  $Y'$  as the continuous-time Markov chain with transition rate matrix  $Q' = (q'_{ij})_{i,j \in E}$ , where

$$q'_{ij} = \frac{\pi_j}{\pi_i} q_{ji}.$$

We will consider time-shifted trajectories  $\gamma$  of the Markov chain  $Y$  killed upon reaching 0. For this purpose, let us formally define a convenient Skorokhod-like space of trajectories. Let  $E'$  denote  $E \cup \{\Delta\}$ , where  $\Delta \notin E$  is arbitrary, and let  $\Gamma$  denote the set of càdlàg functions  $\gamma : \mathbb{R} \rightarrow E'$  such that for all  $t$  with  $|t|$  large enough we have  $\gamma(t) = \Delta$ . With a slight abuse, for any  $a < b \in \mathbb{R}$  and  $\gamma : [a, b[ \rightarrow E$  we identify  $\gamma$  with  $t \mapsto \gamma(t)\mathbb{1}_{\{t \in [a, b]\}} + \Delta\mathbb{1}_{\{t \notin [a, b]\}}$ . The space  $\Gamma$  can be endowed with the metric  $d$  defined by:

$$d(\gamma, \gamma') = \begin{cases} 1 \wedge \int_{\mathbb{R}} \mathbb{1}_{\{\gamma(t) \neq \gamma'(t)\}} dt & \text{if } \gamma \text{ and } \gamma' \text{ have the same number of jumps,} \\ 1 & \text{otherwise.} \end{cases}$$

Note that time-shifts are then continuous in  $\Gamma$ .

Recall from Notation 2.2 that for two random variables  $A$  and  $B$  defined on the same probability space, we write  $\mathcal{L}(A \dagger B)$  for the distribution of  $A$  biased by  $B$ , that is, under the probability  $\mathbb{P}(\cdot \dagger B) = \mathbb{E}(\mathbb{1}_{\{\cdot\}} B) / \mathbb{E}(B)$ . Conditional on  $T_0$ , let  $U$  be a uniform random variable on  $[\tau, T_0]$ . We are interested in the random trajectory

$$Z \sim \mathcal{L}((Y_{U+t})_{\tau-U \leq t < T_0-U} \dagger T_0 - \tau),$$

seen as a random variable in  $\Gamma$ . This trajectory can be conveniently described by decomposing it into its left and right parts. For this, recall the “back-to-back pasting” operation introduced in Definition 2.3, which to two càdlàg functions  $f : [0, T_f[ \rightarrow E$  and  $g : [0, T_g[ \rightarrow E$  associates the function  $f \wr g : [-T_f, T_g[ \rightarrow E$  defined by

$$f \wr g : t \mapsto \begin{cases} \lim_{s \downarrow t} f(-s) & \text{if } t < 0 \\ g(t) & \text{if } t \geq 0. \end{cases}$$

**Proposition A.1.** *With the definitions above, we have*

$$Z \stackrel{d}{=} Y' \wr Y'',$$

where  $Y'$  is the reversed chain and  $Y''$  has the same transitions as  $Y$ . Both chains are started from  $Y'_0 = Y''_0 \sim \pi^*$ , where  $\pi_i^* = \frac{\pi_i}{1-\pi_0}$  for all  $i \in E \setminus \{0\}$ , and stopped upon reaching 0. Conditional on their common starting point, they are independent.

As discussed above, the proof is left to the reader.

Assume now that the transitions of the process  $Y$  are associated with weights: each transition  $i \rightarrow j$  has weight  $w_{ij} \geq 0$ . Define a random measure  $\mathcal{W}$  by

$$\mathcal{W} = \sum_{i \rightarrow j \text{ at } t} w_{ij} \delta_t,$$

where the sum is over all (finitely many)  $(i, j, t)$  such that  $Y$  jumps from  $i$  to  $j$  at time  $t$  along a trajectory started from 0 and stopped upon reaching 0. Define  $W = \int d\mathcal{W}$  as the total weight accumulated along the trajectory and assume that  $0 < \mathbb{E}(W) < \infty$ . We are now interested in the distribution of

$$Z^w \sim \mathcal{L}((Y_{U+t})_{\tau-U \leq t < T_0-U} \dagger W),$$

where the conditional distribution of  $U$  given  $\mathcal{W}$  is  $\frac{1}{W}\mathcal{W}$ . In other words, the distribution of  $Z^w$  is characterized by

$$\mathbb{E}(F(Z^w)) = \frac{1}{\mathbb{E}(W)} \mathbb{E} \left[ \sum_{i \rightarrow j \text{ at } u} w_{ij} F((Y_{u+t})_{\tau-u \leq t < T_0-u}) \right],$$

for any measurable bounded functional  $F$ .

**Proposition A.2.** *We have*

$$Z^w \stackrel{d}{=} Y' \wr Y'',$$

where  $Y'$  is the reversed chain and  $Y''$  has the same transitions as  $Y$ . Conditional on their starting points,  $Y'$  and  $Y''$  are independent. They are started from a pair of states  $(Y'_0, Y''_0)$  chosen according to the probability

$$\nu(i, j) := \frac{\pi_i w_{ij} q_{ij}}{\pi_0 \mathbb{E}(W)}, \quad i \neq j$$

and stopped upon reaching 0. If one of the chains is started from 0, its trajectory is reduced to a single point.

*Proof.* The proof is a series of elementary Markov chain calculations. We use the standard notation  $q_i = \sum_j q_{ij}$ .

Consider two starting states  $i \neq j$ , a trajectory  $f$  from  $i$  to 0 and a trajectory  $g$  from  $j$  to 0. Let  $\gamma_0, \dots, \gamma_{n_f+1}$  and  $\xi_0, \dots, \xi_{n_g+1}$  be the successive states visited by  $f$  and  $g$ , respectively, and let  $x_0, \dots, x_{n_f}$  and  $y_0, \dots, y_{n_g}$  be the corresponding holding times. Writing  $\mathbb{P}(Z^w \in dh)$  for the probability density of  $Z^w$  evaluated in a specific trajectory  $h$ , by definition of  $Z^w$ , we have

$$\begin{aligned} & \mathbb{E}(W) \mathbb{P}(Z^w \in d(f \wr g)) \\ &= w_{ij} \left( \prod_{k=0}^{n_f} q_{\gamma_{k+1} \gamma_k} \right) q_{ij} \left( \prod_{k=0}^{n_g} q_{\xi_k \xi_{k+1}} \right) \exp \left( - \sum_{k=0}^{n_f} q_{\gamma_k} x_k - \sum_{k=0}^{n_g} q_{\xi_k} y_k \right) dx dy, \end{aligned}$$

where  $dx = dx_0 \cdots dx_{n_f}$  and  $dy = dy_0 \cdots dy_{n_g}$ . Rearranging the terms, we get

$$\begin{aligned} & \mathbb{E}(W) \mathbb{P}(Z^w \in d(f \wr g)) \\ &= \frac{\pi_i}{\pi_0} w_{ij} q_{ij} \times \left( \prod_{k=0}^{n_f} \frac{\pi_{\gamma_{k+1}}}{\pi_{\gamma_k}} q_{\gamma_{k+1} \gamma_k} \right) \exp \left( - \sum_{k=0}^{n_f} q_{\gamma_k} x_k \right) dx \\ & \quad \times \left( \prod_{k=0}^{n_g} q_{\xi_k \xi_{k+1}} \right) \exp \left( - \sum_{k=0}^{n_g} q_{\xi_k} y_k \right) dy \\ &= \frac{\pi_i}{\pi_0} w_{ij} q_{ij} \mathbb{P}_i(Y' \in df) \mathbb{P}_j(Y'' \in dg), \end{aligned}$$

which concludes the proof. □

We close this appendix by proving Proposition 2.4 from the main text, whose statement we reproduce here for convenience. Recall that  $X^m$  denotes the process  $X$  “as seen from a uniform mutation time”, that is,

$$X^m \sim \mathcal{L}((X_{U+t})_{-U \leq t < T-U} \uparrow M),$$

where  $U$  is chosen uniformly at random among the atoms of the point process  $\mathcal{M}$  giving the times of the mutations associated to the trajectory of  $X$ .

**Proposition 2.4.** *Let  $\nu_\circ$  be the probability distribution on the positive integers defined by*

$$\nu_\circ(n) = C \prod_{k=1}^n \frac{1}{\rho_k},$$

with  $C$  the corresponding normalizing constant. Let  $K \sim \nu_\circ$  and, conditional on  $K$ , let  $X'$  and  $X''$  be two independent realizations of the logistic branching process  $X$  started from  $X'_0 = K$  and  $X''_0 = K - 1$ . Then,

$$X^m \stackrel{d}{=} X' \wr X''.$$

*Proof.* Let  $X^\circ$  be a Markov chain started from 0 with the same transition rates as  $X$ , except for an additional “rebirth” transition from 0 to 1 at an arbitrary positive rate. Thus,  $X^\circ$  is positive recurrent, and it is straightforward to check that its stationary distribution  $(\pi_i)_{i \geq 0}$  satisfies, for  $i \geq 1$ ,

$$\pi_i \propto i^{-1} \prod_{k=1}^i \frac{1}{\rho_k}. \tag{A.1}$$

Now, since the excursions of  $X^\circ$  away from 0 are distributed as the restriction of  $X$  to  $[0, T]$ , we have

$$X^m \sim \mathcal{L}((X_{U+t}^\circ)_{\tau-U \leq t < T^\circ-U} \dagger M^\circ),$$

where  $\tau$  is the first jump time of  $X^\circ$ ;  $T^\circ$  its time of first return to 0;  $M^\circ$  its number of mutations on  $[\tau, T^\circ]$ ; and  $U$  the time of a mutation chosen uniformly at random among the mutations on  $[\tau, T^\circ]$ . Moreover, each downward jump of  $X^\circ$  from  $i$  to  $i - 1$  corresponds to a mutation with probability  $\mu/\rho_i$ , independently. Therefore, if conditional on  $X^\circ$  we let  $\mathcal{W}$  be the measure defined by

$$\mathcal{W} = \sum_{i \searrow_x \text{ at } t} \mu \rho_i^{-1} \delta_t,$$

where the sum is over all  $(i, t)$  such that  $X^\circ$  goes from  $i$  to  $i - 1$  at time  $t \in [\tau, T^\circ]$ , then, conditional on  $X^\circ$ ,  $\mathcal{W}$  is the intensity measure of  $\mathcal{M}^\circ$ , the Poisson point process of mutations on  $[\tau, T^\circ]$ . Thus,  $\mathbb{E}(W) = \mathbb{E}(M)$ , where  $W = \int d\mathcal{W}$ , and for any bounded measurable functional  $f$ ,

$$\mathbb{E} \left( \sum_{t \in \mathcal{M}^\circ} f(X^\circ, t) \middle| X^\circ \right) = \int f(X^\circ, t) \mathcal{W}(dt) = \sum_{i \searrow_x \text{ at } t} \mu \rho_i^{-1} f(X^\circ, t),$$

As a result, conditional on  $X^\circ$ , letting  $V \sim \frac{1}{W} \mathcal{W}$  we have

$$\mathcal{L}((X_{U+t}^\circ)_{\tau-U \leq t < T^\circ-U} \dagger M^\circ) = \mathcal{L}((X_{V+t}^\circ)_{\tau-V \leq t < T^\circ-V} \dagger W).$$

Therefore, by applying Proposition A.2 to  $(X^\circ, \mathcal{W})$ , we get that  $X^m \stackrel{d}{=} X' \wr X''$ , where:

- $X'$  and  $X''$  are independent,  $X''$  has the same transition rates as  $X^\circ$ , and  $X'$  has the same transition rates as the time-reversed chain of  $X^\circ$ .
- $(X', X'')$  is started from  $(i, i - 1)$  with probability

$$\frac{\pi_i (\mu \rho_i^{-1}) (i \rho_i)}{\pi_0 \mathbb{E}(W)} = C \prod_{k=1}^i \frac{1}{\rho_k}.$$

where  $C$  is the normalization constant (and we have used the expression of  $\pi_i$  given in (A.1) to get the right-hand side).

Finally, since every positive recurrent birth-death chain is reversible – a standard fact that follows from Kolmogorov’s criterion for time-reversibility –  $X'$  in fact has the same transition rates as  $X^\circ$ ; and since  $X'$  and  $X''$  are both killed upon reaching 0, these two chain also have the same transitions rates as  $X$ .  $\square$

### A.2 Proofs for Section 3.1

In this appendix, we prove Lemma 3.2 and Proposition 3.4, whose statements we will reproduce below for convenience.

Let us start by recalling how the notions of correspondence and distortion can be used to tackle Gromov–Hausdorff–Prokhorov convergence more conveniently than by working directly with the definition of the metric. Note that, in order to deal with the Prokhorov component of the metric, we will use definitions that differ slightly from those traditionally used in the Gromov–Hausdorff setting.

Let  $(\mathcal{X}, r, d, \lambda)$  and  $(\mathcal{X}', r', d', \lambda')$  be two rooted compact metric probability spaces. Since we view a subset  $R \subset \mathcal{X} \times \mathcal{X}'$  as a binary relation, we write  $xRx'$  to indicate that  $(x, x') \in R$ . For any  $A \subset \mathcal{X}$ , we let  $AR = \{x' \in \mathcal{X}' : \exists x \in A \text{ with } xRx'\}$  and we define  $RB$  similarly for any subset  $B \subset \mathcal{X}'$ . In what follows, we use the term *correspondence from  $\mathcal{X}$  to  $\mathcal{X}'$*  to refer to any nonempty subset  $R \subset \mathcal{X} \times \mathcal{X}'$ . Note that it is sometimes required that  $R$  satisfies  $\mathcal{X}R = \mathcal{X}'$  and  $\mathcal{X} = R\mathcal{X}'$  to be called a correspondence, but in our setting it will be more convenient to drop this restriction.

We now introduce a modified version of the notion of distortion of a correspondence. In what follow,  $A^\varepsilon$  denotes the  $\varepsilon$ -neighborhood of a set  $A$ .

**Definition A.3.** *The Prokhorov distortion of a correspondence  $R$  from a compact metric probability space  $(\mathcal{X}, d, \lambda)$  to another  $(\mathcal{X}', d', \lambda')$ , which we denote by  $\underline{\text{dis}}(R)$ , is the infimum of the  $\varepsilon > 0$  such that:*

- (i) For all  $(x, x') \in R$  and  $(y, y') \in R$ ,  $|d(x, y) - d'(x', y')| \leq \varepsilon$ .
- (ii)  $(\mathcal{X}R)^{\varepsilon/2} = \mathcal{X}'$  and  $\mathcal{X} = (R\mathcal{X}')^{\varepsilon/2}$ .
- (iii) For any Borel set  $A \subset \mathcal{X}$ ,  $\lambda'((AR)^{\varepsilon/2}) + \varepsilon \geq \lambda(A)$ .

The usual notion of distortion only takes (i) into account: (ii) is added to be able to relax the usual definition of correspondence, as discussed above; and (iii) controls the Prokhorov part of the Gromov–Hausdorff–Prokhorov topology. It may seem that by replacing the  $\varepsilon/2$  with  $\varepsilon$  would yield a more natural definition; however, this  $\varepsilon/2$  makes several calculations neater. Finally, note that because we have only imposed one inequality in (iii), this definition is not symmetric: if we let  $R^{-1} = \{(x', x) : (x, x') \in R\}$  then a priori  $\underline{\text{dis}}(R) \neq \underline{\text{dis}}(R^{-1})$ .

As the next lemma shows, correspondences and their distortions provide a simple characterization of the (rooted) Gromov–Hausdorff–Prokhorov convergence.

**Lemma A.4.** *Let  $(\mathcal{X}, r, d, \lambda)$  and  $(\mathcal{X}', r', d', \lambda')$  be two rooted compact metric probability spaces. If there exists a correspondence  $R \subset \mathcal{X} \times \mathcal{X}'$  satisfying  $rRr'$  and  $\underline{\text{dis}}(R) \leq \varepsilon$ , then  $d_{\text{GHP}}(\mathcal{X}, \mathcal{X}') \leq \varepsilon$ .*

*Proof.* The proof is a straightforward adaptation of the classic analogous result for the usual notion of correspondence and the Gromov–Hausdorff metric, see for instance [22, Theorem 4.11]

Let us define a metric  $\delta$  on the disjoint union  $\mathcal{X} \sqcup \mathcal{X}'$  by:

- $\forall x, y \in \mathcal{X}, \delta(x, y) = d(x, y)$ ;
- $\forall x', y' \in \mathcal{X}', \delta(x', y') = d'(x, y)$ ;
- $\forall (x, x') \in \mathcal{X} \times \mathcal{X}', \delta(x, x') = \varepsilon/2 + \inf\{d(x, y) + d'(x', y') : (y, y') \in R\}$ .

It is readily checked that  $\delta$  is indeed a metric. Moreover, for any  $(x, x') \in R$ , we have  $\delta(x, x') = \varepsilon/2$ . This implies that  $\delta(r, r') \leq \varepsilon/2$ , and that for each Borel set  $A \subset \mathcal{X}$ ,  $(AR)^{\varepsilon/2} \subset A^\varepsilon$ . Therefore, it follows from point (ii) of Definition A.3 that the Hausdorff

distance between  $\mathcal{X}$  and  $\mathcal{X}'$  in  $(\mathcal{X} \sqcup \mathcal{X}', \delta)$  is at most  $\varepsilon$ . Similarly, it follows from point (iii) of Definition A.3 that the Prokhorov distance between the extensions of  $\lambda$  and  $\lambda'$  to  $\mathcal{X} \sqcup \mathcal{X}'$  is also at most  $\varepsilon$ .

Therefore,  $d_{\text{GHP}}(\mathcal{X}, \mathcal{X}') \leq \varepsilon$  and the proof is complete. □

We are now ready to prove Lemma 3.2. First, recall from Definition 3.1 how to obtain a rooted compact metric probability space  $\mathcal{T}_h$  from a nonnegative càdlàg function  $h$  such that  $h(0) = 0$ .

**Lemma 3.2.** *The map  $h \in D \mapsto \mathcal{T}_h \in \mathbb{M}$  is continuous. In other words, if  $h_1, h_2, \dots$  and  $h$  satisfy the hypotheses of Definition 3.1, then*

$$h_n \rightarrow h \text{ in } D \implies \mathcal{T}_{h_n} \rightarrow \mathcal{T}_h \text{ in } (\mathbb{M}, d_{\text{GHP}}).$$

*Proof.* It is classic [10] that the Skorokhod topology can be metrized by the following metric: for two càdlàg functions  $f$  and  $g : [0, 1] \rightarrow \mathbb{R}$ , define

$$d_{\text{Sk}}(f, g) = \inf_{\theta} ((\text{Lip}(\theta) - 1) \vee \|f - g \circ \theta\|_{\infty}),$$

where  $\theta$  runs over the set of continuous increasing bi-Lipschitz bijections from  $[0, 1]$  into itself, and

$$\text{Lip}(\theta) := \sup_{0 \leq x < y \leq 1} \left( \frac{\theta(y) - \theta(x)}{y - x} \vee \frac{y - x}{\theta(y) - \theta(x)} \right).$$

Note that, since  $\theta(0) = 0$  and  $\theta(1) = 1$  for every such bijection  $\theta$ , if  $\text{Lip}(\theta) < 1 + \varepsilon$  then  $\|\theta - \text{Id}\|_{\infty} < \varepsilon$ , where  $\text{Id}$  is the identity map.

Let  $f$  and  $g$  be two nonnegative càdlàg functions such that  $f(0) = g(0) = 0$ , and let  $(\mathcal{T}_f, r_f, d_f, \lambda_f)$  and  $(\mathcal{T}_g, r_g, d_g, \lambda_g)$  be the corresponding metric spaces. We will show that  $d_{\text{Sk}}(f, g) < \varepsilon \implies d_{\text{GHP}}(\mathcal{T}_f, \mathcal{T}_g) \leq 4\varepsilon$ . By definition of  $d_{\text{Sk}}$ , let us choose a continuous increasing bijection  $\theta : [0, 1] \rightarrow [0, 1]$  such that  $\text{Lip}(\theta) < 1 + \varepsilon$  and  $\|f - g \circ \theta\|_{\infty} < \varepsilon$ .

Let  $\phi : [0, 1] \rightarrow \mathcal{T}_f$  denote the quotient map from  $[0, 1]$  to  $\mathcal{T}_f$ . Note that, because we have completed the quotient space  $[0, 1] / \sim_{d_f}$  in order to obtain  $\mathcal{T}_f$ , this function  $\phi$  may be non-surjective, but that  $\phi([0, 1])$  is dense in  $\mathcal{T}_f$ . Define  $\psi : [0, 1] \rightarrow \mathcal{T}_g$  similarly.

Let  $R \subset \mathcal{T}_f \times \mathcal{T}_g$  be the correspondence defined by:

$$xR x' \iff \exists t \in [0, 1] \text{ s.t. } x = \phi(t) \text{ and } x' = \psi(\theta(t)).$$

By Lemma A.4, to show that  $d_{\text{GHP}}(\mathcal{T}_f, \mathcal{T}_g) \leq 4\varepsilon$ , it is sufficient to check that  $\text{dis}(R) \leq 4\varepsilon$  and that  $r_f R r_g$ . The latter point is trivial since  $r_f = \phi(0)$  and  $r_g = \psi(0) = \psi(\theta(0))$ , therefore we need to check that the following three points hold:

- (i)  $(\mathcal{T}_f R)^{2\varepsilon} = \mathcal{T}_g$  and  $\mathcal{T}_f = (R \mathcal{T}_g)^{2\varepsilon}$ ; this is also immediate since  $\mathcal{T}_f R = \psi([0, 1])$  is dense in  $\mathcal{T}_g$  and since  $R \mathcal{T}_g = \phi([0, 1])$  is dense in  $\mathcal{T}_f$ .
- (ii) For all  $(x, x'), (y, y') \in R$ , we have  $|d_f(x, y) - d_g(x', y')| \leq 4\varepsilon$ .
- (iii) For any Borel subset  $A \subset \mathcal{T}_f$ , we have  $\lambda_g((AR)^{2\varepsilon}) + 4\varepsilon \geq \lambda_f(A)$ .

To prove (ii), consider  $s < t \in [0, 1]$ . We need to show that

$$|d_f(\phi(s), \phi(t)) - d_f(\psi \circ \theta(s), \psi \circ \theta(t))| \leq 4\varepsilon.$$

This is readily seen, since

$$\begin{aligned} & |d_f(\phi(s), \phi(t)) - d_f(\psi \circ \theta(s), \psi \circ \theta(t))| \\ &= |f(s) + f(t) - 2 \inf_{[s,t]} f - (g(\theta(s)) + g(\theta(t)) - 2 \inf_{[s,t]} g \circ \theta)| \\ &\leq 4\|f - g \circ \theta\|_{\infty} \end{aligned}$$

To show (iii), consider a Borel subset  $A \subset \mathcal{T}_f$ , and let  $\ell$  denote the Lebesgue measure on  $[0, 1]$ , so that  $\lambda_f(A) = \ell(\phi^{-1}(A))$  and  $\lambda_g(A) = \ell(\psi^{-1}(A))$ . Notice that, by definition,  $AR = \psi \circ \theta(\phi^{-1}(A))$ . Therefore,

$$\lambda_g(AR) = \ell\left(\psi^{-1}(\psi \circ \theta(\phi^{-1}(A)))\right) \geq \ell(\theta(\phi^{-1}(A))).$$

Now, it suffices to notice that  $\text{Lip}(\theta) < 1 + \varepsilon$  implies that  $\ell(\theta(B)) \geq (1 + \varepsilon)\ell(B)$  for all Borel subsets  $B \subset [0, 1]$ . Indeed, this is easily checked for intervals, and extended to Borel sets by a monotone class argument. From this, we obtain

$$\lambda_g(AR) \geq \ell(\theta(\phi^{-1}(A))) \geq (1 - \varepsilon)\ell(\phi^{-1}(A)) = (1 - \varepsilon)\lambda_f(A) \geq \lambda_f(A) - \varepsilon.$$

Finally,  $\lambda_g((AR)^{2\varepsilon}) + 4\varepsilon \geq \lambda_g(AR) + \varepsilon \geq \lambda_f(A)$ . Therefore,  $d_{\text{GHP}}(\mathcal{T}_f, \mathcal{T}_g) \leq 4\varepsilon$  and we have proved that  $d_{\text{Sk}}(f, g) < \varepsilon \implies d_{\text{GHP}}(\mathcal{T}_f, \mathcal{T}_g) \leq 4\varepsilon$ , finishing the proof.  $\square$

Let us now turn to Proposition 3.4. Recall that an admissible parametrization of  $(\mathcal{X}, r, d, \lambda)$  is a càdlàg function  $\phi: [0, 1] \rightarrow \mathcal{X}$  such that  $\phi([0, 1])$  is dense in  $\mathcal{X}$  and that  $t \mapsto \lambda(\phi([0, t]))$  and  $t \mapsto d(r, \phi(t))$  are well-defined random variables.

**Proposition 3.4.** *Let  $(\mathcal{X}_n, r_n, d_n, \lambda_n)_{n \geq 1}$  be a sequence of random rooted compact metric probability spaces such that, for each  $n \geq 1$ , there exists an admissible parametrization  $\phi_n: [0, 1] \rightarrow \mathcal{X}_n$ . Assume that, setting  $h_n(t) = d_n(r_n, \phi_n(t))$ :*

- (i)  $\sup_{s, t \in [0, 1]} |d_n(\phi_n(s), \phi_n(t)) - d_{h_n}(s, t)| \xrightarrow{d} 0$ .
- (ii)  $\sup_{t \in [0, 1]} |\lambda_n(\phi_n([0, t])) - t| \xrightarrow{d} 0$ .
- (iii)  $(h_n(t))_{t \in [0, 1]} \xrightarrow{d} (h(t))_{t \in [0, 1]}$  for the Skorokhod topology, where  $(h(t))_{t \in [0, 1]}$  is a random càdlàg function.

Then,  $\mathcal{X}_n \xrightarrow{d} \mathcal{T}_h$  for the rooted Gromov–Hausdorff–Prokhorov topology.

*Proof.* Since by Lemma 3.2 the assumption (iii) ensures that  $d_{\text{GHP}}(\mathcal{T}_{h_n}, \mathcal{T}_h) \rightarrow 0$ , to prove the proposition it suffices to show that the assumptions (i) and (ii) imply that  $d_{\text{GHP}}(\mathcal{X}_n, \mathcal{T}_{h_n}) \rightarrow 0$ .

First, by a straightforward extension of Skorokhod’s representation theorem (namely Theorem A.5 in Appendix A.3 below), we can assume that the convergences in assumptions (i–iii) hold almost surely, rather than in distribution. Note that the fact that  $\phi_n$  is assumed to be an admissible parametrization allows to view the pairs  $(\mathcal{X}_n, \phi_n)$  as random variables valued on a Polish space, which is the key to applying Theorem A.5.

Now, as previously, let  $\psi_n: [0, 1] \rightarrow \mathcal{T}_{h_n}$  be the quotient map in the construction of  $\mathcal{T}_{h_n}$ . Remember that, because we have completed  $[0, 1]/\sim_{d_{h_n}}$  to obtain  $\mathcal{T}_{h_n}$ , the map  $\psi_n$  is not surjective; however,  $\psi_n([0, 1])$  is dense in  $\mathcal{T}_{h_n}$ . Let then  $R_n$  be the correspondence from  $\mathcal{X}_n$  to  $\mathcal{T}_{h_n}$  defined as

$$R_n = \{(\phi_n(t), \psi_n(t)) : t \in [0, 1]\}.$$

Since  $\phi_n(0) = r_n$  (because  $\phi_n$  is an admissible parametrization of  $\mathcal{X}_n$ ) and since the root of  $\mathcal{T}_{h_n}$  is by construction  $r'_n := \psi_n(0)$ , we have  $r_n R_n r'_n$ . Thus, by Lemma A.4 to show that  $d_{\text{GHP}}(\mathcal{X}_n, \mathcal{T}_{h_n}) \rightarrow 0$  it suffices to show that  $\underline{\text{dis}}(R_n) \rightarrow 0$ .

First,  $\mathcal{X}_n R_n = \psi_n([0, 1])$  is dense in  $\mathcal{T}_{h_n}$  and  $R_n \mathcal{T}_{h_n} = \phi_n([0, 1])$  is dense in  $\mathcal{X}_n$ . Therefore,

$$\inf\left\{\varepsilon > 0 : (\mathcal{X}_n R_n)^{\varepsilon/2} \subset \mathcal{T}_{h_n} \text{ and } (R_n \mathcal{T}_{h_n})^{\varepsilon/2} \subset \mathcal{X}_n\right\} = 0. \tag{A.2}$$

Second, let  $(x, x')$  and  $(y, y')$  be any two elements of  $R_n$ , i.e. let  $s, t \in [0, 1]$  and set  $(x, x') = (\phi_n(s), \psi_n(s))$  and  $(y, y') = (\phi_n(t), \psi_n(t))$ . Then, by assumption (i),

$$|d_n(x, y) - d_{\mathcal{F}_{h_n}}(x', y')| = |d_n(\phi_n(s), \phi_n(t)) - d_{h_n}(s, t)| \xrightarrow[n \rightarrow \infty]{\text{unif. in } s, t} 0. \tag{A.3}$$

Let us fix  $\varepsilon > 0$ ; we now show that for all  $n$  large enough, for all Borel subset  $A \subset \mathcal{X}_n$ , we have

$$\lambda_{\mathcal{F}_{h_n}}((AR_n)^\varepsilon) + \varepsilon \geq \lambda_n(A). \tag{A.4}$$

First, note that because  $h_n$  converges in  $D$ , there exists  $k \geq 1$  such that for all  $n$  large enough, there exists  $t_0^n = 0 < t_1^n < \dots < t_k^n = 1$  such that for all  $j \in \{1, \dots, k\}$ ,  $\text{diam}(\psi_n(I_j^n)) < \varepsilon$ , where  $I_j^n = [t_{j-1}^n, t_j^n[$  for  $j < k$  and  $I_k^n = [t_{k-1}^n, 1]$ . Fix  $n_0 \geq 1$  so that for all  $n \geq n_0$ , for all  $j$ ,

$$|\lambda_n(\phi_n([0, t_j^n]) - t_j^n| < \frac{\varepsilon}{4k}, \quad \text{so that} \quad |\lambda_n(\phi_n(I_j^n)) - (t_j^n - t_{j-1}^n)| < \frac{\varepsilon}{2k}.$$

Now consider  $n \geq n_0$  and choose a Borel subset  $A \subset \mathcal{X}_n$ . Writing  $B = AR_n$ , define

$$J = \{j : B \cap \psi_n(I_j^n) \neq \emptyset\} \supset \{j : A \cap \phi_n(I_j^n) \neq \emptyset\},$$

and notice that because  $\text{diam}(\psi_n(I_j^n)) < \varepsilon$ , we have  $B^\varepsilon \supset \bigcup_{j \in J} \psi_n(I_j^n)$ . Then we have

$$\lambda_{\mathcal{F}_{h_n}}((AR_n)^\varepsilon) \geq \sum_{j \in J} \lambda_{\mathcal{F}_{h_n}}(\psi_n(I_j^n)) \geq \sum_{j \in J} (t_{j+1} - t_j).$$

But  $(t_{j+1} - t_j) \geq \lambda_n(\phi_n(I_j^n)) - \frac{\varepsilon}{2k}$ , so

$$\sum_{j \in J_B} (t_{j+1} - t_j) \geq \sum_{j \in J_B} \lambda_n(\phi_n(I_j^n)) - \frac{\varepsilon}{2} \geq \lambda_n(A \cap \phi_n([0, 1])) - \frac{\varepsilon}{2}.$$

Finally, note that we fixed  $n$  large enough so that  $\lambda_n(\phi_n([0, 1])) \geq 1 - \frac{\varepsilon}{2k} \geq 1 - \frac{\varepsilon}{2}$ , hence

$$\lambda_n(A \cap \phi_n([0, 1])) - \frac{\varepsilon}{2} \geq \lambda_n(A) - \varepsilon.$$

Putting the last three displays together, we have proved (A.4).

Finally, combining (A.2), (A.3) and (A.4), and recalling the Definition A.3 of the Prokhorov distortion, we see that  $\underline{\text{dis}}(R_n) \rightarrow 0$ . This concludes the proof.  $\square$

### A.3 An extension of Skorokhod's representation theorem

In this appendix, we discuss the extension of Skorokhod's representation theorem used at the beginning of the proof of Proposition 3.4. For the sake of rigour and completeness, we give a formal statement and a proof.

**Theorem A.5.** *Let  $X$  and  $Y$  be two Polish spaces endowed with their Borel  $\sigma$ -field,  $f: X \rightarrow Y$  be a Borel function and  $(\mu_n)_{n \geq 1}$  be a sequence of probability measures on  $X$  such that  $f_*\mu_n$  converges weakly. Then, there exists a probability measure  $\theta$  on  $X^{\mathbb{N}}$  whose  $n$ -th marginal is  $\mu_n$  and which is such that, letting  $(x_n) \sim \theta$ ,  $f(x_n)$  converges almost surely.*

*Proof.* Let  $(\varepsilon_m)_{m \geq 1}$  be a sequence of positive numbers such that  $\sum_m \varepsilon_m < \infty$  and let  $\nu$  be the weak limit of  $f_*\mu_n$ . Since  $Y$  is Polish, for all  $m \geq 1$  we can find a finite, measurable partition  $B_1^m, \dots, B_{k_m}^m$  of  $Y$  satisfying:

- (i)  $\nu(\partial B_i^m) = 0$  for all  $1 \leq i \leq k_m$ .
- (ii)  $\sum_i \nu(B_i^m) < \varepsilon_m$ , where the sum runs over the indices  $i$  s.t.  $\text{diam}(B_i^m) \geq \varepsilon_m$ .

Furthermore, we can assume that these partitions are refining as  $m$  increases; more specifically, that for each  $m \geq 1$ , there exist  $1 \leq j_1 < j_2 < \dots < j_{k_m} = k_{m+1}$  such that, for all  $1 \leq i \leq k_m$ ,

$$B_i^m = \bigcup_{j=j_{i-1}+1}^{j_i} B_j^{m+1},$$

with the convention  $j_0 = 0$ . Now, let us define a measurable partition  $A_1^m, \dots, A_{k_m}^m$  of  $X$  by setting  $A_i^m = f^{-1}(B_i^m)$ . Since  $f_*\mu_n \rightarrow \nu$  weakly, by the Portmanteau theorem  $\mu_n(A_i^m) \rightarrow \nu(B_i^m)$  for all  $i$  as  $n \rightarrow \infty$ . Therefore, for any sequence  $\delta_m > 0$  we can find an increasing sequence  $(N_m)_{m \geq 1}$  such that for all  $m \geq 1$  we have: for all  $n \geq N_m$  and  $1 \leq i \leq k_m$ ,

$$\left| \sum_{j=1}^i \mu_n(A_j^m) - \sum_{j=1}^i \nu(B_j^m) \right| < \delta_m. \tag{A.5}$$

Let us define, for all positive integers  $n, m$  and for all  $1 \leq i \leq k_m$ ,

$$\begin{cases} I_{n,i}^m = \left[ \sum_{j=1}^{i-1} \mu_n(A_j^m), \sum_{j=1}^i \mu_n(A_j^m) \right], \\ J_i^m = \left[ \sum_{j=1}^{i-1} \nu(B_j^m) + \delta_m, \sum_{j=1}^i \nu(B_j^m) - \delta_m \right], \end{cases}$$

where, by convention,  $J_i^m = \emptyset$  whenever  $\sum_{j=1}^{i-1} \nu(B_j^m) + \delta_m \geq \sum_{j=1}^i \nu(B_j^m) - \delta_m$ . Note that, by construction,  $J_i^m \subset I_{n,i}^m$  for all  $n \geq N_m$ .

We now build a sequence of random variables  $(x_n)$  such that  $x_n \sim \mu_n$ . For  $n \geq N_1$ , let  $m_n$  be the integer such that  $N_{m_n} \leq n < N_{m_n+1}$ . It is classic that Borel subsets of a Polish space are standard Borel (see e.g. [45, Proposition 3.3.7]), and this implies – either by a direct construction if the subset is countable, or using the Borel isomorphism theorem [45, Theorem 3.3.13] otherwise – that for each  $1 \leq i \leq k_{m_n}$  such that  $\mu_n(A_i^{m_n}) > 0$ , there exists a measurable map  $\phi_{n,i} : I_{n,i}^{m_n} \rightarrow A_i^{m_n}$  such that when  $U$  is drawn according to the normalized Lebesgue measure on  $I_{n,i}^{m_n}$ , the distribution of  $\phi_{n,i}(U)$  is  $\mu_n(\cdot | A_i^{m_n})$ . Thus, defining  $\phi_n : [0, 1[ \rightarrow X$  by

$$\phi_n(t) = \sum_{i=1}^{k_{m_n}} \phi_{n,i}(t) \mathbb{1}_{\{t \in I_{n,i}^{m_n}\}},$$

and then taking  $x_n = \phi_n(U)$ , where  $U$  is a uniform variable on  $[0, 1[$ , we get a random sequence  $(x_n)$  such that  $x_n \sim \mu_n$  for all  $n$ . Furthermore, notice that, by construction, if  $n \geq N_m$  and  $U \in I_{n,i}^m$ , then  $f(x_n) \in B_i^m$ .

Assume without loss of generality that the sequence  $(\delta_m)$  satisfies  $\sum_m \delta_m k_m < +\infty$ . We will now show that this implies that  $(f(x_n))$  is almost surely a Cauchy sequence. We say that the interval  $J_i^m$  is  $m$ -good if  $\text{diam}(B_i^m) \leq \varepsilon_m$ . By construction, for a fixed integer  $m \geq 1$ , the union  $G_m = \bigcup_i J_i^m$  of  $m$ -good intervals has Lebesgue measure at least  $1 - \varepsilon_m - 2(k_m + 1)\delta_m$ . Therefore, we have

$$\sum_{m \geq 1} \mathbb{P}(U \notin G_m) \leq \sum_{m \geq 1} (\varepsilon_m + 2(k_m + 1)\delta_m) < \infty,$$

and so, by the Borel–Cantelli lemma, there almost surely exists  $m^*$  such that for all  $m \geq m^*$ , there is a unique index  $i_m$  such that  $U$  is in the  $m$ -good interval  $J_{i_m}^m$ . This implies that, almost surely, for all  $n' \geq n \geq N_{m^*}$  and writing  $i = i_{m_n}$  to avoid clutter,  $U \in I_{n,i}^{m_n} \cap I_{n',i}^{m_{n'}}$  and so  $f(x_n), f(x_{n'}) \in B_i^{m_n}$  with  $\text{diam}(B_i^{m_n}) \leq \varepsilon_{m_n}$ . This shows that  $(f(x_n))$  is almost surely a Cauchy sequence, concluding the proof.  $\square$

**A.4 Tail of the size of critical Galton–Watson trees**

In order to make this article as self-contained as possible, we provide a short proof of Proposition 3.5 for the asymptotic equivalent of the probability that a critical Galton–Watson tree has size  $n$ , which is a key element in our study. As previously, we repeat the statement of the proposition here for convenience.

**Proposition 3.5.** *Let  $\hat{\mathcal{T}}$  be a critical Galton–Watson tree whose offspring distribution has a finite variance  $\sigma^2 > 0$  and is not supported on  $k\mathbb{N}$ , for any  $k \geq 2$ . Let  $\hat{A}_n$  denote the event  $\{\hat{\mathcal{T}} \text{ has } n \text{ vertices}\}$ . Then,*

$$\mathbb{P}(\hat{A}_n) \underset{n \rightarrow \infty}{\sim} \frac{1}{\sqrt{2\pi\sigma^2}} n^{-3/2}.$$

*Proof.* Let  $\hat{M}$  denote the offspring distribution of  $\hat{\mathcal{T}}$ , let  $\xi_1, \xi_2, \dots$  be i.i.d. copies of  $\hat{M} - 1$  and set  $S_n = \sum_{i=1}^n \xi_i$ . As was first noted by Dwass in [21],

$$\begin{aligned} \mathbb{P}(\hat{A}_n) &= \mathbb{P}(S_i \geq 0 \text{ for } 1 \leq i \leq n - 1 \text{ and } S_n = -1) \\ &= \frac{1}{n} \mathbb{P}(S_n = -1). \end{aligned}$$

The first equality is easily seen by marking the root as *to-visit*, and then at each step removing a vertex from the *to-visit* pile and adding its children to it: the procedure ends where there are no vertices left to visit – which happens after exactly  $n$  steps, where  $n$  is the total number of vertices in the tree; and if we let  $\xi_i$  denote the number of vertices added/removed from the pile at step  $i$ , then the number of vertices on the pile after step  $i$  is exactly  $S_i + 1$ .

The fact that  $\mathbb{P}(S_i \geq 0 \text{ for } 1 \leq i \leq n - 1, S_n = -1) = \frac{1}{n} \mathbb{P}(S_n = -1)$  has become folklore and is a special case of a result sometimes known as Kemperman’s formula or as the hitting time theorem – see e.g. [49, Theorem 3.14]. We mention a simple proof based on Dvoretzky and Motzkin’s cycle lemma [20]: let  $\mathcal{S}$  denote the set of vectors  $\mathbf{k} = (k_1, \dots, k_n)$  of increments of  $(S_n)_{n \geq 0}$  that are such that  $S_n = -1$ . For every permutation  $\sigma$  of  $\{1, \dots, n\}$ , if  $\mathbf{k} \in \mathcal{S}$  then  $\mathbf{k}_\sigma := (k_{\sigma(1)}, \dots, k_{\sigma(n)}) \in \mathcal{S}$  and  $\mathbb{P}(\xi_1 = k_1, \dots, \xi_n = k_n) = \mathbb{P}(\xi_1 = k_{\sigma(1)}, \dots, \xi_n = k_{\sigma(n)})$ .

Now, let  $\mathfrak{C}(n)$  denote the set of cyclic shifts of  $\{1, \dots, n\}$ , and define an equivalence relation  $\circ$  on  $\mathcal{S}$  by saying that  $\mathbf{k} \circ \mathbf{k}'$  when there exists  $\sigma \in \mathfrak{C}(n)$  such that  $\mathbf{k}' = \mathbf{k}_\sigma$ . By the cycle lemma, each equivalence class of  $\mathcal{S}/\circ$  has exactly one member such that the corresponding trajectory of  $(S_n)$  satisfies  $S_i \geq 0$  for  $1 \leq i \leq n - 1$  and  $S_n = -1$ . Let that member be the representative of its class, and denote by  $\mathcal{S}^*$  the set of those representatives. The cycle lemma also implies that each equivalence class of  $\mathcal{S}/\circ$  has cardinal  $n$ : indeed, if there existed  $\mathbf{k} \in \mathcal{S}^*$  and  $\sigma, \sigma' \in \mathfrak{C}(n)$  such that  $\sigma \neq \sigma'$  and  $\mathbf{k}_\sigma = \mathbf{k}_{\sigma'}$ , then by taking  $\rho = \sigma^{-1} \circ \sigma'$  we would have  $\rho \neq \text{Id}$  and yet  $\mathbf{k}_\rho \in \mathcal{S}^*$ , contradicting the lemma. As a result,

$$\begin{aligned} \mathbb{P}(S_n = -1) &= \sum_{\sigma \in \mathfrak{C}(n)} \sum_{\mathbf{k} \in \mathcal{S}^*} \prod_{i=1}^n \mathbb{P}(\xi_i = k_{\sigma(i)}) \\ &= n \mathbb{P}(S_i \geq 0, 1 \leq i \leq n - 1; S_n = -1). \end{aligned}$$

Finally, applying a local limit theorem – see e.g. [19, Theorem 3.5.2] – to  $(S_n)_{n \geq 0}$  yields  $\mathbb{P}(S_n = -1) \sim 1/\sqrt{2\pi\sigma^2 n}$ , thereby concluding the proof.  $\square$

**References**

[1] Primitivo B Acosta-Humánez, José A Capitán, and Juan J Morales-Ruiz, *Integrability of stochastic birth-death processes via differential Galois theory*, Mathematical Modelling of Natural Phenomena **15** (2020), 70. MR4183252

- [2] Louigi Addario-Berry, Luc Devroye, and Svante Janson, *Sub-gaussian tail bounds for the width and height of conditioned Galton–Watson trees*, *The Annals of Probability* **41** (2013), no. 2, 1072–1087. MR3077536
- [3] David Aldous, *Asymptotic fringe distributions for general families of random trees*, *The Annals of Applied Probability* **1** (1991), no. 2, 228–266. MR1102319
- [4] David Aldous, *The continuum random tree I*, *The Annals of Probability* **19** (1991), no. 1, 1–28. MR1085326
- [5] Vincent Bansaye, Sylvie Méléard, and Mathieu Richard, *Speed of coming down from infinity for birth-and-death processes*, *Advances in Applied Probability* **48** (2016), no. 4, 1183–1210. MR3595771
- [6] Eric Bapteste, Leo van Iersel, Axel Janke, Scot Kelchner, Steven Kelk, James O. McInerney, David A. Morrison, Luay Nakhleh, Mike Steel, Leen Stougie, and James Whitfield, *Networks: expanding evolutionary thinking*, *Trends in Genetics* **29** (2013), no. 8, 439–441.
- [7] Itai Benjamini and Oded Schramm, *Recurrence of distributional limits of finite planar graphs*, *Electronic Journal of Probability* **6** (2001), 1 – 13. MR1873300
- [8] Ulfar Bergthorsson, Keith L Adams, Brendan Thomason, and Jeffrey D Palmer, *Widespread horizontal transfer of mitochondrial genes in flowering plants*, *Nature* **424** (2003), 197–201.
- [9] François Bienvenu, Amaury Lambert, and Mike Steel, *Combinatorial and stochastic properties of ranked tree-child networks*, *Random Structures & Algorithms* **60** (2022), 653–689. MR4429840
- [10] Patrick Billingsley, *Convergence of probability measures*, second ed., *Wiley Series in Probability and Statistics*, John Wiley & Sons, Inc., New York, 1999. MR1700749
- [11] Denis Bosq and Hung T Nguyen, *A course in stochastic processes: Stochastic models and statistical inference*, Springer Science+Business Media, 1996. MR1426825
- [12] Alessandra Caraceni, Michael Fuchs, and Guan-Ru Yu, *Bijections for ranked tree-child networks*, *Discrete Mathematics* **345** (2022), no. 9, 112944. MR4418928
- [13] Adrián González Casanova, Juan Carlos Pardo, and José Luis Pérez, *Branching processes with interactions: Subcritical cooperative regime*, *Advances in Applied Probability* **53** (2021), no. 1, 251–278. MR4232756
- [14] Kenny S Crump and Charles J Mode, *A general age-dependent branching process. I*, *Journal of mathematical analysis and applications* **24** (1968), no. 3, 494–508. MR0237005
- [15] Kenny S Crump and Charles J Mode, *A general age-dependent branching process. II*, *Journal of mathematical analysis and applications* **25** (1969), no. 1, 8–17. MR0237005
- [16] Nicolas Curien, *Random graphs: the local convergence point of view*, *Lecture notes* (2018), <https://www.imo.universite-paris-saclay.fr/~curien/cours/cours-RG.pdf>.
- [17] Tal Dagan and William Martin, *Getting a better picture of microbial evolution en route to a network of genomes*, *Philosophical Transactions of the Royal Society B: Biological Sciences* **364** (2009), no. 1527, 2187–2196.
- [18] W. Ford Doolittle and Eric Bapteste, *Pattern pluralism and the Tree of Life hypothesis*, *Proceedings of the National Academy of Sciences* **104** (2007), no. 7, 2043–2049.
- [19] Rick Durrett, *Probability: Theory and examples*, Cambridge University Press, Cambridge, 2010. MR2722836
- [20] Aryeh Dvoretzky and Theodore Motzkin, *A problem of arrangements*, *Duke Mathematical Journal* **14** (1947), no. 2, 305–313. MR0021531
- [21] Meyer Dwass, *The total progeny in a branching process and a related random walk*, *Journal of Applied Probability* **6** (1969), no. 3, 682–686. MR0253433
- [22] Steven Neil Evans, *Probability and real trees*, *École d’Été de Probabilités de Saint-Flour XXXV-2005*, Springer Berlin, Heidelberg, 2008. MR2351587
- [23] Philippe Flajolet and Robert Sedgewick, *Analytic combinatorics*, Cambridge University Press, 2009. MR2876111
- [24] Clément Foucart, Pei-Sen Li, and Xiaowen Zhou, *On the entrance at infinity of Feller processes with no negative jumps*, *Statistics & Probability Letters* **165** (2020), 108859. MR4118938

- [25] Michael Fuchs, Hexuan Liu, and Tsan-Cheng Yu, *Limit theorems for patterns in ranked tree-child networks*, arXiv preprint (2022), arXiv:2204.07676. MR4418928
- [26] Weilong Hao and GB Golding, *Patterns of bacterial gene movement*, *Molecular biology and evolution* **21** (2004), no. 7, 1294–1307.
- [27] Peter Jagers, *Branching processes with biological applications*, Wiley, 1975, (note: out-of-print, not available online). MR0488341
- [28] Svante Janson, *Simply generated trees, conditioned Galton–Watson trees, random allocations and condensation*, *Probability Surveys* **9** (2012), 103–252. MR2908619
- [29] Aleksandr Vyacheslavovich Kalinkin, *Markov branching processes with interaction*, *Russian Mathematical Surveys* **57** (2002), no. 2, 241–304. MR1918194
- [30] Samuel Karlin and James McGregor, *The classification of birth and death processes*, *Transactions of the American Mathematical Society* **86** (1957), no. 2, 366–400. MR0094854
- [31] Samuel Karlin and James L McGregor, *The differential equations of birth-and-death processes, and the Stieltjes moment problem*, *Transactions of the American Mathematical Society* **85** (1957), no. 2, 489–546. MR0091566
- [32] Sungsik Kong, Joan Carles Pons, Laura Kubatko, and Kristina Wicke, *Classes of explicit phylogenetic networks and their biological and mathematical significance*, *Journal of Mathematical Biology* **84** (2022), no. 47. MR4417412
- [33] Igor Kortchemski, *Invariance principles for Galton–Watson trees conditioned on the number of leaves*, *Stochastic Processes and Their Applications* **122** (2012), no. 9, 3126–3172. MR2946438
- [34] Amaury Lambert, *The branching process with logistic growth*, *The Annals of Applied Probability* **15** (2005), no. 2, 1506–1535. MR2134113
- [35] Jean-François Le Gall, *Random trees and applications*, *Probability Surveys* **2** (2005), 245–311. MR2203728
- [36] C Randal Linder and Loren H Rieseberg, *Reconstructing patterns of reticulate evolution in plants*, *American journal of botany* **91** (2004), no. 10, 1700–1708.
- [37] James Mallet, *Hybrid speciation*, *Nature* **446** (2007), 279–283.
- [38] Jean-François Marckert and Abdelkader Mokkadem, *The depth first processes of Galton–Watson trees converge to the same Brownian excursion*, *The Annals of Probability* **31** (2003), no. 3, 1655–1678. MR1989446
- [39] Grégory Miermont, *Tessellations of random maps of arbitrary genus*, *Annales scientifiques de l’École normale supérieure* **42** (2009), no. 5, 725–781. MR2571957
- [40] Jeffrey P Mower, Saša Stefanović, Gregory J Young, and Jeffrey D Palmer, *Gene transfer from parasitic to host plants*, *Nature* **432** (2004), 165–166.
- [41] Olle Nerman, *On the convergence of supercritical general (C-M-J) branching processes*, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **57** (1981), no. 3, 365–395. MR0629532
- [42] Gabriel Berzunza Ojeda and Juan Carlos Pardo, *Branching processes with pairwise interactions*, arXiv preprint (2020), arXiv:2009.11820.
- [43] Todd L Parsons, *Invasion probabilities, hitting times, and some fluctuation theory for the stochastic logistic process*, *Journal of Mathematical Biology* **77** (2018), no. 4, 1193–1231. MR3856872
- [44] Emmanuel Schertzer and Florian Simatos, *Height and contour processes of Crump-Mode-Jagers forests (I): General distribution and scaling limits in the case of short edges*, *Electronic Journal of Probability* **23** (2018), 1–43. MR3835473
- [45] S. M. Srivastava, *A course on Borel sets*, *Graduate Texts in Mathematics*, vol. 180, Springer Berlin Heidelberg, Berlin, Heidelberg, 1998. MR1619545
- [46] Benedikt Stufler, *Local limits of large Galton–Watson trees rerooted at a random vertex*, *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, vol. 55, 2019, pp. 155–183. MR3901644
- [47] Benedikt Stufler, *Limits of random tree-like discrete structures*, *Probability Surveys* **17** (2020). MR4132643

- [48] Benedikt Stufler, *A branching process approach to level- $k$  phylogenetic networks*, *Random Structures & Algorithms* **61** (2022), no. 2, 397–421. MR4456034
- [49] Remco van der Hofstad, *Random graphs and complex networks*, vol. 1, Cambridge University Press, 2016. MR3617364

**Acknowledgments.** FB was supported by Dr. Max Rössler, the Walter Haefner Foundation and the ETH Zürich Foundation.