

A Critical Branching Process Model for Biodiversity ^{*}

David Aldous[†]

Lea Popovic

Department of Statistics
University of California
367 Evans Hall # 3860
Berkeley, CA 94720-3860

IMA
University of Minnesota
400 Lind Hall
Minneapolis, MN 55455

September 24, 2005

Abstract

Motivated as a null model for comparison with data, 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 $(0, \infty)$. After that origin, the process of extinctions and speciations is a continuous-time critical branching process of constant rate, conditioned on having the prescribed number n of species at the present time. We study various mathematical properties of this model as $n \rightarrow \infty$ limits: time of origin and of most recent common ancestor; pattern of divergence times within lineage trees; time series of numbers of species; number of extinct species in total, or ancestral to extant species; and “local” structure of the tree itself. We emphasize several mathematical techniques: associating walks with trees, a point process representation of lineage trees, and Brownian limits.

***MSC 2000 subject classification.** Primary: 60J85; Secondary: 60J65, 92D15

Key words and phrases. biodiversity, Brownian excursion, contour process, critical branching process, genealogy, local weak convergence, phylogenetic tree, point process

[†]Research supported by N.S.F. Grant DMS-0203062

1 Introduction

1.1 A big picture

This paper forms part of a larger project, which we first outline. There is a substantial literature on comparing data on different aspects of *biodiversity* or *macroevolution* – the evolutionary history of speciations and extinctions – with the predictions of simple “pure chance” stochastic models. Available data includes

- fossil time series – fluctuations in number of taxa over time;
- shapes of phylogenetic trees on extant species (Mooers and Heard [14] provide an extensive survey);
- the distribution of number of species per genus.

The fit of simple models, and of more elaborate models incorporating conjectured biological process, have been studied in these contexts. While data-motivated models are scientifically natural, a mathematical aesthetic suggests a somewhat different approach: start with a “pure chance” model which encompasses simultaneously all the kinds of data that one might hope to find. Here are two instances of what one would like such a model to provide.

- Joint description of the phylogenetic tree on an extant clade of species, its extension to the tree on an observed small proportion of extinct species, and the (unobserved) entire tree on all extinct species.
- Joint description of fossil time series at different levels of the taxonomic hierarchy.

We emphasize the latter because paleontology literature tends to assume that a model can be applied at any level, without enquiring whether this assumption is logically self-consistent.

Our purpose in the larger project is to present what is arguably the mathematically fundamental such model. The underlying model at the species level is simple – a critical branching process conditioned to have n lineages at the present time. This is the subject of the present paper, which is the part of the project most closely related to classical and contemporary applied probability. In a subsequent paper aimed at a more biological audience we will describe how the model extends to higher-order taxa by assuming each new species has some probability of founding a new higher-order taxon; we will consider several explicit classification schemes emphasizing desiderata such as monophyletic groups.

Conceptually, this is a *neutral* model which does not incorporate conjectured biological process such as intrinsic tendency for species numbers to

increase, differential speciation or extinction rates, or ecological constraints on numbers of species. For well-understood mathematical reasons (see section 6) neutral models like ours are implausible for large clades. In a sense, the model seems most appropriate as a “null hypothesis” for small clades, at the recent fringe of the Tree of Life, or for a geological period free of mass extinctions and their aftermath.

Biological questions motivating our model, and suggested by the results of analysis of our model, will be treated in detail elsewhere, so we give just a brief mention here. Phylogenetic trees on extant species are nowadays based on molecular data; technical aspects of tree reconstruction form a large and important subject [10, 22]. But that is not our focus. Let us assume that in the near future we will have a large database of essentially correct phylogenetic trees, and also assume these include the time points of divergence of lineages (rather than just the “shape” of the tree). How might one use such a database?

(i) *Inference about a particular clade.* If we have no direct knowledge about extinct species, then we cannot observe past fluctuations of number of species with time, and cannot observe the time of origin of the clade (typically longer ago than the observable time of most recent common ancestor of extant species). Inference about such quantities requires some stochastic model; given a model, one can use the observed phylogenetic tree to make inferences.

(ii) *Statistical properties of phylogenetic trees in general.* In what systematic way do real phylogenetic trees differ from predictions of a simple model like ours which treats macroevolution as a purely random process, and what is the biological significance of such differences?

1.2 Standard models

Ours is, roughly speaking, the *third* simplest model one might devise, so let us first recall the two simpler models.

The Yule model. Yule [24] proposed the basic model for speciations without extinctions. Initially there is one species. Thereafter, independently for each existing species, new species originate as “daughter” species at constant rate λ (i.e. at the times of a Poisson (rate λ) process). So for given n one can get a model for an n -species tree by taking the present as a random time at which the number of species equals n . (The associated continuous-time Markov chain counting number of species is often called the *Yule process*, though its origin as a model for species is often forgotten.)

The Moran/coalescent model. These models, developed and extensively used in population genetics, can also be applied to macroevolution (see e.g. [11]). In the Moran model ([8] sec. 3.3) the number of coexisting species is fixed at n . At successive discrete times, one randomly-chosen species goes extinct and another randomly chosen one speciates. Implicit in this model (run from the indefinite past until the present) is a model for the phylogenetic tree on the n extant species; for large n , with suitable rescaling of the time unit, the phylogenetic tree approximates the much-studied continuous-time *coalescent* model. To describe the coalescent model, we run time backwards from the present, starting with n “lines of descent”; in a time interval dt , each *pair* of lines of descent has chance dt to merge (“coalesce”) into one line, and we continue until reaching a single *most recent common ancestor*. See [13] for a recent survey.

Why a third model? Obviously many basic inference questions mentioned earlier – about fluctuations in past numbers of (extinct) species, for instance – are not satisfactorily handled within the models above. Biologists have studied more elaborate models, mostly in one of two categories. (We will give a more detailed account elsewhere, but the bottom line is that the actual fit of real-world data to parametric models has not been studied as definitively as one might have expected.) *Exponential growth* models are exemplified by the linear birth-and-death chain model for species numbers ($\lambda_i = \lambda i, \mu_i = \mu i$). This leads to a model [15] with 3 parameters (λ, μ, t_*) where t_* is time of origin of clade. *Logistic* stochastic models posit a logistic-shaped curve for species numbers, and also require 3 or 4 parameters to specify. In contrast, the model we study (described carefully in Section 2) has only 1 parameter (mean species lifetime). It is this simplicity, and the desire to avoid the particular biological presumptions underlying exponential growth or logistic type models, that motivates our particular model.

1.3 Outline of results

A *clade* is the set of all species which are descendants of some (typically extinct) species. The succinct description (to be elaborated in Section 2) of our model for a phylogenetic tree \mathcal{T}_n on a clade with n extant species is as follows.

The origin of the clade is a random time in the past, whose (improper) distribution is uniform on $(0, \infty)$. After that origin, the process of extinctions and speciations is a continuous-time

critical branching process of constant rate, conditioned on having the prescribed number n of species at the present time.

The conditioning is conceptually important: given a real phylogenetic tree on 23 extant species, we want to be able to compare it to the predictions of a stochastic model generating trees on exactly 23 extant species. The uniform prior (for time of origin) avoids the necessity to introduce a second parameter into the model.

There is a vast mathematical literature on branching processes, but we haven't found detailed discussion of any very similar model. In the biological literature, Wollenberg et al [23] give a simulation study of a similar model. On the other hand, this model is clearly open to analysis by the known techniques of applied probability. We exploit one particular modern approach to classical branching process theory: representing trees as walks. See Pitman [19] for a recent survey. This both leads to an exact "point process" description of lineage tree distributions (Proposition 1) and permits us to study asymptotics via weak convergence to Brownian motion. This methodology is known to specialists in other aspects of random trees but is perhaps less familiar in the subject of phylogenetic trees, so we try to explain the key ideas carefully even though they are not entirely new.

Our results describe distributional properties of various aspects of the tree \mathcal{T}_n .

- The lineage tree, via exact formulas (Proposition 1), "global limits" (Corollaries 3 and 4), and "local limits" (Corollaries 6 and 7).
- The time series of number of species (Lemma 8), the maximum number of coexisting species (Corollary 9), and the total number of extinct species (Corollary 10).
- The local limit structure of the complete (i.e. including extinct species) phylogenetic tree, relative to either a typical extant species (Proposition 19) or a typical extinct species (Proposition 18).
- The joint distribution of time of origin of clade and time of most recent common ancestor (Corollary 5), joint also with the number of species alive at the time of most recent common ancestor (Corollary 14).
- The number of extinct species ancestral to some extant species (Corollary 17).

Finally we should admit that the whole paradigm of studying $n \rightarrow \infty$ asymptotics is rather unnatural, because the model is biologically unrealistic

for large n , but one can hope that the approximations implicit in asymptotic results are qualitatively correct for smaller values of n . Our web site [1] shows Monte Carlo simulations for $n = 8, 12, 20$ with 10 repetitions. One can check that numerical values are broadly consistent with the asymptotic predictions.

2 Model and notation

In stating and deriving mathematical results we use the traditional language of branching processes (*individuals, children, births, . . .*) even though we are envisaging species and so should be writing (*species, daughter species, speciations, . . .*).

Let \mathcal{T} be a continuous time critical branching process (CBP) starting with one individual. According to this process each individual lives for an Exponential (rate λ) time, for some $\lambda > 0$, during which it gives birth at times of an independent Poisson (rate λ) process. After birth all individuals behave independently of each other. We can and will scale time so that $\lambda = 1$; so the time unit is interpreted as mean species lifetime.

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 ([9] §XVII.10.11) gives a modified Geometric distribution

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

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. Within a process like $\mathcal{T}_{t,n}$ or \mathcal{T}_n below, we use the notational convention that “time s ” means time s before 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 time 0, the present time. Our previous verbal definition of our model \mathcal{T}_n as a Bayes posterior (for \mathcal{T} started at a uniform past time $t \in (0, \infty)$ and conditioned on having n individuals at the present time 0) now becomes the following rigorous definition. Fix $n \geq 1$.

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

Using (1) and the calculus fact

$$\int_0^\infty \frac{s^{n-1}}{(1+s)^{n+1}} ds = n^{-1}$$

turns this into

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

Within the random tree \mathcal{T}_n , the time parameter s decreases from a random “time of origin” T_n^{or} to 0, where by the formula above, T_n^{or} has density function

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

We shall refer to \mathcal{T}_n and $\mathcal{T}_{t,n}$ as the *complete trees*. Returning to biological terminology, a complete tree records the birth times and every (extinct or extant) species in a clade and the extinction time of extinct species. Every realization of a complete tree also uniquely determines a realization of a *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 *most recent common ancestor* of the n extant species, to the present time 0. (The lineage tree is what is usually called the *phylogenetic tree*, though logically all the trees under consideration are different kinds of phylogenetic tree.)

3 Point process representations of lineage trees

3.1 An exact description

It is perhaps remarkable that there is a useful exact description of the lineage tree $\mathcal{A}_{t,n}$, based on a certain *point process representation* illustrated in Figure 1. Consider an arbitrary lineage tree on n species. Draw the tree as in Figure 1; recursively from the top down, at each divergence point of lineages choose randomly which branch is drawn on the left and which on the right. After drawing the tree, label species as $1, 2, \dots, n$ in left-to-right order. Each divergence of lineages involves adjacent contiguous blocks of species, say $\{i, i+1, \dots, j\}$ and $\{j+1, j+2, \dots, k\}$, and occurs at some time s . We mark the occurrence of this divergence by a mark \times at coordinates $(j + \frac{1}{2}, s)$

and then draw the combined lineage as a vertical line upwards from the mark.

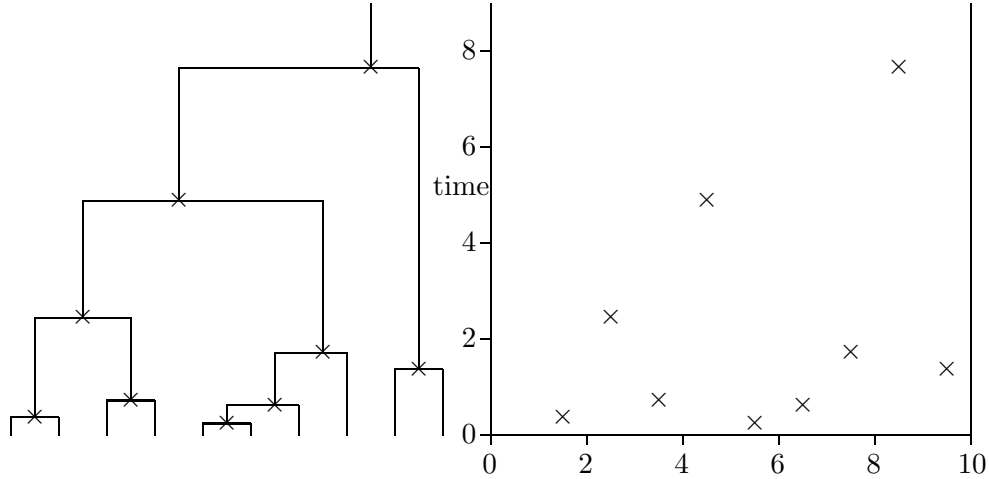


Figure 1. The point process representation of a lineage tree on $n = 10$ species.

The advantage of this precise way of drawing the tree is that one can clearly reconstruct the tree from the coordinates $\{(i + \frac{1}{2}, s_i), 1 \leq i \leq n - 1\}$ of the marks. So the distribution of the point process of marks serves to specify the distribution of the lineage tree.

Proposition 1 ([20] Lemma 3). Fix $n \geq 2$ and $t > 0$. The point process $\{(i + \frac{1}{2}, h_i), 1 \leq i \leq n - 1\}$ where the (h_i) are i.i.d. with density function

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

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

The derivation of this result will be explained in Section 5, where the underlying *contour process* is exploited further.

We are mostly concerned with the lineage tree \mathcal{A}_n , which by (2) has a mixture representation

$$\mathbb{P}(\mathcal{A}_n \in \cdot) = \int_0^\infty \mathbb{P}(\mathcal{A}_{t,n} \in \cdot) q_n(t) dt \quad (5)$$

where $q_n(t)$ is the density function (3) of T_n^{or} . One can get exact formulas for various attributes of \mathcal{A}_n . Consider for instance the number of lineages

at time s . Because each divergence time creates one extra lineage, it is clear that within $\mathcal{A}_{t,n}$ this number of lineages is distributed as

$$1 + \text{Binomial}(n - 1, \bar{F}_t(s))$$

for

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

Thus within \mathcal{A}_n the distribution is the mixture of Binomials implied by (5). Similarly, the exact distribution of the time T_n^{mrca} of most recent common ancestor is

$$\text{P}(T_n^{\text{mrca}} \leq u) = \int_0^\infty (1 - \bar{F}_t(u))^{n-1} q_n(t) dt.$$

In this paper we focus on $n \rightarrow \infty$ asymptotics, which may give more conceptual insight than do complicated exact formulas. As we see below, it is useful to distinguish two kinds of asymptotics: *global limits* refer to times of order n , whereas *local limits* refer to times of order 1.

3.2 The global limit point process

From the formula (3) for $q_n(t)$ we calculate: if $t_n/n \rightarrow t > 0$ then

$$nq_n(t_n) = \frac{n^2}{(1 + t_n)^2} \left(1 - \frac{1}{1 + t_n}\right)^{n-1} \xrightarrow{n \rightarrow \infty} t^{-2} e^{-\frac{1}{t}}.$$

The limit is the density function of the *Inverse Exponential* IE(1) distribution, that is to say of $1/\xi$ where ξ has Exponential(1) distribution. So we have shown

Lemma 2. $n^{-1}T_n^{\text{or}} \xrightarrow{d} T^{\text{or}}$, say, where the limit T^{or} has IE(1) distribution.

Now reconsider Figure 1. To obtain a global limit we want to rescale time by a factor n and we want to rescale the left-to-right positions of marks to fit into a unit interval $[0, 1]$, implying they also must be rescaled by a factor n . Thus the original point process of marks $\{(i + \frac{1}{2}, s_i), 1 \leq i \leq n - 1\}$ is rescaled to $\{(\frac{i + \frac{1}{2}}{n}, \frac{s_i}{n}), 1 \leq i \leq n - 1\}$. In the setting of Proposition 1, the relevant calculation is:

$$\text{if } s_n/n \rightarrow s > 0 \text{ and } t_n/n \rightarrow t > 0 \text{ then } n^2 f_{t_n}(s_n) \rightarrow s^{-2}$$

and the following limit behavior is intuitively clear.

Corollary 3 ([20] Lemma 4, Theorem 5). *Let $t_n/n \rightarrow t > 0$. The rescaled point process $\{(\frac{i+\frac{1}{2}}{n}, \frac{h_i}{n}), 1 \leq i \leq n-1\}$ associated with the lineage tree $\mathcal{A}_{t_n, n}$ converges in distribution to the Poisson point process $(\pi_{1,t}, \text{say})$ whose intensity measure is $\nu(dl \times ds) = dl s^{-2} ds \mathbf{1}_{[0,1] \times (0,t)}$.*

The limit $\pi_{1,t}$, 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 regions away from this boundary. Figure 2 indicates visually how the Poisson point process limit defines a limit random tree which is a kind of “continuum tree” with a lineage for each real $l \in (0, 1)$, though we do not seek to formalize this idea.

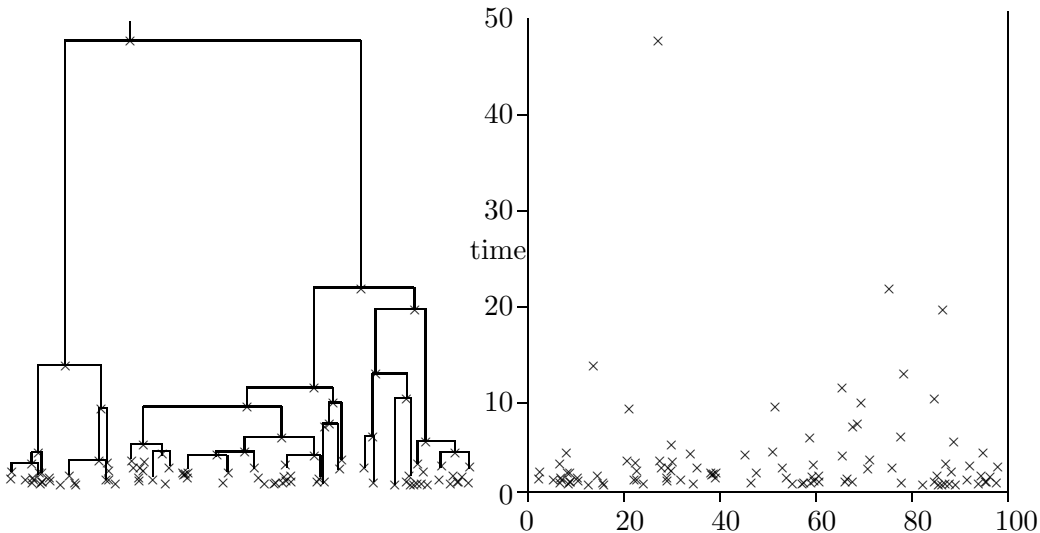


Figure 2. The point process $\pi_{1,t}$ on the right represents the lineage tree of a continuum of species on the left.

The mixture representation (5) and Corollary 3 immediately imply a global limit theorem for \mathcal{A}_n . To state it, let T^{or} have $\text{IE}(1)$ distribution. Define a Cox point process (π_1, say) on $(0, 1) \times (0, \infty)$ as follows. Given $T^{\text{or}} = t$, let π_1 be a Poisson point process with the law of $\pi_{1,t}$.

Corollary 4. *The rescaled point process $\{(\frac{i+\frac{1}{2}}{n}, \frac{s_i}{n}), 1 \leq i \leq n-1\}$ 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 quick application of this global limit theorem.

Corollary 5. *The limit joint behavior of T_n^{or} and T_n^{mrca} is given by*

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

where the limit law has joint density

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

The marginal density of T^{mrca} is

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

The limit joint distribution can alternatively be expressed as $(T^{\text{or}}, T^{\text{mrca}}) \stackrel{d}{=} (\frac{1}{\xi_1}, \frac{1}{\xi_1 + \xi_2})$ where ξ_1, ξ_2 are i.i.d. $\text{Exponential}(1)$.

Proof. Corollary 4 implies convergence in distribution to the limit $(T^{\text{or}}, T^{\text{mrca}})$ in which T^{mrca} is defined as the maximum height (that is, maximum second coordinate) of any point of π_1 . Given that $T^{\text{or}} = t$, π_1 is distributed as a Poisson point process $\pi_{1,t}$ with intensity measure $\nu(dl \times ds) = dl s^{-2} ds \mathbf{1}_{[0,1] \times (0,t)}$. Consequently, for the conditional law of T^{mrca} given $T^{\text{or}} = t$ we have

$$\begin{aligned} \mathbb{P}(T^{\text{mrca}} \leq s | T^{\text{or}} = t) &= \mathbb{P}(\{\pi_{1,t} \cap [0, 1] \times (s, t)\} = \emptyset) \\ &= \exp\left(-\int_s^t u^{-2} du\right) \\ &= e^{\frac{1}{t} - \frac{1}{s}}, \quad 0 < s < t. \end{aligned}$$

So

$$\mathbb{P}(T^{\text{mrca}} \leq s, T^{\text{or}} \in dt) = e^{\frac{1}{t} - \frac{1}{s}} \mathbb{P}(T^{\text{or}} \in dt) = t^{-2} e^{-\frac{1}{s}} dt, \quad 0 < s < t$$

implying the formula for joint density. The remaining calculations are straightforward. \square

3.3 The local limit point process

There is a different limit regime in which time is not rescaled. This 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. The relevant calculation is that, in the setting of Proposition 1, if $t_n \rightarrow \infty$ then

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

Consider the point process on $(\mathbf{Z} + \frac{1}{2}) \times (0, \infty)$ consisting of points $\{(i + \frac{1}{2}, \eta_i), i \in \mathbf{Z}\}$ for i.i.d. (η_i) with density $f(s) = (1 + s)^{-2}$. As illustrated in Figure 3, the point process defines an infinite tree, \mathcal{A}_∞ say, on lineages labeled by \mathbf{Z} .

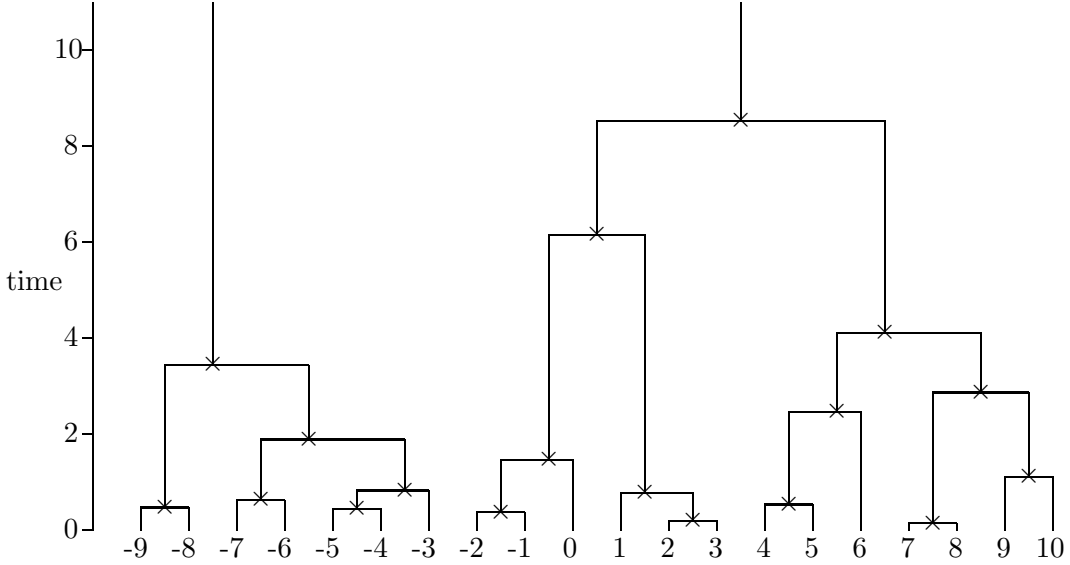


Figure 3. A realization of part of \mathcal{A}_∞ , approximating the local structure of \mathcal{A}_n for large n . The 2 visible ancestral lineages diverged at around time 16.

Proposition 1 and the calculation above easily imply the first assertion below; the second assertion follows from the mixture representation (5), where in this setting the mixing makes no difference.

Corollary 6. *Let $t_n \rightarrow \infty$ and let U_n be uniform on $\{1, 2, \dots, n\}$ independent of $\mathcal{A}_{t_n, n}$. Write $\{(U_n + i + \frac{1}{2}, s_{U_n+i}), i \in \mathbf{Z}\}$ for the point process associated with the lineage tree $\mathcal{A}_{t_n, n}$, centered at lineage U_n , where $s_j = 0$ for j outside $[1, n]$. Then as $n \rightarrow \infty$ this point process converges in distribution to the point process $\{(i + \frac{1}{2}, \eta_i), i \in \mathbf{Z}\}$ defining \mathcal{A}_∞ . The same result holds for \mathcal{A}_n .*

Less formally, the structure of \mathcal{A}_∞ around lineage 0 provides an asymptotic approximation to the structure of \mathcal{A}_n around a random lineage.

3.4 Some local calculations

We record some elementary calculations within \mathcal{A}_∞ , reflecting aspects of the $n \rightarrow \infty$ behavior of the lineage trees \mathcal{A}_n . For a lineage at time s we call the present (time - 0) number of species descending from this lineage the *size* of this lineage. The $n \rightarrow \infty$ limit of n^{-1} (number of lineages in \mathcal{A}_n at time s) is what we will call the *density of lineages* in \mathcal{A}_∞ at time s .

Corollary 7. [*Some calculations for \mathcal{A}_∞ .*]

(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 Geometric($(1 + s)^{-1}$) distribution;*

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

(c) *as s decreases (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 branchpoint has 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 branching points at times greater than s

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

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

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

since this is the distribution of distances between branchpoints at heights greater than s .

(b) As s increases (time runs backwards) the probability of a lineage merging with another lineage is

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

because such a merger occurs in $[s, s + ds]$ when one of the two branchpoints separating the given lineage from its neighboring lineages, which must be at

height $\geq s$, occurs during $[s, s+ds]$, and this has chance $f(s)ds/G(s)$ for each branchpoint. Moreover, if a lineage merges at s then (independent of the size of the first lineage) the size of the second lineage has Geometric($(1+s)^{-1}$) distribution above.

(c) As s decreases (time runs forwards), the unconditional rate of mergers of clades of sizes k_1, k_2 at time t (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 observe by considering the required heights of branchpoints for this event to occur. Similarly the number of size k_1+k_2 lineages at time t , relative to number of species, equals

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

Thus the rate of splitting of a size k_1+k_2 lineage into two lineages of sizes k_1, k_2 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

$$b_k(s) = \frac{k-1}{s(1+s)}$$

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 subclade lineage is Uniform on $\{1, 2, \dots, k-1\}$. \square

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 which is not stationary, or which is conditioned on a terminal value, the time-reversal is typically *non-homogeneous*. So Lemma 8 below highlights a special feature of our processes.

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

$$q_{i,i+1} = q_{i,i-1} = i. \tag{6}$$

Recall the definition of the complete tree \mathcal{T}_n . Write $(N_n(s), T_n^{\text{or}} \geq s \geq 0)$ for the associated process which counts the number of species at time s before present. The next lemma makes precise a sense in which the process $(N_n(s))$ is a time-reversal of the chain (6) started at 0.

Lemma 8. Let $(\widehat{N}_n(s), 0 \leq s \leq T_n^0)$ be the continuous-time chain (6) with $\widehat{N}_n(0) = n$, run until the first hitting time T_n^0 on state 0. Then

$$(N_n(s), T_n^{\text{or}} \geq s \geq 0) \stackrel{d}{=} (\widehat{N}_n(s), 0 \leq s \leq T_n^0).$$

Proof. We verify that $(\widehat{N}_n(s), 0 \leq s \leq T_n^0)$ is the time-reverse of the population size process by checking probabilities of primitive events (see Section 5.4 for more sophisticated views). Fix $s_M > s_{M-1} > \dots > s_1 > s_0 = 0$ and positive integers $1 = k_M, k_{M-1}, \dots, k_1 = n$ 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 $[s_m, s_m + ds_m]$ ($\forall M \geq m \geq 1$) and makes no other jumps

has measure

$$ds_M \times \prod_{m=M}^2 \left(e^{-k_m(s_m - s_{m-1})} k_m ds_{m-1} \right) \times e^{-k_1 s_1}$$

where the first term ds_M comes from the uniform Bayes prior. For the reversed process, the event

as s increases, $\widehat{N}_n(s)$ jumps from k_m to k_{m+1} during $[s_m, s_m + ds_m]$ ($\forall 1 \leq m \leq M$) and makes no other jumps

has probability

$$\prod_{m=1}^M \left(e^{-k_m(s_m - s_{m-1})} k_m ds_m \right).$$

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

We now observe two simple consequences of this time-reversal identity. The process $(\widehat{N}_n(s), 0 \leq s \leq T_n^0)$ is a skip-free martingale started at n and run until hitting 0, so by the hitting time formula for martingales

$$\mathbb{P} \left(\max_{0 \leq s \leq T_n^0} \widehat{N}_n(s) \geq c \right) = \frac{n}{c}, \quad c \geq n.$$

So Lemma 8 implies

Corollary 9.

$$\mathbb{P} \left(\max_{T_n^{\text{or}} \geq s \geq 0} N_n(s) \geq c \right) = \frac{n}{c}, \quad c \geq n.$$

Second, 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 $\widehat{N}_n(s)$ 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 $\widehat{N}_n(\cdot)$, which is just discrete-time simple symmetric random walk.

Corollary 10. *Within the model \mathcal{T}_n of a clade on n extant species, the total number N_n^{ext} of extinct species is distributed as $(D_n - n)/2$, where D_n is the hitting time to 0 for simple symmetric random walk started at n . In particular*

$$n^{-2}D_n \xrightarrow{d} \frac{1}{2}\tau_1$$

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

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

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

5 Exploiting the contour process

The results so far answer some, but not all, questions one might ask about the complete tree \mathcal{T}_n and the lineage tree \mathcal{A}_n in our model. For instance, the time-reversed process $(\widehat{N}_n(s))$ in Lemma 8 has a $n \rightarrow \infty$ rescaled limit, the well-known *Feller branching diffusion*, which therefore is the limit of the population size process $(N_n(s), T_n^{\text{or}} \geq s \geq 0)$. But this doesn't tell us anything about the relationship between $(N_n(s))$ and the lineage tree \mathcal{A}_n . For instance, a conceptually interesting question in the species context concerns $N_n(T_n^{\text{mrca}})$, the total number of species in the clade alive at the time of most recent common ancestor of the extant species. Recall also that the results of Section 3 were all based on the exact formula in Proposition 1, but we have not yet given any indication of its proof. It turns out that both these matters, and the local limit structure of the complete tree, can be studied using the *contour process*, described next.

5.1 The contour process

For any deterministic population process in continuous time, starting at the birth of one single individual, in which individuals have birth times, death times, and may give birth to children at distinct times, there is a particular

representation as a *rooted planar tree* which we now describe. Each individual is represented by an edge whose length equals that individual’s lifetime. The birth of an offspring corresponds to a branchpoint from its parent’s edge, and the length of the parent’s edge up to this branchpoint equals the age of the parent at this offspring’s birth time. From the branchpoint the offspring’s edge is drawn to the right of the parent’s edge. If the total population is finite, then we can label the individuals in a “depth-first” search order. This is illustrated in Figure 4 where tree edges have been drawn as full vertical lines and the branchpoints have been indicated by horizontal dotted lines.

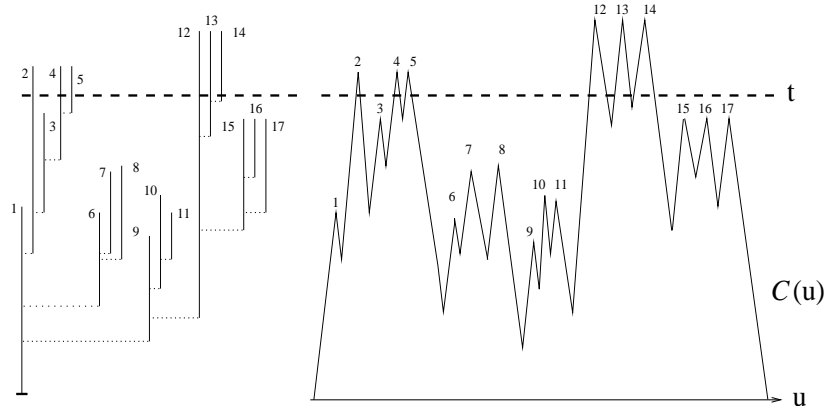


Figure 4. A realization of a tree $\mathcal{T}_{t,n}$ with $n = 6$ extant individuals (labeled $\{2, 4, 5, 12, 13, 14\}$) and its contour process representation $\mathcal{C}(u)$.

Associated with such a rooted planar tree is its *contour process* defined as follows (these ideas go back to Neveu and Pitman [17] and their broader significance can be seen in the lecture notes of Pitman [19]). 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, traverses each edge completely once upwards and once downwards, following the depth-first order (intuitively: clockwise around the edges of the tree) ending back at the root. So the contour process consists of alternating line segments of slopes $+1$ (“rises”) and slopes -1 (“falls”). The unit speed convention implies that heights in the contour process match the times in the population process (birth and death times are matched respectively by the local minima and local maxima in the contour process).

5.2 Contour processes associated with random trees

Recall that \mathcal{T} denotes the continuous-time critical branching process started at time 0 with one individual and continued until extinction. The relevance of contour processes is indicated in the next result of Neveu-Pitman-Le Gall [12, 17].

Proposition 11. *In the contour process of \mathcal{T} the sequence of rises and falls $(\xi_1, -\xi_2, \xi_3, -\xi_4, \dots, \xi_{M-1})$, excluding the last fall, has the distribution derived from a sequence $(\xi_i)_{i \geq 1}$ of independent Exponential(1) variables, for $M := \min\{m : \xi_1 - \xi_2 + \xi_3 - \xi_4 + \dots - \xi_m < 0\}$.*

Call this contour process $(\xi_1, -\xi_2, \dots, \xi_{M-1})$ an *ERW excursion*, for Exponential random walk. Accordingly call the infinite sequence $(\xi_1, -\xi_2, \xi_3, -\xi_4, \dots)$ an *ERW process*. Here is a classical result

Lemma 12. *Let H be the maximum height in an ERW excursion, or equivalently (by Proposition 11) the extinction time of \mathcal{T} . Then*

$$P(H > h) = (1 + h)^{-1}, \quad 0 < h < \infty.$$

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. By (1) the probability of this is $1 - h(1 + h)^{-1} = (1 + h)^{-1}$. \square

Before proceeding to new results let us indicate the proof [20] of Proposition 1, because our arguments in subsequent sections will use similar ideas. Fix t and n . 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 $(\mathcal{T}_{t,n}^+, \text{ say})$ which is the CBP conditioned on having 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 “direction of time parameter” convention, and except for the fact that in $\mathcal{T}_{t,n}$ the process terminates with the n individuals at the present time, whereas in $\mathcal{T}_{t,n}^+$ the process of descendants of these n individuals continues until extinction. But the latter difference plays no role in the following argument. The heights of the minima between each pair of successive upcrossings in Figure 4 match the divergence of lineages of that pair of extant individuals. Marking these heights at regular horizontal interval spacings gives exactly the point process $\mathcal{A}_{t,n}$ as in Figure 1 except for reflecting 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 moreover are distributed exactly as the reflection of the original ERW excursion that is 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 as the maximum height H in Lemma 12 conditioned on $\{H < t\}$. This conditioned distribution is the distribution (4), as required for Proposition 1.

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

Recall that $N_n(T_n^{\text{mrca}})$ stands for 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 up-crossings (which equals the number of down-crossings) in the contour process at height s . If the time since the origin of \mathcal{T}_n is $T_n^{\text{or}} = t$ then the contour process has n up- and down-crossings at height t . If the time of the most recent common ancestor is $T_n^{\text{mrca}} = s$ then the maximal depth of the subexcursions below height t , measured away from t , is s ; see Figure 5. In other words, 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 $t - s$.

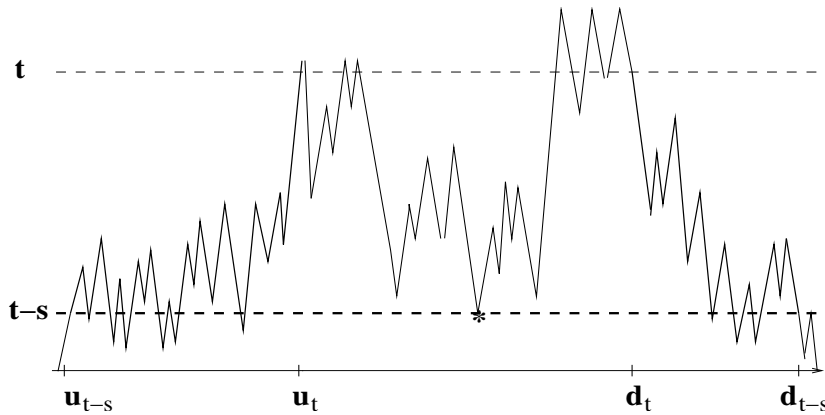


Figure 5. Parts of the contour process between $[u_{t-s}, u_t]$ and $[d_t, d_{t-s}]$ describe the number of species alive at the time of the most recent common ancestor.

In the contour process mark by u_s the horizontal coordinate of the first upcrossing of a height s and d_s the coordinate of the last downcrossing of

this height. There are no up- or down-crossings of $t - s$ before the first upcrossing of $t - s$ and after the last downcrossing of $t - s$. And if $t - s$ is the height of the T_n^{mrca} , as in Figure 5, then there are no up- or downcrossings of $t - s$ between the first upcrossing of t and the last downcrossing of t . So, $N_n(T_n^{\text{mrca}})$ 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 that is conditioned to have n upcrossings and downcrossings at height T_n^{or} , we can now calculate

Lemma 13. *Conditional on $(T_n^{\text{or}}, T_n^{\text{mrca}}) = (t, s)$, $N_n(T_n^{\text{mrca}})$ 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 part of the process between the first upcrossings u_{t-s} of $t - s$ and u_t of t when considered from height $t - s$ upwards: $\mathcal{C}(u) - (t - s)$, $u_{t-s} \leq u \leq u_t$, is distributed as an ERW process conditioned to reach height s before it reaches a 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: $\mathcal{C}(u) - (t - s)$, $d_{t-s} \geq u \geq d_t$, is also distributed as a 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 an ERW process reaches s before it reaches $-(t - s)$, by the law of maximum height H in Lemma 12, is

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

The probability an ERW process makes k upcrossings of 0 until it first hits s , provided its height stays below s and its depth above $-(t - s)$, is for $k = 1, 2, \dots$

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

So the number of upcrossings of $t - s$ $\mathcal{C}(u)$ makes during $[u_{t-s}, u_t]$ has a Geometric $\left(1 - \frac{t-s}{1+t-s} \frac{s}{1+s}\right)$ distribution. \square

Since by Corollary 5 $(n^{-1}T_n^{\text{or}}, n^{-1}T_n^{\text{mrca}}) \xrightarrow{d} (T^{\text{or}}, T^{\text{mrca}})$, as $n \rightarrow \infty$

$$np_n = 1 - \frac{T_n^{\text{or}} - T_n^{\text{mrca}}}{1 + T_n^{\text{or}} - T_n^{\text{mrca}}} \frac{T_n^{\text{mrca}}}{1 + T_n^{\text{mrca}}} \xrightarrow{d} \frac{1}{T^{\text{or}} - T^{\text{mrca}}} + \frac{1}{T^{\text{mrca}}}$$

and the above two Geometric(p_n) variables, when rescaled by n^{-1} , converge to independent Exponential($\lambda(T^{\text{or}}, T^{\text{mrca}})$) variables, where $\lambda(t, s) = (t - s)^{-1} + s^{-1}$. Consequently, the conditional law of $n^{-1}N_n(T_n^{\text{mrca}})$ given $(T_n^{\text{or}}, T_n^{\text{mrca}})$ converges to a Gamma variable with shape parameter 2 and scale parameter $\lambda(T^{\text{or}}, T^{\text{mrca}})$. Combining this with the result of Corollary 5 we have established assertion (7) below.

Corollary 14. *The joint limit behavior of the triple $T_n^{\text{or}}, T_n^{\text{mrca}}, N_n(T_n^{\text{mrca}})$ is given by*

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

where the limit has the joint density

$$\begin{aligned} f_{T^{\text{or}}, T^{\text{mrca}}, N^{\text{mrca}}}(t, s, r) &= t^{-2} s^{-2} \lambda(t, s)^2 r e^{-\frac{1}{s} - \lambda(t, s)r} \\ &= (t - s)^{-2} s^{-4} r e^{-\frac{1}{s} - \frac{tr}{s(t-s)}}, \quad 0 < s < t, 0 < r. \end{aligned} \quad (7)$$

The marginal density of N^{mrca} is

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

The marginal density formula follows from (7) via a calculus exercise. Note that while the distribution of N^{mrca} has mean 1 it has infinite variance.

Remark. The contour process of $\mathcal{T}_{t_n, n}$ (illustrated in Figure 5), in the limit $t_n/n \rightarrow t \in (0, \infty)$, converges after rescaling to a Brownian excursion, conditioned on total local time at height t being equal to 1. Results like Corollary 14 may be reinterpreted as providing exact formulas for quantities defined in terms of such conditioned Brownian excursions.

5.4 Extinct species

Textbooks (e.g. [18] page 24) often say

the probability that a given fossil is actually part of an ancestral lineage [of some extant species] is actually rather remote.

Various calculations relevant to this issue can be done within our model.

Consider some species v that originated at time h before the present. In the limit as $n \rightarrow \infty$, the distribution of the clade of species descending from v is given by the local limit structure of the complete tree \mathcal{T}_∞ . As stated in Section 5.5, in the limit, the descendants of a species v evolve, as time runs forwards, as in an ordinary critical branching process \mathcal{T} . Then, the chance that some descendant of v (or v itself) is extant at present equals the chance of the survival of its descendant tree \mathcal{T} for time h or longer. By Lemma 12 this is precisely $(1+h)^{-1}$, so we have

Corollary 15. *For any species alive at time h before the present, the chance that some of its descendant species (or the species itself) is extant is, in the limit $n \rightarrow \infty$,*

$$1/(1+h).$$

Now consider the total number N_n^{anc} of species that are ancestral to the n extant ones. (Precisely, we exclude the extant species, and go back to the time of origin of the clade). Intuitively, because (Lemma 8) the number of species at time h is $N_n(h) \approx n$ for $h = o(n)$, and because (Corollary 5) the time of origin T_n^{or} is of order $O(n)$, we expect from Corollary 15 that

$$\mathbb{E}[N_n^{\text{anc}}] \approx \int_0^{O(n)} \frac{n}{1+h} dh \approx n \log n.$$

We shall prove a precise result as Corollary 17, based on the following lemma.

Lemma 16. *Conditional on $T_n^{\text{or}} = t$, the total number of ancestral individuals N_n^{anc} in \mathcal{T}_n satisfies*

$$N_n^{\text{anc}} \stackrel{d}{=} \sum_{i=1}^n X_i$$

where $X_i, 1 \leq i \leq n$ are independent, X_1 has Poisson(t) distribution and X_2, \dots, X_n have the law, with $f_t(\cdot)$ as in (4),

$$\mathbb{P}(X_i = k) = \int_0^t \frac{e^{-s} s^k}{k!} f_t(s) ds, \quad k \geq 0.$$

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 i th extant individual, without including any of the ones previously counted in $X_j, j < i$.

Suppose $T^{\text{or}} = t$, 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 sub-excursions below t , which we label $e_i, 1 \leq i \leq n - 1$, and the part of $\mathcal{C}(\cdot)$ before the first up-crossing and after the last down-crossing of t ; we label the former part as $e_{0,R}$. See Figure 6.

Let $h_i, 1 \leq i \leq n - 1$ be the depths of the sub-excursions e_i , so that $t - h_i$ are the heights of the lowest points of e_i . These match the times of divergence of lineages of extant individuals. Their law was given by (4) of Proposition 1. Now, partition the excursions e_i at their lowest points and let $e_{i,R}, 1 \leq i \leq n - 1$ denote the parts on the right. Then Figure 6 shows that the ancestors of the 1st extant individual correspond in the piece of the contour process $e_{0,R}$ to the levels of constancy of the process $\varsigma_{0,R}(u) = \inf_{v \geq u}(e_{0,R}(v))$. These levels of constancy of $\varsigma_{0,R}$ match the times of lineage divergence of the ancestors of individual 1. Similarly for the i th, $2 \leq i \leq n$, extant individual Figure 6 shows that the additional ancestors of individual i (excluding those appearing as ancestors of extant individuals $j < i$) correspond in the piece of the contour process $e_{i-1,R}$ to the levels of constancy of the process $\varsigma_{i-1,R}(u) = \inf_{v \geq u}(e_{i-1,R}(v))$.

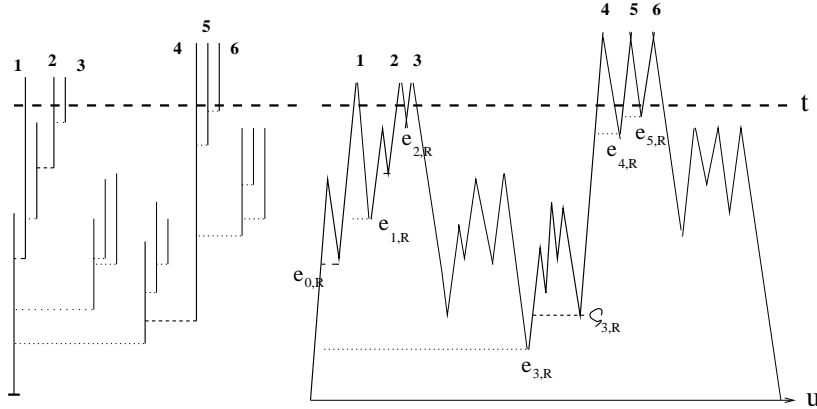


Figure 6. Ancestral lineages of the extant individuals (labeled $\{1, 2, 3, 4, 5, 6\}$) are matched in the contour process by the levels of constancy of the processes $\varsigma_{i-1,R}$, for $1 \leq i \leq n$.

So the number of ancestors X_i of the i th extant individual is the number of levels of constancy of the process $\varsigma_{i-1,R}(\cdot)$. It is clear that the piece $e_{0,R}$ is distributed as an ERW process conditioned to hit t before 0, and stopped the first time it hits t . It is less obvious but none the less true (see Lemma 6 of [20]) that, given h_i , the piece $e_{i,R}$ is also distributed as an ERW process conditioned to reach h_i before 0, and stopped the first time it hits h_i .

For such a conditioned ERW, the levels of constancy of its future infimum process form a Poisson process, restricted to, respectively, the set $[0, t]$ for $e_{0,R}$, and $[0, h_i]$ for $e_{i-1,R}$, $2 \leq i \leq n$. This can easily be seen for levels of constancy of the past supremum process of a conditioned ERW, (Lemma 6 of [20]), then time reversibility of an ERW excursion implies the rest. So the number of ancestors of the 1st extant individual is $\text{Poisson}(t)$, and the number of additional ancestors of the extant individuals i , $2 \leq i \leq n$, is $\text{Poisson}(h_i)$. Combining this with the distributions of the depths h_i , given by (4) in Proposition 1, we have proved the claim. \square

For the limit of the number of ancestors N_n^{anc} we have

Corollary 17. *As $n \rightarrow \infty$ $\frac{N_n^{\text{anc}}}{n \log n} \xrightarrow{p} 1$.*

Proof. Fix (t_n) such that $t_n/n \rightarrow t \in (0, \infty)$. Because (Corollary 5) $n^{-1}T_n^{\text{or}}$ has a distributional limit on $(0, \infty)$, it suffices to prove the following: conditional on $\{T_n^{\text{or}} = t_n\}$ we have $\frac{N_n^{\text{anc}}}{n \log n} \xrightarrow{p} 1$.

We shall prove this using the representation $N_n^{\text{anc}} = \sum_{i=1}^n X_i$ from Lemma 16, where in the following argument we are always conditioning on $\{T_n^{\text{or}} = t_n\}$. Note that the contribution to the sum from X_1 is negligible (because X_1 has $\text{Poisson}(t_n)$ distribution), so we may assume X_1 has the same distribution as the X_i , $2 \leq i \leq n$. We now calculate

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

and a similar calculation shows

$$\text{var}[X_2] = O(n).$$

Thus

$$E[N_n^{\text{anc}}] \sim n \log n; \quad \text{var}[N_n^{\text{anc}}] = O(n^2)$$

and the desired result $\frac{N_n^{\text{anc}}}{n \log n} \xrightarrow{p} 1$ follows via Chebyshev's inequality. \square

5.5 Local limit structure of the complete tree

The contour process makes it conceptually easy to see a result, complementing Corollary 6 for the lineage tree \mathcal{A}_n , concerning the local limit behavior of the complete tree \mathcal{T}_n . It turns out that the local structure relative to a given typical individual in \mathcal{T}_n , converges to the local structure relative to the root in an infinite tree that can be easily defined from a CBP tree.

There are two versions of such results, depending on whether for the typical individual we choose a random *extant* species or a random species from the entire history of the clade (which will be *extinct*, with probability $\rightarrow 1$ as $n \rightarrow \infty$). The details in obtaining the results are rather fussy, so we only outline the proofs.

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 time is in the time interval $[b(i) - \sigma, b(i) + \sigma]$ and for whom the divergence time of their lineage from that of i is in the time interval $[b(i) - \sigma, b(i) + \sigma]$. See Figure 7. Call i the *distinguished individual* in $\tilde{\mathcal{T}}_n(i, [b(i) - \sigma, b(i) + \sigma])$.

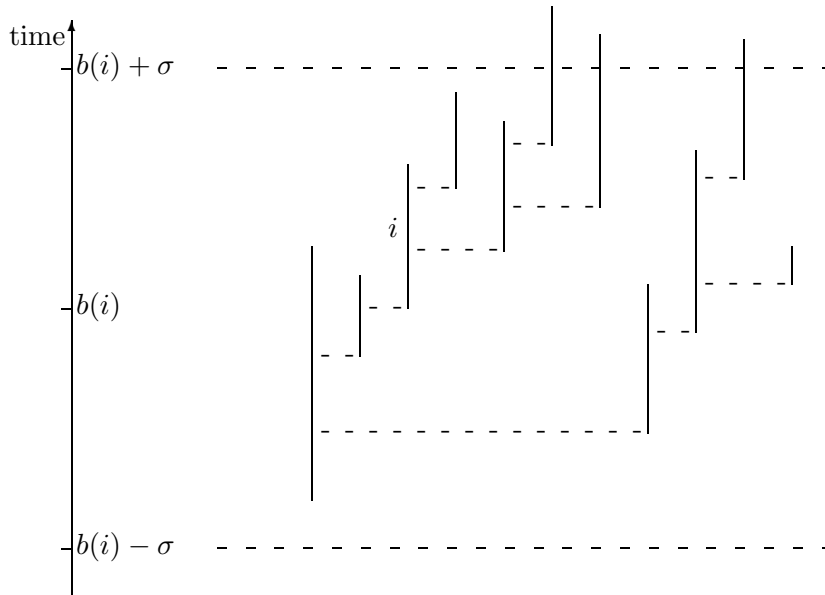


Figure 7. The local structure of the complete tree, relative to individual i .

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 it and its descendants be distributed as the CBP tree \mathcal{T} . Let the parent of this individual have Exponential(1) age at time 0 and have an independent Exponential(1) lifetime after time 0. Inductively, let the grandparent have Exponential(1) age at the birth of the parent, and have an independent Exponential(1)

lifetime after that birth time; and so on. For each of these ancestors, let them have other children during their lifetimes at the times of a Poisson (rate 1) process and let the trees of such children and their descendants be distributed as independent CBP trees. This completes the description of $\tilde{\mathcal{T}}$.

Recall the construction (Proposition 11) of 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 individuals j whose birth time is in the time interval $[-\sigma, \sigma]$ and for whom the divergence time of the lineages of j and of the distinguished individual is in the time interval $[-\sigma, \sigma]$. Note that

$$\tilde{\mathcal{T}}[-\sigma, \sigma] \text{ is determined by } (\xi_i, M^- \leq i \leq M^+) \quad (8)$$

where $(\xi_{M^-}, -\xi_{M^-+1}, \dots, -\xi_0, \xi_1, \dots, \xi_{M^+-1}, -\xi_{M^+})$ is the excursion of the two-sided ERW process above height $-\sigma$.

Here is the result for the convergence of the local structure of \mathcal{T}_n as seen relative to a random (extinct) individual, to that of the local structure of $\tilde{\mathcal{T}}$.

Proposition 18. *Let I_n denote a uniform random species from the clade \mathcal{T}_n . Then as $n \rightarrow \infty$ for fixed $\sigma > 0$,*

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

Remark. 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. We outline the proof, omitting 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 $\{(\xi_{2i+1}, \xi_{2i+2}, \dots, \xi_{2i+2m}); 0 \leq i \leq N-1\}$. So $\theta_{m,N}$ is a random probability distribution, which (using the Glivenko-Cantelli Theorem on \mathbb{R}^{2m}) converges in probability, as $N \rightarrow \infty$, to the non-random probability distribution $\mu_m = \text{dist}(\xi_1, \dots, \xi_{2m})$. By large deviation theory (see [6] §6.3) this convergence remains true conditional on events A_N for which $1/P(A_N) = O(\beta^N)$ for all $\beta > 1$.

To prove the proposition, recall (Lemma 2) that T_n^{or} is order n . So we can fix (t_n) such that $t_n/n \rightarrow t \in (0, \infty)$ and it is sufficient to prove the Proposition for $\mathcal{T}_{t_n, n}$. Fix also integers N_n such that $N_n/n^2 \rightarrow v \in (0, \infty)$. 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 over height t_n . (Then n^2 is precisely the right scaling for the number of rises and falls of an excursion with n upcrossings of a level t_n of order 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 the total number of individuals equal to N_n . Let us first prove the proposition for $\mathcal{T}_{t_n, n, N_n}^+$.

One can show that the probability $P(A_{N_n})$ decreases not faster than polynomially in $1/N_n$, so by our “large deviation” result 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 $\{2, 4, 6, \dots, 2N_n\}$,

$$(\xi_{J_n-m+1}, \dots, \xi_{J_n}, \dots, \xi_{J_n+m}) \xrightarrow{d} \mu_m \quad (9)$$

where the left side is conditioned on A_{N_n} . But this says 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 which is determined by the contour process segment 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 large and appealing to (8), we see that the proposition holds for $\mathcal{T}_{t_n, n, N_n}^+$.

To complete the proof it is enough 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 non-negligible 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 origin. Instead, fix small $0 < \delta_1 < \delta_2$ and consider the segments of the contour process \mathcal{C}^+ of $\mathcal{T}_{t_n, n, N_n}^+$ defined by:

- s_1 is the segment of \mathcal{C}^+ until its first upcrossing of $(1 - \delta_1)t_n$,
- s_2 is the segment of \mathcal{C}^+ from the subsequent downcrossing of $(1 - \delta_2)t_n$ until the next upcrossing of $(1 - \delta_1)t_n$,
- s_3 is the segment of \mathcal{C}^+ from the subsequent downcrossing of $(1 - \delta_2)t_n$ until the next upcrossing of $(1 - \delta_1)t_n$,
- ...
- s_N is the final segment of \mathcal{C}^+ after 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 all these positions, each segment is distributed as an ERW process conditioned on having the first upcrossing of a certain level after a prescribed number of steps. The number of these segments is stochastically bounded as $n \rightarrow \infty$, so the probability of the conditioning event for each segment is still only polynomially small in $1/\text{length of the segment}$. Thus separately on each segment we can show as

above that the contour process satisfies (9) for J_n uniform on that segment. Since these segments comprise (in the $n \rightarrow \infty$ limit) a proportion $1 - \varepsilon(\delta_1, \delta_2)$ of the entire contour process of $\mathcal{T}_{t_n, n, N_n}$, where $\varepsilon \rightarrow 0$ as $\delta_1, \delta_2 \rightarrow 0$, we can deduce the proposition for the stopped process $\mathcal{T}_{t_n, n, N_n}$. \square

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

Proposition 19. *Let I_n denote a uniform random extant species from the clade \mathcal{T}_n . Then as $n \rightarrow \infty$ for fixed $\sigma > 0$,*

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

One can now make exact calculations of probabilities for the distinguished individual in $\tilde{\mathcal{T}}$, which represent the $n \rightarrow \infty$ limit results for a random extant individual in \mathcal{T}_n . Here is a simple example of possible calculations within $\tilde{\mathcal{T}}$.

Corollary 20. *For the distinguished individual in $\tilde{\mathcal{T}}$:*

- (a) *the probability that its parent is extant equals 1/2;*
- (b) *the probability that some ancestor of it is extant equals $1 - e^{-1}$.*

Proof. (a) The probability that the parent of the distinguished individual is alive at time 0 is simply $P(\xi_1 < \xi_2)$, 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 $P(\xi_1 < \xi_2) = 1/2$, by symmetry.

(b) To calculate the probability that no ancestor of the distinguished individual is still alive, one only need to note that the times at which some ancestor originates form a Poisson process of rate 1, and an ancestor originating at time s before present has chance e^{-s} to be extant, so the random number of extant ancestors has Poisson distribution with mean $\int_0^\infty e^{-s} \times 1 ds = 1$, and thus takes value 0 with probability e^{-1} . \square

6 Final remarks

1. Our model of \mathcal{T}_n and \mathcal{A}_n has considerable variability between realizations. This can be seen mathematically in our distributional formulas (Corollary 14

in particular) and visually on our web site [1]. In one sense this variability is an artifact of the uniform prior on time of origin, but serves a useful purpose in emphasizing that radically different appearance of real-world trees might logically be just chance variation without biological significance.

2. Wollenberg et al [23] study via simulation a model similar to ours – critical branching conditioned on n extant species – but handle the issue of time of origin in a different way, by taking it as the deterministic time t_n which is the maximum likelihood estimator of origin time. In a sense this is unrealistic in the opposite sense to that of the previous remark, by underestimating variability. Our model extends more naturally to higher-order taxa.

3. Our model is qualitatively similar (in the sense of orders of magnitude) to the Moran model, for quantities which can be studied in the latter model. In fact the results involving local weak limits (sections 3.3 and 3.4) are exactly the same in our model as in a continuized Moran model, because our model converges (in the $n \rightarrow \infty$ limit) to the continuized Moran model over time intervals (backwards from present) of length $t = o(n)$.

4. Neutral models like ours are unrealistic for large clades, by the following reasoning. For an n -species clade, our model gives (Corollary 5) the time of origin of a clade as order n time units ago. The time unit is mean species lifetime, typically estimated as a few million years. Thus our model predicts the origin of a n -species clade as being at least n million years ago, which is known to be an overestimate for most clades of size $n \geq 100$.

5. The local point process limit in Corollary 6 is a simple instance of a general notion of *local weak convergence* of graphical structures associated with point processes on \mathbf{R}^d or abstract spaces. See [5, 4] for more sophisticated examples. In particular, Proposition 18 fits the general setting of *asymptotic fringe distributions* which exist for many different models of random trees [2].

6. Mathematicians traditionally tend to regard pictures as mere visual aids to illustrate a logical argument. But the *graphical representations* we use in Figures 1 and 4 really comprise the essence of the mathematical argument, by relating our model of random trees to well-understood models of point processes or random walks.

Acknowledgement We thank Maxim Krikun for helpful comments.

References

- [1] D. J. Aldous. Stochastic models for phylogenetic trees. Web site www.stat.berkeley.edu/users/aldous/Phylo/Pindex.html.
- [2] D.J. Aldous. Asymptotic fringe distributions for general families of random trees. *Ann. Appl. Probab.*, 1:228–266, 1991.
- [3] D.J. Aldous. The continuum random tree III. *Ann. Probab.*, 21:248–289, 1993.
- [4] D.J. Aldous and R. Lyons. Processes on unimodular random networks. Unpublished, 2004.
- [5] D.J. Aldous and J.M. Steele. The objective method: Probabilistic combinatorial optimization and local weak convergence. In H. Kesten, editor, *Probability on Discrete Structures*, volume 110 of *Encyclopaedia of Mathematical Sciences*, pages 1–72. Springer-Verlag, 2003.
- [6] A. Dembo and O. Zeitouni. *Large Deviations and Applications*. Jones and Bartlett, Boston MA, second edition, 1992.
- [7] R. Durrett. *Probability: theory and examples*. Duxbury Press, Belmont, CA, second edition, 1996.
- [8] W.J. Ewens. *Mathematical Population Genetics*. Springer, 1979.
- [9] W. Feller. *An Introduction to Probability and its Applications, Vol I, 3rd edn*. Wiley, New York, 1968.
- [10] J. Felsenstein. *Inferring Phylogenies*. Sinauer, 2003.
- [11] J. Hey. Using phylogenetic trees to study speciation and extinction. *Evolution*, 46:627–640, 1992.
- [12] J.F. Le Gall. Marches Aléatoires, mouvement Brownien et processus de branchement. In *Séminaire de Probabilités XXIII*, volume 1372 of *Lecture Notes in Math.*, pages 258–274. Springer, 1989.
- [13] M. Möhle. Ancestral processes in population genetics. *J. Theor. Biol.*, 204:629–638, 2000.
- [14] A.Ø. Mooers and S. B. Heard. Inferring evolutionary process from phylogenetic tree shape. *Quarterly Rev. Biology*, 72:31–54, 1997.

- [15] S. Nee, R.M. May, and P.H. Harvey. The reconstructed evolutionary process. *Philos. Trans. Roy. Soc. London Ser. B*, 344:305–311, 1994.
- [16] J. Neveu and J. Pitman. The branching process in a Brownian excursion. In *Séminaire de Probabilités XXIII*, volume 1372 of *Lecture Notes in Math.*, pages 248–257. Springer, 1989.
- [17] J. Neveu and J. Pitman. Renewal property of the extrema and tree property of a one-dimensional Brownian motion. In *Séminaire de Probabilités XXIII*, volume 1372 of *Lecture Notes in Math.*, pages 239–247. Springer, 1989.
- [18] R.D.M. Page and E.C. Holmes. *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science, 1998.
- [19] J. Pitman. Combinatorial stochastic processes. Technical Report 621, Dept. Statistics, U.C. Berkeley, 2002. Lecture notes for St. Flour course, July 2002.
- [20] L. Popovic. Asymptotic genealogy of a critical branching process. Technical Report 628, U.C. Berkeley Statistics Dept., 2002. To appear in *Ann. Appl. Probab.*
- [21] L. C. G. Rogers and D. Williams. *Diffusions, Markov Processes and Martingales, Vol. II: Itô Calculus*. Wiley, 1987.
- [22] S. Semple and M. Steel. *Phylogenetics*. Oxford Univ. Press, 2003.
- [23] K. Wollenberg, J. Arnold, and J.C. Avise. Recognizing the forest for the trees: Testing temporal patterns of cladogenesis using a null model of stochastic diversification. *Mol. Biol. Evol.*, 13:833–849, 1996.
- [24] G.U. Yule. A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis. *Philos. Trans. Roy. Soc. London Ser. B*, 213:21–87, 1924.