ARTICLE IN PRESS

Theoretical Population Biology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

The coalescent of a sample from a binary branching process

Amaury Lambert [*]

Laboratoire de Probabilités, Statistique & Modélisation (LPSM), Sorbonne Université, CNRS, Paris, France
Center for Interdisciplinary Research in Biology (CIRB), Collège de France, CNRS, INSERM, PSL Research University, Paris, France

ARTICLE INFO

Article history:
Available online xxxx

In memoriam Paul Joyce

MSC:
primary 60J80
secondary 60-06
60J85
60G09
92D25
92D10
92D15

Keywords:
Splitting tree
Random tree
Birth–death process
Incomplete sampling
Coalescent point process
Finite exchangeable sequence.

ABSTRACT

At time 0, start a time-continuous binary branching process, where particles give birth to a single particle independently (at a possibly time-dependent rate) and die independently (at a possibly time-dependent and age-dependent rate). A particular case is the classical birth–death process. Stop this process at time $T > 0$. It is known that the tree spanned by the N tips alive at time T of the tree thus obtained (called a reduced tree or coalescent tree) is a coalescent point process (CPP), which basically means that the depths of interior nodes are independent and identically distributed (iid). Now select each of the N tips independently with probability y (Bernoulli sample). It is known that the tree generated by the selected tips, which we will call the Bernoulli sampled CPP, is again a CPP. Now instead, select exactly k tips uniformly at random among the N tips (a k-sample). We show that the tree generated by the selected tips is a mixture of Bernoulli sampled CPPs with the same parent CPP, over some explicit distribution of the sampling probability y. An immediate consequence is that the genealogy of a k-sample can be obtained by the realization of k random variables, first the random sampling probability Y and then the $k-1$ node depths which are iid conditional on $Y = y$.

1. Introduction

1.1. Model and objective of the paper

In this work, we consider a binary branching process in continuous time, possibly non-Markovian, that has the following properties, further denoted ($\star$).

- The process starts with one particle at time 0;
- At any time t, particles give birth independently at rate $\lambda(t)$, to a single daughter particle at each birth event;
- At any time t, particles with age x independently die at rate $\mu(t,x)$;
- The process is stopped at time $T > 0$ and is conditioned to have $N \geq 1$ particles alive at time T.

This process generates a discrete metric tree, called a *splitting tree* (Geiger and Kersting, 1997; Lambert, 2010), with origin at time 0 and N tips at distance T from the root, that we call *extant tips*. The inherent asymmetry between mother and daughter endows the splitting tree with a natural *plane orientation*, where daughters sprout to the right of their mother; see Fig. 1a.

An oriented, ultrametric tree with N tips is characterized by its *node depths*, or *coalescence times*, $H_1, \ldots, H_{N-1}$, as in Fig. 1b. The orientation of the tree implies that H_i ($1 \leq i \leq n-1$) is the coalescence time between extant tip $i-1$ and extant tip i, where tips are labelled $0, \ldots, n-1$ from left to right in the plane orientation, and also that $\max\{H_{i+1}, \ldots, H_j\}$ is the coalescence time between tip i and tip j.

When $\mu(t,x)$ does not depend on age x, the branching process is merely a birth–death process with *per capita* birth rate $\lambda(t)$ and death rate $\mu(t)$. In this case, it is actually equivalent to select uniformly at each birth event which lineage is the mother and which lineage is the daughter.

We are interested in the so-called *reduced tree*, also called *coalescent tree* in population genetics and *reconstructed tree* in phylogenetics, i.e., the tree generated by the extant tips of the splitting tree, which is the genealogy of the particles alive at time T. This tree is said to be *ultrametric* with height T, in the sense that all its tips are at the same distance T to the root (i.e., all its tips are extant tips), so that the metric induced by the tree metric on its tip set is ultrametric (see e.g. Lambert (2017, 2018)).

[*] Correspondence to: Laboratoire de Probabilités, Statistique & Modélisation (LPSM), Sorbonne Université, 4 place Jussieu, 75252 Paris Cedex 05, France.
E-mail address: amaury.lambert@upmc.fr.
URL: http://www.lpsm.paris/pageperso/amaury.lambert/.

https://doi.org/10.1016/j.tpb.2018.04.005
0040-5809/© 2018 Published by Elsevier Inc.

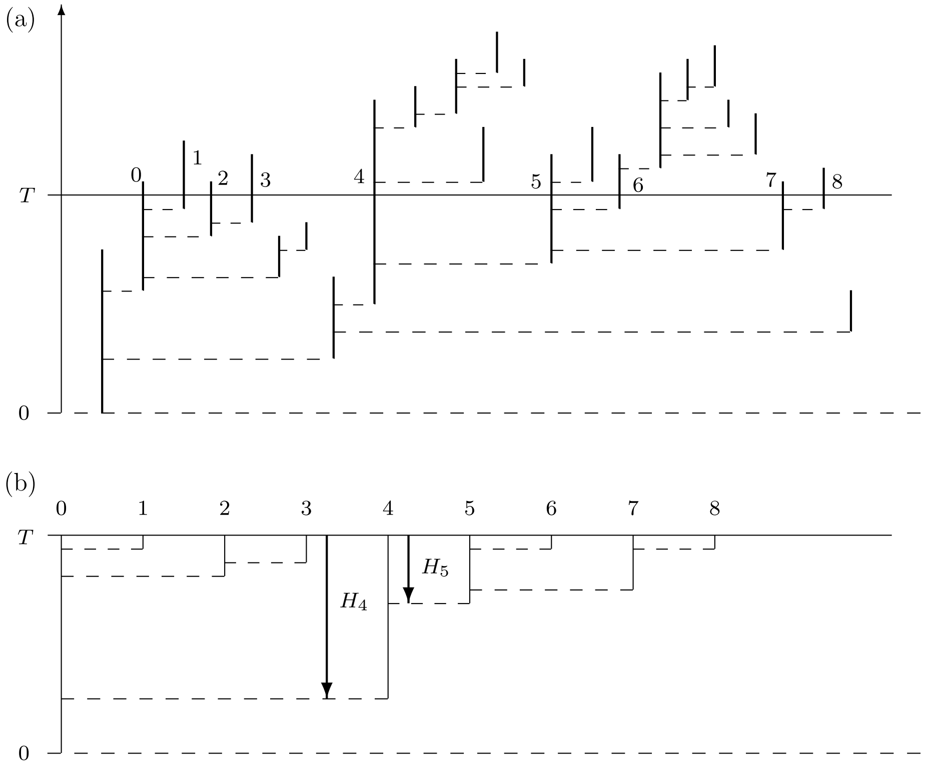

Fig. 1. (a) A plane oriented tree generated by a branching process; the $N = 9$ particles extant at T are labelled $0, 1, \dots, 8$ from left to right; (b) The reduced tree obtained from the full tree in (a) by throwing away all subtrees extinct by time T or starting after T. The figure shows the coalescence times H_4, between extant tips 3 and 4, and H_5 between extant tips 4 and 5. The structure of the reduced tree is characterized by the $N - 1 = 8$ coalescence times H_i, which are iid under our assumptions ($\star$).

More specifically, we are interested in the tree generated by a sample of the extant tips of the splitting tree, or equivalently, in the tree generated by a sample from (the tips of) the reduced tree. In the biology literature, there are mainly two classical sampling schemes (but other sampling schemes can be useful, like diversified sampling or higher-level sampling, see Lambert and Stadler (2013)). The first scheme, called the *Bernoulli sampling* scheme, consists of selecting each extant tip independently with the same probability, say y. The second scheme, called *k-sampling* scheme, consists in drawing uniformly k tips among the extant tips of the splitting tree conditioned upon $N \geq k$. The goal of the paper is to gather some known results about Bernoulli samples and to present new results for the genealogy of a k-sample, including an explicit de Finetti representation of node depths, that is, as a mixture of sequences of independent and identically distributed (iid) random variables (Aldous, 1985; Finetti, 1931; Hewitt and Savage, 1955).

1.2. Coalescent point processes

A coalescent point process (CPP) with height T is a random, oriented ultrametric tree with height T, whose node depths $H_1, \dots, H_{N-1}$ form a sequence of independent copies of some random variable (rv) $H > 0$, stopped at its first value larger than T. Throughout the paper, we will assume that H has a density denoted f and we will use the notation

$$F(t) := \frac{1}{P(H > t)},$$

so that

$$f = \frac{F'}{F^2},$$

and we will say that the CPP has *inverse tail distribution* F. A consequence is that the number (again denoted) N of extant tips in a CPP is always a shifted geometric rv, namely $P(N = n) = (1 - a)a^{n-1}$, where

$$a := P(H < T).$$

Now the likelihood of an ultrametric tree τ with n tips and node depths $x_1 < \cdots < x_{n-1}$ under the CPP distribution is simply

$$\mathcal{L}(\tau) = \frac{C(\tau)}{F(T)} \prod_{i=1}^{n-1} f(x_i), \tag{1}$$

where $C(\tau)$ is a constant that depends whether τ is oriented or not (Lambert, 2017; Lambert and Stadler, 2013). Specifically, $C(\tau) = 1$ if τ is oriented and $C(\tau) = 2^{n-1-\alpha(\tau)}$ if τ is non-oriented, where $\alpha(\tau)$ is the number of cherries of τ (a cherry is a pair of tips which are the only tips descending from their most recent common ancestor in τ). Notice that the likelihood of τ conditional on $N = n$ is obtained by dividing $\mathcal{L}(\tau)$ by $P(H < T)^{n-1}P(H > T)$.

In Lambert (2010) and Lambert and Stadler (2013), it was shown that the reduced tree of a splitting tree is a CPP with inverse tail distribution F that can be characterized from the knowledge of the rates λ and μ, as in the following statement which merges Theorem 3 and Proposition 4 from Lambert and Stadler (2013) (see also Hallinan (2012), Höhna (2013), Nee et al. (1994) and Popovic (2004) for similar, but partial results). We first need to define for any $s \geq t$, the density $g(t, s)$ at time s of the death time of a particle born at time t. Elementary properties of Poisson processes entail the following formula:

$$g(t, s) = \mu(s, s - t) \exp\left\{-\int_t^s \mu(u, u - t)\, du\right\}. \tag{2}$$

Theorem 1. *The reduced tree at height T of a splitting tree satisfying the properties ($\star$) stated in the introduction is a CPP whose inverse*