HAL open science

A limit theorem for trees of alleles in branching processes with rare neutral mutations

Jean Bertoin

To cite this version:

Jean Bertoin. A limit theorem for trees of alleles in branching processes with rare neutral mutations. 2009. hal-00373262v2

HAL Id: hal-00373262

https://hal.science/hal-00373262v2

Preprint submitted on 25 Jun 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A limit theorem for trees of alleles in branching processes with rare neutral mutations

Jean Bertoin

Laboratoire de Probabilités, Université Pierre et Marie Curie
175, rue du Chevaleret, F-75013 Paris, France,
and DMA, Ecole Normale Supérieure, Paris

Summary. We are interested in the genealogical structure of alleles for a Bienaymé-Galton-Watson branching process with neutral mutations (infinite alleles model), in the situation where the initial population is large and the mutation rate small. We shall establish that for an appropriate regime, the process of the sizes of the allelic sub-families converges in distribution to a certain continuous state branching process (i.e. a Jiřina process) in discrete time. Itô's excursion theory and the Lévy-Itô decomposition of subordinators provide fundamental insights for the results.

Key words. Weak convergence, branching process, neutral mutations, allelic partition, Lévy-Itô decomposition.

A.M.S. Classification. 60 J 80, 60 J 05

e-mail. `jean.bertoin@upmc.fr`

1 Introduction

Poisson point processes are a cornerstone of at least two fundamental contributions of Professor Kiyoshi Itô to Probability Theory, namely the Lévy-Itô decomposition of Lévy processes (Chapter 1 in [8]) and Itô's excursion theory [9]. The law of rare events, which stresses that Poisson variables arise as limiting distributions for the number of successes in a large number of independent trials where each trial has the same small probability of success, explains their prominent role amongst stochastic processes. For instance, Lévy processes can be viewed as weak limits of rescaled random walks and their jumps correspond to rare large steps of the latter. Informally, the law of rare events thus suggests the Poissonian structure of the jumps of Lévy processes, and this is indeed the core of the Lévy-Itô decomposition. A related but more delicate heuristic also applies to Itô's description of the excursions of Markov processes, as in discrete time, the succession of the excursions of a Markov chain away from a recurrent point forms an i.i.d. sequence of paths.

In this paper, we shall point out the relevance of this paradigm to study a question motivated by genetics. Recall that Bienaymé-Galton-Watson branching processes [1, 7, 10] model a population in which at every generation each individual begets according to a fixed offspring distribution and independently of the other individuals, and then dies. Imagine that neutral mutations may happen, so that a child can be either a clone of its parent or a mutant, and the reproduction laws of clones and of mutants are identical. We shall further suppose that each time a mutation occurs, it produces a mutant with a genetic type (allele) which has never been observed before; this setting has been referred to as the infinite alleles model by Kimura and Crow .

The allelic partition consists in decomposing the entire population into sub-families of individuals carrying the same allele. One important issue in the study of random population models with mutations (cf. the celebrated sampling formula of Ewens [5] for the Wright-Fisher model) concerns statistics of this allelic partition: what is the probability of observing allelic clusters of certain sizes, how to describe the random genealogical structure connecting these clusters to each other, ... Our main concern here will be to investigate asymptotics when the size of the population is large (typically because the number of ancestors is large) and mutations rare. We shall see that, under some mild conditions and for an appropriate regime, a non-degenerate limit exists and is conveniently described in terms of a certain continuous state branching process in discrete time [11]. It is well-known that continuous state branching processes bear close connexions to certain infinitely divisible distributions; in particular we shall provide a representation of the limiting allelic partitions in terms of Poisson point measures appearing in the Lévy-Itô decomposition of the jumps of an underlying Lévy process.

Let us give a rough idea of the orders of magnitude of the quantities involved. We shall consider a fixed reproduction law with unit mean and finite variance, and let the Galton-Watson process start from n ancestors having all the same genetic type. It is well-known that if n generations represent one unit of time and if we rescale the population at each generation by a factor $1/n$, then the rescaled Galton-Watson process converges in distribution as n tends to infinity to a Feller diffusion. We also suppose that neutral mutations affect each child with probability $1/n$. The scaling between population sizes, generations and mutation rates should not come as a surprise since it is precisely the regime of interest for other standard population models, such as the Wright-Fisher model and Kingman coalescent [13]. Recall that such a critical Galton-Watson process becomes extinct after roughly n generations, and that the total population is of order n^2. So there are only a few mutations at each generation and thus about n different alleles; furthermore the largest allelic sub-families have size of order n^2.

Our main result can be described as follows. We use the universal tree $\mathbb{U}$, that is the set of finite sequences of integers (including the empty sequence $\varnothing$ that serves as the root of $\mathbb{U}$) to record the genealogy of alleles, and define the tree of alleles as a random process $\mathcal{A}$ on $\mathbb{U}$, such that the values at vertices are given by the sizes of the corresponding allelic sub-families, with the convention that sizes are ranked in the decreasing order on each sibling. We consider a fixed reproduction law which is critical and has finite variance, and for every integer n, a Galton-Watson process with this reproduction law, started from n ancestors and in which mutations occur at random with rate $1/n$. We write $\mathcal{A}^{(n)}$ for the process on $\mathbb{U}$ that describes the corresponding tree of alleles. Then as n tends to infinity, the rescaled tree of alleles $n^{-2}\mathcal{A}^{(n)}$ converges in the sense of finite dimensional distributions towards a process $\mathcal{A}$ on $\mathbb{U}$ with values

in $(0, \infty)$. The latter describes the genealogy of a continuous state branching process in discrete time with an inverse Gaussian reproduction law. We stress that its law only depends on the variance of the offspring distribution of the Galton-Watson process, and hence may be viewed as a universal tree of alleles.

The plan of this paper is as follows. In Section 2, we first present the general setting, stressing the role of the general branching property for the study of Galton-Watson processes with neutral mutations. Then we compute explicitly reproduction laws related to allelic sub-families and point at a connexion with certain downward-skip-free random walks. Such questions have been addressed from a different point of view in [2] to which the present work can be viewed as a complement and a sequel. Section 3 provides some background on continuous state branching processes and convergence of rescaled Galton-Watson processes. The main asymptotic results, namely Proposition 2 and Theorem 1 are stated and then proved in Section 4.

2 Galton-Watson processes with neutral mutations

2.1 Basic definitions and branching properties

In a Galton-Watson process with neutral mutations, every individual reproduces according to the same distribution and independently of the other individuals, no matter whether it is a mutant or a clone. Of course, a clone child of a mutant bears the same allele as its parent. Recall also that we are working in the infinite alleles setting, i.e. the same genetic type cannot be recovered from a cycle of mutations. Our basic data are hence provided by a pair of non-negative integer-valued random variables

$$\xi = (\xi^{(\mathrm{c})}, \xi^{(\mathrm{m})})$$

which describes the number of clone-children and the number of mutant-children of a typical individual. In this paper, we shall mainly be interested in a special situation which appears commonly as a model in population genetics, namely where mutations affect each child according to a fixed probability and independently of the other children (in other words, the conditional distribution of $\xi^{(\mathrm{m})}$ given $\xi^{(\mathrm{c})} + \xi^{(\mathrm{m})} = \ell$ is binomial with parameter (ℓ, p)). However the first steps of the analysis can be carried on without difficulties using the general framework. We assume throughout this work that

$$\mathbb{E}(\xi^{(\mathrm{c})}) \leq 1\,,$$

i.e. the process of clones is critical or sub-critical [1]; and we further implicitly exclude the degenerate cases when $\xi^{(\mathrm{c})} \equiv 0$, or $\xi^{(\mathrm{m})} \equiv 0$. For every integer $a \geq 1$, we denote by $\mathbb{P}_a$ the law of a Galton-Watson process with neutral mutations, started from a ancestors having the same genetic type and with reproduction law given by that of $\xi = (\xi^{(\mathrm{c})}, \xi^{(\mathrm{m})})$.

The basic *branching property* states that for every fixed generation, conditionally on the number of individuals at that generation, the descents of those individuals are given by independent copies of the initial process, independently of the preceding generations. It is natural

[1] Note that this is weaker than assuming that $\mathbb{E}(\xi^{(\mathrm{c})} + \xi^{(\mathrm{m})}) \leq 1$ which was required in [2].

to expect that this branching property should hold more generally for certain stopping rules, and that this is indeed the case will play an important role in our analysis. For the sake of simplicity, we shall now present such an extension in a rather informal way, referring to Chauvin [3] for technical details.

The genealogy of each ancestor is conveniently described by a planar rooted tree, with edges connecting parents to children. More precisely, this requires an additional ordering of the children of each individual, and in this direction we may decide to rank siblings uniformly at random. A line is defined as a family of edges such that every branch from the root (i.e. the ancestor) contains at most one edge in that family. For instance, the edges between parents at generation $k \in \mathbb{Z}_+$ and children at generation $k+1$ form a line. A stopping line should be thought of as a random line such that for every edge in the tree, the event that this edge is part of the line only depends on the marks found on the path from the root to that edge. Recall that every edge of the tree corresponds to a pair of individuals (parent,child), and denote by C_τ the subset of children in the family of edges of some stopping line τ. By removing the edges of τ, we disconnect the genealogical tree into sub-trees whose roots are formed on the one hand by the ancestor, and on the other hand by the individuals in C_τ. The *general branching property* then states that conditionally on C_τ, the sub-trees rooted at the individuals in C_τ are independent copies of the initial genealogical tree, and also independent of the initial tree pruned along τ.

We now take into account mutations by assigning marks to the edges between parents and their mutant children. Since we are interested by the genealogy of alleles (or equivalently, of mutants), it is convenient to say that an individual has the k-th type if its genotype has been affected by k mutations, that is if its ancestral line comprises exactly k marks. Plainly, the family $\tau(k)$ of edges connecting a parent of the $(k-1)$-th type to a mutant child is a stopping line, and the set $C_{\tau(k)}$ coincides with that of the mutants of the k-th type. We denote by T_k the total population of individuals of the k-th type and by M_k the total number of mutants of k-th type, agreeing that mutants of the 0-th type are the ancestors (so $M_0 = a$, $\mathbb{P}_a$-a.s.). The general branching property should make the following statement obvious; we refer the reader to e.g. Chapter Ten in Taib [15] for a rigorous argument.

Lemma 1 *Under* $\mathbb{P}_a$,

$$(M_k, k \in \mathbb{Z}_+)$$

is a standard Galton-Watson process with reproduction law $\mathbb{P}_1(M_1 \in \cdot)$. *More generally,*

$$((T_k, M_{k+1}), k \in \mathbb{Z}_+)$$

is a Markov chain with transition probabilities

$$\mathbb{P}_a(T_k = n', M_{k+1} = m' \mid T_{k-1} = n, M_k = m) \;=\; \mathbb{P}_m(T_0 = n', M_1 = m')\,.$$

Remark. We stress the fact that the chain $(T_k, k \in \mathbb{Z}_+)$ of the sizes of sub-populations with given types is not Markov; nonetheless it can be viewed as a *hidden* Markov chain. In this direction, we also point out that $((T_k, M_k), k \in \mathbb{Z}_+)$ is Markovian, since the transition probabilities of the chain $((T_k, M_{k+1}), k \in \mathbb{Z}_+)$ only depend on the second coordinate. Indeed,

by a straightforward application of the general branching property, one gets (assuming implicitly that the events on which we condition have positive probability)

$$\begin{aligned} & \mathbb{P}_a(T_{k+1} = n', M_{k+1} = m' \mid M_k = m, T_k = n) \\ = \quad & \mathbb{P}_{m'}(T_0 = n') \frac{\mathbb{P}_a(T_k = n, M_{k+1} = m' \mid M_k = m)}{\mathbb{P}_a(T_k = n \mid M_k = m)} \\ = \quad & \frac{\mathbb{P}_{m'}(T_0 = n')}{\mathbb{P}_m(T_0 = n)} \mathbb{P}_m(T_0 = n, M_1 = m') \,. \end{aligned}$$

Next, observe that in an infinite alleles model, the genealogy of individuals naturally induces a genealogy for the alleles in that population. Indeed, we may identify alleles and mutants, which enables us to use the set of new mutants plus a root corresponding to the ancestors of the population (recall that we assume that all the ancestors have the same genetic type) as the set of vertices. We draw an edge between the root and mutants of the 1st type, and for every $k \geq 1$ we also draw an edge between a mutant of the k-th type and a mutant of the $(k+1)$-type if and only if the path connecting these individuals in the genealogical tree does not contains other mutants. Hence the set of alleles has a natural structure of rooted tree. Note that for $k \geq 1$, M_k corresponds to the number of vertices at the k-th level [2] in the tree of alleles.

Our main goal in this paper is to establish asymptotic features on the genealogy of allelic sub-families, and in this direction, it will be convenient to view the latter as random processes indexed by the universal tree. More precisely, introduce the set of finite sequences of positive integers

$$\mathbb{U} := \bigcup_{k \in \mathbb{Z}_+} \mathbb{N}^k \,,$$

where $\mathbb{N} = \{1, 2, \ldots\}$ and $\mathbb{N}^0 = \{\varnothing\}$. Let us briefly recall some standard notation in this setting: if $u = (u_1, \ldots, u_k)$ is vertex at level $k \geq 0$ in $\mathbb{U}$, then the children of u are $uj := (u_1, \ldots, u_k, j)$ for $j \in \mathbb{N}$. We also denote by $|u|$ the level of the vertex u, with the convention that the root has level 0, i.e. $|\varnothing| = 0$. We now take advantage of the natural tree structure of $\mathbb{U}$ to record the genealogy of allelic sub-families together with their sizes.

Given a Galton-Watson process with neutral mutations, we construct recursively a process $\mathcal{A} = (\mathcal{A}_u : u \in \mathbb{U})$ as follows. First, $\mathcal{A}_\varnothing = T_0$ is the size of the sub-population without mutation. Next, recall that M_1 denotes the number of mutants of the first type. We enumerate the M_1 allelic sub-populations of the first type in the decreasing order of their sizes, with the convention that in the case of ties, sub-populations of the same size are ranked uniformly at random. We denote by $\mathcal{A}_j$ the size of the j-th allelic sub-populations of the first type, agreeing that $\mathcal{A}_j = 0$ if $j > M_1$. We then complete the construction at all levels by iteration in an obvious way. Specifically, if $\mathcal{A}_u = 0$ for some $u \in \mathbb{U}$, then $\mathcal{A}_{uj} = 0$ for all $j \in \mathbb{N}$. Otherwise, we enumerate in the decreasing order of their sizes the allelic sub-populations of type $|u| + 1$ which descend from the allelic sub-family indexed by the vertex u, and then $\mathcal{A}_{uj}$ is the size of this j-th sub-family (as before, in the case of ties, sub-families are ordered uniformly at random, and empty sub-families have size 0). See Figure 1 for an example. We call the process $\mathcal{A} = (\mathcal{A}_u : u \in \mathbb{U})$ the *tree of alleles*.

[2] For the sake of clarity we shall keep the name *generation* for the distance to the root of individuals in the genealogical tree, and use the name *level* when dealing with the structure of alleles.

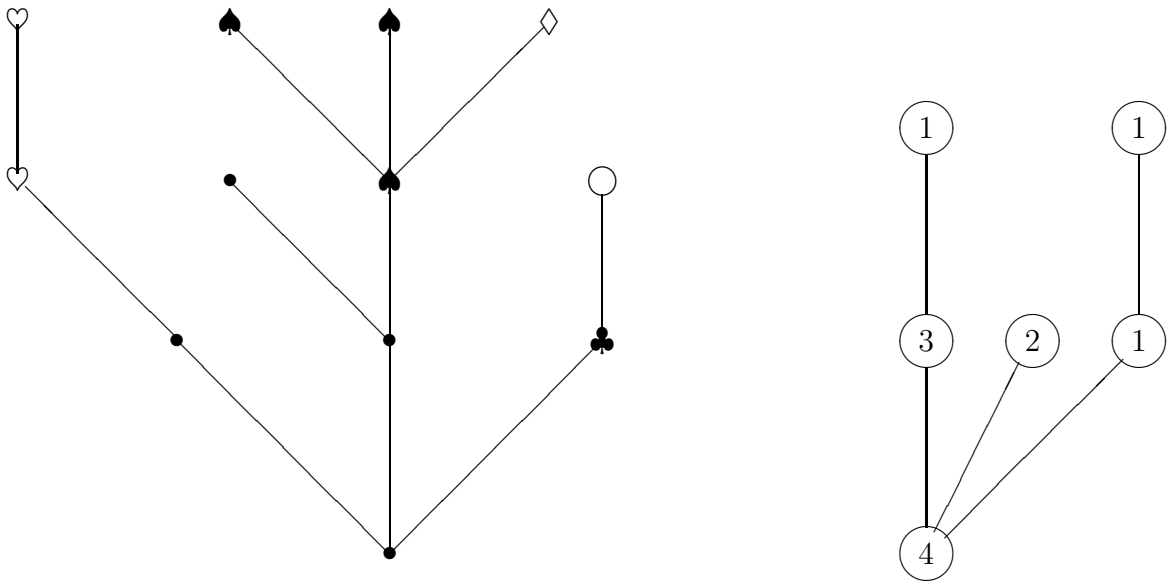

Figure 1: *Genealogical tree with mutations (left) and tree of alleles (right). The symbols* $\bullet, \spadesuit, \heartsuit, \diamondsuit, \clubsuit, \bigcirc$ *represent the different alleles. The labels on the tree of alleles are the sizes of the corresponding allelic sub-families; sub-families with zero size (i.e. which are empty) are omitted.*

It is important to observe that the transition probabilities of the chain $((T_k, M_{k+1}), k \in \mathbb{Z}_+)$ in Lemma 1 depend only on the second coordinate, and that the latter alone is a Galton-Watson process. This suggests that the tree of alleles should enjoy some kind of branching property. In order to give a formal statement, is convenient to define first the (outer) degree of the tree of alleles $\mathcal{A}$ at some vertex $u \in \mathbb{U}$ as

$$d_u := \max\{j \geq 1 : \mathcal{A}_{uj} > 0\}\,,$$

where we agree that $\max \varnothing = 0$. In words, d_u is the number of allelic sub-populations of type $|u|+1$ which descend from the allelic sub-family indexed by the vertex u; in particular $d_\varnothing = M_1$. We shall also need the following notation. Let γ be a random variable in $\mathbb{N}^2$, $d \geq 1$ an integer, and $\gamma^{(d)} = (\gamma_1, \ldots, \gamma_d)$ where the γ_i are independent copies of γ. We then denote by $\gamma^{(d\downarrow)}$ the rearrangement of $\gamma^{(d)}$ in the decreasing order of the first coordinate, with the convention that in the case of ties, the variables γ_i with the same first coordinate are ranked uniformly at random.

The characterization of the probabilistic structure of the tree of alleles that we are now ready to present stems again easily from the general branching property by iteration.

Lemma 2 *For every integers $a \geq 1$ and $k \geq 0$, the tree of alleles fulfills the following properties under $\mathbb{P}_a$ conditionally on $((\mathcal{A}_v, d_v) : |v| \leq k)$:*

(i) *the families of variables*

$$((\mathcal{A}_{uj}, d_{uj}) : 1 \leq j \leq d_u) \ , \qquad u \text{ vertex at level } k \text{ such that } \mathcal{A}_u > 0 \, ,$$

are independent,

(ii) *for each vertex u at level k with $\mathcal{A}_u > 0$, the d_u-tuple $((\mathcal{A}_{uj}, d_{uj}) : 1 \leq j \leq d_u)$ is distributed as $(T_0, M_1)^{(d_u \downarrow)}$ under $\mathbb{P}_1$.*

Of course Lemma 2 of much more informative than the sole Markovian description of the chain $((T_k, M_{k+1}), k \in \mathbb{Z}_+)$ in Lemma 1 as it retains the information about the genealogy of the allelic sub-families and not merely the sizes of populations of a given type. In this direction, observe that

$$T_k = \sum_{|u|=k} \mathcal{A}_u \quad \text{and} \quad M_{k+1} = \sum_{|u|=k} d_u \, .$$

2.2 Calculation of reproduction laws

We shall now determine the transition probabilities that appear in Lemma 1. Essentially, this has been achieved recently in [2] using an approach that largely relies on Harris connexion between downward-skip-free random walks and standard Galton-Watson processes, extended to encompass the situation where neutral mutations occur. Here, we shall use a different route, developing calculations that involve generating functions in the case when mutants are supposed to be sterile.

We denote the law of $\xi = (\xi^{(\mathrm{c})}, \xi^{(\mathrm{m})})$ by $\pi = (\pi_{k,\ell} : k, \ell \in \mathbb{Z}_+)$, that is

$$\pi_{k,\ell} := \mathbb{P}(\xi^{(\mathrm{c})} = k, \xi^{(\mathrm{m})} = \ell) \, .$$

We also introduce the generating function

$$g(x, y) := \sum_{k,\ell=0}^{\infty} x^k y^\ell \pi_{k,\ell} = \mathbb{E}(x^{\xi^{(\mathrm{c})}} y^{\xi^{(\mathrm{m})}}) \, , \qquad x, y \in [0, 1] \, .$$

As we are interested in the joint distribution of the total number of individuals of the 0-th type and the number of mutants of the 1-st type, we may imagine a two-type branching process such that clones reproduce independently of each other according to the same distribution π, while mutants are sterile, i.e. have no progeny a.s. We write φ for the generating function of the total population of 0-th type and the number of mutants when there is a single ancestor, i.e.

$$\varphi(x, y) := \mathbb{E}_1(x^{T_0} y^{M_1}) \, , \qquad x, y \in [0, 1] \, ,$$

so that by the branching property, the generating function of (T_0, M_1) under $\mathbb{P}_a$ is φ^a. The following result is a slight extension of Theorem 1(ii) of [2] (recall that here we only assume that $\mathbb{E}(\xi^{(\mathrm{c})}) \leq 1$ and have an arbitrary number of ancestors, while in [2] we worked with a single ancestor and assumed that $\mathbb{E}(\xi^{(\mathrm{c})} + \xi^{(\mathrm{m})}) \leq 1$). It can be viewed as a generalization of the well-known Dwass formula [4] for the distribution of the total population in standard Galton-Watson processes.

Proposition 1 (i) *The generating function φ is determined by the equation*

$$\varphi(x,y) = xg(\varphi(x,y),y)\,, \qquad x,y \in [0,1].$$

(ii) *The distribution of (T_0, M_1) is given by*

$$\mathbb{P}_a(T_0 = n, M_1 = \ell) = \frac{a}{n}\pi^{*n}_{n-a,\ell}\,, \qquad n \geq a \geq 1 \text{ and } \ell \geq 0\,,$$

where π^{*n} denotes the n-th convolution power of π (i.e. π^{*n} is the distribution of the sum of n i.i.d. copies of ξ).

Proof: (i) A standard application of the branching property at the first generation gives

$$\begin{aligned}\varphi(x,y) &= \mathbb{E}_1(x^{T_0}y^{M_1}) \\ &= x\sum_{i,j=0}^{\infty}\mathbb{P}(\xi^{(\mathrm{c})}=i,\xi^{(\mathrm{m})}=j)\mathbb{E}_1(x^{T_0}y^{M_1})^i y^j \\ &= xg(\varphi(x,y),y)\,.\end{aligned}$$

This invites us to consider the equation in the variable $z \in [0,1]$

$$\frac{g(z,y)}{z} = \frac{1}{x}\,, \tag{1}$$

where $x, y \in (0,1]$ are fixed. Our assumptions $\mathbb{E}(\xi^{(\mathrm{c})}) \leq 1$ and $\xi^{(\mathrm{c})} \not\equiv 1$ imply that $g(0,y) > 0$, and hence $\lim_{z\to 0+} g(z,y)/z = \infty$. On the other hand, the derivative of $z \to z^{-1}g(z,y)$ is $z \to z^{-2}(z\partial_z g(z,y) - g(z,y))$, and this derivative is strictly negative when $z > 0$ is sufficiently small. This ensures that for each fixed $y \in [0,1]$ and $x > 0$ small enough, the equation (1) has a unique solution $z = \varphi(x,y)$, and this suffices to determine the law of (T_0, M_1).

(ii) We shall now derive explicitly the law of (T_0, M_1) under $\mathbb{P}_a$ from its generating function φ^a using the classical Lagrange inversion formula. For each fixed $y \in [0,1]$, the function $x \to g(x,y)$ is analytic with $g(0,y) \neq 0$. More precisely, we have

$$g(x,y) = \sum_{k=0}^{\infty} a_k(y)x^k \quad \text{with} \quad a_k(y) := \sum_{\ell=0}^{\infty} y^\ell \pi_{k,\ell}\,.$$

According to Lagrange inversion formula (see for instance Section 5.1 in [16]), the a-th power of the solution to the equation (1) with $y \in [0,1]$ fixed and $x > 0$ sufficiently small, can be expressed in the form

$$\varphi^a(x,y) = \sum_{n=1}^{\infty}\frac{a}{n}\alpha^{*n}_{n-a}\,x^n\,,$$

where α^{*n} stands for the n-th convolution power of the finite measure $\alpha = (a_k(y) : k \in \mathbb{Z}_+)$. Observe that the generating function of α is $x \to g(x,y)$, so that of α^{*n} is

$$x \to g(x,y)^n = \sum_{k=0}^{\infty} x^k \left(\sum_{\ell=0}^{\infty} y^\ell \pi^{*n}_{k,\ell}\right)\,.$$

Hence we have

$$\alpha_k^{*n} = \sum_{\ell=0}^{\infty} y^\ell \pi_{k,\ell}^{*n} ,$$

and we conclude that

$$\varphi^a(x,y) = \sum_{n=1}^{\infty}\sum_{\ell=0}^{\infty} \frac{a}{n} \pi_{n-a,\ell}^{*n} x^n y^\ell ,$$

which completes the proof of (ii). □

As generating functions easily yield moments of variables, one immediately deduces from Proposition 1 simple criteria to decide whether the number of mutant children M is critical, sub-critical, or super-critical, or has a finite second moment.

Corollary 1 (i) *Suppose that the mean number of clone children is sub-critical, i.e.* $\mathbb{E}(\xi^{(\mathrm{c})}) < 1$. *Then*

$$\mathbb{E}_a(M_1) = a\frac{\mathbb{E}(\xi^{(\mathrm{m})})}{1-\mathbb{E}(\xi^{(\mathrm{c})})} = \mathbb{E}(\xi^{(\mathrm{m})})\mathbb{E}_a(T_0) ,$$

and in particular

$$\mathbb{E}_1(M_1) \begin{cases} < 1 \\ = 1 \\ > 1 \end{cases} \quad \Longleftrightarrow \quad \mathbb{E}(\xi^{(\mathrm{c})} + \xi^{(\mathrm{m})}) \begin{cases} < 1 \\ = 1 \\ > 1 \end{cases}$$

Further

$$\mathbb{E}_1(M_1^2) < \infty \quad \Longleftrightarrow \quad \mathbb{E}((\xi^{(\mathrm{c})} + \xi^{(\mathrm{m})})^2) < \infty .$$

(ii) *If* $\mathbb{E}(\xi^{(\mathrm{c})}) = 1$, *then* $\mathbb{E}_1(M_1) = \infty$.

Proof: Recall that the first moment of an integer-valued variable is given by the left-derivative at 1 of its generating function. We get from Proposition 1(i)

$$\frac{\partial\varphi}{\partial y}(1,y) = \frac{\partial\varphi}{\partial y}(1,y)\frac{\partial g}{\partial x}(\varphi(1,y),y) + \frac{\partial g}{\partial y}(1,y) .$$

Since $\varphi(1,1) = 1$, this identity forces

$$\frac{\partial\varphi}{\partial y}(1,1) = \mathbb{E}_1(M_1) = \infty$$

when

$$\frac{\partial g}{\partial x}(1,1) = \mathbb{E}(\xi^{(\mathrm{c})}) = 1$$

(recall that $\mathbb{E}(\xi^{(\mathrm{m})}) > 0$ by assumption), whereas it entails

$$\mathbb{E}_1(M_1) = \frac{\mathbb{E}(\xi^{(\mathrm{m})})}{1-\mathbb{E}(\xi^{(\mathrm{c})})}$$

whenever $\mathbb{E}(\xi^{(\mathrm{c})}) < 1$.

Observe further that the process of the number of clone children is a branching process with offspring distribution given by the law of $\xi^{(\mathrm{c})}$. In particular, in the sub-critical case $\mathbb{E}(\xi^{(\mathrm{c})}) < 1$, the total population of clones has a finite expectation given by $\mathbb{E}_a(T_0) = a/(1-\mathbb{E}(\xi^{(\mathrm{c})}))$. The first equivalence in (i) follows readily. Similar calculations involving the second derivative of generating functions yield the second equivalence in (i). □

2.3 Construction from a random walk

The starting point of this section is the observation that the transition probabilities of the Markov chain $((T_k, M_{k+1}) : k \in \mathbb{Z}_+)$ have a simple interpretation in terms of random walks. In this direction, let us first introduce some notation. We consider a sequence $(\xi_n = (\xi_n^{(\mathrm{c})}, \xi_n^{(\mathrm{m})}) : n \in \mathbb{N})$ of i.i.d. variables with law π, and then the random walk started from $a \geq 1$ and with steps $\xi^{(\mathrm{c})} - 1$,

$$S_n^{(\mathrm{c})} := a + \xi_1^{(\mathrm{c})} + \cdots + \xi_n^{(\mathrm{c})} - n\,, \qquad n \in \mathbb{Z}_+\,.$$

It is convenient to use the (slightly abusive) notation $\mathbb{P}_a$ for the law of $(S_n^{(\mathrm{c})} : n \in \mathbb{Z}_+)$. We also define the first hitting times

$$\varsigma(j) := \inf\{n \in \mathbb{Z}_+ : S_n^{(\mathrm{c})} = -j\}\,, \qquad j \in \mathbb{Z}_+\,,$$

and

$$\Sigma(j) := \sum_{i=1}^{\varsigma(j)} \xi_i^{(\mathrm{m})}\,.$$

We stress that our basic assumption $\mathbb{E}(\xi^{(\mathrm{c})}) \leq 1$ ensures that the random walk $S^{(\mathrm{c})}$ does not drift to $+\infty$, and hence the passage times $\varsigma(j)$ are finite a.s. The first identity in next lemma can be viewed as a two-dimensional extension of the well-known result of Otter and Dwass (see e.g. Section 6.2 in [14]) which relates the distribution of the total population in a Galton-Watson process to that of the first hitting time of 0 of a random walk.

Lemma 3 *The pairs of random variables*

$$(\varsigma(0), \Sigma(0)) \quad \textit{and} \quad (T_0, M_1)$$

have the same distribution under $\mathbb{P}_a$. Further, the shifted sequence $(\xi_{\varsigma(0)+j} : j \in \mathbb{N})$ consists of i.i.d. variables with law π and is independent of $(\varsigma(0), \Sigma(0))$.

Proof: Introduce for $a = 1$ the generating function

$$\tilde{\varphi}(x, y) := \mathbb{E}_1(x^{\varsigma(0)} y^{\Sigma(0)})\,, \qquad x, y \in [0, 1]\,.$$

Because $(S_n : n \in \mathbb{Z}_+)$ is a downwards skip free random walk, an application of the strong Markov property at its first downward passage times shows readily that for an arbitrary integer $a \geq 1$

$$\mathbb{E}_a(x^{\varsigma(0)} y^{\Sigma(0)}) \;=\; \tilde{\varphi}(x, y)^a\,, \qquad x, y \in [0, 1]\,.$$

Now we return to the case $a = 1$; by conditioning on the first step of the random walk, we get the obvious identity

$$\begin{aligned}
\tilde{\varphi}(x, y) &= \mathbb{E}_1(x^{\varsigma(0)} y^{\Sigma(0)}) \\
&= x \sum_{k,\ell=0}^{\infty} \tilde{\varphi}(x, y)^k y^\ell \pi_{k,\ell} \\
&= x g(\tilde{\varphi}(x, y), y)\,,
\end{aligned}$$

where g denotes the generating function of $\xi = (\xi^{(\mathrm{c})}, \xi^{(\mathrm{m})})$. Thus $\tilde{\varphi}$ solves the equation of Proposition 1(i), which establishes our first claim. As the hitting time $\varsigma(0)$ is a stopping time, an application of the strong Markov property then yields the second assertion. $\square$

Next, set $\tilde{T}_0 := \varsigma(0)$, $\tilde{M}_1 := \Sigma(0)$ and define for every $k \in \mathbb{N}$ by an implicit recurrence

$$\tilde{T}_0 + \cdots + \tilde{T}_k = \varsigma(\tilde{M}_1 + \cdots + \tilde{M}_k)$$

and

$$\tilde{M}_1 + \cdots + \tilde{M}_{k+1} = \Sigma(\tilde{T}_0 + \cdots + \tilde{T}_k) = \Sigma(\varsigma(\tilde{M}_1 + \cdots + \tilde{M}_k))\,.$$

Figure 2 below depicts these quantities.

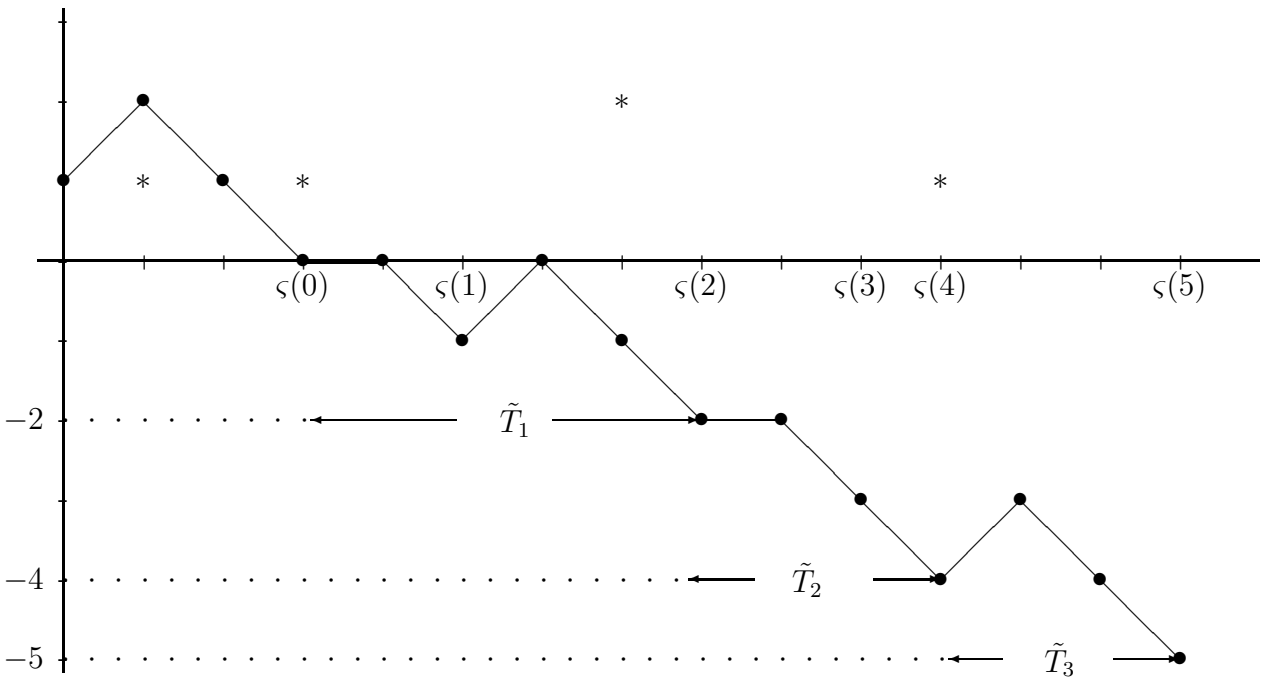

Figure 2: *The graph of the random walk $S^{(\mathrm{c})}$; the $*$ represent the non-zero values of the variables $\xi^{(\mathrm{m})}$. Here $\tilde{M}_1 = 2$, $\tilde{M}_2 = 2$, $\tilde{M}_3 = 1$ and $\tilde{M}_4 = 0$.*

Corollary 2 *For every $a \geq 1$, the chains $((T_k, M_{k+1}) : k \in \mathbb{Z}_+)$ and $((\tilde{T}_k, \tilde{M}_{k+1}) : k \in \mathbb{Z}_+)$ have the same distribution under $\mathbb{P}_a$.*

Proof: It is immediately checked by induction that each $\tau_k := \tilde{T}_0 + \cdots + \tilde{T}_k$ is a stopping time in the natural filtration $(\mathcal{G}(n))_{n\in\mathbb{N}}$ generated by the i.i.d. sequence $(\xi_n : n \in \mathbb{N})$, and that M_{k+1} is $\mathcal{G}(\tau_k)$-measurable. By an application of the strong Markov property, we get that $((\tilde{T}_k, \tilde{M}_{k+1}) : k \in \mathbb{Z}_+)$ is a homogeneous Markov chain. More precisely, the conditional distribution of $(\tilde{T}_k, \tilde{M}_{k+1})$ given $\tilde{T}_{k-1} = t$ and $\tilde{M}_k = m$ is that of $(\varsigma(0), \Sigma(0))$ under $\mathbb{P}_m$. Combining these observation with Lemmas 1 and 3 completes the proof. $\square$

More generally, we can apply Lemma 3 to construct from the i.i.d. variables ξ_n a random process $\mathcal{A}'$ indexed by $\mathbb{U}$ with the same distribution as tree of alleles $\mathcal{A}$, by making use of the characterization of the law of the latter in Lemma 2. To start with, the process $\mathcal{A}'$ fulfills the following two requirements. First, if $\mathcal{A}'_u = 0$ for some $u \in \mathbb{U}$, then $\mathcal{A}'_{uj} = 0$ for all $j \in \mathbb{N}$. Second, for every vertex $u \in \mathbb{U}$ such that $\mathcal{A}'_u > 0$, the (outer) degree of $\mathcal{A}'$ at u,

$$d'_u := \#\{j \in \mathbb{N} : \mathcal{A}'_{uj} > 0\}\,,$$

is a finite number and $\mathcal{A}'_{uj} > 0$ if and only if $j \leq d'_u$. We set $\mathcal{A}'_\varnothing = \varsigma(0)$ and $d'_\varnothing = \Sigma(0)$. Next, consider the increments

$$\lambda(j) := \varsigma(j) - \varsigma(j-1) \;\text{ and }\; \delta(j) = \Sigma(j) - \Sigma(j-1))\,, \qquad j \geq 1\,,$$

For vertices at the first level, $((\mathcal{A}'_j, d'_j) : 1 \leq j \leq d'_\varnothing)$ is given by the rearrangement of the sequence $((\lambda(j), \delta(j)) : 1 \leq j \leq d'_\varnothing)$ in the decreasing order the first coordinate $\lambda(j)$ (with the usual convention in case of ties). We may then continue with vertices of the next levels by an iteration which should be obvious (but which would also be quite intricate to state explicitly). Figure 3 below may help visualizing the construction.

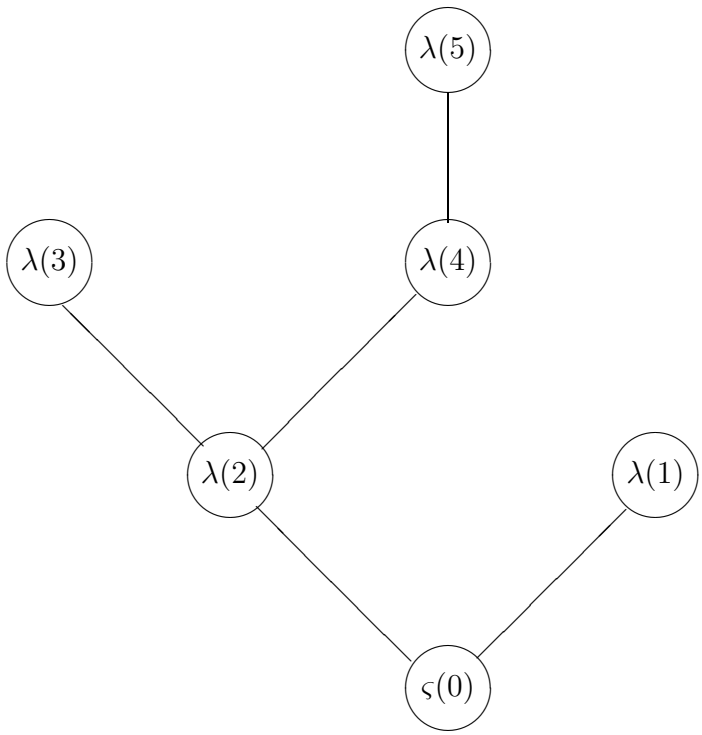

Figure 3: *Tree of alleles constructed from the random walk $S^{(\mathrm{c})}$ and the variables $\xi^{(\mathrm{m})}$ of Figure 2. The labels on the vertices are the lengths of the excursions of $S^{(\mathrm{c})}$ above its current minimum, they correspond to the sizes of the allelic sub-families (again sub-families with size 0 are omitted).*

3 Background on continuous state branching processes

Before describing our main limit results for trees of alleles, we need to develop some basic material about limits of rescaled Galton-Watson processes. The Lévy-Itô decomposition of subordinators plays a crucial role for the representation of the genealogical structure of the continuous state limits.

We start with the classical convergence to Feller diffusions [6, 11], i.e. the solutions $(X(x,t), t \geq 0)$ to stochastic differential equations of the type

$$X(x,t) = x + \int_0^t \sigma \sqrt{X(x,s)} \mathrm{d}B_s + b \int_0^t X(x,s) \mathrm{d}s \,, \qquad t \geq 0 \,, \tag{2}$$

where $x \geq 0$ is the initial value, $b \in \mathbb{R}$ and $\sigma^2 > 0$ are parameters, and $(B_t : t \geq 0)$ denotes a standard Brownian motion. For every $n \in \mathbb{N}$, consider a Galton-Watson process $(Z_k^{(n)} : k \in \mathbb{Z}_+)$ which starts from $Z_0^{(n)} = a(n)$ ancestors and has reproduction law $\rho^{(n)}$, where $\rho^{(n)}$ is some probability measure on $\mathbb{Z}_+$ and $a(n)$ a positive integer. Write

$$\mathrm{m}(\rho^{(n)}) := \sum_{i=0}^{\infty} i \rho_i^{(n)} \quad \text{and } \mathrm{var}(\rho^{(n)}) := \sum_{i=0}^{\infty} (i - \mathrm{m}(\rho^{(n)}))^2 \rho_i^{(n)}$$

for the first moment and the variance of $\rho^{(n)}$. In the situation where

$$a(n) \sim nx \ , \ \mathrm{m}(\rho^{(n)}) - 1 \sim bn^{-1} \ \text{ and } \mathrm{var}(\rho^{(n)}) \sim \sigma^2 \qquad \text{as } n \to \infty \tag{3}$$

for some $x \in (0, \infty)$, $b \in \mathbb{R}$ and $\sigma^2 > 0$, it is well known that

$$(n^{-1} Z_{\lfloor nt \rfloor}^{(n)} : t \geq 0) \implies (X(x,t) : t \geq 0) \tag{4}$$

where the notation $\Rightarrow$ refers to convergence in distribution as $n \to \infty$ and $X(x,t)$ is the Feller diffusion specified by (2).

We next turn our attention to the simpler situation where one only rescales the number of individuals and uses the generations as a discrete time parameter. For the sake of clarity, we shall deal with a framework that is slightly less general than it could be. We denote the tail distribution of $\rho^{(n)}$ by $\bar{\rho}^{(n)}(y) := \rho^{(n)}((y, \infty))$ for $y > 0$ and now assume that

$$\lim_{n \to \infty} n^{-1} a(n) = x \ \text{ and } \ \lim_{n \to \infty} n \bar{\rho}^{(n)}(ny) = \bar{\nu}(y) \quad \text{in } L^1_{\mathrm{loc}}([0, \infty), \mathrm{d}y) \,, \tag{5}$$

where $\bar{\nu}$ is some locally integrable non-increasing function on $[0, \infty)$ with $\bar{\nu}(\infty) = 0$. We may thus think of $\bar{\nu}$ as the tail of a Radon measure ν on $(0, \infty)$ with $\int (1 \wedge y) \nu(\mathrm{d}y) < \infty$; ν will be often referred to as a Lévy measure. Our assumptions ensure that

$$n^{-1} Z_1^{(n)} \implies Z_1 \,, \tag{6}$$

where Z_1 is a random variable with values in $[0, \infty)$ which is infinitely divisible. Indeed, we

have for any $q > 0$

$$\begin{aligned}
\mathbb{E}(\exp(-qn^{-1}Z_1^{(n)})) &= \left(1 - \int_{[0,\infty)} (1 - \mathrm{e}^{-qy/n})\rho_n(\mathrm{d}y)\right)^{a(n)} \\
&= \left(1 - \frac{q}{n}\int_0^\infty \mathrm{e}^{-qy/n}\bar{\rho}_n(y)\mathrm{d}y\right)^{a(n)} \\
&= \left(1 - q\int_0^\infty \mathrm{e}^{-qy}\bar{\rho}_n(ny)\mathrm{d}y\right)^{a(n)} ,
\end{aligned}$$

and (5) ensures that the latter quantity converges as $n \to \infty$ towards the Laplace transform of an infinitely divisible variable

$$\mathbb{E}(\exp(-qZ_1)) = \exp(-x\kappa(q)) ,$$

where the cumulant κ is given by the Lévy-Khintchine formula

$$\kappa(q) = \int_{(0,\infty)} (1 - \mathrm{e}^{-qy})\nu(\mathrm{d}y) . \tag{7}$$

We underline the fact that the drift coefficient is 0; this will play an important role in the sequel. An application of the Markov property now shows that more generally

$$(n^{-1}Z_k^{(n)} : k \in \mathbb{Z}_+) \implies (Z_k : k \in \mathbb{Z}_+) \tag{8}$$

where $(Z_k : k \in \mathbb{Z}_+)$ is a Markov chain with values in $\mathbb{R}_+$, started from $Z_0 = x$ and whose transition probabilities are characterized as follows: for every $k \in \mathbb{Z}_+$ and $q, y \geq 0$,

$$\mathbb{E}(\mathrm{e}^{-qZ_{k+1}} \mid Z_k = y) = \exp(-y\kappa(q)) .$$

One refers to the limiting chain Z as a (discrete time) continuous state branching process, in short CSBP, with reproduction measure ν and started from x.

It is interesting to recast the preceding convergence in the framework of the law of rare events. In this direction, recall that the Lévy-Itô decomposition of the infinitely divisible variable Z_1 reads

$$Z_1 = \sum_{i=1}^{\infty} \mathtt{a}_i , \tag{9}$$

where $\mathtt{a}_1 \geq \mathtt{a}_2 \geq \ldots$ are the atoms ranked in the decreasing order of a Poisson random measure on $(0, \infty)$ with intensity $x\nu$, with the convention that atoms are repeated according to their multiplicity and that when the Poisson random measure is finite (which occurs if and only if $\nu((0, \infty)) < \infty$), then $\mathtt{a}_i = 0$ whenever the index i exceeds the total mass of the Poisson measure. Consider for every $n \in \mathbb{N}$ a family $(\xi_i^{(n)} : 1 \leq i \leq a(n))$ of i.i.d. variables with law $\rho^{(n)}$; we should think of $\xi_i^{(n)}$ as the number of children of the i-th ancestor in the Galton-Watson process $Z^{(n)}$. Denote by $\mathtt{a}_1^{(n)} \geq \mathtt{a}_2^{(n)} \geq \ldots \geq \mathtt{a}_{a(n)}^{(n)}$ the decreasing reordering of the rescaled

variables $(n^{-1}\xi_i^{(n)} : 1 \leq i \leq a(n))$. In the regime (5), the law of rare events for null arrays (e.g. Theorem 14.18 in [12]) ensures that

$$(\mathsf{a}_1^{(n)}, \mathsf{a}_2^{(n)}, \ldots, \mathsf{a}_{a(n)}^{(n)}) \Longrightarrow (\mathsf{a}_1, \mathsf{a}_2, \ldots), \tag{10}$$

in the sense of finite dimensional distributions. Note also that (6) can be re-written in this setting as

$$\sum_{i=1}^{a(n)} \mathsf{a}_i^{(n)} \Longrightarrow \sum_{i\in\mathbb{N}} \mathsf{a}_i\,;$$

however the latter does not follow from (10).

This invites us to describe the convergence of rescaled Galton-Watson processes to (discrete time) CSBP from another point of view that takes into account the genealogy, and not merely the total sizes of populations at given generations. In this direction, we use a representation of the latter as random processes indexed by the universal tree $\mathbb{U}$. For simplicity, suppose for a while that the Lévy measure ν is infinite, so a Poisson random measure with intensity $c\nu$ with $c > 0$ has infinitely many atoms a.s. Recall from the Lévy-Itô decomposition (9) that almost all the individuals at the first generation in a CSBP descend from only countably many ancestors (we stress again that we are dealing with cumulants κ with no drift component), and plainly the same feature holds for the subsequent generations. Roughly speaking, vertices $u \in \mathbb{U}$ at level $|u| = k \geq 1$ represent the sizes of the sub-populations at generation k in the CSBP which descent from the same parent at generation $k-1$. We construct a random process $(\mathcal{Z}_u : u \in \mathbb{U})$ related to the CSBP Z, where $\mathcal{Z}_{uj}$ is the size of the j-th largest sub-population at generation $|u|+1$ which descents from a parent in the sub-population represented by u. We stress the process $\mathcal{Z}$ is by definition non-increasing on each sibling, i.e. the map $j \to \mathcal{Z}_{uj}$ is non-increasing on $\mathbb{N}$ for every $u \in \mathbb{U}$. More precisely, conditionally on $\mathcal{Z}_u = z$, the Lévy-Itô decomposition (9) suggests that $(\mathcal{Z}_{uj} : j \in \mathbb{N})$ should be given by the sequence of the atoms of a Poisson random measure on $(0, \infty)$ with intensity $z\nu$, where atoms are repeated according to their multiplicity and ranked in the decreasing order. We make the construction formal in the following definition.

Definition 1 *Fix $x > 0$ and ν a measure on $(0, \infty)$ with $\int (1 \wedge y)\nu(\mathrm{d}y) < \infty$. A tree-indexed CSBP with reproduction measure ν and initial population of size x is a process $(\mathcal{Z}_u : u \in \mathbb{U})$ with values in $\mathbb{R}_+$ and indexed by the universal tree, whose distribution is characterized by induction on the levels as follows:*

(i) *$\mathcal{Z}_\varnothing = x$ a.s.;*

(ii) *for every $k \in \mathbb{Z}_+$, conditionally on $(\mathcal{Z}_v : v \in \mathbb{U}, |v| \leq k)$, the sequences $(\mathcal{Z}_{uj})_{j\in\mathbb{N}}$ for the vertices $u \in \mathbb{U}$ at generation $|u| = k$ are independent, and each sequence $(\mathcal{Z}_{uj})_{j\in\mathbb{N}}$ is distributed as the family of the atoms of a Poisson random measure on $(0, \infty)$ with intensity $\mathcal{Z}_u\nu$, where atoms are repeated according to their multiplicity, ranked in the decreasing order, and completed by an infinite sequence of 0 if the Poisson measure is finite.*

It should be plain that if $\mathcal{Z}$ is a tree-indexed CSBP with reproduction measure ν and initial population of size x, then $\left(\sum_{|u|=k} \mathcal{Z}_u : k \in \mathbb{Z}_+\right)$ is a CSBP with reproduction measure ν started

from x. We also point out that for every integer n, we can represent similarly the genealogy for the Galton-Watson process $Z^{(n)}$ as a process $\mathcal{Z}^{(n)}$ indexed by the universal tree $\mathbb{U}$, and one can check that under the regime (5), the following extension of (8) holds:

$$n^{-1}\mathcal{Z}^{(n)} \implies \mathcal{Z}$$

in the sense of finite dimensional distributions. This should be viewed as a variation of the law of rare events (10); the easy proof is left to the interested reader.

We now conclude this section by underlying the connexion between discrete time CSBP and subordinators (i.e. Lévy processes with values in $\mathbb{R}_+$). Consider a subordinator $\tau = (\tau_t : t \geq 0)$ with no drift and Lévy measure ν. Its cumulant κ is given by the Lévy-Khintchine formula (7) and we have

$$\mathbb{E}(\mathrm{e}^{-q\tau_t}) = \exp(-t\kappa(q)) \,, \qquad \text{for all } q, t \geq 0 \,.$$

Fix $x > 0$ and define a sequence $(\zeta_k : k \in \mathbb{Z}_+)$ by implicit iteration as follows:

$$\zeta_0 = x \,,\ \zeta_1 = \tau_x \,,\ \zeta_1 + \zeta_2 = \tau_{x+\zeta_1} \,,\ \ldots \,,\ \zeta_1 + \cdots + \zeta_{k+1} = \tau_{x+\zeta_1+\cdots+\zeta_k} \,.$$

Observe by an easy induction that the random times $x + \zeta_1 + \cdots + \zeta_k$ are stopping times in the natural filtration of τ, so that the strong Markov property can be applied. It is then immediate to check that $(\zeta_k : k \in \mathbb{Z}_+)$ is a CSBP with reproduction measure ν and initial population of size x.

More generally, the tree-indexed CSBP $\mathcal{Z}$ can be constructed from the subordinator τ by making full use of the Lévy-Itô decomposition. Specifically, we know from the latter that the Stieltjes measure $d\tau$ on the random interval

$$I_k := (x + \zeta_1 + \cdots + \zeta_{k-1}, x + \zeta_1 + \cdots + \zeta_k]$$

is purely atomic, and conditionally on $|I_k|$, the sequence of the atomic masses has the same distribution as the family of the atoms in a Poisson point measure on $(0, \infty)$ with intensity $|I_k|\nu$. These atoms should be viewed as the sizes of sub-families at level k, so it remains to identify the siblings and rank atoms corresponding to a same sibling in the decreasing order. This is straightforward for the first levels but becomes increasingly intricate for larger levels. Specifically, we let $\mathcal{Z}_\varnothing = x$ and declare that $(\mathcal{Z}_j : j \in \mathbb{N})$ is given by the sequence of the jumps of τ on $(0, x]$ ranked in the decreasing order. Next $(\mathcal{Z}_{1j} : j \in \mathbb{N})$ corresponds to the ranked sequence of the jumps of τ on the interval $(x, x + \tau_{\mathcal{Z}_1}]$, $(\mathcal{Z}_{2j} : j \in \mathbb{N})$ to those on the interval $(x + \tau_{\mathcal{Z}_1}, x + \tau_{\mathcal{Z}_1+\mathcal{Z}_2}]$ and so on. The algorithm may be thought of as a variant of the breadth first search in which each sibling is ordered according to the size of its progeny.

4 Asymptotic for rare mutations

This section contains our main results on limits of trees of alleles; we shall first present and discuss the general framework, then state the results, and finally prove the latter.

4.1 Framework and main results

We consider a fixed probability measure $\pi^{(+)}$ on $\mathbb{Z}_+$ which serves as reproduction law for a standard Galton-Watson process denoted by $Z^{(+)}$. We assume that $Z^{(+)}$ is critical, i.e.

$$\sum_{i=0}^{\infty} i\pi_i^{(+)} = 1\,,$$

and has a finite variance

$$\sum_{i=0}^{\infty} (i-1)^2\pi_i^{(+)} = \sigma^2 < \infty\,.$$

Further, we suppose that mutations affect each child according to a fixed probability $p \in (0,1)$ and independently of the other children. That is to say that the probability measure π on $\mathbb{Z}_+ \times \mathbb{Z}_+$ which gives the law of the number of clone children and the number of mutant children of a typical individual is given by

$$\pi_{k,\ell} = \pi_{k+\ell}^{(+)} \binom{k+\ell}{k} (1-p)^k p^\ell\,.$$

We will use the notation $\mathbb{P}_a^p$ for the probability measure under which the Galton-Watson process $Z^{(+)}$ has a ancestors and the mutation rate is p, and $\mathcal{L}\left(\cdot, \mathbb{P}_a^p\right)$ will then refer to the distribution of a random variable or a process under $\mathbb{P}_a^p$

We are interested in the situation where the mutation rate $p = p(n)$ is small and the number of ancestors $a = a(n)$ large when the parameter n goes to infinity. Specifically, we consider the regime

$$a(n) \sim nx \quad \text{and} \quad p(n) \sim cn^{-1}\,, \tag{11}$$

where c, x are some positive constants. Let us start by mentioning some results of convergence in distribution for Galton-Watson processes in this setting.

First, we know from (4) that the Galton-Watson process $Z^{(+)}$ properly rescaled converges to a Feller diffusion on $\mathbb{R}_+$; specifically

$$\mathcal{L}\left((n^{-1}Z_{\lfloor nt\rfloor}^{(+)} : t \geq 0), \mathbb{P}_{a(n)}^{p(n)}\right) \implies (X_t^{(+)} : t \geq 0)\,, \tag{12}$$

where $(X_t^{(+)} : t \geq 0)$ solves the SDE (2) for the parameter $b = 0$. In the same direction, the marginal law of $\xi^{(\mathrm{c})}$ under $\mathbb{P}^{p(n)}$ has first moment $1 - p(n)$ and variance close to σ^2 when n is large. Hence, if $Z^{(\mathrm{c})}$ denotes the Galton-Watson process of clones (i.e. we only consider individuals of the 0-th type), then

$$\mathcal{L}\left((n^{-1}Z_{\lfloor nt\rfloor}^{(\mathrm{c})} : t \geq 0), \mathbb{P}_{a(n)}^{p(n)}\right) \implies (X_t^{(\mathrm{c})} : t \geq 0) \qquad \text{as } n \to \infty, \tag{13}$$

where $(X_t^{(\mathrm{c})} : t \geq 0)$ is another Feller diffusion solution to the SDE (2) for the parameter $b = -c$.

On the other hand, recall from Lemma 1 and Corollary 1(i) that the process of the number of mutants of given types $(M_k : k \in \mathbb{Z}_+)$ is a critical Galton-Watson process with finite variance. In view of the classical limit theorem stated as (4) in Section 3, one might suspect that the

rescaled process $(n^{-1}M_{\lfloor nt\rfloor} : t \geq 0)$ could converge to some Feller diffusion. However this is not the case; indeed an easy calculation shows that the variance of the reproduction law of $M_\cdot$ under $\mathbb{P}^{p(n)}$ is of order n, and thus the requirement (3) fails. Nonetheless one can deduce from a few lines of calculations based on Proposition 1 that the condition (5) is fulfilled by the reproduction law of M under $\mathbb{P}^{(n)}_{a(n)}$, and hence

$$\mathcal{L}\left((n^{-1}M_k) : k \in \mathbb{Z}_+), \mathbb{P}^{p(n)}_{a(n)}\right)$$

converges weakly when $n \to \infty$ towards the law of some discrete time CSBP started from a. We do not give a formal statement as the forthcoming Proposition 2 is a stronger result.

The asymptotics (12) and (13) point to the fact that in the regime (11), the total size of the population of the Galton-Watson process should be rescaled by a factor n^{-2}, and in particular the asymptotic behavior of the number $T_0 = \sum_{k=0}^{\infty} Z_k^{(c)}$ of individuals of 0-th type is given by

$$\mathcal{L}\left(n^{-2}T_0, \mathbb{P}^{p(n)}_{a(n)}\right) \implies \int_0^\infty X_t^{(c)} \mathrm{d}t\,.$$

More generally, we have the following joint convergence in distribution for the rescaled process of the sizes of sub-populations and the number of mutants of a given type.

Proposition 2 *In the regime* (11)*, we have*

$$\mathcal{L}\left(((n^{-2}T_k, n^{-1}M_{k+1}) : k \in \mathbb{Z}_+), \mathbb{P}^{p(n)}_{a(n)}\right) \implies ((Z_{k+1}, cZ_{k+1}) : k \in \mathbb{Z}_+)$$

where $(Z_k : k \in \mathbb{Z}_+)$ *is a CSBP with reproduction measure*

$$\nu(\mathrm{d}y) = \frac{c}{\sqrt{2\pi\sigma^2 y^3}} \exp\left(-\frac{c^2 y}{2\sigma^2}\right) \mathrm{d}y\,, \qquad y > 0\,,$$

and initial population of size x/c.

The Lévy-Itô decomposition now suggests that conditionally on $n^{-2}T_k \sim y$, the sequence of the sizes of the sub-populations carrying a same allele of the $(k+1)$-type and normalized by a factor n^{-2} should converge in distribution to the sequence of the atoms of a Poisson random measure on $\mathbb{R}_+$ with intensity specified in Proposition 2. Recall also that d_u denotes the outer degree at the vertex $u \in \mathbb{U}$ in the tree of alleles, and observe from Lemma 2 that for a Galton-Watson process with neutral mutations, the process $((\mathcal{A}_u, d_u) : u \in \mathbb{U})$ has a simpler Markovian structure than $(\mathcal{A}_u : u \in \mathbb{U})$ alone. This leads us to our main limit theorem for the tree of alleles.

Theorem 1 *In the regime* (11)*, the rescaled tree of alleles* $n^{-2}\mathcal{A}$ *under* $\mathbb{P}^{p(n)}_{a(n)}$ *converges in the sense of finite dimensional distributions to the tree indexed CSBP* $(\mathcal{Z}_u : u \in \mathbb{U})$ *with reproduction measure* ν *given in Proposition 2 and random initial population with inverse Gaussian distribution:*

$$\frac{\mathbb{P}(\mathcal{Z}_\varnothing \in \mathrm{d}y)}{\mathrm{d}y} = \frac{x}{\sqrt{2\pi\sigma^2 y^3}} \exp\left(-\frac{(cy-x)^2}{2\sigma^2 y}\right)\,, \qquad y > 0\,.$$

More precisely, if we also take into account the outer degrees, then we have the joint convergence in the sense of finite dimensional distributions:

$$\mathcal{L}\left(((n^{-2}\mathcal{A}_u, n^{-1}d_u) : u \in \mathbb{U}), \mathbb{P}^{p(n)}_{a(n)}\right) \implies (\mathcal{Z}_u, c\mathcal{Z}_u) : u \in \mathbb{U}) \ .$$

4.2 Proofs

Let us first present informally some intuitions for the proofs, which rely on the connexion with random walks in Section 2.3. Roughly speaking, we shall observe that in the regime (11), the random walk $S^{(c)}$ suitably rescaled converges to a Brownian motion with negative drift. As the lengths of the excursions of $S^{(c)}$ above its current minimum correspond to the sizes of sub-populations with the same allele, this suggests that in the limit, the lengths of the excursions of a Brownian motion with drift above its current minimum should describe the limit of rescaled sub-populations. According to Itô's excursion theory, these lengths can be described in terms of a Poisson point process. The comparison with the construction of the tree indexed CSBP presented in Section 3.2 should then make Theorem 1 more intuitive.

The proofs of Proposition 2 and Theorem 1 both rely on the following technical lemma.

Lemma 4 *In the regime* (11)*, we have:*

(i) *Let* $(\tau_x : x \geq 0)$ *be a inverse Gaussian subordinator with cumulant*

$$\kappa(q) = \sigma^{-2}\left(\sqrt{c^2 + 2q\sigma^2} - c\right) = c^{-1}\int_0^\infty (1 - \mathrm{e}^{-qy})\nu(\mathrm{d}y)\,, \qquad q \geq 0\,,$$

i.e. with zero drift and Lévy measure $c^{-1}\nu$ *where* ν *given in Proposition 2. Then*

$$\mathcal{L}\left((n^{-2}T_0, n^{-1}M_1), \mathbb{P}^{p(n)}_{a(n)}\right) \implies (\tau_x, c\tau_x)\,.$$

(ii) *The behavior of the joint tail distribution of* T_0 *and* M_1 *under* $\mathbb{P}^{p(n)}_1$ *is given by*

$$\lim_{n\to\infty} n\mathbb{P}^{p(n)}_1\left(n^{-2}T_0 > t \text{ or } n^{-1}M_1 > m\right) = c^{-1}\bar{\nu}(\min(t, m/c)) \qquad \text{in } L^1_{\mathrm{loc}}(\mathbb{R}_+ \times \mathbb{R}_+, \mathrm{d}t\,\mathrm{d}m)\,,$$

where $\bar{\nu}$ *denotes the tail function of the Lévy measure* ν*.*

Proof: One could establish these limits from the explicit expressions in Proposition 1; however a probabilistic argument based on the construction in Section 2.3 circumvents the somewhat tedious calculations.

(i) Recall that the fixed reproduction law $\pi^{(+)}$ has unit mean and variance σ^2. For each $n \in \mathbb{N}$, consider a random walk $(S^{(n)}_k : k \in \mathbb{Z}_+)$ started from $S^{(n)}_0 = a(n)$ and with step distribution that of $\xi^{(+)} - 1$. By Donsker's invariance principle and Skorohod's representation, we may suppose that with probability one

$$\lim_{n\to\infty} n^{-1}S^{(n)}_{\lfloor n^2 t\rfloor} = x + \sigma B_t\,,$$

where $(B_t : t \geq 0)$ is a standard Brownian motion and the convergence holds uniformly on every compact time-interval.

For every fixed n, we now decompose each variable $\xi_i^{(+)}$ as the sum $\xi_i^{(+)} = \xi_i^{(\mathrm{c}n)} + \xi_i^{(\mathrm{m}n)}$ by using a Bernoulli sampling; that is conditionally on $\xi_i^{(+)} = \ell$, $\xi_i^{(\mathrm{m}n)}$ has the binomial distribution with parameter $(\ell, p(n))$. Of course, we use independent Bernoulli sampling for the different indices i, so that the pairs $(\xi_i^{(\mathrm{c}n)}, \xi_i^{(\mathrm{m}n)})$ are i.i.d. and have the law of ξ under $\mathbb{P}^{p(n)}$. If we define

$$S_k^{(\mathrm{m}n)} := \xi_1^{(\mathrm{m}n)} + \cdots + \xi_k^{(\mathrm{m}n)}, \qquad k \in \mathbb{Z}_+,$$

then $\mathbb{E}(\xi_1^{(\mathrm{m}n)}) = p(n) \sim c/n$ and $\mathrm{var}(\xi_1^{(\mathrm{m}n)}) = O(1/n)$, and it is easy to verify that with probability one

$$\lim_{n\to\infty} n^{-1} S_{\lfloor n^2 t\rfloor}^{(\mathrm{m}n)} = ct,$$

uniformly on every compact time-interval. Hence the random walk

$$S_k^{(\mathrm{c}n)} := a(n) + \xi_1^{(\mathrm{c}n)} + \cdots + \xi_k^{(\mathrm{c}n)} = S_k^{(n)} - S_k^{(\mathrm{m}n)}$$

fulfills

$$\lim_{n\to\infty} n^{-1} S_{\lfloor n^2 t\rfloor}^{(\mathrm{c}n)} = x + \sigma B_t - ct,$$

where again the convergence holds a.s., uniformly on every compact time-interval.

Now recall the framework of Section 2.3 and introduce

$$\varsigma^{(n)}(0) := \inf\{k \in \mathbb{Z}_+ : S_k^{(\mathrm{c}n)} = 0\} \quad \text{and} \quad \Sigma^{(n)}(0) := \sum_{i=1}^{\varsigma^{(n)}(0)} \xi_i^{(\mathrm{m}n)} = S_{\varsigma^{(n)}(0)}^{(\mathrm{m}n)}.$$

It follows readily from the preceding observations that with probability one

$$\lim_{n\to\infty} n^{-2}\varsigma^{(n)}(0) = \tau_x \quad \text{and} \lim_{n\to\infty} n^{-1}\Sigma^{(n)}(0) = c\tau_x$$

where τ denotes the process of first passage times for a Brownian motion with drift,

$$\tau_y := \inf\{t \geq 0 : ct - \sigma B_t > y\}, \qquad y \geq 0.$$

It is well-known that the latter is a subordinator with cumulant κ as given in the statement, and the first claim is established by an appeal to Lemma 3.

(ii) The branching property shows that the law of (T_0, M_1) under $\mathbb{P}_{a(n)}^{p(n)}$ is that of the sum of $a(n)$ i.i.d. variables distributed as (T_0, M_1) under $\mathbb{P}_1^{p(n)}$. This observation enables to deduce (ii) from (i) by an argument similar to that we use to establish (6). Indeed, write

$$\bar{\mu}_n(t, m) = \mathbb{P}_1^{p(n)}\left(T_0 > t \text{ or } M_1 > m\right)$$

for the bivariate tail distribution of the pair (T_0, M_1) under $\mathbb{P}_1^{p(n)}$. By an elementary calculation, we have that for every $q, r > 0$

$$\begin{aligned} &\mathbb{E}_{a_n}^{p(n)}\left(\exp\left(-\frac{q}{n^2}T_0 - \frac{r}{n}M_1\right)\right) \\ = &\left(1 - qr\int_0^\infty\int_0^\infty \mathrm{e}^{-qt}\mathrm{e}^{-rm}\bar{\mu}_n(n^2t, nm)\mathrm{d}t\,\mathrm{d}m\right)^{a_n}. \end{aligned}$$

We know from (i) that this quantity converges as $n \to \infty$ towards

$$\mathbb{E}(\exp(-(q+cr)\tau_x)) = \exp\left(-x\int_0^\infty (1-\mathrm{e}^{-(q+cr)y})c^{-1}\nu(\mathrm{d}y)\right),$$

so that taking logarithms, we get

$$\begin{aligned}
&\lim_{n\to\infty} qra_n \int_0^\infty\int_0^\infty \mathrm{e}^{-qt}\mathrm{e}^{-rm}\bar{\mu}_n(n^2t, nm)\mathrm{d}t\,\mathrm{d}m \\
= \;& x\int_0^\infty (1-\mathrm{e}^{-(q+cr)y})c^{-1}\nu(\mathrm{d}y) \\
= \;& xqr\int_0^\infty\int_0^\infty \mathrm{e}^{-qt}\mathrm{e}^{-rm}c^{-1}\bar{\nu}(\min(t, m/c))\mathrm{d}t\,\mathrm{d}m\,.
\end{aligned}$$

This entails our claim. □

Proposition 2 immediately follows from Lemma 1 and Lemma 4(i), so we turn our attention to the proof of Theorem 1.

Proof of Theorem 1: Recall that $\mathcal{A}_\varnothing = T_0$ and $d_\varnothing = M_1$. On the one hand, we know from Lemma 4(i) that

$$\mathcal{L}\left((n^{-2}\mathcal{A}_\varnothing, n^{-1}d_\varnothing), \mathbb{P}^{p(n)}_{a(n)}\right) \implies (\tau_x, c\tau_x)\,.$$

On the other hand, Lemma 4(ii) and the law of rare events for null arrays (e.g. Theorem 14.18 in [12]) entails that for any sequence of integers $b(n)$ such that $b(n) \sim bn$ for some $b > 0$,

$$\mathcal{L}\left((n^{-2}T_0, n^{-1}M_1)^{(b(n)\downarrow)}, \mathbb{P}^{p(n)}_1\right) \implies ((\mathtt{a}_1, c\mathtt{a}_1), (\mathtt{a}_2, c\mathtt{a}_2), \ldots)\,,$$

where the notation $\gamma^{(d\downarrow)}$ has been defined just before Lemma 2 and $(\mathtt{a}_1, \mathtt{a}_2, \ldots)$ stands for the sequence ranked in the decreasing order of the atoms of a Poisson measure on $(0, \infty)$ with intensity $bc^{-1}\nu$.

Denote by $\mathbb{Q}_x$ the law of a tree-indexed CSBP and initial population distributed as τ_x and reproduction measure ν. We now see from Lemma 2 that

$$\mathcal{L}\left(((n^{-2}\mathcal{A}_u, n^{-1}d_u) : |u| \le 1), \mathbb{P}^{p(n)}_{a(n)}\right) \implies \mathcal{L}\left(((\mathcal{A}_u, c\mathcal{A}_u) : |u| \le 1), \mathbb{Q}_x\right)$$

in the sense of finite dimensional convergence. Lemma 2 enables us to iterate the argument to the subsequent levels of vertices, which establishes our claim. □

Acknowledgments. The question of describing the asymptotic shape of the tree of alleles for large populations with rare mutations was raised by Matthias Winkel during a lecture based on [2] that I delivered at the University of Oxford. I would like to thank Matthias for having stimulated this work. This work has been supported by ANR-08-BLAN-0220-01.

References

[1] Athreya, K.B. and Ney, P.E. *Branching processes*. Springer-Verlag, Berlin, 1972.

[2] Bertoin, J. The structure of the allelic partition of the total population for Galton-Watson processes with neutral mutations. To appear in *Ann. Probab.*

[3] Chauvin, B. Sur la propriété de branchement. *Ann. Inst. Henri Poincaré B* **22** (1986), 233-236.

[4] Dwass, M. The total progeny in a branching process. *J. Appl. Probab.* **6** (1969), 682-686.

[5] Ewens, W. J. The sampling theory of selectively neutral alleles, *Theoret. Popul. Biol.* **3** (1972), 87-112.

[6] Feller, W. Diffusion processes in genetics. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 227-246. Univ. California Press, Berkeley, Calif., 1951.

[7] Harris, Th. E. *The theory of branching process.* Springer-Verlag, Berlin, 1963.

[8] Itô, K. *Stochastic processes.* Lectures given at Aarhus University. Springer-Verlag, Berlin, 2004.

[9] Itô, K. Poisson point processes attached to Markov processes. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. III, pp. 225-239. Univ. California Press, Berkeley, Calif., 1972. Available via: `http://projecteuclid.org/euclid.bsmsp/1200514340`

[10] Jagers, P. *Branching processes with biological applications.* John Wiley & Sons, 1975.

[11] Jiřina, M. Stochastic branching processes with continuous state space. *Czechoslovak Math. J.* **8** (1958), 292-313. Available via `http://dml.cz/dmlcz/100304`

[12] Kallenberg, O. *Foundations of modern probability.* Second edition. Probability and its Applications (New York). Springer-Verlag, New York, 2002.

[13] Kingman, J. F. C. The coalescent. *Stochastic Process. Appl.* **13** (1982), 235-248.

[14] Pitman, J. *Combinatorial stochastic processes. École d'été de Probabilités de St-Flour*, Lect. Notes in Maths **1875**, Springer-Verlag, Berlin, 2006. Available via :

`http://stat-www.berkeley.edu/users/pitman/`

[15] Taïb, Z. *Branching processes and neutral evolution.* Lecture Notes in Biomathematics **93**. Springer-Verlag, Berlin, 1992.

[16] Wilf, H. S. : *Generatingfunctionology.* Academic Press, 1994. Also available via : `http://www.math.upenn.edu/~wilf/gfology2.pdf`