The Annals of Probability
2009, Vol. 37, No. 4, 1502–1523
DOI: 10.1214/08-AOP441

THE STRUCTURE OF THE ALLELIC PARTITION OF THE TOTAL POPULATION FOR GALTON–WATSON PROCESSES WITH NEUTRAL MUTATIONS

BY JEAN BERTOIN

Université Pierre et Marie Curie

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.

1. Introduction.

We consider a Galton–Watson process, that is, 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 a clone, that is, of the same type (or allele) as its parent, or a mutant, that is, of a new type. 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 allele. 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., [5, 14, 23, 31, 34, 39]); we also refer to [1, 6, 7, 28, 30, 37] for interesting variations (these references are of course far from being exhaustive). Note also that Galton–Watson processes with mutations can be viewed as a special instance of multitype branching processes (see Chapter V in Athreya and Ney [8] or Chapter 7 in Kimmel and Axelrod [26]).

We are interested in the partition of the population into clusters of individuals having the same allele, which will be referred to as the *allelic partition*. Statistics of the allelic partition of a random population model with neutral mutations have been first determined in a fundamental work of Ewens [20] for the Wright–Fisher

Received January 2008; revised August 2008.

AMS 2000 subject classifications. 60J80, 60J10.

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

model (more precisely this concerns the partition of the population at a fixed generation). Kingman [27] provided a deep analysis of this framework, in connection with the celebrated coalescent process that depicts the genealogy of the Wright–Fisher model. We refer to [9, 10, 15, 16, 33] 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 Galton–Watson processes with neutral mutations. We will always assume that the Galton–Watson process is critical or subcritical, so the descent of any individual becomes eventually extinct, and in particular the allelic clusters are finite a.s. We suppose that every ancestor (i.e., individual in the initial population) bears a different allele; it is convenient to view each ancestor as a mutant of the zeroth kind. We then call mutant of the first kind a mutant-child of an individual of the allelic cluster of an ancestor, and the set of all its clones (including that mutant) a cluster of the first kind. By iteration, we define mutants and clusters of the kth kind for any integer $k \geq 0$.

In order to describe the statistics of the allelic partition, we distinguish an ancestor which will then be referred to as *Eve*, and focus on its descent. The set of all individuals bearing the same allele as Eve is called the *Eve cluster*. 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 kth 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 forest 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 who stem from the Eve cluster 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 [24, 25] between ordinary Galton–Watson processes and sequences of i.i.d. integer-valued random variables. Specifically, we incorporate neutral mutations in Harris representation and by combination with the celebrated ballot theorem (which is another classical tool in this area as it is expounded, e.g., by Pitman; see Chapter 6 in [36]), we obtain expressions for the joint distribution of various natural variables (size of the total descent of an ancestor, number of alleles, 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 mutant-children of an individual are independent (and some further technical conditions), the sequence of the relative sizes of the allelic clusters in a typical tree has a limiting conditional distribution when the size of the tree and the number of types both tend to infinity according to some appropriate regime. The limiting distribution that arises has already appeared in the study of the standard additive coalescent by Aldous and Pitman [6]. We also point at limit theorems for allelic partitions of Galton–Watson forests, where, following Duquesne and Le Gall [17, 18], the limits are described in terms of certain Lévy trees. In particular, this provides an explanation to a rather 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 [6], and on the other hand by splitting the unit-interval at instants when the standard Brownian excursion with a negative drift reaches new infima [11].

2. Allelic partitions in a Galton–Watson forest.

We first develop some material and notation about Galton–Watson forests with neutral mutations, referring to Chapter 6 in Pitman [36] for background in the case without mutations.

2.1. *Basic setting.*

Let

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

be a pair of nonnegative 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 also write

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

for the total number of children, and assume throughout this work that

$$\mathbb{E}(\xi^{(+)}) \le 1,$$

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

We write $\mathbb{Z}_+$ and $\mathbb{N}$ for the sets of nonnegative integers and positive integers, respectively. A pair $(g, n) \in \mathbb{Z}_+ \times \mathbb{N}$ is then used to identify an individual in an infinite population model, where the first coordinate g refers to the generation and the second coordinate n to the rank of the individual of that generation (we stress that each generation consists of an infinite sequence of individuals). We assume that each individual at generation $g + 1$ has a unique parent at generation g. We consider a family

$$(\xi_{g,n} : g \in \mathbb{Z}_+ \text{ and } n \in \mathbb{N})$$

of i.i.d. copies of ξ which we use to define the Galton–Watson process with neutral mutations. Specifically, $\xi_{g,n} = (\xi_{g,n}^{(\mathrm{c})}, \xi_{g,n}^{(\mathrm{m})})$ is the pair given by the number of clone-children and mutant-children of the nth individual at generation g. We may assume that the offspring of each individual is ranked, which induces a natural order at the next generation by requiring further that if (g, n) and (g, n') are two individuals at the same generation g with $n < n'$, then at generation $g + 1$ the children of (g, n) are all listed before those of (g, n').

2.2. *Encoding the Galton–Watson forest with mutations.* Next, we enumerate as follows the individuals of the entire population (i.e., of all generations) by a variation of the well-known depth-first search algorithm that takes mutations into account. We associate to each individual a label (a, m, s), where $a \in \mathbb{N}$ is the rank of the ancestor in the initial population, m the number of mutations and s a finite sequence of positive integers which keeps track of the genealogy of the individual. Specifically, the label of the ath individual in the initial generation $g = 0$ is $(a, 0, \varnothing)$. If an individual at the gth generation has the label $(a, m, (i_1, \dots, i_g))$, and if this individual has $j^{(\mathrm{c})}$ clone-children and $j^{(\mathrm{m})}$ mutant-children, then the labels assigned to its clone-children are

$$(a, m, (i_1, \dots, i_g, 1)), \dots, (a, m, (i_1, \dots, i_g, j^{(\mathrm{c})})),$$

whereas the labels assigned to its mutant-children are

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

Clearly, any two distinct individuals have different labels. We then introduce the (random) map

$$\rho : \mathbb{N} \to \mathbb{Z}_+ \times \mathbb{N},$$

which consists in ranking the individuals in the lexicographic order of their labels; see Figure 1. That is to say that $\rho(i) = (g, n)$ if and only if the ith individual in the lexicographic order of labels corresponds to the nth individual at generation g. This procedure for enumerating the individuals will be referred to as the *depth-first search algorithm with mutations*. We shall also use the notation

$$\xi_i = \xi_{\rho(i)}, \qquad i \in \mathbb{N},$$

and whenever no generation is specified, the terminology ith individual will implicitly refer to the rank of that individual induced by depth-first search with mutation, that is, the ith individual means the nth individual at generation g where $\rho(i) = (g, n)$.

LEMMA 1. (i) *The variables $\xi_1, \xi_2, \dots$ are i.i.d. with the same law as ξ.*

(ii) *The sequence $(\xi_{g,n} : g \in \mathbb{Z}_+$ and $n \in \mathbb{N})$ can be recovered from $(\xi_i : i \in \mathbb{N})$ a.s.*

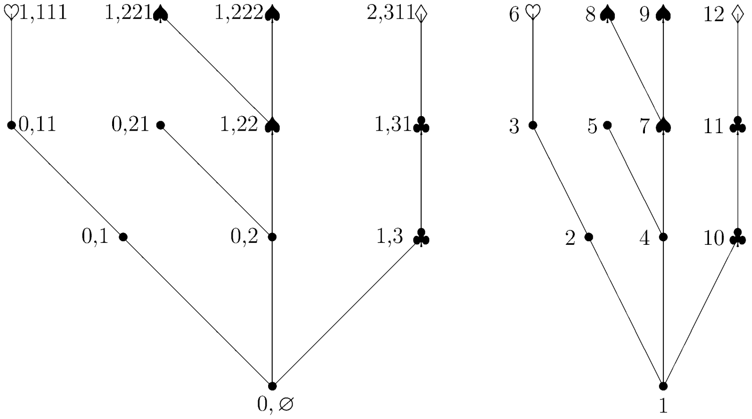

FIG. 1. *Depth-first search with mutations on a genealogical tree. The symbols* •, ♠, ♡, ♢, ♣ *represent the different alleles. Left*: *the label* (m,s) *of an individual is given by the number* m *of mutations and the sequence* s *that specifies its genealogy*; *for the sake of simplicity*, *the rank* a *of the ancestor has been omitted. Right*: *the same tree with individuals ranked by the depth-first search algorithm with mutations.*

PROOF. It should be plain from the definition of the depth-first search algorithm with mutations that for every $i \in \mathbb{N}$, $\rho(i+1)$ is a deterministic function of $\xi_1, \dots, \xi_i$ which takes values in $(\mathbb{Z}_+ \times \mathbb{N}) \setminus \{\rho(1), \dots, \rho(i)\}$. Since $(\xi_{g,n} : g \in \mathbb{Z}_+ \text{ and } n \in \mathbb{N})$ is a sequence of i.i.d. variables with the same law as ξ, this yields the first claim by induction. The second claim follows from the fact that each individual has a finite descent a.s. [because the Galton–Watson process is (sub-)critical], which easily entails that the map ρ is bijective. Further, it is readily seen that the inverse bijection is a function of the sequence $(\xi_i : i \in \mathbb{N})$. □

Henceforth, we shall therefore encode the Galton–Watson process with neutral mutations by a sequence $(\xi_i : i \in \mathbb{N})$ of i.i.d. copies of ξ. We denote by $(\mathcal{F}_i)_{i\in\mathbb{N}}$ the natural filtration generated by this sequence.

We next briefly describe the genealogy of the Galton–Watson process as a forest of i.i.d. genealogical trees. Denote for every $n \in \mathbb{N}$ by

$$\alpha_n = \rho^{-1}(0,n),$$

so that $\alpha_1 = 1 < \alpha_2 < \cdots$ is the increasing sequence of the ranks of ancestors induced by the depth-first search algorithm with mutations. For example, $\alpha_2 = 13$ in the situation described by Figure 1. The procedure for labeling individuals ensures that the descent of the ith ancestor α_i corresponds to the integer interval

$$[\alpha_i, \alpha_{i+1}[:= \{\alpha_i, \alpha_i + 1, \dots, \alpha_{i+1} - 1\}$$

(that is to say, if we index the population model using generations, then the descent of $(0,i)$ is the image of $[\alpha_i, \alpha_{i+1}[$ by the inverse bijection ρ^{-1}).

We write

$$\mathbb{T}_i := (\xi_{\alpha_i-1+\ell} : 1 \le \ell \le \alpha_{i+1} - \alpha_i)$$

for the finite sequence of the numbers of clone-children and mutant-children of the individuals in the descent of the ith ancestor. So $\mathbb{T}_i$ encodes (by the depth-first search algorithm with mutations) the genealogical tree of the ith ancestor, and it should be intuitively clear that the family $(\mathbb{T}_i : i \in \mathbb{N})$ is a forest consisting in a sequence of i.i.d. genealogical trees. To give a rigorous statement, it is convenient to introduce the downward skip-free (or left-continuous) random walk

$$S_n^{(+)} := \xi_1^{(+)} + \cdots + \xi_n^{(+)} - n, \qquad n \in \mathbb{Z}_+, \tag{1}$$

and the passage times

$$T_i^{(+)} := \inf\{n \ge 0 : S_n^{(+)} = -i\}, \qquad i \in \mathbb{Z}_+. \tag{2}$$

We stress that the $T_i^{(+)}$ form an increasing sequence of $(\mathcal{F}_n)$-stopping times.

LEMMA 2. *There is identity*

$$\alpha_i - 1 = T_{i-1}^{(+)}$$

for every $i \in \mathbb{N}$ and, as a consequence, the sequence $\mathbb{T}_1, \dots$ is i.i.d.

PROOF. This formula is a close relative of the classical identity of Dwass [19] and would be well known if individuals were enumerated by the usual depth-first search algorithm (i.e., without taking care of mutations), see, for example, Lemma 63 in [36] or [29]. The proof in the present case is similar. Indeed the formula is obvious for $i = 1$, and for $i = 2$, we have on the one hand that

$$\alpha_2 - 1 = 1 + \xi_1^{(+)} + \cdots + \xi_{\alpha_2-1}^{(+)}$$

by expressing the fact that the predecessor of the second ancestor found by depth-first search with mutations has a rank given by the size of the population generated by Eve, that is, Eve herself and her descendants. On the other hand, we must have $1 + \xi_1^{(+)} + \cdots + \xi_n^{(+)} > n$ when $n < \alpha_2 - 1$, since otherwise the depth-first search algorithm with mutations would explore the second ancestor before having completed the exploration of the entire descent of Eve. This proves the identity for $i = 2$, and the general case then follows by iteration. Finally, the last claim is an immediate consequence of Lemma 1(i) and the strong Markov property. □

2.3. *Allelic partitions.* We can now turn our attention to defining allelic partitions. In this direction, recall that every ancestor has a different type (i.e., bears a different allele), and thus should be viewed as an initial mutant. More generally, we call *mutant* an individual which either belongs to the initial generation or is the mutant-child of some individual, and then write

$$1 = \mu_1 < \mu_2 < \cdots$$

for the ranks of mutants in the depth-first search algorithm with mutations. For example, $\mu_2 = 6$, $\mu_3 = 7$, $\mu_4 = 10$, $\mu_5 = 12$ and $\mu_6 = \alpha_2 = 13$ in the situation depicted by Figure 1. The upshot of this algorithm is that the set of individuals that bear the same allele as the jth mutant μ_j corresponds precisely to the integer interval $[\mu_j, \mu_{j+1}[$. In this direction, it is therefore natural to introduce for every $j \in \mathbb{N}$ the jth *allelic cluster*

$$\mathbb{C}_j := (\xi_{\mu_j - 1 + \ell} : 1 \leq \ell \leq \mu_{j+1} - \mu_j),$$

that is, $\mathbb{C}_j$ is the finite sequence of the numbers of clone-children and mutant-children of the individuals bearing the same allele as the jth mutant. The sequence $(\mathbb{C}_j)_{j \in \mathbb{N}}$ encodes the allelic partition of the entire population.

REMARKS. 1. Each allelic cluster $\mathbb{C}_j$ is naturally endowed with a structure of rooted planar tree which is induced by the Galton–Watson process. More precisely, the latter is encoded via the usual depth-first search algorithm by the sequence $(\xi^{(c)}_{\mu_j - 1 + \ell} : 1 \leq \ell \leq \mu_{j+1} - \mu_j)$; in particular the jth mutant μ_j is viewed as the root (i.e., ancestor) of the cluster $\mathbb{C}_j$. In other words, the depth-first search algorithm with mutations for the Galton–Watson process induces precisely the usual depth-first search applied to the forest of allelic clusters viewed as a sequence of planar rooted trees.

2. We also stress that the initial Galton–Watson process can be recovered from the allelic partition $(\mathbb{C}_j)_{j \in \mathbb{N}}$. Indeed, the previous observation shows how to construct the portion of the genealogical tree corresponding to the allelic cluster generated by an initial mutant, and the latter also contains the information which is needed to identify the mutant-children of the first kind. Mutant-children of the first kind are the roots of the subtrees corresponding to the allelic clusters of the second kind, and by iteration the entire genealogical forest can be recovered.

Just as above, it is now convenient to introduce the downward skip-free random walk

$$S^{(c)}_n := \xi^{(c)}_1 + \cdots + \xi^{(c)}_n - n, \qquad n \in \mathbb{Z}_+, \tag{3}$$

and the passage times

$$T^{(c)}_j := \inf\{n \geq 0 : S^{(c)}_n = -j\}, \qquad j \in \mathbb{Z}_+. \tag{4}$$

Again, the $T^{(c)}_j$ form an increasing sequence of $(\mathcal{F}_i)$-stopping times.

LEMMA 3. *There is identity*

$$\mu_j - 1 = T^{(c)}_{j-1}$$

for every $j \in \mathbb{N}$. *As a consequence, for every* $j \in \mathbb{N}$, $\mathbb{C}_j$ *is adapted to the sigma-field* $\mathcal{F}_{T^{(c)}_j}$, *whereas* $\mathbb{C}_{j+1}$ *is independent of* $\mathcal{F}_{T^{(c)}_j}$ *and has the same distribution as* $\mathbb{C}_1$. *In particular the sequence of the allelic clusters* $\mathbb{C}_1, \mathbb{C}_2, \dots$ *is i.i.d.*

The proof is similar to that of Lemma 2 and therefore omitted.

We also introduce the number of alleles, that is, of different types, which are present in the ith tree $\mathbb{T}_i$:

$$A_i := \mathrm{Card}\{j \in \mathbb{N} : \mu_j \in [\alpha_i, \alpha_{i+1}[\};$$

for example, $A_1 = 5$ in the situation described by Figure 1. Note that there is the alternative expression

$$A_i = 1 + \sum_{\alpha_i \le \ell < \alpha_{i+1}} \xi^{(m)}_\ell.$$

COROLLARY 1. (i) *For every* $i \in \mathbb{Z}_+$, *we have*

$$\alpha_{i+1} = \mu_{A_1 + \cdots + A_i + 1};$$

equivalently, there is the identity

$$T^{(+)}_i = T^{(c)}_{A_1 + \cdots + A_i}.$$

(ii) *The allelic partition of the tree* $\mathbb{T}_i$, *which is induced by restricting the allelic partition of the entire population to* $\mathbb{T}_i$, *is given by*

$$(\mathbb{C}_{A_1 + \cdots + A_{i-1} + \ell} : 1 \le \ell \le A_i).$$

As a consequence, the sequence of the allelic partitions of the trees $\mathbb{T}_i$ *for* $i \in \mathbb{N}$, *is i.i.d.*

PROOF. (i) The first identity should be obvious from the definition of the depth-first search with mutations, as $A_1 + \cdots + A_i$ is the number of alleles which have been found after completing the exploration of the i first trees and the next mutant is then the $(i+1)$th ancestor. The second then follows from Lemmas 2 and 3.

(ii) The first assertion is immediately seen from (i) and the definitions of the trees and of the allelic clusters. Then observe that the number A_i of alleles in the tree $\mathbb{T}_i$ is a function of that tree, and so is the allelic partition. The second assertion thus derives from Lemma 2. □

It may be interesting to point out that $(T_i^{(+)}, i \geq 0)$ and $(T_j^{(\mathrm{c})}, j \geq 0)$ are both increasing random walks. The range

$$\mathcal{R}^{(+)} := \{T_i^{(+)} : i \geq 0\}$$

is the set of predecessors of ancestors (in the depth-first search algorithm with mutations), whereas

$$\mathcal{R}^{(\mathrm{c})} := \{T_j^{(\mathrm{c})} : j \geq 0\}$$

corresponds to predecessors of mutants. These are two regenerative subsets of $\mathbb{Z}_+$, in the sense that each can be viewed as the set of renewal epochs of some recurrent event (cf. Feller [21, 22]). Observe that both yield a partition of the set of positive integers into disjoint intervals:

$$\mathbb{N} = \bigcup_{i \geq 1}]T_{i-1}^{(+)}, T_i^{(+)}] = \bigcup_{j \geq 1}]T_{j-1}^{(\mathrm{c})}, T_j^{(\mathrm{c})}],$$

that correspond respectively to the trees in the Galton–Watson forest and to the allelic clusters. By Corollary 1(i), there is the embedding

$$\mathcal{R}^{(+)} \subseteq \mathcal{R}^{(\mathrm{c})}$$

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 $\mathcal{R}^{(+)} \circ \theta_k := \{i \geq 0 : k + i \in \mathcal{R}^{(+)}\}$ and $\mathcal{R}^{(\mathrm{c})} \circ \theta_k := \{j \geq 0 : k + j \in \mathcal{R}^{(\mathrm{c})}\}$ are independent of the sigma-field $\mathcal{F}_k$ generated by $(\xi_1, \ldots, \xi_k)$ and their joint law is the same as that of $(\mathcal{R}^{(+)}, \mathcal{R}^{(\mathrm{c})})$. We refer to [11] for applications of this notion. Roughly speaking, this implies that the allelic split of each interval $]T_{i-1}^{(+)}, T_i^{(+)}]$ produces smaller intervals $]T_{j-1}^{(\mathrm{c})}, T_j^{(\mathrm{c})}]$ in a random way that only depends on the length $T_i^{(+)} - T_{i-1}^{(+)}$ (i.e., the size of $\mathbb{T}_i$), independently of its location and of the other integer intervals. This can be thought of as a fragmentation property (see [13]) for the sizes of the trees.

2.4. *Allelic trees and forest.* In order to analyze the structure of allelic partitions, we introduce some related notions. The genealogy of the population model naturally induces a structure of forest on the set of different alleles. More precisely, we enumerate this set by declaring that the jth allele is that of the jth cluster $\mathbb{C}_j$, and define a planar graph on the set of alleles (which is thus identified as $\mathbb{N}$) by drawing an edge between two integers $j < k$ if and only if the parent of the kth mutant μ_k is an individual of the jth allelic cluster $\mathbb{C}_j$. This graph is clearly a forest (i.e., it contains no cycles), which we call the *allelic forest*, and more precisely the ith allelic tree is that induced by the mutant descent of the ith ancestor α_i. In other words, the ith allelic tree is the genealogical tree of the different alleles present in $\mathbb{T}_i$. In particular, the sequence of allelic trees is i.i.d. and their sizes are given by $(A_i, i \in \mathbb{N})$.

Recall that the *breadth-first search* in a forest consists in enumerating individuals in the lexicographic order of their labels, where the label of the nth individual at generation g is now given by the triplet (a, g, n), with a the rank of the ancestor at the initial generation. After a (short) moment of thought, we see that the definition of depth-first search with mutations for the Galton–Watson process ensures that the labeling of alleles by integers agrees with breadth-first search on the allelic forest, in the sense that the jth allele is found at the jth step of the breadth-first search on the allelic forest.

For every $j \in \mathbb{N}$, we consider the number of new mutants who are generated by the jth allelic cluster, viz.

$$M_j := \sum_{\mu_j \le \ell < \mu_{j+1}} \xi_\ell^{(\mathrm{m})}.$$

For instance, we have $M_1 = 3$, $M_4 = 1$ and $M_2 = M_3 = M_5 = 0$ in the situation depicted by Figures 1 and 2. The allelic forest is thus encoded by breadth-first search via the sequence $(M_j, j \in \mathbb{N})$.

LEMMA 4. *The sequence $(M_j, j \in \mathbb{N})$ is i.i.d., and therefore the allelic forest is a Galton–Watson forest with reproduction law the distribution of M_1. As a consequence, the size A_1 of the first allelic tree is given by the identity*

$$A_1 = \min\{j \ge 1 : M_1 + \cdots + M_j = j - 1\},$$

showing that A_1 is an $(\mathcal{F}_{T_j^{(\mathrm{c})}})$-stopping time.

PROOF. Recall from Lemma 3 that the sequence $\mathbb{C}_1, \mathbb{C}_2, \dots$ of the allelic clusters is i.i.d. Clearly, each variable M_j only depends on $\mathbb{C}_j$, which entails our first claim. The second follows from the well-known fact that breadth-first search induces a bijective transformation between the distributions of (sub-)critical Galton–Watson forests and those of i.i.d. sequences of integer-valued variables with mean less than or equal to one (see, e.g., Section 6.2 in [36]).

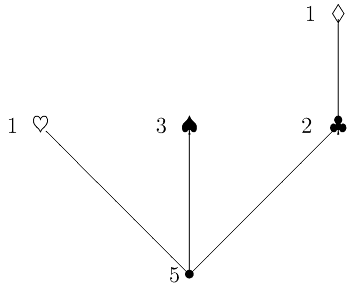

FIG. 2. *Allelic tree corresponding to the genealogical tree with mutations in Figure 1. The labels represent the sizes of the allelic clusters.*

Finally, the identity for the number A_1 of alleles present in the tree $\mathbb{T}_1$ follows from the preceding observations and again a variation of the celebrated formula of Dwass [19] (see, e.g., Lemma 2 in the present work), as plainly, A_1 coincides with the total size of the first tree in the allelic forest. □

3. 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 [40]. 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

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

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

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

and assume that

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

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

LEMMA 5 (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*

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

3.1. *Distribution of the allelic tree.* We have now introduced all the tools which are needed for describing some statistics of the allelic partition of a Galton–Watson tree with neutral mutations. We only need one more notation. We write

$$\pi_{k,\ell} = \mathbb{P}(\xi^{(\mathrm{c})} = k, \xi^{(\mathrm{m})} = \ell), \qquad k, \ell \in \mathbb{Z}_+, \tag{5}$$

for the probability function of the reproduction law of the Galton–Watson process with mutations. For every integer $n \ge 1$, we also write π^{*n} for the nth convolution product of that law, that is,

$$\pi^{*n}_{k,\ell} = \mathbb{P}(\xi^{(\mathrm{c})}_1 + \cdots + \xi^{(\mathrm{c})}_n = k, \xi^{(\mathrm{m})}_1 + \cdots + \xi^{(\mathrm{m})}_n = \ell).$$

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 forest is obtained by pruning or percolation on the genealogical forest of the Galton–Watson process, cutting each edge with probability p and independently of the other edges. See, for example, Aldous and Pitman [5] or Chapter 4 in Lyons and Peres [31]. Analytically, this means that if ξ is a random variable with law ϱ, then the conditional distribution of $(\xi^{(\mathrm{c})}, \xi^{(\mathrm{m})})$ 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

$$\pi^{*n}_{k,\ell} = \binom{k+\ell}{k} (1-p)^k p^\ell \varrho^{*n}_{k+\ell} \tag{6}$$

with ϱ^{*n} denoting the nth convolution power of ϱ. This expression is entirely explicit when ϱ is, for example, 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 $\xi^{(\mathrm{c})}$ and $\xi^{(\mathrm{m})}$ are assumed to be independent or when $\xi^{(\mathrm{c})} = \beta\xi$ and $\xi^{(\mathrm{m})} = (1-\beta)\xi$ where β stands for a Bernoulli variable which is independent of ξ.

Corollary 1 enables us to restrict our attention to the allelic partition of the tree generated by a typical ancestor, say for simplicity, Eve. Recall that $T_1^{(+)}$ denotes the size of the genealogical tree $\mathbb{T}_1$ of Eve, that A_1 is the number of alleles found in $\mathbb{T}_1$ and that the jth allelic cluster $\mathbb{C}_j$ generates M_j mutant-children. Further, we know from Lemma 4 that the first allelic tree is encoded by breadth-first search via the finite sequence $(M_j, 1 \le j \le A_1)$. The latter only retains partial information about the structure of the allelic partition of $\mathbb{T}_1$, and thus it is natural to enrich it by considering more generally the sequence of pairs $((|\mathbb{C}_j|, M_j), 1 \le j \le A_1)$, where

$$|\mathbb{C}_j| := \mu_{j+1} - \mu_j$$

denotes the size of the jth allelic cluster, that is, the number of individuals having the jth type. In other words, we enrich the allelic tree by assigning to each allele the size of the corresponding allelic cluster. We may now state our main result, which can be viewed as a generalization of a celebrated identity due to Dwass [19].

THEOREM 1. (i) *The joint law of the size of* $\mathbb{T}_1$ *and its number of alleles is given by*

$$\mathbb{P}(T_1^{(+)} = n, A_1 = k) = \frac{1}{n}\pi^{*n}_{n-k,k-1}, \qquad 1 \le k \le n.$$

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

$$\mathbb{P}(|\mathbb{C}_1| = n, M_1 = \ell) = \frac{1}{n}\pi^{*n}_{n-1,\ell}, \qquad n \ge 1 \textit{ and } \ell \ge 0.$$

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

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

we have

$$\mathbb{P}(|\mathbb{C}_1| = n_1, M_1 = \ell_1, \dots, |\mathbb{C}_k| = n_k, M_k = \ell_k) = \prod_{i=1}^{k} \frac{1}{n_i} \pi^{*n_i}_{n_i - 1, \ell_i}.$$

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

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

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

2. In particular, the law of the number A_1 of alleles is given by

$$\mathbb{P}(A_1 = k) = \sum_{n=1}^{\infty} n^{-1} \pi^{*n}_{n-k, k-1}, \qquad k \geq 0.$$

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

$$\mathbb{P}(A_1 = k) = \frac{1}{k} \nu^{*k}_{k-1},$$

where

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

and ν^{*k} the kth convolution power of ν. Indeed, this alternative formulation is seen from Lemma 4 and Dwass formula [19].

PROOF. Recall that $(\xi_1^{(\mathrm{c})}, \xi_1^{(\mathrm{m})}), \dots, (\xi_n^{(\mathrm{c})}, \xi_n^{(\mathrm{m})})$ is a sequence of n i.i.d. copies of $(\xi^{(\mathrm{c})}, \xi^{(\mathrm{m})})$ and consider the partial sums of coordinates

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

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

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

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

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

Plainly, there is the identity

$$\{T_1^{(+)} = n, A_1 = k\} = \Lambda_{n-k,k-1} \cap \{\min\{j \geq 1 : \Sigma_j = j-1\} = n\}$$

as, according to Lemma 2,

$$\min\{j \geq 1 : \Sigma_j = j-1\} = \min\{j \geq 1 : S_j^{(+)} = -1\} = T_1^{(+)}.$$

By the ballot theorem [take $f(x^{(\mathrm{c})}, x^{(\mathrm{m})}) = x^{(\mathrm{c})} + x^{(\mathrm{m})} - 1$ in Lemma 5], we have

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

which yields (i).

The proof of (ii) is similar, observing that

$$\{|\mathbb{C}_1| = n, M_1 = \ell\} = \Lambda_{n-1,\ell} \cap \{\min\{j \geq 1 : \Sigma_j^{(\mathrm{c})} = j-1\} = n\}.$$

Finally (iii) follows by iteration from (ii) and the fact that conditionally on $A_1 \geq j+1$, the $(j+1)$th allelic cluster $\mathbb{C}_{j+1}$ is independent of $(\mathbb{C}_k, 1 \leq k \leq j)$ and has the same distribution as the Eve cluster $\mathbb{C}_1$ (see Lemma 3). □

3.2. *Conditioning on the population size and the number of alleles.* In the rest of this section, we will be interested in the relative sizes of clusters in the allelic partition of the first tree $\mathbb{T}_1$, ignoring their connections. We start with a description 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 2. *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 collection of positive integers $n_1, \dots, n_k$ with $n_1 + \cdots + n_k = n$, we have*

$$\mathbb{P}(|\mathbb{C}_{\sigma(1)}| = n_1, \dots, |\mathbb{C}_{\sigma(k)}| = n_k \mid T_1^{(+)} = n, A_1 = k)$$
$$= \frac{n}{k\pi^{*n}_{n-k,k-1}} \sum \prod_{i=1}^{k} \frac{1}{n_i} \pi^{*n_i}_{n_i-1,\ell_i},$$

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

PROOF. A classical application of the ballot theorem shows that the conditional distribution of $(\mathbb{C}_{\sigma(1)}, \dots, \mathbb{C}_{\sigma(k)})$ given $T_1^{(+)} = n$ and $A_1 = k$ is the same as that of $(\mathbb{C}_1, \dots, \mathbb{C}_k)$ conditioned on $\sum_{i=1}^k |\mathbb{C}_i| = n$ and $\sum_{i=1}^k M_i = k-1$. Then note that

$$\sum_{i=1}^k |\mathbb{C}_i| = n \quad \text{and} \quad \sum_{i=1}^k M_i = k-1 \quad \Longleftrightarrow \quad T_k^{(\mathrm{c})} = n \quad \text{and} \quad \sum_{i=1}^n \xi_i^{(\mathrm{m})} = k-1$$

and an application of the ballot theorem (much in the same way as in the proof of Theorem 1) shows that the probability of that event equals

$$\frac{k}{n}\pi^{*n}_{n-k,k-1}.$$

Theorem 1(ii) completes the proof. □

Next, we normalize the size $|\mathbb{C}_i|$ of each cluster by the size $T_1^{(+)}$ of the total population (recall we focus on the descent of a single ancestor, namely Eve), and write

$$\Gamma_1 \geq \Gamma_2 \geq \cdots \geq \Gamma_{A_1}$$

for the sequence which is obtained by ranking the ratios $|\mathbb{C}_i|/T_1^{(+)}$ in the decreasing order. So $\Gamma = (\Gamma_1, \dots, \Gamma_{A_1})$ 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 supremum distance, which yields a compact metric space; see Section 2.1 in [13] 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 tree $\mathbb{T}_1$ and the number $A_1 = k$ of alleles, when $n, k \to \infty$. 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 $\mathrm{a}_1 > \mathrm{a}_2 > \cdots > 0$ of the atoms ranked in the decreasing order of a Poisson point measure on $]0, \infty[$ with intensity $ba^{-3/2}\,da$. Roughly speaking, we then get a random proper mass-partition by conditioning on $\sum_{i=1}^{\infty} \mathrm{a}_i = 1$; see, for example, [35] or Proposition 2.4 in [13] 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 [6], more precisely it arose by logging the Continuum Random Tree according to Poissonian cuts along its skeleton; see also [3, 7, 12, 32] for related works. In the present setting, we may interpret such cuts as mutations which induce an allelic partition. As we know from Aldous [4] that the Continuum Random Tree can be viewed as the limit when $n \to \infty$ of Galton–Watson trees conditioned to have total size n, the fact that the preceding random mass-partitions appear again in the framework of this work should not come as a surprise.

For the sake of simplicity, we shall focus on the case when the number of clone-children $\xi^{(\mathrm{c})}$ and the number of mutant-children $\xi^{(\mathrm{m})}$ 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 $\mathbb{E}(\xi^{(\mathrm{c})}) < 1$. We shall work under the hypothesis that by a suitable exponential tilting, this subcritical 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

$$\mathbb{E}(\xi^{(\mathrm{c})}\theta^{\xi^{(\mathrm{c})}}) = \mathbb{E}(\theta^{\xi^{(\mathrm{c})}}) \quad \text{and} \quad \sigma_\theta^2 := \mathbb{E}((\xi^{(\mathrm{c})})^2\theta^{\xi^{(\mathrm{c})}})/\mathbb{E}(\theta^{\xi^{(\mathrm{c})}}) - 1 < \infty. \tag{7}$$

It can be readily checked that (7) then specifies θ uniquely.

PROPOSITION 1. *Suppose that $\xi^{(\mathrm{c})}$ and $\xi^{(\mathrm{m})}$ are independent, that neither distribution is supported by a strict subgroup of $\mathbb{Z}$ and that* (7) *holds. Fix $b > 0$ and let $n, k \to \infty$ 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 alleles $A_1 = k$ converges weakly on the space of mass-partitions to the sequence* $(\mathrm{a}_1, \mathrm{a}_2, \dots)$ *of the atoms of a Poisson random measure on* $]0, \infty[$ *with intensity*

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

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

REMARK. The special case when $\xi^{(\mathrm{c})}$ and $\xi^{(\mathrm{m})}$ are two independent Poisson variables, say with rates $r^{(\mathrm{c})}$ and $r^{(\mathrm{m})}$ can also be viewed as an instance of the situation where mutations affect children independently with probability $p = r^{(\mathrm{m})}/(r^{(\mathrm{c})} + r^{(\mathrm{m})})$ (cf. the example discussed before Theorem 1). More precisely the reproduction law of the standard Galton–Watson process is then Poisson with rate $r^{(\mathrm{c})} + r^{(\mathrm{m})}$. 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 replacements 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 $\xi_1^{(\mathrm{c})}, \dots$ are i.i.d. under $\tilde{\mathbb{P}}$ with law given by

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

where z_θ is the normalization factor, namely,

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

As in the proof of Corollary 2, we see from an application of the ballot theorem that the conditional distribution of $(n\Gamma_1, \dots, n\Gamma_{A_1})$ given $T_1^{(+)} = n$ and $A_1 = k$ is the same as that obtained from the i.i.d. sequence $|\mathbb{C}_1|, \dots, |\mathbb{C}_k|$ by ranking in the decreasing order and conditioning on $\sum_{i=1}^k |\mathbb{C}_i| = n$ and $\sum_{i=1}^k M_i = k - 1$. Observe that the latter is equivalent to conditioning on $\sum_{i=1}^k |\mathbb{C}_i| = n$ and $\sum_{i=1}^n \xi_i^{(\mathrm{m})} = k - 1$. Further, recall from Lemma 3 that $|\mathbb{C}_j| = T_j^{(\mathrm{c})} - T_{j-1}^{(\mathrm{c})}$ and hence, on this event, the variables $|\mathbb{C}_1|, \dots, |\mathbb{C}_k|$ are functions of $\xi_1^{(c)}, \dots, \xi_n^{(c)}$. 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 a conditional law, in the sense that the sequence $|\mathbb{C}_1|, \dots, |\mathbb{C}_k|$ has the same distribution under $\mathbb{P}(\cdot \mid T_1^{(+)} = n)$ as under $\tilde{\mathbb{P}}(\cdot \mid T_1^{(+)} = n)$.

We then estimate the distribution of the size of the Eve cluster under $\tilde{\mathbb{P}}$, which is given again according to the Dwass formula [19] by

$$\tilde{\mathbb{P}}(|\mathbb{C}_1| = n_1) = \frac{1}{n_1}\tilde{\mathbb{P}}(\xi_1^{(c)} + \cdots + \xi_{n_1}^{(c)} = n_1 - 1) = \frac{1}{n_1}\tilde{\mathbb{P}}(S_{n_1}^{(c)} = -1).$$

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

$$\tilde{\mathbb{P}}(|\mathbb{C}_1| = n_1) \sim \frac{1}{\sqrt{2\pi\sigma_\theta^2 n_1^3}} \qquad \text{as } n_1 \to \infty.$$

Putting the pieces together, we get that the conditional distribution of $(n\Gamma_1, \dots, n\Gamma_{A_1})$ given $T_1^{(+)} = n$ and $A_1 = k$ 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

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

An application of Corollary 2.2 in [13] completes the proof of our claim. □

4. 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., downward 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 [17] 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 $\xi^{(+)}(n) := \xi^{(c)}(n) + \xi^{(m)}(n) - 1$ and $\xi^{(c)}(n) - 1$, respectively. Let also $X = (X_t, t \in \mathbb{R}_+)$ denote a Lévy process with no negative jumps and Laplace exponent ψ, namely,

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

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

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

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 $(\mathbb{T}_\ell, \ell \in \mathbb{N})$. That is, for $i \geq 0$, $H_i(n)$ denotes the generation of the $(i+1)$th individual found by the usual depth-first search (i.e., mutations are discarded) on the Galton–Watson forest. 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 [17] for precise definitions and further references.

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

$$\lim_{n\to\infty} n^{-1} S^{(+)}_{n\gamma_n}(n) = X_1 \qquad \text{in law;} \tag{8}$$

we also assume that the technical condition (2.27) in [17] is fulfilled. Then the rescaled height function

$$(\gamma_n^{-1} H_{[tn\gamma_n]}(n) : t \geq 0)$$

converges in distribution, in the sense of weak convergence on Skorohod space $\mathbb{D}(\mathbb{R}_+, \mathbb{R}_+)$ as $n \to \infty$ toward the height process $(\mathcal{H}_t : t \geq 0)$ which is constructed from the Lévy process $X = (X_t, t \geq 0)$; see Theorem 2.3.1 in [17].

Similarly, we write $H^{(\mathrm{c})}(n) = (H_i^{(\mathrm{c})}(n), i \in \mathbb{N})$ for the height function of the Galton–Watson forest $(\mathbb{C}_j, j \in \mathbb{N})$, where each allelic cluster $\mathbb{C}_j$ is endowed with the genealogical tree structure induced by the population model (see Remark, item 1 in Section 2.3).

PROPOSITION 2. *Suppose that the preceding assumptions hold, and also that*

$$\lim_{n\to\infty} \gamma_n \mathbb{E}(\xi^{(\mathrm{m})}(n)) = d \quad \textit{and} \quad \lim_{n\to\infty} n^{-1}\gamma_n \operatorname{Var}(\xi^{(\mathrm{m})}(n)) = 0 \tag{9}$$

for some $d \geq 0$. Then the rescaled height function

$$(\gamma_n^{-1} H^{(\mathrm{c})}_{[tn\gamma_n]}(n) : t \geq 0)$$

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

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

which is constructed from the Lévy process $X^{(d)} = (X_t^{(d)} := X_t - dt, t \geq 0)$.

REMARK. More recently, Duquesne and Le Gall [18] (see also the survey [29]) 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. Proposition 2 can also be restated in this setting.

PROOF OF PROPOSITION 2. The assumption (8) ensures the convergence in distribution

$$\big(n^{-1}S^{(+)}_{[tn\gamma_n]}(n): t \geq 0\big) \Longrightarrow (X_t : t \geq 0),$$

see Theorem 2.1.1 in [17] and (2.3) there. On the other hand, by a routine argument based on martingales, the assumption (9) entails that

$$\lim_{n\to\infty} n^{-1}\big(S^{(+)}_{[tn\gamma_n]}(n) - S^{(\mathrm{c})}_{[tn\gamma_n]}(n)\big) = dt,$$

uniformly for t in compact intervals, in $L^2(\mathbb{P})$. The convergence in distribution

$$\big(n^{-1}S^{(\mathrm{c})}_{[tn\gamma_n]}(n): t \geq 0\big) \Longrightarrow (X_t - dt : t \geq 0)$$

follows. Recall that depth-first search with mutations on the initial forest yields the usual depth-first search for the forest of allelic clusters (cf. Remark, item 1 in Section 2.3). We can then complete the proof as in Theorem 2.3.1 in [17]. □

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

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

is the same for all n. For the sake of simplicity, we assume also that $\mathbb{E}(\xi) = 1$ and $\mathrm{Var}(\xi) = 1$. We may then take $\gamma_n = n$, so by the central limit theorem, (8) 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 $\xi^{(+)}(n) := \xi^{(\mathrm{c})}(n) + \xi^{(\mathrm{m})}(n) = k$, the number $\xi^{(\mathrm{m})}(n)$ of mutant-children of a typical individual has the binomial distribution $B(k, d/n)$. In other words, in the nth population model, mutations affect each child with probability d/n, independently of the other children. Then (9) clearly holds. Roughly speaking, Theorem 2.3.1 of [17] implies in this setting that the initial Galton–Watson forest associated with the nth population model, converges in law after a suitable renormalization to the Brownian forest, whereas Proposition 2 of the present work shows that the allelic forest renormalized in the same way, converges in law to the forest generated by a Brownian motion 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 [6] by logging the Continuum Random Tree according to a Poisson point process on its skeleton, and the fragmentation process constructed in [12] 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 [38] already pointed at several applications of the (continuous) ballot theorem in this framework. More generally, the transformation $X \to X^{(d)}$ of Lévy processes with no negative jumps also appeared in an article by Miermont [32] on certain eternal additive coalescents, whereas Aldous and Pitman [7] showed that the latter arise asymptotically from independent pruning of certain sequences of birthday trees. Finally, we also refer [2] for another interesting recent work on pruning Lévy random trees.

Acknowledgment. I would like to thank two anonymous referees for their careful check of this work.

REFERENCES

[1] ABRAHAM, R. and DELMAS, J.-F. (2008). Williams' decomposition of the Lévy continuous random tree and simultaneous extinction probability for populations with neutral mutations. Preprint. Available at http://hal.archives-ouvertes.fr/.

[2] ABRAHAM, R., DELMAS, J.-F. and VOISIN, G. (2009). Pruning a Lévy continuum random tree. *Stochastic Process. Appl.* To appear.

[3] ABRAHAM, R. and SERLET, L. (2002). Poisson snake and fragmentation. *Electron. J. Probab.* **7** 15 (electronic). MR1943890

[4] ALDOUS, D. (1993). The continuum random tree. III. *Ann. Probab.* **21** 248–289. MR1207226

[5] ALDOUS, D. and PITMAN, J. (1998). Tree-valued Markov chains derived from Galton–Watson processes. *Ann. Inst. H. Poincaré Probab. Statist.* **34** 637–686. MR1641670

[6] ALDOUS, D. and PITMAN, J. (1998). The standard additive coalescent. *Ann. Probab.* **26** 1703–1726. MR1675063

[7] ALDOUS, D. and PITMAN, J. (2000). Inhomogeneous continuum random trees and the entrance boundary of the additive coalescent. *Probab. Theory Related Fields* **118** 455–482. MR1808372

[8] ATHREYA, K. B. and NEY, P. E. (1972). *Branching Processes*. Springer, New York. MR0373040

[9] BASDEVANT, A.-L. and GOLDSCHMIDT, C. (2008). Asymptotics of the allele frequency spectrum associated with the Bolthausen–Sznitman coalescent. *Electron. J. Probab.* **13** 486–512. MR2386740

[10] BERESTYCKI, J., BERESTYCKI, N. and SCHWEINSBERG, J. (2007). Beta-coalescents and continuous stable random trees. *Ann. Probab.* **35** 1835–1887. MR2349577

[11] BERTOIN, J. (1999). Renewal theory for embedded regenerative sets. *Ann. Probab.* **27** 1523–1535. MR1733158

[12] BERTOIN, J. (2000). A fragmentation process connected to Brownian motion. *Probab. Theory Related Fields* **117** 289–301. MR1771665

[13] BERTOIN, J. (2006). *Random Fragmentation and Coagulation Processes*. *Cambridge Studies in Advanced Mathematics* **102**. Cambridge Univ. Press, Cambridge. MR2253162

[14] CRUMP, K. S. and GILLESPIE, J. H. (1976). The dispersion of a neutral allele considered as a branching process. *J. Appl. Probab.* **13** 208–218. MR0408877

[15] DELMAS, J.-F., DHERSIN, J.-S. and SIRI-JEGOUSSE, A. (2008). Asymptotic results on the length of coalescent trees. *Ann. Appl. Probab.* **18** 997–1025. MR2418236

[16] DONG, R., GNEDIN, A. and PITMAN, J. (2007). Exchangeable partitions derived from Markovian coalescents. *Ann. Appl. Probab.* **17** 1172–1201. MR2344303

[17] DUQUESNE, T. and LE GALL, J.-F. (2002). Random trees, Lévy processes and spatial branching processes. *Astérisque* **281** vi–147. MR1954248

[18] DUQUESNE, T. and LE GALL, J.-F. (2005). Probabilistic and fractal aspects of Lévy trees. *Probab. Theory Related Fields* **131** 553–603. MR2147221

[19] DWASS, M. (1969). The total progeny in a branching process and a related random walk. *J. Appl. Probab.* **6** 682–686. MR0253433

[20] EWENS, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Population Biology* **3** 87–112. MR0325177

[21] FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications* **I**. Wiley, New York. MR0228020

[22] FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **II**. Wiley, New York. MR0270403

[23] GRIFFITHS, R. C. and PAKES, A. G. (1988). An infinite-alleles version of the simple branching process. *Adv. in Appl. Probab.* **20** 489–524. MR955502

[24] HARRIS, T. E. (1952). First passage and recurrence distributions. *Trans. Amer. Math. Soc.* **73** 471–486. MR0052057

[25] HARRIS, T. E. (1963). *The Theory of Branching Processes*. Springer, Berlin. MR0163361

[26] KIMMEL, M. and AXELROD, D. E. (2002). *Branching Processes in Biology. Interdisciplinary Applied Mathematics* **19**. Springer, New York. MR1903571

[27] KINGMAN, J. F. C. (1980). *Mathematics of Genetic Diversity. CBMS-NSF Regional Conference Series in Applied Mathematics* **34**. SIAM, Philadelphia, PA. MR591166

[28] LAMBERT, A. (2008). Spine decompositions and allelic partitions of splitting trees. In preparation.

[29] LE GALL, J.-F. (2005). Random trees and applications. *Probab. Surv.* **2** 245–311 (electronic). MR2203728

[30] LIGGETT, T. M., SCHINAZI, R. B. and SCHWEINSBERG, J. (2008). A contact process with mutations on a tree. *Stochastic Process. Appl.* **118** 319–332. MR2389047

[31] LYONS, R. and PERES, Y. (2008). Probability on trees and networks. Available at http://php.indiana.edu/~rdlyons/prbtree/book.pdf.

[32] MIERMONT, G. (2001). Ordered additive coalescent and fragmentations associated to Lévy processes with no positive jumps. *Electron. J. Probab.* **6** 33 (electronic). MR1844511

[33] MÖHLE, M. (2006). On sampling distributions for coalescent processes with simultaneous multiple collisions. *Bernoulli* **12** 35–53. MR2202319

[34] NERMAN, O. (1987). Branching processes and neutral mutations. In *Proceedings of the First World Congress of the Bernoulli Society, Vol.* 2 (*Tashkent*, 1986) 683–692. VNU Sci. Press, Utrecht. MR1092502

[35] PERMAN, M., PITMAN, J. and YOR, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields* **92** 21–39. MR1156448

[36] PITMAN, J. (2006). *Combinatorial Stochastic Processes. Lecture Notes in Math.* **1875**. Springer, Berlin. MR2245368

[37] SCHINAZI, R. B. and SCHWEINSBERG, J. (2008). Spatial and nonspatial stochastic models for immune response. *Markov Process. Related Fields* **14** 255–276. MR2437531

[38] SCHWEINSBERG, J. (2001). Applications of the continuous-time ballot theorem to Brownian motion and related processes. *Stochastic Process. Appl.* **95** 151–176. MR1847096

[39] TAÏB, Z. (1992). *Branching Processes and Neutral Evolution. Lecture Notes in Biomathematics* **93**. Springer, Berlin. MR1176317

[40] TAKÁCS, L. (1967). *Combinatorial Methods in the Theory of Stochastic Processes*. Wiley, New York. MR0217858

LABORATOIRE DE PROBABILITÉS
UNIVERSITÉ PIERRE ET MARIE CURIE
175, RUE DU CHEVALERET
F-75013 PARIS
FRANCE
AND
DMA
ECOLE NORMALE SUPÉRIEURE
PARIS
FRANCE
E-MAIL: jean.bertoin@upmc.fr