Asymptotic Genealogy of a Branching Process and a Model of Macroevolution

by

Lea Popovic

Hon.B.Sc. (University of Toronto) 1997

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Statistics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor David J. Aldous, Chair
Professor Steven N. Evans
Professor Vaughan R.F. Jones

Fall 2003

The dissertation of Lea Popovic is approved:

Chair Date

Date

Date

University of California, Berkeley

Fall 2003

Asymptotic Genealogy of a Branching Process and a Model of Macroevolution

Abstract

Asymptotic Genealogy of a Branching Process and a Model of Macroevolution

by

Lea Popovic
Doctor of Philosophy in Statistics

University of California, Berkeley

Professor David J. Aldous, Chair

We consider a stochastic model for an evolutionary process that allows us to model phylogenies and fossil time series in a coherent manner. The model includes both data on the number of extant taxa, and the likelihood of the appearance of extinct taxa in the historical record. We study this problem within a branching process model. Consider a continuous-time binary branching process conditioned to have population size n at some time t, and with a chance p for recording each extinct individual in the process. Within the family tree of this process, we consider the smallest subtree containing the genealogy of the extant individuals together with the genealogy of the recorded extinct individuals. We introduce a novel representation of such subtrees in terms of a point-process, and provide asymptotic results on the distribution of this point-process as the number of extant individuals increases. We motivate the study within the scope of a coherent analysis for an a-priori model for macroevolution.

Professor David J. Aldous
Dissertation Committee Chair

Contents

Acknowledgments

I would like to thank my advisor David J.Aldous for all his help during the work on this thesis, his suggestions and comments that pushed forward its progress. I thank Steven N.Evans and Vaughan R.F. Jones for agreeing to serve on my dissertation committee. I also would like to thank Anita Winter for her interest and comments on an earlier draft of this thesis.

Chapter 1

Introduction

1.1 Phylogeny and the Fossil Record

The use of stochastic models in the theory of macroevolution (origin and extinction of species) has been common practice for many years now. Stochastic models have been used to recreate phylogenetic trees of extant taxa from molecular data, and to recreate the time series of the past number of taxa from the fossil record. However, only few attempts have been made to make the two analyses consistent with each other. Instead of studying data-motivated models (which are scientifically more realistic for specific applications), the first purpose of this thesis is to study a purely random model that can accommodate such a coherent analysis. We study a mathematically fundamental stochastic model which allows for inclusion of both extant and fossil types of data in one analysis.

The model we propose is the continuous time critical branching process. The reasons for our choice are the following. If one is to consider a model in which extinctions and speciations are random without systematic tendencies for the number of species to increase or decrease, then for a branching process this translates into the criticality of the process (the average number of offspring of each individual is 1). Such a model corresponds to one general view in evolutionary biology that (except for mass extinctions and their aftermath) the overall number of species does not have exponential growth nor an exponential decrease.

The fundamental critical branching processes previously employed in evolutionary models have drawbacks that exclude their use in our proposed study. The basic evolution model is the Yule process [24], the elementary continuous-time pure birth process. This process starts with one individual, each individual gives birth to offspring according to a Poisson(rate 1) process. One can clearly not employ this model, as it a priori does not involve the extinction of species, hence does not allow for inclusion of the fossil record. The next candidate model which includes the extinction of individuals, is the basic neutral model used in population genetics. The Moran model [9], is the process of uniformly random speciations and extinctions of individuals in a population of a fixed size. In this process the total number of individuals is a fixed number, each individual lives for an Exponential(mean 1) lifetime, at the end of which it is replaced by an offspring chosen uniformly at random from the total population including itself. One can consider this process as having persisted from a distant past to the present, giving implicitly a genealogical tree of the extant individuals. Asymptotically in the total population size (with suitable rescaling) this genealogical process (backwards in time) is the Kingman's coalescent model. Although it is possible to make modifications of this model to allow for non-constant population size [14], this unfortunately requires an a priori assumption on the evolution of the total population size in time.

We are interested in considering a group of species that have some common ancestor at their origin. This corresponds to the practice in evolutionary biology of considering monophyletic groups. In this sense, the critical continuous-time binary branching process, in which individuals live for an Exponential(mean 1) time during which they produce offspring at Poisson(rate 1) times, is the natural basic model for the given purpose. Our first goal is to study the genealogical structure of the process conditioned on its population size at a given time t. By genealogical structure we mean a particular subtree of the branching process family tree. We consider all the extant individuals at time t, and the subset of the extinct individuals each having independently a chance p of being sampled into the record. The subtree we are interested in is the smallest one containing all the common ancestors of the extant individuals and all the sampled extinct individuals. We introduce a point-process representation of this subtree, which has a convenient graphical interpretation, and derive its law.

The main result of the first two chapters is the asymptotic behavior of such point-processes (as the number of extant individuals increases, appropriately rescaled), and their connection to a conditioned Brownian excursion. We further want to avoid assuming that the time of origin of the branching process is known (giving time t of today), and to rely only on the number of extant species as known. We hence incorporate our results for given t in a Bayesian model which randomizes the time of origin based on the number of extant individuals. We also consider some statistics of interest describing this genealogical structure, such as: the time of the last common ancestor of all extant individuals, the number of individuals present at the time of the last common ancestor, etc. We derive their distributions in the asymptotic setting.

As a last remark on the choice of the branching process, we note that, as implied by general convergence results on critical branching processes ([3] and many others), the same asymptotic process representing the genealogical and fossil structure as obtained here, should hold in general for any critical branching process with finite offspring variance.

The relationship between random trees and Brownian excursions has been much explored in the literature. We note only a small selection that is directly relevant to the work in this paper. Neveu-Pitman [19],[18] and Le Gall [15] noted the appearance of continuous-time critical branching processes embedded in the structure of a Brownian excursion. Abraham [1] and Le Gall [16] considered the construction of an infinite tree within a Brownian excursion, which is in some sense a limit of the trees from the work of Neveu-Pitman. The convergence of critical branching processes conditioned on total population size to a canonical tree within a Brownian excursion (the *continuum random tree*) was introduced by Aldous [3]. We state a connection of the asymptotic results in this paper with the above mentioned results.

In the mathematical literature, some aspects of the genealogy of critical Galton-Watson trees conditioned on non-extinction have been studied by Durrett [8], without the use of random trees. The genealogy of branching processes which are "barely" super-critical was studied by O'Connell [20] for which he explored questions of last common ancestry of all extant individuals. Geiger [12] introduced a different point-process representation of the genealogy of a critical branching process. He considered branching processes that are size-biased according to the number of individuals extant at some time t, and represented the genealogy of the extant individuals relative to their degree of relationship with a distinguished individual chosen randomly from this extant set. The genealogy of critical branching processes has also been studied within the context of super-processes (for an excellent survey see [16]). In particular, Le Gall, Le Jan, and Duquesne [17],[7] have considered Galton-Watson branching processes with offspring distributions μ_n conditioned on total progeny, that when suitably rescaled converge to a continuous-state branching process with some branching mechanism ψ. They have shown that the genealogies of these conditioned branching processes then also converge to a continuous branching structure coded by a ψ-height process, constructed as a local time functional of the Levy process with Laplace exponent ψ).

1.2 Higher Order Taxa

The use of stochastic models of evolution has also been extensively applied on each level of taxonomy (species, genera, etc.) separately. However, it is certainly desirable to insure hierarchical consistency between them, so that the phylogenetic tree on species is consistent with the phylogenetic tree on genera consisting of these species. The second purpose of this thesis is to extend our model on species to encompass a consistent model on higher order taxa. A natural way to extend our analysis to the next taxonomic level is to superimpose on the branching process a random process of marks distinguishing some species as sufficiently different as to be originators of a new genus. The way one defines what constitutes a new genus from these marks is subject to different constraints that produce different degrees of coarseness in the next taxonomic level.

We shall consider the coarsest definition that makes each genus a clade. In biological terms this translates into a monophyletic property of higher order taxa. We provide an analysis for tree on genera thus generated from the tree on species. We derive the distribution of the number of lineages at a time s in the past, the merge-rate of lineages, as well as the number of species per genus. We also derive the relevant statistics for the shape of the tree on genera. As implied by Aldous' discussion of several shape statistics from the biological literature ([4]), we consider here the distribution of split-rates as a mathematically optimal way of describing the tree shape.

A further analysis of models on higher order taxa, as well as a list of useful references to the biological literature, can be found in [5], a survey paper of Aldous in collaboration with the author of this thesis, which aims to provide a comprehensive discussion of coherent and consistent stochastic models for macroevolution.

1.3 Overview

The analysis of a the genealogical and fossil structure of the model on species are presented in Chapters 2- 3. In Section 2.2 we give a precise definition of the genealogical point-process representing the common ancestry of the extant individuals. We provide its exact law, as well as its asymptotic behavior in Section 2.3. Then, in Section 3.1 we give the definition of the corresponding genealogical point-process that includes the sampled extinct individuals as well. We provide its exact law, and in Sectione 3.2 derive its asymptotic behavior as well.

The Bayesian calculations randomizing the time of origin of the process, are given in 2.4. Furtermore, the distribution of statistics describing the geneaological structure are given in 2.5.

The analysis of the model on higher order taxa is presented in Chapter 4. We give the definition of a higher order taxon, say a genus, based on a process of changes on the species, and analyze the superimposed process on genera. We provide the distribution for the lifetime of a genus containing a typical extant species in Section 4.3, as well as the distribution of the number of other extant species contained in it in Section 4.4. Lastly, in Sections4.5-4.6 we analyze the shape of the tree on species as well as the shape of the tree on genera, via probabilities of different types of branching points in them.

Chapter 2

Genealogy of the Extant Individuals

In this chapter we define precisely the branching process model for the evolution of species and recall equivalent ways of representing the genealogical history of this process as a random tree with edge-lengths and by the contour process of this tree. We introduce a novel way of representing the genealogy of extant individuals with a point-process (named the *genealogical point-process*), and derive the law of this process (Lemma 3). We further introduce a point-process defined from a conditioned Brownian excursion (named the *continuum genealogical point-process*), and show that this is precisely the asymptotic process of the rescaled genealogical point-processes as the number of extant individuals increases (Theorem 5). We also give the associated Bayesian asymptotic result with the time of the of the origin of the process randomized (Corollary 7). In the last section we derive the asymptotic distributions of some statistics of interest which can be used to describe the genealogy.

2.1 Critical Branching Process

Let $\mathcal{T}$ be a continuous-time critical branching process, with initial population size 1. In such a process each individual has an Exponential(rate 1) lifetime, in the course of which it gives birth to new individuals at Poisson(rate 1) times, with all the individuals living and reproducing independently of each other. Let $\mathcal{T}_{t,n}$ be the process $\mathcal{T}$ conditioned to have population size n at a given time t. We shall use the same notation ($\mathcal{T}$ and $\mathcal{T}_{t,n}$) for the random trees with edge-lengths that are the family trees of these processes.

We depict these family trees as rooted planar trees with the following conventions. Each individual is represented with a set of edges whose total length is equal to that individual's lifetime. Each birth time of an offspring corresponds to a branch-point in the parent's edge, with the total length of the parent's edge until the branch-point equal to the parent's age at this time. The new individual is then represented by the edge on the right, while the parent continues in the edge on the left. Such trees are identified by their shape and by the collection of the birth times and lifetimes of individuals. We shall label the vertices in the tree in a depth-first search manner. An example of a random tree realization of $\mathcal{T}_{t,n}$ is shown in Figure 2.1(a).

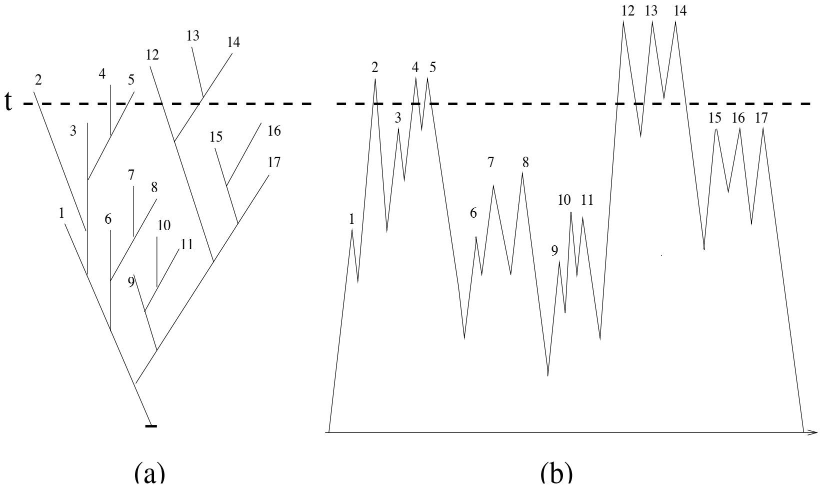

Figure 2.1: (a) A realization of the tree $\mathcal{T}_{t,n}$ whose population at time t is $n = 5$; the leaves are labeled in depth-first search manner; (b) The contour $\mathcal{C}_{\mathcal{T}_{t,n}}$ process of the tree $\mathcal{T}_{t,n}$; each local maximum of $\mathcal{C}_{\mathcal{T}_{t,n}}$ corresponds to the height of a leaf of $\mathcal{T}_{t,n}$.

Remark. The random tree $\mathcal{T}$ we defined is almost the same as the family tree of a continuous-time critical binary-branching Galton-Watson process. The difference between the two is only in the identities of the individuals. If, in the Galton-Watson process, at each branching event with two offspring we were to impose the identification of the left offspring with its parent, the resulting random tree would be the same as the family tree of our branching process $\mathcal{T}$.

Let $\mathcal{C}_{\mathcal{T}}$ be the contour process induced by the random tree $\mathcal{T}$. The contour process of a rooted planar tree is a continuous function giving the distance from the root of a unit-speed depth-first search of the tree. Such a process starts at the root of the tree, traverses each edge of the tree once upwards and once downwards following the depth-first search order of the vertices, and ends back at the root of the tree. The contour process consists of line segments of slope $+1$ (the rises), and line segments of slope -1 (the falls). The unit speed of the traversal insures that the height levels in the process are equivalent to distances from the root in the tree, in other words to the times in the branching process. In particular, the local maxima of $\mathcal{C}_{\mathcal{T}}$ correspond in height to the leaves of $\mathcal{T}$ (in other words, to the times of death of individuals), while the local minima of $\mathcal{C}_{\mathcal{T}}$ correspond in height to the branching points of $\mathcal{T}$ (to times of births of new individuals). The contour process induced by the random tree $\mathcal{T}_{t,n}$ depicted in Figure 2.1(a) is $\mathcal{C}_{\mathcal{T}_{t,n}}$ shown in Figure 2.1(b). For a formal definition of planar trees with edge lengths, contour processes and their many useful properties one can consult the recent lecture notes of Pitman [21] §6.1.

2.2 Genealogical Point-process

Let the *genealogy* of extant individuals at a given time t be defined as the smallest subtree of the family tree which contains all the edges representing the ancestry of the extant individuals. The genealogy of extant individuals at t in $\mathcal{T}_{t,n}$ is thus an n-leaf tree, which we denote by $\mathcal{G}(\mathcal{T}_{t,n})$. Figure 2.2(a) shows the genealogical subtree of the tree from Figure 2.1(a). We next introduce a novel point-process representation of this genealogical tree $\mathcal{G}(\mathcal{T}_{t,n})$. We thus get an object that is much simpler to analyze, and gives much clearer asymptotic results than if made in the original space of trees with edge-lengths.

Informally, think of forming this point-process by taking the heights of the branching points of the genealogical tree $\mathcal{G}(\mathcal{T}_{t,n})$ in the order they have as vertices in the tree. For convenience reasons (in considering asymptotics with t increasing) we keep track of the heights of the branching points in terms of their distances from level t. The vertical coordinate of each branching point is thus its distance below level t, while its horizontal coordinate is just its index. The point-process representation of $\mathcal{G}(\mathcal{T}_{t,n})$ from Figure2.2(a) is shown in Figure 2.2(b). Formally, let $A_i, 1 \leq i \leq n-1$, be the times (distance to the root) of branch-points in the tree $\mathcal{G}(\mathcal{T}_{t,n})$, indexed in order induced from the depth-first search of the vertices in $\mathcal{T}_{t,n}$, let $\tau_i = t - A_i$ be their distance below level t, and let $\ell_i = i$.

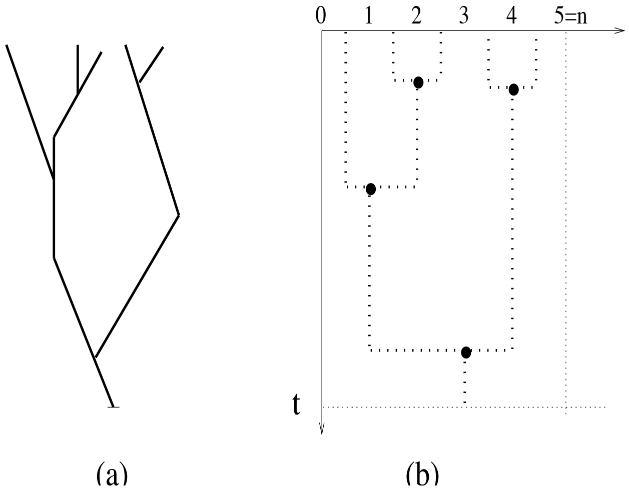

Figure 2.2: (a) The genealogical tree $\mathcal{G}(\mathcal{T}_{t,n})$ of the extant individuals at time t; (b) The point-process $\mathit{\Pi}_{t,n}$ representation of $\mathcal{G}(\mathcal{T}_{t,n})$ (the dotted lines show the simple reconstruction of $\mathcal{G}(\mathcal{T}_{t,n})$ from its point-process).

Definition. The *genealogical point-process* $\mathit{\Pi}_{t,n}$ is the random finite set

$$\mathit{\Pi}_{t,n} = \{(\ell_i, \tau_i) : 1 \leq i \leq n-1,\, 0 < \tau_i < t\} \tag{2.1}$$

For practical purposes it is most useful to exploit the bijection between a random tree and its contour process. We can obtain the point-process $\mathit{\Pi}_{t,n}$ equivalently from the contour process $\mathcal{C}_{\mathcal{T}_{t,n}}$ as follows. The ith individual extant at t corresponds to the pair (U_i, D_i), consisting of an up-crossing U_i and the subsequent down-crossing D_i of the level t. The branch-points $A_i, 1 \leq i \leq n-1$, of $\mathcal{G}(\mathcal{T}_{t,n})$ correspond to the levels of lowest local minima of the excursions of $\mathcal{C}_{\mathcal{T}_{t,n}}$ below level t, in other words $A_i = \inf\{\mathcal{C}_{\mathcal{T}_{t,n}}(u) : D_i < u < U_{i+1}\}$.

We next use this observation together with the description of the law of $\mathcal{C}_{\mathcal{T}_{t,n}}$ in order to obtain the law of $\mathit{\Pi}_{t,n}$. We first recall the result of Neveu-Pitman-Le Gall, regarding the law of the contour process $\mathcal{C}_{\mathcal{T}}$ of an unconditioned random tree $\mathcal{T}$ (one can consult either [15] or [19] for its proof).

Lemma 1. *In the contour process $\mathcal{C}_{\mathcal{T}}$ of a critical branching process $\mathcal{T}$ the sequence of rises and falls (up to the last fall) has the same distribution as a sequence of independent Exponential(rate* 1*) variables stopped one step before the sum of successive rises and falls becomes negative (the last fall is then set to equal this sum).*

The following corollary is an immediate consequence of Lemma 1 and the memoryless property of the exponential distribution.

Corollary 2. *For the contour process $\mathcal{C}_{\mathcal{T}}$ the process $X_{\mathcal{T}} = (\mathcal{C}_{\mathcal{T}}, slope[\mathcal{C}_{\mathcal{T}}])$ is a time-homogeneous strong Markov process on $\mathbb{R}^+ \times \{+1, -1\}$ stopped when it first reaches $(0, -1)$.*

The law of the genealogical point-process $\Pi_{t,n}$ can now easily be derived using some standard excursion theory of Markov processes. Note that the contour process of a whole class of binary branching processes can be shown to be a time-homogeneous Markov process as well (see [11]). In the following Lemma we show that the distances of the $n-1$ branching points below level t are independent and identically distributed, with the same law as that of the height of a random tree $\mathcal{T}$ conditioned on its height being less than t.

Lemma 3. *For any fixed $t > 0$, the random set $\Pi_{t,n}$ is a simple point-process on $\{1, \ldots, n-1\} \times (0, t)$ with intensity measure*

$$\nu_{t,n}\big(\{i\} \times d\tau\big) = \frac{1}{2} \frac{d\tau}{(1+\tau)^2} \frac{1+t}{t} \tag{2.2}$$

In other words, $\tau_i, 1 \leq i \leq n-1$ are i.i.d. variables on $(0, t)$ with the law (2.2).

Proof. In short the proof relies on the following. The contour process $\mathcal{C}_{\mathcal{T}}$ of an unconditioned tree $\mathcal{T}$ is, by the previous Corollary, a Markov process considered until a certain stopping time. Hence, its excursions below some level t are independent and identically distributed. Conditioning of the tree $\mathcal{T}_{t,n}$ translates simply in terms of its contour process, into conditioning this Markov process to have exactly $n-1$ excursions below t until this stopping time. Further, for the law of these excursions it will follow, by the sign invariance of the law of $\mathcal{C}_{\mathcal{T}}$, that their law is the same as that of a copy of $\mathcal{C}_{\mathcal{T}}$ conditioned to have a height less than t.

Consider the Markov process $X_{\mathcal{T}} = (\mathcal{C}_{\mathcal{T}}, \text{slope}[\mathcal{C}_{\mathcal{T}}])$ until the first hitting time $U_{(0,-1)} = \inf\{u \geq 0 : X_{\mathcal{T}}(u) = (0, -1)\}$, and consider its excursions from the point $(t, +1)$ using the distribution of $\mathcal{C}_{\mathcal{T}}$ given by Lemma 1. For $i \geq 1$ let U_i be the times of the up-crossings of level t by $\mathcal{C}_{\mathcal{T}}$

$$U_0 = 0,\ U_i = \inf\{u > U_{i-1} : X_{\mathcal{T}}(u) = (t, +1)\}, i \geq 1.$$

Clearly $\mathbf{P}_{(t,+1)}\big[\inf\{u > 0 : X_{\mathcal{T}}(u) = (t, +1)\} > 0\big] = 1$, hence the set of all visits to $(t, +1)$ at times $\{U_i, i \geq 1\}$ is discrete. The excursions of $X_{\mathcal{T}}$ from level t are, for $i \geq 1$

$$e_i(u) = X_{\mathcal{T}}(U_i + u),\ \text{for } u \in [0, U_{i+1} - U_i),\ \text{and } e_i(u) = (0, +1) \text{ else.}$$

The number of visits in an interval $[0, u]$ is

$$\ell(0) = 0,\ \ell(u) = \sup\{i > 0 : u > U_i\}, u > 0,$$

and the total number prior to $U_{(0,-1)}$ is $L = \sup\{i \geq 0 : U_{(0,-1)} > U_i\} = \ell(U_{(0,-1)})$. If $\mathbf{n}$ is the $\mathbf{P}_{(t,+1)}$-law of e_i, and if $\mathbf{e}^{<t}$ is the set of excursions from $(t, +1)$ that return to $(t, +1)$ without reaching $(0, -1)$, and $\mathbf{e}^{>t}$ the set of all others, then it is clear that (e.g.[23] Vol.2 §VI.50.)

- $\mathbf{P}_{(t,+1)}\big[L \geq i\big] = \big[\mathbf{n}(\mathbf{e}^{<t})\big]^{i-1}$, $i \geq 1$, and $e_1, e_2, \ldots$ are independent
- given that $L \geq i$: the law of $e_1, e_2, .., e_{i-1}$ is $\mathbf{n}(\cdot \cap \mathbf{e}^{<t})/\mathbf{n}(\mathbf{e}^{<t})$
- given that $L = i$: the law of e_i is $\mathbf{n}(\cdot \cap \mathbf{e}^{>t})/\mathbf{n}(\mathbf{e}^{>t})$

This makes $\{(\ell(U_i), e_i), 1 \leq i \leq L - 1\}$ a simple point-process, (note that $\ell(U_i) = i$, and $\ell(\infty) = L$), whose number of points has a Geometric($\mathbf{n}(\mathbf{e}^{>t})$) law, and with each e_i having the law $\mathbf{n}(\cdot \cap \mathbf{e}^{<t})/\mathbf{n}(\mathbf{e}^{<t})$.

This observation is particularly convenient for analyzing the law of $\mathcal{C}_{\mathcal{T}_{t,n}}$. Since $\mathcal{C}_{\mathcal{T}_{t,n}}$ is just $\mathcal{C}_{\mathcal{T}}$ conditioned on $L = n$, the $n - 1$ excursions of $\mathcal{C}_{\mathcal{T}_{t,n}}$ below t are independent identically distributed with the law $\mathbf{n}(\cdot \cap \mathbf{e}^{<t})/\mathbf{n}(\mathbf{e}^{<t})$. We next derive the law of their depth A_i measured as distance from level t by $\tau_i = t - A_i$.

For each up-crossing time U_i of level t, we have a down-crossing time

$$D_i = \inf\{u > U_i : X_{\mathcal{T}}(u) = (t, -1)\}, i \geq 1.$$

For the values of $A_i, i \geq 1$ we are only interested in the part of the excursions from $(t, +1)$ below level t

$$e_i^{<t} = e_i(D_i + u), u \in [0, U_{i+1} - D_i), \text{ and } e_i^{<t}(u) = (0, +1) \text{ else.}$$

We note that the shift and reflection invariance of the transition function of $\mathcal{C}_{\mathcal{T}}$, as well as its strong Markov property, applied to the law $\mathbf{n}$ for $e_i^{<t}$ imply that the law of $e_i^{+} = t - e_i^{<t}$ is the same as the law of $X_{\mathcal{T}}$. Consequently the law of $t - \inf(e_i^{<t}) = \sup(e_i^{+})$ is the same law as that of $\sup(\mathcal{C}_{\mathcal{T}})$.

To explicitly express the law of $\sup(\mathcal{C}_{\mathcal{T}})$ we now recall classical results for the branching process $\mathcal{T}$ (e.g. [10] §XVII.10.11.), by which the law of the population size $N(t)$ of $\mathcal{T}$ at time t is given by

$$\mathbf{P}\big[N(t) = 0\big] = \frac{t}{1+t}; \;\; \mathbf{P}\big[N(t) = k\big] = \frac{t^{k-1}}{(1+t)^{k+1}}, \text{ for } k \geq 1. \tag{2.3}$$

Hence

$$\mathbf{P}\big[\sup(\mathcal{C}_{\mathcal{T}}) > t\big] = \mathbf{P}\big[N(t) > 0\big] = \frac{1}{1+t}, \text{ for } t \geq 0. \tag{2.4}$$

Now for $\mathcal{C}_{\mathcal{T}_{t,n}}$ and for each $1 \leq i \leq n-1$ we have that $A_i = \inf(e_i^{<t})$, and the $e_i^{<t}$ are independent with $e_i^{<t} \sim \mathbf{n}(\cdot \cap \mathbf{e}^{<t})/\mathbf{n}(\mathbf{e}^{<t})$, hence then each $\tau_i = t - A_i$ has the law

$$\begin{aligned} \mathbf{P}[\tau_i \in d\tau] &= \mathbf{P}[\sup(\mathcal{C}_{\mathcal{T}}) \in d\tau | \sup(\mathcal{C}_{\mathcal{T}}) < t] \\ &= \frac{d\tau}{(1+\tau)^2} \frac{1+t}{t}, \text{ for } 0 \leq \tau \leq t. \end{aligned} \tag{2.5}$$

Since for the genealogical point-process $\mathit{\Pi}_{t,n}$ we consider only the excursions below level t, we have that the rate of these points is 1/2. □

2.3 Continuum Genealogical Point-process

We could establish the asymptotics for $\mathit{\Pi}_{t,n}$ now with a routine calculation. However, instead of considering this result in isolation, it is far more natural to view it as part of the larger picture connecting critical branching processes and Brownian excursions. Let us recall the asymptotic results for critical Galton-Watson processes conditioned on a "large" total populations size. A result of Aldous [3] (Thm 23) says that its contour process (when appropriately rescaled) converges as the total population size increases, to a Brownian excursion (doubled in height) conditioned to be of length 1. Note that, if N_{tot} is the total population size of a critical Galton-Watson process, and $N(t)$ its population size at some given time t, then the events $\{N_{tot} = n\}$ and $\{N(t) = n|N(t) > 0\}$ are both events of "small" probabilities. The first has asymptotic chance $cn^{-3/2}$ as $n \to \infty$, and for $t/n \to \mathrm{t}$ as $n \to \infty$ the second has asymptotic chance $c(\mathrm{t})\, n^{-1}$ [3]. While the total population N_{tot} size corresponds to the total length of the contour process, the population size $N(t)$ at a particular time t corresponds to the occupation time of the contour process at level t. Hence, it is natural to expect that the contour process of a critical Galton-Watson process conditioned on a "large" population at time t (when appropriately rescaled) converges, when $t/n \to \mathrm{t}$ as $n \to \infty$, to a Brownian excursion conditioned to have local time 1 at given level t.

We will show the following. Consider a Brownian excursion conditioned to have local time 1 at level t, as a "contour process" of an infinite tree (in the sense of the bijection between continuous functions and trees established in [3]). Consider defining a "genealogical" point-process from this Brownian excursion, using the depths of its excursions below level t, in the same manner as used in defining $\Pi_{t,n}$ from the contour process $\mathcal{C}_{\mathcal{T}_{t,n}}$, except that the excursions are now indexed by the amount of local time at level t at their beginning. The state-space of such a point-process can be simply described, and we show that it has quite a simple law as well. It is then easy to show that this point-process is precisely the asymptotic process of appropriately rescaled processes $\Pi_{t,n}$ as $n \to \infty$.

We construct a point-process from a Brownian excursion conditioned to have local time 1 at level t, in the same manner in which $\Pi_{t,n}$ was constructed from the contour process $\mathcal{C}_{\mathcal{T}_{t,n}}$. Let $\mathcal{B}(u), u \geq 0$ be a Brownian excursion. For a fixed $\mathrm{t} > 0$, let $\ell_{\mathrm{t}}(u), u \geq 0$, be its local time at level t up to time u (with standard normalization of local time as occupation density relative to Lebesgue measure). Let $i_{\mathrm{t}}(\ell), \ell \geq 0$, be the inverse process of ℓ_{t}, in other words $i_{\mathrm{t}}(\ell) = \inf\{u > 0 : \ell_{\mathrm{t}}(u) > \ell\}$. Let $\mathcal{B}_{\mathrm{t},1}(u), u \geq 0$, then be the excursion $\mathcal{B}$ conditioned to have total local time ℓ_{t} equal to 1, where $\ell_{\mathrm{t}} = \ell_{\mathrm{t}}(\infty)$ is the total local time at t. Consider excursions $e_{\ell}^{<\mathrm{t}}$ of $\mathcal{B}_{\mathrm{t},1}$ below level t indexed by the amount of local time ℓ at the time $i_{\mathrm{t}}(\ell^{-})$ of their beginning. For each such excursion let a_{ℓ} be its infimum, and let t_{ℓ} be the depth of the excursion measured from level t, $t_{\ell} = t - a_{\ell}$. Ito's excursion theory then insures that the process $\{(\ell, t_\ell) : i_{\mathrm{t}}(\ell^-) \neq i_{\mathrm{t}}(\ell)\}$ is well defined.

Definition. The *continuum genealogical point-process* $\pi_{\mathrm{t},1}$ is the random countably infinite set

$$\pi_{\mathrm{t},1} = \{(\ell, t_\ell) : i_{\mathrm{t}}(\ell^-) \neq i_{\mathrm{t}}(\ell)\} \tag{2.6}$$

Remark. The name of the process will be justified by establishing it as the limit of genealogical point-processes.

For the state-space of the continuum genealogical process we introduce the notion of a nice point-process, (see [3]§2.8.). A *nice point-process on* $[0,1] \times (0,\infty)$ is a countably infinite set of points such that:

- for any $\delta > 0$: $[0,1] \times [\delta, \infty)$ contains only finitely many points
- for any $0 \leq x < y \leq 1, \delta > 0$: $[x,y] \times (0,\delta)$ contains at least one point.

We now show that the state-space for $\pi_{\mathrm{t},1}$ is the set of nice point-processes, and establish the law of this process using standard results of Levy-Ito-Williams on excursion theory.

Lemma 4. *The random set* $\pi_{\mathrm{t},1}$ *is a Poisson point-process on* $[0,1] \times (0,\mathrm{t})$ *with intensity measure*

$$\nu(d\ell \times d\tau) = \frac{d\ell}{2}\frac{d\tau}{\tau^2} \tag{2.7}$$

In particular, the random set $\pi_{\mathrm{t},1}$ *is a.s. a nice point-process.*

Proof. The crux of the proof lies in the following observations. An unconditioned Brownian excursion $\mathcal{B}$ observed from the first time it reaches level t, is just t−a standard Brownian motion observed until the first time it reaches t. The excursions of $\mathcal{B}$ below level t are thus the positive excursions of the Brownian motion. By a standard result, the process of excursions of Brownian motion from 0, indexed by the amount of local time at 0 at the time of their beginning, is a Poisson point-process with intensity measure $d\ell \times \mathbf{n}$, where $\mathbf{n}$ is Ito's excursion measure. One can show that the condition on $\mathcal{B}$ to have local time 1 at level t, is equivalent to the condition that the shifted Brownian motion has all its excursions until local time 1 of height lower than t and has one excursion at local time 1 higher than t. This then, by the independence properties of Poisson processes, allows for a simple description of the point-process of the depths of excursions below t of $\mathcal{B}_{\mathrm{t},1}$ as a Poisson process itself, except restricted to the set $[0,1] \times (0,\mathrm{t})$.

Consider the path of an (unconditioned) Brownian excursion $\mathcal{B}$ after the first hitting time of t, $U_{\mathrm{t}} = \inf\{u \geq 0 : \mathcal{B}(u) = \mathrm{t}\}$, shifted and reflected about the u-axis

$$\beta(u) = \mathrm{t} - \mathcal{B}(U_{\mathrm{t}} + u), \text{ for } u \geq 0 \tag{2.8}$$

Let $\ell_0^\beta(u), u \geq 0$ be the local time of β at level 0 up to time u, and let $i_0^\beta(\ell), \ell \geq 0$ be the inverse process of this local time, in other words $i_0^\beta(\ell) = \inf\{u > 0 : \ell_0^\beta(u) > \ell\}$. Then the process $\beta(u), u \geq 0$ is a standard Brownian motion stopped at the first hitting time of t, $U_{\mathrm{t}}^\beta = \inf\{u \geq 0 : \beta(u) = \mathrm{t}\}$.

Next, the excursions of β from 0 are (with a change of sign), precisely the excursions of $\mathcal{B}$ from t, and the local time process ℓ_0^β of β is equivalent to the local time process ℓ_{t} of $\mathcal{B}$. We are only interested in the excursions of $\mathcal{B}$ below t, which are the positive excursions of β, for $i_0^\beta(\ell^-) \neq i_0^\beta(\ell)$ and $\beta(i_0^\beta(\ell)^+) > 0$

$$e_\ell^+ = \beta(i_0^\beta(\ell^-) + u), u \in [0, i_0^\beta(\ell) - i_0^\beta(\ell^-)), \text{ and } e_\ell^+(u) = 0 \text{ else}$$

Note that we thus have that the infimum of an excursion of $\mathcal{B}$ below t to be simply $\inf(e_\ell^{<\mathrm{t}}) = \mathrm{t} - \sup(e_\ell^+)$.

Standard results of Ito's excursion theory (e.g. [23] Vol.2 §VI.47.) imply that for a standard Brownian motion β the random set for its positive excursions $\{(\ell, \sup(e_\ell^+)) : i_0^\beta(\ell^-) \neq i_0^\beta(\ell), \beta(i_0^\beta(\ell)^+) > 0\}$ is a Poisson point-process on $\mathbb{R}^+ \times \mathbb{R}^+$ with intensity measure $d\ell/2 \times d\tau/\tau^2$.

Now let $\mathrm{L} = \inf\{\ell \geq 0 : \sup(e_\ell^+) \geq \mathrm{t})\}$. Then stopping ℓ_0^β at the hitting time L is equivalent to stopping β at its hitting time U_{t}^β. Let π_{t} be a random set defined from the unconditioned Brownian excursion $\mathcal{B}$, in the same manner in which we defined $\pi_{\mathrm{t},1}$ from a conditioned Brownian excursion $\mathcal{B}_{\mathrm{t},1}$. Then, using the relationship (2.8) of $\mathcal{B}$ and β, we observe that π_{t} is equivalent to a restriction of $\{(\ell, \sup(e_\ell^+)) : i_0^\beta(\ell^-) \neq i_0^\beta(\ell), \beta(i_0^\beta(\ell)^+) > 0\}$ on the random set $[0, \mathrm{L}] \times (0, \mathrm{t})$. The Poisson point-process description of $\{(\ell, \sup(e_\ell^+)) :$

$i^{\beta}_{0}(\ell^{-}) \neq i^{\beta}_{0}(\ell)$, $\beta(i^{\beta}_{0}(\ell)^{+}) > 0\}$ now implies that π_{t} is a Poisson point-process on $\mathbb{R}^{+} \times \mathbb{R}^{+}$ with intensity measure $d\ell/2 \times d\tau/\tau^{2}$ restricted to the random set $[0, \mathrm{L}] \times (0, \mathrm{t})$.

Next, note that the condition $\{\ell_{\mathrm{t}} = 1\}$ for $\mathcal{B}$ is equivalent to the condition $\{\ell^{\beta}_{0}(U^{\beta}_{\mathrm{t}}) = 1\}$ for β, which is further equivalent to the condition $\{\mathrm{L} = 1\}$ for π_{t}. We have thus established that $\pi_{\mathrm{t},1} \overset{d}{=} \pi_{\mathrm{t}}|\{\mathrm{L} = 1\}$.

Further, the condition $\{\mathrm{L} = 1\}$ on π_{t} is equivalent to the condition that π_{t} has no points in $[0, 1)\times[\mathrm{t}, \infty)$ and has a point in $\{1\}\times[\mathrm{t}, \infty)$. But since π_{t} is Poisson, independence of Poisson random measures on disjoint sets implies that conditioning π_{t} on $\{\mathrm{L} = 1\}$ will not alter its law on the set $[0, 1] \times (0, \mathrm{t})$. However, since $\pi_{\mathrm{t},1}$ is supported precisely on $[0, 1] \times (0, \mathrm{t})$, the above results together imply that $\pi_{\mathrm{t},1}$ is a Poisson point-process on $[0, 1] \times (0, \mathrm{t})$ with intensity measure $d\ell/2 \times d\tau/\tau^{2}$.

It is now easy to see from the intensity measure of $\pi_{\mathrm{t},1}$ that its realizations are a.s. nice point-processes, namely

- for any $\delta > 0$: $\iint_{[0,1]\times[\delta,\infty)} d\ell/2\times d\tau/\tau^{2} = 1/2\delta < \infty$
- for any $0\leq x<y\leq 1$, and $\delta>0$: $\iint_{[x,y]\times(0,\delta)} d\ell/2\times d\tau/\tau^{2} = (y-x)/2\cdot\infty$

And since $\pi_{\mathrm{t},1}$ is Poisson, finiteness of its intensity measure on $[0, 1] \times [\delta, \infty)$ implies that it has a.s. only finitely many points in the set $[0, 1] \times [\delta, \infty)$, while infiniteness of its intensity measure on $[x, y]\times(0, \delta)$ implies that it has a.s. at least one point on the set $[x, y]\times(0, \delta)$. □

Having thus obtained the description of the continuum genealogical point-process induced by a conditioned Brownian excursion, it is now an simple task to confirm that it indeed arises as the limit of genealogical processes. The right rescaling for $\mathcal{T}_{t,n}$ is to speed up the time by n and to assign mass n^{-1} to each extant individual, which implies the appropriate rescaling of each coordinate of $\mathit{\Pi}_{t,n}$ by n^{-1}. We hence define the rescaled genealogical point-process as

$$n^{-1}\mathit{\Pi}_{t,n} = \{(n^{-1}\ell_{i}, n^{-1}\tau_{i}) : (\ell_{i}, \tau_{i}) \in \mathit{\Pi}_{t,n}\} \tag{2.9}$$

and establish its asymptotic behavior as $n \to \infty$.

Theorem 5. *For any* $\{t_n > 0\}_{n\geq 1}$ *such that* $t_n/n \underset{n\to\infty}{\to} \mathrm{t}$ *we have*

$$n^{-1}\mathit{\Pi}_{t_n,n} \underset{n\to\infty}{\overset{d}{\Longrightarrow}} \pi_{\mathrm{t},1} \tag{2.10}$$

Remark. The notation $\overset{d}{\Longrightarrow}$ is used to mean weak convergence of processes.

Proof. The proof of the Theorem is a just consequence of the fact that weak convergence of Poisson point-processes follows from the weak convergence of their intensity measures.

By Lemma 3 and the rescaling (2.9) we have that $n^{-1}\mathit{\Pi}_{t_n,n}$ is a simple point-process on $\{1/n,\ldots,1-1/n\}\times(0,t_n/n)$ with intensity measure

$$\frac{1}{n}\sum_{i=1}^{n-1}\frac{\delta_{\{i\}}(\ell)}{2}\,\frac{nd\tau}{(1+n\tau)^2}\,\frac{1+t_n}{t_n} \tag{2.11}$$

If $\{t_n\}_{n\geq 1}$ is such that $t_n/n \to \mathrm{t}$ as $n\to\infty$, then it is clear that the support set of the process $n^{-1}\mathit{\Pi}_{t_n,n}$ converges to $[0,1]\times(0,\mathrm{t})$, the support set of the process $\pi_{\mathrm{t},1}$. It is also clear that the intensity measure (2.11) converges to $d\ell/2\times d\tau/\tau^2$ which, by Lemma 4, is the intensity measure of $\pi_{\mathrm{t},1}$. For simple point-processes this is sufficient (e.g.[6] §12.3.) to insure weak convergence of the processes $n^{-1}\mathit{\Pi}_{t_n,n}$ to a Poisson point-process on $[0,1]\times(0,\mathrm{t})$ with intensity measure $d\ell/2\times d\tau/\tau^2$. By Lemma 4, we thus have that $n^{-1}\mathit{\Pi}_{t_n,n} \underset{n\to\infty}{\overset{d}{\Longrightarrow}} \pi_{\mathrm{t},1}$. □

2.4 Randomization of Time of Origin

We would now like to incorporate our results in a Bayesian model which randomizes the time of origin of the process. In the use of stochastic models of macroevolution one mostly estimates the time of origin of the process is estimated along with the genealogy of extant species. For that reason we want to avoid the assumption that time t of today is known. Instead, we assume that the prior distribution for t is Uniform on $(0,\infty)$ and make use of the posterior distribution on t given that the in $\mathcal{T}$ there are n extant individuals at time t. Let q_n denote the density of this posterior distribution of t. The following Lemma establishes the density q_n, as well as its asymptotics.

Lemma 6. *For a Uniform$((0,\infty))$ prior on t_n, given that $\mathcal{T}$ has n extant individuals at time t_n, the posterior distribution of t_n has the density q_n which satisfies*

$$nq_n(t_n) \underset{\substack{n\to\infty \\ t_n/n\to \mathrm{t}}}{\rightarrow} q(\mathrm{t}), \quad q(t) = \frac{1}{t^2}\exp(-\frac{1}{t}), t > 0 \tag{2.12}$$

with q being the density of the inverse Exponential (*rate* 1) *law.*

Proof. The proof is a straightforward calculation using the law of the population size of the branching process $\mathcal{T}$ time t after its origin. Let $N(t_n)$ be the population size of $\mathcal{T}$ at time t_n. Then recall (see (2.3)) that $\mathbf{P}\big[N(t_n) = n\big] = t_n^{n-1}/(1+t_n)^{n+1}$, for $n \geq 1$. Since the prior on t_n is Uniform on $(0,\infty)$, the posterior distribution is continuous with the density

$$q_n(t_n) = \frac{1}{c(n)}\mathbf{P}\big[N(t_n) = n\big] = \frac{1}{c(n)}\frac{t_n^{n-1}}{(1+t_n)^{n+1}}, \text{ for } t_n > 0,$$

where

$$c(n) = \int_0^\infty \mathbf{P}\big[N(s) = n\big] ds = \int_0^\infty \frac{s^{n-1}}{(1+s)^{n+1}} ds = \frac{1}{n}.$$

We are interested in the asymptotics as $n \to \infty$, and $t_n/n \to \mathrm{t}$, hence for the posterior density for the rescaled t_n we have that

$$nq_n(t_n) = \frac{n^2}{(1+t_n)^2}\Big(1 - \frac{1}{1+t_n}\Big)^{n-1} \underset{\substack{n\to\infty \\ t_n/n\to \mathrm{t}}}{\rightarrow} \frac{1}{\mathrm{t}^2}\exp(-\frac{1}{\mathrm{t}}) = q(\mathrm{t})$$

as claimed. □

Remark. Improper (σ-finite) prior distributions often lead to proper (in other words, probability) posterior distributions. With our choice of the prior this is the case with our Bayesian model.

We now incorporate our earlier results on the genealogical point-process into this Bayesian model.

Definition. $\mathit{\Pi}_n$ is the point-process specified by the posterior law of the genealogical point-process under a Uniform$((0,\infty))$ prior on t and given that $\mathcal{T}$ has n extant individuals at time t. Also, π_1 is the point-process specified by first choosing t according to the inverse Exponential(rate 1) law, then choosing a point-process according to the law of the continuum genealogical point-process $\pi_{\mathrm{t},1}$.

We define the rescaling of the process $\mathit{\Pi}_n$ in the same manner as for the process $\mathit{\Pi}_{t,n}$. The next Corollary establishes the asymptotic behavior of the rescaled process $n^{-1}\mathit{\Pi}_n$ as $n \to \infty$.

Corollary 7. *For any* $\{t_n > 0\}_{n \geq 1}$ *such that* $t_n/n \underset{n\to\infty}{\to} \mathrm{t}$ *we have*

$$n^{-1}\mathit{\Pi}_n \underset{n\to\infty}{\overset{d}{\Longrightarrow}} \pi_1 \tag{2.13}$$

Proof. The proof is a direct consequence of our earlier result on the asymptotics of the rescaled process $\mathit{\Pi}_{t,n}$, together with the Lemma above.

Namely, it is clear that the law of the point-process $\mathit{\Pi}_n$ is the same as the law obtained by first choosing t_n according to the posterior distribution q_n, and then choosing a point-process according to the law of the genealogical point-process $\mathit{\Pi}_{t_n,n}$. We express this by the notation

$$\mathcal{L}(\mathit{\Pi}_n) = q_n(t_n)\mathcal{L}(\mathit{\Pi}_{t_n,n})$$

which after rescaling of the point-processes becomes

$$\mathcal{L}(n^{-1}\mathit{\Pi}_n) = nq_n(t_n)\mathcal{L}(n^{-1}\mathit{\Pi}_{t_n,n})$$

Whenever $t_n/n \to \mathrm{t}$, by Theorem 5 we have that $\mathcal{L}(n^{-1}\mathit{\Pi}_{t_n,n}) \underset{n\to\infty}{\to} \mathcal{L}(\pi_{\mathrm{t},1})$, and by Lemma 6 we have that $nq_n(t_n) \underset{n\to\infty}{\to} q(\mathrm{t})$. Since the point-process π_1 was defined by specifying $\mathcal{L}(\pi_1) = q(t)\mathcal{L}(\pi_{\mathrm{t},1})$, it follows that

$$\mathcal{L}(n^{-1}\mathit{\Pi}_n) \underset{n\to\infty}{\to} \mathcal{L}(\pi_1)$$

as claimed. □

2.5 Statistics of Genealogy

Let us next consider how we can use our asymptotic results to make approximate conclusions about the qualitative properties and the distributions of some statistics describing the genealogy. One can use the law of the continuum genealogical point-process to approximate a realization of the n-extant species genealogical point-process. This allows us to write down simple steps for simulating (an approximate) genealogical subtree of the extant species. For describing the genealogy with a few statistics, one is generally interested in the time of the last common ancestor for the extant species, and the number of species in the process at this time. We derive the distributions of the joint law of these statistics in the asymptotic setting which can be used to approximate the true distributions.

One can construct an approximate realization of a genealogical subtree of n-extant species under our Bayesian model as follows (using Lemma 6 and Lemma 3):

- Choose t from the inverse Exponential law $q(t) = \frac{1}{t^2}\exp(-\frac{1}{t}), t > 0$, let $t_n = n\mathrm{t}$;
- Choose $n-1$ values $\tau_i, 1 \leq i \leq n-1$, independently according to the same law $f_{\tau_i}(\tau) = \frac{1}{\tau^2}\frac{t_n}{t_n+1}, 0 < \tau < t_n$;
- Construct a genealogical subtree by letting the points $(i, 0),\ 1 \leq i \leq n,$ represent the extant species, and then letting the point $(\frac{1}{2} + i, \tau_i),\ 1 \leq i \leq n-1,$ represent the branching point that is the last common ancestor of the species i and $i+1$ (see Figure 2.2(b)).

In the construction we have only used the asymptotic distribution for the time of origin of the process, however in the rest of this section we shall make more use of the nice structure of the asymptotic law of the genealogical point-process as well.

Let t_n^{lca} denote the last common ancestor of all the extant species in the n-species model. Then given the time of origin of the process, we have that in the genealogical point-process $\mathit{\Pi}_{t_n,n}$

$$t_n^{lca} = \sup\{\tau_i : 1 \leq i \leq n-1, (\ell_i, \tau_i) \in \mathit{\Pi}_{t_n,n}\}.$$

As $n \to \infty$, $t_n/n \to \mathrm{t}$ we get $t_n^{lca}/n \to \mathrm{t}^{lca}$, where $\mathrm{t}^{lca} = \sup\{t_\ell : (\ell, t_\ell) \in \pi_{\mathrm{t},1}\}$. Since $\pi_{\mathrm{t},1}$ is a Poisson point-process with intensity measure $dl \times d\tau/\tau^2$ (Lemma 4), we have for the conditional law of t^{lca} given t:

$$\mathbf{P}[\mathrm{t}^{lca} \leq s|\mathrm{t}] = \mathbf{P}\big[\{\pi_{\mathrm{t},1} \cap (0,1) \times (s, \mathrm{t})\} = \emptyset\big] = \exp(\frac{1}{\mathrm{t}} - \frac{1}{s}),\ 0 < s < \mathrm{t}.$$

Hence the joint law of $(\mathrm{t}, \mathrm{t}^{lca})$ is

$$f_{\mathrm{t},\mathrm{t}^{lca}}(t, s) = \frac{1}{t^2}\frac{1}{s^2}\exp(-\frac{1}{s}),\ 0 < s < t. \tag{2.14}$$

And also the marginal law of t^{lca} is

$$f_{\mathrm{t}^{lca}}(s) = \frac{1}{s^3}\exp(-\frac{1}{s}),\ 0 < s < \infty. \tag{2.15}$$

Let $N(t_n^{lca})$ denote the number of species in the process at the time of the last common ancestor of all the extant species in the n-species model. Then given t_n and t_n^{lca}, we have that $N(t_n^{lca})$ is the occupation time at level t_n^{lca} of $\mathcal{C}_{\mathcal{T}_{t_n,n}}$. Asymptotically, in place of $\mathcal{C}_{\mathcal{T}_{t_n,n}}$ and its occupation time $N(t_n^{lca})$, we have $\mathcal{B}_{\mathrm{t},1}$, a Brownian excursion conditioned to have local time at level t equal to 1, and its local time $\ell_{\mathrm{t}^{lca}}$ at level t^{lca} ($\ell_{\mathrm{t}^{lca}}$ is shorthand for the total local time $\ell_{\mathrm{t}^{lca}}(\infty)$). Let $U_{\mathrm{t}} = \inf\{u \geq 0 : \mathcal{B}_{\mathrm{t},1}(u) = \mathrm{t}\}$ be the first visit time of level t, and let $D_{\mathrm{t}} = \sup\{u \geq 0 : \mathcal{B}_{\mathrm{t},1}(u) = \mathrm{t}\}$ be its last visit time. Note that $\mathrm{t}^{lca} = \inf\{\mathcal{B}_{\mathrm{t},1}(u) : U_{\mathrm{t}} < u < D_{\mathrm{t}}\}$, so that we have the decomposition $\ell_{\mathrm{t}^{lca}} = \ell_{\mathrm{t}^{lca}}(U_{\mathrm{t}}) + (\ell_{\mathrm{t}^{lca}}(\infty) - \ell_{\mathrm{t}^{lca}}(D_{\mathrm{t}}))$ (see Figure 2.3).

Now, let $\beta_U(u) = \mathcal{B}_{\mathrm{t},1}(u), u \leq U_{\mathrm{t}}$, and let $\beta_D(u) = \mathrm{t} - \mathcal{B}_{\mathrm{t},1}(u), u \geq D_{\mathrm{t}}$. Since we have that $\ell_{\mathrm{t}^{lca}}$ depends only on $\mathcal{B}_{\mathrm{t},1}(u), u \in (0, U_{\mathrm{t}}) \cup (D_{\mathrm{t}}, \infty)$, while ℓ_{t} depends only on $\mathcal{B}_{\mathrm{t},1}(u), u \in (U_{\mathrm{t}}, D_{\mathrm{t}})$, the strong Markov property and the time reversibility imply that β_U and β_D are two independent Brownian motions started at 0 stopped when they first hit t, and conditioned to actually reach t before they first hit 0. Also $\ell_{\mathrm{t}^{lca}}(U_{\mathrm{t}})$ and $\ell_{\mathrm{t}^{lca}}(\infty) - \ell_{\mathrm{t}^{lca}}(D_{\mathrm{t}})$ are independent identically distributed as $\ell^{\beta}_{\mathrm{t}^{lca}}(U_{\mathrm{t}})$.

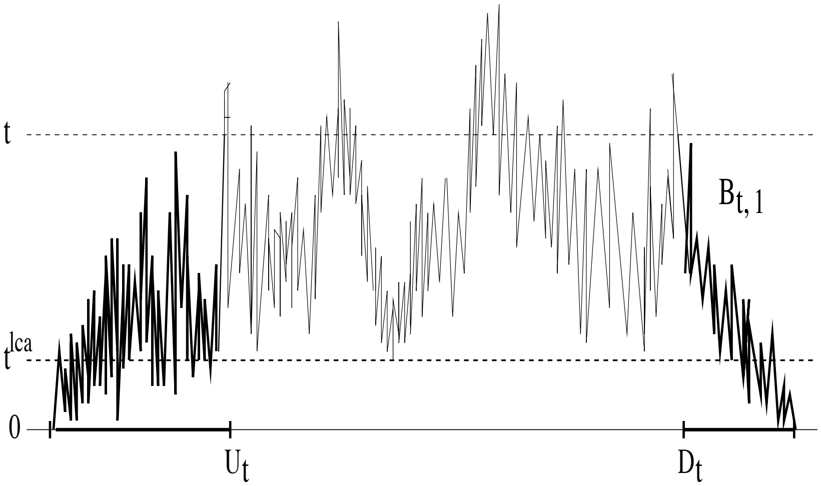
t
$\mathrm{B}_{\mathrm{t},1}$
$\mathrm{t}^{\mathrm{lca}}$
0
U_{t}
D_{t}

Figure 2.3: The conditioned Brownian excursion $\mathcal{B}_{\mathrm{t},1}$ and its local time at level $\mathrm{t}^{lca} = \inf\{\mathcal{B}_{\mathrm{t},1}(u) : U_{\mathrm{t}} < u < D_{\mathrm{t}}\}$

We establish the law of $\ell^{\beta}_{\mathrm{t}^{lca}}(U_{\mathrm{t}})$ with a variation of the standard argument for local times of Markov processes. Let $U_{\mathrm{t}^{lca}} = \inf\{u \geq 0 : \beta(u) = \mathrm{t}^{lca}\}$ and consider the shifted process

$\beta(U_{\mathrm{t}^{lca}} + u) - \mathrm{t}^{lca}$ which is a Brownian motion started at 0 and stopped when it first hits $\mathrm{t} - \mathrm{t}^{lca}$, conditioned to reach $\mathrm{t} - \mathrm{t}^{lca}$ before it first hits $-\mathrm{t}^{lca}$. Note that if we define $U^{\beta}_{\mathrm{t}-\mathrm{t}^{lca}} = \inf\{u \geq U_{\mathrm{t}^{lca}} : \beta(U^{lca}_{\mathrm{t}} + u) - \mathrm{t}^{lca} = \mathrm{t} - \mathrm{t}^{lca}\}$ we have that $U^{\beta}_{\mathrm{t}-\mathrm{t}^{lca}} = U_{\mathrm{t}} - U_{\mathrm{t}^{lca}}$. If we further define $U^{\beta}_{-\mathrm{t}^{lca}} = \inf\{u \geq U_{\mathrm{t}^{lca}} : \beta(U_{\mathrm{t}^{lca}} + u) - \mathrm{t}^{lca} = -\mathrm{t}^{lca}\}$ and we let $\ell^{\beta}_{0}(u)$ to be the local time at 0 of the shifted process $\beta(U_{\mathrm{t}^{lca}} + u) - \mathrm{t}^{lca}$, then we have that $\ell^{\beta}_{\mathrm{t}^{lca}}(U_{\mathrm{t}}) = \ell^{\beta}_{0}(U^{\beta}_{\mathrm{t}-\mathrm{t}^{lca}} \wedge U^{\beta}_{-\mathrm{t}^{lca}})$. For establishing the law of $\ell^{\beta}_{0}(U^{\beta}_{\mathrm{t}-\mathrm{t}^{lca}} \wedge U^{\beta}_{-\mathrm{t}^{lca}})$ we now appeal to a standard result on local times for Brownian motion (see e.g.[23]) which says that

$$\mathbf{P}\big[\ell^{\beta}_{0}(U^{\beta}_{\mathrm{t}-\mathrm{t}^{lca}} \wedge U^{\beta}_{-\mathrm{t}^{lca}}) > x\big] = \exp\big(-(\frac{1}{\mathrm{t} - \mathrm{t}^{lca}} + \frac{1}{\mathrm{t}^{lca}})x\big), \text{ for } x > 0.$$

In other words, $\ell^{\beta}_{0}(U^{\beta}_{\mathrm{t}-\mathrm{t}^{lca}} \wedge U^{\beta}_{-\mathrm{t}^{lca}})$ has the distribution of an Exponential random variable with parameter $1/(\mathrm{t} - \mathrm{t}^{lca}) + 1/\mathrm{t}^{lca}$. It immediately follows that $\ell_{\mathrm{t}^{lca}}$ has the distribution of a Gamma random variable with parameters $\lambda(\mathrm{t}, \mathrm{t}^{lca}) = 1/(\mathrm{t} - \mathrm{t}^{lca}) + 1/\mathrm{t}^{lca}$ and 2.

Hence the conditional law of $\ell_{\mathrm{t}^{lca}}$, given t and t^{lca}, is

$$f_{\ell_{\mathrm{t}^{lca}}|\mathrm{t},\mathrm{t}^{lca}}(r) = \lambda(\mathrm{t}, \mathrm{t}^{lca})^{2} r \exp\big(-r\lambda(\mathrm{t}, \mathrm{t}^{lca})\big), \; r > 0.$$

So that the joint law of $\mathrm{t}, \mathrm{t}^{lcs}$, and $\ell_{\mathrm{t}^{lca}}$ is

$$f_{t,t^{lca},\ell_{t^{lca}}}(t, s, r) = \frac{1}{(t-s)^2}\frac{1}{s^4}\, r \exp(-\frac{1}{s} - \frac{tr}{s(t-s)}), \; 0 < s < t, 0 < r. \tag{2.16}$$

And also the marginal law of $\ell_{\mathrm{t}^{lca}}$ is

$$f_{\ell_{\mathrm{t}^{lca}}}(r) = \frac{2}{(1+r)^3}, \; r > 0. \tag{2.17}$$

We can now draw approximate values for the statistics of the genealogy of n-extant species : $t_n = n\mathrm{t}, t^{lca}_{n} = nt^{lca}$, and $N(t^{lca}_{n}) = n\ell_{\mathrm{t}^{lca}}$, using these asymptotic distributions.

Chapter 3

Genealogy of Sampled Extinct Individuals

In this chapter we define the sampling process model for the fossil record, and consider the genealogical history of the recorded extinct individuals together with the extant ones. We introduce a way of representing this joint genealogical history with a point-process that includes the genealogical point-process (named the *p-sampled historical point-process*), and derive the law of this process (Lemma 8). We further introduce a similar point-process defined from a conditioned Brownian excursion (named the p-*sampled continuum historical point-process*), and show that this is precisely the asymptotic process of the rescaled p-sampled historical point-processes as the number of extant individuals increases, while the number of recorded extinct individuals remains finite (Theorem 11). In the course, we also prove a result about the asymptotic behavior of critical branching processes of known extinction time whose individuals are p-sampled (Lemma 10).

3.1 p-Sampled Historical Point-process

We now consider extending the analysis of the ancestry of extant individuals to include some proportion of the extinct individuals as well. Suppose that each individual in the past has independently had a given chance p of appearing in the historical record. We indicate such sampling of extinct individuals by putting a star mark on the leaf of $\mathcal{T}_{t,n}$ corresponding to the recorded individual. An example of a realization of such p-sampling is shown in Figure 3.1(a), and the induced sampling in the contour process in Figure 3.1(b).

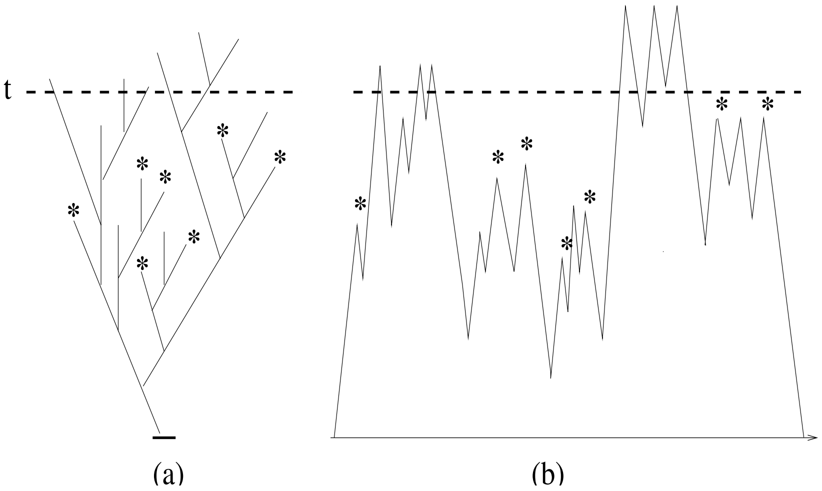

Figure 3.1: (a) The tree $\mathcal{T}_{t,n}$ with p-sampling on its individuals (the sampled individuals are represented by $*$'s); (b) The contour process of this tree with the sampling on the corresponding local maxima.

The goal is to combine the information on the sampled extinct individuals, with our analysis of the ancestry of the extant ones. In order to do so we extend our earlier notions of the genealogy of the extant individuals and of the genealogical point-process.

Let the *p-sampled history* of extant individuals at time t be defined as the smallest subtree of the family tree which contains all the edges representing both the ancestry of the extant individuals as well as of all of the p-sampled extinct individuals. We denote the p-sampled history of extant individuals at t in $\mathcal{T}_{t,n}$ by $\mathcal{G}_p(\mathcal{T}_{t,n})$. Note that by definition $\mathcal{G}_p(\mathcal{T}_{t,n})$ contains the genealogy $\mathcal{G}(\mathcal{T}_{t,n})$ (which would correspond to a 0-sampled history). It is in fact convenient to think of $\mathcal{G}_p(\mathcal{T}_{t,n})$ as consisting of the "main genealogical tree" $\mathcal{G}(\mathcal{T}_{t,n})$, and a collection of "$p$-sampled subtrees" attached to this main tree linking with additional branches the ancestry of p-sampled extinct individuals. Figure 3.2(a) shows the p-genealogical subtree of the tree from Figure 3.1(a). We next extend the notion of the genealogical point-process to represent this enriched p-sampled genealogy. We construct a point-process representation of $\mathcal{G}_p(\mathcal{T}_{t,n})$ so that it contains $\Pi_{t,n}$ as its "main points".

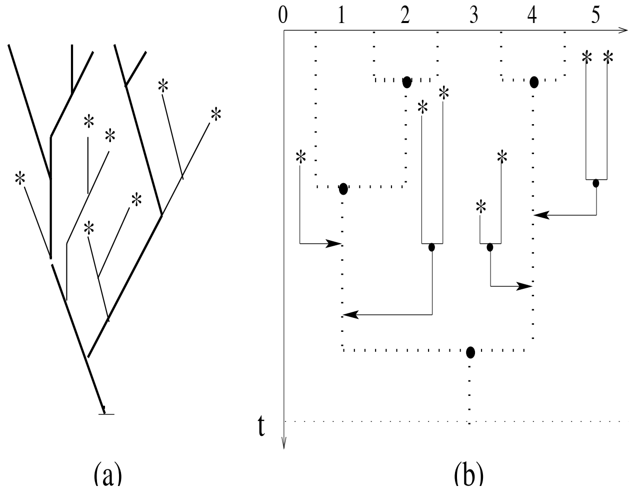

Figure 3.2: (a) The p-sampled tree $\mathcal{G}_p(\mathcal{T}_{t,n})$; the "main tree" (in bold) has the "p-sampled subtrees" attached to it; (b) The point-process representation $\Xi^p_{t,n}$ of $\mathcal{G}_p(\mathcal{T}_{t,n})$; each of the "main points" (large dots) have an associated left set and a right set representing the p-sampled subtrees attaching to the left and right of that branch-point.

Informally, think of extending the point-process $\Pi_{t,n}$ (representing $\mathcal{G}(\mathcal{T}_{t,n})$), by adding sets representing the p-sampled subtrees as follows. At each branch-point of the main tree there is a set of p-sampled subtrees attached to the edges of the main tree on the left of this branching point, and a set of p-sampled subtrees attached on the right of this branching point (see Figure 3.2(a)). We associate to each branch-point at height A_i, a left set $\mathcal{L}_i$ and a right set $\mathcal{R}_i$, which shall represent these sets of subtrees. Each such $\mathcal{L}_i$ and $\mathcal{R}_i$ needs to contain the following information: the heights $a_{i,L}(j)$ and $a_{i,R}(j)$ at which the p-sampled subtrees get attached to the edges of the main tree (as before we shall keep track of these heights as distances from level t in terms of $t_{i,L}(j) = t - a_{i,L}(j)$ and $t_{i,R}(j) = t - a_{i,R}(j)$); and the shape of the subtrees $\Upsilon_{i,L}(j)$ and $\Upsilon_{i,R}(j)$ themselves (the indexing $j \geq 0$ on the subtrees is induced by a depth-first search forwards to the branch-point at A_i for the left sets and a depth-first search backwards to the branch-point at A_i for the right sets). The point-process representing the p-sampled genealogical tree from Figure 3.2(a) is shown in Figure 3.2(b). In order to describe the law of the p-subtrees it will also be convenient to keep track of the height $h_{i,L}(j)$ and $h_{i,L}(j)$ of the subtrees $\Upsilon_{i,L}(j)$ and $\Upsilon_{i,R}(j)$.

Formally, we define the point-process of $\mathcal{G}_p(\mathcal{T}_{t,n})$ from the contour process $\mathcal{C}_{\mathcal{T}_{t,n}}$. The p-sampling on the tree is represented by the sampling of the local maxima of $\mathcal{C}_{\mathcal{T}_{t,n}}$. From the definition of $\mathit{\Pi}_{t,n}$, we have the heights of the branch-points of $\mathcal{G}(\mathcal{T}_{t,n})$ to be $A_i = \inf\{\mathcal{C}_{\mathcal{T}_{t,n}}(u) : D_i < u < U_{i+1}\}$, occurring in the contour process $\mathcal{C}_{\mathcal{T}_{t,n}}$ at times $B_i = \mathrm{argmin}\{\mathcal{C}_{\mathcal{T}_{t,n}}(u) : u \in (D_i, U_{i+1})\}$. The set $\mathcal{L}_i$, representing the set of p-subtrees attaching to the edges of $\mathcal{G}(\mathcal{T}_{t,n})$ on the left of the branch-point A_i, is defined from the part of the excursion of $\mathcal{C}_{\mathcal{T}_{t,n}}$ below t before time B_i. In other words if, for $X_{\mathcal{T}_{t,n}} = (\mathcal{C}_{\mathcal{T}_{t,n}}, \mathrm{slope}[\mathcal{C}_{\mathcal{T}_{t,n}}])$, we define

$$e^{<t}_{i,L}(u) = X_{\mathcal{T}_{t,n}}(D_i + u), u \in [0, B_i - D_i),$$

then $\mathcal{L}_i$ is completely defined by $e^{<t}_{i,L}$. Analogously $\mathcal{R}_i$ is defined from the part of the excursion of $\mathcal{C}_{\mathcal{T}_{t,n}}$ below t after time B_i, in other words if we define

$$e^{<t}_{i,R}(u) = X_{\mathcal{T}_{t,n}}(U_{i+1} - u), u \in [0, U_{i+1} - B_i)$$

then it is completely defined by $e^{<t}_{i,R}$ (the subscripts $_L$ and $_R$ reflect whether the entities are involved in defining $\mathcal{L}_i$ or $\mathcal{R}_i$). Note that the $e^{<t}_{i,L}$ runs forwards up to time B_i, while $e^{<t}_{i,R}$ runs backwards. On the extreme ends, we have the set of p-subtrees on the far left of the main tree defined by the part of $\mathcal{C}_{\mathcal{T}_{t,n}}$ prior to the first up-crossing time U_1, $e^{<t}_{0,L}(u) = X_{\mathcal{T}_{t,n}}(U_1 - u), u \in [0, U_1)$. Analogously, the set of p-subtrees on the right of the last branching point is defined by the part of $\mathcal{C}_{\mathcal{T}_{t,n}}$ after the last down-crossing time D_n, $e^{<t}_{n,R}(u) = X_{\mathcal{T}_{t,n}}(D_n + u), u \in [0, U_{(0,-1)} - D_n)$, where $U_{(0,1)} = \inf\{u \geq 0 : X_{\mathcal{T}_{t,n}} = (0, -1)\}$.

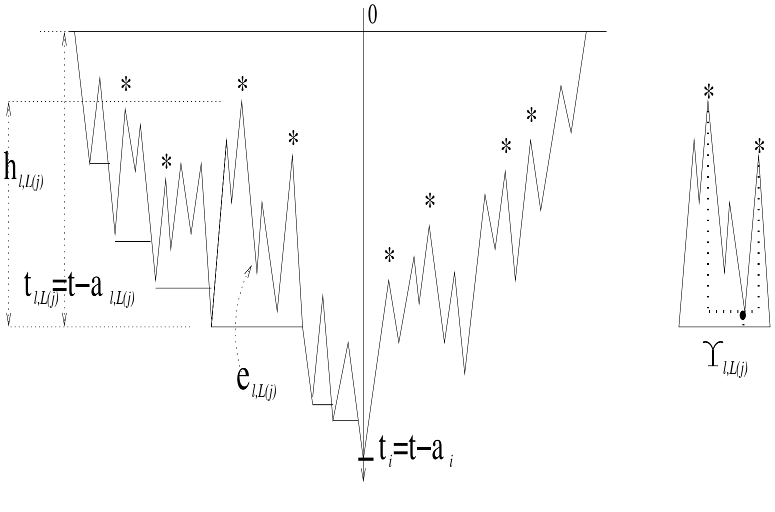

Figure 3.3: The left half $e^{<t}_{i,L}$ of an excursion of $\mathcal{C}_{\mathcal{T}_{t,n}}$ below t, with its infimum process $\varsigma_{i,L}$ whose levels of constancy are $\{a_{i,L}(j)\}_j$, above which lie the p-marked subtrees $\{\Upsilon_{i,L}(j)\}_j$ of heights $\{h_{i,L}(j)\}_j$.

In order to define the sets $\mathcal{L}_i$ and $\mathcal{R}_i$ we also need to define the processes

$$\begin{aligned} \varsigma_{i,L}(u) &= \inf_{0\leq v\leq u} e^{<t}_{i,L}(v), \quad u \in [0, B_i - D_i), \quad \text{and} \\ \varsigma_{i,R}(u) &= \inf_{0\leq v\leq u} e^{<t}_{i,R}(v), \quad u \in [0, U_{i+1} - B_i). \end{aligned}$$

The bijection between the tree $\mathcal{T}_{t,n}$ and its contour process $\mathcal{C}_{\mathcal{T}_{t,n}}$ implies that the heights at which the p-subtrees are attached to the edges of the main tree, are precisely the levels of constancy of the processes $\varsigma_{i,L}$ and $\varsigma_{i,R}$. Furthermore, the p-subtrees themselves have as their contour processes the excursions of $e^{<t}_{i,L} - \varsigma_{i,L}$ and $e^{<t}_{i,R} - \varsigma_{i,R}$ above these levels of constancy (see [21] for a detailed description). Figure 3.3 shows $e^{<t}_{i,L}$ together with its infimum process $\varsigma^{<t}_{i,L}$.

We define $a_{i,L}(j), j \geq 0$ to be the successive levels of constancy of $\varsigma_{i,L}$, and let $t_{i,L}(j) = t - a_{i,L}(j)$ be their distance form level t. For each level of constancy $a_{i,L}(j)$, let $e^{<t}_{i,L}(j)$ be the excursion of $e^{<t}_{i,L} - \varsigma_{i,L}$ that lies above the level $a_{i,L}(j)$. Let $h_{i,L}(j)$ be the height of this excursion, $h_{i,L}(j) = \sup(e^{<t}_{i,L}(j))$, and let $\Upsilon_{i,L}(j)$ be the tree whose contour process is the excursion $e^{<t}_{i,L}(j)$. Figure 3.3 shows an excursion $e^{<t}_{i,L}(j)$ with the p-subtree $\Upsilon_{i,L}(j)$ it defines. Note that all the star marks due to p-sampling are contained in the excursions $e^{<t}_{i,L}(j)$, hence are contained in the subtrees $\Upsilon_{i,L}(j)$. An analogous definition leads to $a_{i,R}(j), j \geq 0$, $h_{i,R}(j)$, and $\Upsilon_{i,L}(R)$ from $e^{<t}_{i,R}(j)$ and $\varsigma_{i,R}(j)$. With each point (ℓ_i, t_i) of $\mathit{\Pi}_{t,n}$ we now associate the sets

$$\mathcal{L}_i = \{(t_{i,L}(j), \Upsilon_{i,L}(j))\}_{j\geq 0}, \text{ and } \mathcal{R}_i = \{(t_{i,R}(j), \Upsilon_{i,R}(j))\}_{j\geq 0}. \tag{3.1}$$

In addition, for extreme ends we define one set $\mathcal{R}_0$ from $e^{<t}_{0,R}$, and we define a set $\mathcal{L}_n$ from $e^{<t}_{n,L}$. For ease of future notation we set $\mathcal{L}_0 = \emptyset$, $\mathcal{R}_n = \emptyset$, $(\ell_0, t) = (1, t)$, and $(\ell_n, t_n) = (n, t)$.

Definition. The *p-sampled historical point-process* $\Xi^p_{t,n}$ is the random set

$$\Xi^p_{t,n} = \{(l_i, t_i, \mathcal{L}_i, \mathcal{R}_i) : (l_i, t_i) \in \mathit{\Pi}_{t,n},\, 0 \leq i \leq n\} \tag{3.2}$$

Remark. We have in fact implicitly defined a point-process representation $\Xi_{t,n}$ of a complete historical point-process (which would correspond to 1-sampling). The difference between $\Xi_{t,n}$ and $\Xi^p_{t,n}$ is only in the $*$'s on the leaves in the latter. It will however be clear that for nice asymptotic behavior we need to consider $\Xi^p_{t,n}$ with $p < 1$; in other words we can only keep track of a proportion of the extinct individuals.

We can now derive the law of the point-process $\Xi^p_{t,n}$. For this we shall also need the law of the p-subtrees appearing in the sets $\mathcal{L}_i$ and $\mathcal{R}_i$. Let $\mathbf{T}$ denote the space of finite rooted binary trees with edge-lengths, and Λ denote the law on $\mathbf{T}$ of the tree $\mathcal{T}$. Then, let Λ^p denote the law on $\mathbf{T}$ induced by the p-sampling on the tree $\mathcal{T}$. Further, for any $h > 0$, let Λ^p_h denote the law induced by restricting Λ^p to the trees $\mathcal{T}$ of height h.

In order to describe the law of $\Xi^p_{t,n}$ we use a more careful and detailed analysis of the structure of the contour process $\mathcal{C}_{\mathcal{T}_{t,n}}$. First we use the result of Lemma 3, which gives us the law of the main points of $\Xi^p_{t,n}$. Then conditional on the location of the main points, we give the law of the sets $\mathcal{L}_i$ and $\mathcal{R}_i$ of p-subtrees. We show that the sets $\mathcal{L}_i$ and $\mathcal{R}_i$ are independent Poisson point-processes. The intensity measure of each such a set is given by the following. First, choose $t_{i,L}(j)$, the distances below t at which the p-sampled subtrees are getting attached, uniformly over t_i, the total distance below t to the ith branch-point. Next, choose $h_{i,L}(j)$, the height for each p-subtree, according to the same law as that of the height of a tree $\mathcal{T}$ whose height is known to be less than $t_{i,L}(j)$. Finally, choose the law of $\Upsilon_{i,L}(j)$, the attaching subtree, according to the law Λ^p_h described above.

Lemma 8. *For any fixed $0 < p < 1$, the law of the random set $\Xi^p_{t,n}$ is given by:*

- $\{(l_i, t_i) : 1 \le i \le n-1\}$ *is the simple point-process $\Pi_{t,n}$ of Lemma 3, and*
- *given $\{(l_i, t_i), 1 \le i \le n-1\}$: the sets $\mathcal{L}_i$ and $\mathcal{R}_i$ are independent; and for each $0 \le i \le n$ the random sets $\mathcal{L}_i$ and $\mathcal{R}_i$ are Poisson point-processes on $\mathbb{R}^+ \times \mathbf{T}$ with intensity measure*
$$1_{\{0<t<t_i\}}dt \;\; 1_{\{0<h<t\}}\frac{dh}{(1+h)^2}\frac{1+t}{t} \;\; \Lambda^p_h \tag{3.3}$$

Proof. The proof relies on the multiple reconnaissance of the appearance of (conditioned versions of) an alternating walk with Exponential(rate 1) steps within the contour process $\mathcal{C}_{\mathcal{T}_{t,n}}$. From earlier we have that the excursions of $\mathcal{C}_{\mathcal{T}_{t,n}}$ below t are independent, and the law of their depths below t is given by Lemma 3. We further show that for such an excursion, given its depth is t_i, the part before its lowest point is independent of the part after it. In fact, if the former is run forwards to the lowest point, and the latter backwards to the lowest point, then the two parts have the same law as well. This law is the same as that of $t-$an alternating exponential step walk, conditioned to reach t_i before it comes back to

0. Its Markovian property further leads to a simple description of the law of the levels of constancy of its infimum process. The excursions above the levels of constancy of the infimum are then shown to be copies of this alternating exponential step walk, conditioned on its maximal height.

The independence of the sets $\mathcal{L}_i$ over the index i follows from the independence of the excursions $e_i^{<t}$ of $X_{\mathcal{T}_{t,n}}$ below level t (the same holds for the sets $\mathcal{R}_i$). The strong Markov property of $X_{\mathcal{T}}$ also gives the independence of $\mathcal{R}_0$ and $\mathcal{L}_n$ from these sets as well. Then given t_i, the conditional independence and the equality in law of $\mathcal{L}_i$ and $\mathcal{R}_i$, follow from the time reversibility and the strong Markov property of $X_{\mathcal{T}}$. Consider the left half $e_{i,L}^{<t}$ of an excursion below level t. By Lemma 3, the conditional law of $t - e_i^{<t}$ given t_i is that of $X_{\mathcal{T}}|\{\sup(\mathcal{C}_{\mathcal{T}}) = t_i\}$. Hence, the law of $t - e_{i,L}^{<t}$ is that of $X_{\mathcal{T}}|\{\tau_{t_i} < \tau_0\}$ where τ_{t_i}, τ_0 are the first hitting times by $X_{\mathcal{T}}$ of $(t_i, +1), (0, -1)$ respectively. Now, consider the levels of constancy $\{a_{i,L}(j)\}_j$ of $\varsigma_{i,L} = \inf(e_{i,L}^{<t})$. If $t_{i,L}(j) = t - a_{i,L}(j)$, then $\{t_{i,L}(j)\}_j$ are levels of constancy of $t - \varsigma_{i,L} = \sup(t - e_{i,L}^{<t})$. The fact that $\mathcal{C}_{\mathcal{T}}$ is an alternating sum of exponential variables implies that $\{t_{i,L}(j)\}_j$ form a Poisson process of rate 1 on the set $(0, t_i)$. It also implies that the excursions $\{e_{i,L}^{<t}(j)\}_j$ of $e_{i,L}^{<t} - \varsigma_{i,L}$ above these levels of constancy have the laws of $X_{\mathcal{T}}|\{\sup(X_{\mathcal{T}}) < t_{i,L}(j)\}$. Hence for each j, given $t_{i,L}(j)$ the law of $h_{i,L}(j) = \sup(e_{i,L}^{<t}(j))$, by (2.4) of Lemma 3, has the density

$$\frac{dh}{(1+h)^2}\,\frac{1+t_{i,L}(j)}{t_{i,L}(j)}$$

on the set $(0, t_{i,L}(j))$. Then given the value of $h_{i,L}(j)$, for each j the excursion $e_{i,L}^{<t}(j)$ has the law of $X_{\mathcal{T}}|\{\sup(X_{\mathcal{T}}) = h_{i,L}(j)\}$, hence the tree defined by $e_{i,L}^{<t}(j)$ as its contour process has the law of of $\mathcal{T}_{\Delta = h_{i,L}(j)}$. Now, the strong Markov property implies that the p-sampling on the local maxima of $\mathcal{C}_{\mathcal{T}_{t,n}}$ is for each $e_{i,L(j)}^{<t}$ again a Bernoulli p-sampling on its local maxima. Thus the law of the p-sampled tree $\Upsilon_{i,L}(j)$ is $\Lambda^p_{h_{i,L}(j)}$. Putting all the above results together we have that the set $\{(t_{i,L}(j), \Upsilon_{i,L}(j))\}_{j\geq 0}$ is a Poisson point-process with intensity measure

$$1_{\{0<t<t_i\}}dt\ \ 1_{\{0<h<t\}}\frac{dh}{(1+h)^2}\ \ \Lambda^p_h$$

□

3.2 p-Sampled Continuum Historical Point-process

Let us now consider the implications that the p-sampling of extinct individuals has in the asymptotic context. In Section 2, the genealogical point-process was defined from the contour process $\mathcal{C}_{\mathcal{T}_{t,n}}$, and its asymptotics process was identified as the continuum genealogical point-process defined from a Brownian excursion $\mathcal{B}_{\mathrm{t},1}$ conditioned to have local time 1 at level t.

The p-sampled historical process is defined from a contour process $\mathcal{C}_{\mathcal{T}_{t,n}}$ whose local maxima are sampled independently with equal chance p. In terms of the (horizontal) u-coordinate of $\mathcal{C}_{\mathcal{T}_{t,n}}$ the p-sampled individuals form a random set of marks on $\mathbb{R}^+$. The fact that $\mathcal{C}_{\mathcal{T}}$ is an alternating sum of independent Exponential(rate 1) random variables implies that the random set formed by the local maxima of $\mathcal{C}_{\mathcal{T}}$ is a Poisson process of rate 1/2 on $\mathbb{R}^+$, and the same still holds for the sets formed by the local maxima of each part of an excursion of $\mathcal{C}_{\mathcal{T}_{t,n}}$ below t. If we further sample these local maxima independently with chance p we have a Poisson process of rate $p/2$ on $\mathbb{R}^+$. For the asymptotics, the appropriate rescaling, as in Section 2, speeds up the time axis of $\mathcal{C}_{\mathcal{T}_{t,n}}$ by n. Hence if we consider p_n such that $np_n \to \mathrm{p}$ as $n \to \infty$, then asymptotically the p_n-sampling on $\mathcal{C}_{\mathcal{T}_{t,n}}$ will converge to a Poisson process of rate p/2. This prompts us to consider for the asymptotics of the p-historical point-process a process similarly defined from a conditioned Brownian excursion $\mathcal{B}_{\mathrm{t},1}$ sampled according to a Poisson(rate p/2) process along its (horizontal) u-coordinate.

Remark. We are interested in obtaining an asymptotic point-process that has a.s. finitely many extinct individuals recorded. It is clear that thus the rate of sampling asymptotically has to satisfy $np_n \to \mathrm{p}$ as $n \to \infty$.

We define a process derived from a conditioned Brownian excursion $\mathcal{B}_{\mathrm{t},1}$ in the same manner that $\Xi^{p}_{t,n}$ was derived from the contour process of the conditioned branching process $\mathcal{C}_{\mathcal{T}_{t,n}}$. Recall that $\mathcal{B}(u), u \geq 0$ denotes a Brownian excursion, for a fixed $\mathrm{t} > 0$ $\ell_{\mathrm{t}}(u), u \geq 0$ is its local time at level t up to time u, $i_{\mathrm{t}}(\ell), \ell > 0$ is the inverse process of ℓ_{t}. Also, $\mathcal{B}_{\mathrm{t},1}(u), u \geq 1$ denotes the excursion $\mathcal{B}$ conditioned to have total local time at t equal to 1, and $(\ell, e^{<\mathrm{t}}_{\ell})$ denotes the set of excursions of $\mathcal{B}_{\mathrm{t},1}$ below level t indexed by the local time ℓ_{t} at the time of their beginning.

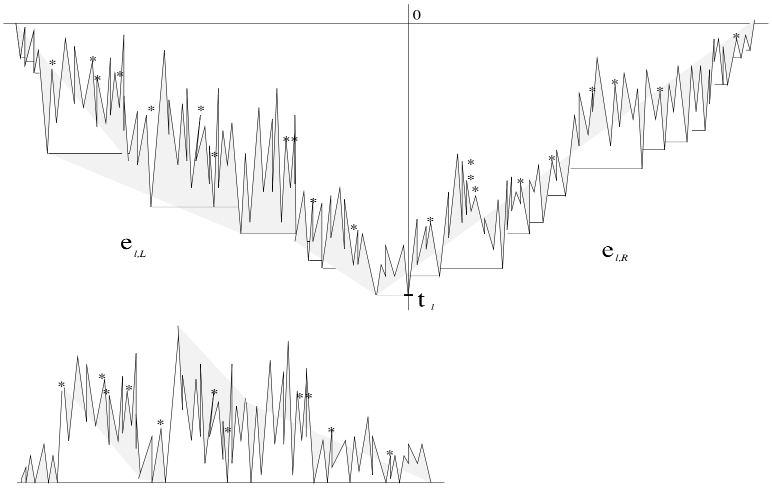

Figure 3.4: Top: An excursion $e_\ell^{<\mathrm{t}}$ of $\mathcal{B}_{\mathrm{t},1}$ below t, its left $e_{\ell,L}^{<\mathrm{t}}$ and right $e_{\ell,R}^{<\mathrm{t}}$ parts, with their infimum processes; Bottom: shows the process $e_{\ell,L}^{<\mathrm{t}} - \varsigma_{\ell,L}$.

Define the p-sampling on $\mathcal{B}_{\mathrm{t},1}$ to be a Poisson(rate p/2) process along the u-axis of $\mathcal{B}_{\mathrm{t},1}$. We indicate this by putting a star mark on the graph of $\mathcal{B}_{\mathrm{t},1}$ at the times of this Poisson process. Let $e_\ell^{<\mathrm{t}}$ be an excursion of $\mathcal{B}_{\mathrm{t},1}$ below level t

$$e_\ell^{<\mathrm{t}}(u) = \mathcal{B}_{\mathrm{t},1}(i_{\mathrm{t}}(\ell^-) + u),\ u \in [0, i_{\mathrm{t}}(\ell) - i_{\mathrm{t}}(\ell^-))$$

Recall that $a_\ell = \inf(e_\ell^{<\mathrm{t}})$ is its lowest point occurring at $u_\ell = \operatorname{argmin}\{e^{<\mathrm{t}}(u)\}$, and that $t_\ell = \mathrm{t} - t_\ell$ denotes its distance from level t. For each $e_\ell^{<\mathrm{t}}$ we define its left and right parts (relative to its lowest point) to be

$$\begin{aligned} e_{\ell,L}^{<\mathrm{t}}(u) &= \mathcal{B}_{\mathrm{t},1}(i_{\mathrm{t}}(\ell^-) + u), u \in [0, u_\ell - i_{\mathrm{t}}(\ell^-)) \text{ and} \\ e_{\ell,R}^{<\mathrm{t}}(u) &= \mathcal{B}_{\mathrm{t},1}(i_{\mathrm{t}}(\ell) - u), u \in [0, i_{\mathrm{t}}(\ell) - u_\ell) \end{aligned}$$

Note that $e_{\ell,L}^{<\mathrm{t}}$ runs forwards to the lowest point of $e_\ell^{<\mathrm{t}}$, while $e_{\ell,R}^{<\mathrm{t}}$ runs backwards in time to it. We shall also need their respective processes of infima

$$\begin{aligned} \varsigma_{\ell,L}(u) &= \inf_{0 \le v \le u} e_{\ell,L}^{<\mathrm{t}}(v), u \in [0, u_\ell - i_{\mathrm{t}}(\ell^-)) \text{ and} \\ \varsigma_{\ell,L}(u) &= \inf_{0 \le v \le u} e_{\ell,L}^{<\mathrm{t}}(v), u \in [0, u_\ell - i_{\mathrm{t}}(\ell^-)) \end{aligned}$$

Figure 3.4 shows $e_{\ell,L}^{<\mathrm{t}}$ and $e_{\ell,L}^{<\mathrm{t}}$ with $\varsigma_{\ell,L}$ and $\varsigma_{\ell,L}$.

We define $a_{\ell,L}(j), j \ge 0$ to be the successive levels of constancy of $\varsigma_{\ell,L}$, and we let $t_{\ell,L}(j) = \mathrm{t} - a_{\ell,L}(j)$ be their distance to level t. For each level of constancy $a_{\ell,L}(j)$, let $e_{\ell,L}^{<\mathrm{t}}(j)$ be the excursion of $e^{<\mathrm{t}}_{\ell,L} - \varsigma_{\ell,L}$ that lies above the level $a_{\ell,L}(j)$. Let $h_{\ell,L}(j) = \sup(e^{<\mathrm{t}}_{\ell,L}(j))$ be the height of this excursion. Note that a.s. all the p-sampled points on $\mathcal{B}_{\mathrm{t},1}$ lie on these excursions $e^{<\mathrm{t}}_{\ell,L}(j)$. We define a tree $\Upsilon_{\ell,L}(j)$ induced by such a p-sampled excursion $e^{<\mathrm{t}}_{\ell,L}(j)$, as the tree whose contour process is the linear interpolation of the sequence of the values of $e^{<\mathrm{t}}_{\ell,L}(j)$ at the p-sampling times, alternating with the sequence of the minima of $e^{<\mathrm{t}}_{\ell,L}(j)$ between the p-sampling times. An analogous definition leads to $a_{\ell,R}(j), j \geq 0$, $t_{\ell,R}(j), j \geq 0$ $h_{\ell,R}(j)$, and $\Upsilon_{\ell,L}(R)$ from $e^{<t}_{\ell,R}(j)$ and $\varsigma_{\ell,R}(j)$.

Remark. This definition of a tree from an excursion path sampled at given times has been explored for different sampling distributions in the literature (for some examples see [21] §6). Since for each $e^{<\mathrm{t}}_{\ell}$ there are a.s. only finitely many p-sampled points the trees $\{\Upsilon_{\ell,L}(j)\}_j, \{\Upsilon_{\ell,R}(j)\}_j$ are a.s. in the space $\mathbf{T}$ of rooted planar trees with edge-lengths and finitely many leaves.

With each point (ℓ, t_ℓ) of $\pi_{\mathrm{t},1}$ we now associate the sets

$$\mathcal{L}_\ell = \{(t_{\ell,L}(j), \Upsilon_{\ell,L}(j))\}_{j\geq 0}, \text{ and} \mathcal{R}_\ell = \{(t_{\ell,R}(j), \Upsilon_{\ell,R}(j))\}_{j\geq 0} \tag{3.4}$$

We also define the first "right" set $\mathcal{R}_0$ and the last "left" set $\mathcal{L}_1$ from paths $e^{<\mathrm{t}}_{0,R}$ of $\mathcal{B}_{\mathrm{t},1}$ before the first hitting time of t, and $e^{<\mathrm{t}}_{1,L}$ of $\mathcal{B}_{\mathrm{t},1}$ after the last hitting time of t. For ease of notation we let $\mathcal{L}_0 = \mathcal{R}_1 = \emptyset$, $\mathrm{t} = t_1 = \mathrm{t}$.

Definition. The p-*sampled continuum historical point-process* $\xi^{\mathrm{p}}_{\mathrm{t},1}$ is the random set

$$\xi^{\mathrm{p}}_{\mathrm{t},1} = \{(\ell, t_\ell, \mathcal{L}_\ell, \mathcal{R}_\ell) : (\ell, t_\ell) \in \pi_{\mathrm{t},1},\, i_{\mathrm{t}}(\ell^-) \neq i_{\mathrm{t}}(\ell)\} \tag{3.5}$$

We next derive law of the point-process $\xi^{\mathrm{p}}_{\mathrm{t},1}$. For this we shall also need the law of the trees induced by the p-sampled excursions of $e^{<\mathrm{t}} - \varsigma$. Let λ^{p} denote the law on the space $\mathbf{T}$ induced by a $\mathcal{B}$ sampled at Poisson(rate p) points (in the sense of the bijection between sampled continuous functions and trees, [3], same as the definition of $\Upsilon_{\ell,L}(j)$ from the p-sampled $e_{\ell,L}(j)$). Then, for any $h > 0$, let λ^{p}_h denote the law induced by restricting λ^{p} to the set of Brownian excursions $\mathcal{B}$ of height h.

In order to derive the law of we exploit in a more detailed manner the nice properties of Brownian excursions. We first use the result of Lemma 4, which gives us the law of the set

$\{(\ell, t_\ell) : i_{\mathrm{t}}(\ell^-) \neq i_{\mathrm{t}}(\ell)\}$. Then conditional on this set we give the law of the sets $\mathcal{L}_\ell$ and $\mathcal{R}_\ell$. We show that $\{\mathcal{L}_\ell, \mathcal{R}_\ell\}_\ell$ are independent Poisson point-processes. the intensity measure of each such set is given by the following. First, choose $t_{\ell,L}(j)$, the distances below t at which the p-sampled subtrees excursions of $e^{<\mathrm{t}}_{\ell,L} - \varsigma_{\ell,L}$ occur uniformly over t_ℓ, the distance below t of the lowest point of $e^{<\mathrm{t}}_{\ell}$. Next, choose $h_{\ell,L}(j)$, the height for each such p-sampled excursion, according to the same law as that of the height of a $\mathcal{B}$ whose height is known to be less than $t_{\ell,L}(j)$. Finally, choose the law of the induced tree $\Upsilon_{\ell,L}(j)$ according to the law λ^{p}_h described above.

Lemma 9. *The random set $\xi^{\mathrm{p}}_{\mathrm{t},1}$ is such that:*

- $\{(\ell, t_\ell) : i_{\mathrm{t}}(\ell^-) \neq i_{\mathrm{t}}(\ell)\}$ *is the Poisson point-process $\pi_{\mathrm{t},1}$ of Lemma 4, and*
- *given $\{(\ell, t_\ell) : i_{\mathrm{t}}(\ell^-) \neq i_{\mathrm{t}}(\ell)\}$ the sets $\mathcal{L}_\ell$ and $\mathcal{R}_\ell$ are independent; and for each $\ell : i_{\mathrm{t}}(\ell^-) \neq i_{\mathrm{t}}(\ell)$ $\mathcal{L}_\ell$ and $\mathcal{R}_\ell$ are Poisson point-processes on $\mathbf{R}^+ \times \mathbf{T}$ with intensity measure*
$$1_{\{0<t<t_\ell\}} dt \;\; 1_{\{0<h<t\}} \frac{dh}{h^2} \;\; \lambda^{\mathrm{p}}_h \tag{3.6}$$

Proof. The proof proceeds in much of the same steps as the one for deriving the law of the p-sampled historical process $\Xi^p_{t,n}$. The notable difference is that we now have to resort to more sophisticated Markovian results on the decomposition of a Brownian path, such as the Williams decomposition of a Brownian excursion given its height, and the Pitman theorem on Bessel processes. In short, we consider the decomposition of the conditioned Brownian excursion $\mathcal{B}_{\mathrm{t},1}$ into its excursions below level t provided by the Lemma 4. For each such excursion below t given its lowest point at distance t_ℓ below t, Williams' decomposition gives us the independence and identity in law of its left and right parts, as well as the description of their laws in terms of a 3-dimensional Bessel process. Furthermore, we can use Pitman's theorem that describes the law of the excursions of this Bessel process above the levels of constancy of its future infimum. After taking care of some conditioning issues, this finally gives us a simple description of these excursions above the levels of constancy as simply Brownian excursions conditioned on their maximal height.

The independence of the sets $\mathcal{L}_\ell$ over the index ℓ (the same holds for the sets $\mathcal{R}_\ell$) follows from the independence of the excursions of $\mathcal{B}_{\mathrm{t},1}$ below level t. This also holds (by the strong

Markov property of $\mathcal{B}$) for the sets $\mathcal{R}_0$ and $\mathcal{L}_\infty$ defined from the parts of the path of $\mathcal{B}_{\mathrm{t},1}$ of its ascent to level t and its descent from it. For each $e_\ell^{<\mathrm{t}}$ excursion of $\mathcal{B}_{\mathrm{t},1}$ below level t, we let $e_\ell^+ = \mathrm{t} - e_\ell^{<\mathrm{t}}$. By Lemma 4, the conditional law of e_ℓ^+ given (ℓ, t_ℓ) is that of a Brownian excursion $\mathcal{B}$ conditioned on the value of its supremum $\mathcal{B}|\{\sup(\mathcal{B}) = t_\ell\}$. Let $\tau_{t_\ell} = \inf\{u > 0 : e_\ell^+(u) = t_\ell\}$, then by Williams' decomposition of a Brownian excursion $\mathcal{B}$ (e.g. [23] Vol.1 §III.49.), the law of $e_{\ell,L}^+ = \mathrm{t} - e_{\ell,L}^{<\mathrm{t}}$ is that of a Bess(3) (3-dimensional Bessel) process ρ stopped the first time $\tau_{t_\ell}^\rho = \inf\{u > 0 : \rho(u) = t_\ell\}$ it hits t_ℓ. By time reversibility of $\mathcal{B}$ the process

$$r_{\ell,L}(u) = t_\ell - e_{\ell,L}^+(\tau_{t_\ell} - u), u \in (0, \tau_{t_\ell})$$

also has the law of the stopped Bess(3) process $\rho(u), u \in (0, \tau_{t_\ell}^\rho)$. Let

$$j_{\ell,L}(u) = \inf_{u \le v \le \tau_{t_\ell}} r_{\ell,L}, u \in (0, \tau_{t_\ell})$$

Then $\{t_\ell - t_{\ell,L}(j)\}_j$ are (in reversed index order) the successive levels of constancy of the process $j_{\ell,L}(u), u \in (0, \tau_{t_\ell})$, $\{h_{\ell,L}(j)\}_j$ (in reversed index order) are the heights of the successive excursions from 0 of the process $r_{\ell,L}(u) - j_{\ell,L}(u), u \in (0, \tau_{t_\ell})$, and $\{\Upsilon_{\ell,L}(j)\}_j$ (in reversed index order) are the trees induced by the p-sampled points on these excursions. To obtain the law of $j_{\ell,L}$ and $r_{\ell,L} - j_{\ell,L}$ consider the Bess(3) process $\rho(u), u \ge 0$ and its future infimum process $\jmath(u) = \inf_{v \ge u} \rho(v), u \ge 0$. We note that the law of $j_{\ell,L}(u), u \in (0, \tau_{t_\ell})$ is equivalent to that of $\jmath(u), u \in (0, \tau_{t_\ell}^\rho)$ if $\jmath(\tau_{t_\ell}^\rho) = t_\ell$, in other words, if $\rho(u), u \ge 0$ after it first reaches t_ℓ never returns to that height again. So,

$$(j_{\ell,L}, r_{\ell,L} - j_{\ell,L}) \overset{d}{=} (\jmath, \rho - \jmath)|\{\jmath(\tau_{t_\ell}^\rho) = t_\ell\} \text{ for } u \in (0, \tau_{t_\ell})$$

By Pitman's theorem, then by Levy's theorem (e.g. [22] VI.§3. and§6.)

$$(\jmath, \rho - \jmath) \overset{d}{=} (\zeta, \zeta - \beta) \overset{d}{=} (\bar{\ell}, |\bar{\beta}|)$$

where β is a standard Brownian motion, ζ its supremum process; $|\bar{\beta}|$ is a reflected Brownian motion, $\bar{\ell}$ its local time at 0 (with the occupation time normalization). Thus, for $\bar{\tau}_{t_\ell} := \inf\{u \ge 0 : |\bar{\beta}|_u + \bar{\ell}_u = t_\ell\}$,

$$(j_{\ell,L}, r_{\ell,L} - j_{\ell,L}) \overset{d}{=} (\bar{\ell}, |\bar{\beta}|)\,|\{\bar{\ell}_{\bar{\tau}_{t_\ell}} = t_\ell\} \quad \text{for } u \in (0, \tau_{t_\ell})$$

The condition $\{\bar{\ell}_{\bar{\tau}_{t_\ell}} = t_\ell\}$ is equivalent to the condition $\{\bar{\ell}_{\bar{\tau}_{t_\ell}} = t_\ell, |\bar{\beta}|_{\bar{\tau}_{t_\ell}} = 0\}$ and $\{u < \bar{\tau}_{t_\ell} : \bar{\ell}_u < t_\ell, |\bar{\beta}|_u < t_\ell - \bar{\ell}_u\}$. Hence,

$$(j_{\ell,L}, r_{\ell,L} - j_{\ell,L}) \overset{d}{=} (\bar{\ell}, |\bar{\beta}|)\,|\{\bar{\ell}_u < t_\ell, |\bar{\beta}|_u < t_\ell - \bar{\ell}_u; \bar{\ell}_{\bar{\tau}_{t_\ell}} = t_\ell, |\bar{\beta}|_{\bar{\tau}_{t_\ell}} = 0\} \tag{3.7}$$

Since $\big(\bar{\ell},\, \sup(|\bar{\beta}|)\big)$ is a Poisson point-process with intensity measure $d\bar{\ell}d\bar{h}/\bar{h}^2$, then using the independence property of a Poisson random measure on disjoint sets in (3.7), we obtain for $t = t_\ell - \bar{\ell}$ that $\big(t_\ell - j_{\ell,L},\, \sup(r_{\ell,L} - j_{\ell,L})\big)$ is a Poisson point-process with intensity measure

$$1_{(0<t<t_\ell)}dt\, 1_{(0<h<t)}\frac{dh}{h^2}$$

Recall the relationship of the values $\{t_{\ell,L}(j), h_{\ell,L}(j), \Upsilon_{\ell,L}(j)\}_j$ of $\mathcal{L}_\ell$ with the processes $j_{\ell,L}$ and $r_{\ell,L} - j_{\ell,L}$. The above result thus implies that $\mathcal{L}_\ell$ is a Poisson point-process with intensity measure

$$1_{(0<t<t_\ell)}dt\;\; 1_{(0<h<t)}\frac{dh}{h^2}\;\; \lambda_h^{\mathrm{p}}$$

where the last factor comes from the fact that $\Upsilon_{\ell,L}(j)$ is just the tree induced by the p-sampled excursion of $|\bar{\beta}|$ of height $h_{\ell,L}(j)$. □

Our next goal is to show that the process $\xi_{\mathrm{t},1}^{\mathrm{p}}$ whose law we have just obtained, is indeed the asymptotic result of the processes $\Xi_{t,n}^{p}$ after appropriate rescaling. In order to do so, we first must show that the laws $\Lambda_h^{p_n}$ on the space of trees converge as $n \to \infty$ to the law λ_h^{p} if $np_n \to \mathrm{p}$.

3.3 p-Sampled Trees $\mathcal{T}$ of Given Height h

We need to consider more closely the trees $\Upsilon_{i,L}(j)$ and $\Upsilon_{\ell,L}(j)$ induced by the sampled excursions appearing in the historical point-processes above. In both cases we have an excursion, $\mathcal{C}_\mathcal{T}$ or $\mathcal{B}$, of a given height and with marks on it produced by a sampling process. Laws of the trees induced by sampled excursions of unrestricted height can be very simply and elegantly described (see [13] for the case of $\mathcal{B}$). However, for the trees from excursions of a given height that we need to consider here, the description is much messier. We shall give next a recursive description that applies equally to define an $\Upsilon_{l,L}(j)$ from $\mathcal{C}_\mathcal{T}$ of a given height, or to define $\Upsilon_{\ell,L}j$ from $\mathcal{B}$ of a given height. A similar recursive description of an infinite tree induced by an unsampled Brownian excursion is given by Abraham [2].

Define the "spine" of the tree to extend from the root of the tree to the point of maximal height in the excursion. An equivalent representation of the tree is one in which the subtrees of the trees on the left and on the right of the axis through the spine are attached to this spine, and example of which is shown in Figure 3.5. We obtain the branch levels at which these subtrees are attached, as well as parameters needed for the description of the subtrees as follows.

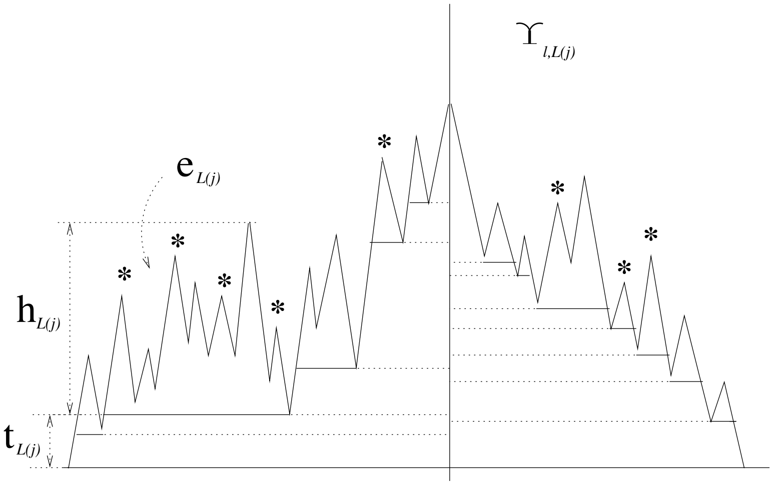

Figure 3.5: The “first” set in the recursive description consists of branch levels $\{t_L(j)\}_j$ at which subtrees induced by sampled excursions of $e_L - \varsigma_L$ are attached to the spine; and the heights $\{h_L(j)\}_j$ of these subtrees.

We denote the excursion function defining this tree by $e(u), u \geq 0$ (in other words $e = \mathcal{C}_{\mathcal{T}}$ or $e = \mathcal{B}$). Let h be its given height, and $U_h = \text{argmax}\{e(u) : u > 0\}$ the time at which it is achieved. Then let $e_L(u), u \in [0, U_h]$ be the left part of the excursion, and we also define its future infimum process $\varsigma_L(u) = \inf_{v \geq u} e(v), u \in [0, U_h]$. Then the subtrees attaching on the left of the spine are defined by the process $e_L - \varsigma_L$ and the set of sampled marks. They are precisely the trees induced by the sampled excursions $e_L(j)$ of $e_L - \varsigma_L$ whose height is some $h_L(j)$. The levels at which they are attached to the spine are the levels of constancy $t_L(j)$ of ς_L at which the excursions of $e_L - \varsigma_L$ occur. Thus the set $\{(t_L(j), h_L(j))\}_{j \geq 0}$ is the “first” set in our recursive definition of tress. The “second” set is derived in the same manner from the sampled excursions $\{e_L(j)\}_j$, and so on. We define these sets analogously for the right part of e.

This recursive procedure is clearly very similar to our definition of the left and right sets, $\mathcal{L}_i, \mathcal{R}_i$ for $e_i^{<t}$ and $\mathcal{L}_\ell, \mathcal{R}_\ell$ for $e_\ell^{<t}$ as defined earlier. The main difference is that the subtrees here are defined from excursions above the levels of constancy of the future infimum process

for e, whereas earlier they were defined from excursions above the levels of constancy of the past infimum process for $e_i^{<t}$ and $e_\ell^{<\mathrm{t}}$. However, time inversion and reflection invariance of the transition function of e will allow us to easily derive the laws of the "first" set of points here from the results of Lemma 8 and Lemma 9.

Recall that $\mathbf{T}$ is the space of finite rooted binary trees with edge-lengths. We defined Λ^p as the law on the space $\mathbf{T}$ induced by the p-sampling of the critical branching process $\mathcal{T}$, and we defined Λ_h^p to be the law induced by restricting Λ^p to the trees $\mathcal{T}$ of height h. Also recall that we defined λ^{p} as the law on the space $\mathbf{T}$ induced by a $\mathcal{B}$ sampled at Poisson(rate p) points, and we defined λ_h^{p} to be the law induced by restricting λ^{p} to the set of Brownian excursions $\mathcal{B}$ of height h.

In the next Lemma we give a recursive description of the law of $\Lambda_h^{p_n}$ and λ_h^{p}, and we show that we do have the convergence of the $\Lambda_h^{p_n}$ (appropriately rescaled) to λ_h^{p} if $np_n \to \mathrm{p}$.

Lemma 10. *The law $\Lambda_h^{p_n}$ of a tree induced by a p_n-sampled contour process $\mathcal{C}_\mathcal{T}$ of a given height h is such that the first sets of points $\{t_L(j), h_L(j)\}_j$ and $\{t_R(j), h_R(j)\}_j$ are independent Poisson point-processes with intensity measure*

$$\frac{1}{\sqrt{p_n}}\, 1_{(0<\tau<h)} d\tau\, 1_{(0<\kappa<h-\tau)} \frac{d\kappa}{(1+\kappa)^2} \frac{1+\tau}{\tau} \tag{3.8}$$

The law λ_h^{p} of a tree induced by a p-sampled Brownian excursion $\mathcal{B}$ of a given height h is such that the first sets of points $\{t_L(j), h_L(j)\}_j$ and $\{t_R(j), h_R(j)\}_j$ are independent Poisson point-processes with intensity measure

$$\frac{1}{\sqrt{\mathrm{p}}}\, 1_{(0<\tau<h)} d\tau\, 1_{(0<\kappa<h-\tau)} \frac{d\kappa}{\kappa^2} \tag{3.9}$$

Let $n^{-1}\Lambda_h^{p_n}$ be the law of the tree induced by a rescaled p_n-sampled contour process $\mathcal{C}_\mathcal{T}$ by n^{-1} in the vertical coordinate.
Then for any $\{p_n \in (0,1)\}_{n\geq 1}$ such that $np_n \underset{n\to\infty}{\to} \mathrm{p}$ we have $n^{-1}\Lambda_h^{p_n} \underset{n\to\infty}{\Longrightarrow} \lambda_h^{\mathrm{p}}$.

Proof. The key for this proof is to observe the following. If $e(u), u \geq 0$ is the p_n-sampled process $X_\mathcal{T}|\{\sup(\mathcal{C}_\mathcal{T}) = h\}$ then $e_L(u) = e(u), u \in [0, U_h]$ has the law of a p_n-sampled $X_\mathcal{T}|\{\tau_h < \tau_0\}$ where τ_h, τ_0 are the first hitting times of $(h, +1), (0, -1)$ respectively by $X_\mathcal{T}$. Then time reversibility and the reflection invariance of the transition function of

$X_{\mathcal{T}}$ imply that $h - e_L(U_h - u), u \in [0, U_h]$ has the same law as $e_L(u), u \in [0, U_h]$. Now the levels of constancy of ς_L, and the corresponding excursions $e_L - \varsigma_L$ above them, are equivalent to the levels of constancy and excursions of a set $\mathcal{L}_i$ considered in Lemma 8, thus giving a Poisson process of intensity measure as in (3.3). The factor $p^{-1/2}$ in the intensity measure (3.8) comes from the fact that here we only consider the excursions of $e_L - \varsigma_L$ that have at least one sampled mark in them. Namely, for the branching process $\mathcal{T}$, if N_{tot} denote the total population size of $\mathcal{T}$, then the generating function of N_{tot} is $\mathbf{E}(x^{N_{tot}}) = 1 - (1-x)^{1/2}$. Hence, the chance of at least one mark in the p_n-sampled point-process of $\mathcal{T}$ is $1 - \mathbb{E}((1-p_n)^{N_{tot}}) = {p_n}^{1/2}$.

A similar argument applies when $e(u), u \geq 0$ is the process $\mathcal{B}|\{\sup(\mathcal{B}) = h\}$ sampled at Poisson(rate p/2) times. Time reversibility and reflection invariance of the transition function of $\mathcal{B}$ allow us to identify that the law of the levels of constancy of of ς_L, and the corresponding excursions $e_L - \varsigma_L$ above them are the same as those for a set $\mathcal{L}_\ell$ considered in Lemma 9, which we know form a Poisson process with intensity measure as in (3.6). The factor $\mathrm{p}^{-1/2}$ in the intensity measure of (3.9) then comes from the rate of excursions with at least one sampled mark. Namely, a Poisson(rate p/2) process of marks on $\mathcal{B}$ along its time coordinate is in its local time coordinate a Poisson(rate $\mathrm{p}^{1/2}$) process of marks (see [23] Vol.2§VI.50.).

Now the law of the first set of the rescaled process with under $n^{-1}\Lambda_h^{p_n}$ converges to the law of the first set of the process with the law λ_h^{p}. This follows from the fact that the former is a sequence of Poisson point-processes whose support set and intensity measure converge to those of the latter Poisson point-process. Since for Poisson random measures the convergence of finite dimensional sets is sufficient to insure weak convergence of the whole process our claim follows for the first sets, and by recursion for the whole process. □

Finally, we can obtain the asymptotic result for the p_n-sampled historical point-processes. The rescaling of $\Xi_{t_n,n}^{p_n}$ is the same as that for $\Pi_{t,n}$. Both coordinates of $\Pi_{t,n}$ are rescaled by n^{-1}, so that the vertical coordinate of the sets $\mathcal{L}_i$,$\mathcal{R}_i$ is also rescaled by n^{-1}, and the sampling rate is rescaled by n. Hence the rescaled process is defined as

$$n^{-1}\Xi_{t_n,n}^{p_n} = \{(n^{-1}l_i, n^{-1}\tau_i, n^{-1}\mathcal{L}_i, n^{-1}\mathcal{R}_i) : (l_i, \tau_i, \mathcal{L}_i, \mathcal{R}_i) \in \Xi_{t_n,n}^{p_n}\} \tag{3.10}$$

The asymptotic properties of the rescaled p-sampled historical process are now easily established from our earlier results.

Theorem 11. *For any* $\{t_n > 0\}_{n\geq 1}$*, and* $\{p_n \in (0,1)\}_{n\geq 1}$ *such that* $t_n/n \underset{n\to\infty}{\to} \mathrm{t}$*, and* $np_n \underset{n\to\infty}{\to} \mathrm{p}$ *we have* $n^{-1}\Xi^{p_n}_{t_n,n} \underset{n\to\infty}{\Longrightarrow} \xi^{\mathrm{p}}_{\mathrm{t},1}$.

Proof. By Theorem 5 we already have that $n^{-1}\mathit{\Pi}_{t_n,n} \underset{n\to\infty}{\Longrightarrow} \pi_{\mathrm{t},1}$. Applying the rescaling to the results of Lemma 8 together with the result of Lemma 10 now implies that the support set and intensity measure of the Poisson point-process of each $\mathcal{L}_i$ after rescaling converges to those of the Poisson point-process $\mathcal{L}_\ell$ as given by Lemma 9. Then the convergence of the support set and intensity measure for the Poisson random measure $\Xi^{p_n}_{t_n,n}$ to those of $\xi^{\mathrm{p}}_{\mathrm{t},1}$ implies the weak convergence of these processes. □

Remark. The randomization of the time of origin (Section 2.4) can be used to extend the results above into Bayesian asymptotic results as well.

Chapter 4

Genealogy of Higher Order Taxa

In this chapter we define a model on higher order taxa, in such a way that it incorporates the model on species within it. We analyze this model from the perspective of a typical species, and the genus containing it. We determine the distribution of the lifetime of this genus (Lemma 12), and the distribution of the number of extant species contained in it (Lemma 14). We also determine the shape of the tree on genera in terms of the probabilities of branching points with different split type appearing within the tree (Lemma 16 and Theorem 17).

4.1 Model on Genera

Let $\gamma \geq 0$ be the parameter of a Poisson process on the edges of the critical branching tree $\mathcal{T}$. The times of this Poisson process represent the occurrences within the lifetime of a species of a change that makes it and its progeny significantly different enough so as to generate a new higher-order taxon. This random process of changes is superimposed on the critical branching process model on the evolution of species. We indicate these changes by putting a cross mark on the place on the edge of the family tree at the time of its occurrence.

For concreteness we shall talk about a model on genera to represent this model on higher order taxa (genera being the next level above species), although we can employ this model at an arbitrary higher order level. The next issue in the above model is deciding, given these changes, how to partition the species into genera. This is really a deterministic issue rather then a stochastic one, with several different ways of resolving it depending on how wide a range within a group of species one is willing to allow a genus include. In other words, there are varying degrees of coarseness that a division into genera within a tree on species can have. Several mathematically reasonable solutions are discussed in [5].

The starting point for all the divisions, is that the marks for changes partition the family of species into classes, where two species are in the same class if the path in the family tree between them contains no mark for such a change. The coarsest of all models is one in which we define each genus to consist of one such class. However, these classes are not necessarily clades, where a *clade* is a set of species consisting of all the descendents of some ancestral species (in biology it is called a *monophyletic group*). Mathematically, a clade in a tree is the complete subtree of all the edges that have a particular common edge on their paths back to the root. For example, in Figure 4.1 the partition $\{5, 6, 7\}$ forms a clade while the partition $\{1, 2, 3, 4, 5\}$ doesn't. In biology, the emphasis is generally on classifications in which higher order taxa are monophyletic groups. An argument for such a choice is that one would like to avoid having the situation occurring in non-monophyletic classifications where: for some species $a_1, a_2 \in A$, and $b_1 \in B$ in two genera, we have that a_1 is more closely related to b_1 in a different genus then to a_2 in its own.

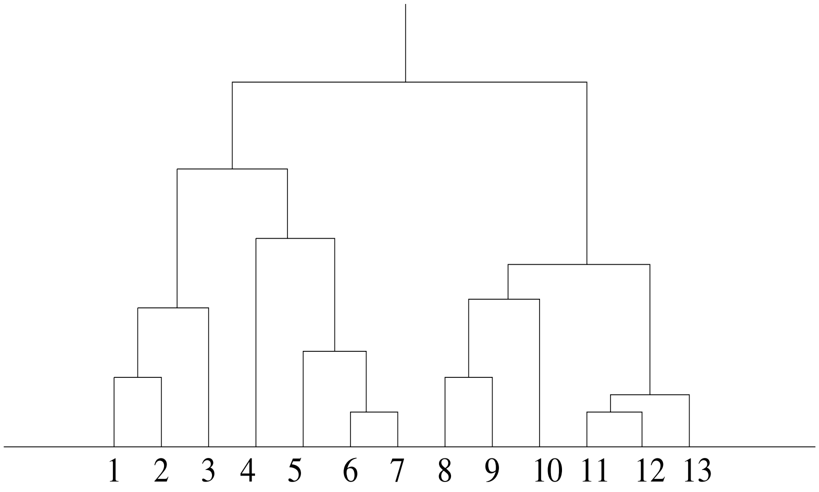

Figure 4.1: A tree (cladogram) on the extant species.

Hence, we impose in our model the additional requirement that the group of species making a genus should be a clade. Given the class partition of species according to the marks for changes, we define a partition into genera to be the coarsest one in which each genus is a clade and each class is a union of one or more genera. Figure 4.2 gives an example of a partition of a tree on species into genera that are monophyletic. Consequently, we have the following rule for constructing genera from a tree on species: two species are in the same genus if on the path in the family tree between them there is no mark for a change or a lineage arising that has a mark for a change in it. Within the tree on species the founders of the genera are: the species that have a mark for change during their lifetime and all of their ancestors.

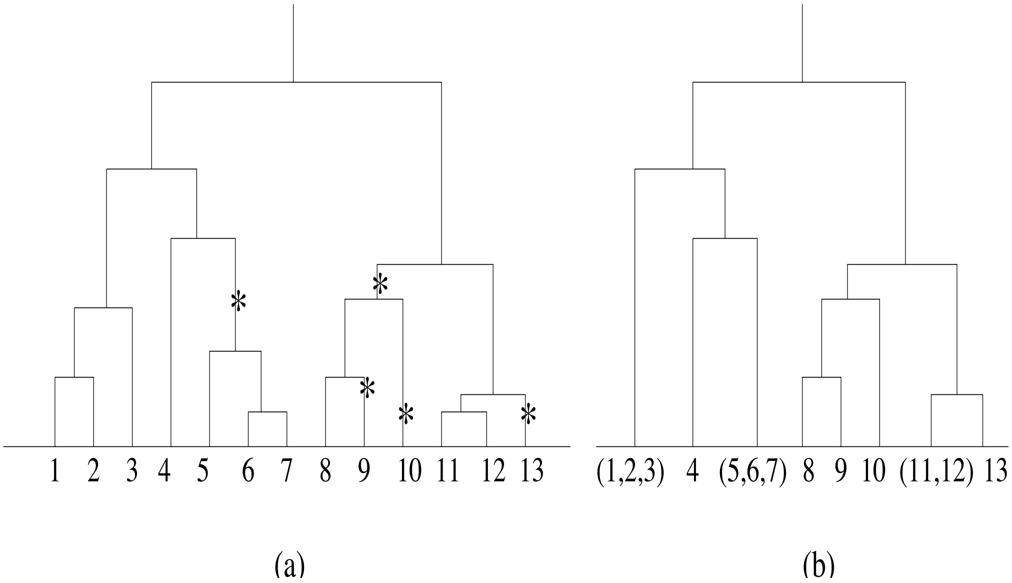

Figure 4.2: (a) The tree on species with marks for change represented by $*$'s; (b) The induced tree on genera.

4.2 Local Structure of the Model

Consider our Bayesian model on species from the point of view of a typical extant species. Our earlier construction provides the genealogy of all extant species with a genealogical point-process. For a large number of extant species, according to (2.2) and (2.12), the law of this point-process and its time of origin is approximately given by

$$\nu_{\mathrm{t}}(\{i\} \times d\tau) = \frac{1}{2} \frac{d\tau}{(1+\tau)^2} \frac{1+\mathrm{t}}{\mathrm{t}} \frac{1}{\mathrm{t}^2} \exp(-\frac{1}{\mathrm{t}}), \;\; \tau \in (0, \mathrm{t}), \, \mathrm{t} > 0$$

for any $i \in \mathbb{N}, i < n$. Let $\nu(\{i\} \times d\tau)$ be the marginal law of this process, integrating out its time of origin. Then, for any $i \in \mathbb{N}, i < n$

$$\nu(\{i\} \times d\tau) = \frac{d\tau}{(1+\tau)^2}, \;\; \tau > 0$$

This gives us a simple description of the local structure of the genealogy of extant species, where by local we mean within order 1 of time from the present. At present there are a large number of extant species. As τ increases (time runs backwards), the lineage of a typical species merges with the lineages of other species according to the density of points in the genealogical point-process

$$f_{\mathrm{s}}(\tau)d\tau = \frac{d\tau}{(1+\tau)^2}, \;\; \tau > 0.$$

Hence, the survival function of a lineage of a typical species is

$$\overline{F}_{\mathrm{s}}(\tau) = \int\limits_{\tau}^{\infty} f_{\mathrm{s}}(s)ds = \frac{1}{1+\tau}, \;\; \tau > 0.$$

and the lineages merge at a rate

$$r_{\mathrm{s}}(\tau)d\tau = \frac{f_{\mathrm{s}}(\tau)d\tau}{\overline{F}_{\mathrm{s}}(\tau)} = \frac{d\tau}{1+\tau}, \;\; \tau > 0.$$

In order to obtain a description of the local structure of the model on genera, we focus on the extant genus that a typical extant species belongs to. To do so, we consider the effect that the occurrences of the marks for change have on the local structure of the genealogy of an extant species. Note that, as time runs backwards, exactly three possible types of events occur within a lineage:

(1) a mark for change occurs on this lineage,

(2) a merge occurs with a lineage which already has a mark for change in it,

(3) a merge occurs with a lineage which yet contains no marks for change.

Figure 4.3 shows the three different possibilities for the lineage of some extant species in a small time interval $(\tau, \tau + d\tau)$.

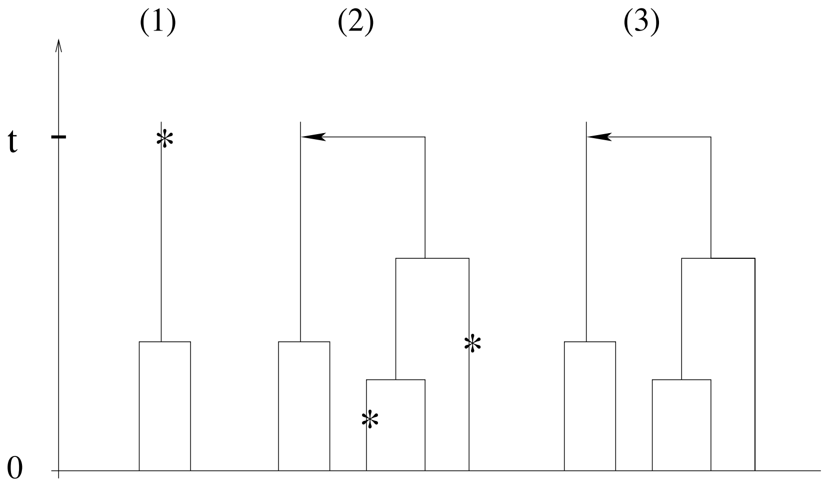

Figure 4.3: The three different types of events (1), (2), and (3) at time t back from the present.

4.3 Lifetime of a Genus

We first consider the time at which the genus containing a typical extant species emerged as a separate genus. According to our definition of a genus which requires it to be a complete clade, an extant species is separated into a new genus the first time, back from the present, that an event of type (1) or of type (2) occurs.

Definition. Let τ_{g} be the lifetime of a genus containing a typical extant species, measured from the present back to its emergence.

The following Lemma establishes the law of the lifetime τ_{g} by considering the possible effects the above three events have on the occurrence of a new genus within a small interval of time.

Lemma 12. *In the model on genera in which species change at rate γ, and a genus is the largest collection of changed species that is a complete clade, the lifetime τ_{g} of a genus of a typical extant species has the law given by*

$$\overline{F}_{\mathrm{g}}(\tau) = \frac{e^{-\gamma\tau}(1+\tau)^{-1}}{1 - \int\limits_0^\tau e^{-\gamma u}(1+u)^{-2}du}, \ \tau > 0 \tag{4.1}$$

where $\overline{F}_{\mathrm{g}}(\tau) = \mathbf{P}[\tau_{\mathrm{g}} > \tau]$.

Proof. The proof is a classical applied probabilistic argument. We consider the effect in an arbitrary small interval of time $(\tau, \tau + d\tau)$ that the three possible events have on the probability of a start of a new genus.

Given that a new genus has not occurred by time τ, the probability of it occurring in $(\tau, \tau + d\tau)$ is precisely the probability of events of type (1) or of type (2) above.

Since the rate of the marks of change along any lineage is γ, the probability of the type(1) event is $\mathbf{P}[(1) \in (\tau, \tau + d\tau)] = \gamma d\tau$.

The probability of a lineage merging with any other lineage in $(\tau, \tau + d\tau)$ is given by the rate in the local structure on species $r_{\mathrm{s}}(\tau) = 1/(1 + \tau)$, so $\mathbf{P}[(2) \cup (3) \in (\tau, \tau + d\tau)] = r_{\mathrm{s}}(\tau)$.

The lineage it merges with already has a mark for change in it, if and only if the occurrence of a new genus has already happened in that lineage by time τ back from the present. Locally the lineages of different extant species evolve independently and with identical law, hence the occurrence of a new genus in the other lineage by time τ is equal to $\mathbf{P}[\tau_{\mathrm{g}} \leq \tau]$ as well. Hence, $\mathbf{P}[(2) \in (\tau, \tau + d\tau)] = r_{\mathrm{s}}(\tau)\mathbf{P}[\tau_{\mathrm{g}} \leq \tau]$.

Now, let $F_{\mathrm{g}}(\tau) = \mathbf{P}[\tau_{\mathrm{g}} \leq \tau]$, for $\tau \geq 0$. Note that the independence of the process of changes and the process of merges implies that the occurrence of both type(1) and type(2) events in $(\tau, \tau + d\tau)$ has a negligible probability. Also, the Markov property of the tree implies the independence of events in $(\tau, \tau + d\tau)$ from those in $[0, \tau]$. Hence, we have that

$$\mathbf{P}[\tau_{\mathrm{g}} \in (\tau, \tau + d\tau)] = \mathbf{P}[\tau_{\mathrm{g}} > \tau] \Big(\gamma d\tau + r_{\mathrm{s}}(\tau) d\tau \mathbf{P}[\tau_{\mathrm{g}} > \tau]\Big),$$

in other words,

$$F_{\mathrm{g}}(\tau + d\tau) - F_{\mathrm{g}}(\tau) = \Big(1 - F_{\mathrm{g}}(\tau)\Big) \Big(\gamma d\tau + r_{\mathrm{s}}(\tau) F_{\mathrm{g}}(\tau) d\tau\Big).$$

Now deriving the law of τ_{g} reduces to solving a differential equation for its cumulative distribution function.

Substituting in $r_{\mathrm{s}}(\tau) = 1/(1 + \tau)$, we obtain that the cumulative distribution function is the solution of the non-linear differential equation

$$\frac{dF_{\mathrm{g}}}{d\tau} = \big(1 - F_{\mathrm{g}}\big)\big(\gamma + \frac{F_{\mathrm{g}}}{1 + \tau}\big), \quad \tau > 0, \tag{4.2}$$

that also satisfies $F_{\mathrm{g}}(0) = 0,\ dF_{\mathrm{g}}/d\tau \geq 0,\ \lim_{\tau\uparrow\infty} F_{\mathrm{g}}(\tau) = 1$.

We first make a transformation of the equation by introducing the function $y = -\log\big(1 - F_{\mathrm{g}}\big)$, so that $F_{\mathrm{g}} = 1 - e^{-y}$, and $dy/dF_{\mathrm{g}} = e^{y} = 1/(1 - F_{\mathrm{g}})$. In terms of the function y the equation becomes

$$\begin{aligned}\frac{dy}{d\tau} &= \frac{1}{(1-F_{\mathrm{g}})}\,\big(1-F_{\mathrm{g}}\big)\big(\gamma + \frac{F_{\mathrm{g}}}{1+\tau}\big)\\ &= \gamma + \frac{F_{\mathrm{g}}}{1+\tau} = \gamma + \frac{1-e^{-y}}{1+\tau}.\end{aligned}$$

We also transform the time parameter by introducing $s = \log\big(1+\tau\big)$, so that $\tau = e^{s} - 1$, and $d\tau/ds = 1 + \tau$. In terms of the parameter s the equation then becomes

$$\begin{aligned}\frac{dy}{ds} &= \big(\gamma + \frac{1-e^{-y}}{1+\tau}\big)\ (1+\tau)\\ &= \gamma(1+\tau) + (1-e^{-y}) = \gamma e^{s} + (1-e^{-y}).\end{aligned}$$

Hence we need the solution of the differential equation

$$\frac{dy}{ds} = \gamma e^{s} + (1-e^{-y}),\quad s > 0,$$

which satisfies the equivalent properties $y(0),\ dy/ds \geq 0,\ \lim_{s\uparrow\infty} y(s) = \infty$.

We now proceed as follows. We first obtain an explicit solution in y of this equation using only the initial condition $y(0) = 0$. We then further show that this solution indeed satisfies the remaining two conditions as well.

First, in order to obtain a solution, we let $z = e^{y}$ so that then $dz/dy = e^{y}$. In terms of z we thus obtain a linear non-homogeneous differential equation

$$\frac{dz}{ds} = e^{y}\ \big(\gamma e^{s} + 1 - e^{-y}\big) = z\big(\gamma e^{s} + 1\big) - 1,\ s > 0$$

which can now be solved using integrating factors.
Namely, multiplying both sides of the last equation by the factor $e^{\int_0^s (\gamma e^{r}+1)dr} = e^{-\gamma e^{s} - s}$, we obtain the equation

$$\frac{d\big(ze^{-\gamma e^{s}-s}\big)}{ds} = -e^{\gamma e^{s}+s},$$

so that

$$z(s) = e^{\gamma e^{s}+s}\ \big(c_0 - \int_0^s e^{-\gamma e^{r}+r} dr\big),\quad s \geq 0.$$

The constant c_0 is determined from the initial condition, $z(0) = e^{y(0)} = 1$, to be $c_0 = e^{-\gamma}$. In terms of the original function $y = \log(z)$ we have

$$y(s) = \big(\gamma e^s + s\big) + \log\big(e^{-\gamma} - \int_0^s e^{-\gamma e^r + r} dr\big), \quad s \geq 0.$$

Second, we show that this solution satisfies $dy/ds \geq 0$ and $\lim_{s\uparrow\infty} y(s) = \infty$. Let $I(s) = \int_0^s e^{-\gamma e^r - r} dr$, so that $y(s) = (\gamma e^s + s) + \log\big(e^{-\gamma} - I(s)\big)$. We next show that $y(s) \geq 0,\ \forall s \geq 0$. It is clear that $I(s) \geq 0,\ dI/ds \geq 0$, and moreover that

$$I(s) = \int_0^s e^{-\gamma e^r + r} dr \leq e^{-\gamma} \int_0^s e^{-r} dr = e^{-\gamma}(1 - e^{-s}).$$

Hence, $e^{-\gamma} - I(s) \geq e^{-\gamma - s}$, and thus $\log\big(e^{-\gamma} - I(s)\big) \geq -\gamma - s$. So then,

$y(s) = (\gamma e^s + s) + \log\big(e^{-\gamma} - I(s)\big) \geq (\gamma e^s + s) - \gamma - s = \gamma(e^s - 1)$, which certainly implies $y(s) \geq 0,\ \forall s \geq 0$.

We now use this fact to show that $dy/ds \geq 0,\ \forall s > 0$. Simply considering the differential equation for y we observe that, because $y \geq 0$,

$dy/ds = \gamma e^s + 1 - e^{-y} \geq \gamma e^s$, which certainly implies $dy/ds \geq 0,\ \forall s \geq 0$.

Finally we show that $\lim_{s\uparrow\infty} y(s) = \infty$. This follows from an earlier result that $y(s) = (\gamma e^s + s) + \log\big(e^{-\gamma} - I(s)\big) \geq \gamma(e^s - 1)$, so clearly $y(s) \uparrow \infty,$ as $s \uparrow \infty$.

Expressing this result in terms of F_{g} and parameter τ this gives us the solution for the distribution of τ_{g}:

$$\begin{aligned} F_{\mathrm{g}}(\tau) &= 1 - e^{-y(\log(1+\tau))} \\ &= 1 - \frac{e^{-\gamma(1+\tau)+\log(1+\tau)}}{e^{-\gamma} - \int_0^{\log(1+\tau)} e^{-\gamma e^r - r} dr} \\ &= 1 - \frac{e^{-\gamma\tau}(1+\tau)^{-1}}{1 - \int_0^\tau e^{-\gamma u}(1+u)^{-2} du} \ \tau \geq 0, \end{aligned}$$

from which we have the formula for the survival function $\overline{F}_{\mathrm{g}}(\tau) = 1 - F_{\mathrm{g}}(\tau)$ of τ_{g} as claimed. □

We can now make some easy estimates for the expected value of τ_{g}.

Theorem 13. *The expected lifetime of a genus of typical species, in the model in which species change with rate γ, satisfies*

$$\frac{1}{2\gamma} \leq \mathbf{E}[\tau_{\text{g}}] \leq \frac{1}{\gamma}. \tag{4.3}$$

Proof. The proof is a simple consequence of the formula for the law of τ_{g} from the above Lemma.

The inequality $1 \leq e^{-\gamma u} \leq e^{-\gamma\tau}$ for $0 \leq u \leq \tau$, yields

$$\int_0^\tau (1+u)^{-2}du \geq \int_0^\tau e^{-\gamma u}(1+u)^{-2}du \geq e^{-\gamma\tau}\int_0^\tau (1+u)^{-2}du,$$

so

$$(1+\tau)^{-1} \leq 1 - \int_0^\tau e^{-\gamma u}(1+u)^{-2}du \leq (1+\tau)^{-1}(1+\tau-\tau e^{-\gamma\tau}),$$

hence

$$e^{-\gamma\tau} \geq \frac{e^{-\gamma\tau}(1+\tau)^{-1}}{1-\int_0^\tau e^{-\gamma u}(1+u)^{-2}du} \geq e^{-\gamma\tau}(1+\tau-\tau e^{-\gamma\tau})^{-1},$$

and since $1+\tau-\tau e^{-\gamma\tau} \leq e^{\gamma\tau}$, finally

$$1 - e^{-\gamma\tau} \leq F_{\text{g}}(\tau) \leq 1 - e^{-2\gamma\tau} \tag{4.4}$$

and the claimed estimates of the expected value immediately follow. □

The bounds we obtained have an intuitively obvious explanation. Suppose we had a process in which the marks for change occurred only along the lineage of the chosen typical species. Then its genus emerges at the time of the first mark, and its lifetime has an Exponential(rate γ) distribution. In light of all the other possibilities contributing to the emergence of the genus in the original process , the lifetime of a genus is clearly stochastically dominates this Exponential(rate γ) distribution. On the other hand, consider a lineage that merges with that of the chosen species. If this lineage has a mark for change at some time then this will cause the emergence of the genus for our chosen species. Suppose we take then the minimum of the times of the first mark in this neighboring lineage and of the first mark of the lineage of the chosen species, which has an Exponential(rate 2γ) distribution. Since the emergence in the event of the mark on the neighboring lineage does not occur until the actual merger, it is follows that the lifetime of a genus is dominated by this Exponential(rate 2γ) distribution.

The implications on τ_{g} from the above bounds on F_{g} for the extreme values of γ $(= 0, \uparrow\infty)$, have clear interpretations as well. For $\gamma = 0$ we have that $F_{\text{g}}(\tau) = 0\ \forall\tau$, in other words $\tau_{\text{g}} = \infty$ a.s., corresponding to the fact that there are no marks for change. For $\gamma \uparrow \infty$ we have that $F_{\text{g}}(\tau) \uparrow 1\ \forall\tau$, in other words $\tau_{\text{g}} = 0$, which corresponds to the fact that a mark for change occurs a.s. immediately after $\tau = 0$. We shall make use of the behavior of this genus for the extreme values of γ in the following Section.

4.4 Size of a Genus

We next consider how the genera partition the set of extant species. We shall hence consider the present size of the clade that contains a typical extant species. In other words, we shall consider the number of extant species that a genus of an arbitrary extant species contains. According to the local structure, the extant species in this genus are those whose lineages merge with the chosen extant species, prior to any type (1) or type (2) event. Thus as time runs runs back from the present, this gives us a partial size of the genus which increases at any time that an event of type (3) occurs. The complete size of this genus is given at time τ_{g}.

Definition. Let $\mathcal{N}_\tau$, for $0 \leq \tau \leq \tau_{\text{g}}$, be the partial size of the genus of a typical extant species, by time τ back from the present, and let $\mathcal{N} = \mathcal{N}_{\tau_{\text{g}}}$ be the complete size of this genus. Size here means the number of extant species in this genus.

The following Lemma establishes the law of the process $(\mathcal{N}_\tau)_{0\leq\tau\leq\tau_{\text{g}}}$ in terms of a set of recursive equations it satisfies.

Lemma 14. *The partial size $\mathcal{N}_\tau$ by time τ of a genus of a typical extant species, given that the genus has not emerged yet has the conditional distribution $f_{\mathcal{N}_\tau}$ defined by the set of recursive equations*

$$f_{\mathcal{N}_\tau}(k) = \frac{1}{(1+\tau)e^{\gamma\tau}\overline{F}_{\text{g}}(\tau)} \int\limits_0^\tau e^{\gamma s}\overline{F}_{\text{g}}(s) \sum_{i=1}^{k-1} f_{\mathcal{N}_s}(i) f_{\mathcal{N}_s}(k-i)ds \tag{4.5}$$

for $k \geq 2$, with

$$f_{\mathcal{N}_\tau}(1) = \frac{1}{(1+\tau)e^{\gamma\tau}\overline{F}_{\mathrm{g}}(\tau)}. \tag{4.6}$$

The complete size $\mathcal{N}$ of this genus then has the law $f_{\mathcal{N}}$ given by

$$f_{\mathcal{N}}(k) = \int_0^\infty \; f_{\mathcal{N}}(\tau)\, dF_{\mathrm{g}}(\tau), \;\; k \geq 1 \tag{4.7}$$

where $\overline{F}_{\mathrm{g}} = 1 - F_{\mathrm{g}}$ is as in (4.1) of Lemma 12.

Proof. The proof once more relies on considering the effect that the three possible events have on the partial size of this genus in an arbitrary small interval of time $(\tau, \tau + d\tau)$.

On the event that emergence of the genus has not yet occurred by time τ, the probability that within this event the partial size increases is precisely the probability of an event of type (3) above. The probability that the partial size remains the same within this event is just the probability that none of the type (1), (2), or (3) events happen.

Let $g(k,\tau) = \mathbf{P}[\mathcal{N}_\tau = k, \tau_{\mathrm{g}} > \tau]$, for $k \geq 1, \tau \geq 0$. Then, since the Markov property of the model again implies the independence of the event in $(\tau, \tau + d\tau)$ from the those in $(0, \tau]$, we have that

$$g(k, \tau + d\tau) = \sum_{i=1}^{k-1} g(i,\tau) p_\tau(k-i) d\tau + g(k,\tau) p_\tau(0) d\tau$$

where we have used the notation $p_\tau(j)$ to mean, for all $j \geq 1$

$p_\tau(j) d\tau = \mathbf{P}[$of a type (3) event with a lineage of size $k - i \in (\tau, \tau + d\tau)]$,

$p_\tau(0) d\tau = 1 - \mathbf{P}[$any event of type (1), (2), or (3) $\in (\tau, \tau + d\tau)]$.

Now, it is clear that $p_\tau(0) d\tau = \big(1 - \gamma - r_{\mathrm{s}}(\tau)\big) d\tau$. On the other hand, for $j \geq 1$ we have that $p_\tau(j) d\tau = r_{\mathrm{s}}(\tau) g(j,\tau) d\tau$, since at a time of a merger in $(\tau, \tau + d\tau)$ the chance that the merging lineage has partial size j and has not had its genus emergence yet is precisely $g(j,\tau)$. This establishes the following recursive relationship for the family of functions $\{g(k,\tau)\}$, over $k \geq 1$, $0 \leq \tau < \tau_{\mathrm{g}}$, as

$$g(k, \tau + d\tau) = \sum_{i=1}^{k-1} r_{\mathrm{s}}(\tau)\, g(i,\tau) g(k-i,\tau)\, d\tau + \big(1 - \gamma - r_{\mathrm{s}}(\tau)\big)\, g(k,\tau)\, d\tau.$$

In other words,

$$\frac{dg(k,\tau)}{d\tau} = \sum_{i=1}^{k-1} r_{\rm s}(\tau)\, g(i,\tau) g(k-i,\tau) - \big(\gamma + r_{\rm s}(\tau)\big)\, g(k,\tau).$$

Note that the initial values are dictated by the one chosen extant species to be $g(1,0) = 1,$ and $g(k,0) = 0$ for $k \geq 1$.

For $k = 1$ the above equation is just

$$\frac{dg(1,\tau)}{d\tau} = -\big(\gamma + r_{\rm s}(\tau)\big)\, g(1,\tau),$$

hence $g(1,\tau) = c_1\, e^{-\int_0^\tau (\gamma + r_{\rm s}(s))ds}$.

Since $r_{\rm s}(\tau) = 1/(1+\tau)$, and $g(1,0) = 1,$ it follows that

$$g(1,\tau) = \frac{e^{-\gamma\tau}}{1+\tau}, \quad 0 \leq \tau < \tau_{\rm g}. \tag{4.8}$$

For $k \geq 1$ using the integrating factor $e^{\int_0^\tau (\gamma + r_{\rm s}(s))ds}$ we obtain

$$\frac{d\big(g(k,\tau) e^{\int_0^\tau (\gamma + r_{\rm s}(s))ds}\big)}{d\tau} = e^{\int_0^\tau (\gamma + r_{\rm s}(s))ds} \sum_{i=1}^{k-1} r_{\rm s}(\tau)\, g(i,\tau) g(k-i,\tau).$$

Since $r_{\rm s}(\tau) = 1/(1+\tau)$, we have that $e^{\int_0^\tau (\gamma + r_{\rm s}(s))ds} = e^{\gamma\tau}(1+\tau),$ and also since $g(k,0) = 0$ for $k \geq 1,$ it follows that

$$g(k,\tau) = \frac{e^{-\gamma\tau}}{1+\tau} \int\limits_0^\tau \sum_{i=1}^{k-1} e^{\gamma s}\, g(i,s) g(k-i,s) ds, \quad 0 \leq \tau < \tau_{\rm g}. \tag{4.9}$$

Now, let $f_{\mathcal{N}_\tau}$ represent the conditional distribution of the partial size $\mathcal{N}_\tau$, given the event $\{\tau_{\rm g} \geq \tau\}$. Then,

$$f_{\mathcal{N}_\tau}(k) = \frac{g(k,\tau)}{\overline{F}_{\rm g}(\tau)},$$

where $\overline{F}(\tau)$ is given by (4.1) from Lemma 12, which immediately yields the claimed recursive definition of the distribution $\{f_{\mathcal{N}_\tau}(k),\, k \geq 1\}$.

To obtain the law of the complete size of the genus, we can now use the conditional distribution of the size $\mathcal{N}_\tau$ as $\tau \uparrow \tau_{\rm g}$, and integrate over the possible values of $\tau_{\rm g}$.

Let $f_{\mathcal{N}}$ represent the distribution of $\mathcal{N}$. Now, at the emergence of the genus there is no size increase, and also we have the Markov property for the tree, hence

$$\begin{aligned}
\mathbf{P}[\mathcal{N}=k,\tau_{\mathrm{g}}=\tau] &= \mathbf{P}[\mathcal{N}_{\tau^-}=k,\tau_{\mathrm{g}}=\tau] \\
&= \mathbf{P}[\mathcal{N}_{\tau^-}=k,\tau_{\mathrm{g}}>\tau^-,\tau_{\mathrm{g}}\in(\tau^-,\tau)] \\
&= \mathbf{P}[\mathcal{N}_{\tau^-}=k,\tau_{\mathrm{g}}>\tau^-]\mathbf{P}[\tau_{\mathrm{g}}\in(\tau^-,\tau)] \\
&= \mathbf{P}[\mathcal{N}_{\tau^-}=k,\tau_{\mathrm{g}}>\tau^-]\;\mathbf{P}[\text{type}\,(1)\,\text{or}\,(2)\,\text{event}\in(\tau^-,\tau)] \\
&= g(k,\tau^-)\;\mathbf{P}[\text{type}\,(1)\,\text{or}\,(2)\,\text{event}\in(\tau^-,\tau)]
\end{aligned}$$

Next note that, for each $k\geq 1$, $g(k,\tau)$ is a continuous function in τ. Also, from Lemma 12 we have $\mathbf{P}[\text{type}\,(1)\,\text{or}\,(2)\,\text{event}\in(\tau^-,\tau)]=dF_{\mathrm{g}}(\tau^-)/\overline{F}_{\mathrm{g}}(\tau^-)$, with $F_{\mathrm{g}}(\tau)$ also continuous in τ, so

$$\mathbf{P}[\mathcal{N}=k,\tau_{\mathrm{g}}=\tau]=\frac{g(k,\tau)}{\overline{F}_{\mathrm{g}}(\tau)}\,dF_{\mathrm{g}}(\tau).$$

Hence,

$$f_{\mathcal{N}}(k)=\mathbf{P}[\mathcal{N}=k]=\int_0^\infty\frac{g(k,\tau)}{\overline{F}_{\mathrm{g}}(\tau)}\,dF_{\mathrm{g}}(\tau)=\int_0^\infty f_{\mathcal{N}_\tau}(k)dF_{\mathrm{g}}(\tau),\tag{4.10}$$

as claimed. □

We can now give a bound on the expected number of species per genus.

Theorem 15. *The expected size of the genus of a typical extant species, in the model in which species change with rate γ, satisfies*

$$1\leq\mathbb{E}[\mathcal{N}]\leq 1+\frac{2}{\gamma}\tag{4.11}$$

Proof. The proof relies on a comparison of the distribution of the partial size of a genus for an arbitrary γ, with that of the same distribution for the particular case when $\gamma=0$.

We start by identifying the distribution of $\mathcal{N}_\tau$, the partial size of a genus at time τ from the present, in the case when the rate $\gamma=0$, so that there are no marks for change. The probabilities for the partial genus size $\mathcal{N}_\tau$ when $\{\tau_{\mathrm{g}}>\tau\}$, are given by (4.8) and a set of recursive equations (4.9). When $\gamma = 0$, these are just

$$\begin{aligned} g(1,\tau) &= \frac{1}{1+\tau}, \\ \text{and}\quad g(k,\tau) &= \frac{1}{1+\tau}\int\limits_0^\tau \sum_{i=1}^{k-1} g(i,s)g(k-i,s)ds,\ k \geq 2. \end{aligned}$$

holding for all $\tau \geq 0$, since $\tau_{\mathrm{g}} = \infty$ a.s. when $\gamma = 0$.

We now observe that

$$g(k,\tau) = \frac{1}{1+\tau}\Big(\frac{\tau}{1+\tau}\Big)^{k-1},\quad k \geq 1. \tag{4.12}$$

are the explicit solutions of the above equations.

This is clearly the case for $k = 1$, and assuming this is the case for all $i \leq k-1$, we have that

$$\begin{aligned} g(k,\tau) &= \frac{1}{1+\tau}\int_0^\tau \sum_{i=1}^{k-1} (\frac{1}{1+s})^2(\frac{s}{1+s})^{i-1+k-i-1}\,ds \\ &= \frac{1}{1+\tau}\int_0^\tau (k-1)(\frac{1}{1+s})^2(1-\frac{1}{1+s})^{k-2}\,ds \\ &= \frac{1}{1+\tau}\Big(1-\frac{1}{1+\tau}\Big)^{k-1} = \frac{1}{1+\tau}\Big(\frac{\tau}{1+\tau}\Big)^{k-1} \end{aligned}$$

holds for k as well.

In other words, when there are no marks for change the number of extant species whose lineages have merged with a randomly chosen one by time τ from the present follows a Geometric($\frac{1}{1+\tau}$) distribution.

We next claim that for an arbitrary $\gamma \geq 0$, the probabilities for the partial genus size $\mathcal{N}_\tau$ when $\tau_{\mathrm{g}} > \tau$, satisfy

$$g(k,\tau) \leq e^{-\gamma\tau}\,\frac{1}{1+\tau}\Big(\frac{\tau}{1+\tau}\Big)^{k-1},\quad k \geq 1.$$

Namely, the partial size of a genus on $\tau_{\mathrm{g}} > \tau$ is due to two types of events, the non-occurrence of the marks for change along its lineage, and the merger with other mark free lineages.

Suppose we had a process in which the marks for change occurred only along the lineage of our chosen extant species, while the other lineages were not subject to being marked. For this process, denote by $\mathcal{N}'_\tau$ the partial genus size by time τ arising from the chosen species, and by $\tau'_{\rm g}$ the time of emergence of this genus. Then we have $\mathcal{N}'_\tau = k, \tau'_{\rm g} > \tau$ if and only if the lineage of the chosen species is mark free and it has merged with $k-1$ lineages of other extant species by time τ. Since the marks for change on the lineage of the chosen species are independent of the mergers, the distribution of the number of lineages merged with the lineage of the chosen one is the same as in the process with no marks for change. In other words

$$\mathbf{P}[\mathcal{N}'_\tau = k, \tau'_{\rm g} > \tau] = e^{-\gamma\tau}\,\frac{1}{1+\tau}\Big(\frac{\tau}{1+\tau}\Big)^{k-1}$$

where the first term is simply the probability of no marks on the chosen lineage by time τ, and the second is the probability of genus size in a mark free process given by (4.12).

In the original process, the possibilities of marks along lineages of other extant species reduce the possibilities of merger with mark free lineages. Compared to the process analyzed above, the rate of mergers with mark free lineages before the time of occurrence of this genus is reduced, and we have that

$$g(k,\tau) = \mathbf{P}[\mathcal{N}_\tau = k, \tau_{\rm g} > \tau] \le \mathbf{P}[\mathcal{N}'_\tau = k, \tau'_{\rm g} > \tau] = e^{-\gamma\tau}\,\frac{1}{1+\tau}\Big(\frac{\tau}{1+\tau}\Big)^{k-1}$$

as claimed.

We can now use this comparison of the probabilities $g(k,\tau)$ for arbitrary γ with those for the case $\gamma = 0$ to establish the bounds for the expected value of $\mathcal{N}$. Namely, the distribution of $\mathcal{N}$ is given by (4.10) as

$$f_{\mathcal{N}}(k) = \int_0^\infty g(k,\tau)\,\frac{dF(\tau)}{\overline{F}(\tau)}$$

hence

$$\begin{aligned}
\mathbf{E}[\mathcal{N}] &= \sum_{k\ge1} k f_{\mathcal{N}}(k) = \sum_{k\ge1}\int_0^\infty k g(k,\tau)\,\frac{dF(\tau)}{\overline{F}(\tau)} \\
&\le \sum_{k\ge1} k\int_0^\infty e^{-\gamma\tau}\frac{1}{1+\tau}(\frac{\tau}{1+\tau})^{k-1}\,\frac{dF(\tau)}{\overline{F}(\tau)} \\
&= \int_0^\infty e^{-\gamma\tau}\Big(\sum_{k\ge1} k\frac{1}{1+\tau}(\frac{\tau}{1+\tau})^{k-1}\Big)\,\frac{dF(\tau)}{\overline{F}(\tau)} \\
&= \int_0^\infty e^{-\gamma\tau}\,(1+\tau)\,\frac{dF(\tau)}{\overline{F}(\tau)}
\end{aligned}$$

since the expected value of the Geometric($\frac{1}{1+\tau}$) distribution of the $\gamma = 0$ case is $1+\tau$.

Further, the differential equation for F_{g} given by (4.2) gives

$$\frac{dF_{\mathrm{g}}(\tau)}{\overline{F}_{\mathrm{g}}(\tau)} = \left(\gamma + \frac{F_{\mathrm{g}}(\tau)}{(1+\tau)}\right)d\tau$$

hence

$$\begin{aligned}\mathbf{E}[\mathcal{N}] &\leq \int_0^\infty e^{-\gamma\tau}(1+\tau)\big(\gamma + \frac{F_{\mathrm{g}}(\tau)}{1+\tau}\big)d\tau \\ &= \int_0^\infty \gamma e^{-\gamma\tau}(1+\tau)d\tau + \int_0^\infty e^{-\gamma\tau}F_{\mathrm{g}}(\tau)d\tau \\ &\leq \int_0^\infty \gamma e^{-\gamma\tau}(1+\tau)d\tau + \int_0^\infty e^{-\gamma\tau}d\tau\end{aligned}$$

since $\forall\tau \geq 0,\ F_{\mathrm{g}}(\tau) \leq 1$. A simple calculation of integrals now yields

$$\mathbf{E}[\mathcal{N}] \leq 1 + \frac{1}{\gamma} + \frac{1}{\gamma} = 1 + \frac{2}{\gamma}$$

as claimed. □

In terms of the extreme values for γ $(\gamma\!=\!0, and\gamma\!\uparrow\!\infty)$, we have the following interpretation. For $\gamma = 0$ we have seen that $\tau_{\mathrm{g}} = \infty$ a.s., and that the partial genus size $\mathcal{N}_\tau$ has a Geometric($\frac{1}{1+\tau}$) distribution. Thus, $\forall k \geq 1$

$$f_{\mathcal{N}}(k) = \mathbf{P}[\mathcal{N} = k] = \lim_{\tau\uparrow\infty} \mathbf{P}[\mathcal{N}_\tau = k, \tau < \infty] = \lim_{\tau\uparrow\infty} \tfrac{1}{1+\tau}\big(1 - \tfrac{1}{1+\tau}\big)^{k-1} = 0$$

In other words $\mathcal{N} = \infty$ a.s., corresponding to the fact that when there are no marks for change all the extant species are in the same genus. On the other hand, for $\gamma\uparrow\infty$ we have that $\tau_{\mathrm{g}} = 0$ a.s., and

$$g(k,\tau) = \mathbf{P}[\mathcal{N}_\tau = k, \tau < \tau_{\mathrm{g}}] = 0,\ \forall\tau \geq 0, k \geq 1,\ \text{ except for } g(1,0) = 1$$

Thus $\mathcal{N} = 1$ a.s., corresponding to the fact that a mark for change occurs a.s. immediately after $\tau = 0$.

4.5 Shape and Branching Structure of the Model

We next consider the issue of tree shape in our models on species and genera. The distribution of the model manifests itself in the shape and the amount of balance (or lack thereof) in the branching structure of the tree. To assess the shape and (im)balance of the tree we consider the probabilities of different types of branching points. The type of each branching point is defined in terms of the type of split it generates, in terms of sizes of its daughter clades. The probabilities with which the different split types appear in the tree then characterize the distribution of its shape.

The motivation for assessing the tree shape in terms of proportions of different split types, is that it is arguably mathematically preferable to assessing the tree balance in terms of a single numerical summary statistic (see ([4]) Sec §4. for a brief discussion of this issue). We here only point out the fact that, using relative proportions of different split types, it is easy to make comparisons of tree balance between trees with different numbers of leaves.

A branching point (of a binary tree) is a split of type (i, j) if its subclades have i, and j number of leaves respectively. For a general assessment of tree balance there is no need to distinguish between the left and the right subclades, and we use a convention whereby in a split of type (i, j), i denotes the size of the smaller subclade, and j the size of the larger subclade. Let $p(i, j)$ denote the probability of a branching point being of split type (i, j), then the set $\{p(i, j),\ i \geq 1, j \geq 1\}$ completely describes the branching structure hence shape of the tree.

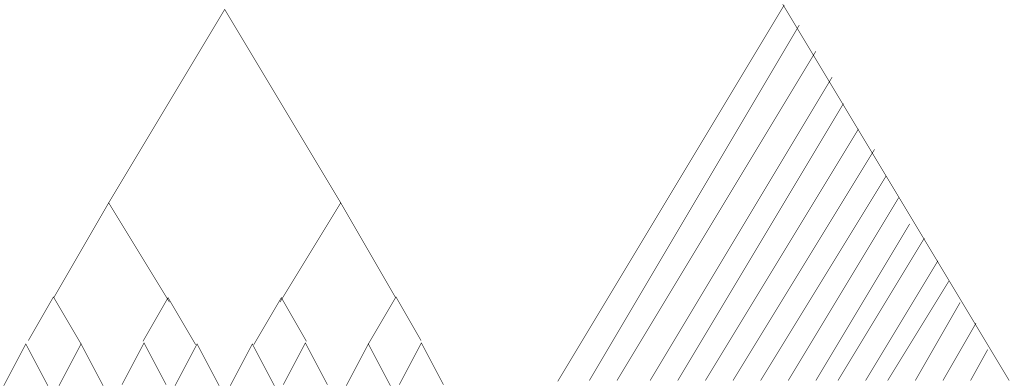

Figure 4.4: The "perfectly balanced" tree (on left), and the "comb" tree (on right).

To illustrate, consider the extreme cases of tree shapes, the "perfectly balanced" tree and the "comb" tree, as shown in Figure 4.4. In a given tree the probabilities $p(i, j)$ are naturally interpreted as relative proportions of branching points of split type (i, j). Let $p(1, 1)$ denote the proportion of splits of type $(1,1)$, $p(1,2^+)$ the proportion of splits of type $(1,j)$ over all $j \geq 2$, and $p(2^+,2^+)$ the proportion of splits of type (i,j) over all $j > i \geq 2$. Then asymptotically for the perfectly-balanced tree we have:

$p(1,1)=1/2$, $p(1,2^+)=0$, and $p(2^+,2^+)=1/2$.

On the other hand asymptotically for the comb tree we have:

$p(1,1)=0$, $p(1,2^+)=1$, and $p(2^+,2^+)=0$.

Note how these extreme values of proportions for the three groups of splits in a tree characterize the amount of (im)balance in a tree, attained in these two extreme non-random cases. We shall use them for a quick comparison with our model on genera.

For the branching structure in the model on species, we first determine the relative proportions of different lineages. At any time $\tau \geq 0$, back from the present, the mean rate of lineages at time τ (the mean number of lineages relative to the number of extant species) is given by the survival function of a lineage of a typical species $\overline{F}_{\mathrm{s}}(\tau) = 1/(1+\tau)$. Let us define the size of any lineage at time τ to be the number of extant species descending from it. Then, for any $\tau \geq 0$ we have that $\overline{F}_{\mathrm{s}}(\tau) = h_1(\tau) + h_{2^+}(\tau)$, where $h_1(\tau)$ is the mean rate at time τ of lineages whose size is exactly 1, and $h_{2^+}(\tau)$ is the mean rate at time τ of lineages whose size is greater than 1.

We consider the effect in an arbitrary small interval of time $(\tau, \tau + d\tau)$ that the merging of lineages within the model on species has on the mean rates $h_1(\tau)$, and $h_{2^+}(\tau)$. Merging of any two lineages generates another lineage of size greater than 1. The probabilities of different mergers of lineages are simply proportional to their mean rates (taking into account possible merging from both the left and the right). This easily yields that the differential equations for $h_1(\tau)$, and $h_{2^+}(\tau)$ are

$$\begin{aligned}\frac{dh_1(\tau)}{d\tau} &= -2h_1^2(\tau) - 2h_1(\tau)h_{2^+}(\tau), \;\; h_1(0) = 1,\\ \frac{dh_{2^+}(\tau)}{d\tau} &= +h_1^2(\tau) - 2h_{2^+}^2(\tau) + h_{2^+}^2(\tau) \;\; h_{2^+}(0) = 0.\end{aligned}$$

(Note that the merger of a size 1 lineage with a size greater than 1 lineage does not alter the rate of the latter, hence is absent from second equation.)

Since $\forall\tau \geq 0,\ h_1(\tau) + h_{2^+}(\tau) = \overline{F}_{\mathrm{s}}(\tau) = 1/(1+\tau),$ the above easily yield solutions

$$\begin{aligned} h_1(\tau) &= e^{-2\int_0^\tau \overline{F}_{\mathrm{s}}(s)ds} = e^{-2\log(1+\tau)} = \frac{1}{(1+\tau)^2},\ \tau \geq 0, \\ h_{2^+}(\tau) &= \overline{F}_{\mathrm{s}}(\tau) - h_1(\tau) = \frac{\tau}{1+\tau},\ \tau \geq 0. \end{aligned}$$

We can now use these two mean rates of lineages to determine the probabilities of three types of branching points according to their split types. We introduce the notation for these probabilities on the tree on species.

Definition. Let the probabilities of split types in the model on species be

$p_{\mathrm{s}}(1\ ,1\ )(\tau) = \mathbf{P}[\text{of a (1,1) split type at time } \tau],$

$p_{\mathrm{s}}(1, 2^+)(\tau) = \mathbf{P}[\text{of any } (1, j),\ j \geq 2 \text{ split type at time } \tau],$

$p_{\mathrm{s}}(2^+,2^+)(\tau) = \mathbf{P}[\text{of any } (i, j),\ i > j \geq 2 \text{ at time } \tau]$

with $p_{\mathrm{s}}(1, 1)$, $p_{\mathrm{s}}(1, 2^+)$, $p_{\mathrm{s}}(2^+, 2^+)$ the respective overall probabilities.

It is clear that the probabilities for branching at some time τ are given in terms of the mean rates of different lineage types by

$$p_{\mathrm{s}}(1\ ,1\ )(\tau) = h_1^2(\tau),\ \ p_{\mathrm{s}}(1, 2^+)(\tau) = 2h_1(\tau)h_{2^+}(\tau),\ \ p_{\mathrm{s}}(2^+,2^+)(\tau) = h_{2^+}^2(\tau)$$

so that the respective overall probabilities are

$$p_{\mathrm{s}}(1\ ,1\ ) = \int_0^\infty p_{\mathrm{s}}(1, 1)(\tau)d\tau = \int_0^\infty \frac{1}{(1+\tau)^4}d\tau = \frac{1}{3} \tag{4.13}$$

$$p_{\mathrm{s}}(1, 2^+) = \int_0^\infty p_{\mathrm{s}}(1, 2^+)(\tau)d\tau = \int_0^\infty 2\frac{1}{(1+\tau)^2}\frac{\tau}{(1+\tau)^2}d\tau = \frac{1}{3} \tag{4.14}$$

$$p_{\mathrm{s}}(2^+,2^+) = \int_0^\infty p_{\mathrm{s}}(2^+,2^+)(\tau)d\tau = \int_0^\infty \frac{\tau^2}{(1+\tau)^4}d\tau = \frac{1}{3}. \tag{4.15}$$

The above three probabilities describe the shape for the genealogical tree of a large number of extant species.

4.6 Shape of Tree on Genera

For the branching structure on the tree on genera we have to consider the rates of the following types of lineages. Note that the marks for change now create the following three types of lineages at time τ:

(a) lineages with no marks for change up to time τ,

(b) lineages with marks for change between their last merger time and time τ,

(c) all other lineages.

Note that the lineages of type (a) are precisely those within which no genus has emerged yet by τ, hence has no complete genus within it, the lineages of type (b) have exactly one emerged genus by time τ, and lineages of type (c) have at time τ already more than one genus within them. Figure 4.5 shows the different types (a), (b), and (c).

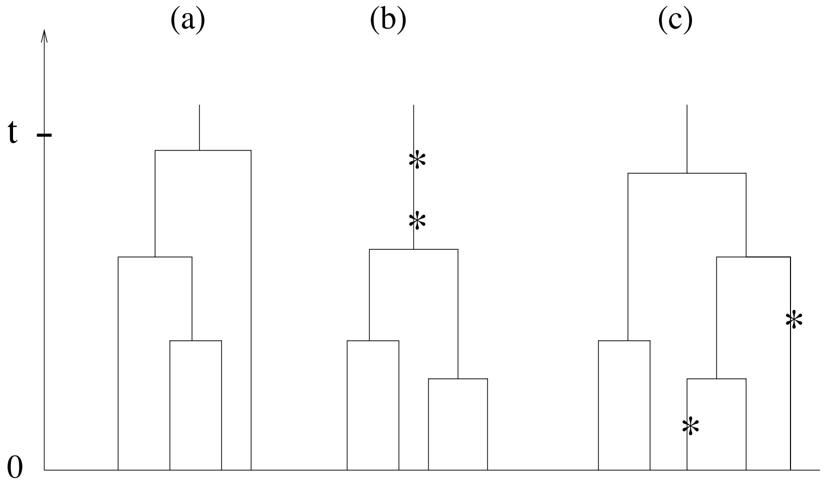

Figure 4.5: The three different types of lineages (a), (b), and (c) at time t back from the present.

We now calculate the mean rate of these types of lineages.

Lemma 16. *At any time $\tau \geq 0$ the mean rates $h_a(\tau), h_b(\tau), h_c(\tau)$ of the lineages of types (a), (b), (c) respectively are given by*

$$h_a(\tau) = (1+\tau)^{-2}\overline{F}_{\mathrm{g}}(\tau) \tag{4.16}$$

$$h_b(\tau) = (1+\tau)^{-2}\gamma \int\limits_0^\tau (1+s)\overline{F}_{\mathrm{g}}(s)ds \tag{4.17}$$

$$h_c(\tau) = (1+\tau)^{-2}\Big((1+\tau)F_{\mathrm{g}}(\tau) - \gamma \int\limits_0^\tau (1+s)\overline{F}_{\mathrm{g}}(s)ds\Big) \tag{4.18}$$

Proof. We consider the effect that the mergers of lineages of different types have on the rates $h_a(\tau), h_b(\tau), h_c(\tau)$ within a small interval of time $(\tau, \tau + d\tau)$. The different types of events in terms of the mergers of (a), (b), and (c) type lineages and of appearance of a mark for change are

- $(a) + (a) \to (a), \quad (a) + (b) \to (c), \quad (a) + (c) \to (c)$
- $(b) + (b) \to (c), \quad (b) + (c) \to (c), \quad (c) + (c) \to (c)$
- $(a) +$ mark $\to (b), \quad (b) +$ mark $\to (b), \quad (c) +$ mark $\to (c)$

This now yields the differential equations for $h_a(\tau), h_b(\tau), h_c(\tau)$ to be

$$\begin{aligned}
\frac{dh_a(\tau)}{d\tau} &= -2h_a^2(\tau) + h_a^2(\tau) - 2h_a(\tau)h_b(\tau) - 2h_a(\tau)h_c(\tau) - \gamma h_a(\tau), \\
\frac{dh_b(\tau)}{d\tau} &= -2h_a(\tau)h_b(\tau) - 2h_b^2(\tau) - 2h_b(\tau)h_c(\tau) + \gamma h_a(\tau), \\
\frac{dh_c(\tau)}{d\tau} &= +2h_a(\tau)h_b(\tau) + h_b^2(\tau) - 2h_c^2(\tau) + h_c^2(\tau),
\end{aligned}$$

(Note that marks for changes on a type (b) lineage do not alter its rate, hence this is absent from the second equation; also, marks for change on a type (c) lineage as well as any mergers of type (c) lineage do not alter its rate hence all these events are absent from the last equation.)

Note that the initial values are $h_a(0) = 1, \;\; h_b(0) = 0, \;\; h_c(0) = 0.$ Using the fact that $h_a + h_b + h_c = \overline{F}_{\mathrm{s}} = 1/1 + \tau,$ the equation for h_a becomes

$$\frac{dh_a}{d\tau} = -2h_a\overline{F}_{\mathrm{s}} + h_a^2 - \gamma h_a = h_a^2 - h_a\big(\frac{2}{1+\tau} + \gamma\big), \;\; h_a(0) = 1.$$

Now using a transform $y_a = 1/h_a$ we obtain an equivalent equation for y_a

$$\frac{dy_a}{d\tau} = y_a\big(\frac{2}{1+\tau} + \gamma\big) - 1, \;\; y_a(0) = 1.$$

We can now use the integrating factor $e^{-\int_0^\tau(\frac{2}{1+s}+\gamma)ds} = (1+\tau)^{-2}e^{-\gamma\tau}$ to solve

$$\frac{d\Big((1+\tau)^{-2}e^{-\gamma\tau}y_a\Big)}{d\tau} = -(1+\tau)^{-2}e^{-\gamma\tau}$$

obtaining back in terms of h_a

$$h_a(\tau) = \frac{1}{y_a(\tau)} = \frac{(1+\tau)^{-2}e^{-\gamma\tau}}{1 - \int\limits_0^\tau (1+s)^{-2}e^{-\gamma s}d\tau}$$

Using the result for the law $\overline{F}_{\rm g}(\tau)$ of the time to the emergence of a genus from Lemma 12, we see that this is precisely $h_a(\tau) = \overline{F}_{\rm g}(\tau)/(1+\tau)$.

The equation for h_b is

$$\frac{dh_b}{d\tau} = -2h_b\overline{F}_{\rm s} + \gamma h_a = -\frac{2h_b}{1+\tau} + \gamma h_a, \;\; h_b(0) = 0.$$

hence

$$h_b(\tau) = (1+\tau)^{-2}\gamma\int\limits_0^\tau (1+s)^2 h_a(s)ds = (1+\tau)^{-2}\gamma\int\limits_0^\tau (1+s)\overline{F}_{\rm g}(s)ds$$

Finally, for h_c we have $h_c = \overline{F}_{\rm s} - h_a - h_b$ that is

$$h_c(\tau) = (1+\tau)^{-2}\Big((1+\tau)F_{\rm g}(\tau) - \gamma\int\limits_0^\tau (1+s)\overline{F}_{\rm g}(s)ds\Big)$$

. □

From the different types of lineages we can now derive the probabilities for different split types within the tree on genera.

Definition. Let the probabilities of split types in the model on genera be

$p_{\rm g}(1\;,1\;)(\tau) = \mathbf{P}$[of a (1,1) split type at time τ],

$p_{\rm g}(1,2^+)(\tau) = \mathbf{P}$[of any $(1,j)$, $j \geq 2$ split type at time τ],

$p_{\rm g}(2^+_;2^+)(\tau) = \mathbf{P}$[of any (i,j), $i > j \geq 2$ at time τ]

with $p_{\rm g}(1,1)$, $p_{\rm g}(1,2^+)$, $p_{\rm g}(2^+,2^+)$ the respective overall probabilities.

Let the ratio of balance be defined as $p_{\mathrm{g}} = p_{\mathrm{g}}(1,2^+)/p_{\mathrm{g}}(1,1)$.

It is clear how that the balance of the tree is determined by the ratio $p_{\mathrm{g}}(1,2^+)/p_{\mathrm{g}}(1,1)$. The higher the ratio is, the more balanced the tree. The value of such a ratio for a perfectly balanced tree is ∞, while for the comb tree it is 0. We are interested in determining how the (im)balance of the tree on genera compares to that of the tree on species, whose ratio of balance is equal to 1. Using the results of Lemma 12 we can now derive estimates for such a comparison.

Theorem 17. *The ratio of balance for the tree on genera satisfies*

$$\frac{l_{(1,2^+)}}{u_{(1,1)}} \leq p_{\mathrm{g}} \leq \frac{u_{(1,2^+)}}{l_{(1,1)}}, \tag{4.19}$$

where $l_{(1,2^+)}$, $u_{(1,2^+)}$, $l_{(1,1)}$, and $u_{(1,1)}$ are given by (4.20)-(4.23).

Proof. The proof relies on using the mean rates of different types of lineages in the model on genera, and then applying our earlier estimates for the law of the time of emergence of a genus $\overline{F}_{\mathrm{g}}$.

We first observe the results the mergers of different types of lineages produce. The merger of two type (a) lineages is not noted in the tree on genera, since both lineages are part of a yet incomplete genus. This merger is only seen as a branching point in the tree on species, and is absent from the tree on genera. We shall denote this type of a “non-existent” branching point by a split type $(0,0)$, an its probability at any time τ by $p_{\mathrm{g}}(0,0)(\tau)$, and overall by $p_{\mathrm{g}}(0,0)$. The merger of a type (a) lineage with a type (b) lineage, as well as a merger of two type (b) lineages gives a branching point of split type $(1,1)$. The merger of a type (c) lineage with either a type (a) lineage or a type (b) lineage generates a branching point of split type $(1,2^+)$. And finally, the merger of two type (c) lineages gives a branching point of split type $(2^+,2^+)$. Thus,

$$\begin{aligned}
&p_{\mathrm{g}}(0,0)(\tau) = h_a^2(\tau),\\
&p_{\mathrm{g}}(1,1)(\tau) = 2h_a(\tau)h_b(\tau) + h_b^2(\tau),\\
&p_{\mathrm{g}}(1,2^+)(\tau) = 2h_a(\tau)h_c(\tau) + 2h_b(\tau)h_c(\tau),\\
&p_{\mathrm{g}}(2^+,2^+)(\tau) = h_c^2(\tau).
\end{aligned}$$

Hence,

$$p_{\mathrm{g}} = \frac{\int_0^\infty \Big(2h_a(\tau)h_c(\tau) + 2h_b(\tau)h_c(\tau)\Big)\,d\tau}{\int_0^\infty \Big(2h_a(\tau)h_b(\tau) + h_b^2(\tau)\Big)\,d\tau}$$

Using the bounds for $\overline{F}_{\mathrm{g}}$ (from Lemma 12), together with the formulae for $h_a(\tau),\ h_b(\tau),\ h_c(\tau)$ (from Lemma 16), we have that

$$\frac{e^{-2\gamma\tau}}{1+\tau} \le h_a(\tau) \le \frac{e^{-\gamma\tau}}{1+\tau},$$

$$\frac{1}{2(1+\tau)^2}\Big(1-(1+\tau)e^{-2\gamma\tau}+\frac{1-e^{-2\gamma\tau}}{2\gamma}\Big) \le h_b(\tau) \le \frac{1}{(1+\tau)^2}\Big(1-(1+\tau)e^{-\gamma\tau}+\frac{1-e^{-\gamma\tau}}{\gamma}\Big),$$

$$\frac{1}{(1+\tau)^2}\Big(\tau-\frac{1-e^{-\gamma\tau}}{\gamma}\Big) \le h_c(\tau) \le \frac{1}{2(1+\tau)^2}\Big(\tau-\frac{1-e^{-2\gamma\tau}}{2\gamma}-(1+\tau)(1-e^{-2\gamma\tau})\Big).$$

Thus, the lower and upper bounds for $p_{\mathrm{g}}(1,2^+)$ are

$$l_{(1,2^+)} = \int\limits_0^\infty 2\quad \frac{1}{(1+\tau)^2}\Big(\tau-\frac{1-e^{-\gamma\tau}}{\gamma}\Big)$$
$$\Big(\frac{e^{-2\gamma\tau}}{1+\tau}+\frac{1}{2(1+\tau)^2}\Big(1-(1+\tau)e^{-2\gamma\tau}+\frac{1-e^{-2\gamma\tau}}{2\gamma}\Big)\Big)d\tau, \qquad (4.20)$$

$$u_{(1,2^+)} = \int\limits_0^\infty 2\quad \frac{1}{2(1+\tau)^2}\Big(\tau-\frac{1-e^{-2\gamma\tau}}{2\gamma}-(1+\tau)(1-e^{-2\gamma\tau})\Big)$$
$$\Big(\frac{e^{-\gamma\tau}}{1+\tau}+\frac{1}{(1+\tau)^2}\Big(1-(1+\tau)e^{-\gamma\tau}+\frac{1-e^{-\gamma\tau}}{\gamma}\Big)\Big)d\tau; \qquad (4.21)$$

and for $p_{\mathrm{g}}(1,1)$ are

$$l_{(1,1)} = \int\limits_0^\infty \quad \frac{1}{2(1+\tau)^2}\Big(1-(1+\tau)e^{-2\gamma\tau}+\frac{1-e^{-2\gamma\tau}}{2\gamma}\Big)$$
$$\Big(\frac{e^{-2\gamma\tau}}{1+\tau}+\frac{1}{2(1+\tau)^2}\Big(1-(1+\tau)e^{-2\gamma\tau}+\frac{1-e^{-2\gamma\tau}}{2\gamma}\Big)\Big)d\tau, \qquad (4.22)$$

$$u_{(1,1)} = \int\limits_0^\infty \quad \frac{1}{(1+\tau)^2}\Big(1-(1+\tau)e^{-\gamma\tau}+\frac{1-e^{-\gamma\tau}}{\gamma}\Big)$$
$$\Big(2\frac{e^{-\gamma\tau}}{1+\tau}+\frac{1}{(1+\tau)^2}\Big(1-(1+\tau)e^{-\gamma\tau}+\frac{1-e^{-\gamma\tau}}{\gamma}\Big)\Big)d\tau. \qquad (4.23)$$

yielding the claim. □

For the extreme values for γ ($= 0$, $\uparrow\infty$), we have the following. For $\gamma = 0$ we have $\forall\tau \geq 0$ all the lineages within one incomplete genus. Hence all the branching points can be seen solely within the tree on species and none within the tree on genera, corresponding to $p_{\mathrm{g}}(0, 0) = 1$, $p_{\mathrm{g}}(1, 1) = p_{\mathrm{g}}(1, 2^+) = p_{\mathrm{g}}(2^+, 2^+) = 0$. On the other hand for $\gamma\uparrow\infty$ we have immediately after $\tau = 0$ that all the extant lineages belong to their own species. This corresponds to $p_{\mathrm{g}}(0, 0) = p_{\mathrm{g}}(1, 1) = p_{\mathrm{g}}(1, 2^+) = 0$, $p_{\mathrm{g}}(2^+, 2^+) = 1$. For an arbitrary $\gamma > 0$, it is hard, though certainly not for lack of trying, to find good estimates on the lower and upper bounds for this ratio. However, numerical simulations for several values of $\gamma = 1/10, 1, 10$, etc. give values of the ratio greater than 1, suggesting that the tree on genera is more unbalanced than its respective tree on species.

Bibliography

[1] R. Abraham. Un arbre aletoire infini associe a l'excursion brownienne. In *Seminaire de Probabilites XXVI*, pages 374–397. Springer Verlag, 1992. Lect. Notes in Math. 1526.

[2] R. Abraham and L. Mazliak. Branching properties of Brownian paths and trees. *Expositiones Mathematicae*, 16:59–74, 1998.

[3] D.J. Aldous. The continuum random tree iii. *Ann. Probab.*, 21:248–289, 1993.

[4] D.J. Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statistical Science*, 16:23–34, 2001.

[5] D.J. Aldous and L. Popovic. Coherent stochastic models for macroevolution. in preparation, 2003.

[6] P. Billingsley. *Convergence of Probability Measrues.* John Wiley & Sons, 3rd edition, 1999.

[7] T. Duquesne. A limit theorem for the contour process of conditioned Galton-Watson trees. *Ann. Probab.*, 31:996–1027, 2003.

[8] R. Durrett. The genealogy of critical branching processes. *Stochastic Process. Appl.*, 8:101–116, 1978.

[9] W.J. Ewens. *Mathematical Population Genetics.* Springer Verlag, 1979.

[10] W. Feller. *An Introduction to Probability Theory and its Applications*, volume 1st. John Wiley & Sons, 3rd edition, 1968.

[11] J. Geiger. Contour processes of random trees. In *Stochastic Partial Differential Equations*, pages 72–96. Cambridge University Press, 1995. London Math. Soc. Lect. Notes 216.

[12] J. Geiger. Poisson point process limits in size-biased Galton-Watson trees. *Electron. J. Probab.*, 5(17):1–12, 2000.

[13] D.G. Hobson. Marked excursions and random trees. In *Seminaire de Probabilites XXXIV*, pages 289–301. Springer Verlag, 2000.

[14] J.F.C Kingman. *Mathematics of Genetic Diversity.* S.I.A.M,Philadelphia, 1980.

[15] J.F. Le Gall. Marches aleatoires, mouvement brownien et processus de branchement. In *Seminaire de Probabilites XXIII*, pages 258–274. Springer Verlag, 1989. Lect. Notes in Math. 1372.

[16] J.F. Le Gall. Brownian excursions, trees and measure-valued branching processes. *Ann. Probab.*, 19:1399–1439, 1991.

[17] J.F. Le Gall and Y. LeJan. Branching processes in Levy processes: the exploration process. *Ann. Probab.*, 26:1407–1432, 1998.

[18] J. Neveu and J.W. Pitman. The branching process in a Brownian excursion. In *Seminaire de Probabilites XXIII*, pages 248–257. Springer Verlag, 1989. Lect. Notes in Math. 1372.

[19] J. Neveu and J.W. Pitman. Renewal property of the extrema and the tree property of a one-dimensional Brownian motion. In *Seminaire de Probabilites XXIII*, pages 239–247. Springer Verlag, 1989. Lect. Notes in Math. 1372.

[20] N. O'Connell. The geneology of branching processes and the age of our most recent common ancestor. *Adv. in Appl. Probab.*, 27:418–442, 1995.

[21] J. Pitman. Combinatorial stochastic processes. Technical Report SD-621, UC Berkeley, 2002. Lect. Notes for St Flour Course July 2002.

[22] D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion.* Springer Verlag, 1st edition, 1991.

[23] L.C.G. Rogers and D. Williams. *Diffusions, Markov processes, and Martingales*, volume 1 &2. John Wiley & Sons, 2nd edition, 1987.

[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.