



HAL
open science

The structure of the allelic partition of the total population for Galton-Watson processes with neutral mutations

Jean Bertoin

► **To cite this version:**

Jean Bertoin. The structure of the allelic partition of the total population for Galton-Watson processes with neutral mutations. 2007. hal-00191132v1

HAL Id: hal-00191132

<https://hal.science/hal-00191132v1>

Preprint submitted on 24 Nov 2007 (v1), last revised 28 Aug 2009 (v3)

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.

The structure of the allelic partition of the total population for Galton-Watson processes with 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 consider a (sub) critical Galton-Watson process with neutral mutations (infinite alleles model), and decompose the entire population into clusters of individuals carrying the same allele. We specify the law of this allelic partition in terms of the distribution of the number of clone-children and the number of mutant-children of a typical individual. The approach combines an extension of Harris representation of Galton-Watson processes and a version of the ballot theorem. Some limit theorems related to the distribution of the allelic partition are also given.

Key words. Branching process, infinite alleles model, allelic partition, ballot theorem.

A.M.S. Classification. 60 J 80, 60 J 10

e-mail. jbe@ccr.jussieu.fr

1 Introduction

We consider a Galton-Watson process, i.e. a population model with asexual reproduction such that at every generation, each individual gives birth to a random number of children according to a fixed distribution, and independently of the other individuals in the population. We are interested in the situation where a child can be either of the same type as its parent (i.e. a clone), or of a new type (i.e. a mutant). We stress that each mutant has a distinct type and in turn gives birth to clones of itself and to new mutants according to the same statistical law as its parent, even though it bears a different type. In other words, we are working with an infinite alleles model where mutations are neutral for the population dynamics. We might as well think of a spatial population model in which children either occupy the same location as their parents or migrate to new places and start growing colonies on their own. This quite basic framework has been often considered in the literature, see e.g. [4, 13, 22, 30, 33, 39]; we also

refer to [1, 5, 6, 27, 29, 36] for interesting variations (these references are of course far from being exhaustive). Note also that such Galton-Watson processes with mutations can be viewed as a special instance of multitype branching processes; see Chapter V in Athreya and Ney [7] or Chapter 7 in Kimmel and Axelrod [26].

We will always assume that the Galton-Watson process is critical or sub-critical, so it becomes eventually extinguished. We shall often focus on the case when it starts from a single ancestor, which will then be called *Eve*. We are interested in the partition of the entire population according to the types of individuals. Specifically, let us call the set of all individuals having the same type as Eve, the *Eve cluster*. Similarly we associate to each mutant the cluster of all individuals that bear the same type as that mutant. This way, the total population of the Galton-Watson process is split into clusters of different types, which will be referred to as the *allelic partition*. It is convenient to view Eve as a mutant of the zero-th kind, then to call mutant of the first kind a mutant-child of an individual of the Eve cluster, and the set of all its clones (including the initial mutant) a cluster of the first kind. By iteration, we define mutants and clusters of the k -th kind for any integer $k \geq 0$, where we agree that $k = 0$ corresponds to Eve.

Statistics of the allelic partition of a random population model with mutations have been first determined in a fundamental work of Ewens [19] for the Wright-Fisher model (more precisely this concerns the partition of the population at a fixed generation). Kingman [25] provided a deep analysis of this framework, in connexion with the celebrated coalescent process that depicts the genealogy of the Wright-Fisher model. We refer to [8, 9, 14, 15, 32] for some recent developments in this area which involve some related population models with fixed generational size and certain exchangeable coalescents. The main purpose of the present work is to describe explicitly the structure of the allelic partition of the entire population for (sub) critical Galton-Watson processes with neutral mutations.

In the framework of branching processes, the Eve cluster has obviously the genealogical structure of a Galton-Watson tree with reproduction law given by the distribution of the number of clone-children of a typical individual. Informally, the branching property indicates that the same holds for the other clusters of the allelic partition. Further, it should be intuitively clear that the process which counts the number of clusters of the k -th kind for $k \geq 0$ is again a Galton-Watson process whose reproduction law is given by the distribution of the number of mutants of the first kind; this phenomenon has already been pointed at in the work of Taïb [39]. That is to say that, in some loose sense, the allelic partition inherits branching structures from the initial Galton-Watson process. Of course, these formulations are only heuristic, and precise statements will be given later on. We also stress that the tree structure which connects clusters of different kinds and the genealogical structure on each cluster are not independent since, typically, the number of mutants of the first kind is statistically related to the size of the Eve cluster.

Our approach essentially relies on a variation of the well-known connection due to Harris [23, 24] between ordinary Galton-Watson processes and sequences of i.i.d. integer valued random variables. Specifically, we incorporate mutations in Harris representation by deciding to enumerate for each sibling, first the clone-children and then the mutant-children. At first sight, the mathematical content of this convention may look insipid; however it has some interesting consequences which might worth being recorded. By combining with the celebrated ballot the-

orem (which is another classical tool in this area as it is amplified e.g. by Pitman; see Chapter 6 in [35]), we obtain expressions for the joint distribution of various natural variables (size of the total population, number of types, size and number of mutant-children of an allelic cluster, ...) in terms of the transition probabilities of the two-dimensional random walk which is generated by the numbers of clone-children and of mutant-children of a typical individual.

We also investigate some limit theorems in law; typically we show that when the numbers of clone-children and of mutant-children of an individual are independent, the sequence of the relative sizes of the allelic clusters has a limiting conditional distribution when the size of the total population and the number of types both tend to infinity according to some appropriate regime. The limiting distribution appears in the study of the standard additive coalescent by Aldous and Pitman [5]. We also point at limit theorems for allelic partitions of Galton-Watson forests (i.e. when there are infinitely many ancestors), where, following Duquesne and Le Gall [16, 17], the limits are described in terms of certain Lévy trees. In particular, this provides an explanation to a striking identity between two self-similar fragmentation processes that were defined on the one-hand by logging the Continuum Random Tree according to a Poisson point process along its skeleton [5], and on the other hand, by splitting the unit-interval at instants when the standard Brownian excursion with a negative drift reaches new infima [10].

2 Allelic partitions for deterministic genealogical trees

The purpose of this section is to introduce the main notions of interest and tools in the deterministic setting.

A population model of total size $T \in \mathbb{N}$ with asexual reproduction and a single ancestor is represented by a rooted tree on a set of T vertices, a combinatorial structure which can be described as follows. One of the T vertices is distinguished and called the root (it represents the ancestor) and the other $T - 1$ vertices are indistinguishable (they correspond to the total progeny of the ancestor). There are $T - 1$ edges between pairs of distinct vertices, which indicate the direct parental links between two individuals. The resulting graph on the set of T vertices is connected; plainly, since there are $T - 1$ edges, this graph contains no loops.

Now if we distinguish and order the progeny of each individual, then we can view the population model as a *planar rooted tree*, i.e. the graph that represents the genealogy of the population can be drawn in the plane. More precisely, an individual at generation k is specified by a finite sequence of integers (i_1, \dots, i_k) , the empty sequence \emptyset corresponding to the ancestor at generation 0, and if this individual has $j \geq 1$ children at generation $k + 1$, these children are represented by the sequences $(i_1, \dots, i_k, 1), \dots, (i_1, \dots, i_k, j)$. This enables us to enumerate the individuals in the lexicographic order, which is the well-known *depth-first-search algorithm*. As a consequence, the population model with a single ancestor and ordered siblings is entirely described by the sequence $(\xi_n, 1 \leq n \leq T)$ where T denotes the total size of the population and ξ_n the number of children of the n -th individual found by the depth-first-search algorithm; see Figure 1. Note the important identity

$$T = \min\{n \geq 1 : \xi_1 + \dots + \xi_n - n = -1\}; \quad (1)$$

indeed, on the one-hand, we have that $1 + \xi_1 + \dots + \xi_T = T$ by expressing the fact that the

entire population consists of Eve and its descendants, and on the other hand, we must have $1 + \xi_1 + \dots + \xi_n > n$ when $n < T$, since otherwise the depth-first-search algorithm would stop before having explored the entire population. In particular we always have $\xi_T = 0$, i.e. the ultimate individual has no progeny. Conversely, any finite sequence $(\xi_n, 1 \leq n \leq T)$ of nonnegative integers which fulfills (1) encodes a finite population model started from a single ancestor and ordered siblings, see e.g. Lemma 63 in [35] or [28].

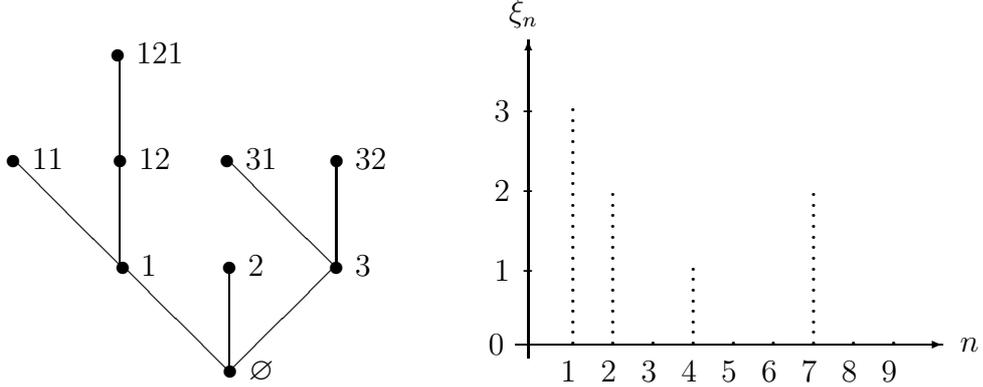


Figure 1 : Genealogical tree labeled by depth-first-search (left) and its representation (right)

In the presence of mutations, we say that a sibling is *well-ordered* if the clone-children are listed before the mutant-children. In other words, if (i_1, \dots, i_k) represents some individual which has $j^{(c)}$ clone-children and $j^{(m)}$ mutant-children, then its clone-children are represented by the sequences

$$(i_1, \dots, i_k, 1), \dots, (i_1, \dots, i_k, j^{(c)})$$

and its mutant-children by

$$(i_1, \dots, i_k, j^{(c)} + 1), \dots, (i_1, \dots, i_k, j^{(c)} + j^{(m)}).$$

We thus represent a finite population model with mutations, well-ordered siblings and started from a single ancestor by a finite sequence $((\xi_n^{(c)}, \xi_n^{(m)}), 1 \leq n \leq T)$ of pairs of nonnegative integers, where $\xi_n^{(c)}$ is the number of clone-children of the n -th individual and $\xi_n^{(m)}$ the number of its mutant-children. Of course, (1) now reads

$$T = \min \left\{ n \geq 1 : \sum_{k=1}^n (\xi_k^{(c)} + \xi_k^{(m)}) = n - 1 \right\}, \quad (2)$$

and more precisely (2) is the necessary and sufficient condition for a finite sequence $((\xi_n^{(c)}, \xi_n^{(m)}), 1 \leq n \leq T)$ of pairs of nonnegative integers to encode a finite population model with mutations, started from a single ancestor and with ordered siblings.

We denote by

$$\tau = 1 + \sum_{n=1}^T \xi_n^{(m)},$$

the number of different types in the total population. It is convenient to think of Eve as the first mutant (so that τ is also the number of mutants), and to write

$$1 = \mu_1 < \cdots < \mu_\tau$$

for the ranks of mutant individuals listed in the lexicographic order, i.e. the j -th mutant is found at the μ_j -th step by the depth-first-search algorithm. In this direction, it is also convenient to declare that $\mu_{\tau+1} = T + 1$, where T is the total size of the population. The upshot of the convention of ordering first the clone-children and then the mutant-children in each sibling is that the set of individuals that bear the same type as the j -th mutant corresponds precisely to the (integer) interval

$$[\mu_j, \mu_{j+1}[:= \{\mu_j, \mu_j + 1, \dots, \mu_{j+1} - 1\}.$$

We now record these notions (see also Figure 2 below for an illustration).

Definition 1 Consider $((\xi_n^{(c)}, \xi_n^{(m)}), 1 \leq n \leq T)$, a finite population model with mutations, well-ordered siblings, and started from a single ancestor. Let

$$\tau = 1 + \sum_{n=1}^T \xi_n^{(m)}$$

denote the number of types in the entire population, and, viewing Eve as the first mutant, let $1 = \mu_1 < \cdots < \mu_\tau$ be the ranks of mutants in the depth-first-search. We further make the convention that $\mu_{\tau+1} = T + 1$.

(i) For every $1 \leq j \leq \tau$, define the j -th allelic cluster by

$$C_j := ((\xi_{\mu_j+\ell-1}^{(c)}, \xi_{\mu_j+\ell-1}^{(m)}) : 1 \leq \ell \leq \mu_{j+1} - \mu_j).$$

The sequence $(C_j)_{1 \leq j \leq \tau}$ is called the allelic partition of the population model.

(ii) The set of mutants is naturally endowed with a structure of planar rooted tree, which we call the genealogical tree of mutants. Specifically, $\{1, \dots, \tau\}$ form the set the vertices, where the vertex i corresponds to the i -th mutant found by the depth-first search algorithm, and of course the distinguished vertex 1 corresponds to Eve and serves as the root. There is an edge connecting a pair of vertices (i, j) with $1 \leq i < j \leq \tau$ if and only if the parent of the j -th mutant μ_j is an individual in the i -th cluster C_i .

(iii) We enrich the structure (ii) by assigning a mark to each vertex; more precisely the mark attached to the vertex $i \in \{1, \dots, \tau\}$ is the size $|C_i|$ of the cluster C_i (i.e. the number of individuals bearing the same type as the i -th mutant). The genealogical tree of mutants endowed with such marks is referred to as the allelic tree.

Remarks : 1. Each allelic cluster C_j also inherits from the population model a structure of rooted planar tree, which is encoded by the sequence $(\xi_{\mu_j+\ell-1}^{(c)} : 1 \leq \ell \leq \mu_{j+1} - \mu_j)$. In particular the j -th mutant μ_j is viewed as the root of the cluster C_j . We stress that the allelic split is compatible with the depth-first-search, in the sense that the step at which an individual is explored by depth-first-search in this initial tree is the same as when we consider this individual as a vertex of the forest formed by the allelic clusters.

2. We further stress that, by construction, the order on the set of vertices $\{1, \dots, \tau\}$ that is induced by the depth-first search algorithm on the genealogical tree of mutants coincides with the natural order; more precisely the vertex j is found at the j -th step of the algorithm.

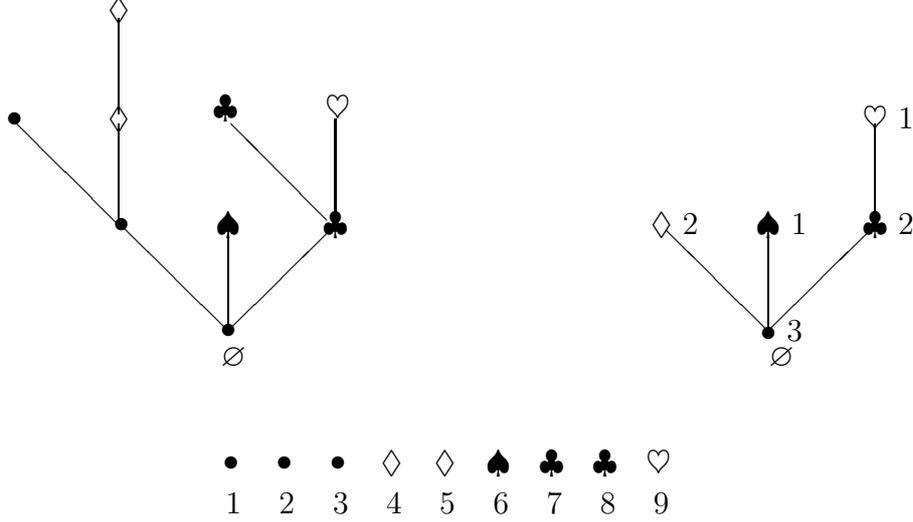


Figure 2 : Genealogical tree with mutations (top-left), the allelic tree (top-right) and the sequence of types listed by depth-first search (bottom). The marks on vertices in the allelic tree are the sizes of the allelic clusters. There are five allelic clusters which are represented by $C_1 = ((1, 2), (1, 1), (0, 0))$, $C_2 = ((1, 0), (0, 0))$, $C_3 = ((0, 0))$, $C_4 = ((1, 1), (0, 0))$ and $C_5 = ((0, 0))$. The ranks of mutants are $\mu_1 = 1$, $\mu_2 = 4$, $\mu_3 = 6$, $\mu_4 = 7$, $\mu_5 = 9$.

Lemma 1 (i) For every $1 \leq j \leq \tau$, the rank μ_j of the j -th mutant is given by

$$\mu_j = 1 + \min\{n \geq 0 : \xi_1^{(c)} + \dots + \xi_n^{(c)} - n = 1 - j\}.$$

(ii) For every $1 \leq j \leq \tau$, let

$$M_j := \sum_{\ell=\mu_j}^{\mu_{j+1}-1} \xi_\ell^{(m)}$$

denote the number of mutant-children generated by the cluster C_j . Then the genealogical tree of mutants is encoded by the sequence

$$(M_j, 1 \leq j \leq \tau).$$

As a consequence, we have

$$\tau = \min\{j \geq 1 : M_1 + \dots + M_j = j - 1\}.$$

Proof: (i) The formula is obvious for $j = 1$. For $j = 2$, the fact that for each individual the clone-children are listed before the mutant-children implies that the rank μ_2 of the second

mutant is given by 1 plus the size of the Eve cluster. Recall that the latter is encoded by the sequence $(\xi_\ell^{(c)} : 1 \leq \ell \leq \mu_2 - 1)$, and we thus know from (1) that there is the identity

$$\mu_2 - 1 = \min\{n \geq 1 : \xi_1^{(c)} + \cdots + \xi_n^{(c)} - n = -1\}.$$

The general case follows by iteration.

(ii) The first claim should be clear from Definition 1(ii). The last identity is then seen from (1) and the fact that τ is the number of allelic clusters. \square

3 Galton-Watson processes with neutral mutations

It is a well-known and useful fact that the depth-first-search algorithm described in the preceding section provides a bijective correspondence between the laws of ordinary (i.e. without mutations) Galton-Watson processes which are critical or sub-critical and the laws of left-continuous random walks (i.e. with steps in $\{-1, 0, 1, 2, \dots\}$) with non-positive drift. Specifically, consider some random variable ξ with values in \mathbb{Z}_+ and mean $\mathbb{E}(\xi) \leq 1$, and a sequence $(\xi_n, n \geq 1)$ of i.i.d. copies of ξ . Define

$$S_n = \xi_1 + \cdots + \xi_n - n \quad , \quad T_j = \min\{n \geq 1 : S_n = -j\} \quad \text{for } j \geq 0.$$

We can think of the random sequence $(\xi_n, 1 \leq n \leq T_1)$ as the population model corresponding to a Galton-Watson process started from a single progenitor and with reproduction law given by the distribution of ξ . We stress that the representation $(\xi_n, 1 \leq n \leq T_1)$ implies that the siblings appearing in the Galton-Watson process have been ordered, but of course this ordering is irrelevant when studying properties of the Galton-Watson process as, loosely speaking, all individuals reproduce independently and with the same law.

Further, it is seen from the strong Markov property that the sequences

$$(\xi_{T_{\ell-1}+n}, 1 \leq n \leq T_\ell - T_{\ell-1}), \quad \ell \in \mathbb{N},$$

are i.i.d. and thus can be used to encode a sequence of i.i.d. Galton-Watson trees, which is referred to as a Galton-Watson forest. We refer e.g. to Section 6.2 in [35] for details; see in particular Equation (308) there. Similar arguments can be developed in the presence of neutral mutations; we shall first discuss trees and then forests.

3.1 Trees

We first introduce some notation. Let $(\xi^{(c)}, \xi^{(m)})$ be a pair of integer-valued random variables which should be thought of respectively as the number of clone-children and the number of mutant-children of a typical individual. We write

$$\pi_{k,\ell} = \mathbb{P}(\xi^{(c)} = k, \xi^{(m)} = \ell), \quad k, \ell \in \mathbb{Z}_+ \tag{3}$$

for their joint probability law and assume throughout this work that

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

i.e. we work in the critical or sub-critical regime. We implicitly exclude the degenerate case when $\xi^{(c)} \equiv 0$ or $\xi^{(m)} \equiv 0$, and, as a consequence, the means $\mathbb{E}(\xi^{(c)})$ and $\mathbb{E}(\xi^{(m)})$ are always less than 1.

We consider a sequence $(\xi_1^{(c)}, \xi_1^{(m)}), (\xi_2^{(c)}, \xi_2^{(m)}), \dots$ of i.i.d. copies of $(\xi^{(c)}, \xi^{(m)})$, and introduce the filtration

$$\mathcal{F}_n := \sigma((\xi_k^{(c)}, \xi_k^{(m)}) : 1 \leq k \leq n).$$

We also define the left-continuous random walks

$$\begin{aligned} S_n &= (\xi_1^{(c)} + \xi_1^{(m)}) + \dots + (\xi_n^{(c)} + \xi_n^{(m)}) - n, \\ S_n^{(c)} &= \xi_1^{(c)} + \dots + \xi_n^{(c)} - n, \quad n \in \mathbb{N}. \end{aligned}$$

Then, recall (2) and introduce for every $j \geq 0$ the (\mathcal{F}_n) -stopping times

$$T_j := \min\{n \geq 1 : S_n = -j\}$$

and

$$T_j^{(c)} := \min\{n \geq 1 : S_n^{(c)} = -j\} \quad \text{for } j \geq 0.$$

As it has been explained in the preceding section, we can view the random sequence

$$((\xi_n^{(c)}, \xi_n^{(m)}) : 1 \leq n \leq T_1)$$

as a Galton-Watson population model with mutations, well-ordered siblings and started from a single ancestor. Note that the number of clone-children and the number of mutant-children of a typical individual are distributed as $(\xi^{(c)}, \xi^{(m)})$, and since the distribution of these quantities is fixed, mutations are neutral in the sense that they do not affect the reproduction law of individuals.

Recall notions that were introduced in this setting, in particular Definition 1 and Lemma 1. It will be convenient for our analysis to extend now some notations to arbitrary indices. In this direction, we observe from Lemma 1(i) that, provided that the number of types

$$\tau_1 = 1 + \sum_{i=1}^{T_1} \xi_i^{(m)}$$

is at least j , the rank of the j -th mutant is

$$\mu_j = 1 + T_{j-1}^{(c)}.$$

This incites us to define for arbitrary $j \geq 1$

$$C_j := ((\xi_{T_{j-1}^{(c)}+\ell}^{(c)}, \xi_{T_{j-1}^{(c)}+\ell}^{(m)}) : 1 \leq \ell \leq T_j^{(c)} - T_{j-1}^{(c)}),$$

and then

$$|C_j| := T_j^{(c)} - T_{j-1}^{(c)} = 1 + \sum_{\ell=T_{j-1}^{(c)}+1}^{T_j^{(c)}} \xi_\ell^{(c)}$$

and

$$M_j := \sum_{\ell=T_{j-1}^{(c)}+1}^{T_j^{(c)}} \xi_\ell^{(m)}.$$

Thus for $j \leq \tau_1$, C_j is the j -th allelic cluster, $|C_j|$ its size, and M_j the number of its mutant-children. We stress that the preceding quantities have been defined also when $j > \tau_1$.

We now record a series of elementary observations.

Lemma 2 (i) *For every $j \in \mathbb{N}$, C_j is adapted to the sigma-field $\mathcal{F}_{T_j^{(c)}}$, whereas C_{j+1} is independent of $\mathcal{F}_{T_j^{(c)}}$ and has the same distribution as C_1 . In particular the sequence C_1, C_2, \dots is i.i.d.*

(ii) *The variable M_j is a functional of C_j and therefore M_1, M_2, \dots is and i.i.d. sequence that is adapted to the filtration $(\mathcal{F}_{T_j^{(c)}})_{j \geq 1}$.*

(iii) *The number of types*

$$\tau_1 = \min\{j \geq 1 : M_1 + \dots + M_j = j - 1\}$$

is an $(\mathcal{F}_{T_j^{(c)}})$ -stopping time.

Proof: Writing as usual θ for the shift-operator on the space of sequences, there is the identity $T_{j+1}^{(c)} = T_1^{(c)} \circ \theta_{T_j^{(c)}}$, which yields the first assertion by an application of the strong Markov property. The second is an obvious consequence of the first and our definitions. The third then follows from Lemma 1(ii). \square

We immediately deduce from Lemma 2 a recursive description of the allelic partition and the statistical structure of the genealogical tree of mutants (see also [33, 39]).

Proposition 1 (i) *For every $j \geq 1$, conditionally on $\tau_1 \geq j + 1$, the $j + 1$ -th allelic cluster C_{j+1} is independent of $(C_k, 1 \leq k \leq j)$ and has the same distribution as the Eve cluster C_1 .*

(ii) *The genealogical tree of mutants has the structure of a (sub)critical Galton-Watson process with reproduction law given by the distribution of M_1 , the number of mutant-children of the Eve cluster.*

Proof: Recall from Lemma 2 that the event $\{\tau_1 \geq j + 1\} = \{\tau_1 \leq j\}^c$ is $\mathcal{F}_{T_j^{(c)}}$ -measurable, and that the clusters $(C_k : 1 \leq k \leq j)$ generate the sigma-field $\mathcal{F}_{T_j^{(c)}}$. Lemma 2(i) then yields the first claim. The second claim is seen by combining Lemma 1(ii) and Lemma 2(ii). \square

3.2 Forests

We now use the infinite sequence $((\xi_1^{(c)}, \xi_1^{(m)}), \dots)$ to encode a Galton-Watson *forest* with neutral mutations by a straightforward generalization of the preceding section. Indeed, it is seen from an application of the strong Markov property that the random finite sequences

$$\mathbb{T}_\ell := ((\xi_{T_{\ell-1}+i}^{(c)}, \xi_{T_{\ell-1}+i}^{(m)}) : 1 \leq i \leq T_\ell - T_{\ell-1}), \quad \ell \in \mathbb{N}$$

are i.i.d. and thus can be used to encode a sequence of i.i.d. planar Galton-Watson trees with neutral mutations, which we call a Galton-Watson forest with mutations. Its distribution is entirely determined by the reproduction law π given by (3).

When we focus on clones, the infinite sequence $(\xi_1^{(c)}, \xi_2^{(c)}, \dots)$ encodes in turn a standard Galton-Watson forest, whose law is determined by the first marginal $\pi^{(c)}$ of π , viz.

$$\pi_k^{(c)} = \mathbb{P}(\xi^{(c)} = k) = \sum_{\ell=0}^{\infty} \pi_{k,\ell}.$$

Specifically, for each $j \in \mathbb{N}$, the finite sequence

$$\mathbb{T}_j^{(c)} = (\xi_{T_{j-1}^{(c)}+i}^{(c)}, 1 \leq i \leq T_j^{(c)} - T_{j-1}^{(c)})$$

encodes a Galton-Watson tree, which can be viewed as the subtree corresponding to a cluster of individuals of the same type (allele) of some tree \mathbb{T}_ℓ in the Galton-Watson forest with mutations.

The interpretation of Lemma 2(i) in that setting is that the allelic cut in a Galton-Watson forest with neutral mutations and reproduction law π produces another a Galton-Watson forest with reproduction law given by the first marginal $\pi^{(c)}$ of π .

It should be plain that both $(T_\ell, \ell \geq 0)$ and $(T_j^{(c)}, j \geq 0)$ are increasing random walks, so their ranges

$$\mathcal{R} := \{T_\ell : \ell \geq 0\} \quad \text{and} \quad \mathcal{R}^{(c)} := \{T_j^{(c)} : j \geq 0\},$$

are two regenerative subsets of \mathbb{Z}_+ , in the sense each can be viewed as the set of renewal epochs of some recurrent event (cf. Feller [20, 21]). Observe that both yield a partition of the set of positive integers (i.e. individuals ranked by depth-first search) into disjoint intervals that correspond to the trees in the forests :

$$\mathbb{N} = \bigcup_{\ell \geq 1}]T_{\ell-1}, T_\ell] = \bigcup_{j \geq 1}]T_{j-1}^{(c)}, T_j^{(c)}].$$

Clearly, there is the embedding

$$\mathcal{R} \subseteq \mathcal{R}^{(c)},$$

and more precisely, this embedding is *compatible with regeneration*, in the sense that for every $k \in \mathbb{Z}_+$, conditionally on $k \in \mathcal{R}$, the shifted sets $\mathcal{R} \circ \theta_k := \{\ell \geq 0 : k + \ell \in \mathcal{R}\}$ and $\mathcal{R}^{(c)} \circ \theta_k := \{j \geq 0 : k + j \in \mathcal{R}^{(c)}\}$ are independent of the sigma-field \mathcal{F}_k and their joint law is the same as that of $(\mathcal{R}, \mathcal{R}^{(c)})$. We refer to [10] for applications of this notion. Roughly speaking, this implies that the allelic split of each interval $]T_{\ell-1}, T_\ell]$ produces smaller intervals $]T_{j-1}^{(c)}, T_j^{(c)}]$ in a random way that only depends on the length $T_\ell - T_{\ell-1}$ (that is the size of \mathbb{T}_ℓ), independently of its location and of the other integer intervals. This can be thought of as a fragmentation property (see [12]) for the sizes of the trees .

4 Some applications of the ballot theorem

We start by stating a version of the classical ballot theorem that will be used in this section; see [38]. Let (X_1, \dots, X_n) be an n -tuple of random variables with values in some space E which

is cyclically exchangeable, in the sense that for every $i \in \mathbb{N}$, there is the identity in law

$$(X_1, \dots, X_n) \stackrel{\mathcal{L}}{=} (X_{i+1}, \dots, X_{i+n}),$$

where we agree that addition of indices is taken modulo n . Consider a function

$$f : E \rightarrow \{-1, 0, 1, 2, \dots\},$$

and assume that

$$\sum_{j=1}^n f(X_j) = -k$$

for some $1 \leq k \leq n$.

Lemma 3 (Ballot Theorem) *Under the assumptions above, the probability that the process of the partial sums of the sequence $f(X_1), \dots, f(X_n)$ remains above $-k$ until the n -step is*

$$\mathbb{P} \left(\min\{j \geq 1 : \sum_1^j f(X_i) = -k\} = n \right) = k/n.$$

4.1 Distribution of the allelic tree

We have now introduced all the tools which are needed for describing the distribution of the allelic tree of a Galton-Watson process with neutral mutations and started from a single ancestor. We only need one more notation. Recall from (3) that π denotes the probability function of the reproduction law of the Galton-process with mutations; for every integer $n \geq 1$, we also write π^{*n} for the n -th convolution product of that law, i.e.

$$\pi_{k,\ell}^{*n} = \mathbb{P}(\xi_1^{(c)} + \dots + \xi_n^{(c)} = k, \xi_1^{(m)} + \dots + \xi_n^{(m)} = \ell).$$

Example : Suppose that the dynamics of the population can be described as follows. We start from a usual Galton-Watson process with reproduction law on \mathbb{Z}_+ , say ϱ , and assume that at each step, mutations affect each child with probability $p \in]0, 1[$, independently of the other children. In other words, the allelic tree is obtained by pruning or percolation on the genealogical tree of the Galton-Watson process, cutting each edge with probability p and independently of the other edges. See e.g. Aldous and Pitman [4], or Chapter 4 in Lyons and Peres [30]. Analytically, this means that if ξ is a random variable with law ϱ , then the conditional distribution of $(\xi^{(c)}, \xi^{(m)})$ given $\xi = k$ is that of $(k - B(k, p), B(k, p))$, where $B(k, p)$ denotes a binomial variable with parameters k and p . In this situation, it is easily seen that

$$\pi_{k,\ell}^{*n} = \binom{k+\ell}{k} (1-p)^k p^\ell \varrho_{k+\ell}^{*n}, \tag{4}$$

with ϱ^{*n} denoting the n -th convolution power of ϱ . This expression is entirely explicit when ϱ is e.g. the Poisson, or binomial, or geometric, distribution as in those cases, there are known formulas for ϱ^{*n} . Of course, there are other natural examples in which the two-dimensional

probability function π^{*n} can be expressed in terms of simpler one-dimensional probability functions, for instance when $\xi^{(c)}$ and $\xi^{(m)}$ are assumed to be independent, or when $\xi^{(c)} = \beta\xi$ and $\xi^{(m)} = (1 - \beta)\xi$ where β stands for a Bernoulli variable which is independent of ξ .

We assume here that the Galton-Watson process starts from a single ancestor. Recall that T_1 denotes the size of the total population, τ_1 the number of types in the population, and for $1 \leq i \leq \tau_1$, that the i -th allelic cluster C_i has size $|C_i|$ and has M_i mutant-children. We may now state our main result, which can be viewed as a generalization of a celebrated identity due to Dwass [18].

Theorem 1 (i) *The joint law of the size of the total population and the number of types is given by*

$$\mathbb{P}(T_1 = n, \tau_1 = k) = \frac{1}{n} \pi_{n-k, k-1}^{*n}, \quad 1 \leq k \leq n.$$

(ii) *The joint law of the size of the Eve population and the number of its mutant-children is given by*

$$\mathbb{P}(|C_1| = n, M_1 = \ell) = \frac{1}{n} \pi_{n-1, \ell}^{*n}, \quad 0 \leq \ell < n.$$

(iii) *For every integers $k \geq 1$, $n_1, \dots, n_k \geq 1$ and $\ell_1, \dots, \ell_k \geq 0$ such that*

$$\sum_{i=1}^j \ell_i > j - 1 \quad \text{whenever} \quad 1 \leq j < k,$$

we have

$$\mathbb{P}(|C_1| = n_1, M_1 = \ell_1, \dots, |C_k| = n_k, M_k = \ell_k) = \prod_{i=1}^k \frac{1}{n_i} \pi_{n_i-1, \ell_i}^{*n_i}.$$

Remarks : 1. Restricting our attention in part (iii) to sequences $\ell_1, \dots, \ell_k \geq 0$ with

$$\inf\{j \geq 1 : \sum_{i=1}^j \ell_i = j - 1\} = k,$$

we stress that the statement describes the law of the entire allelic tree.

2. In particular, the law of the number τ_1 of types is given by

$$\mathbb{P}(\tau_1 = k) = \sum_{n=1}^{\infty} n^{-1} \pi_{n-k, k-1}^{*n}, \quad k \geq 0.$$

It may be interesting to point out that there is also the formula

$$\mathbb{P}(\tau_1 = k) = \frac{1}{k} \nu_{k-1}^{*k}$$

where

$$\nu_\ell = \mathbb{P}(M_1 = \ell) = \sum_{n=1}^{\infty} n^{-1} \pi_{n-1, \ell}^{*n}$$

and ν^{*k} the k -th convolution power of ν . Indeed, this alternative formulation is seen from Proposition 1(ii) and Dwass formula [18].

Proof: Consider the sequence $(\xi_1^{(c)}, \xi_1^{(m)}), \dots, (\xi_n^{(c)}, \xi_n^{(m)})$ of n i.i.d. copies of $(\xi^{(c)}, \xi^{(m)})$ and the partial sums of coordinates

$$\Sigma_j^{(c)} = \sum_{i=1}^j \xi_i^{(c)} \quad , \quad \Sigma_j^{(m)} = \sum_{i=1}^j \xi_i^{(m)} \quad \text{and} \quad \Sigma_j = \Sigma_j^{(c)} + \Sigma_j^{(m)} .$$

Introduce for every $1 \leq k \leq n$ the event

$$\Lambda_{n-k, k-1} = \{ \Sigma_n^{(c)} = n - k, \Sigma_n^{(m)} = k - 1 \} = \{ \Sigma_n = n - 1, \Sigma_n^{(m)} = k - 1 \} ,$$

and observe that the sequence $(\xi_1^{(c)}, \xi_1^{(m)}), \dots, (\xi_n^{(c)}, \xi_n^{(m)})$ is (cyclically) exchangeable conditionally on $\Lambda_{n-k, k-1}$. Further, we have by definition that

$$\mathbb{P}(\Lambda_{n-k, k-1}) = \pi_{n-k, k-1}^{*n} .$$

Plainly, there is the inclusion $\{T_1 = n, \tau_1 = k\} \subseteq \Lambda_{n-k, k-1}$, and more specifically, we deduce from (2) the identity

$$\{T_1 = n, \tau_1 = k\} = \Lambda_{n-k, k-1} \cap \{ \min\{j \geq 1 : \Sigma_j = j - 1\} = n \} .$$

By the ballot theorem (more precisely take $f(x^{(c)}, x^{(m)}) = x^{(c)} + x^{(m)} - 1$ in Lemma 3), we have

$$\mathbb{P}(\min\{j \geq 1 : \Sigma_j = j - 1\} = n \mid \Lambda_{n-k, k-1}) = 1/n ,$$

which yields the formula (i).

The proof of (ii) is similar, observing that

$$\{|C_1| = n, M_1 = \ell\} = \Lambda_{n-1, \ell} \cap \{ \min\{j \geq 1 : \Sigma_j^{(c)} = j - 1\} = n \} .$$

Finally (iii) follows from (ii) and Proposition 1(i) by iteration. \square

4.2 Conditioning on the population size and the number of types

In the rest of this section, we will be interested in the relative sizes of the allelic clusters, ignoring the genealogical tree structure that connects them. We start with a description in that vein which is essentially a variation of that in Theorem 1(iii). Recall that a random uniform cyclic permutation of $\{1, \dots, k\}$, say σ , is given by $\sigma(i) = U + i$ where U is uniform on $\{1, \dots, k\}$ and the addition is taken modulo k .

Corollary 1 *Fix $1 \leq k \leq n$ and let σ be a random uniform cyclic permutation of $\{1, \dots, k\}$ which is independent of the Galton-Watson process. Then for every positive integers n_1, \dots, n_k with $n_1 + \dots + n_k = n$, we have*

$$\mathbb{P}(|C_{\sigma(1)}| = n_1, \dots, |C_{\sigma(k)}| = n_k \mid T_1 = n, \tau_1 = k) = \frac{1}{\pi_{n-k, k-1}^{*n}} \sum \prod_{i=1}^k \frac{1}{n_i} \pi_{n_i-1, \ell_i}^{*n_i} ,$$

where in the right-hand side, the sum is taken over the sequences ℓ_1, \dots, ℓ_k in \mathbb{Z}_+ such that $\ell_1 + \dots + \ell_k = k - 1$.

Proof: An immediate application of the ballot theorem shows that the conditional distribution of $(C_{\sigma(1)}, \dots, C_{\sigma(k)})$ given $T_1 = n$ and $\tau_1 = k$ is the same as that of (C_1, \dots, C_k) conditioned on $\sum_{i=1}^k |C_i| = n$ and $\sum_{i=1}^n \xi_i^{(m)} = k - 1$. Theorem 1(ii) yields the claim. \square

Next, we normalize the size $|C_i|$ of each cluster by the size T_1 of the total population (we assume again that there is a single ancestor), and write

$$\Gamma_1 \geq \Gamma_2 \geq \dots \geq \Gamma_{\tau_1}$$

for the sequence which is obtained by ranking these quantities in the decreasing order. So $\Gamma = (\Gamma_1, \dots, \Gamma_{\tau_1})$ is a proper partition of the unit mass, in the sense that it is given by a ranked sequence of positive real numbers with sum 1. The space of mass partitions (possibly with infinitely many strictly positive terms and sum less than 1) is endowed with the uniform distance, which yields a compact metric space; see Section 2.1 in [12] for details.

Our purpose now is to investigate the asymptotic behavior of the random mass partition Γ , under the conditional probability given the size $T_1 = n$ of the total population and the number $\tau_1 = k$ of types, when $n, k \rightarrow \infty$. We shall show that, under appropriate hypotheses, one can establish convergence in distribution, where the limit can be described as follows. For some fixed parameter $b > 0$, consider the sequence $\mathbf{a}_1 > \mathbf{a}_2 > \dots > 0$ of the atoms ranked in the decreasing order of a Poisson point measure on $]0, \infty[$ with intensity $ba^{-3/2}da$. We can then condition by $\sum_{i=1}^{\infty} \mathbf{a}_i = 1$ and get a random proper mass-partition; see e.g. [34] or Proposition 2.4 in [12] for a rigorous definition of this conditioning by a singular event. This family of random mass-partitions has appeared previously in a remarkable work by Aldous and Pitman [5], more precisely it arose by logging the Continuum Random Tree according to Poissonian cuts along its skeleton (see also [2, 6, 11, 31] for related works). As we know from Aldous [3] that the Continuum Random Tree can be viewed as the limit when $n \rightarrow \infty$ of Galton-Watson trees conditioned to have total size n , the fact that these random mass-partitions appears again in the present setting should not come as a surprise.

For the sake of simplicity, we shall focus on the case when the number of clone-children $\xi^{(c)}$ and the number of mutant-children $\xi^{(m)}$ are independent, although it seems likely that our argument should also apply to more general situations. Recall that the expected number of clone-children of a typical individual is $\mathbb{E}(\xi^{(c)}) < 1$. We shall work under the hypothesis that by a suitable exponential tilting, this sub-critical random variable can be turned into a critical one with finite variance. That is, we shall assume that there exists a real number $\theta > 1$ such that

$$\mathbb{E}(\xi^{(c)}\theta^{\xi^{(c)}}) = \mathbb{E}(\theta^{\xi^{(c)}}) \quad \text{and} \quad \sigma_{\theta}^2 := \mathbb{E}((\xi^{(c)})^2\theta^{\xi^{(c)}})/\mathbb{E}(\theta^{\xi^{(c)}}) < \infty. \quad (5)$$

It can be readily checked that (5) then specifies θ uniquely.

Proposition 2 *Suppose that $\xi^{(c)}$ and $\xi^{(m)}$ are independent, that neither distribution is supported by a strict sub-group of \mathbb{Z} and that (5) holds. Fix $b > 0$ and let $n, k \rightarrow \infty$ according to the regime $k \sim b\sqrt{n}$. Then the conditional law of Γ given that the size of the total population is $T_1 = n$ and the number of types $\tau_1 = k$ converges weakly on the space of mass-partitions to the sequence $(\mathbf{a}_1, \mathbf{a}_2, \dots)$ of the atoms of a Poisson random measure on $]0, \infty[$ with intensity*

$$\frac{b}{\sqrt{2\pi\sigma_{\theta}^2}a^3} da, \quad a > 0,$$

ranked in the decreasing order and conditioned by $\sum_{i=1}^{\infty} \mathbf{a}_i = 1$.

Remark : The special case when $\xi^{(c)}$ and $\xi^{(m)}$ are two independent Poisson variables, say with rates $r^{(c)}$ and $r^{(m)}$ can also be viewed as an instance of the situation where mutations affect children independently with probability $p = r^{(m)}/(r^{(c)} + r^{(m)})$ (cf. the example discussed before Theorem 1). More precisely the reproduction law of the standard Galton-Watson process is then Poisson with rate $r^{(c)} + r^{(m)}$. This special case has some importance, as it is well-known that conditioning a Galton-Watson tree with Poisson(1) reproduction law to have a size n and then assigning to each individual a distinct label in $\{1, \dots, n\}$ by uniform sampling without replacement yields the uniform distribution on the set of rooted trees with n labeled vertices.

Proof: Let $\tilde{\mathbb{P}}$ denote the probability measure which is obtained from \mathbb{P} by exponential tilting, and more precisely, in such a way that the variables $\xi_1^{(c)}, \dots$ are i.i.d. under $\tilde{\mathbb{P}}$ with law given by

$$\tilde{\mathbb{P}}(\xi^{(c)} = j) = \theta^j \mathbb{P}(\xi^{(c)} = j) / z_\theta, \quad j \in \mathbb{Z}_+,$$

where z_θ is the normalization factor, viz.

$$z_\theta = \mathbb{E}(\theta^{\xi^{(c)}}).$$

As in the proof of Corollary 1, we see from an application of the ballot theorem that the conditional distribution of $(n\Gamma_1, \dots, n\Gamma_{\tau_1})$ given $T_1 = n$ and $\tau_1 = k$ is the same as that obtained from the i.i.d. sequence $|C_1|, \dots, |C_k|$ by ranking in the decreasing order and conditioning on $\sum_{i=1}^k |C_i| = n$ and $\sum_{i=1}^n \xi_i^{(m)} = k - 1$. Further, recall from Lemma 1(i) that the variables $|C_1|, \dots, |C_k|$ are functions of $\xi_1^{(c)}, \dots, \xi_n^{(c)}$, and thus the assumption of independence between $\xi^{(c)}$ and $\xi^{(m)}$ enables us to ignore the conditioning on $\sum_{i=1}^n \xi_i^{(m)} = k - 1$. Finally, it should be clear that the exponential tilting does not affect such conditional law, in the sense that the sequence $|C_1|, \dots, |C_k|$ has the same distribution under $\mathbb{P}(\cdot | T_1 = n)$ as under $\tilde{\mathbb{P}}(\cdot | T_1 = n)$.

We then estimate the distribution of the size of the Eve cluster under $\tilde{\mathbb{P}}$, which is given again according to Dwass formula [18] by

$$\tilde{\mathbb{P}}(|C_1| = n_1) = \frac{1}{n_1} \tilde{\mathbb{P}}(\xi_1^{(c)} + \dots + \xi_{n_1}^{(c)} = n_1 - 1).$$

Recall that by assumption, $\xi^{(c)}$ is critical with variance σ_θ^2 under $\tilde{\mathbb{P}}$, so an application of Gnedenko's local central limit theorem gives

$$\tilde{\mathbb{P}}(|C_1| = n_1) \sim \frac{1}{\sqrt{2\pi\sigma_\theta^2 n_1^3}} \quad \text{as } n_1 \rightarrow \infty.$$

Putting the pieces together, we get that the conditional distribution of $(n\Gamma_1, \dots, n\Gamma_{\tau_1})$ given $T_1 = n$ and $\tau_1 = k$ is the same as that obtained from an i.i.d. sequence Y_1, \dots, Y_k by ranking in the decreasing order and conditioning on $\sum_{i=1}^k Y_i = n$, where

$$\mathbb{P}(Y_1 = n_1) \sim \frac{1}{\sqrt{2\pi\sigma_\theta^2 n_1^3}} \quad \text{as } n_1 \rightarrow \infty.$$

An application of Corollary 2.2 in [12] completes the proof of our claim. □

5 Lévy forests with mutations

The purpose of this section is to point at an interpretation of a standard limit theorem involving left-continuous (i.e. downwards skip-free) random walks and Lévy processes with no negative jumps, in terms of Galton-Watson and Lévy forests in the presence of neutral mutations. We first introduce some notation and hypotheses in this area, referring to the monograph by Duquesne and Le Gall [16] for details.

For every integer $n \geq 1$, let $(\xi^{(c)}(n), \xi^{(m)}(n))$ be a pair of integer valued random variables with

$$\mathbb{E}(\xi^{(c)}(n) + \xi^{(m)}(n)) = 1.$$

We consider two left-continuous random walks

$$S(n) = (S_i(n) : i \in \mathbb{Z}_+) \quad \text{and} \quad S^{(c)}(n) = (S_i^{(c)}(n) : i \in \mathbb{Z}_+)$$

whose steps are (jointly) distributed as $\xi^{(c)}(n) + \xi^{(m)}(n) - 1$ and $\xi^{(c)}(n) - 1$.

Let also $X = (X_t, t \in \mathbb{R}_+)$ denote a Lévy process with no negative jumps and Laplace exponent ψ , viz.

$$\mathbb{E}(\exp -\lambda X_t) = \exp t\psi(\lambda) \quad \text{for every } \lambda, t \geq 0.$$

We further suppose that X does not drift to $+\infty$, which is equivalent to $\psi'(0+) \geq 0$, and that

$$\int_1^\infty \frac{d\lambda}{\psi(\lambda)} < \infty.$$

The key hypothesis is the existence of a nondecreasing sequence of positive integers $(\gamma_n, n \in \mathbb{N})$ converging to ∞ and such that

$$\lim_{n \rightarrow \infty} n^{-1} S_{\gamma_n}(n) = X_1 \quad \text{in law} \quad (6)$$

and

$$\lim_{n \rightarrow \infty} n^{-1} \gamma_n \mathbb{E}(\xi^{(m)}(n)) = d \quad (7)$$

for some $d \geq 0$.

We also need to introduce a different procedure for encoding forests by paths, which is more convenient to work with when discussing continuous limits of discrete structures. For each $n \geq 1$, we write $H(n) = (H_i(n), i \in \mathbb{N})$ for the (discrete) *height function* of the Galton-Watson forest $(\mathbb{T}_\ell, \ell \in \mathbb{Z}_+)$. That is, for $0 \leq i < |\mathbb{T}_1|$, $H_i(n)$ denotes the generation of the $i + 1$ -th individual found by depth-first search on the tree \mathbb{T}_1 , and more generally, when $|\mathbb{T}_1| + \dots + |\mathbb{T}_\ell| \leq i < |\mathbb{T}_1| + \dots + |\mathbb{T}_{\ell+1}|$, $H_i(n)$ is the generation of the $(1 + i - |\mathbb{T}_1| - \dots - |\mathbb{T}_\ell)$ -th individual found by depth-first search on the tree $\mathbb{T}_{\ell+1}$. Similarly, we write $H^{(c)}(n) = (H_i^{(c)}(n), i \in \mathbb{N})$ for the height function of the Galton-Watson forest $(\mathbb{T}_j^{(c)}, j \in \mathbb{N})$. In the continuous setting, trees and forests can be defined for a fairly general class of Lévy processes with no negative jumps, and in turn are encoded by (continuous) height functions; cf. Chapter 1 in [16] for precise definitions and further references.

Theorem 2 *Suppose that (6) and (7) hold, and also that the technical condition (2.27) in [16] is fulfilled. Then the pair of rescaled height functions*

$$\left(\gamma_n^{-1}(H_{[t\gamma_n]}(n), H_{[t\gamma_n]}^{(c)}(n)), t \geq 0 \right)$$

converges in distribution the sense of weak convergence on Skorohod space $\mathbb{D}(\mathbb{R}_+, \mathbb{R}_+)$ as $n \rightarrow \infty$ towards

$$\left((\mathcal{H}_t, \mathcal{H}_t^{(d)}), t \geq 0 \right),$$

where $\mathcal{H} = (\mathcal{H}_t, t \geq 0)$ is the height process constructed from the Lévy process $X = (X_t, t \geq 0)$ and $\mathcal{H}^{(d)} = (\mathcal{H}_t^{(d)}, t \geq 0)$ is the height process constructed from the Lévy process $X^{(d)} = (X_t^{(d)} = X_t - dt, t \geq 0)$.

Remark : More recently, Duquesne and Le Gall [17] (see also the survey [28]) have developed the framework when Lévy trees are viewed as random variables with values in the space of real trees, endowed with the Gromov-Hausdorff distance. Theorem 2 can also be re-stated in this setting.

Proof: The assumption (6) ensures the convergence in distribution

$$(n^{-1}S_{[t\gamma_n]}(n), t \geq 0) \implies (X_t, t \geq 0),$$

see Theorem 2.1.1 in [16] and Equation (2.3) there. On the other hand, by the law of large numbers, the assumption (7) entails that

$$\lim_{n \rightarrow \infty} n^{-1}(S_{[t\gamma_n]}(n) - S_{[t\gamma_n]}^{(c)}(n)) = dt$$

uniformly for t in compact intervals, a.s. The joint convergence in distribution

$$\left((n^{-1}S_{[t\gamma_n]}(n), n^{-1}S_{[t\gamma_n]}^{(c)}(n)), t \geq 0 \right) \implies ((X_t, X_t - dt), t \geq 0),$$

follows. We complete the proof as in Theorem 2.3.1 in [16]. □

We now conclude this work by discussing a natural example. Specifically, we suppose that the distribution of

$$\xi^{(c)}(n) + \xi^{(m)}(n) = \xi(n) := \xi$$

is the same for all n . For the sake of simplicity, we assume also that $\mathbb{E}(\xi) = 1$ and $\text{Var}(\xi) = 1$. We may then take $\gamma_n = n^2$, so by the central limit theorem, (6) holds and the Lévy process X is a standard Brownian motion. We fix an arbitrary $d > 0$ and consider the independent pruning model where for each integer $n > d$, conditionally on the total number of children $\xi^{(c)}(n) + \xi^{(m)}(n) = k$, the number $\xi^{(m)}(n)$ of mutant-children of a typical individual has the binomial distribution $B(k, d/n)$. In other words, in the n -th population model, mutations affect each child with probability d/n , independently of the other children. Then (7) clearly holds. Roughly speaking, Theorem 2 implies in this setting that the initial Galton-Watson forest associated with the n -th population model, converges in law after a suitable re-normalization to the Brownian forest, whereas the forest which is induced by the allelic partition and re-normalized in the same way, converges to the forest generated by the same Brownian motion but now with drift $-d$.

This provides an explanation to the rather intriguing relation which identifies two seemingly different fragmentation processes: the fragmentation process constructed by Aldous and Pitman [5] by logging the Continuum Random Tree according to a Poisson point process on its skeleton, and the fragmentation process constructed in [11] by splitting the unit interval at instants when a Brownian excursion with negative drift reaches a new infimum. It is interesting to mention that Schweinsberg [37] already pointed at several applications of the (continuous) ballot theorem in this framework. More generally, the transformation $X \rightarrow X^{(d)}$ of Lévy processes with no negative jumps also appeared in an article by Miermont [31] on certain eternal additive coalescents, whereas Aldous and Pitman [6] showed that the latter arise asymptotically from independent pruning of certain sequences of birthday trees.

References

- [1] Abraham, R. and Delmas, J.-F. Williams' decomposition of the Lévy continuous random tree and simultaneous extinction probability for populations with neutral mutations. Preprint
- [2] Abraham, R. and Serlet, L. Poisson snake and fragmentation. *Electron. J. Probab.* **7** (2002), Paper No.17. Available via : <http://www.math.washington.edu/~ejpecp/>
- [3] Aldous, D. J. The continuum random tree. III. *Ann. Probab.* **21** (1993), 248-289.
- [4] Aldous, D. J. and Pitman, J. Tree-valued Markov chains derived from Galton-Watson processes. *Ann. Inst. Henri Poincaré* **34** (1998), 637-686.
- [5] Aldous, D. J. and Pitman, J. The standard additive coalescent. *Ann. Probab.* **26** (1998), 1703-1726.
- [6] Aldous, D. J. and Pitman, J. Inhomogeneous continuum random trees and the entrance boundary of the additive coalescent. *Probab. Theory Related Fields* **118** (2000), 455-482.
- [7] Athreya, K.B. and Ney, P.E. *Branching processes*. Springer-Verlag, Berlin-Heidelberg-New York, 1972.
- [8] Basdevant, A.-L. and Goldschmidt, C. Asymptotics of the allele frequency spectrum associated with the Bolthausen-Sznitman coalescent, arXiv:0706.2808 (2007).
- [9] Berestycki, J., Berestycki, N. and Schweinsberg, J. Beta-coalescents and continuous stable random trees. *Ann. Probab.* **35** (2007), 1835-1887.
- [10] Bertoin, J. Renewal theory for embedded regenerative sets. *Ann. Probab.* **27** (1999), 1523-1535.
- [11] Bertoin, J. A fragmentation process connected to Brownian motion. *Probab. Theory Related Fields* **117** (2000), 289-301.
- [12] Bertoin, J. *Random fragmentation and coagulation processes*. Cambridge Studies in Advanced Mathematics **102**, Cambridge University Press, 2006.

- [13] Crump, K. S. and Gillespie, J. H. The dispersion of a neutral allele considered as a branching process. *J. Appl. Probab.* **13** (1976), 208-218.
- [14] Delmas, J.-F., Dhersin, J.-S. and Siri-Jegousse, A. Asymptotic results on the length of coalescent tree. To appear in *Ann. Appl. Probab.* (2007).
- [15] Dong, R., Gnedin, A. and Pitman, J. Exchangeable partitions derived from Markovian coalescents, arXiv:math.PR/0603745 (2006).
- [16] Duquesne, T. and Le Gall, J.-F. *Random trees, Lévy processes and spatial branching processes*. Astérisque **281** (2002).
- [17] Duquesne, T. and Le Gall, J.-F. Probabilistic and fractal aspects of Lévy trees. *Probab. Theory Relat. Fields* **131** (2005), 553-603.
- [18] Dwass, M. The total progeny in a branching process. *J. Appl. Probab.* **6** (1969), 682-686.
- [19] Ewens, W. J. The sampling theory of selectively neutral alleles, *Theoret. Popul. Biol.* **3** (1972), 87-112.
- [20] Feller, W. *An introduction to probability theory and its applications*. Vol I. 3rd ed. Wiley Series in Probability and Mathematical Statistics. New York, 1968.
- [21] Feller, W. *An introduction to probability theory and its applications*. Vol II. 2nd ed. Wiley Series in Probability and Mathematical Statistics. New York, 1971.
- [22] Griffiths, R.C. and Pakes, A. G. An infinite-alleles version of the simple branching process. *Adv. Appl. Probab.* **20** (1988), 489-524.
- [23] Harris, T. E. First passage and recurrence distributions. *Trans. Amer. Math. Soc.* **73** (1952), 471-486.
- [24] Harris, T.E. *The theory of branching processes*. Springer-Verlag, New York, 1963.
- [25] Kingman, J. F. C. *Mathematics of genetic diversity*. CBMS-NSF Regional Conference Series in Applied Mathematics, 34. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., 1980.
- [26] Kimmel, M. and Axelrod, D. E. *Branching Processes in Biology*. New York, Springer, 2002.
- [27] Lambert, A. Spine decompositions and allelic partitions of splitting trees. In preparation.
- [28] Le Gall, J.-F. Random trees and applications. *Probab. Surv.* **2** (2005), 245-311. Available via <http://www.i-journals.org/ps/>
- [29] Liggett, T.M., Schinazi, R.B. and Schweinsberg, J. A contact process with mutations on a tree. To appear in *Stochastic Process. Appl.*
- [30] Lyons, R. and Peres, Y. *Probability on trees and networks*. Available via : <http://php.indiana.edu/~rdlyons/prbtree/book.pdf>

- [31] Miermont, G. Ordered additive coalescent and fragmentations associated to Lévy processes with no positive jumps. *Electron. J. Probab.* **6** (2001). Available via :
<http://www.math.washington.edu/~ejpecp/>
- [32] Möhle, M. On sampling distributions for coalescent processes with simultaneous multiple collisions. *Bernoulli* **12** (2006), 35-53.
- [33] Nerman, O. Branching processes and neutral mutations. In : *Mathematical statistics, theory and applications*, Proc. World Congr. Bernoulli Soc., Tashkent/USSR 1986, Vol. 2 (1987), 683-692.
- [34] Perman, M., Pitman, J. and Yor, M. Size-biased sampling of Poisson point processes and excursions. *Probab. Th. Rel. Fields* **92** (1992) 21-39.
- [35] Pitman, J. *Combinatorial stochastic processes. École d'été de Probabilités de St-Flour*, Lect. Notes in Maths **1875**, Berlin, Springer, 2006. Available via :
<http://stat-www.berkeley.edu/users/pitman/>
- [36] Schinazi, R.B. and Schweinsberg, J. Spatial and non-spatial stochastic models for immune response. To appear in *Markov Process. Related Fields*.
- [37] Schweinsberg, J. Applications of the continuous-time ballot theorem to Brownian motion and related processes. *Stochastic Process. Appl.* **95** (2001), 151-176.
- [38] Takács, L. *Combinatorial methods in the theory of stochastic processes*. John Wiley and Sons, New York, 1967.
- [39] Taïb, Z. *Branching processes and neutral evolution*. Lecture Notes in Biomathematics **93**. Springer-Verlag, Berlin, 1992.