Adv. Appl. Prob. **37,** 1094–1115 (2005)
Printed in Northern Ireland
© *Applied Probability Trust* 2005

A CRITICAL BRANCHING PROCESS MODEL FOR BIODIVERSITY

DAVID ALDOUS,* *University of California, Berkeley*
LEA POPOVIC,** *University of Minnesota*

Abstract

We study the following model for a phylogenetic tree on n extant species: the origin of the clade is a random time in the past whose (improper) distribution is uniform on $(0, \infty)$; thereafter, the process of extinctions and speciations is a continuous-time critical branching process of constant rate, conditioned on there being the prescribed number n of species at the present time. We study various mathematical properties of this model as $n \to \infty$: namely the time of origin and of the most recent common ancestor, the pattern of divergence times within lineage trees, the time series of the number of species, the total number of extinct species, the total number of species ancestral to the extant ones, and the 'local' structure of the tree itself. We emphasize several mathematical techniques: the association of walks with trees; a point process representation of lineage trees; and Brownian limits.

Keywords: Biodiversity; Brownian excursion; contour process; critical branching process; genealogy; local weak convergence; phylogenetic tree; point process

2000 Mathematics Subject Classification: Primary 60J85
Secondary 60J65; 92D15

1. Introduction

There is a substantial literature on comparing data on different aspects of *biodiversity* or *macroevolution* with the predictions of stochastic models. Available data range from time series for the number of species to shapes of phylogenetic trees on extant species. While data-motivated models are scientifically natural, a mathematical aesthetic suggests the following approach: start with a stochastic model that serves as a 'null hypothesis' model, and compare the properties of this pure-chance model with those derived from the data. Our focus will be on formulating a model that does not incorporate specific conjectured biological hypotheses, and on studying its mathematical properties. The use of such a model does not mean we believe macroevolution really did proceed according to this particular model; rather, the ultimate goal is to assess in what systematic way real phylogenetic trees differ from the predictions of a pure-chance model. The model (Section 2) is *neutral* in the sense that speciations and extinctions occur with equal probability. Placing a uniform prior on the time of origin of the process, and conditioning to have a given number n of extant species, gives a model $\mathcal{T}_n$ for the past macroevolution of a clade with n extant species that could in principle be compared with real

Received 27 October 2004; revision received 23 June 2005.

* Postal address: Department of Statistics, University of California, 367 Evans Hall #3860, Berkeley, CA 94720-3860, USA. Email address: aldous@stat.berkeley.edu

** Current address: Department of Mathematics, 310 Malott Hall, Cornell University, Ithaca, NY 14853-4201, USA.

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

data for such clades. Our results describe distributional properties of various aspects of the tree $\mathcal{T}_n$.

- The lineage tree, via exact formulae (Proposition 1), ‘global limits’ (Corollaries 1 and 2), and ‘local limits’ (Corollaries 4 and 5).
- The time series of the number of species (Lemma 2), the maximum number of coexisting species (Corollary 6), and the total number of extinct species (Corollary 7).
- The local limit structure of the complete phylogenetic tree (i.e. including extinct species), relative to either a typical extant species (Proposition 4) or a typical extinct species (Proposition 3).
- The joint distribution of the time of origin of the clade and the time of the most recent common ancestor (Corollary 3), joint also with the number of species extant at the time of the most recent common ancestor (Corollary 8).
- The number of extinct species ancestral to some extant species (Corollary 10).

Implicit in several formulae is that the model has high variability – several features may vary wildly from one realization to another – and this suggests caution when arguing that some given real-world phylogenetic tree could not have arisen by a chance process. We should admit that the whole paradigm of studying $n \to \infty$ asymptotics is rather unnatural, because the model is biologically unrealistic for large n, but we can hope that the approximations implicit in asymptotic results are qualitatively correct for smaller values of n. The authors' website (www.stat.berkeley.edu/users/aldous/Research/Phylo/index.html) shows Monte Carlo simulations for $n = 8, 12, 20$ with ten repetitions; these verify that numerical values are broadly consistent with the asymptotic predictions, and vividly illustrate variability between realizations.

In Section 6, we comment on other models in the literature.

2. Model and notation

Note 1. We use the traditional language of branching processes (*individuals, children, births, deaths*) instead of the specific terms for evolution of species (*species, daughter species, speciations, extinctions*).

Let $\mathcal{T}$ be a continuous-time critical branching process (CBP) starting with one individual. In this process, each individual lives for an exponential time with rate λ, $\lambda > 0$, during which it gives birth at times of an independent Poisson process with rate λ. After birth, individuals behave independently of one another. We scale time so that $\lambda = 1$ and, thus, a time unit represents the mean lifetime of an individual. Write $N_{\mathcal{T}}(t) \geq 0$ for the number of individuals alive at time t after the origin of $\mathcal{T}$. A classical result [7, p. 480, Equation (10.4)] gives a modified geometric distribution for this number:

$$\mathrm{P}(N_{\mathcal{T}}(t) = 0) = \frac{t}{1+t}, \qquad \mathrm{P}(N_{\mathcal{T}}(t) = n) = \frac{t^{n-1}}{(1+t)^{n+1}}, \quad n \geq 1. \tag{1}$$

Write $\mathcal{T}_{t,n}$ for the process $\mathcal{T}$ originating at time t in the past and conditioned on having exactly n individuals at the present time.

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

Note 2. Within a process like $\mathcal{T}_{t,n}$ or $\mathcal{T}_n$, we use the convention that 'time s' means time s before the present. Thus, within $\mathcal{T}_{t,n}$, the time parameter s decreases from t to 0, meaning that time 'increases' from time t before present to the present time 0.

Let the time of origin of $\mathcal{T}$ have a prior that is uniform on $(0, \infty)$. Then, for a fixed $n \geq 1$, let $\mathcal{T}_n$ denote the posterior distribution of $\mathcal{T}$ conditioned on having n individuals at the present time. Rigorously,

$$\mathrm{P}(\mathcal{T}_n \in \cdot) = \frac{\int_0^\infty \mathrm{P}(\mathcal{T}_{t,n} \in \cdot)\, \mathrm{P}(N_{\mathcal{T}}(t) = n)\, \mathrm{d}t}{\int_0^\infty \mathrm{P}(N_{\mathcal{T}}(t) = n)\, \mathrm{d}t}.$$

Using (1) and the calculus result $\int_0^\infty s^{n-1}/(1+s)^{n+1}\, \mathrm{d}s = 1/n$, the distribution of $\mathcal{T}_n$ becomes

$$\mathrm{P}(\mathcal{T}_n \in \cdot) = \int_0^\infty \mathrm{P}(\mathcal{T}_{t,n} \in \cdot) \frac{nt^{n-1}}{(1+t)^{n+1}}\, \mathrm{d}t. \tag{2}$$

Within the random tree $\mathcal{T}_n$, the 'time of origin' T_n^{or} is a random time and, by the formula above, has density function

$$q_n(t) = \frac{nt^{n-1}}{(1+t)^{n+1}}, \qquad t > 0. \tag{3}$$

We refer to $\mathcal{T}_n$ and $\mathcal{T}_{t,n}$ as *complete trees*. In biological terminology, a complete tree records the birth times of all the (extinct or extant) species in a clade, as well as the extinction times of all the extinct species. A realization of a complete tree determines a realization of the *lineage tree* of the extant species. This is the smallest subtree of the complete tree that contains all the divergence times for pairs of lineages of extant species, without recording which ancestral species contain the lineage. We let $\mathcal{A}_{t,n}$ and $\mathcal{A}_n$ denote the lineage trees of $\mathcal{T}_{t,n}$ and $\mathcal{T}_n$, respectively. The time parameter s within $\mathcal{A}_n$ decreases from the time T_n^{mrca} of the *most recent common ancestor* of the n extant species to the present time 0 (in biology, the lineage tree is usually called the *phylogenetic tree*).

The continuous-time branching model $\mathcal{T}_{t,n}$ conditioned on having n extant individuals has previously been explored in [17]. The technique of representing random trees as walks was used to give an exact distribution for the lineage tree $\mathcal{A}_{t,n}$ (see Lemma 3 of [17]) via a convenient 'point process representation'. It was also used to describe the distribution of the limit structure of this tree (see Lemma 4 and Theorem 5 of [17]) via weak convergence of random walks to Brownian motion. In the present paper we draw upon these results, in particular when describing the distribution of the lineage tree $\mathcal{A}_n$ and its 'global' and 'local' limit structures. In Section 3, we recall the result for the distribution of $\mathcal{A}_{t,n}$ from [17] and use it, together with the distribution of the random time T_n^{or}, to obtain the distribution of $\mathcal{A}_n$. Later, in Section 5, we use the representation [17] of the tree $\mathcal{T}_{t,n}$ by a *contour process* in order to derive the distributions of various quantities associated with extinct species.

3. Point process representations of lineage trees

3.1. An exact description

The *point process representation* illustrated in Figure 1 is a useful exact description of the lineage tree $\mathcal{A}_{t,n}$. Consider an arbitrary lineage tree on n species. Draw this tree recursively from the top down, at each lineage divergence point randomly choosing which branch is drawn on the left and which on the right (see Figure 1, left-hand diagram). Label the extant species as $1, 2, \ldots, n$ from left to right. Each divergence of lineages involves adjacent contiguous blocks

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

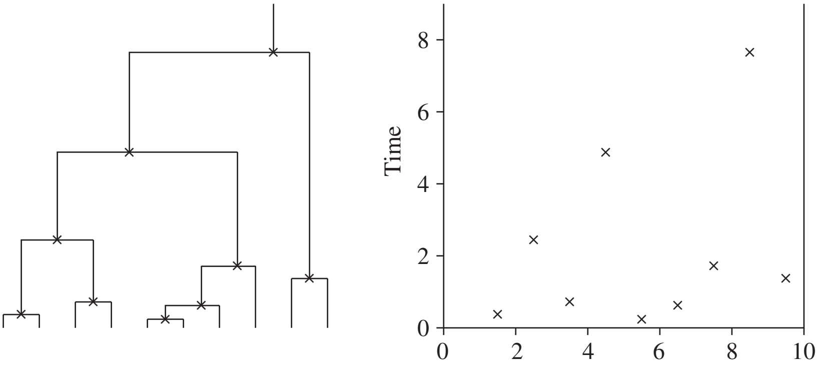

FIGURE 1: The point process representation (*right*) of a lineage tree (*left*) on $n = 10$ species.

of species, say $\{i, i+1, \dots, j\}$ and $\{j+1, j+2, \dots, k\}$, and occurs at some time s. The point process representation consists of marks at coordinates $(j + \frac{1}{2}, s)$ for each divergence.

The advantage of this method of representing the lineage tree is that we can clearly reconstruct the tree from the coordinates $\{(i + \frac{1}{2}, s_i),\ 1 \le i \le n-1\}$ of the point process: each combined lineage can be drawn upwards from the mark for its divergence as a vertical line. The distribution of the point process thus specifies the distribution of the lineage tree.

Proposition 1. (Lemma 3 of [17].) *Fix an* $n \ge 2$ *and a* $t > 0$. *The point process*

$$\{(i + \tfrac{1}{2}, h_i),\ 1 \le i \le n-1\},$$

where the (h_i) *are independent and identically distributed with density function*

$$f_t(s) = (1 + t^{-1})(1+s)^{-2}, \qquad 0 < s < t, \tag{4}$$

represents the lineage tree $\mathcal{A}_{t,n}$ *within the complete tree* $\mathcal{T}_{t,n}$.

By (2), the lineage tree $\mathcal{A}_n$ has a mixture representation

$$\mathrm{P}(\mathcal{A}_n \in \cdot) = \int_0^\infty \mathrm{P}(\mathcal{A}_{t,n} \in \cdot) q_n(t)\,\mathrm{d}t, \tag{5}$$

where $q_n(t)$ is the density function (3) of T_n^{or}. We can obtain exact formulae for various attributes of $\mathcal{A}_n$. Consider, for instance, the number of lineages at time s. Because each divergence creates one extra lineage, it is clear that within $\mathcal{A}_{t,n}$ this number of lineages is distributed as

$$1 + \mathrm{binomial}(n-1, \bar{F}_t(s)),$$

where

$$\bar{F}_t(s) = \int_s^t f_t(u)\,\mathrm{d}u = \frac{t-s}{t(1+s)}.$$

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

In the lineage tree $\mathcal{A}_n$, the distribution of the number of lineages is the mixture of binomial distributions implied by (5). The exact distribution of the time T_n^{mrca} for $\mathcal{A}_n$ is

$$\begin{aligned}\mathrm{P}(T_n^{\text{mrca}} \leq s) &= \int_0^\infty (1 - \bar{F}_t(s))^{n-1} q_n(t)\,\mathrm{d}t \\ &= \int_0^s q_n(t)\,\mathrm{d}t + n\Big(\frac{s}{1+s}\Big)^{n-1} \int_s^\infty (1+t)^{-2}\,\mathrm{d}t \\ &= \int_0^s q_n(t)\,\mathrm{d}t + \frac{ns^{n-1}}{(1+s)^n}, \qquad s > 0.\end{aligned}$$

Taking the derivative with respect to s shows that T_n^{mrca} has density

$$f_{T_n^{\text{mrca}}}(s) = \frac{n(n-1)s^{n-2}}{(1+s)^{n+1}}, \qquad s > 0.$$

In particular,

$$\mathrm{E}[T_n^{\text{mrca}}] = \int_0^\infty s f_{T_n^{\text{mrca}}}(s)\,\mathrm{d}s = (n-1)\int_0^\infty q_n(s)\,\mathrm{d}s = n-1. \tag{6}$$

We thank a referee for pointing out this elegant formula. In this paper, we mainly focus on asymptotic results rather than seeking more complicated exact formulae for other quantities. It is useful to distinguish two kinds of asymptotics: *global limits*, which refer to asymptotics for times of order n, and *local limits*, which refer to asymptotics for times of order 1.

3.2. The global limit point process

From (3), we calculate that if $t_n/n \to t > 0$, then

$$nq_n(t_n) = \frac{n^2}{(1+t_n)^2}\Big(1 - \frac{1}{1+t_n}\Big)^{n-1} \to t^{-2}\mathrm{e}^{-1/t}.$$

The limit is the density function of the *inverse exponential* IE(1) distribution, that is, the distribution of $1/\xi$ for a random variable ξ with an exponential(1) distribution. We summarize this in the following lemma, where '$\xrightarrow{\mathrm{D}}$' denotes convergence in distribution.

Lemma 1. *As $n \to \infty$, $n^{-1}T_n^{\text{or}} \xrightarrow{\mathrm{D}} T^{\text{or}}$, where the limit T^{or} has an IE(1) distribution.*

Now reconsider Figure 1. To obtain a global limit, we need to rescale both time and the left-to-right positions of the marks by a factor of n so as to fit the latter into a unit interval $[0, 1]$. Thus, the original point process of marks $\{(i + \frac{1}{2}, s_i),\ 1 \leq i \leq n-1\}$ is rescaled to $\{((i + \frac{1}{2})/n, s_i/n),\ 1 \leq i \leq n-1\}$. Given Proposition 1, the relevant result is that $n^2 f_{t_n}(s_n) \to s^{-2}$ as $s_n/n \to s > 0$ and $t_n/n \to t > 0$. The following limit behavior is intuitively clear.

Corollary 1. (Lemma 4 and Theorem 5 of [17].) *Let $t_n/n \to t > 0$. The rescaled point process $\{((i + \frac{1}{2})/n, h_i/n),\ 1 \leq i \leq n-1\}$ associated with the lineage tree $\mathcal{A}_{t_n,n}$ converges in distribution to the Poisson point process $\pi_{1,t}$ whose intensity measure is $\nu(\mathrm{d}l \times \mathrm{d}s) = \mathrm{d}l s^{-2}\,\mathrm{d}s \mathbf{1}_{[0,1]\times(0,t)}$.*

The limit process $\pi_{1,t}$ (illustrated in Figure 2) has an infinite number of points close to the lower boundary, but weak convergence on the open interval $(0, t)$ means convergence over

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

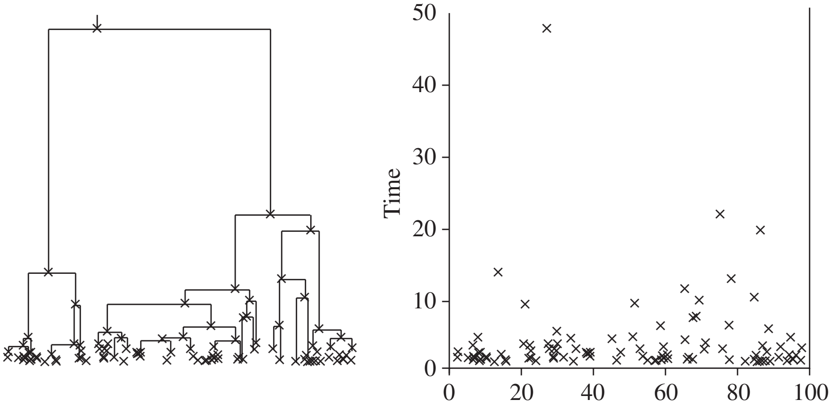

FIGURE 2: The point process $\pi_{1,t}$ (*right*) represents the lineage tree of a 'continuum tree' of species (*left*).

regions away from this boundary. Figure 2 indicates visually how the Poisson point process limit defines a limit random tree that is a kind of 'continuum tree' [2] with a lineage for each real $l \in (0, 1)$.

The mixture representation (5) and Corollary 1 immediately imply a global limit theorem for $\mathcal{A}_n$. To state it, let the random time T^{or} have an IE(1) distribution. Define a Cox point process π_1 on $(0, 1) \times (0, \infty)$ as follows: given that $T^{\text{or}} = t$, let π_1 be a Poisson point process with the law of $\pi_{1,t}$.

Corollary 2. *The rescaled point process* $\{((i + \frac{1}{2})/n, s_i/n),\ 1 \le i \le n-1\}$ *associated with the lineage tree* $\mathcal{A}_n$*, considered jointly with* T_n^{or}*, converges in distribution to the Cox point process* π_1*, considered jointly with* T^{or}*.*

Here is a brief application of this global limit theorem.

Corollary 3. *The limit joint behavior of* T_n^{or} *and* T_n^{mrca} *is given by*

$$(n^{-1}T_n^{\text{or}}, n^{-1}T_n^{\text{mrca}}) \xrightarrow{\text{D}} (T^{\text{or}}, T^{\text{mrca}}),$$

where the limit law has the joint density

$$f_{T^{\text{or}},T^{\text{mrca}}}(t, s) = t^{-2}s^{-2}\mathrm{e}^{-1/s}, \qquad 0 < s < t.$$

The marginal density of the random time T^{mrca} *is*

$$f_{T^{\text{mrca}}}(s) = s^{-3}\mathrm{e}^{-1/s}, \qquad s > 0.$$

The limit joint distribution can alternatively be expressed as

$$(T^{\text{or}}, T^{\text{mrca}}) \overset{\text{D}}{=} \left(\frac{1}{\xi_1}, \frac{1}{\xi_1 + \xi_2}\right),$$

where ξ_1 *and* ξ_2 *are* exponential(1) *independent, identically distributed random variables, and '*$\overset{\text{D}}{=}$*' denotes equality in distribution.*

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

Proof. Corollary 2 implies the required convergence in distribution to the limit $(T^{\mathrm{or}}, T^{\mathrm{mrca}})$, in which T^{mrca} is defined as the maximum height (that is, maximum second coordinate) of any point of π_1. Given that $T^{\mathrm{or}} = t$, the process π_1 is distributed as a Poisson point process $\pi_{1,t}$ with intensity measure $\nu(\mathrm{d}l \times \mathrm{d}s) = \mathrm{d}l s^{-2}\,\mathrm{d}s \mathbf{1}_{[0,1]\times(0,t)}$. Therefore, for the conditional law of T^{mrca} given $T^{\mathrm{or}} = t$, we have

$$\begin{aligned}\mathrm{P}(T^{\mathrm{mrca}} \le s \mid T^{\mathrm{or}} = t) &= \mathrm{P}(\{\pi_{1,t} \cap [0,1] \times (s,t)\} = \varnothing) \\ &= \exp\biggl(-\int_s^t u^{-2}\,\mathrm{d}u\biggr) \\ &= \mathrm{e}^{1/t-1/s}, \qquad 0 < s < t.\end{aligned}$$

Hence,

$$\mathrm{P}(T^{\mathrm{mrca}} \le s,\ T^{\mathrm{or}} \in \mathrm{d}t) = \mathrm{e}^{1/t-1/s}\,\mathrm{P}(T^{\mathrm{or}} \in \mathrm{d}t) = t^{-2}\mathrm{e}^{-1/s}\,\mathrm{d}t, \qquad 0 < s < t,$$

implying the joint density formula. The remaining calculations are straightforward.

Note 3. It follows that $\mathrm{E}[T^{\mathrm{mrca}}] = 1, \mathrm{var}[T^{\mathrm{mrca}}] = \infty$, and $\mathrm{E}[T^{\mathrm{or}}] = \mathrm{var}[T^{\mathrm{or}}] = \infty$. Different realizations of the lineage tree $\mathcal{A}_n$ vary greatly from one another.

3.3. The local limit point process

There is a different limit regime in which time is not rescaled. This limit tells us the local structure of the lineage tree relative to a given typical species, where ‘local’ refers to lineages merging with the given lineage within bounded time. Given Proposition 1, the relevant result is that

$$f_{t_n}(s) \to f(s) := (1+s)^{-2}, \qquad 0 < s < \infty, \quad \text{as } t_n \to \infty.$$

Consider the point process on $(\mathbb{Z}+\frac{1}{2}) \times (0,\infty)$ consisting of points $\{(i+\frac{1}{2}, \eta_i),\ i \in \mathbb{Z}\}$, where $(\eta_i)_{i\in\mathbb{Z}}$ are independent and identically distributed with density $f(s) = (1+s)^{-2}$. This point process (illustrated in Figure 3) defines an infinite tree $\mathcal{A}_\infty$ on an infinite set of lineages labeled by $\mathbb{Z}$.

Proposition 1 and the calculation above clearly imply the first assertion of the following corollary; the second assertion follows from the mixture representation (5), because in the local limit the mixing makes no difference.

Corollary 4. *Let $t_n \to \infty$ and let U_n be uniform random variables on $\{1,2,\dots,n\}$, independent of $\mathcal{A}_{t_n,n}$. Write $\{(U_n+i+\frac{1}{2}, s_{U_n+i}),\ i \in \mathbb{Z}\}$ for the point process associated with the lineage tree $\mathcal{A}_{t_n,n}$, centered at lineage U_n, where $s_j = 0$ for those j outside $[1,n]$. Then as $n \to \infty$, this point process converges in distribution to the point process $\{(i+\frac{1}{2}, \eta_i),\ i \in \mathbb{Z}\}$ defining $\mathcal{A}_\infty$. The same result holds for $\mathcal{A}_n$.*

Informally, the structure of $\mathcal{A}_\infty$ around lineage 0 provides an asymptotic approximation to the structure of $\mathcal{A}_n$ around a random lineage.

3.4. Some local calculations

We next give some elementary calculations within $\mathcal{A}_\infty$ that reflect the approximate behavior of the lineage trees $\mathcal{A}_n$ for large n. For a lineage at time s, we call the present $(t = 0)$ number of species descending from this lineage the *size* of the lineage. We call the $n \to \infty$ limit of $n^{-1} \times$ (number of lineages in $\mathcal{A}_n$ at time s) the *density of lineages* in $\mathcal{A}_\infty$ at time s.

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

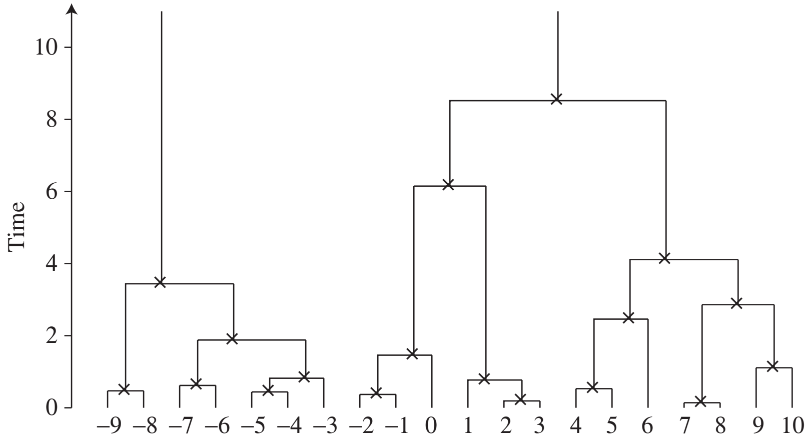

FIGURE 3: A realization of a part of $\mathcal{A}_\infty$ that approximates the local structure of $\mathcal{A}_n$ for large n (the two main ancestral lineages diverged at around time $t = 16$).

Corollary 5. (Some calculations for $\mathcal{A}_\infty$.) (a) *The density of ancestral lineages at time s in the past equals $(1+s)^{-1}$, and the size of a random lineage at time s has a* geometric$((1+s)^{-1})$ *distribution.*

(b) *The rate of lineages merging as s increases (i.e. as time runs backwards) is $m(s) = 2(1+s)^{-1}$ and, given that this event occurs at s for some lineage, the size of the lineage it merges with has a* geometric$((1+s)^{-1})$ *distribution.*

(c) *As s decreases (i.e. as time runs forward), the rate at which a lineage of size $k \geq 1$ branches is $b_k(s) = (k-1)(s(1+s))^{-1}$ at time s, and the size of the lineage produced on the left of the branch-point has a uniform distribution on $\{1, \dots, k-1\}$.*

Proof. (a) The density of ancestral lineages at time s in the past is just the density of branch-points at times greater than s, given by

$$G(s) = \int_s^\infty f(u)\,\mathrm{d}u = (1+s)^{-1}.$$

Hence, the number of extant species descended from a 'typical' lineage at time s has a geometric$((1+s)^{-1})$ distribution

$$p_s(i) = \left(\frac{1}{1+s}\right)\left(\frac{s}{1+s}\right)^{i-1}, \qquad i \geq 1,$$

as this is the distribution of distances between branch-points at heights greater than s.

(b) As s increases (i.e. as time runs backwards), the probability of a lineage merging with another lineage is

$$m(s) = 2\frac{f(s)}{G(s)} = \frac{2}{1+s}.$$

This holds because such a merging occurs in $[s, s + \mathrm{d}s]$ when one of the two branch-points separating the given lineage from its neighboring lineages, which must be at a height greater

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

than or equal to s, occurs during $[s, s+\mathrm{d}s]$; this event has probability $f(s)\,\mathrm{d}s/G(s)$ for each branch-point. Moreover, if a lineage merges at s then (independent of the size of the first lineage) the size of the second lineage has the geometric$((1+s)^{-1})$ distribution above.

(c) As s decreases (i.e. as time runs forwards), the unconditional rate of mergers of clades of sizes k_1 and k_2 at time s (per unit time, relative to number of species) equals

$$G(s)(1-G(s))^{k_1-1}f(s)(1-G(s))^{k_2-1}G(s),$$

which we derive by considering the heights of branch-points required for this event to occur. Similarly, the number of size-(k_1+k_2) lineages at time s, relative to the number of species, equals

$$G(s)(1-G(s))^{k_1+k_2-1}G(s).$$

The rate of splitting of a size-(k_1+k_2) lineage into two lineages of sizes k_1 and k_2 therefore equals

$$\frac{G(s)(1-G(s))^{k_1-1}f(s)(1-G(s))^{k_2-1}G(s)}{G(s)(1-G(s))^{k_1+k_2-1}G(s)} = \frac{1}{s(1+s)}.$$

Thus, if a lineage is of size k then at time s the stochastic rate of branching is $b_k(s) = (k-1)/[s(1+s)]$. Since the rate of splitting is independent of the choice of partition of k into k_1 and k_2, the size of a left-hand side subclade lineage is uniform on $\{1, 2, \dots, k-1\}$.

4. Time reversal and consequences

Recall that for a *stationary* Markov process, its time-reversal is also a stationary Markov process. For a Markov process that is not stationary, or which is conditioned on a terminal value, the time-reversal is typically *inhomogeneous*. Lemma 2, below, highlights a special feature of our processes.

In the critical branching process underlying our model (see Section 2), the population size is the continuous-time Markov chain with transition rates

$$q_{i,i+1} = q_{i,i-1} = i. \tag{7}$$

Recall the definition of the complete tree $\mathcal{T}_n$. Write $\{N_n(s),\ T_n^{\mathrm{or}} \ge s \ge 0\}$ for the associated process that counts the number of species at time s before present (regard this process as making the transition $0 \to 1$ at time $s = T_n^{\mathrm{or}}$).

Lemma 2. *Let $\{\hat{N}_n(s),\ 0 \le s \le T_n^0\}$ be the continuous-time chain with transition rates given by (7) and $\hat{N}_n(0) = n$, run until the first hitting time T_n^0 at state* 0. *Then*

$$\{N_n(s),\ T_n^{\mathrm{or}} \ge s \ge 0\} \overset{\mathrm{D}}{=} \{\hat{N}_n(s),\ 0 \le s \le T_n^0\}$$

and, so, in particular, $T_n^{\mathrm{or}} \overset{\mathrm{D}}{=} T_n^0$.

Proof. We will verify that $\{\hat{N}_n(s),\ 0 \le s \le T_n^0\}$ is the time-reversal of the population size process by checking probabilities of primitive events. Fix $s_0, \dots, s_M$ with $s_M > s_{M-1} > \cdots > s_1 > s_0 = 0$ and positive integers $k_M = 1, k_{M-1}, \dots, k_1 = n$ with $|k_m - k_{m-1}| = 1$. Set $k_{M+1} = 0$. The event

{as s decreases, $N_n(s)$ jumps from k_{m+1} to k_m during $[s_m, s_m + \mathrm{d}s_m]$
(for all m, $M \ge m \ge 1$), and makes no other jumps}

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

has measure

$$\mathrm{d}s_M \times \prod_{m=M}^{2} (\mathrm{e}^{-k_m(s_m - s_{m-1})} k_m \, \mathrm{d}s_{m-1}) \times \mathrm{e}^{-k_1 s_1},$$

where the first term, $\mathrm{d}s_M$, comes from the uniform Bayes prior. For the reversed process, the event

$$\{\text{as } s \text{ increases, } \hat{N}_n(s) \text{ jumps from } k_m \text{ to } k_{m+1} \text{ during } [s_m, s_m + \mathrm{d}s_m]$$
$$(\text{for all m, } 1 \le m \le M), \text{ and makes no other jumps}\}$$

has probability

$$\prod_{m=1}^{M} (\mathrm{e}^{-k_m(s_m - s_{m-1})} k_m \, \mathrm{d}s_m).$$

By inspection, the first measure is exactly $1/n$ times the second probability, so, after conditioning, the probability measures are equal.

We now observe two simple consequences of this time-reversal identity. The process $\{\hat{N}_n(s),\ 0 \le s \le T_n^0\}$ is a martingale with steps of ± 1 started at n and let run until 0 is hit. Hence, by the exit place formula for such martingales,

$$\mathrm{P}\Big(\max_{0 \le s \le T_n^0} \hat{N}_n(s) \ge c\Big) = \frac{n}{c}, \qquad c \ge n.$$

Lemma 2 therefore implies the following corollary.

Corollary 6. *For the process* $\{N_n(s),\ T_n^{\mathrm{or}} \ge s \ge 0\}$,

$$\mathrm{P}\Big(\max_{T_n^{\mathrm{or}} \ge s \ge 0} N_n(s) \ge c\Big) = \frac{n}{c}, \qquad c \ge n.$$

Furthermore, every extinction within the process $\mathcal{T}_n$ corresponds to a downwards step in $N_n(s)$ as s decreases, and, hence, to an upwards step in $\hat{N}_n(s)$ as s increases. The number of such upward steps equals $(D_n - n)/2$, where D_n is the number of steps of the embedded jump chain of $\hat{N}_n(\cdot)$, which is just a discrete-time simple symmetric random walk.

Corollary 7. *Within the model $\mathcal{T}_n$ of a clade on n extant species, the total number N_n^{ext} of extinct species is distributed in the same way as $(D_n - n)/2$, where D_n is the hitting time at* 0 *of a simple symmetric random walk started at n. In particular,*

$$n^{-2} D_n \xrightarrow{\mathrm{D}} \tfrac{1}{2}\tau_1,$$

where τ_1 is the first passage time from 1 *to* 0 *of a standard Brownian motion with density function*

$$f_{\tau_1}(x) = (2\pi x^3)^{-1/2} \mathrm{e}^{-1/(2x)}, \qquad 0 < x < \infty.$$

The second assertion follows, of course, from the weak convergence of a simple random walk to a Brownian motion.

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

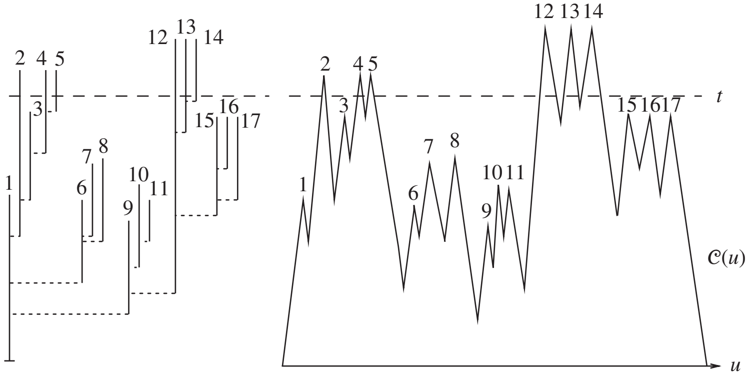

FIGURE 4: A realization of a tree $\mathcal{T}_{t,n}$ with $n = 6$ extant individuals (labeled $\{2, 4, 5, 12, 13, 14\}$), and its contour process representation $\mathcal{C}(u)$.

5. Exploiting the contour process

The results so far answer some, but not all, questions we might ask about the complete tree $\mathcal{T}_n$ and the lineage tree $\mathcal{A}_n$. For instance, the rescaled limit as $n \to \infty$ of the time-reversed process $\{\hat{N}_n(s)\}$ in Lemma 2 is the well-known *Feller branching diffusion*, which therefore is the limit of the total population size process $\{N_n(s),\ T_n^{\text{or}} \geq s \geq 0\}$. However, this does not tell us anything about the relationship between $\{N_n(s)\}$ and the lineage tree $\mathcal{A}_n$. We might also be interested in questions about the number of extinct species, for instance the total number or the number alive at the time of the most recent common ancestor of the extant species. All of these matters, as well as the local limit structure of the complete tree, can be studied using a representation of a tree given by its *contour process*.

5.1. The contour process

For any deterministic population process in continuous time, starting at the birth of a single individual there is a particular representation of its family history as a *rooted planar tree* (illustrated in Figure 4). Each individual is represented by an edge whose length equals the individual's lifetime. The birth of an offspring corresponds to a branch-point in its parent's edge, and the length of the parent's edge up to this branch-point equals the age of the parent at this offspring's birth time. From the branch-point, the offspring's edge is drawn to the right of the parent's edge. (In Figure 4, the tree edges have been drawn as solid vertical lines and the branch-points have been indicated by horizontal dotted lines.) If the total population is finite then we can label the individuals in a 'depth-first' search order.

Associated with such a rooted planar tree is its contour process, defined as follows (this idea goes back to Neveu and Pitman [14], and a recent survey is given in [16]). The contour process $\mathcal{C}(u)$ is a continuous function giving the distance from the root at time u in a unit-speed, depth-first walk around the tree. Such a walk starts at the root, and completely traverses each edge once upwards and once downwards, following the depth-first order (intuitively, clockwise around the edges of the tree), ending back at the root. The contour process consists of alternating line segments of slopes $+1$ ('rises') and slopes -1 ('falls'). The convention of unit speed implies that heights in the contour process match the times in the population process; birth and death times are respectively matched by the local minima and local maxima of the contour process.

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

5.2. Contour process of a critical branching tree

Recall that $\mathcal{T}$ denotes the continuous-time critical branching process started with one individual at time 0 and continued until extinction. The next result, due to Le Gall [11] and Neveu and Pitman [14], gives a simple description of the contour process of $\mathcal{T}$.

Proposition 2. *In the contour process of $\mathcal{T}$, the sequence of rises and falls*

$$(\xi_1, -\xi_2, \xi_3, -\xi_4, \dots, \xi_{M-1}),$$

which excludes the last fall, has a distribution derived from a sequence $(\xi_i)_{i\geq 1}$ of independent exponential(1) *variables, for*

$$M := \min\{m \colon \xi_1 - \xi_2 + \xi_3 - \xi_4 + \cdots - \xi_m < 0\}.$$

We call this contour process $(\xi_1, -\xi_2, \dots, \xi_{M-1})$ an *ERW excursion*: an excursion of an exponential random walk. Accordingly, we call the infinite sequence $(\xi_1, -\xi_2, \xi_3, -\xi_4, \dots)$ an *ERW process*. Note the brief proof of the following classical result.

Lemma 3. *Let H be the maximum height in an ERW excursion, or, equivalently (by Proposition 2), the extinction time of $\mathcal{T}$. Then* $\mathrm{P}(H > h) = (1+h)^{-1}$, $0 < h < \infty$.

Proof. This follows directly from the law of the population size process of $\mathcal{T}$ given in (1). The extinction time of $\mathcal{T}$ is greater than h if and only if the population size of $\mathcal{T}$ at time h is strictly greater than 0, which by (1) has probability $1 - h(1+h)^{-1} = (1+h)^{-1}$.

Before proceeding to new results, let us give the proof, from [17], of Proposition 1, because our arguments in subsequent sections will use similar ideas. Fix $t > 0$ and $n \geq 2$. Condition the contour process $\mathcal{C}(\cdot)$ to have exactly n upcrossings over height t (see Figure 4). This gives the contour process of the random tree ($\mathcal{T}_{t,n}^{+}$, say), which is the CBP conditioned on there being exactly n individuals alive at time t. This $\mathcal{T}_{t,n}^{+}$ is the same as our model $\mathcal{T}_{t,n}$, except for the convention regarding the direction of the time parameter and the fact that, in $\mathcal{T}_{t,n}$, the process terminates with the n individuals extant at the present time, whereas, in $\mathcal{T}_{t,n}^{+}$, the process of descendants of these n individuals continues until extinction. The latter difference plays no role in the following argument. The height of the minimum between each pair of successive upcrossings in Figure 4 matches the divergence of lineages (i.e. the branch-point) of that pair of extant individuals. Marking these heights at regular horizontal interval spacings gives the point process $\mathcal{A}_{t,n}$ exactly as in Figure 1, except for a reflection of the vertical time-scale. Since $\mathcal{C}(\cdot)$ is strong Markov and stationary, the parts of an ERW excursion between a downcrossing of t and the next upcrossing of t are mutually independent, and are distributed in exactly the same way as the reflection of the original ERW excursion conditioned not to have height greater than t. Thus, these heights of lineage divergence, when measured on the reflected time-scale (i.e. downwards from t), are distributed in the same way as the maximum height H in Lemma 3 conditioned on $\{H < t\}$. This conditioned distribution is the distribution (4), proving Proposition 1.

5.3. Species numbers at the time of most recent common ancestor and weak convergence of the contour process

Recall that $N_n(T_n^{\mathrm{mrca}})$ denotes the number of species alive at the time of the most recent common ancestor. In the contour process, the number of species at any time s after its origin is the number of upcrossings (which equals the number of downcrossings) of the contour process at height s. If the time since the origin of $\mathcal{T}_n$ is $T_n^{\mathrm{or}} = t$, then the contour process has n

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

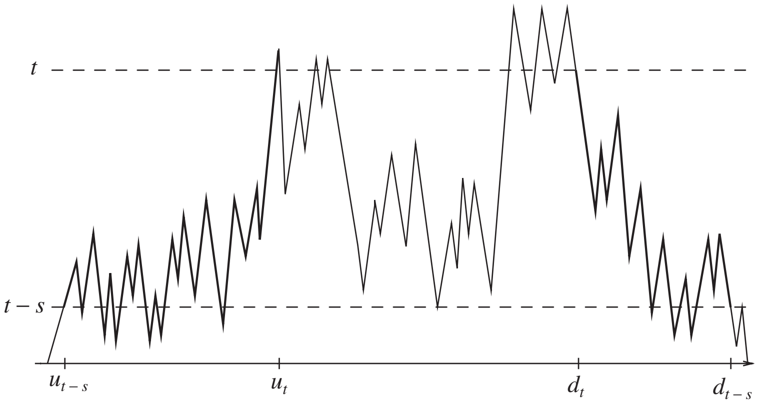

Figure 5: The contour process: certain parts in the intervals $[u_{t-s}, u_t]$ and $[d_t, d_{t-s}]$ represent the number of species alive at the time of the most recent common ancestor.

upcrossings and downcrossings at height t. If the time of the most recent common ancestor is $T_n^{\mathrm{mrca}} = s$, then the maximal depth of the subexcursions below height t, measured away from t, is s (see Figure 5). The lineage divergence of the most recent common ancestor is the lowest local minimum between the first and last upcrossing of t and occurs at height $t - s$.

In the contour process, mark by u_s the horizontal coordinate of the first upcrossing of a height s, and mark by d_s the coordinate of the last downcrossing of this height. As shown in Figure 5, there are no upcrossings or downcrossings of $t - s$ before the first upcrossing of $t - s$ or after the last downcrossing of $t - s$. If $t - s$ is the height of T^{mrca} then there are no up- or downcrossings of $t - s$ between the first upcrossing of t and the last downcrossing of t. Hence, $N_n(T_n^{\mathrm{mrca}})$ is the number of upcrossings of $t - s$ between u_{t-s} and u_t plus the number of downcrossings between d_t and d_{t-s}. Since the contour process is an ERW excursion conditioned to have n upcrossings and downcrossings at height T_n^{or}, we can now perform the following calculation.

Lemma 4. *Conditional on* $(T_n^{\mathrm{or}}, T_n^{\mathrm{mrca}}) = (t, s)$*,* $N_n(T_n^{\mathrm{mrca}})$ *is distributed as a sum of two independent* geometric(p_n) *random variables, where*

$$p_n = 1 - \frac{t-s}{1+t-s}\frac{s}{1+s}.$$

Proof. Since the contour process $\mathcal{C}(\cdot)$ is strong Markov and stationary, the increment of the process between the first upcrossings u_{t-s} of $t - s$ and u_t of t, that is,

$$\mathcal{C}(u) - (t-s), \qquad u_{t-s} \le u \le u_t,$$

is distributed in the same way as an ERW process conditioned to reach height s before it reaches depth $-(t - s)$, and stopped when it first hits s. Since $\mathcal{C}(\cdot)$ has the same law when its u coordinate is run in reverse, the part of the contour process between the last downcrossings d_t of $t - s$ and d_{t-s} of t (when run backwards in the u coordinate), that is,

$$\mathcal{C}(u) - (t-s), \qquad d_{t-s} \ge u \ge d_t,$$

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

is also distributed in the same way as an ERW process conditioned to reach height s before it reaches a depth $-(t-s)$, and stopped when it first hits s. Additionally, these two parts of the contour process are independent. The probability that an ERW process reaches s before it reaches $-(t-s)$ is, by the law of maximum height H in Lemma 3,

$$\frac{\mathrm{P}(H>s)}{\mathrm{P}(H>s)+\mathrm{P}(H>t-s)-\mathrm{P}(H>s)\,\mathrm{P}(H>t-s)}=\frac{1+t-s}{1+t}.$$

The probability that an ERW process makes k upcrossings of 0 before it first hits s, provided that its height stays below s and its depth above $-(t-s)$, is

$$(\mathrm{P}(H<t-s)\,\mathrm{P}(H<s))^{k-1}\,\mathrm{P}(H>s)=\left(\frac{t-s}{1+t-s}\frac{s}{1+s}\right)^{k-1}\frac{1}{1+s}$$

for $k=1,2,\dots$. Hence, the number of upcrossings of $t-s$ that $\mathcal{C}(u)$ makes during $[u_{t-s},u_t]$ has a

$$\text{geometric}\left(1-\frac{t-s}{1+t-s}\frac{s}{1+s}\right)$$

distribution.

Since, by Corollary 3, $(n^{-1}T_n^{\mathrm{or}},n^{-1}T_n^{\mathrm{mrca}})\xrightarrow{\mathrm{D}}(T^{\mathrm{or}},T^{\mathrm{mrca}})$, as $n\to\infty$ we have

$$np_n=1-\frac{T_n^{\mathrm{or}}-T_n^{\mathrm{mrca}}}{1+T_n^{\mathrm{or}}-T_n^{\mathrm{mrca}}}\frac{T_n^{\mathrm{mrca}}}{1+T_n^{\mathrm{mrca}}}\xrightarrow{\mathrm{D}}\frac{1}{T^{\mathrm{or}}-T^{\mathrm{mrca}}}+\frac{1}{T^{\mathrm{mrca}}}.$$

Hence, the two geometric(p_n) variables in Lemma 4, when rescaled by n^{-1}, converge to two independent exponential($\lambda(T^{\mathrm{or}},T^{\mathrm{mrca}})$) variables, where $\lambda(t,s)=(t-s)^{-1}+s^{-1}$. Consequently, the conditional law of $n^{-1}N_n(T_n^{\mathrm{mrca}})$ given $(T_n^{\mathrm{or}},T_n^{\mathrm{mrca}})$ converges to a gamma-distributed variable with shape parameter 2 and scale parameter $\lambda(T^{\mathrm{or}},T^{\mathrm{mrca}})$. Combining this with the result of Corollary 3 establishes assertion (8) of the following corollary.

Corollary 8. *The joint limit behavior of T_n^{or}, T_n^{mrca}, and $N_n(T^{\mathrm{mrca}})$ is given by*

$$(n^{-1}T_n^{\mathrm{or}},n^{-1}T^{\mathrm{mrca}},n^{-1}N_n(T_n^{\mathrm{mrca}}))\xrightarrow{\mathrm{D}}(T^{\mathrm{or}},T^{\mathrm{mrca}},N^{\mathrm{mrca}}),$$

where the limit has the joint density

$$\begin{aligned}f_{T^{\mathrm{or}},T^{\mathrm{mrca}},N^{\mathrm{mrca}}}(t,s,r)&=t^{-2}s^{-2}\lambda(t,s)^2r\exp\left(-\frac{1}{s}-\lambda(t,s)r\right)\\&=(t-s)^{-2}s^{-4}r\exp\left(-\frac{1}{s}-\frac{tr}{s(t-s)}\right),\qquad 0<s<t,\ 0<r.\end{aligned}\tag{8}$$

The marginal density of N^{mrca} is

$$f_{N^{\mathrm{mrca}}}(r)=2(1+r)^{-3},\qquad r>0.$$

The marginal density formula follows from (8) via a calculus exercise.

Note 4. Observe that $\mathrm{E}[N^{\mathrm{mrca}}]=1$ and $\mathrm{var}[N^{\mathrm{mrca}}]=\infty$, further indicating the high variability between tree realizations in our model.

Note 5. In the limit $t_n/n\to t\in(0,\infty)$, after rescaling the contour process of $\mathcal{T}_{t_n,n}$ (illustrated in Figure 5) converges to a Brownian excursion conditioned on the total local time (i.e. the occupation measure) at height t being equal to 1. Results like Corollary 8 may be reinterpreted as providing exact formulae for quantities defined in terms of such conditioned Brownian excursions.

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

5.4. Extinct species

Textbooks on evolution often contain comments to the effect that [15, p. 24] 'the probability that a given fossil is actually part of an ancestral lineage [of some extant species] is actually rather remote'. Using our model allows several calculations relevant to this issue to be performed.

Consider some species v originating at time h before the present. Proposition 3, below, will verify the intuitively natural fact that the process of descendants of v is, in the $n \to \infty$ limit, just the unconditioned CBP $\mathcal{T}$. Given this result, the chance that some descendant of v (or v itself) is extant at the present time is just the chance of its descendant tree $\mathcal{T}$ surviving for at least time h. By Lemma 3, this is precisely $(1+h)^{-1}$, from which we have the following result.

Corollary 9. *For a species alive at time h before the present, the chance that some of its descendant species (or the species itself) is extant tends to $1/(1+h)$ as $n \to \infty$, with h fixed.*

Next consider the total number N_n^{anc} of species ancestral to the n extant ones. In this number we include every species in $\mathcal{T}_n$ that has an extant species as an ultimate descendant, but exclude the extant species themselves. We know from Lemma 2 that the number of species at time h is $N_n(h) \approx n$, for $h = o(n)$, and from Corollary 3 that the time of origin T_n^{or} is of order $O(n)$. Thus, informally, Corollary 9 leads us to expect that

$$\mathrm{E}[N_n^{\mathrm{anc}}] \approx \int_0^{O(n)} \frac{n}{1+h}\,\mathrm{d}h \approx n \log n.$$

In Corollary 10, we will prove a precise result for N_n^{anc} based on the following lemma.

Lemma 5. *Conditional on $T_n^{\mathrm{or}} = t$, the total number of ancestral individuals N_n^{anc} in $\mathcal{T}_n$ satisfies*

$$N_n^{\mathrm{anc}} \xrightarrow{\mathrm{D}} \sum_{i=1}^{n} X_i,$$

where the random variables X_i, $1 \le i \le n$, are independent, X_1 has a Poisson(t) *distribution, and $X_2, \dots, X_n$ have the law*

$$\mathrm{P}(X_i = k) = \int_0^t \frac{\mathrm{e}^{-s} s^k}{k!} f_t(s)\,\mathrm{d}s, \qquad k \ge 0,$$

with $f_t(\cdot)$ as in (4).

Proof. Label the extant individuals $\{1, 2, \dots, n\}$ from left to right as they appear in the contour process. Let X_i be the number of ancestors of the ith extant individual, without including any of those previously counted in X_j, $j < i$.

Suppose that $T^{\mathrm{or}} = t$; then the ancestry of the extant individuals is described by the part of the contour process $\mathcal{C}(\cdot)$ below height t. Recall that the part of $\mathcal{C}(\cdot)$ below t consists of $n-1$ independent subexcursions below t, which we label e_i, $1 \le i \le n-1$, and the parts of $\mathcal{C}(\cdot)$ before the first upcrossing and after the last downcrossing of t; we label the former part $e_{0,R}$ (see Figure 6).

Let h_i, $1 \le i \le n-1$, be the depths below t of the subexcursions e_i, meaning that $t - h_i$ are the heights of the lowest points of e_i. These heights match the times of lineage divergence of extant individuals. Their law is given by (4). Now partition the excursions e_i at their lowest points and let $e_{i,R}$, $1 \le i \le n-1$, denote the parts on the right. Figure 6 then shows that the ancestors of the first extant individual correspond in $e_{0,R}$ to the levels of constancy of the

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

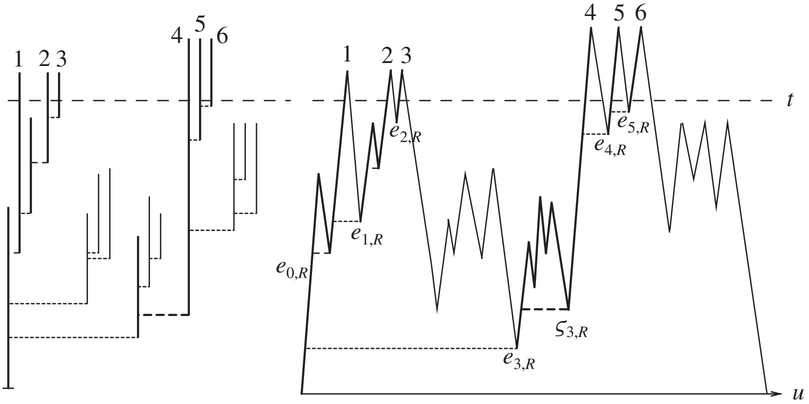

FIGURE 6: Ancestral lineages of the extant individuals (labeled $\{1, 2, 3, 4, 5, 6\}$) are matched in the contour process by the levels of constancy of the processes $\varsigma_{i-1,R}$, $1 \le i \le n$.

process $\varsigma_{0,R}(u) = \inf_{v\ge u}(e_{0,R}(v))$. These levels of constancy of $\varsigma_{0,R}$ match the times of lineage divergence of the ancestors of the first individual. Similarly, for the ith, $2 \le i \le n$, extant individual, Figure 6 shows that the additional ancestors of individual i (excluding those appearing as ancestors of extant individuals $j < i$) correspond in $e_{i-1,R}$ to the levels of constancy of the process $\varsigma_{i-1,R}(u) = \inf_{v\ge u}(e_{i-1,R}(v))$.

Thus, the number of ancestors X_i of the ith extant individual equals the number of levels of constancy of the process $\varsigma_{i-1,R}(\cdot)$. It is clear that $e_{0,R}$ is distributed in the same way as an ERW process conditioned to hit t before 0 and stopped the first time it does so. It is less obvious that, given h_i, $e_{i,R}$ is distributed in the same way as an ERW process conditioned to reach h_i before 0 and stopped the first time it does so (see Lemma 6 of [17]). For such an ERW process, the levels of constancy of its future infimum process form a Poisson process restricted to the set $[0, t]$ for $e_{0,R}$ and the set $[0, h_i]$ for $e_{i-1,R}$, $2 \le i \le n$. This can easily be seen for levels of constancy of the past supremum process of a conditioned ERW process (again see Lemma 6 of [17]), and the time reversibility of ERW excursions implies the rest. Hence, the number of ancestors of the first extant individual is Poisson(t) distributed, and the number of additional ancestors of the extant individuals i, $2 \le i \le n$, is Poisson(h_i) distributed. Combining this with the distributions (4) of the depths h_i proves the claim.

Corollary 10. *As $n \to \infty$, we have $N_n^{\text{anc}}/(n \log n) \xrightarrow{\text{P}} 1$, where '$\xrightarrow{\text{P}}$' denotes convergence in probability.*

Proof. Fix a sequence (t_n) such that $t_n/n \to t \in (0, \infty)$. Corollary 3 shows that $n^{-1}T_n^{\text{or}}$ has a distributional limit on $(0, \infty)$, so it suffices to prove that $N_n^{\text{anc}}/(n \log n) \xrightarrow{\text{P}} 1$ conditional on $\{T_n^{\text{or}} = t_n\}$.

We prove this using the representation $N_n^{\text{anc}} = \sum_{i=1}^n X_i$ from Lemma 5, and throughout the argument we condition on $\{T_n^{\text{or}} = t_n\}$. Note that the contribution to the sum from X_1 is negligible (because X_1 has a Poisson(t_n) distribution), so we may assume that X_1 has the same distribution as the X_i, $2 \le i \le n$. We now calculate

$$\text{E}[X_2] = \int_0^{t_n} s f_{t_n}(s)\,\text{d}s \sim \int_0^{t_n} s(1+s)^{-2}\,\text{d}s \sim \log t_n \sim \log n.$$

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

A similar calculation gives $\operatorname{var}[X_2] = O(n)$. Therefore,

$$\mathrm{E}[N_n^{\mathrm{anc}}] \sim n \log n, \qquad \operatorname{var}[N_n^{\mathrm{anc}}] = O(n^2),$$

and the desired result, $N_n^{\mathrm{anc}}/(n \log n) \xrightarrow{\mathrm{P}} 1$, follows via Chebyshev's inequality.

5.5. Local limit structure of the complete tree

We can also use the contour process to derive local limit results for the complete tree $\mathcal{T}_n$ that are analogous to those of Corollary 4 for the lineage tree $\mathcal{A}_n$. We show below that the local structure of $\mathcal{T}_n$ relative to a given typical individual converges to the local structure relative to the root of an infinite tree that can be easily defined from a CBP tree. There are two versions of such results, depending on whether the typical individual is chosen as a random *extant* species, or as a random species from the entire history of the clade.

Let i be an individual in the complete tree $\mathcal{T}_n$, with birth time $b(i)$, say. Within this section, our convention for the time parameter in $\mathcal{T}_n$ is that it increases as time increases. For $\sigma > 0$, let $\tilde{\mathcal{T}}_n(i, [b(i) - \sigma, b(i) + \sigma])$ denote the subtree of $\mathcal{T}_n$ comprised of all the individuals j whose birth times are in the interval $[b(i) - \sigma, b(i) + \sigma]$ and for whom the divergence times of their lineages from that of i are in the interval $[b(i) - \sigma, b(i) + \sigma]$ (see Figure 7). We call i the *distinguished individual* in $\tilde{\mathcal{T}}_n(i, [b(i) - \sigma, b(i) + \sigma])$.

We now describe an infinite random tree $\tilde{\mathcal{T}}$ derived from the CBP. Take a distinguished individual born at time 0 and let the tree of its descendants be distributed in the same way as a CBP tree $\mathcal{T}$. Let the parent of this individual have an exponential(1) age at time 0 and an independent exponential(1) lifetime after time 0. In turn, let the grandparent have an exponential(1) age at the birth of the parent and an independent exponential(1) lifetime after that birth time, and so on for the other ancestors. Let each of the ancestors have other children at the times of a rate-1 Poisson process, and let the trees of such children and their descendants be distributed in the same way as independent CBP trees.

From Proposition 2, recall the construction of a CBP tree $\mathcal{T}$ from the ERW excursion $(\xi_1, -\xi_2, \xi_3, \dots)$. It is easy to check that, given a two-sided ERW process

$$(\dots, -\xi_{-2}, \xi_{-1}, -\xi_0, \xi_1, -\xi_2, \xi_3, \dots),$$

an analogous construction produces the infinite tree $\tilde{\mathcal{T}}$. Write $\tilde{\mathcal{T}}[-\sigma, \sigma]$ for the subtree of $\mathcal{T}$ comprised of those individuals j whose birth times are in the interval $[-\sigma, \sigma]$, and for whom the divergence times of their lineages from that of the distinguished individual are in the interval $[-\sigma, \sigma]$. Note that

$$\tilde{\mathcal{T}}[-\sigma, \sigma] \quad \text{is determined by} \quad (\xi_i,\ M^- \le i \le M^+). \tag{9}$$

The sequence

$$(\xi_{M^-}, -\xi_{M^-+1}, \dots, -\xi_0, \xi_1, \dots, \xi_{M^+-1}, -\xi_{M^+})$$

is the excursion of the two-sided ERW process above height $-\sigma$.

The following proposition shows the convergence of the local structure of $\mathcal{T}_n$, relative to a random (most likely extinct) individual, to the local structure of $\tilde{\mathcal{T}}$.

Proposition 3. *Let I_n denote a uniform random species from the clade $\mathcal{T}_n$. Then, as $n \to \infty$ for fixed $\sigma > 0$,*

$$\tilde{\mathcal{T}}_n(I_n, [b(I_n) - \sigma, b(i) + \sigma]) \xrightarrow{\mathrm{D}} \tilde{\mathcal{T}}[-\sigma, \sigma].$$

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

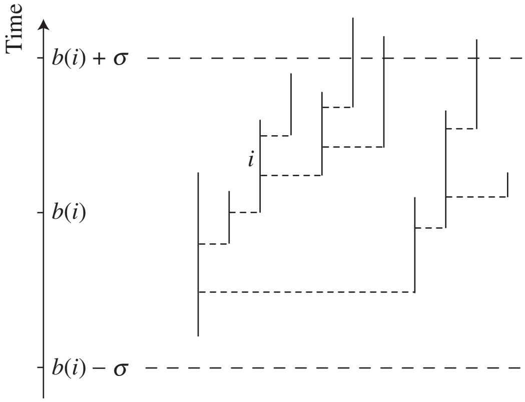

FIGURE 7: The local structure of the complete tree, relative to the individual i.

Note 6. The underlying notion of *convergence* of finite trees is the natural one, which can be formalized in several equivalent ways, e.g. via a point process representation.

Proof of Proposition 3. We outline the proof, omitting small details. Write $(\xi_i,\ i \geq 1)$ for the ERW process. Fix an integer $m \geq 2$. Let $\theta_{m,N}$ be the empirical distribution of the N $2m$-vectors

$$\{(\xi_{2i+1}, \xi_{2i+2}, \dots, \xi_{2i+2m}),\ 0 \leq i \leq N-1\}.$$

Then $\theta_{m,N}$ is a random probability distribution and (by the Glivenko–Cantelli theorem on R^{2m}) converges in probability, as $N \to \infty$, to the nonrandom probability distribution

$$\mu_m = \mathrm{dist}(\xi_1, \dots, \xi_{2m})$$

(where $\mathrm{dist}(\cdots)$ denotes distribution). By large deviation theory (see [4, Section 6.3, pp. 272–278]) this convergence remains true conditional on events A_N for which $1/\,\mathrm{P}(A_N) = O(\beta^N)$ for some $\beta > 1$.

To prove the proposition, recall from Lemma 1 that T_n^{or} is of order n. We can therefore fix a sequence (t_n) such that $t_n/n \to t \in (0,\infty)$; it is sufficient to prove the proposition for $\mathcal{T}_{t_n,n}$. Also fix integers N_n such that $N_n/n^2 \to v \in (0,\infty)$. Let A_{N_n} be the event that an ERW process has an excursion above 0 with exactly N_n rises and falls, and that this excursion has exactly n upcrossings of level t_n (here n^2 is the correct scaling for the number of rises and falls of an excursion with n upcrossings of level t_n). Conditioned on this event, the ERW excursion is the contour process of a random tree $\mathcal{T}^+_{t_n,n,N_n}$, which is the tree $\mathcal{T}_{t_n,n}$ continued until extinction – that is, conditioned to have a total number of individuals equal to N_n. Let us first prove the proposition for $\mathcal{T}^+_{t_n,n,N_n}$.

We can show that the probability $\mathrm{P}(A_{N_n})$ decreases no faster than polynomially in $1/N_n$. Hence, by the 'large deviation' result mentioned earlier, the empirical distribution θ_{m,N_n} of $2m$-tuples conditioned on A_{N_n} converges to μ_m. This implies the weaker result that, for J_n uniform on $\{2, 4, 6, \dots, 2N_n\}$,

$$(\xi_{J_n-m+1}, \dots, \xi_{J_n}, \dots, \xi_{J_n+m}) \xrightarrow{\mathrm{D}} \mu_m, \tag{10}$$

where the left-hand side is conditioned on A_{N_n}. However, this implies that, relative to a uniform random individual I_n in $\mathcal{T}^+_{t_n,n,N_n}$, any aspect of the 'local structure' of the tree that is determined

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

by the segment of the contour process of length $2m$ centered on that individual will converge in distribution to the same aspect of the local structure of $\tilde{\mathcal{T}}$. By taking m to be large and appealing to (9), we see that the proposition holds for $\mathcal{T}^{+}_{t_n,n,N_n}$.

To complete the proof, it suffices to show that the proposition holds for the stopped tree $\mathcal{T}_{t_n,n,N_n}$. Unfortunately, this does not follow directly from the unstopped case, because a nonnegligible fraction of all individuals in $\mathcal{T}^{+}_{t_n,n,N_n}$ will be descendants of the n individuals alive at time t_n after the origin. Instead, fix a pair of small numbers δ_1 and δ_2, $0 < \delta_1 < \delta_2$, and consider the segments of the contour process $\mathcal{C}^+$ of $\mathcal{T}^{+}_{t_n,n,N_n}$ defined as follows:

- s_1 is the segment of $\mathcal{C}^+$ ending at the first upcrossing of $(1-\delta_1)t_n$;
- s_2 is the segment of $\mathcal{C}^+$ extending from the subsequent downcrossing of $(1-\delta_2)t_n$ to the next upcrossing of $(1-\delta_1)t_n$;
- s_3 is the segment of $\mathcal{C}^+$ extending from the subsequent downcrossing of $(1-\delta_2)t_n$ to the next upcrossing of $(1-\delta_1)t_n$;
- s_i, $i = 4, \dots, N-1$, are defined similarly;
- s_N is the segment of $\mathcal{C}^+$ beginning at the final downcrossing of t_n.

Conditional on the event A_{N_n}, there is some conditional distribution of starting and ending positions for each segment. Given these positions, each segment is distributed in the same way as an ERW process conditioned on having the first upcrossing of a certain level occur after a prescribed number of steps. The number of the segments is stochastically bounded as $n \to \infty$, meaning that the probability of the conditioning event for each segment is still only polynomially small in $1/(\text{length of the segment})$. Thus, separately for each segment, we can show, as above, that the contour process satisfies (10) for J_n uniform on that segment. Since, in the $n \to \infty$ limit, these segments comprise a proportion $1 - \varepsilon(\delta_1, \delta_2)$ of the entire contour process of $\mathcal{T}_{t_n,n,N_n}$, where $\varepsilon \to 0$ as $\delta_1, \delta_2 \to 0$, we can deduce the proposition for the stopped process $\mathcal{T}_{t_n,n,N_n}$, completing the proof.

We now state the parallel local limit result for $\mathcal{T}_n$ relative to a random *extant* individual, omitting a similar proof. In this setting, the relevant limit infinite tree, which we again call $\tilde{\mathcal{T}}$, is the following variation of the $\tilde{\mathcal{T}}$ above. The distinguished individual has an exponential(1) age at time 0. Its ancestors and their descendants are all as described above, except that now the infinite tree $\tilde{\mathcal{T}}$ is stopped at time 0.

Proposition 4. *Let I_n denote a uniform random extant species from the clade $\mathcal{T}_n$. Then, as $n \to \infty$ for fixed $\sigma > 0$,*

$$\tilde{\mathcal{T}}_n(I_n, [-\sigma, 0]) \xrightarrow{\mathrm{D}} \tilde{\mathcal{T}}[-\sigma, 0].$$

We can now perform exact calculations of probabilities for the distinguished individual in $\tilde{\mathcal{T}}$ that represent the $n \to \infty$ asymptotic results for a random extant individual in $\mathcal{T}_n$. Here is a simple example of possible calculations within $\tilde{\mathcal{T}}$.

Corollary 11. *For the distinguished individual in $\tilde{\mathcal{T}}$,*

(a) *the probability that its parent is extant equals $\frac{1}{2}$, and*

(b) *the probability that at least one of its ancestors is extant equals $1 - \mathrm{e}^{-1}$.*

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

Proof. (a) The probability that the parent of the distinguished individual is alive at time 0 is simply $\mathrm{P}(\xi_1 < \xi_2)$, where ξ_1 is the age of the distinguished individual and ξ_2 is the subsequent lifetime of its parent after the birth. Because ξ_1 and ξ_2 are independent exponential(1) random times, we have $\mathrm{P}(\xi_1 < \xi_2) = \frac{1}{2}$ by symmetry.

(b) To calculate the probability that no ancestor of the distinguished individual is still alive, we need only note that the times at which some ancestor originates form a rate-1 Poisson process, and that an ancestor originating at time s before present has a chance equal to e^{-s} to be extant at present. Therefore, the number of extant ancestors has a Poisson distribution with mean

$$\int_0^\infty (\mathrm{e}^{-s} \times 1)\,\mathrm{d}s = 1$$

and, thus, takes the value 0 with probability e^{-1}.

6. Final notes

Note 7. There are two well-known other models that, like ours, are simple in the specific sense of having no scale-free parameters. A basic model for speciations without extinctions is the Yule process [19], which is the continuous-time pure-birth process with constant rate. The Moran model [6, Section 3.3, pp. 84–89], the basic neutral model used in population genetics, can also be applied to macroevolution [9]; this model involves a process of uniform random speciations and extinctions in a population of fixed size. In the large population limit, this model run backwards in rescaled time has the continuous-time coalescent model as its lineage tree (see [12] for a recent survey). The essentially different aspect of our model is that species numbers are allowed to fluctuate freely, permitting a much broader range of quantities to be studied. Biologists have studied more elaborate models, mostly in one of two categories. *Exponential growth* models are exemplified by the linear birth–death chain model for species numbers, where the transition numbers are $\lambda_i = \lambda i$ and $\mu_i = \mu i$. This leads to a model [13] with three parameters, λ, μ, and t_*, where t_* is time of origin of the clade. *Logistic* stochastic models posit a logistic-shaped curve for species numbers, and also require three or four parameters for their specification. In contrast, it is the simplicity of a one-parameter model, and the desire to avoid the particular biology presumptions underlying exponential growth or logistic-type models, that motivate our particular model.

Note 8. As already mentioned, our model of $\mathcal{T}_n$ and $\mathcal{A}_n$ has considerable variability between realizations. This variability is partly an artefact of the uniform prior on the time of origin of the process, but it serves a useful purpose in emphasizing that radically different appearances of real-world trees might logically be just chance variation without biological significance. Wollenberg *et al.* [18] studied a model similar to ours (critical branching conditioned on n extant species) via simulation, but handled the issue of the time of origin in a different way, by taking for it the deterministic time t_n that is the maximum likelihood estimator of the time of origin. This assumption understates variability.

Note 9. Our model is qualitatively similar (in the sense of orders of magnitude) to the Moran model for all quantities which can be studied in that model. Our results involving local weak limits, in Sections 3.3 and 3.4, are exactly the same as in a continuized Moran model, because our model converges (in the $n \to \infty$ limit) to the continuized Moran model over size-$o(n)$ time intervals extending backwards from the present.

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

Note 10. Branching processes conditioned on the *total* population have been studied extensively [5]. Conditioning on population size n at time t after origination occurs implicitly in works such as [8], but our device of having a uniform prior for the origination time has apparently not been studied before. It would be mathematically natural to study the quantities in this paper with our equal-rates linear birth–death process replaced by the Yule process or by the unequal-rates linear process. A referee has observed that many of the results in this paper have parallels for the Yule process, albeit with different orders of magnitude. For instance, $\mathrm{E}[T_n^{\mathrm{mrca}}] \sim \log n$ and $T_n^{\mathrm{mrca}} - \log n$ has a distributional limit with density $\exp(-2s - \mathrm{e}^{-s})$.

Note 11. Neutral models like ours are unrealistic for large clades, by the following reasoning. For an n-species clade, in our model its time of origin is of order n time units before the present, as shown in Corollary 3. The time unit represents the mean species lifetime, typically estimated to be a few million years. Thus, our model predicts the origin of an n-species clade to be at least n million years ago, which is known from fossil data be an overestimate for most clades of size $n \geq 100$.

Note 12. The local point process limit in Corollary 4 is a simple instance of a general notion of *local weak convergence* of graphical structures associated with point processes on $\mathbb{R}^d$ or abstract spaces; see [3] for more sophisticated examples. In particular, Proposition 3 fits the general setting of *asymptotic fringe distributions*, which exist for many different models of random tree [1].

Note 13. A sequel [10] will treat phylogenetic trees on higher order taxa, emphasizing how the choice of classification scheme may affect the distribution of tree shape.

Acknowledgements

We thank Maxim Krikun for helpful comments and a referee for suggesting expository improvements and formula (6).

References

[1] Aldous, D. J. (1991). Asymptotic fringe distributions for general families of random trees. *Ann. Appl. Prob.* **1,** 228–266.

[2] Aldous, D. J. (1993). The continuum random tree. III. *Ann. Prob.* **21,** 248–289.

[3] Aldous, D. J. and Steele, J. M. (2004). The objective method: probabilistic combinatorial optimization and local weak convergence. In *Probability on Discrete Structures* (Encyclopaedia Math. Sci. **110**), ed. H. Kesten, Springer, Berlin, pp. 1–72.

[4] Dembo, A. and Zeitouni, O. (1992). *Large Deviations and Applications*, 2nd edn. Jones and Bartlett, Boston, MA.

[5] Duquesne, T. and Le Gall, J.-F. (2002). Random trees, Lévy processes and spatial branching processes. *Astérisque* **281,** vi+147.

[6] Ewens, W. J. (1979). *Mathematical Population Genetics*. Springer, Berlin.

[7] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd edn. John Wiley, New York.

[8] Geiger, J. (2000). Poisson point processes in size-biased Galton–Watson trees. *Electron. J. Prob.* **5,** 12pp.

[9] Hey, J. (1992). Using phylogenetic trees to study speciation and extinction. *Evolution* **46,** 627–640.

[10] Krikun, M., Popovic, L. and Aldous, D. J. (2005). Stochastic models for phylogenetic trees on higher order taxa. In preparation.

[11] Le Gall, J.-F. (1989). Marches aléatoires, mouvement brownien et processus de branchement. In *Séminaire de Probabilités XXIII* (Lecture Notes Math. **1372**), Springer, Berlin, pp. 258–274.

[12] Möhle, M. (2000). Ancestral processes in population genetics. *J. Theoret. Biol.* **204,** 629–638.

[13] Nee, S., May, R. M. and Harvey, P. H. (1994). The reconstructed evolutionary process. *Philos. Trans. R. Soc. London B* **344,** 305–311.

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press

[14] Neveu, J. and Pitman, J. (1989). Renewal property of the extrema and tree property of a one-dimensional Brownian motion. In *Séminaire de Probabilités XXIII* (Lecture Notes Math. **1372**), Springer, Berlin, pp. 239–247.

[15] Page, R. D. M. and Holmes, E. C. (1998). *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science, Oxford.

[16] Pitman, J. (2002). Combinatorial stochastic processes. Tech. Rep. 621, Department of Statistics, University of California, Berkeley.

[17] Popovic, L. (2004). Asymptotic genealogy of a critical branching process. *Ann. Appl. Prob.* **14,** 2120–2148.

[18] Wollenberg, K., Arnold, J. and Avise, J. C. (1996). Recognizing the forest for the trees: testing temporal patterns of cladogenesis using a null model of stochastic diversification. *Molec. Biol. Evol.* **13,** 833–849.

[19] Yule, G. U. (1924). A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis. *Philos. Trans. R. Soc. London B* **213,** 21–87.

https://doi.org/10.1239/aap/1134587755 Published online by Cambridge University Press