Technische Universität München
Zentrum Mathematik

Evolving trees: Models for Speciation and Extinction in Phylogenetics

Tanja Stadler

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. Rupert Lasser
Prüfer der Dissertation:	1.	Univ.-Prof. Dr. Anuschirawan Taraz
	2.	Prof. Dr. Jotun Hein, University of Oxford / United Kingdom
	3.	Prof. Dr. Mike Steel, University of Canterbury / New Zealand

Die Dissertation wurde am 10.09.2008 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 10.12.2008 angenommen.

Nothing in Biology Makes Sense Except in the Light of Evolution

Theodosius Dobzhansky

Abstract

A phylogeny represents the evolutionary relationship between species. In this thesis, we answer questions arising in the process of reconstructing and analyzing phylogenies. We develop and discuss a general class of neutral models for speciation and extinction. These results are used in our novel algorithm for dating supertrees as well as for drawing lineages-through-time plots. Further, using the analytic results, we can calculate p-values of our introduced statistic to test the evolutionary hypothesis of lineage-specific bursting. We provide accurate simulation tools for models which cannot or have not been analyzed. Widely used simulation algorithms are shown to be wrong for these general models. We end the thesis by providing a complexity result for dating phylogenies with reticulation events. In general, a phylogeny with reticulations does not necessarily have a valid temporal dating. We prove that adding missing species in an optimal way, such that the altered phylogeny has a dating, is NP-complete. The main mathematical tools used in this thesis come from stochastics, statistics, combinatorics and complexity theory.

Zusammenfassung

Eine Phylogenie repräsentiert die Verwandtschaftsbeziehungen zwischen Spezies. In der vorliegenden Arbeit beantworten wir Fragestellungen, die bei der Rekonstruktion und Analyse von Phylogenien auftreten. Wir entwickeln eine allgemeine Klasse von neutralen Modellen für Speziation und Aussterben. Die Ergebnisse werden in unserem neuen Algorithmus zur Datierung von Supertrees sowie zur Erstellung von so genannten lineages-through-time Plots verwendet. Des Weiteren benutzen wir unsere theoretischen Ergebnisse, um p-Werte für die entwickelte Statistik zum Testen von Phylogenien auf Linien-spezifische Bursts zu berechnen. Für Modelle, die nicht analysiert werden können oder nicht analysiert wurden, stellen wir Simulationsalgorithmen zur Verfügung. Wir zeigen, dass weit verbreitete Simulationsprogramme selbst unter gängigen Modellen fehlerhaft arbeiten. Zum Abschluss der Arbeit beschäftigen wir uns mit der Datierung von Phylogenien, in denen Retikulationsereignisse auftreten. Eine Phylogenie mit Retikulationen besitzt nicht immer eine zulässige zeitliche Datierung. Wir beweisen, dass das optimale Hinzufügen von fehlenden Spezies, so dass die neue Phylogenie eine Datierung besitzt, NP-vollständig ist. Die in dieser Arbeit verwendeten mathematischen Methoden kommen hauptsächlich aus der Stochastik, Statistik, Kombinatorik und Komplexitätstheorie.

Acknowledgements

Completing my PhD thesis has been a wonderful experience. My advisor Anusch Taraz has been very supportive throughout the whole time. He helped me to establish my direction in research and always gave me very valuable advice.

The visits to New Zealand would not have been possible without Mike Steel providing fantastic support. Talking to Mike was very inspiring and led to creative new ideas. Mike always encouraged me to continue following my interests and goals.

Erick Matsen was a great office mate in New Zealand and Berkeley. Working with him on the runs statistic was a lot of fun. And Erick, thanks for fixing my laptop! Klaas Hartmann is a great collaborator asking the right questions regarding useful applications. On top, he is the best mountaineer guide.

Collaborations and discussions with Vincent Berry, James Degnan, Alexei Drummond, Daniel Ford, Simone Linz, Dirk Metzler, Noah Rosenberg, Roland Schultheiß, Charles Semple, Tom Wilke and Dennis Wong led to new exciting projects and helped me to see my research in a broader scientific perspective.

Further, I would like to thank the people of the group M9 in the Mathematics department at the Technische Universität München, the people of the Biomathematics research center at the University of Canterbury in New Zealand and the people of the Evolutionary Biology department at the Ludwig-Maximilian-Universität in München for providing a great atmosphere to work in.

Financial support by the Deutsche Forschungsgemeinschaft through the graduate program "Angewandte Algorithmische Mathematik" at the Technische Universität München and by the Allan Wilson Center through several summer studentships is gratefully acknowledged.

Last but not least I want to thank my husband and my family for the fantastic support during all the ups and downs while working on the thesis.

List of Symbols

Symbol	*Meaning*	*page*
$\mathcal{A}_n^{i,j}$	time between i-th and j-th speciation event in tree with n leaves	23
$\mathcal{A}_n^k$	date of k-th speciation event in tree with n leaves	23
$\mathcal{A}_{\mathcal{T}}^e$	length of edge e in tree $\mathcal{T}$	23
$\mathcal{A}_{\mathcal{T}}^u$	date of vertex u in tree $\mathcal{T}$	23
λ	birth rate	6
μ	death rate	6
$\mathcal{N}$	reticulation network	108
$p_u(i)$	probability of vertex u having rank i in $\mathcal{T}$	19
$p_{u,v}(i,j)$	probability of vertex u having rank i and v having rank j in $\mathcal{T}$	20
r	rank function of an oriented tree $\mathcal{T}$	7
$r(\mathcal{T})$	set of rank functions on $\mathcal{T}$	7
$\mathcal{R}(\mathcal{T})$	number of runs in tree $\mathcal{T}$	77
t_{or}	time of origin of a reconstructed oriented tree	5
τ	tree shape	8
$\mathcal{T}, \mathcal{T}_n$	(reconstructed / ranked) oriented tree on n leaves	6
$\mathcal{T}_v$	subtree of $\mathcal{T}$ induced by the descendants of vertex v	10
$\mathring{V}$	set of interior vertices of an (oriented) tree	7
BDA	constant rate birth-death approach	99
BDP	constant rate birth-death process	2, 32
cBDP	BDP conditioned on obtaining n leaves today	32
CBP	constant rate critical branching process	6, 30
cCBP	CBP conditioned on obtaining n leaves today	30
CRP	constant relative probability (model)	15
DNA	desoxyribonucleic acid	1
ET	evolving tree (model)	5
GSA	general sampling approach	96
HCV	Hepatitis C virus	75
HGT	horizontal gene transfer	106
LSB	lineage specific bursting	73
LTT	lineages-through-time (plot)	2

MCMC	Markov chain Monte Carlo	1
mrca	most recent common ancestor	7
PBMSA	pure-birth memoryless sampling approach	93
PDA	proportional-to-distinguishable-arrangements	6
SSA	simple sampling approach	91
UR	uniform rank (model)	2, 17

Contents

Chapter 1

Introduction

In phylogenetics, the central goal is to reconstruct the evolutionary relationship between known species – the phylogeny of the species. The commonly used reconstruction methods are distance-based methods, maximum parsimony, maximum likelihood, and Markov chain Monte Carlo (MCMC) methods, for a detailed discussion see [21]. Reconstruction methods assume some model for the evolution of species, a *macro-evolutionary model*. In this thesis, we will extensively discuss such models.

Species evolve because of changes of the four nucleotides (adenine, cytosine, guanine, thymine) in the desoxyribonucleic acid (DNA). There are various models describing these changes. The simplest model is the Jukes-Cantor model [43], where each base in the DNA sequence has an equal chance of undergoing a substitution event. Under this model, when a base changes, it *mutates* to one of the three other bases with equal probability. More general models have been proposed in the literature [49, 55, 21].

Branching processes are used for modeling speciation and extinction. A mutation model then evolves on the branching process: A sequence of nucleotides mutates along a branch of the species tree. At a speciation event, the sequence is copied, and one sequence is evolving on each descending branch. A branching process with a mutation model specifies a macro-evolutionary model. A variety of branching processes as models for speciation and extinction have been proposed. The simplest model is the *Yule model* [101], where each species evolves independently and produces new species at a constant rate λ. The Yule model is extended to models with extinction by introducing a constant death rate μ, such a model is called *constant rate birth-death process* [20, 47]. The case $\mu = \lambda$, a *critical branching process*, has been studied in detail as a model for speciation and extinction in [74, 2]. Departing from this classical birth-death model, there have been proposed a variety of more complex models where the speciation rate is a function of the age of the species or the size of the population. Further, speciation rates might change over time by a random process. For an overview of these models see [64]. Other models have been proposed where the rates depend on a character state of the species [60].

However, even the simple models are not well-understood, which becomes ap-

parent in incorrect sampling algorithms for these models (see Chapter 5). We will define and investigate a broad class of neutral models for speciation and extinction (Chapter 2 and 3), this class includes the constant rate birth-death process. In Chapter 2, we model trees without specifying a time between speciation events. We call the considered class of models *uniform rank* (UR) models. Note that we model the ordering of the speciation events, but not the time between speciation events. This generalizes the α model [23] and β model [1], which are inducing a distribution only on tree shapes, without an ordering of the speciation events. In Chapter 3, we specify the speciation times; the considered model is a constant rate birth-death process (BDP) [20, 47].

The derived analytic results for the BDP in Section 3.4 are used in the MCMC inference program BEAST [16]: Given an alignment of sequences, and a prior distribution on the evolutionary trees, BEAST calculates the posterior distribution on trees with the given sequences being the leaves. The coalescent is used as a prior for micro-evolution (i.e. evolution within a species). For macro-evolution, the uniform distribution on trees has been used as a prior. Recently, our derived prior under the constant rate birth-death process has been implemented into BEAST .

Reconstruction methods depend to a great extent on the model which is assumed. In order to develop realistic models, we need to understand the biological process of speciation and extinction. Known phylogenies are compared to the prediction of commonly used models. A powerful statistic is required in order to decide if the model is reasonable for the phylogeny.

Standard statistics are the *Colless statistic* [11] and the γ *statistic* [76] with its visualization through *lineages-through-time* (LTT) plots [35]. The Colless statistic performs on tree shapes, i.e. phylogenies without edge lengths. The tree shape is a purely combinatorial structure. The distribution of the Colless values under the BDP is known [81]. The γ statistic and the corresponding LTT plot consider the time between speciation events, but not the tree shape. We derive analytic results for the LTT plot of trees with n species under the BDP in Section 3.6 and 3.8.

In 2004, Joseph Felsenstein pointed out in his fundamental book on inferring phylogenies [21] the lack of a statistical test which incorporates tree shape and timing information. He wrote, "At present we are lacking the following: (i) Any method that uses both of these kinds of information (tree shapes and branch lengths). (ii) Any framework that takes into account the uncertainty of branch lengths and tree shapes."

Chapter 4 introduces a new statistic on ranked trees which combines relative ordering of speciation events and tree shape, the runs statistic. A statistic on relative timing is more robust towards uncertainties in edge lengths than a statistic on the actual edge lengths. Further, using relative timing instead of edge lengths allows us to apply the statistic not only to speciation models, but also to the coalescent [53, 51, 52] with varying population size, the standard model in population genetics. We derive analytic results for the distribution of the runs statistic under the BDP and the UR model. The statistic is applied to different data sets to investigate the underlying evolutionary process for these data.

When reconstructing a phylogeny for a clade, we might have missing species, i.e. not all species are sampled. Incomplete sampling is modeled by choosing uniformly at random some leaves from the tree on all extant species. This is called *random taxon sampling*. The Colless statistic, as well as the runs statistic, is invariant under random taxon sampling. For LTT plots, we discuss the effect of random taxon sampling (Section 3.7). In particular, we show that expected LTT plots of trees with random taxon sampling look like LTT plots of trees with complete taxon sampling and a smaller death rate. Therefore, from the LTT plots, incomplete taxon sampling cannot be detected. Further, since the ranked tree distribution is invariant under random taxon sampling, missing taxa cannot be detected by the shape of the phylogeny.

Analytic results for the BDP also find application in supertree methods. A supertree is a phylogeny on lots of species, inferred by combining many trees on subsets of the species. In supertree reconstruction, we are usually not able to date all speciation events. For undated nodes, estimates are required. In [98], an undated speciation event is estimated via the expectation of the time of that vertex under a BDP without extinction, the Yule model. The expectation is obtained via simulations. The author of [98] asked for a fast analytic approach (personal communication). In Section 3.1.1 we calculate the distribution and the expectation for the time of a vertex in polynomial time for any BDP model. The analytic method for dating supertrees is coded as part of my Python package Cass [89]. Currently, Jonathan Davies is using Cass for dating his Carnivora supertree. In Section 3.9, the ideas are generalized for dating trees where the time of some speciation events is known.

If we consider more complex models than the BDP, it is often difficult to obtain analytic solutions, or analytic solutions might not exist for the considered problem. In order to get a better understanding of the model, we need to simulate trees on n species. This has some pitfalls, and commonly used simulation tools actually do not produce the correct distribution. We develop correct algorithms for arbitrary models of speciation and extinction. Klaas Hartmann coded the algorithms in Perl and provides a stand-alone application [33].

The evolution within a species, i.e. the evolution of a population, is modeled by a *micro-evolutionary model*. A lot of the results and methods developed in this thesis for macro-evolutionary models apply to micro-evolutionary models as well. We will describe the coalescent as the standard micro-evolutionary model, and point out the relationships to macro-evolutionary models in the different chapters.

So far, we assumed that evolution is tree-like, i.e. species pass on their genetic material to daughter species. However, *reticulate evolution* is observed for some classes of species. Reticulate evolution means that genetic material is passed to co-existing lineages (horizontal gene transfer), or two lineages join to a hybrid (hybridization). Horizontal gene transfer is observed in bacteria, hybridization in plants and fish. Reticulate evolution is modeled as a binary network. Note that in a tree, a valid dating for the speciation events always exists; we provide estimates for the dates under a BDP in Chapter 3. However, reticulation networks might not have a valid dating. By adding (non-sampled) species, we can alter the network such that a valid

dating exists [4]. In Chapter 6, we show that determing the minimum number of species to add such that the network has a valid dating is NP-complete.

This thesis is organized as follows. In the remainder of this chapter, we formally introduce the main objects studied in the thesis, models for speciation and extinction, as well as the concept of neutrality and the terminology used in the thesis. In Chapter 2, a broad class of neutral models for speciation and extinction is introduced; the time between speciation events is not specified. These results are fundamental for the later chapters in the thesis. Chapter 3 derives analytic results for the BDP, a neutral model with an exponential distributed lifetime of a species. The analytic results are used for understanding the effect of random taxon sampling (Section 3.7), obtaining LTT plots (Section 3.6 and 3.8), and dating phylogenies (Section 3.1.1). The methods for dating phylogenies are extended to trees where some vertices have a known date (Section 3.9) – we condition on the known dates to calculate the expectation of the other dates. Chapter 4 introduces the new statistic on phylogenies, the runs statistic. The statistic is applied to a data set of the Hepatitis C virus [58, 78], to ant phylogenies [66, 67] and to the genus *Dina* (family leech) in the ancient Lake Ohrid [96]. Algorithms for simulating trees under an arbitrary model for speciation are provided in Chapter 5. In Chapter 6, we discuss some issues arising in the case of reticulate evolution.

Most of Chapter 2 and Chapter 4 is joint work with Daniel Ford and Erick Matsen. The application of the statistic to the genus *Dina* in Lake Ohrid is joint work with the group of Tom Wilke at the University of Gießen. Chapter 5 is joint work with Klaas Hartmann and Dennis Wong. I published or submitted the results of this thesis in the following articles:

T. Gernhard. New analytic results for speciation times in neutral models. *Bull. Math. Biol.*, 70(4): 1082-1097, 2008.

T. Gernhard. The conditioned reconstructed process. *J. Theo. Biol.*, 253(4): 769-778, 2008.

T. Gernhard, K. Hartmann, M. Steel. Stochastic properties of generalised Yule models, with biodiversity applications. *J. Math. Biol.*, 57: 713-735, 2008.

T. Stadler. Lineages-through-time plots of neutral models for speciation. *Math. Biosci.*, 216: 163-171, 2008.

D. Ford, E. Matsen, T. Stadler. A method for investigating relative timing information on phylogenetic trees. *Under review,* 2008.

K. Hartmann, T. Stadler, D. Wong. Sampling trees from evolutionary models. *Under review,* 2008.

S. Trajanovski, C. Albrecht, R. Schultheiß, T. Gernhard, M. Benke, T. Wilke. Testing the temporal framework of speciation in an ancient lake species flock: the genus Dina (Hirudinea: Erpobdellidae) in Lake Ohrid. *Under review,* 2008.

The major methods introduced in Chapter 2–4 are implemented in Python in the package CASS [89], the package can be downloaded from my website http://www.tb.ethz.ch/people/tstadler. The algorithms in Chapter 5 have been implemented in PERL by Klaas Hartmann [33].

1.1 Evolving tree (ET) models

The main part of this thesis deals with tree-like evolution, i.e. it is assumed that species pass on genetic material only to their descendants (and not to sister species). A phylogeny connecting species through a tree is called a *phylogenetic tree.* Tree-like evolution is modeled in a very general way by a stochastic process where the leaves in a tree bifurcate and produce two new leaves. The time of bifurcation of a leaf is determined by the stochastic process. If we have an extinction event, the leaf which goes extinct will not speciate further. We call this class of bifurcation and extinction models the *evolving tree (ET)* models. To distinguish between the two new leaves after a bifurcation event, we label one edge descending from a bifurcation with *left* and the other edge with *right*. Any evolutionary process on binary trees can be modeled by an ET model: leaves bifurcate in an arbitrary fashion and produce new leaves. A *pure-birth* ET model is an ET model without extinction.

We condition the ET models to have n extant leaves today. This is crucial for comparing the ET model to a given phylogeny on n species. Let today be time 0 and the origin of the tree be time $t_{or} > 0$, so time is increasing going into the past. If t_{or} is not known, we assume a uniform prior on $(0, \infty)$ for the time of origin as it has been done in [2, 74]. Note that this prior does not integrate to 1. For any constant function, the integral is ∞. Therefore the prior is not a density. Such a prior is called *improper*; a discussion and justification is found e.g. in [5]; the idea is that a uniform prior is defined on $(0, T)$ and then we take the limit $T \to \infty$. Conditioning t_{or} on obtaining n species today yields,

$$f(t_{or}|n) = \frac{f(n|t_{or})f(t_{or})}{\int_0^\infty f(t_{or}, n)dt_{or}} = \frac{f(n|t_{or})}{\int_0^\infty f(n|t_{or})dt_{or}}, \tag{1.1}$$

i.e. t_{or} conditioned on n is a well-defined distribution if $\infty > \int_0^\infty f(n|t_{or})dt_{or} > 0$. In [28], it is shown that $\infty > \int_0^\infty f(n|t_{or})dt_{or}$ if the expected lifetime of each species is finite. For the models we consider, we have an expected finite lifetime. Further, for our models, we have $f(n|t_{or}) > 0$ for $t_{or} \in (0, \infty)$, which yields $\int_0^\infty f(n|t_{or})dt_{or} > 0$.

1.2 Neutral models

Neutral models are pure chance models, where no prior assumptions are made [3]. We introduce uniform rank (UR) models as a general neutral pure-birth model (Chapter 2). In an UR model, each ordering (or rank function, as defined in Section 1.3) of speciation events in a phylogenetic tree is equally likely. This means that the UR models are neutral in the sense that no ordering on a given tree is favored. A

more restrictive though widely used class of neutral models is the class of entirely homogeneous models; it is assumed that throughout time, whenever a speciation (or extinction) event occurs, each species is equally likely to be the one undergoing that event. We call the class of homogeneous models the *constant across lineages* (CAL) class of models. We will see that in this class each "ranked oriented tree" (as defined in Section 1.3) on n species is equally likely. Note that this is different from the PDA (proportional-to-distinguishable-arrangements) model [82]. Under the PDA model, each oriented (or labeled) tree is equally likely. For the CAL models, each ranked oriented tree is equally likely, i.e. each oriented tree with a rank function. The pure-birth CAL models are a subset of the UR models.

Note that a model in the CAL class can be formulated without specifying speciation times. In Chapter 3 we investigate the constant rate birth-death process [20, 47] as it is probably the most accepted and most popular homogeneous model with specified speciation times. A constant rate birth-death process is a stochastic process which starts with an initial species. A species gives birth to a new species after an exponential (rate λ) waiting time and dies after an exponential (rate μ) waiting time, where $0 \leq \mu \leq \lambda$. The species evolve independently. Special cases of the birth-death process are the Yule model [101] where $\mu = 0$ and the critical branching process (CBP) [2, 74] where $\mu = \lambda$. Note that for all birth-death processes where $\mu < \lambda$, the expected number of extant species is increasing exponentially. For $\mu = \lambda$, we do not have an expected exponential increase, which is more reasonable biologically [3]. However, the process goes extinct with probability 1. So if a BDP is assumed, the death rate μ should be chosen slightly smaller than λ. In [80, 61], the extinction rate is estimated from data to be $\mu = 0.9\lambda$. For an overview of the connection between the discussed models in the thesis, see Figure 1.1.

Of course speciation is not just random – lineages will differ in their expected diversification rates for both instrinsic and extrinsic factors [64]. However, for rejecting a neutral model for some data set, we need to know characteristics of the neutral models, for example the tree balance distribution or the speciation times distribution. We then test whether the data set also has these characteristics. On the other hand, if we may assume a neutral model for some phylogeny, we can make further statements about the evolution of the phylogeny – inferred from the characteristics of the neutral model. Chapter 2 and 3 discuss some characteristics of neutral models. Chapter 4 uses the analytic results of the neutral models for calculating the p-values of the runs statistic.

1.3 Terminology

In this section, we introduce the terminology for trees used throughout the thesis.

Definition 1.3.1. An *oriented tree* is a rooted binary tree where the descending edges of each vertex have distinct orientations, *left / right*. Above the *root*, there is an additional edge without orientation, the root edge. The parent vertex of the root edge is the *origin* of the tree. For an example see Figure 1.2, right.

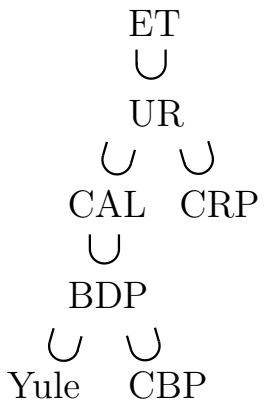

Figure 1.1: Overview of the discussed classes of evolving tree models in this thesis. Note that the PDA model is not listed, since it is not described as an evolving tree model but defined as each labeled tree being equally likely.

Definition 1.3.2. A *complete oriented tree* is an oriented tree where each edge has a positive real value assigned, the *length* of the edge. For an example see Figure 1.2, left.

Definition 1.3.3. A *reconstructed oriented tree* is a complete oriented tree where the length of any path from the root to a leaf is the same. For an example see Figure 1.2, middle.

Note that a realization of the ET model is a complete oriented tree. A complete oriented tree induces a reconstructed oriented tree in the following way: We delete the extinct leaves together with the attached edges from the complete oriented tree. This results in degree-two vertices in the remaining tree. Replace iteratively each degree-two vertex and its two incident edges by a single edge, with orientation inherited from the deleted edge which was closer to the origin. This results in a binary tree where each vertex (except of the origin) has a descending edge *left* and *right*. An oriented tree is induced by the reconstructed oriented tree when dropping the edge lengths. We use the following notation for a (complete / reconstructed) oriented tree induced by an ET model.

Definition 1.3.4. The *most recent common ancestor (mrca)* in a (complete / reconstructed) oriented tree which is induced by an ET model is the "earliest" bifurcation in the tree where the left and right branch have extant species descending. Note that for a (reconstructed) oriented tree, the *mrca* coincides with the root of the tree. The sum of edge lengths on a path from the origin to the extant leaves is the *age* of the tree or the *time since origin*. *Today* or *the present* is the time of the extant leaves. The time of today shall be zero, time is increasing going into the past.

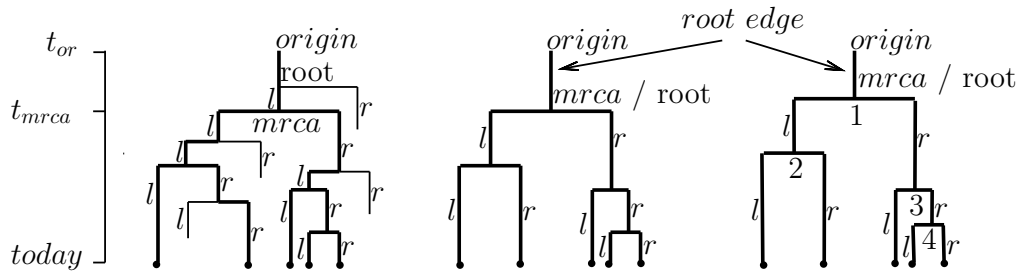

Figure 1.2: A complete oriented tree (left) and its reconstructed, oriented tree (middle). The length of an edge is defined by the vertical time axis. Below each bifurcation, the branch on the left has orientation l (for $left$) and the branch on the right orientation r (for $right$). The right tree is the ranked, oriented tree induced by the reconstructed, oriented tree.

Definition 1.3.5. We denote the set of interior vertices in an oriented tree – i.e. the set of all vertices except of the leaves and the origin of the tree – by $\mathring{V}$. A *rank function* r [85] on an oriented tree is a bijection from $\mathring{V} \to \{1, 2, \ldots, |\mathring{V}|\}$ where the ranks are increasing on any path from the root to the leaves. A ranked oriented tree is an oriented tree with a rank function. For an example see Figure 1.2, right. The set of all rank functions on an oriented tree $\mathcal{T}$ is denoted by $r(\mathcal{T})$.

A rank function induces a total order on $\mathring{V}$ which can be interpreted as the order of speciation events in the oriented tree. A vertex with rank k is the k-th speciation event in the oriented tree.

Definition 1.3.6. A *labeled oriented tree* is an oriented tree where the leaves are uniquely labeled. A *labeled tree* is a labeled oriented tree without the orientation, i.e. a binary rooted labeled tree with root edge. A *tree shape* is an oriented tree without the orientation. The shape of an oriented tree $\mathcal{T}$ is the tree shape induced by $\mathcal{T}$ via dropping the orientation. Note that the shape of a tree is also called the topology of a tree in the literature. A *complete / reconstructed / ranked labeled tree* and a *complete / reconstructed / ranked tree shape* is defined analogue.

When inferring a tree from some data, we obtain a complete tree if the fossil record is included. For data consisting only of extant species, we obtain a reconstructed tree. Since most methods infer trees on extant species, it is crucial to understand the distribution on reconstructed trees induced by ET models.

In the biological application, when inferring trees on the extant species, we obtain reconstructed labeled trees. In an evolving labeled tree, we assign a unique label for each new leaf and eliminate the label in case of extinction. For trees on n leaves evolving under a speciation and extinction model, however, we cannot guarantee that all such trees on n species have the same leaf labels (due to random extinction events). Therefore, it will be convenient to discuss oriented trees rather than labeled

trees. This allows us to distinguish the children of each vertex without having to explicitly label species which may later become extinct.

Chapter 2

Neutral models for ranked trees

In this chapter, we will discuss properties of ET models and introduce neutral classes of ET models. Throughout this chapter, we do not specify the time between successive speciation events. The results hold for any specification of waiting time between speciation events. We discuss the distribution on ranked trees which are induced by an ET model. We introduce *tree shuffles* which are equivalent to rank functions. For many arguments, shuffles are more convenient to use than rank functions, because shuffles define a rank function recursively. The results in this chapter are extensively used in the later chapters.

2.1 Shuffles and ranks

There is a bijection between the rank function of a tree, and a *tree shuffle* as defined in this section. We will need the notation of shuffles, since shuffles define a rank function recursively. This recursive formulation will be of particular use in Chapter 4.

For an internal node $v \in \mathring{V}$ of an oriented tree $\mathcal{T}$, define $\mathcal{T}_v$ to be the subtree of $\mathcal{T}$ rooted in v containing all the descendants of v. The *daughter trees* of v are the two subtrees of $\mathcal{T}_v$ which we obtain by deleting v and its two incident edges.

A *total order* on a set is a binary relation (usually written $<$) such that for any two distinct elements a and b of the set either $a < b$ or $b < a$. Note that a rank function on a set is equivalent to a total order on that set: given a total order one can rank the elements in increasing order of rank, and given a rank function one can define a total order by numerical inequality of rank. Thus a ranked labeled tree is exactly a tree equipped with a total order on its internal nodes.

An (m, n) *shuffle* on symbols p and q is a sequence of length $m + n$ containing m p's and n q's. For example $pqppq$ is a $(3, 2)$ shuffle on p and q. The usefulness of these shuffles in the present context is summarized in the following lemma.

Lemma 2.1.1. *Given totally-ordered sets P and Q, the total orderings of $P \cup Q$ respecting the given orderings of P and Q are in one-to-one correspondence with the $(|P|, |Q|)$ shuffles on symbols p and q.*

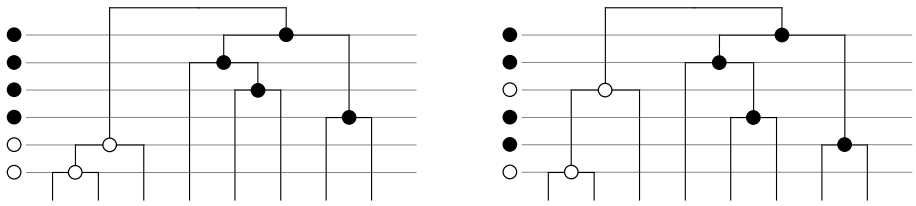

Figure 2.1: A shuffle at a given internal node. Bifurcations on the left subtree are marked with a hollow circle, and those on the right subtree are marked with a solid circle. The relative timing for these events is shown beside the tree; we call this sequence of symbols a *shuffle*. A set of shuffles for every internal node of a phylogenetic tree exactly determines the relative order of speciation events. Similar type symbols occurring together as in the left tree is evidence of lineage-specific bursts.

Proof. Assume total orders $p_1 < p_2 < \cdots < p_m$ on P and $q_1 < q_2 < \cdots < q_n$ on Q are given, along with an (m, n) shuffle on p and q. The required total ordering on $P \cup Q$ is obtained by progressing along the shuffle and substituting p_i and q_j for p and q in order: for example the shuffle *pqppq* uniquely defines the total order $p_1 < q_1 < p_2 < p_3 < q_2$ when $p_1 < p_2 < p_3$ and $q_1 < q_2$. In the other direction, a total ordering on $P \cup Q$ uniquely defines a $(|P|, |Q|)$ shuffle and a total ordering on each of P and Q. □

We will now define shuffles on a tree. Assume that v is an internal node of a tree and that the tree $\mathcal{T}_v$ containing the descendants of v is composed of two daughter subtrees L_v and R_v. Assume L_v and R_v have m and n internal nodes, respectively. We define a *shuffle at the internal node* v to be an (m, n) shuffle on symbols ℓ and r. The intuition behind the shuffle idea at vertices is presented in Figure 2.1. As shown in this figure, the relative order of speciation events for an internal node of a tree is determined by the sequence of full and hollow circles on the left side of each tree. This sequence induces a shuffle at node v.

We can use shuffles to develop a recursive formulation of ranked oriented trees. By Lemma 2.1.1, a total ordering on the internal nodes of $\mathcal{T}_v$ respecting the orderings on the internal nodes of L_v and R_v is equivalent to a shuffle at the internal node v. Therefore we can recursively reconstruct the rank function for any ranked tree given a shuffle at each internal node. We define a *tree shuffle* to be such a choice of shuffles. With Lemma 2.1.1, we have the following result, which is crucial to our analysis:

Lemma 2.1.2. *Each rank function on a given tree being equally likely is equivalent to the statement: For each internal node v, each shuffle at v is equally likely and these shuffles are independent.*

Shuffles also have a natural interpretation in terms of evolutionary history. Namely, *bursting* diversification leads to symbols of a shuffle clustering together. The opposite situation, where there is a post-diversification delay before a lineage

can diversify again, can be recognized by the interspersing of different symbols. This latter situation has been called *refractory* diversification [57].

2.2 Ranked oriented trees

We will discuss the distribution on ranked oriented trees and on ranked tree shapes. This will be useful when introducing the runs statistic in Chapter 4.

Proposition 2.2.1. *There are* $(n-1)!$ *ranked oriented trees on* n *leaves.*

Proof. Proceed by induction on n; for $n = 2$ the statement is obviously true, there is only one ranked oriented tree. Suppose there are $(n-1)!$ ranked oriented trees on n leaves. A ranked oriented tree on $n+1$ leaves is uniquely determined by a tree on n leaves and an additional leaf which evolved at the n-th bifurcation event. There are n possibilities to attach the additional leaf. Thus there will be $n(n-1)! = n!$ ranked oriented trees on $n+1$ leaves. □

Lemma 2.2.2. *Given a ranked oriented tree with* n *leaves, there are* $n(n+1)$ *ways to add an additional leaf.*

Proof. First, decide which rank the new internal node will have, from 1 (earliest) to n (latest). If the new internal node has rank k then there are k choices at that level for the edge to add it to, and then 2 choices for which side of this edge the new pendant leaf will sit. This gives a total of $2\sum_{i=1}^{n} i = 2\frac{n(n+1)}{2} = n(n+1)$ ways to insert the new leaf edge. □

In Section 4.3 we consider, for computational convenience, the likelihood of rank functions on tree shapes rather than on oriented trees. The following proposition and corollary show that the uniform distributions on rank functions on a given oriented tree induce the uniform distributions on rank functions on a given tree shape when orientation is forgotten. In particular, p-values for such rank functions may be computed over either oriented trees or tree shapes.

First, define a *symmetric vertex* to be a vertex for which the unoriented shapes of the two descending subtree are the same (isomorphic as tree shapes). A *non-trivial symmetric vertex* is a symmetric vertex with more than two leaves below.

Proposition 2.2.3. *A uniform distribution on rank functions on a given oriented tree induces a uniform distribution on rank functions on its corresponding tree shape.*

Proof. Let $\mathcal{T}$ be an oriented tree with n leaves and τ its corresponding tree shape. Let q denote the number of non-trivial symmetric vertices of $\mathcal{T}$. We prove the following statement for oriented trees by induction on n: for each ranking r on τ there are exactly 2^q rankings on $\mathcal{T}$ with r and τ being the corresponding ranked tree shape. This establishes the proposition.

For $n = 2$, which implies $q = 0$, we have for the ranking on τ exactly $1 = 2^0$ ranking on $\mathcal{T}$. For $n > 2$, the induction breaks into three cases.

Case 1: Suppose the two subtrees descending the root branch-point of τ are non-isomorphic tree shapes, each having more than 1 leaf. The subtrees may therefore be distinguished from each other, and given a ranking on τ, the shuffle at the root node of $\mathcal{T}$ is determined. Call the two daughter subtrees $\mathcal{T}_1, \mathcal{T}_2$, with corresponding shapes τ_1, τ_2 and q_1, q_2 non-trivial symmetric vertices, respectively. By the inductive assumption, there are 2^{q_1} rankings for τ_1 and 2^{q_2} rankings for τ_2. This gives $2^{q_1+q_2}$ total since there is no choice for the shuffle at the root branch-point of $\mathcal{T}$. $q_1 + q_2$ is the number of non-trivial symmetric vertices of τ.

Case 2: Suppose the two subtrees descending the root branch-point of τ are non-isomorphic tree shapes, one of the children being a leaf. The subtrees may therefore be distinguished from each other, and given a ranking on τ the shuffle at the root node of $\mathcal{T}$ is determined. The bigger subtree $\mathcal{T}_b$ has q_b non-trivial symmetry vertices. By the inductive assumption, there are 2^{q_b} rankings for $\mathcal{T}_b$ which map to the corresponding rank function on its shape. Attaching a leaf to $\mathcal{T}_b$ to obtain $\mathcal{T}$ does not change the number of rankings for $\mathcal{T}$ or τ. Therefore, there are 2^{q_b} rank functions on $\mathcal{T}$ which map to the given rank function on τ, and q_b is the number of non-trivial symmetric vertices of τ.

Case 3: Suppose that the two children of τ are isomorphic. Therefore they may not be distinguished except by the ranking. Therefore the shuffle at the root branch-point of $\mathcal{T}$ is only determined up to swapping the left and right subtrees. After this choice the two subtrees are distinguished: which subtree of τ is "left" and which is "right" is determined by the shuffle. The rest of the argument proceeds as before, except that this time there are $2^{q_1+q_2+1}$ rank functions on $\mathcal{T}$ which map to the given rank function on τ, and $q_1 + q_2 + 1$ is the number of non-trivial symmetric vertices of τ.

The result now follows by induction. □

The proposition directly yields:

Corollary 2.2.4. *If a probability function on ranked oriented trees is uniform on rank functions conditioned on the oriented tree, then it is also uniform on rank functions of an (unoriented) tree shape when conditioned on that (unoriented) tree shape.*

This corollary allows us to apply our runs statistic in Chapter 4 to trees which are given without orientation – ranked tree shapes.

2.3 Neutral ET models

We now define pure-chance ET models, the CAL, CRP and UR models. Some models in the CAL class have been widely used in the literature as neutral models for speciation. The CRP class allows clades to evolve with different rates. The UR class generalizes the pure-birth CAL class and the CRP class. We will use the CAL and UR class of models as neutral models for the runs statistic in Chapter 4.

2.3.1 Constant across lineages (CAL) models

We define a *constant across lineages* (CAL) model to be an ET model such that any new (speciation or extinction) event is equally likely to occur in any extant lineage. You may also think of the projection of this process onto ranked tree shapes, by forgetting the orientation of children at each internal vertex. Any model described in terms of rates is a CAL model if the speciation and extinction rates are equal between lineages at any given time. However, these rates may vary in an arbitrary fashion depending on time or the current state of the process. This class of models includes the Yule model [101], the critical branching process [2] and the constant rate birth-death process [71] which will be discusssed in detail in Chapter 3.

However, the CAL class is more general. It includes macroevolutionary models that have global speciation and extinction rate variation (i.e. the rate changes simultaneously in all species), for example due to global environmental conditions. Furthermore, it is also possible to incorporate models which take into account incomplete random taxon sampling, which is equivalent to the deletion of k species uniformly at random from the complete tree. Indeed, if the complete tree evolved under a CAL model then we simply run the model for longer with the probability of speciation set to zero and the extinction probability non-zero (and uniform across taxa). This extended model is clearly still within the CAL class.

The CAL class also includes microevolutionary models such as the coalescent with arbitrary population size history. This very simple but important fact means that the tests for non-neutral diversification described in Chapter 4 are not fooled by ancestral population size variation (as are a number of other tests in the literature).

Proposition 2.3.1. *At all times, the distribution of ranked oriented trees with n leaves is uniform under a CAL model.*

Proof. Assume that after k events, all $(m-1)!$ ranked oriented trees of size m are equally likely. If the next event is a speciation then, because the result of each (tree, speciation event) pair is distinct, after this event all $m!$ ranked oriented trees with $m+1$ leaves are equally likely. Similarly, if the next event is an extinction then for each of the $(m-1)!$ equally likely trees there are m equally likely choices for which leaf to extinguish, giving $m!$ possibilities in all. By Lemma 2.2.2 each ranked oriented tree with $m-1$ leaves results from $m(m-1)$ of these tree-plus-leaf choices. Thus each ranked oriented tree with $m-1$ leaves is equally likely, with probability $m(m-1)/m! = 1/(m-2)!$.

Since this is true for any such sequence of speciations and extinctions it is true at all times. □

Of course, any model giving the uniform distribution on ranked oriented trees with n leaves gives the uniform distribution on rank assignments given an oriented tree with n leaves. Thus we have the following corollary,

Corollary 2.3.2. *Any CAL model gives the uniform distribution on rank assignments (and thus tree shuffles) given an oriented tree.*

We have the following limited converse of Proposition 2.3.1.

Proposition 2.3.3. *Pure-birth CAL models are precisely the set of pure-birth ET models which, for any $n \geq 1$, give the uniform distribution on ranked oriented trees with n taxa when halted as soon as n taxa are present.*

Proof. By the proof of Proposition 2.3.1, pure-birth CAL models result in a uniform distribution on ranked oriented trees of size n (since there have been exactly $n-1$ events).

Now consider a model which does not satisfy the CAL condition. Assume that the k-th speciation event was the first speciation event not picked uniformly among lineages, i.e. there is a ranked tree $\mathcal{T}_0$ with lineages l_1 and l_2 which have probabilities $p_1 \neq p_2$ to speciate. Let $\mathcal{T}_1$ (respectively $\mathcal{T}_2$) be the ranked tree produced if l_1 (respectively l_2) speciates. In a pure birth process, $\mathcal{T}_1$ and $\mathcal{T}_2$ may only be reached in this way. Now

$$\mathbb{P}[\mathcal{T}_1] = \mathbb{P}[\mathcal{T}_0] \cdot p_1 \neq \mathbb{P}[\mathcal{T}_0] \cdot p_2 = \mathbb{P}[\mathcal{T}_2]$$

which shows that this model cannot give the uniform distribution on ranked trees when the process is halted after the k-th speciation event. There is only one way to build each ranked oriented tree with n leaves so the distribution on these cannot be uniform. Thus, by contradiction, there is no such k and so no such model. □

Note that in the last proposition, the restriction to a pure-birth process is needed. Consider a process with extinction where speciation is equally likely for each species but extinction is history dependent: whenever an extinction event occurs, it undoes the most recent speciation event. This model clearly does not belong to the class of CAL models. However, it gives a uniform distribution on ranked trees of some fixed size.

2.3.2 Constant relative probability (CRP) models

The motivation for the *constant relative probability* (CRP) models comes from considering the models on ranked trees which might emerge from non-selective diversification, perhaps based on physical or reproductive barriers. For example, assume we could watch a set of species emerge via allopatric (geographic) speciation, and the fundamental geographic barrier is a mountain range dividing land into two regions, A and B. These regions may differ in size or fecundity, so there may be some difference in the rate of diversification in A versus B. However, our neutral assumption for the CRP class is that the *relative* rate stays constant over time. In contrast, non-neutral models might dictate that a speciation in one region will shift the equilibrium such that further diversification in that region will become more likely ("bursting" diversification) or less likely ("refractory" diversification). Again, for convenience, we work with ranked oriented trees so we may distinguish the two children of any speciation event.

A *constant relative probability* (*CRP*) model is a pure-birth ET model with a density P on the unit interval $[0, 1]$, where each internal vertex has a real number,

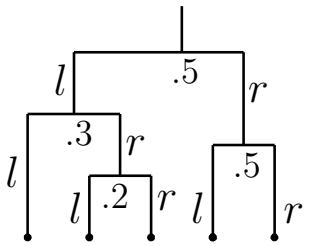

Figure 2.2: A tree which evolved under the CRP model with the bifurcation probabilities p_v for each internal vertex v. Given we see a speciation event, the probability for the leaf on path l, l (from root to leaf) to speciate next is $.5 \times .3 = .15$. Note that under the CAL model, each leaf is equally likely to speciate with probability $1/5$.

p_v associated with it. Each new speciation occurring in the clade below v occurs in L_v with probability p_v, and occurs in R_v with probability $1 - p_v$. For each new speciation event (internal vertex), v, choose the value p_v by an independent draw from P. As with CAL models, there is no constraint of any kind on waiting times between speciation events. For an example see Figure 2.2.

Proposition 2.3.4. *At all times, the distribution of rank functions on a given oriented tree is the uniform distribution under a CRP model.*

Proof. Consider the distribution of ranked oriented trees resulting from the stopped CRP. Consider a particular oriented tree, $\mathcal{T}$, with k internal vertices $v_1, \ldots, v_k$. Let n_i and m_i denote the number of internal vertices below the left and right subtrees, respectively, of vertex v_i. Consider a ranking on the tree $\mathcal{T}$. We now compute the probability of this ranked oriented tree under the model (conditional on the total number of leaves). Consider an assignment of p_{v_i} to each internal vertex v_i. Given this choice, the probability of the given ranked oriented tree is the product of the probabilities of each speciation event. For a speciation at vertex v_i, the probability of this event is the product of p_{v_j} for all v_j for which v_i lies on its left subtree times the product of $(1-p_{v_j})$ for all v_j for which v_i lies on its right subtree. In the product of these probabilities over all v_i, the term p_{v_j} occurs exactly n_j times (once for each internal vertex on the left subtree of v_j) and the term $(1 - p_{v_j})$ occurs exactly m_j times (once for each internal vertex on the right subtree of v_j). Thus, the probability of this ranked oriented tree (given the choice of p_v) is:

$$\prod_{j=1}^{k} p_{v_j}^{n_j} (1 - p_{v_j})^{m_j}$$

Note that the probability is independent of the ranking. Since the p_{v_i} are picked independently from a distribution P, the probability of the ranked oriented tree $\mathcal{T}$ is

$$\int_{p_{v_1}} \cdots \int_{p_{v_k}} \prod_{j=1}^{k} p_{v_j}^{n_j} (1 - p_{v_j})^{m_j} \, dP \cdots dP$$

which is again independent of the ranking. Therefore, all rankings of this oriented tree are equally likely. □

2.3.3 Uniform ranking (UR) models

In Proposition 2.3.3 we established that the class of pure-birth CAL models is precisly the class of pure-birth ET models which induces a uniform distribution on ranked, oriented trees.

In Proposition 2.3.4, it is shown that the CRP models induce a uniform distribution on rankings given the oriented tree. In the following, we will characterize the whole class of pure-birth models in the pure-birth ET class which induce a uniform distribution on rankings given the oriented tree, we call that class the *uniform rank (UR)* class of model. Note that each ranking being equally likely on an oriented tree is equivalent to each ranking being equally likely on a tree shape (Corollary 2.2.4). Each ranking on an oriented tree being equally likely means that each shuffle on an interior vertex is equally likely (Lemma 2.1.2). So UR models are neutral in the sense that no shuffle at an interior vertex is favored. The runs statistic introduced in Chapter 4 can test for deviation from the UR class of models.

First we note that any ET model is equivalent to the following process. If we have a tree on n leaves and add another leaf (i.e. a species speciates), we choose in which of the two subtrees below the root this leaf should be attached (call the subtree descending from the left branch the left subtree and the other one the right subtree.) In the root of the chosen subtree, we again choose one of the two subtrees. We continue until we end in a leaf. This leaf shall speciate.

We will describe the necessary and sufficient condition for the probabilities in choosing the left/right subtree below a vertex in order to obtain a uniform distribution on shuffles. This characterizes the set of UR models. A *root shuffle* of an oriented tree $\mathcal{T}$ is a shuffle at the root of $\mathcal{T}$.

Let ρ be the root (i.e. the first speciation event) in our evolving tree, with N speciation events below; n events in the left subtree and $N-n$ events in the right subtree. The sequence of l's and r's (for left and right) of successive speciation events is called the $(n, N-n)$ shuffle on ρ, we write $s_{n,N-n}$ for such a sequence. Let $p^\rho(s_{n,N-n})$ be the probability, that given N shuffle elements, we have n l's (and therefore $(N-n)$ r's) in the order $s_{n,N-n}$. In the following, we determine the set of models such that each $(n, N-n)$ shuffle on the vertex ρ is equally likely for fixed n $(0 \leq n \leq N)$, that probability is $p^\rho_{n,N-n} := p^\rho(s_{n,N-n})$ where $s_{n,N-n}$ is an arbitrary $(n, N-n)$ shuffle.

Now, let s_1, s_2 be $(n, N-n)$ shuffles, both having the same probability. Let $p^\rho_{l|s_1}$ (resp. $p^\rho_{l|s_2}$) be the probability that an l is the next shuffle element after s_1 (resp. s_2). If $p^\rho_{l|s_1} \neq p^\rho_{l|s_2}$ then the shuffles $(s_1 l)$ and $(s_2 l)$ have different probabilities. Therefore we do not have a uniform distribution on $(n+1, N-n)$ shuffles. So we assume $p^\rho_{l|s_1} = p^\rho_{l|s_2}$ for all s_1, s_2. Let $p^\rho_{l|n,N-n}$ (resp. $p^\rho_{r|n,N-n}$) be the probability that an l (resp. r) is added to an $(n, N-n)$ shuffle. For $N = 1$, each $(n, N-n-1)$ shuffle has equal probability, since we only have one $(0, 0)$ shuffle. Now let $N > 1$. For fixed

n, assume that each $(n, N-n-1)$ shuffle has equal probability $(0 \leq n \leq N-1)$. Under this assumption, any $(n, N-n)$ shuffle having equal probability is equivalent to requiring:

$$\begin{aligned} p^{\rho}_{l|n-1,N-n} p^{\rho}_{n-1,N-n} &= p^{\rho}_{r|n,N-n-1} p^{\rho}_{n,N-n-1} \\ &= (1 - p^{\rho}_{l|n,N-n-1}) p^{\rho}_{n,N-n-1}. \end{aligned}$$

Therefore the following condition is necessary and sufficient for each $(n, N-n)$ shuffle to have the same probability:

$$p^{\rho}_{l|n,N-n-1} = 1 - \frac{p^{\rho}_{n-1,N-n}}{p^{\rho}_{n,N-n-1}} p^{\rho}_{l|n-1,N-n}. \tag{2.1}$$

We may determine $p^{\rho}_{l|n,N-n-1}$ for one n where $0 \leq n < N$. W.l.o.g. determine $p^{\rho}_{l|0,N-1}$, i.e. the probability that we add a vertex to the left subtree, given the tree has zero interior vertices in the left subtree and $N-1$ in the right subtree.

The recursion in Equation (2.1) and determining $p^{\rho}_{l|0,N}$ for all N defines precisely the set of models which have a uniform distribution on the root shuffles. The probabilities for the other interior vertices are determined in the same way (each interior vertex is root of a subtree).

Since a uniform distribution on shuffles for each interior vertex is equivalent to a uniform distribution on rankings (given the oriented tree), we characterized the set of models with uniform rankings given the oriented tree.

Theorem 2.3.5. *Defining $p^{\rho}_{l|0,N} := p^{\rho}_{l}$ for all N and each $\rho \in \mathring{V}$ induces the CRP class of models.*

Proof. We proof the theorem by induction on N. Set $p^{\rho}_{r} = 1 - p^{\rho}_{l}$. For $N = 0$, we have $p^{\rho}_{l|0,0} = p^{\rho}_{l}$ by definition. Assume for all $k < N, n \leq k$ that $p^{\rho}_{l|n,k-n} = p^{\rho}_{l}$. Note that this implies $p^{\rho}_{n,k-n} = (p^{\rho}_{l})^{n} (p^{\rho}_{r})^{k-n}$ for all $k \leq N$. For $N = k$, we have for $1 \leq n \leq N$,

$$\begin{aligned} p^{\rho}_{l|n,N-n} &= 1 - \frac{p^{\rho}_{n-1,N-n+1}}{p^{\rho}_{n,N-n}} p^{\rho}_{l|n-1,N-n+1} \\ &= 1 - \frac{(p^{\rho}_{l})^{n-1} (p^{\rho}_{r})^{N-n+1}}{(p^{\rho}_{l})^{n} (p^{\rho}_{r})^{N-n}} p^{\rho}_{l|n-1,N-n+1}. \end{aligned}$$

We proceed with an induction on n. We have $p^{\rho}_{l|0,N} = p^{\rho}_{l}$ by definition. Assume $p^{\rho}_{l|m,N-m} = p^{\rho}_{l}$ for $m < n$, therefore,

$$\begin{aligned} p^{\rho}_{l|n,N-n} &= 1 - \frac{(p^{\rho}_{l})^{n-1} (p^{\rho}_{r})^{N-n+1}}{(p^{\rho}_{l})^{n} (p^{\rho}_{r})^{N-n}} p^{\rho}_{l} \\ &= 1 - p^{\rho}_{r} = p^{\rho}_{l} \end{aligned}$$

which proves the theorem. □

Theorem 2.3.6. *Defining $p^{\rho}_{l|0,N} = \frac{1}{N+2}$ for all N and each $\rho \in \mathring{V}$ induces the CAL class of models.*

Proof. Let the oriented tree $\mathcal{T}$ have $N+2$ leaves, and the left subtree under the root having $n+1$ leaves. Under the CAL model, each leaf is equally likely to speciate next, so the probability for a leaf speciating in the left subtree next is

$$p^{\rho}_{l|n,N-n} = \frac{n+1}{N+2}. \tag{2.2}$$

We will show that the UR models with $p^{\rho}_{l|0,N} = \frac{1}{N+2}$ fulfil Equation (2.2). For $N = 0$, $p^{\rho}_{l|0,0} = \frac{1}{2}$ by definition. Assume that $p_{l|n,k-n} = \frac{n+1}{k+2}$ for $k < N, n \leq k$. This implies

$$p^{\rho}_{n,k-n} = \frac{n!(k-n)!}{(k+1)!} = \frac{1}{\binom{k}{n}(k+1)}$$

for all $k \leq N, n \leq k$. For $N = k$,

$$\begin{aligned} p^{\rho}_{l|n,N-n} &= 1 - \frac{p^{\rho}_{n-1,N-n+1}}{p^{\rho}_{n,N-n}} p^{\rho}_{l|n-1,N-n+1} \\ &= 1 - \frac{N-n+1}{n} p^{\rho}_{l|n-1,N-n+1}. \end{aligned}$$

We proceed with an induction on n. We have $p^{\rho}_{l|0,N} = \frac{1}{N+2}$ by definition. Assume $p^{\rho}_{l|m,N-m} = \frac{m+1}{N+2}$ for $m < n$, therefore,

$$p^{\rho}_{l|n,N-n} = 1 - \frac{N-n+1}{n} \frac{n}{N+2} = \frac{n+1}{N+2}$$

which establishes the theorem. □

2.4 Calculating the rank distribution for a vertex

The considered neutral models for speciation in the thesis, the CAL, CRP and UR models, induce a uniform distribution on rank functions. In the next chapters, we will need the probability of a rank of a vertex in a given oriented tree $\mathcal{T}$ which evolved under one of these models. In this section, we explain how to calculate the required probability. Let r be a rank function on the oriented tree $\mathcal{T}$. For an interior vertex u of $\mathcal{T}$, define $p_u := (\mathbb{P}[r(u) = i])_{i=1,\ldots,n-1}$. In [27], we gave a formula for calculating p_u: Label the vertices on the path from the vertex u to the most recent common ancestor $mrca$ with $u = x_1, x_2, \ldots, x_m = mrca$, see Fig. 2.3. Define λ_j as the number of leaves below x_j minus 1. With that notation, we get from [27] that

$$p_u = \frac{M_{m-1} M_{m-2} \ldots M_1 e_1}{|M_{m-1} M_{m-2} \ldots M_1 e_1|_1} \tag{2.3}$$

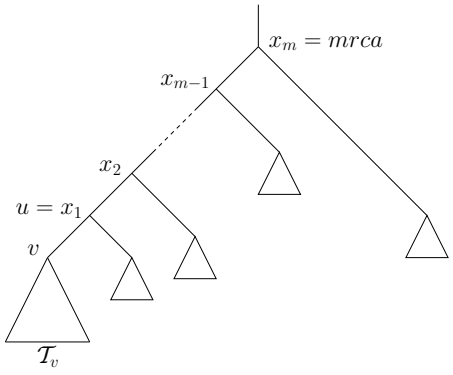

Figure 2.3: Labeling of the tree for calculating the probability for the rank of a vertex.

where $|\cdot|_1$ is the 1-norm, $e_1 = (1, 0, 0, \ldots, 0)^T$ and the matrix M_j is defined as follows,

$$(M_k)_{i,j} = \begin{cases} 0 & \text{if } j < i - 1 - (\lambda_{k+1} - \lambda_k), \\ 0 & \text{if } j > i - 1, \\ \binom{\lambda_{k+1}-i}{\lambda_{k+1}-\lambda_k-i+j+1}\binom{i-2}{i-j-1} & \text{else.} \end{cases}$$

The algorithm RankProb in [27] calculates p_u according to Equation (2.3).

For an edge $e = (u, v)$ in $\mathcal{T}$, we will need the probability $p_{u,v}(i, j) := \mathbb{P}[r(u) = i, r(v) = j]$. First, let e be an interior edge. In [27], we calculate $p_{u,v}(i, j), 1 \leq i < j \leq n - 1$ by running RankProb on different subtrees of $\mathcal{T}$. In the following, we give an expression to calculate $p_{u,v}(i, j)$ directly from $p_u(i)$ which makes the calculations faster. Let $\mathcal{T}_v$ be the daughter tree of v, see Fig. 2.3. The subtree $\mathcal{T}_v$ has n_v leaves. Let $r(\mathcal{T})$ be the set of rank functions on $\mathcal{T}$.

The number of rank functions where $r(u) = i$ is $p_u(i) \cdot |r(\mathcal{T})|$. Assume we fix the rank of the first i interior nodes, with u being the i-th node. There are $\binom{n-1-i}{n_v-1}$ possibilities to shuffle the interior vertices in $\mathcal{T}_v$ with the remaining interior vertices. Only $\binom{n-1-j}{n_v-2}$ of those shuffles assign rank j to vertex v (note that we have $n_v - 1$ interior vertices in $\mathcal{T}_v$, but we do not count vertex v since it has rank j). Overall, we therefore get for the number of rank functions with $r(u) = i$ and $r(v) = j$:

$$p_u(i) \cdot |r(\mathcal{T})| \frac{\binom{n-1-j}{n_v-2}}{\binom{n-1-i}{n_v-1}}.$$

For the probability $p_{u,v}(i, j)$, we have to divide the previous equation by the number of rank functions. Therefore

$$p_{u,v}(i, j) = p_u(i) \cdot \frac{\binom{n-1-j}{n_v-2}}{\binom{n-1-i}{n_v-1}}.$$

This is equivalent to

$$p_{u,v}(i, j) = \begin{cases} p_u(i) \frac{n_v-1}{n-n_v-i+1} \prod_{k=1}^{n_v-2} \frac{n-j-k}{n-i-k}, & \text{if } n - j + 1 \geq n_v, 1 \leq i < j < n; \\ 0, & \text{else.} \end{cases} \quad (2.4)$$

We will extend the distribution $p_{u,v}$ for leaves. Since the leaves are after the $(n-1)$-st speciation event, we can assume that all leaves have rank n. So for pendant edges, we have

$$p_{u,v}(i,n) = \begin{cases} p_u(i), & \text{if } v \text{ is a leaf;} \\ 0, & \text{else.} \end{cases} \tag{2.5}$$

Further, we will define $p_\rho, p_{\rho,v}$ for the origin ρ. The origin is always the very first vertex, the most recent common ancestor ($mrca$) is its descendant. Therefore, we define,

$$p_\rho(0) = 1, \qquad p_{\rho,v}(0,i) = p_v(i). \tag{2.6}$$

Chapter 3

Neutral models with constant rate

In the previous section, we discussed models which induce distributions on ranked trees. The time between speciation events has not been modeled. In this chapter, we discuss the most widely used model where the speciation times are specified – the CAL model with an exponential distributed time between speciation/extinction events – the constant rate birth-death process (BDP). Special cases of a BDP are the Yule model where no extinction occurs (Section 3.2) and the critical branching process where the birth and death rate are equal (Section 3.3). Since the BDP is a CAL model, the distribution on ranked oriented trees is the uniform distribution. We will now calculate the time of the speciation events. This completely describes the distribution of reconstructed oriented trees induced by the BDP.

These theoretic results on reconstructed oriented trees under the BDP can be applied to a variety of applications. For example, the latest version of the MCMC program Beast for inferring phylogenies uses our derived densities for calculating the prior distribution under the BDP (Section 3.4.3).

When reconstructing phylogenetic supertrees, we combine small phylogenetic trees to one single big phylogenetic tree. In supertrees, we often cannot infer the speciation times due to the absence of a molecular clock or due to incomplete sequences. Our formulae for the expected time of speciation events are used for dating supertrees (Section 3.1, 3.4.7 and 3.9).

Lineages-through time (LTT) plots are a popular graphical tool for comparing the reconstructed phylogeny with models. Using the analytic results for the expected time of speciation events allows us to draw these LTT plots without simulations (Section 3.6). Since the data is often incomplete, i.e. some species of a clade are missing, our method for drawing LTT plots is extended to the scenario of incomplete taxon sampling (Section 3.7).

Even though we have a lot of analytic results for the BDP, some complex questions can still only be answered via simulations. In Chapter 5 we provide accurate algorithms to sample under general speciation models. With the theoretic results of this chapter, we develop a more efficient algorithm for the BDP models (Section 5.1.5).

We start this chapter by giving the general idea of calculating the time of speci-

ation events under the BDP.

3.1 Calculating the time of speciation events

Let an oriented tree $\mathcal{T}$ evolve under a CAL model with a specified time between speciation and extinction events. Let $\mathcal{A}_{\mathcal{T}}^e$ be the random variable 'length of edge e in the oriented tree $\mathcal{T}$' with density function $f_{\mathcal{A}_{\mathcal{T}}^e}$. Let $\mathcal{A}_{\mathcal{T}}^u$ be the random variable 'time of the speciation event u in the oriented tree $\mathcal{T}$' with density function $f_{\mathcal{A}_{\mathcal{T}}^u}$. In an oriented tree with n species, let $\mathcal{A}_n^{i,j}$ be the random variable 'time between the $i-$th and the $j-$th speciation event' with density function $f_{\mathcal{A}_n^{i,j}}$. Let $\mathcal{A}_n^k$ be the random variable 'time of the $k-$th speciation event' with density function $f_{\mathcal{A}_n^k}$. With $k = 0$, we denote the origin of the tree. With $k = n$, we denote the present, i.e. the time of the leaves. Time is set zero at the present, and is increasing going into the past. In the following, we calculate the density and expectation for $\mathcal{A}_{\mathcal{T}}^e, \mathcal{A}_{\mathcal{T}}^u, \mathcal{A}_n^{i,j}, \mathcal{A}_n^k$.

The CAL models induces a uniform distribution on ranked oriented trees on n species as shown in Proposition 2.3.1. We use the notation from Section 2.4: Let $p_u(i)$ be the proability that vertex u in the oriented tree $\mathcal{T}$ has rank i; let $p_{u,v}(i,j)$ be the probability that vertex u has rank i and vertex v has rank j. For a CAL model, where the speciation times $\mathcal{A}_n^k$ are independent of a given ranked oriented tree on n species, we can calculate the density and expectation of the random variables $\mathcal{A}_n^k, \mathcal{A}_n^{i,j}, \mathcal{A}_{\mathcal{T}}^u, \mathcal{A}_{\mathcal{T}}^e$ in the following way. For the density, we have, with $e = (u, v)$,

$$f_{\mathcal{A}_{\mathcal{T}}^u}(s) = \sum_{i=0}^{n-1} f_{\mathcal{A}_n^i}(s) p_u(i); \tag{3.1}$$

$$f_{\mathcal{A}_{\mathcal{T}}^e}(s) = \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} f_{\mathcal{A}_n^{i,j}}(s) p_{u,v}(i,j). \tag{3.2}$$

Note that for interior edges, $p_{u,v}(i,n) = 0$; for pendant edges (except of the root edge), $p_{u,v}(i,j) = 0$ for $j < n$; and for the root edge, $p_{u,v}(0,1) = 1$. For the expectation, we obtain from $\mathbb{E}[\mathcal{A}_n^k]$,

$$\mathbb{E}[\mathcal{A}_{\mathcal{T}}^u] = \sum_{i=0}^{n-1} \mathbb{E}[\mathcal{A}_n^i] p_u(i); \tag{3.3}$$

$$\mathbb{E}[\mathcal{A}_n^{i,j}] = \mathbb{E}[\mathcal{A}_n^i] - \mathbb{E}[\mathcal{A}_n^j], \qquad 0 \le i < j \le n; \tag{3.4}$$

$$\mathbb{E}[\mathcal{A}_{\mathcal{T}}^e] = \mathbb{E}[\mathcal{A}_{\mathcal{T}}^u] - \mathbb{E}[\mathcal{A}_{\mathcal{T}}^v] = \sum_{i=0}^{n} \mathbb{E}[\mathcal{A}_n^i](p_u(i) - p_v(i)). \tag{3.5}$$

Note that $\mathcal{A}_n^n$ is the present. Hence, $\mathbb{E}[\mathcal{A}_n^n] = 0$.

We can calculate the probabilities $p_u(i), p_{u,v}(i,j)$ as described in Section 2.4, Equations (2.3)–(2.6), so it is left to calculate $f_{\mathcal{A}_n^{i,j}}(s)$, $f_{\mathcal{A}_n^k}(s)$ and $\mathbb{E}[\mathcal{A}_n^k]$. The three values will be computed both conditioning and not conditioning on the age of the tree. This is done for the Yule model in Section 3.2 and for the BDP in Section 3.4.

Knowing the expectation of $\mathcal{A}_n^k$, we can calculate expected LTT plots which is done in Section 3.6. Further, we can date phylogenies as explained in the next section.

3.1.1 Application: Estimating divergence times

Analytic results for the expectation of $\mathcal{A}_T^v$ find application in supertree methods. Supertree methods infer a phylogeny by combining many small phylogenies. With the supertree methods, we are usually not able to date all speciation events. For undated nodes, estimates are required. Standard procedures assume the Yule model as the model for speciation in the tree. In the primate phylogeny [75] and the mammal phylogeny [7], the time of the undated speciation events is estimated – assuming the Yule model – in the following way. Let u be the undated vertex, and let the time of birth of the direct ancestor of u be t_a. Let the clade size of the ancestor be c_a and the clade size of u be c_u. Then we estimate the date of u, t_u, as $t_u = t_a \log c_u / \log c_a$. The motivation for this estimate is given in [75]. This estimate has a bias though. Iteratively estimating nodes with the described procedure biases the model to have a slow-down in the diversification rate [98]. In [98], an undated speciation event is estimated via the expectation of the time of that vertex under the Yule model. The expectation is obtained via simulations. The advantage of taking the expectation of the speciation event as an estimate is, that it is not biased toward a slow-down in the diversification rate [98]. Arne Mooers and Rutger Vos asked for an analytic approach (personal communication). In Section 3.4.5, we calculate the expectation of $\mathcal{A}_n^k$ under any BDP which yields the expectation of $\mathcal{A}_T^v$ with Equation (3.3). This analytic method for dating supertrees is coded as part of the Python package Cass [89]. In Section 3.9, the method for dating trees is extended to trees where some dates are known; we use the known dates in a tree in order to estimate the unknown dates.

3.2 The Yule model

The simplest and most widely used null model for speciation is the Yule model [3, 18, 101] which G.U. Yule introduced in 1924 for modeling the size of genera. Under the Yule model, no extinction occurs and each species speciates after an exponential (rate λ) waiting time. The Yule model is often used as a null model, even though extinction clearly occurs in nature. But being a pure-birth model with exponential waiting times, the Yule model is relatively simple to analyze which makes it attractive to use. For example, common procedures for estimating the time of undated divergence times in supertrees assume the Yule model [7, 75, 98], even though extinction clearly occurred in the considered phylogenies.

Under the Yule model, each species has an exponential (rate λ) lifetime. Since the exponential distribution is memoryless, this is equivalent to having an exponential (rate $k\lambda$) waiting time in a tree with k species until the next speciation event: After the $(k-1)$-th speciation event, we have k extant species with independent

exponential (rate λ) distributed lifetimes $S_1, S_2, \ldots, S_k$. The random variable $\mathcal{A}^{k-1,k}$, the time between the $(k-1)$-th and the k-th speciation event (without conditioning on observing n species today), is distributed as follows,

$$\mathbb{P}[\mathcal{A}^{k-1,k} \geq t] = \mathbb{P}[S_1, S_2, \ldots, S_k \geq t] = \prod_{j=1}^{k} \mathbb{P}[S_j \geq t] = e^{-\lambda k t}.$$

The density function $f_{\mathcal{A}^{k-1,k}}(t)$ is therefore

$$f_{\mathcal{A}^{k-1,k}}(t) = \frac{d}{dt}(1 - \mathbb{P}[\mathcal{A}^{k-1,k} \geq t]) = \lambda k e^{-\lambda k t}$$

which is the exponential (rate λk) distribution. This yields

$$\mathbb{E}[\mathcal{A}^{k-1,k}] = \frac{1}{\lambda k}, \qquad \mathrm{Var}[\mathcal{A}^{k-1,k}] = \frac{1}{(\lambda k)^2}.$$

Whenever a speciation event occurs, each extant species is equally likely to speciate next. Therefore the Yule model belongs to the class of CAL models. We will set $\lambda = 1$ in the following. Note that from results for $\lambda = 1$, we get results for a general λ via the following property. Let $F_\lambda(t)$ be the exponential (rate λ) distribution, the lifetime of a species. We have

$$F_\lambda(t) = e^{-\lambda t} = F_1(\lambda t).$$

Therefore, changing from rate λ to 1 is scaling time by a factor of λ.

Since the exponential distribution is memoryless, at any point in time of a Yule process, the time until the next speciation event is independent of the current ranked oriented tree. This establishes the following lemma,

Lemma 3.2.1. *For a given Yule tree on n species, the ranked, oriented tree is independent of the time of the speciation events.*

Therefore, we can use Equations (3.1),(3.2),(3.3),(3.4),(3.5) in order to calculate the density of $\mathcal{A}^e_{\mathcal{T}}, \mathcal{A}^u_{\mathcal{T}}$ and the expectation of $\mathcal{A}^e_{\mathcal{T}}, \mathcal{A}^u_{\mathcal{T}}, \mathcal{A}^{i,j}_n$. In the following, we calculate the distribution of $\mathcal{A}^k_n, \mathcal{A}^{i,j}_n$, and the expectation of $\mathcal{A}^k_n$. In Section 3.2.1, we assume a uniform prior for the time since origin, and in Section 3.2.2, we condition on the age t of the tree.

3.2.1 Unknown age of the tree

First, we do not condition on the age of the tree, t, but assume a uniform prior for the age of the tree. A first ancestor species was created at any point in the past with equal probability. Since we want to obtain n species today, the time of origin has to be conditioned to obtain n species today.

This is equivalent to growing a tree and stopping the process at a time between the $(n-1)$-th and n-th speciation event. Therefore, the waiting time between the

$(k-1)$-th speciation event and the k-th speciation event in a tree which has n species today is $\mathcal{A}_n^{k-1,k} = \mathcal{A}^{k-1,k}$ for $k < n$. The stopping time is independent of the realization up to the $(n-1)$-th speciation event due to the memoryless exponential distribution. We need to determine $\mathcal{A}_n^{n-1}$, the time between the $(n-1)$-th speciation event and today. The time between the $(n-1)$-th and n-th speciation event, $\mathcal{A}^{n-1,n}$, is exponential (rate n) distributed. Note that the probability of observing a particular tree on n species with $\mathcal{A}^{n-1,n} = \sigma_n$ is proportional to σ_n (if σ_n is very small, the n-th speciation event occured very soon after the $(n-1)$-th speciation event, and therefore, it is unlikely to observe the tree with exactly n species being present).

Since we assume a uniform prior, we stop at a time chosen uniformly at random between the $(n-1)$-th and n-th speciation event, each stopping time has probability $1/\sigma_n$.

Overall, we therefore obtain by integrating over all possible σ_n,

$$f_{\mathcal{A}_n^{n-1}}(s) \propto \int_s^\infty \sigma_n f_{\mathcal{A}^{n-1,n}}(\sigma_n) \frac{1}{\sigma_n} d\sigma_n = e^{-ns},$$

therefore $\mathcal{A}_n^{n-1}$ is the exponential (rate n) distribution. The time between the $(n-1)$-th speciation event and today equals the time between the $(n-1)$-th speciation event and the n-th speciation event. The same result is derived in a different way in Remark 3.4.12. This issue is discussed in detail for general models in Chapter 5.

This establishes, $\mathcal{A}_n^k = \sum_{i=k+1}^n \mathcal{A}^{i-1,i}$ and,

Theorem 3.2.2. *The expectation of $\mathcal{A}_n^k$ is ($0 \le i \le n-1$),*

$$\mathbb{E}[\mathcal{A}_n^k] = \mathbb{E}[\sum_{i=k+1}^n \mathcal{A}^{i-1,i}] = \sum_{i=k+1}^n \frac{1}{i}. \tag{3.6}$$

For the variance, we get, since $\mathcal{A}^{i-1,i}, \mathcal{A}^{j-1,j}$ are independent for $i \neq j$,

$$\mathrm{Var}[\mathcal{A}_n^k] = \mathrm{Var}[\sum_{i=k+1}^n \mathcal{A}^{i-1,i}] = \sum_{i=k+1}^n \frac{1}{i^2}.$$

For the second moment, we get

$$\mathbb{E}[(\mathcal{A}_n^k)^2] = \mathrm{Var}[\mathcal{A}_n^k] + (\mathbb{E}[\mathcal{A}_n^k])^2 = \sum_{i=k+1}^n \frac{1}{i^2} + \sum_{i=k+1}^n \sum_{j=k+1}^n \frac{1}{ij}.$$

Define $Y_i^j := \sum_{k=i}^j \mathcal{A}_n^{k-1,k}$. Note that $\mathcal{A}_n^k = Y_{k+1}^n$ and $\mathcal{A}_n^{i,j} = Y_{i+1}^j$. In the following, we will obtain the density function for Y_i^j.

Let X, Y be independent non-negative random variables. Then the density of $Z = X + Y$ is the convolution of X, Y:

$$f_Z(s) = \int_0^s f_X(\tau) f_Y(s-\tau) d\tau.$$

In [59], a formula for the convolution of n exponential distributed random variables is established. Note that $Y_i^j := \sum_{k=i}^{j} \mathcal{A}_n^{k-1,k}$ is a convolution of $j-i+1$ exponential distributed random variables. From the general formula for the convolution in [59], we obtain for our setting

$$f_{Y_i^j}(s) = i \cdot (i+1) \cdot \ldots \cdot j e^{-js} \varphi_{j-i+1}(s) \tag{3.7}$$

where $\varphi_n(s) = \int_0^s e^x \varphi_{n-1}(x)dx$ and $\varphi_1(s) = 1$. We need the following lemma for obtaining a closed form for the density function $f_{Y_i^j}(s)$ in our setting.

Lemma 3.2.3. *With the notation above, we have,*

$$\varphi_n(s) = \frac{1}{(n-1)!}(e^s - 1)^{n-1}. \tag{3.8}$$

Proof. We prove this lemma by induction on n. The formula is true for $n = 1$, and if it holds for an arbitrary specific value of n, then it also holds for the next value, since

$$\begin{aligned} \varphi_{n+1}(s) &= \int_0^s e^x \varphi_n(x)dx = \frac{1}{(n-1)!} \int_0^s e^x (e^x - 1)^{n-1} dx \\ &= \frac{1}{(n-1)!} \left[\frac{(e^x - 1)^n}{n} \right]_0^s = \frac{(e^s - 1)^n}{n!}. \end{aligned}$$

□

Lemma 3.2.4. *The density of Y_i^j is*

$$f_{Y_i^j}(s) = i \binom{j}{i} e^{-js} (e^s - 1)^{j-i}.$$

Proof. With Equation (3.7) and (3.8) we obtain

$$\begin{aligned} f_{Y_i^j}(s) &= i \cdot (i+1) \cdot \ldots \cdot j e^{-js} \frac{1}{(j-i)!} (e^s - 1)^{j-i} \\ &= i \binom{j}{i} e^{-js} (e^s - 1)^{j-i}. \end{aligned}$$

□

Since $\mathcal{A}_n^k = Y_{k+1}^n$, we have established the following theorem.

Theorem 3.2.5. *The density of $\mathcal{A}_n^k$ is,*

$$f_{\mathcal{A}_n^k}(s) = (k+1) \binom{n}{k+1} e^{-ns} (e^s - 1)^{n-k-1},$$

where $0 \leq i \leq n-1$.

Since $\mathcal{A}_n^{i,j} = Y_{i+1}^j$, we have established the following theorem.

Theorem 3.2.6. *The density of $\mathcal{A}_n^{i,j}$ is,*

$$f_{\mathcal{A}_n^{i,j}}(s) = (i+1)\binom{j}{i+1}e^{-js}(e^s-1)^{j-i-1}, \tag{3.9}$$

where $0 \leq i < j \leq n$.

Remark 3.2.7. Note that $\mathcal{A}_n^{k-1,k}, \mathcal{A}_n^{l-1,l}$ are independent for $k \neq l$. This implies that $\mathcal{A}_n^{i,j}$ and $\mathcal{A}_n^{k,l}$ are independent for $i < j \leq k < l$.

3.2.2 Known age of the tree

Next we calculate the density and expectation of $\mathcal{A}_n^k$ and the density of $\mathcal{A}_n^{i,j}$ conditioned on the age t of the tree.

Theorem 3.2.8. *The random variable $\mathcal{A}_n^k$ ($1 \leq k \leq n-1$) conditioned on the age t of the Yule tree has the density function:*

$$f_{\mathcal{A}_n^k}(s|t) = k\binom{n-1}{k}(1-e^{-t})^{1-n}e^{-ks}(1-e^{-s})^{n-k-1}(1-e^{-(t-s)})^{k-1}.$$

For $k = 0$, we have $f_{\mathcal{A}_n^0}(s|t) = \delta(s-t)$ were δ is the Dirac delta function.

Proof. The distribution for $k = 0$ is obvious, since we condition on the age t of the tree. So now let $k > 0$. Note that $\mathcal{A}_n^{0,k}$ and $\mathcal{A}_n^{k,n}$ are independent (Remark 3.2.7). Denote the joint density function of $\mathcal{A}_n^{0,k}, \mathcal{A}_n^{k,n}$ as $f_{\mathcal{A}_n^{0,k},\mathcal{A}_n^{k,n}}$, and apply Equation (3.9) to obtain:

$$\begin{aligned}
f_{\mathcal{A}_n^k}(s|t) &= f_{\mathcal{A}_n^{k,n}}(s|\mathcal{A}_n^{0,n} = t) \\
&= \frac{f_{\mathcal{A}_n^{k,n},\mathcal{A}_n^{0,k}}(s, t-s)}{f_{\mathcal{A}_n^{0,n}}(t)} \\
&= \frac{f_{\mathcal{A}_n^{k,n}}(s) f_{\mathcal{A}_n^{0,k}}(t-s)}{f_{\mathcal{A}_n^{0,n}}(t)} \\
&= \frac{(k+1)\binom{n}{k+1}e^{-ns}(e^s-1)^{n-k-1}ke^{-k(t-s)}(e^{t-s}-1)^{k-1}}{ne^{-nt}(e^t-1)^{n-1}} \\
&= k\binom{n-1}{k}(1-e^{-t})^{1-n}e^{-ks}(1-e^{-s})^{n-k-1}(1-e^{-(t-s)})^{k-1}.
\end{aligned} \tag{3.10}$$

□

Theorem 3.2.9. *The expectation of $\mathcal{A}_n^k$ (for $1 \leq k \leq n-1$) conditioned on t is:*

$$\mathbb{E}[\mathcal{A}_n^k|t] = \sum_{k_1=0}^{n-k-1}\sum_{k_2=0}^{k-1} B_{k_1,k_2}(1-e^{-t})^{1-n}(e^{-k_2 t} - ((k+k_1-k_2)t+1)e^{-(k+k_1)t})$$

with $B_{k_1,k_2} := k\binom{n-1}{k}\binom{n-k-1}{k_1}\binom{k-1}{k_2}(-1)^{k_1+k_2}(k+k_1-k_2)^{-2}$.

For $k = 0$, we have $\mathbb{E}[\mathcal{A}_n^0|t] = t$.

Proof. The expectation for $k = 0$ is obvious since we condition on the age t of the tree. Now let $k > 0$. We have

$$\begin{aligned}\mathbb{E}[\mathcal{A}_n^k|t] &= \int_0^t s f_{\mathcal{A}_n^k}(s|t)ds \\ &= k\binom{n-1}{k}(1-e^{-t})^{1-n}\sum_{k_1=0}^{n-k-1}\sum_{k_2=0}^{k-1}\binom{n-k-1}{k_1}\binom{k-1}{k_2}\times \\ &\quad (-1)^{k_1+k_2}e^{-k_2 t}\int_0^t s e^{-(k+k_1-k_2)s}ds.\end{aligned}$$

Using integration by parts,

$$\int_a^b se^{-cs}ds = [-\frac{s}{c}e^{-cs}]_a^b + \frac{1}{c}\int_a^b e^{-cs}ds = -\frac{1}{c^2}[(sc+1)e^{-cs}]_a^b.$$

Therefore, with $B_{k_1,k_2} := k\binom{n-1}{k}\binom{n-k-1}{k_1}\binom{k-1}{k_2}(-1)^{k_1+k_2}(k+k_1-k_2)^{-2}$, we have:

$$\mathbb{E}[\mathcal{A}_n^k|t] = \sum_{k_1=0}^{n-k-1}\sum_{k_2=0}^{k-1} B_{k_1,k_2}(1-e^{-t})^{1-n}(e^{-k_2 t} - ((k+k_1-k_2)t+1)e^{-(k+k_1)t})$$

which establishes the theorem. □

Theorem 3.2.10. *The random variable $\mathcal{A}_n^{i,j}$ $(1 \le i < j \le n-1)$ conditioned on the age t has the density function:*

$$f_{\mathcal{A}_n^{i,j}}(s|t) = \sum_{k_1=0}^{i-1}\sum_{k_2=0}^{n-j-1} B_{k_1,k_2}e^{(n-j)s}\frac{(e^s-1)^{j-i-1}}{(e^t-1)^{n-1}}(e^{(n-i+k_1)(t-s)} - e^{k_2(t-s)})$$

with $B_{k_1,k_2} = i(i+1)\binom{j}{i+1}\binom{n-1}{j}\binom{i-1}{k_1}\binom{n-j-1}{k_2}\frac{(-1)^{n+i-j-k_1-k_2}}{n-i+k_1-k_2}$.

For $\mathcal{A}_n^{i,n}$ with $i < n$, we have $f_{\mathcal{A}_n^{i,n}}(s|t) = f_{\mathcal{A}_n^i}(s|t)$.

For $\mathcal{A}_n^{0,j}$ we have $f_{\mathcal{A}_n^{0,j}}(s|t) = f_{\mathcal{A}_n^j}(t-s|t)$.

Proof. First, consider $1 \le i < j \le n-1$. We can write $f_{\mathcal{A}_n^{i,j}}(s|t)$ as

$$\begin{aligned}f_{\mathcal{A}_n^{i,j}}(s|t) &= f_{\mathcal{A}_n^{i,j}}(s|\mathcal{A}_n^{0,n}=t) \\ &= \frac{f_{\mathcal{A}_n^{i,j},\mathcal{A}_n^{0,n}}(s,t|n)}{f_{\mathcal{A}_n^{0,n}}(t)} \\ &= \frac{f_{\mathcal{A}_n^{i,j},\mathcal{A}_n^{0,i}+\mathcal{A}_n^{j,n}}(s,t-s)}{f_{\mathcal{A}_n^{0,n}}(t)} \\ &= \frac{f_{\mathcal{A}_n^{i,j}}(s)f_{\mathcal{A}_n^{0,i}+\mathcal{A}_n^{j,n}}(t-s)}{f_{\mathcal{A}_n^{0,n}}(t)}. \qquad (3.11)\end{aligned}$$

The last equality holds since $\mathcal{A}_n^{i,j}$ and $\mathcal{A}_n^{0,i} + \mathcal{A}_n^{j,n}$ are independent, because the $\mathcal{A}_n^{k-1,k}$ are independent (see Remark 3.2.7). We will now obtain an expression for $f_{\mathcal{A}_n^{0,i}+\mathcal{A}_n^{j,n}}(t-s)$. The random variables $\mathcal{A}_n^{0,i}, \mathcal{A}_n^{j,n}, j \geq i$ are independent (see Remark 3.2.7). So $f_{\mathcal{A}_n^{0,i}+\mathcal{A}_n^{j,n}}(t-s)$ is the convolution of $\mathcal{A}_n^{0,i}, \mathcal{A}_n^{j,n}$,

$$\begin{aligned}
f_{\mathcal{A}_n^{0,i}+\mathcal{A}_n^{j,n}}(t-s) &= \int_0^{t-s} f_{\mathcal{A}_n^{0,i}}(u) f_{\mathcal{A}_n^{j,n}}(t-s-u)du \\
&= \int_0^{t-s} ie^{-iu}(e^u-1)^{i-1}(j+1) \times \\
&\quad \binom{n}{j+1} e^{-n(t-s-u)}(e^{t-s-u}-1)^{n-j-1}du \\
&= i(j+1)\binom{n}{j+1} e^{-n(t-s)} \sum_{k_1=0}^{i-1} \sum_{k_2=0}^{n-j-1} \binom{i-1}{k_1}\binom{n-j-1}{k_2} \times \\
&\quad \frac{(-1)^{n+i-j-k_1-k_2}}{n-i+k_1-k_2} e^{k_2(t-s)}(e^{(n-i+k_1-k_2)(t-s)}-1). \qquad (3.12)
\end{aligned}$$

Equation (3.11) combined with Equations 3.12 and 3.9 gives the formula described in Theorem 3.2.10.

For $j = n$, we have by definition $f_{\mathcal{A}_n^{i,n}}(s|t) = f_{\mathcal{A}_n^i}(s|t)$. The time between $i = 0$ and j is the age of the tree minus the time from today to the j-th speciation event. This is $t - \mathcal{A}_n^j$, which completes the proof. □

3.3 The critical branching process (CBP)

The *constant rate critical branching process* (CBP) as a model for speciation and extinction has first been introduced in [74]. This process generalizes the Yule model by including extinction. The CBP is a one-parameter process operating as follows. We start with one species at some time t in the past, the time of origin. A species has an exponential (rate λ) lifetime, in the course of which it gives birth to new species at Poisson (rate λ) times. Different branches of the tree behave independently. Since the time until speciation (extinction) of an extant species is exponentially (rate λ) distributed, an individual is equally likely to die or to speciate next. Further, since the exponential distribution is memoryless, each species is equally likely to be the next undergoing a speciation (extinction) event. Therefore the CBP belongs to the CAL class of models.

The CBP with conditioning on having n extant individuals today is called a conditioned constant rate critical branching process (cCBP). The asymptotic behavior of the cCBP for large n has been analyzed [2, 74]. The authors point out that the model is of biological significance mainly for small n (i.e. for small clades). They calculate the distribution and expectation of $\mathcal{A}_n^1$, the time of the most recent common ancestor of n extant species. We extend this idea and calculate the distribution and all moments of $\mathcal{A}_n^k$ $(k = 1, \ldots, n-1)$ in Section 3.4 as a special case of the cBDP model.

The probability of a CBP going extinct is 1 (since the process is critical). In the remainder of the section, we investigate the realizations of the CBP which went extinct. For CBP trees on n extant species, we have a uniform distribution on ranked, oriented trees (Proposition 2.3.1). For extinct CBP trees on n extinct species, we establish a uniform distribution on oriented trees.

3.3.1 Extinct trees under the CBP

We calculate the distribution on oriented trees on n extinct species and no extant species which evolved under the CBP. For the remainder of this section, let T be an extinct realization of the CBP where we do not keep track of the timing of the speciation and extinction events. Therefore T is an oriented tree. Let V_T be the number of leaves in T.

Lemma 3.3.1. *Let T be an extinct realization of the CBP as defined above. We have,*

$$\mathbb{P}[V_T = n] = (1/2)^{2n-1} c_{n-1}$$

with $c_n = \frac{1}{n+1}\binom{2n}{n}$ being the n-th Catalan number.

Proof. We prove the statement by induction on n. The probability that the tree goes extinct with $n = 1$ lineages has probability $1/2$ (since the probability of extinction of a lineage is $1/2$ under the CBP) .

Let the lemma be true for all $k < n$. For an extinct tree T with $V_T = n$, let L be the left daughter tree, and R be the right daugther tree. The number of leaves of L shall be k, where $1 \leq k < n$. We have,

$$\mathbb{P}[V_T = n] \;=\; \frac{1}{2}\sum_{k=1}^{n-1} \mathbb{P}[V_L = k]\mathbb{P}[V_R = n-k],$$

which is the probability of the first lineage speciating $(1/2)$ times the probability that the left tree has k leaves and the right tree has $n-k$ leaves. The product is summed over all possible k. With our induction assumption, we have,

$$\begin{aligned}\mathbb{P}[V_T = n] &= (1/2)^{2n-1}\sum_{k=1}^{n-1} c_{k-1}c_{n-k-1} \\ &= (1/2)^{2n-1} c_{n-1}\end{aligned}$$

where the last equation follows from the recursion for Catalan numbers (see for example [83]),

$$c_{n+1} = \sum_{k=0}^{n} c_k c_{n-k}.$$

□

Theorem 3.3.2. *We have*

$$\mathbb{P}[T|V_T = n] = \frac{n}{\binom{2(n-1)}{n-1}}$$

which is the uniform distribution on oriented trees with n leaves.

Proof. We will establish the theorem by induction on n. For $n = 1$, we only have one tree, an edge from the root to the leaf. Its probability is 1.

For $n > 1$, denote the left daughter tree of T with L and the right daughter tree with R, we define $k := V_L$ and therefore $V_R = n - k$. We have,

$$\begin{aligned}\mathbb{P}[T|V_T = n] &= \mathbb{P}[L, R|V_T = n] \\ &= \mathbb{P}[L, R|V_L = k, V_R = n-k]\mathbb{P}[V_L = k, V_R = n-k|V_T = n].\end{aligned}$$

Since L and R evolve independently,

$$\mathbb{P}[T|V_T = n] = \mathbb{P}[L|V_L = k]\mathbb{P}[R|V_R = n-k]\frac{\mathbb{P}[V_L = k, V_R = n-k, V_T = n]}{\mathbb{P}[V_T = n]}.$$

We have

$$\mathbb{P}[V_L = k, V_R = n-k, V_T = n] = \mathbb{P}[V_L = k, V_R = n-k] = \frac{1}{2}\mathbb{P}[V_L = k]\mathbb{P}[V_R = n-k]$$

where $1/2$ is the probability that the first speciation event occurs such that L and R can evolve. Overall, by the induction assumption and Lemma 3.3.1,

$$\begin{aligned}\mathbb{P}[T|V_T = n] &= \frac{1}{2}\mathbb{P}[L|V_L = k]\mathbb{P}[R|V_R = n-k]\frac{\mathbb{P}[V_L = k]\mathbb{P}[V_R = n-k]}{\mathbb{P}[V_T = n]} \\ &= \frac{n}{\binom{2(n-1)}{n-1}}\end{aligned}$$

which proves the theorem. □

Note that for CBP trees on n extant species, we have the uniform distribution on ranked oriented trees since the CBP belongs to the CAL class of models. For CBP trees on n extinct species, we now established the uniform distribution on oriented trees. Note that the same distribution is induced under the PDA model (see Section 1.2).

3.4 The constant rate birth-death process (BDP)

In this section, we consider a generalization of the Yule model and the CBP model, the *constant rate birth-death process (BDP) model*. The BDP model is a two-parameter model, where each species has an exponential (rate μ) lifetime during which it produces new species according to an exponential (rate λ) distribution

(with $\lambda \geq \mu \geq 0$). In a tree with k species, this is equivalent to an exponential distributed (rate $k\lambda$) waiting time until the next speciation event and an exponential distributed (rate $k\mu$) waiting time until the next extinction event. If we observe a speciation / extinction event, each species is equally likely to be the one speciating / going extinct. This is due to the memoryless exponential distribution (see Section 3.2 for a detailed discussion). Therefore, the BDP belongs to the class of CAL models. We condition the BDP to have n species at the present, this is called the conditioned birth-death process (cBDP). Since the BDP and therefore also the cBDP belong to the CAL class of models, we establish with Proposition 2.3.1 that the distribution on ranked oriented trees induced by the cBDP is the uniform distribution. In [95], it is established, that the distribution on speciation times is independent of the ranked, oriented tree. This is due to the memoryless exponential distribution. Therefore Equations (3.1),(3.2),(3.3),(3.4),(3.5) hold for the cCBP. In this section, we calculate the density of $\mathcal{A}_n^k, \mathcal{A}_n^{i,j}$ and the expectation of $\mathcal{A}_n^k$ for the cCBP. In order to do so, we introduce the *point process representation* of a cBDP.

In [71], the reconstructed tree of a BDP after time t is discussed. We condition on having n extant species, since this allows us to compare the model with phylogenies on n extant species. The age t of the tree is either fixed, or a uniform prior is assumed. We will obtain the density for each speciation event in a tree with n species. The joint probability for all speciation times and the shape as well as conditioning on the shape has been established in [95]. However, no individual probabilities have been established. For establishing the individual probabilities, we introduce the point process representation for reconstructed trees (Section 3.4.1). This had been done for the critical branching process in [2, 74].

The results are used for dating phylogenies (Section 3.1.1), for obtaining LTT plots (Section 3.6), as a prior in the MCMC program Beast (Section 3.4.3) and for simulating trees (Chapter 5).

3.4.1 The point process

It will be very convenient to use a *point process representation* for reconstructed trees in order to establish various theorems in this chapter. The following point process has first been considered in connection with trees in [2, 74].

Definition 3.4.1. A point process in $\mathbb{R}^2$ of size n and age t_{or} is defined as follows. We have n points $(0,1),(0,2),\ldots,(0,n)$. Further, we have $n-1$ points at $(i+1/2, s_i)$, $i = 1,2,\ldots,(n-1)$ with $0 < s_i < t_{or}$ and $s_i \neq s_j$ for $i \neq j$.

Lemma 3.4.2. *We have a bijection between reconstructed oriented trees of age t_{or} and the point process of age t_{or}.*

Proof. "Draw" the given reconstructed oriented tree from the top (origin) to the bottom (leaves). At each speciation event, choose the branch with label *right* to be on the right and with label *left* on the left. The leaves are located at position $(0,1),(0,2),\ldots(0,n)$ from left to right. Define s_i to be the time of the speciation event having leaf at position $(0,i)$ as a descendant in the left daughter tree and

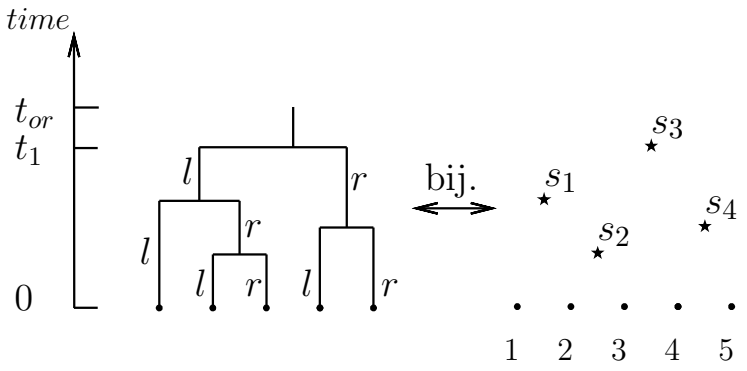

Figure 3.1: Reconstructed oriented tree and the corresponding point process. The time of origin of the process is t_{or}, the time of the most recent common ancestor is t_1.

leaf at position $(0, i+1)$ in the right daughter tree. This defines $(i+1/2, s_i)$, $i = 1, 2, \ldots, (n-1); 0 < s_i < t_{or}$, the $n-1$ points of the point process, see also Figure 3.1. The mapping to the point process is obviously injective and surjective, i.e. bijective. □

For completeness, we give the mapping from the point process to the reconstructed oriented trees. Consider a realization of the point process. Connect the most recent "speciation event" (the smallest s_i) with its two neighboring leaves (i.e. if $s_j < s_i$ for all $i \neq j$ then connect s_j with $(0, j), (0, j+1)$). This most recent speciation event replaces the two neighboring leaves in the leaf set. Continue in this way until all points are connected. This gives us the corresponding reconstructed oriented tree. An example of the point process is given in Figure 3.1.

Corollary 3.4.3. *Each permutation of the $n-1$ speciation points $s_1, \ldots, s_{n-1}$ of the point process induced by the cBDP has equal probability.*

Proof. We have a bijection between the reconstructed oriented trees and the point process (Lemma 3.4.2). Choosing the $n-1$ speciation points $s_1, \ldots, s_{n-1}$ induces a ranked oriented tree, let the probability of that tree be p. Now permute the $n-1$ speciation points s_i arbitrary. This induces a different ranked oriented tree. Since we have a uniform distribution on ranked oriented trees (Proposition 2.3.1), the probability of the new tree is again p. So each permutation is equally likely. □

For obtaining the density of the speciation time s_i, we need the following results. Under a birth-death process, the probability that a lineage leaves n descendants

after time t is $p_n(t)$. From [48], we know for $\lambda > \mu \geq 0$,

$$\begin{aligned} p_0(t) &= \frac{\mu(1-e^{-(\lambda-\mu)t})}{\lambda-\mu e^{-(\lambda-\mu)t}}, \\ p_1(t) &= \frac{(\lambda-\mu)^2 e^{-(\lambda-\mu)t}}{(\lambda-\mu e^{-(\lambda-\mu)t})^2}, \\ p_n(t) &= (\lambda/\mu)^{n-1} p_1(t)[p_0(t)]^{n-1}. \end{aligned} \tag{3.13}$$

Let $\mathcal{T}$ be the oriented tree with n leaves and $x_1 > x_2 > \ldots > x_{n-1}$ the time of the speciation events. Note that the $x_i, i = \{1, 2, \ldots, n-1\}$ is the order statistic of the values $s_i, i = \{1, 2, \ldots, n-1\}$ of the point process.

Let t_1 be the time of the *mrca* in a reconstructed oriented tree. In [95], page 56, the density g of the ordered speciation times, $x_2 > x_3 > \ldots > x_{n-1}$, given n and $x_1 = t_1$ is derived,

$$g(x_2, x_3, \ldots, x_{n-1}|t_1 = t, n) = (n-2)! \prod_{i=2}^{n-1} \mu \frac{p_1(x_i)}{p_0(t)}.$$

This joint density is used in [100] in order to infer reconstructed trees with Bayesian methods. We will calculate the density for each speciation event separately; this will enable us to estimate the speciation times separately. The variables $x_2, x_3, \ldots, x_n$ are the order statistic of say $s_2, s_3, \ldots, s_{n-1}$. Each permutation of the $n-2$ random variables $s_2, s_3, \ldots, s_{n-1}$ has equal probability (Corollary 3.4.3), and therefore the density f of the speciation times is,

$$f(s_2, \ldots, s_{n-1}|t_1 = t, n) = \frac{g(x_2, x_3, \ldots, x_n|t_1 = t, n)}{(n-2)!} = \prod_{i=2}^{n-1} \mu \frac{p_1(s_i)}{p_0(t)}$$

which (by definition of independence) shows that the s_i are i.i.d., and therefore,

$$f(s_i|t_1 = t, n) = \mu \frac{p_1(s_i)}{p_0(t)} = (\lambda-\mu)^2 \frac{e^{-(\lambda-\mu)s_i}}{(\lambda-\mu e^{-(\lambda-\mu)s_i})^2} \frac{\lambda-\mu e^{-(\lambda-\mu)t}}{1-e^{-(\lambda-\mu)t}}. \tag{3.14}$$

Note that the expression for the density of s_i does not depend on n, we have the same distribution for any n. Therefore, we do not need to condition on n. For the distribution, we obtain by integrating Equation (3.14) w.r.t. s_i,

$$F(s_i|t_1 = t) = \frac{1-e^{-(\lambda-\mu)s_i}}{\lambda-\mu e^{-(\lambda-\mu)s_i}} \frac{\lambda-\mu e^{-(\lambda-\mu)t}}{1-e^{-(\lambda-\mu)t}}. \tag{3.15}$$

Note that the probabilities are conditioned on t_1, the time of the most recent common ancestor (*mrca*). It is of interest to condition on t_{or} instead, the time since *origin* of the tree. We have the property that

$$f(s_i|t_1 = t) = f(s_i|t_{or} = t). \tag{3.16}$$

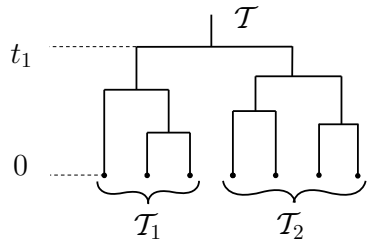

Figure 3.2: Reconstructed tree $\mathcal{T}$ with daughter trees $\mathcal{T}_1$, $\mathcal{T}_2$. We have $mrca(\mathcal{T}) = origin(\mathcal{T}_1) = origin(\mathcal{T}_2) = t_1$.

The following argument verifies Equation (3.16). Suppose we have a tree where the *mrca* was at time t_1. The daughter trees $\mathcal{T}_n, \mathcal{T}_m$ of the *mrca* have n, m extant species. The speciation times in $\mathcal{T}_n, \mathcal{T}_m$ occured according to Equation (3.14). On the other hand, since the two daughter trees of the *mrca* evolve independently, the tree $\mathcal{T}_n$ can be regarded as a birth-death process which is conditioned to have n species today and the time of origin was $t_{or} = t$. Therefore $f(s_i|t_1 = t) = f(s_i|t_{or} = t)$, see also Figure 3.2. This establishes the following theorem.

Theorem 3.4.4. *The speciation times $s_1, \ldots, s_{n-1}$ in a reconstructed oriented tree with n species conditioned on the age of the tree are i.i.d. The speciation times $s_2, \ldots, s_{n-1}$ in a reconstructed oriented tree with n species conditioned on the mrca are i.i.d. The time s of a speciation event given (i) the time since the origin of the tree is t_{or}, or (ii) the time since the mrca is t_1, has the following density and distribution for $\lambda > \mu \geq 0$,*

$$f(s|t) := f(s|t_{or} = t) = f(s|t_1 = t) = \begin{cases} \frac{(\lambda-\mu)^2 e^{-(\lambda-\mu)s}}{(\lambda-\mu e^{-(\lambda-\mu)s})^2} \frac{\lambda-\mu e^{-(\lambda-\mu)t}}{1-e^{-(\lambda-\mu)t}} & \text{if } s \leq t, \\ 0 & \text{else,} \end{cases}$$

$$F(s|t) := F(s|t_{or} = t) = F(s|t_1 = t) = \begin{cases} \frac{1-e^{-(\lambda-\mu)s}}{\lambda-\mu e^{-(\lambda-\mu)s}} \frac{\lambda-\mu e^{-(\lambda-\mu)t}}{1-e^{-(\lambda-\mu)t}}, & \text{if } s \leq t, \\ 1 & \text{else.} \end{cases}$$

Since conditioning a tree to have the *mrca* at time t can be interpreted as conditioning the two daughter trees $\mathcal{T}_n$ and $\mathcal{T}_m$ of $\mathcal{T}$ to have the origin at time t, we will only condition on the origin of the tree in the following.

Special models

The Yule model

For the special case of a pure birth process, i.e. $\mu = 0$, which is the Yule model, Equation (3.14) simplifies to

$$f(s|t) = \frac{\lambda e^{-\lambda s}}{1 - e^{-\lambda t}}$$

$$F(s|t) = \frac{1 - e^{-\lambda s}}{1 - e^{-\lambda t}}$$

which has already been established in [70] – the author conditions on the time since the *mrca* though.

The conditioned critical branching process
In a cCBP, we have $\lambda = \mu$. As $\mu \to \lambda$, we get in the limit using Equation (3.14), (3.15) and (3.16), and the property $e^{-\epsilon} \sim 1 - \epsilon$ for $\epsilon \to 0$,

$$\begin{aligned} f(s|t) &= \frac{1}{(1+\lambda s)^2}\frac{1+\lambda t}{t}, \\ F(s|t) &= \frac{s}{1+\lambda s}\frac{1+\lambda t}{t}. \end{aligned}$$

This has already been established in a different way for $\lambda = 1$ in [2, 74].

Remark 3.4.5. The point process representation will be used in Chapter 5 in order to simulate cBDP trees on n species efficiently.

3.4.2 The time of origin

Suppose nothing is known about t, the time of origin of a tree. As in [2, 74], we then assume a uniform prior on $(0, \infty)$, i.e. a tree is equally likely to originate at any point in time. This is an improper prior as mentioned in Section 1.1. Assuming this uniform prior, we will establish the density for t given n extant species. From Equation (3.13), we have the probability of n extant species given the time of origin is t,

$$\mathbb{P}_{or}[n|t] = \lambda^{n-1}(\lambda-\mu)^2\frac{(1-e^{-(\lambda-\mu)t})^{n-1}e^{-(\lambda-\mu)t}}{(\lambda-\mu e^{-(\lambda-\mu)t})^{n+1}}$$

with $\lambda > \mu \geq 0$. In order to derive the density for t given n, we need the following lemma.

Lemma 3.4.6. *Let $\mathbb{P}_{or}[n|t]$ be the probability that a tree has n extant species given the time of origin t. We have for $\lambda > \mu \geq 0$,*

$$\int_0^\infty \mathbb{P}_{or}[n|t]dt = \frac{1}{n\lambda}.$$

Proof. The derivative of $\left(\frac{1-e^{-(\lambda-\mu)t}}{\lambda-\mu e^{-(\lambda-\mu)t}}\right)^n$ is, using the quotient rule,

$$\frac{d}{dt}\left(\frac{1-e^{-(\lambda-\mu)t}}{\lambda-\mu e^{-(\lambda-\mu)t}}\right)^n = n\frac{(1-e^{-(\lambda-\mu)t)^{n-1}}}{(\lambda-\mu e^{-(\lambda-\mu)t})^{n+1}}(\lambda-\mu)^2e^{-(\lambda-\mu)t}$$

and therefore,

$$\begin{aligned} \int_0^\infty \mathbb{P}_{or}[n|t]dt &= \frac{\lambda^{n-1}}{n}\left[\left(\frac{1-e^{-(\lambda-\mu)t}}{\lambda-\mu e^{-(\lambda-\mu)t}}\right)^n\right]_0^\infty \\ &= \frac{\lambda^{n-1}}{n}\left(\frac{1}{\lambda^n}-0\right) = \frac{1}{\lambda n} \end{aligned}$$

which establishes the lemma.

□

Theorem 3.4.7. *We assume the uniform prior on $(0,\infty)$ for the time of origin of a tree. Conditioning the tree on having n species today, the time of origin has density function,*

$$q_{or}(t|n) = n\lambda^n(\lambda-\mu)^2\frac{(1-e^{-(\lambda-\mu)t})^{n-1}e^{-(\lambda-\mu)t}}{(\lambda-\mu e^{-(\lambda-\mu)t})^{n+1}} \tag{3.17}$$

where $\lambda > \mu \geq 0$.

Proof. With Bayes' law, we have

$$\begin{aligned} q_{or}(t|n) &= \frac{\mathbb{P}_{or}[n|t]q_{or}(t)}{\mathbb{P}_{or}[n]} = \frac{\mathbb{P}_{or}[n|t]q_{or}(t)}{\int_0^\infty \mathbb{P}_{or}[n,t]dt} \\ &= \frac{\mathbb{P}_{or}[n|t]q_{or}(t)}{\int_0^\infty \mathbb{P}_{or}[n|t]q_{or}(t)dt} = \frac{\mathbb{P}_{or}[n|t]}{\int_0^\infty \mathbb{P}_{or}[n|t]dt} \\ &\overset{\text{Lemma 3.4.6}}{=} \lambda n\mathbb{P}_{or}[n|t] \\ &= n\lambda^n(\lambda-\mu)^2\frac{(1-e^{-(\lambda-\mu)t})^{n-1}e^{-(\lambda-\mu)t}}{(\lambda-\mu e^{-(\lambda-\mu)t})^{n+1}}. \end{aligned}$$

□

Remark 3.4.8. For $\lambda = \mu$, we have

$$q_{or}(t|n) = \frac{nt^{n-1}}{(1+t)^{n+1}} \tag{3.18}$$

which is obtained from Theorem 3.4.7 by taking the limit $\mu \to \lambda$. The density has already been established in [2].

3.4.3 Properties of the speciation times

In Section 3.4.1, we showed that under the cBDP, a reconstructed tree of age t can be interpreted as a point process on $n-1$ points which are i.i.d. We will see in this section, that the same is not true if we do not condition on the age of the tree but assume a uniform prior.

From Theorem 3.4.4 we obtain the density function for $x = (x_1,\ldots,x_{n-1})$, the order statistic of the speciation times, conditioned on the time of origin, t,

$$f(x|t,n) = (n-1)!\prod_{i=1}^{n-1}\frac{(\lambda-\mu)^2e^{-(\lambda-\mu)x_i}}{(\lambda-\mu e^{-(\lambda-\mu)x_i})^2}\frac{\lambda-\mu e^{-(\lambda-\mu)t}}{1-e^{-(\lambda-\mu)t}}.$$

With the uniform prior on the time of origin, we obtain the density for x, given we have n extant species, $f(x|n)$, for $0 \leq \mu < \lambda$,

$$\begin{aligned} f(x|n) &= \int_{x_1}^{\infty} f(x|t,n) q_{or}(t|n) dt \\ &= n!\lambda^n(\lambda-\mu)^2 \left(\prod_{i=1}^{n-1} \frac{(\lambda-\mu)^2 e^{-(\lambda-\mu)x_i}}{(\lambda-\mu e^{-(\lambda-\mu)x_i})^2} \right) \int_{x_1}^{\infty} \frac{e^{-(\lambda-\mu)t}}{(\lambda-\mu e^{-(\lambda-\mu)t})^2} dt \\ &\overset{\mu \neq 0}{=} n!\lambda^n(\lambda-\mu)^2 \left(\prod_{i=1}^{n-1} \frac{(\lambda-\mu)^2 e^{-(\lambda-\mu)x_i}}{(\lambda-\mu e^{-(\lambda-\mu)x_i})^2} \right) \left[\frac{1}{-\mu(\lambda-\mu)(\lambda-\mu e^{-(\lambda-\mu)t})} \right]_{x_1}^{\infty} \\ &= n!\lambda^{n-1}(\lambda-\mu) \frac{e^{-(\lambda-\mu)x_1}}{\lambda-\mu e^{-(\lambda-\mu)x_1}} \prod_{i=1}^{n-1} \frac{(\lambda-\mu)^2 e^{-(\lambda-\mu)x_i}}{(\lambda-\mu e^{-(\lambda-\mu)x_i})^2}. \end{aligned} \tag{3.19}$$

If the $n-1$ speciation points were i.i.d. with density function g, we would have $f(x|n) = (n-1)! \prod_{i=1}^{n-1} g(x_i|n)$. Such a function g does not exist due to the x_1, i.e. the s_i are not i.i.d. However, since each permutation of the s_i is equally likely (Corollary 3.4.3), the s_i are distributed identical. If we condition on the time of the *mrca*, x_1, we again have independent points, as stated in Theorem 3.4.4.

For $\mu = 0$, i.e. for the Yule model, we establish the analogous result,

$$\begin{aligned} f(x|n) &= \int_{x_1}^{\infty} f_{or}(x|t,n) q_{or}(t|n) dt = n!\lambda^n \prod_{i=1}^{n-1} e^{-\lambda x_i} \int_{x_1}^{\infty} e^{-\lambda t} dt \\ &= n!\lambda^{n-1} e^{-\lambda x_1} \prod_{i=1}^{n-1} e^{-\lambda x_i}. \end{aligned}$$

For the limit $\mu \to \lambda$, we obtain an analog result with the property $e^{-\epsilon} \sim 1-\epsilon$ for $\epsilon \to 0$,

$$f(x|n) = \frac{n!}{1+\lambda x_1} \prod_{i=1}^{n-1} \frac{\lambda}{(1+\lambda x_i)^2}$$

Remark 3.4.9. In the latest implementation of BEAST [16], an MCMC-program for inferring trees, the BDP model is implemented as a possible tree prior. The density of a tree assuming the BDP prior is calculated in BEAST using the above results, i.e. with the formula for $f(x|n)$ given in Equation (3.19).

Remark 3.4.10. Let us now consider the joint probability of the $n-1$ speciation events and the time of origin, $f(x,t|n)$. With $t := x_0$, we have,

$$f(x_0, x_1, \ldots x_{n-1}|n) = f(x_1, \ldots x_{n-1}|t,n) q_{or}(t|n) = n! \prod_{i=0}^{n-1} \lambda \frac{(\lambda-\mu)^2 e^{-(\lambda-\mu)x_i}}{(\lambda-\mu e^{-(\lambda-\mu)x_i})^2},$$

i.e. $x_0, x_1, \ldots x_{n-1}$ is the order statistic of n i.i.d. random variables.

3.4.4 The time of speciation events

In this section, we calculate the density for the time of the k-th speciation event under the cBDP given we have n species today. We condition on the age of the tree, t, as well as assuming a uniform prior for t.

Known age of the tree

Let $\mathcal{A}_n^k$ be again the time of the k-th speciation event in a reconstructed tree with n extant species. We condition on the age of the tree t. The $n-1$ speciation events (conditioned on t) are i.i.d. and have the density function $f(s|t)$, see Theorem 3.4.4. The density of $\mathcal{A}_n^k$ is therefore the $(n-k)$-th order statistic, which is (see e.g. [14], Theorem 9.17),

$$f_{\mathcal{A}_n^k}(s|t) \;=\; (n-k)\binom{n-1}{n-k}F(s|t)^{n-k-1}(1-F(s|t))^{k-1}f(s|t), \qquad (3.20)$$

for $s \leq t$ and $f_{\mathcal{A}_n^k}(s|t) = 0$ else. The distribution function of $\mathcal{A}_n^k$ conditioned on t is

$$F_{\mathcal{A}_n^k}(s|t) = \sum_{i=0}^{k-1}\binom{n-1}{i}F(s|t)^{n-i-1}(1-F(s|t))^{i} \qquad (3.21)$$

for $s \leq t$ and $F_{\mathcal{A}_n^k}(s|t) = 1$ else.

Unknown age of the tree

If the time of origin is unknown, we assume a uniform prior for the time of origin. Using this assumption, we will calculate the density function for $\mathcal{A}_n^k$, the time of the k-th speciation event in a tree with n extant species.

Theorem 3.4.11. *Let $\mathcal{A}_n^k$ be the time of the k-th speciation event in a tree with n extant species. We have for $0 \leq \mu < \lambda$,*

$$f_{\mathcal{A}_n^k}(s) = (k+1)\binom{n}{k+1}\lambda^{n-k}(\lambda-\mu)^{k+2}e^{-(\lambda-\mu)(k+1)s}\frac{(1-e^{-(\lambda-\mu)s})^{n-k-1}}{(\lambda-\mu e^{-(\lambda-\mu)s})^{n+1}}.$$

Proof. For a fixed time t of origin, we have the density $f_{\mathcal{A}_n^k}(s|t)$ for the time of the k-th speciation event (Equation 3.20). With the uniform prior, the time of origin

has density function $q_{or}(t|n)$. The density $f_{\mathcal{A}_n^k}(s)$ is therefore

$$\begin{aligned}
f_{\mathcal{A}_n^k}(s) &= \int_s^\infty f_{\mathcal{A}_n^k}(s|t) q_{or}(t|n) dt \\
&= \int_s^\infty k\binom{n-1}{k}(\lambda-\mu)^{k+1}(e^{-(\lambda-\mu)s} - e^{-(\lambda-\mu)t})^{k-1} e^{-(\lambda-\mu)s} \times \\
&\quad \frac{(\lambda-\mu e^{-(\lambda-\mu)t})^{n-k}(1-e^{-(\lambda-\mu)s})^{n-k-1}}{(\lambda-\mu e^{-(\lambda-\mu)s})^n (1-e^{-(\lambda-\mu)t})^{n-1}} \times \\
&\quad n\lambda^n(\lambda-\mu)^2 \frac{(1-e^{-(\lambda-\mu)t})^{n-1} e^{-(\lambda-\mu)t}}{(\lambda-\mu e^{-(\lambda-\mu)t})^{n+1}} dt \\
&= nk\binom{n-1}{k}(\lambda-\mu)^{k+3}\lambda^n \frac{e^{-(\lambda-\mu)ks}(1-e^{-(\lambda-\mu)s})^{n-k-1}}{(\lambda-\mu e^{-(\lambda-\mu)s})^n} \times \\
&\quad \int_s^\infty \frac{(1-e^{-(\lambda-\mu)(t-s)})^{k-1} e^{-(\lambda-\mu)t}}{(\lambda-\mu e^{-(\lambda-\mu)t})^{k+1}} dt \\
&= nk\binom{n-1}{k}(\lambda-\mu)^{k+3}\lambda^n \frac{e^{-(\lambda-\mu)ks}(1-e^{-(\lambda-\mu)s})^{n-k-1}}{(\lambda-\mu e^{-(\lambda-\mu)s})^n} \times \\
&\quad \frac{e^{-(\lambda-\mu)s}}{k(\lambda-\mu)(\lambda-\mu e^{-(\lambda-\mu)s})} \left[\left(\frac{1-e^{-(\lambda-\mu)(t-s)}}{\lambda-\mu e^{-(\lambda-\mu)t}}\right)^k\right]_s^\infty \\
&= (k+1)\binom{n}{k+1}\lambda^{n-k}(\lambda-\mu)^{k+2} e^{-(\lambda-\mu)(k+1)s} \frac{(1-e^{-(\lambda-\mu)s})^{n-k-1}}{(\lambda-\mu e^{-(\lambda-\mu)s})^{n+1}}
\end{aligned}$$

which establishs the theorem. □

Remark 3.4.12. Under the Yule model, i.e. setting $\mu = 0$ and λ arbitrary in Theorem 3.4.11, we have,

$$f_{\mathcal{A}_n^k}(s) = (k+1)\binom{n}{k+1}\lambda \frac{(e^{\lambda s}-1)^{n-k-1}}{e^{\lambda s n}}$$

which has been established in Section 3.2 for $\lambda = 1$ in a different way.

Under the cCBP the birth rate equals the death rate, $\lambda = \mu$. Taking the limit $\mu \to \lambda$, we obtain from Theorem 3.4.11 using the fact $e^{-\epsilon} \sim 1 - \epsilon$,

$$f_{\mathcal{A}_n^k}(s) = (k+1)\binom{n}{k+1}\lambda^{n-k} \frac{s^{n-k-1}}{(1+\lambda s)^{n+1}}$$

which can also be obtained with the functions $F(s|t), f(s|t), q_{or}(t|n)$ for the cCBP directly (see [26]).

3.4.5 Expected speciation times

In this section, we calculate the expected time of the k-th speciation event in a reconstructed tree with n species analytically. Our Python implementation CASS for dating trees uses the analytic results. Higher moments are calculated numerically.

Known age of the tree

Theorem 3.4.13. *The expectation of $\mathcal{A}_n^k$ conditioned on t is, for $0 < \mu < \lambda$,*

$$\begin{aligned}\mathbb{E}[\mathcal{A}_n^k|t] &= t - \sum_{i=0}^{k-1}\sum_{j=0}^{i}\binom{n-1}{i}\binom{i}{j}(-1)^{i+j}\left(\frac{\lambda-\mu e^{-(\lambda-\mu)t}}{1-e^{-(\lambda-\mu)t}}\right)^{n-j-1} \times \\ &\quad \left[g(j) + \sum_{l=1}^{n-j-1}\sum_{m=0}^{l-1}\binom{n-j-1}{l}\binom{l-1}{m}(-1)^{l+m}\frac{\lambda^{l-1-m}}{(\lambda-\mu)\mu^l}h(j,m)\right]\end{aligned}$$

where

$$\begin{aligned}g(j) &= \frac{1}{(\lambda-\mu)\lambda^{n-j-1}} \times \\ &\quad \left[\ln\left(\frac{\lambda e^{(\lambda-\mu)t}-\mu}{\lambda-\mu}\right) - \sum_{m=1}^{n-j-2}\binom{n-j-2}{m}\frac{\mu^m}{m}\left(\lambda e^{(\lambda-\mu)t}-\mu\right)^{-m} - (\lambda-\mu)^{-m}\right]\end{aligned}$$

and

$$h(j,m) = \begin{cases} \ln\left(\frac{\lambda-\mu e^{-(\lambda-\mu)t}}{\lambda-\mu}\right), & \text{if } m+j+1-n=-1, \\ \frac{(\lambda-\mu e^{-(\lambda-\mu)t})^{m+j+2-n}-(\lambda-\mu)^{m+j+2-n}}{m+j+2-n} & \text{else.}\end{cases}$$

For $\mu = 0$, we have,

$$\begin{aligned}\mathbb{E}[\mathcal{A}_n^k|t] &= \sum_{i=0}^{n-k-1}\sum_{j=0}^{k-1}\frac{k\binom{n-1}{k}\binom{n-k-1}{i}\binom{k-1}{j}(-1)^{i+j}}{\lambda(k+i-j)^2} \times \\ &\quad (1-e^{-\lambda t})^{1-n}(e^{-j\lambda t} - ((k+i-j)\lambda t+1)e^{-(k+i)\lambda t}).\end{aligned}$$

For $\mu = \lambda$, we have

$$\begin{aligned}\mathbb{E}[\mathcal{A}_n^k|t] &= t - \sum_{i=0}^{k-1}\sum_{j=0}^{i}\binom{n-1}{i}\binom{i}{j}\frac{(-1)^{i+j}}{\lambda^{n-j}}\left(\frac{1+\lambda t}{t}\right)^{n-j-1} \times \\ &\quad \left[\lambda t - (n-j-1)\ln(1+\lambda t) + \sum_{l=2}^{n-j-1}\binom{n-j-1}{l}(-1)^l\frac{(1+\lambda t)^{-l+1}-1}{1-l}\right].\end{aligned}$$

Proof. Under the Yule model, i.e. $\mu = 0$, the expectation of $\mathcal{A}_n^k$ has been calculated in Theorem 3.2.9 for $\lambda = 1$,

$$\begin{aligned}\mathbb{E}[\mathcal{A}_n^k(\lambda=1)|t] &= \sum_{i=0}^{n-k-1}\sum_{j=0}^{k-1}\frac{k\binom{n-1}{k}\binom{n-k-1}{i}\binom{k-1}{j}(-1)^{i+j}}{(k+i-j)^2} \times \\ &\quad (1-e^{-t})^{1-n}(e^{-jt} - ((k+i-j)t+1)e^{-(k+i)t}).\end{aligned}$$

For general λ, since

$$f_{\mathcal{A}_n^k}(s|t) = k\binom{n-1}{k}\lambda(e^{-\lambda s}-e^{-\lambda t})^{k-1}e^{-\lambda s}\frac{(1-e^{-\lambda s})^{n-k-1}}{(1-e^{-\lambda t})^{n-1}},$$

we have

$$\mathbb{E}[\mathcal{A}_n^k(\lambda)|t] = \int_0^t k\binom{n-1}{k}\lambda s(e^{-\lambda s}-e^{-\lambda t})^{k-1}e^{-\lambda s}\frac{(1-e^{-\lambda s})^{n-k-1}}{(1-e^{-\lambda t})^{n-1}}ds.$$

Substituting $x=\lambda s$ yields

$$\begin{aligned}\mathbb{E}[\mathcal{A}_n^k(\lambda)|t] &= \int_0^{\lambda t}\frac{k}{\lambda}\binom{n-1}{k}x(e^{-x}-e^{-\lambda t})^{k-1}e^{-x}\frac{(1-e^{-x})^{n-k-1}}{(1-e^{-\lambda t})^{n-1}}dx\\ &= \frac{\mathbb{E}[\mathcal{A}_n^k(\lambda=1)|\lambda t]}{\lambda}.\end{aligned}$$

For $0<\mu<\lambda$, we have,

$$\begin{aligned}\mathbb{E}[\mathcal{A}_n^k|t] &= \int_0^t s f_{\mathcal{A}_n^k}(s|t)ds = [sF_{\mathcal{A}_n^k}(s|t)]_0^t - \int_0^t F_{\mathcal{A}_n^k}(s|t)ds\\ &= t-\sum_{i=0}^{k-1}\sum_{j=0}^{i}\binom{n-1}{i}\binom{i}{j}(-1)^{i+j}\int_0^t F(s|t)^{n-j-1}ds \qquad (3.22)\\ &= t-\sum_{i=0}^{k-1}\sum_{j=0}^{i}\binom{n-1}{i}\binom{i}{j}(-1)^{i+j}\left(\frac{\lambda-\mu e^{-(\lambda-\mu)t}}{1-e^{-(\lambda-\mu)t}}\right)^{n-j-1}\\ &\quad\int_0^t\left(\frac{1-e^{-(\lambda-\mu)s}}{\lambda-\mu e^{-(\lambda-\mu)s}}\right)^{n-j-1}ds\\ &= t-\sum_{i=0}^{k-1}\sum_{j=0}^{i}\sum_{l=0}^{n-j-1}\binom{n-1}{i}\binom{i}{j}\binom{n-j-1}{l}(-1)^{i+j+l}\\ &\quad\left(\frac{\lambda-\mu e^{-(\lambda-\mu)t}}{1-e^{-(\lambda-\mu)t}}\right)^{n-j-1}\int_0^t\frac{e^{-(\lambda-\mu)ls}}{(\lambda-\mu e^{-(\lambda-\mu)s})^{n-j-1}}ds.\end{aligned}$$

With the substitution $x=\lambda-\mu e^{-(\lambda-\mu)s}$, we obtain for $l>0$,

$$\begin{aligned}\int_0^t\frac{e^{-(\lambda-\mu)ls}}{(\lambda-\mu e^{-(\lambda-\mu)s})^{n-j-1}}ds &= \frac{1}{\mu(\lambda-\mu)}\int_{\lambda-\mu}^{\lambda-\mu e^{-(\lambda-\mu)t}}\frac{\left(\frac{\lambda-x}{\mu}\right)^{l-1}}{x^{n-j-1}}dx\\ &= \frac{1}{(\lambda-\mu)\mu^l}\sum_{m=0}^{l-1}\binom{l-1}{m}(-1)^m\lambda^{l-1-m}\int_{\lambda-\mu}^{\lambda-\mu e^{-(\lambda-\mu)t}}x^{m+j+1-n}dx\\ &= \frac{1}{(\lambda-\mu)\mu^l}\sum_{m=0}^{l-1}\binom{l-1}{m}(-1)^m\lambda^{l-1-m}h(j,m)\end{aligned}$$

where

$$h(j,m) = \begin{cases} \ln\left(\frac{\lambda-\mu e^{-(\lambda-\mu)t}}{\lambda-\mu}\right), & \text{if } m+j+1-n=-1, \\ \frac{(\lambda-\mu e^{-(\lambda-\mu)t})^{m+j+2-n}-(\lambda-\mu)^{m+j+2-n}}{m+j+2-n} & \text{else.} \end{cases}$$

For $l=0$, we have with the substitution $x=\lambda e^{(\lambda-\mu)s}-\mu$,

$$\begin{aligned} g(j) &:= \int_0^t \frac{1}{\left(\lambda-\mu e^{-(\lambda-\mu)s}\right)^{n-j-1}}ds \\ &= \int_0^t \frac{e^{(\lambda-\mu)(n-j-1)s}}{(\lambda e^{(\lambda-\mu)s}-\mu)^{n-j-1}} \\ &= \frac{1}{(\lambda-\mu)\lambda}\int_{\lambda-\mu}^{\lambda e^{(\lambda-\mu)t}-\mu} \frac{\left(\frac{x+\mu}{\lambda}\right)^{n-j-2}}{x^{n-j-1}}dx \\ &= \frac{1}{(\lambda-\mu)\lambda^{n-j-1}}\sum_{m=0}^{n-j-2}\binom{n-j-2}{m}\mu^m\int_{\lambda-\mu}^{\lambda e^{(\lambda-\mu)t}-\mu} x^{-m-1}dx \\ &= \frac{1}{(\lambda-\mu)\lambda^{n-j-1}}\times \\ &\quad \left[\ln\left(\frac{\lambda e^{(\lambda-\mu)t}-\mu}{\lambda-\mu}\right)-\sum_{m=1}^{n-j-2}\binom{n-j-2}{m}\frac{\mu^m}{m}\left(\lambda e^{(\lambda-\mu)t}-\mu\right)^{-m}-(\lambda-\mu)^{-m}\right)\right]. \end{aligned}$$

So overall,

$$\begin{aligned} \mathbb{E}[\mathcal{A}_n^k|t] &= t-\sum_{i=0}^{k-1}\sum_{j=0}^{i}\binom{n-1}{i}\binom{i}{j}(-1)^{i+j}\left(\frac{\lambda-\mu e^{-(\lambda-\mu)t}}{1-e^{-(\lambda-\mu)t}}\right)^{n-j-1}\times \\ &\quad \left[g(j)+\sum_{l=1}^{n-j-1}\sum_{m=0}^{l-1}\binom{n-j-1}{l}\binom{l-1}{m}(-1)^{l+m}\frac{\lambda^{l-1-m}}{(\lambda-\mu)\mu^l}h(j,m)\right]. \end{aligned}$$

For $\mu=\lambda$, we obtain from Equation (3.22),

$$\mathbb{E}[\mathcal{A}_n^k|t] = t-\sum_{i=0}^{k-1}\sum_{j=0}^{i}\binom{n-1}{i}\binom{i}{j}(-1)^{i+j}\left(\frac{1+\lambda t}{t}\right)^{n-j-1}\int_0^t\left(\frac{s}{1+\lambda s}\right)^{n-j-1}ds.$$

Substituting $x=1+\lambda s$, we get,

$$\begin{aligned} &\int_0^t\left(\frac{s}{1+\lambda s}\right)^{n-j-1}ds \\ &= \frac{1}{\lambda^{n-j}}\int_1^{1+\lambda t}\left(\frac{x-1}{x}\right)^{n-j-1}dx \\ &= \sum_{l=0}^{n-j-1}\binom{n-j-1}{l}\frac{(-1)^l}{\lambda^{n-j}}\int_1^{1+\lambda t}x^{-l}dx \\ &= \frac{1}{\lambda^{n-j}}\left[\lambda t-(n-j-1)\ln(1+\lambda t)+\sum_{l=2}^{n-j-1}\binom{n-j-1}{l}(-1)^l\frac{(1+\lambda t)^{-l+1}-1}{1-l}\right]. \end{aligned}$$

This finally yields,

$$\mathbb{E}[\mathcal{A}_n^k|t] = t - \sum_{i=0}^{k-1}\sum_{j=0}^{i}\binom{n-1}{i}\binom{i}{j}\frac{(-1)^{i+j}}{\lambda^{n-j}}\left(\frac{1+\lambda t}{t}\right)^{n-j-1} \times \left[\lambda t - (n-j-1)\ln(1+\lambda t) + \sum_{l=2}^{n-j-1}\binom{n-j-1}{l}(-1)^l\frac{(1+\lambda t)^{-l+1}-1}{1-l}\right].$$

□

Unknown age of the tree

A closed form solution for the first and second moment (for all k) of $\mathcal{A}_n^k$ under the Yule model (for $\lambda = 1$) is given in Theorem 3.2.2,

$$\mathbb{E}^{Yule}[\mathcal{A}_n^k] = \sum_{i=k+1}^{n}\frac{1}{i}, \tag{3.23}$$

$$\mathbb{E}^{Yule}[(\mathcal{A}_n^k)^2] = \sum_{i=k+1}^{n}\frac{1}{i^2} + \sum_{i=k+1}^{n}\sum_{j=k+1}^{n}\frac{1}{ij}. \tag{3.24}$$

Theorem 3.4.14. *Under the cCBP (with $\lambda = 1$), the expectation and variance of $\mathcal{A}_n^k$ are*

$$\mathbb{E}[\mathcal{A}_n^k] = \frac{n-k}{k}, \qquad \mathrm{Var}[\mathcal{A}_n^k] = \frac{n(n-k)}{k^2(k-1)}.$$

In general, the expectation of the m-th moment of $\mathcal{A}_n^k$ is

$$\mathbb{E}[(\mathcal{A}_n^k)^m] = \begin{cases} \frac{\binom{n-k+m-1}{m}}{\binom{k}{m}}, & \text{if } k \geq m; \\ \infty, & \text{else.} \end{cases} \tag{3.25}$$

Note that for $k = 1$, we have $\mathrm{Var}[\mathcal{A}_n^1] = \infty$.

Proof. For calculating the moments, we need the following result which can be found e.g. in [56],

$$\int_0^\infty \frac{s^a}{(1+s)^b}ds = \begin{cases} \frac{1}{(b-a-1)\binom{b-1}{a}} & \text{if } b > a+1, \\ \infty & \text{else,} \end{cases} \tag{3.26}$$

where $a, b \in \mathbb{N}_0$. With Remark 3.4.12, we have for the moments of $\mathcal{A}_n^k$,

$$\mathbb{E}[(\mathcal{A}_n^k)^m] = (k+1)\binom{n}{k+1}\int_0^\infty \frac{s^{n-k+m-1}}{(1+s)^{n+1}}ds.$$

For $k < m$, the value of the integral is infinite by Equation (3.26) and therefore $\mathbb{E}[(\mathcal{A}_n^k)^m] = \infty$. For $k \geq m$, we obtain with Equation (3.26),

$$\begin{aligned}\mathbb{E}[(\mathcal{A}_n^k)^m] &= (k+1)\binom{n}{k+1}\frac{1}{\binom{n}{n-k+m-1}(k-m+1)} \\ &= \frac{n!(k-m)!(n-k+m-1)!}{k!(n-k-1)!n!} = \frac{\binom{n-k+m-1}{m}}{\binom{k}{m}}\end{aligned}$$

which completes the proof of Theorem 3.4.14. □

Corollary 3.4.15. *Recursively, we have for the moments of $\mathcal{A}_n^k$,*

$$\mathbb{E}[(\mathcal{A}_n^k)^m] = \mathbb{E}[(\mathcal{A}_n^k)^{m-1}]\mathbb{E}[(\mathcal{A}_n^{k-m+1})^1].$$

Proof. From Theorem 3.4.14, we obtain for $k \geq m$

$$\mathbb{E}[(\mathcal{A}_n^k)^{m-1}] = \frac{\binom{n-k+m-2}{m-1}}{\binom{k}{m-1}}, \qquad \mathbb{E}[(\mathcal{A}_n^{k-m+1})^1] = \frac{n-k+m-1}{k-m+1}.$$

Multiplying those expectations yields the formula for $\mathbb{E}[(\mathcal{A}_n^k)^m]$ in Theorem 3.4.14. □

For general λ, μ, we have the following analytic expression for the expectation.

Theorem 3.4.16. *For $0 < \mu < \lambda$, the moments of $\mathcal{A}_n^k$ are, with $\rho := \mu/\lambda$,*

$$\begin{aligned}\mathbb{E}[\mathcal{A}_n^k] &= \frac{k+1}{\lambda}\binom{n}{k+1}(-1)^k \sum_{i=0}^{n-k-1}\binom{n-k-1}{i}\frac{1}{(k+i+1)\rho}\left(\frac{1}{\rho}-1\right)^{k+i} \times \\ &\quad \left[\log\left(\frac{1}{1-\rho}\right) - \sum_{j=1}^{k+i}\binom{k+i}{j}\frac{(-1)^j}{j}\left(1-\left(\frac{1}{1-\rho}\right)^j\right)\right].\end{aligned}$$

For $\mu = 0$ we have

$$\mathbb{E}[\mathcal{A}_n^k] = \sum_{i=k+1}^{n}\frac{1}{\lambda i}$$

and for $\mu = \lambda$ we have

$$\mathbb{E}[\mathcal{A}_n^k] = \frac{n-k}{\lambda k}.$$

In particular, the expectations basically only depend on ρ. Different λ just scale time by $1/\lambda$.

Proof. For $\mu = 0$ and for $\mu = \lambda$, the expectation is established with Remark 3.4.12 and Equations (3.23) and (3.25). For $\mu \neq 0$ and $\mu \neq \lambda$ we have with Theorem 3.4.11,

$$\mathbb{E}[\mathcal{A}_n^k] = \int_0^\infty (k+1)\binom{n}{k+1}\lambda^{n-k}(\lambda-\mu)^{k+2}e^{-(\lambda-\mu)(k+1)s}\frac{(1-e^{-(\lambda-\mu)s})^{n-k-1}}{(\lambda-\mu e^{-(\lambda-\mu)s})^{n+1}}sds.$$

Set

$$\begin{aligned} C_1 &:= (k+1)\binom{n}{k+1}\lambda^{n-k}(\lambda-\mu)^{k+2}, \\ f(s) &:= e^{-(\lambda-\mu)(k+1)s}\frac{(1-e^{-(\lambda-\mu)s})^{n-k-1}}{(\lambda-\mu e^{-(\lambda-\mu)s})^{n+1}}. \end{aligned}$$

Therefore,

$$\mathbb{E}[\mathcal{A}_n^k] = C_1\int_0^\infty f(s)sds = C_1[F(s)s]_0^\infty - C_1\int_0^\infty F(s)ds$$

where $F(s) := \int f(s)ds$.

In the following, we calculate $F(s)$. We use the following substitution:

$$x = \frac{e^{-(\lambda-\mu)s}}{\lambda-\mu e^{-(\lambda-\mu)s}} \qquad \frac{dx}{ds} = -\frac{\lambda(\lambda-\mu)e^{-(\lambda-\mu)s}}{(\lambda-\mu e^{-(\lambda-\mu)s})^2} \qquad e^{-(\lambda-\mu)s} = \frac{\lambda x}{1+\mu x}$$

This yields

$$\begin{aligned} F(s) &= -\frac{1}{\lambda(\lambda-\mu)}\int x^{n-1}\left(\frac{1-(\lambda-\mu)x}{\lambda x}\right)^{n-k-1}dx \\ &= -\frac{1}{\lambda^{n-k}(\lambda-\mu)}\int x^k(1-(\lambda-\mu)x)^{n-k-1}dx \\ &= -\frac{1}{\lambda^{n-k}(\lambda-\mu)}\sum_{i=0}^{n-k-1}\binom{n-k-1}{i}(-(\lambda-\mu))^i\int x^{k+i}dx \\ &= \frac{1}{\lambda^{n-k}}\sum_{i=0}^{n-k-1}\binom{n-k-1}{i}\frac{(-(\lambda-\mu))^{i-1}}{k+i+1}\left(\frac{e^{-(\lambda-\mu)s}}{\lambda-\mu e^{-(\lambda-\mu)s}}\right)^{k+i+1}. \end{aligned}$$

We have $\lim_{s\to\infty}F(s)s = 0$ and $F(0)\cdot 0 = 0$ and therefore,

$$\mathbb{E}[\mathcal{A}_n^k] = -C_1\int_0^\infty F(s)ds.$$

Substitute $x = \lambda - \mu e^{-(\lambda-\mu)s}$,

$$\begin{aligned} F_2(s) &:= \int_0^\infty F(s)ds \\ &= \frac{1}{\lambda^{n-k}}\sum_{i=0}^{n-k-1}\binom{n-k-1}{i}\frac{(-(\lambda-\mu))^{i-1}}{k+i+1}\int_0^\infty\left(\frac{e^{-(\lambda-\mu)s}}{\lambda-\mu e^{-(\lambda-\mu)s}}\right)^{k+i+1}ds \\ &= \frac{1}{\lambda^{n-k}}\sum_{i=0}^{n-k-1}\binom{n-k-1}{i}\frac{(-(\lambda-\mu))^{i-1}}{k+i+1}\int_{\lambda-\mu}^{\lambda}\frac{\left(\frac{\lambda-x}{\mu}\right)^{k+i}}{\mu(\lambda-\mu)x^{k+i+1}}dx \\ &= \frac{1}{\lambda^{n-k}}\sum_{i=0}^{n-k-1}\binom{n-k-1}{i}\frac{(\lambda-\mu)^{i-2}}{k+i+1}\frac{(-1)^{i-1}}{\mu^{k+i+1}}\sum_{j=0}^{k+i}\binom{k+i}{j}\lambda^j(-1)^{k+i-j}\int_{\lambda-\mu}^{\lambda}x^{-(j+1)}dx. \end{aligned}$$

Evaluating the integral yields

$$F_2(s) = \frac{(-1)^k}{\lambda^{n-k}} \sum_{i=0}^{n-k-1} \binom{n-k-1}{i} \frac{(\lambda-\mu)^{i-2}}{k+i+1} \frac{1}{\mu^{k+i+1}} \times \left[-\log(\frac{\lambda}{\lambda-\mu}) + \sum_{j=1}^{k+i} \binom{k+i}{j} \lambda^j \frac{(-1)^j}{j} [\lambda^{-j} - (\lambda-\mu)^{-j}] \right].$$

Therefore, with $\rho := \mu/\lambda$,

$$\begin{aligned}
\mathbb{E}[\mathcal{A}_n^k] &= (k+1)\binom{n}{k+1}(-1)^k \sum_{i=0}^{n-k-1} \binom{n-k-1}{i} \frac{(\lambda-\mu)^{k+i}}{k+i+1} \frac{1}{\mu^{k+i+1}} \times \\
&\quad \left[\log\left(\frac{\lambda}{\lambda-\mu}\right) - \sum_{j=1}^{k+i} \binom{k+i}{j} \frac{(-1)^j}{j} \left(1 - \left(\frac{\lambda}{\lambda-\mu}\right)^j\right) \right] \\
&= \frac{k+1}{\lambda}\binom{n}{k+1}(-1)^k \sum_{i=0}^{n-k-1} \binom{n-k-1}{i} \frac{1}{(k+i+1)\rho} \left(\frac{1}{\rho} - 1\right)^{k+i} \times \\
&\quad \left[\log\left(\frac{1}{1-\rho}\right) - \sum_{j=1}^{k+i} \binom{k+i}{j} \frac{(-1)^j}{j} \left(1 - \left(\frac{1}{1-\rho}\right)^j\right) \right]
\end{aligned}$$

which establishes the theorem. □

3.4.6 The time between speciation events

Under the birth-death process, species speciate and die with exponential waiting times. However, we condition the process to obtain a reconstructed tree with n extant species today. We will determine the time between two speciation events in the reconstructed tree on n species.

Theorem 3.4.17. *Consider the backwards process of the conditioned birth-death process – going back in time, the n extant species coalesce. A pair of species coalesces according to the density function $f(s|t)$ from Theorem 3.4.4.*

Proof. Consider a fixed pair of species out of the n species. Obviously, we can put them next to each other on the x-axis of the point process, at location $(i, i+1)$. Their coalescent point is $(i+1/2, s_i)$, see Theorem 3.4.4. The time s_i has the distribution with density function $f(s|t)$ from Theorem 3.4.4. □

In a reconstructed tree with n species, the time until the last speciation event, i.e. the time between the $(n-1)$-st speciation event and today is $\mathcal{A}_n^{n-1}$. The time between the k-th and the $(k+1)$-th speciation event can be calculated as follows. First note that since the $n-1$ points in the point process are i.i.d. with density function $f(s|t)$, the density function g of point j_1 being at time s_{j_1} and j_2 being at time s_{j_2} is,

$$g(s_{j_1}, s_{j_2}|t) = f(s_{j_1}|t) f(s_{j_2}|t).$$

Assume the k-th speciation event is at time τ and the $(k+1)$-st speciation event is at time $\tau - s$. We have $n-1$ possibilities choosing the point for the k-th speciation event from the $n-1$ points, and $n-2$ possibilities to choose the point for the $(k+1)$-st speciation event. The density function for having a speciation event at time τ and $\tau - s$ is therefore $(n-1)(n-2)f(\tau|t)f(\tau - s|t)$. The probability for $k-1$ speciation points of the remaining $n-3$ speciation points being earlier than τ is $\binom{n-3}{k-1}(1-F(\tau|t))^{k-1}$. The probability that the remaining $n-k-2$ speciation points occured after $\tau - s$ is $F(\tau - s|t)^{n-k-2}$. Overall,

$$f_{\mathcal{A}_n^{k,k+1}}(s|t) = \int_s^t (n-1)(n-2)\binom{n-3}{k-1}(1-F(\tau|t))^{k-1}F(\tau-s|t)^{n-k-2}f(\tau|t)f(\tau-s|t)d\tau.$$

The time between the k-th and l-th speciation event ($k < l$) in a tree of age t can be obtained in the same way. In addition to above, we require $l-k-1$ points to be between τ and $\tau - s$,

$$\begin{aligned} f_{\mathcal{A}_n^{k,l}}(s|t) &= \int_s^t (n-1)(n-2)\binom{n-3}{k-1}\binom{n-k-2}{l-k-1}(1-F(\tau|t))^{k-1} \times \\ &\quad (F(\tau|t)-F(\tau-s|t))^{l-k-1}F(\tau-s|t)^{n-l-1}f(\tau|t)f(\tau-s|t)d\tau. \end{aligned}$$

Note that $\mathcal{A}_n^{k-1,k}$ is the time until a coalescent event for k species in the reconstructed tree. Recall that we found analytic solutions for the above integrals under the Yule model (Section 3.2). For $\mu \neq 0$, the densities can be obtained with numerical integration using the Cass package. However, obtaining the expectation for $\mathcal{A}_n^{k,l}$ can be done analytically, $\mathbb{E}[\mathcal{A}_n^{k,l}] = \mathbb{E}[\mathcal{A}_n^k] - \mathbb{E}[\mathcal{A}_n^l]$. If assuming a uniform prior for the time of origin, we additionally need to integrate the above densities over t, weighted by $q_{or}(t|n)$. Again, for the Yule model, we have analytic solutions (Section 3.2).

Remark 3.4.18. For obtaining the density of the time of the k-th speciation event, $\mathcal{A}_n^k$, under the Yule model, we have the point process approach in this section, as well as the approach in Section 3.2. However, it is difficult to obtain a simple expression for the moments of $\mathcal{A}_n^k$ from the point process approach. Further, there is no obvious way to obtain a simple expression for the density of $\mathcal{A}_n^{k,l}$ from the point process approach. In Section 3.2, we obtained simple expressions for these quantities.

3.4.7 Comparing the extreme neutral models: Yule and CBP

As established in Proposition 2.3.1 the Yule model and the cCBP induce the same distribution on ranked oriented trees, the uniform distribution. However, the speciation times differ. In Equation (3.6) we established $\mathbb{E}_{Yule}[\mathcal{A}_n^k] = \sum_{i=k+1}^n \frac{1}{i}$. Since

$$\lim_{n\to\infty}\left(\sum_{i=1}^n \frac{1}{i} - \ln n\right) = \gamma$$

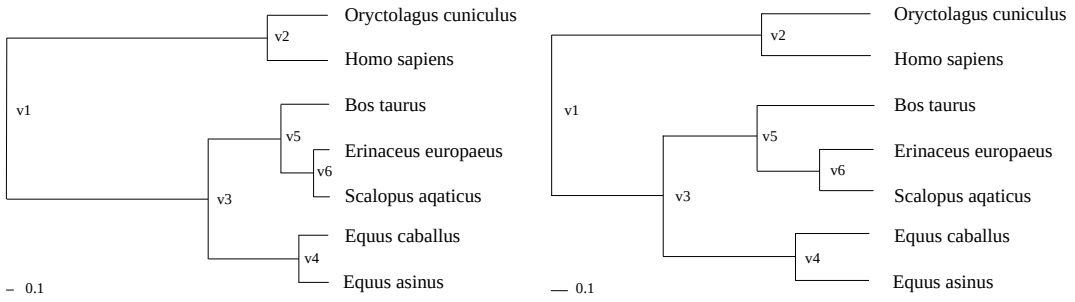

Figure 3.3: Given the labeled tree, we obtained the displayed expected edge lengths under the cCBP (left) and Yule model (right).

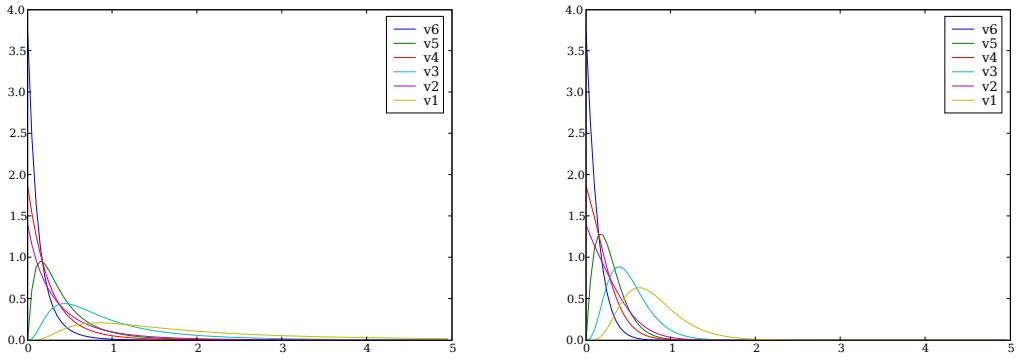

Figure 3.4: Density functions of the time of each interior vertex in the labeled tree shown in Figure 3.3 under the cCBP (left) and Yule model (right). Since $f_{\mathcal{A}_n^k}(0) = 0$ for $k < n-1$, but $f_{\mathcal{A}_n^{n-1}}(0) \neq 0$, we obtain $f_{\mathcal{A}_v}(0) \neq 0$ if and only if $\mathbb{P}[r(v) = n-1] > 0$.

with γ being the Euler constant, we have

$$\sum_{i=1}^{n} \frac{1}{i} = \ln n + \gamma + o(1)$$

as $n \to \infty$ and therefore, for fixed k, $\sum_{i=k+1}^{n} \frac{1}{i} = \ln n + O(1)$. Asymptotically, this is $\sum_{i=k+1}^{n} \frac{1}{i} \sim \ln n$. From Theorem 3.4.14, we have $\mathbb{E}_{cCBP}[\mathcal{A}_n^k] = \frac{n-k}{k} \sim \frac{n}{k}$. So

$$\mathbb{E}_{Yule}[\mathcal{A}_n^k] \sim \ln\left(\mathbb{E}_{cCBP}[\mathcal{A}_n^k]\right)$$

for fixed k. In particular, the root of the tree is expected to be at time $n-1$ under cCBP, but at time $\ln n$ under Yule.

We will now compare the speciation times in a given tree for the Yule and cCBP model. Consider as an example the phylogenetic tree found in Figure 3.3, available on TreeBase [91]. The data tree had no time scale, only the tree shape was inferred.

We will calculate the speciation times for the interior vertices in the given tree shape with the theory established in this chapter. Note that we consider this tree to show how our methods work – we will not discuss if for that particular phylogeny, a neutral assumption is reasonable.

We calculated for each interior vertex the density function for the time of speciation under the cCBP and under the Yule model with Equation (3.1), see Figure 3.4. Further, we calculated the expected speciation times with Equation (3.3), see Figure 3.3. Note that not only the dates, but also the ranking of the obtained expected tree can be different for those two models (even though the distribution on ranked trees is the same for both models). We have $\mathbb{E}_{cCBP}[v2] > \mathbb{E}_{cCBP}[v5]$ but $\mathbb{E}_{Yule}[v2] < \mathbb{E}_{Yule}[v5]$. The expectations with the standard deviation are listed below. All calculations are done with our Cass package.

$$\begin{aligned}
\mathbb{E}_{cCBP}[v1] &= 6.0000 \pm \infty & \mathbb{E}_{Yule}[v1] &= 1.5929 \pm 0.7154\\
\mathbb{E}_{cCBP}[v2] &= 1.0300 \pm 1.6968 & \mathbb{E}_{Yule}[v2] &= 0.5629 \pm 0.4759\\
\mathbb{E}_{cCBP}[v3] &= 2.2667 \pm 2.7439 & \mathbb{E}_{Yule}[v3] &= 1.0262 \pm 0.5072\\
\mathbb{E}_{cCBP}[v4] &= 0.6178 \pm 0.8059 & \mathbb{E}_{Yule}[v4] &= 0.4084 \pm 0.3474\\
\mathbb{E}_{cCBP}[v5] &= 0.9133 \pm 0.9794 & \mathbb{E}_{Yule}[v5] &= 0.5695 \pm 0.3667\\
\mathbb{E}_{cCBP}[v6] &= 0.3222 \pm 0.4063 & \mathbb{E}_{Yule}[v6] &= 0.2473 \pm 0.2343
\end{aligned}$$

3.5 Connections to the coalescent

3.5.1 The point process of the coalescent

The coalescent is the standard neutral model for population genetics [53, 51, 52]. The n individuals in a population are assumed to coalesce as follows. For the most recent coalescent event, pick two of the n individuals uniformly at random, the time between today and their coalescence is exponentially distributed (rate $\binom{n}{2}\lambda$) where λ encodes the population size. Note that if considering the coalescent in forward time, it is a CAL model: each lineage is equally likely to bifurcate next. We will show that the coalescent – even though it is very similar to the Yule process – does not have a point process representation with i.i.d. coalescent points.

Let $x = (x_1, x_2, \ldots, x_{n-1})$ be the order statistic of the coalescent times (with $x_1 > x_2 \ldots > x_{n-1}$). Note that $x_i - x_{i+1}$ is distributed exponentially with rate $\binom{i+1}{2}\lambda$. The density function for x is therefore,

$$\begin{aligned}
f(x|n) &= \left(\lambda\binom{n}{2}e^{-\lambda\binom{n}{2}x_{n-1}}\right)\prod_{i=1}^{n-2}\lambda\binom{i+1}{2}e^{-\lambda\binom{i+1}{2}(x_i-x_{i+1})}\\
&= \frac{n!(n-1)!}{2^{n-1}}\prod_{i=1}^{n-1}\lambda e^{-\lambda i x_i}.
\end{aligned}$$

Conditioning on the time of the most recent common ancestor, x_1, we get,

$$f(x|n,x_1) = \frac{f(x|n)}{f(x_1|n)} = \frac{n!(n-1)!}{f(x_1|n)2^{n-1}} \prod_{i=1}^{n-1} \lambda e^{-\lambda i x_i} = h(x_1,n) \prod_{i=2}^{n-1} \lambda e^{-\lambda i x_i}.$$

where h is a function only depending on x_1, n. If the $n-2$ coalescent points were i.i.d. with density function g, we would have $f(x|n,x_1) = (n-2)! \prod_{i=2}^{n} g(x_i, x_1, n)$. However, due to the factor i in $e^{-\lambda i x_i}$, this property is not satisfied, therefore the $n-2$ points are not i.i.d. However, in the coalescent, also each ranked oriented tree is equally likely (Proposition 2.3.1), therefore each permutation of the s_i has the same probability. That means that the s_i are identically distributed – but not independent.

3.5.2 The CBP and the coalescent

We will show a surprising connection between the cCBP and the coalescent. Under the coalescent setting, the random variable $\mathcal{A}_n^{k-1,k}$, 'waiting time between the k-th and the $(k-1)$-th coalescent event' is exponential $\left(\lambda\binom{k}{2}\right)$ distributed. For the first moment of $\mathcal{A}_n^k$ under the coalescent, we have for $\lambda = 1$,

$$\begin{aligned}\mathbb{E}_{Coal}[\mathcal{A}_n^k] &= \sum_{i=k+1}^{n} \mathcal{A}_n^{i-1,i} = \sum_{i=k+1}^{n} \frac{2}{i(i-1)} \\ &= 2(1-1/n) - 2(1-1/k) = \frac{2}{n}\frac{n-k}{k} = \frac{2}{n}\mathbb{E}_{cCBP}[\mathcal{A}_n^k].\end{aligned}$$

So in expectation, the coalescent with rate 1 is equivalent to the cCBP with rate $\lambda = \frac{n}{2}$. The ranked trees under the coalescent are distributed uniformly at random since the coalescent in forward-time is a CAL model – so the distribution is the same as under the cCBP. Therefore the cCBP and the coalescent are alike when only considering tree shapes and the expected time of the interior vertices. However, when considering higher moments, the models differ, since,

$$\begin{aligned}\mathbb{E}_{Coal}[(\mathcal{A}_n^k)^2] &= \mathrm{Var}_{Coal}[\mathcal{A}_n^k] + (\mathbb{E}_{Coal}[\mathcal{A}_n^k])^2 = \sum_{i=k+1}^{n} \mathrm{Var}_{Coal}[\mathcal{A}_n^{i-1,i}] + (\mathbb{E}_{Coal}[\mathcal{A}_n^k])^2 \\ &= \sum_{i=k+1}^{n} \frac{1}{\binom{i}{2}^2} + \left(\frac{2(n-k)}{nk}\right)^2,\end{aligned}$$

i.e. the second moments of $\mathcal{A}_n^k$ are finite under the coalescent, whereas under the cCBP model, the second moment of $\mathcal{A}_n^1$ is ∞.

3.6 Horizontal LTT plots

Knowing the expected time of the k-th speciation event allows us to draw a lineages-through-time (LTT) plot [71] analytically. In an LTT plot, the time vs. the number

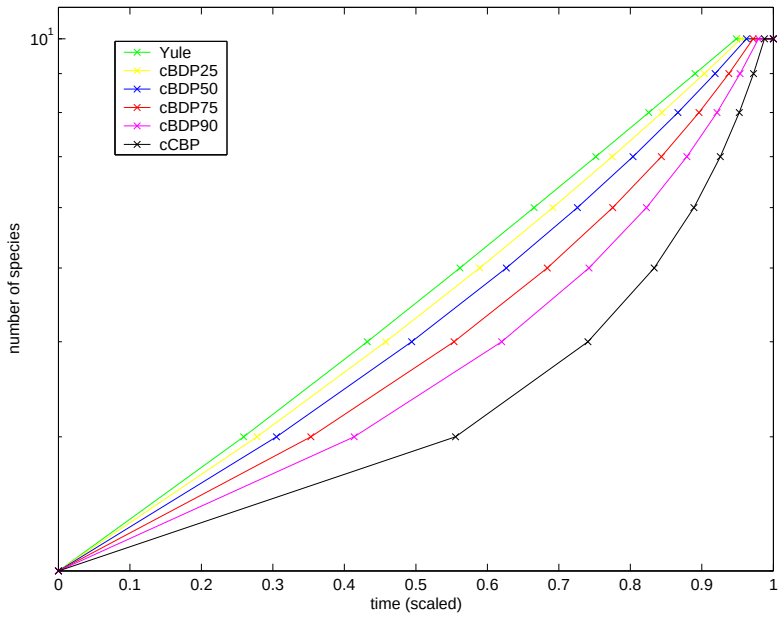

Figure 3.5: Lineages-through-time plot for $n = 10$ species. Time is scaled such that the *mrca* is at 0 and today is 1. We have $\rho = 0, 0.25, 0.5, 0.75, 0.9, 1$ from top to bottom. Note that for variing λ, time is scaled by $1/\lambda$ (compared to $\lambda = 1$). That means, since we scale time, the plots are the same for any λ.

of species at that time (on a logarithmic scale) is drawn. LTT plots are a popular graphical tool to compare the data (i.e. the reconstructed tree) with arbitrary models. For example, in [7], LTT plots are used to investigate if the rise of mammals coincided with the Cretaceous/Tertiary boundary. Using an almost complete phylogeny of the present-day mammals, the authors postulate that the mass extinction occurring at the Cretaceous/Tertiary boundary did not have a major influence on the rise of mammals.

Commonly, the LTT plots for different models are obtained via simulations. For the Yule model, we simulate until we reach n species. If extinction occurs, we would have to simulate forever, since n species can always reoccur. Therefore an analytic approach for drawing LTT plots is of special interest.

Since we know the expected time of the speciation events under the cBDP model analytically, we can plot the LTT plot without simulations, see Figure 3.5. The plot is "horizontal", i.e. we fix the k-th speciation event, and have a distribution for the time; we plot the expected time. Between events, we interpolate with a straight line.

Often, when collecting data, some species are missing in our data set. If each existing species has equal probability of not being sampled, then we can model the scenario via *random taxon sampling*. In Section 3.7, we will determine the LTT plots for cBDP models under random taxon sampling. We will see that random taxon sampling cannot be detected by LTT plots.

Note that it is also interesting to consider the "vertical" expected LTT plot: fix

a time and calculate the expected number of species at that time. Analytic results for the vertical plots are established in Section 3.8.

3.7 Speciation times under random taxon sampling

In this section, we discuss the influence of random taxon sampling on reconstructed oriented trees. We first discuss the distribution on ranked oriented trees and then the distribution on speciation times under random taxon sampling. These two distributions are independent for the cBDP [95] and determine the overall distribution on reconstructed oriented trees. We will see that LTT plots are affected by random taxon sampling. However, LTT plots under random taxon sampling look like LTT plots with complete taxon sampling but a different extinction rate. Therefore, methods for estimating the birth- and death rates of a phylogeny from an LTT plot [71] are biased under random taxon sampling, and extra care has to be taken.

First, we formally define random taxon sampling. Consider a reconstructed oriented tree on n leaves, $\mathcal{T}$. *Random taxon sampling* is choosing uniformly at random a subset of some size k from the set of leaves in $\mathcal{T}$. We consider the subtree $\mathcal{T}'$ of $\mathcal{T}$ which is the minimal subtree of $\mathcal{T}$ containing the k chosen leaves. Note that obtaining such a subtree is equivalent to deleting $n-k$ leaves uniformly at random from $\mathcal{T}$. In $\mathcal{T}'$, degree-two vertices (except of the root) are suppressed: The degree-two vertex with its two adjacent edges is replaced by a single edge. Therefore, $\mathcal{T}'$ is again a reconstructed oriented tree. We investigate the distribution on reconstructed oriented trees which are the result of random taxon sampling.

Note that the results in Section 3.7.1 and Section 3.7.2 apply to any speciation model which induces a uniform distribution on ranked oriented trees (we do not require a constant rate of speciation). Section 3.7.3 applies the results to the cBDP.

3.7.1 Ranked oriented tree distribution

The following theorem is implicitly proven in Proposition 2.3.1 and has been established before in [31, 95, 3]:

Theorem 3.7.1. *Consider the uniform distribution on ranked oriented trees with n leaves. Deleting one leaf uniformly at random induces a uniform distribution on ranked oriented trees with $n-1$ leaves.*

Therefore the ranked oriented tree distribution for trees with n leaves after random taxon sampling is the same as the ranked oriented tree distribution for trees with n leaves and complete taxon sampling. This means that the ranked oriented tree distribution is invariant under random taxon sampling.

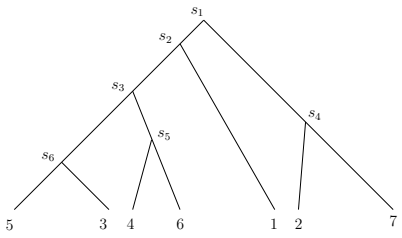

Figure 3.6: Since s_2 has a leaf attached, we observe event $s_2(\mathcal{L})$. Further, we observe event $s_2(5)$, since s_2 has 5 leaves descending. If we delete leaf 1, we delete s_2, i.e. we observe $s_2(\mathcal{D})$.

3.7.2 Speciation times

We established that the ranked oriented tree distribution is invariant under random taxon sampling. In this section, we will see that the time of the k-th speciation event changes under random taxon sampling. The aim of this section is to calculate the time of the k-th speciation event in a tree where m out of the n leaves are sampled uniformly at random, $\mathcal{A}^k_{n,m}$ (the age of the tree might be fixed, or a uniform prior is assumed).

From a ranked oriented tree on n leaves, we delete one leaf uniformly at random. We calculate the probability that we delete a leaf which is attached to the k-th speciation event, s_k ($k = 1, \ldots, n-1$). This can be considered as deleting the speciation event s_k, call that deletion event "$s_k(\mathcal{D})$". Note that s_k can only be deleted if a leaf is attached to s_k. Call the event that s_k is attached to a leaf "$s_k(\mathcal{L})$" and the event that s_k has j descendant leaves "$s_k(j)$". For an example, see Figure 3.6.

We will calculate the probability of $s_k(\mathcal{D})$ in order to obtain the distribution of $\mathcal{A}^k_{n,m}$. We will need the following lemma [87].

Lemma 3.7.2. *Given a uniform distribution on ranked oriented trees on n species, the probability for one subtree below the root having k leaves, $k = 1, 2, \ldots, n-1$, is*

$$\frac{2}{(n-1)} \text{ for } k \neq n/2, \qquad \frac{1}{(n-1)} \text{ for } k = n/2.$$

Therefore, the probability for the left (resp. right) subtree below the root having k leaves, $k = 1, 2, \ldots, n-1$, is

$$\frac{1}{(n-1)}.$$

Theorem 3.7.3. *Pick a ranked oriented tree on n leaves uniformly at random. Delete a leaf uniformly at random from that tree. The probability that the deleted leaf was attached to the interior vertex with rank k is*

$$\mathbb{P}[s_k(\mathcal{D})] = \frac{2k}{n(n-1)}$$

for $1 \leq k \leq n-1$.

Proof. The probability of $s_k(\mathcal{D})$, $2 \leq k \leq n-2$ is,

$$\begin{aligned}
\mathbb{P}[s_k(\mathcal{D})] &= \mathbb{P}[s_k(\mathcal{D})|s_k(\mathcal{L})]\mathbb{P}[s_k(\mathcal{L})] \\
&= \sum_{j=2}^{n-k+1} \mathbb{P}[s_k(\mathcal{D})|s_k(\mathcal{L})]\mathbb{P}[s_k(\mathcal{L})|s_k(j)]\mathbb{P}[s_k(j)] \\
&= \sum_{j=2}^{n-k+1} \frac{2}{n(j-1)}\mathbb{P}[s_k(j)].
\end{aligned}$$

The last equality holds, since, for $j > 2$ the probability of having a leaf attached to the root in an oriented tree with j leaves is $2/(j-1)$ (Lemma 3.7.2) and the probability of choosing a specific leaf in a tree with n leaves is $1/n$ since we assume random taxon sampling. For $j = 2$, the probability of a leaf being attached to s_k is 1. The probability of choosing one of the leaves below s_k is $2/n$.

To calculate $\mathbb{P}[s_k(j)]$, note the following. Each ranked oriented tree on n leaves is equally likely. We will now count the number of trees with $s_k(j)$. There are $(j-1)!$ ranked oriented trees on j leaves. There are $(n-j-1)!$ trees on $n-j$ leaves. For fixed ranked oriented trees $\mathcal{T}_j, \mathcal{T}_{n-j}$ on $j, n-j$ leaves, the root s_k of $\mathcal{T}_j$ has to be attached to $\mathcal{T}_{n-j}$ in order to obtain a ranked oriented tree with $s_k(j)$ on n leaves. The ancestor a of s_k has possible ranks $i = 1, \ldots, k-1$. There are i different lineages in $\mathcal{T}_{n-j}$ to attach the vertex a with rank i and two orientations. So overall, there are $2\sum_{i=1}^{k-1} i = k(k-1)$ possibilities to attach s_k. The number of ways to order the $j-2$ interior vertices in $\mathcal{T}_j$ below s_k and the $n-j-1-(k-2)$ vertices in $\mathcal{T}_{n-j}$ of rank bigger than k is $\binom{n-j-k+1+j-2}{j-2} = \binom{n-k-1}{j-2}$. Therefore, there are $k(k-1)(j-1)!(n-j-1)!\binom{n-k-1}{j-2}$ possible ranked oriented trees on n leaves where s_k has j leaves descending. Since each ranked oriented tree is equally likely, and we have $(n-1)!$ ranked oriented trees on n leaves overall, we have

$$\mathbb{P}[s_k(j)] = k(k-1)\frac{(j-1)!(n-j-1)!}{(n-1)!}\binom{n-k-1}{j-2}.$$

Overall, with $\sum_{j=k}^{n}\binom{j}{k} = \binom{n+1}{k+1}$ (see e.g. [93], p.32), we have,

$$\begin{aligned}
\mathbb{P}[s_k(\mathcal{D})] &= \sum_{j=2}^{n-k+1} \frac{2}{n(j-1)}k(k-1)\frac{(j-1)!(n-j-1)!}{(n-1)!}\binom{n-k-1}{j-2} \\
&= \sum_{j=2}^{n-k+1} 2k(k-1)\frac{(n-j-1)!}{n!}\frac{(n-k-1)!}{(n-k-j+1)!} \\
&= 2k!\frac{(n-k-1)!}{n!}\sum_{j=2}^{n-k+1}\binom{n-j-1}{k-2} \\
&= 2k!\frac{(n-k-1)!}{n!}\sum_{j=k-2}^{n-3}\binom{j}{k-2} \\
&= 2k!\frac{(n-k-1)!}{n!}\binom{n-2}{k-1} = \frac{2k}{n(n-1)}.
\end{aligned}$$

So far we considered the cases $k = 2, \ldots, n-2$. Since the probability of having a leaf attached to the root in an oriented tree with n leaves is $2/(n-1)$, we have

$$\mathbb{P}[s_1(\mathcal{D})] = \frac{2}{n(n-1)}.$$

For $k = n-1$, we know that two leaves are attached to s_{n-1} (since it is the most recent speciation event), and therefore

$$\mathbb{P}[s_{n-1}(\mathcal{D})] = \frac{2}{n}.$$

□

Remark 3.7.4. The distribution established in the last theorem also appears in a different context. Consider a labeled tree on n species, and let x be a leaf label. Let N be the number of leaves in the subtree descending the root which contains x. The distribution of N under the Yule model is established in [92]:

$$\mathbb{P}[N = k] = \frac{2k}{n(n-1)}.$$

Therefore, we have the equality $\mathbb{P}[N = k] = \mathbb{P}[s_k(\mathcal{D})]$.

Consider an arbitrary model for speciation which induces a uniform distribution on ranked oriented trees. Under the chosen model, let $\mathcal{A}_n^k$ be the random variable "time of k-th speciation event in an oriented tree with n species," $1 \leq k \leq n-1$. Now delete one leaf uniformly at random from a tree with n leaves. Let $\mathcal{A}_{n,n-1}^k$ be the random variable "time of k-th speciation event in an oriented tree with $n-1$ uniformly sampled species, out of n species overall", $1 \leq k \leq n-2$. With probability $\mathbb{P}(s_j(\mathcal{D}))$, we delete speciation event j. If $j \leq k$ then $\mathcal{A}_{n,n-1}^k = \mathcal{A}_n^{k+1}$; if $j > k$ then $\mathcal{A}_{n,n-1}^k = \mathcal{A}_n^k$. Therefore, the density for the time of the k-th speciation event in the oriented tree with $n-1$ leaves after random taxon sampling is,

$$\begin{aligned} f_{\mathcal{A}_{n,n-1}^k}(s) &= \sum_{j=k+1}^{n-1} \mathbb{P}(s_j(\mathcal{D})) f_{\mathcal{A}_n^k}(s) + \sum_{j=1}^{k} \mathbb{P}(s_j(\mathcal{D})) f_{\mathcal{A}_n^{k+1}}(s), \qquad 1 \leq k \leq n-2 \\ &= \sum_{j=k+1}^{n-1} \frac{2j}{n(n-1)} f_{\mathcal{A}_n^k}(s) + \sum_{j=1}^{k} \frac{2j}{n(n-1)} f_{\mathcal{A}_n^{k+1}}(s) \\ &= \left(1 - \frac{(k+1)k}{n(n-1)}\right) f_{\mathcal{A}_n^k}(s) + \frac{(k+1)k}{n(n-1)} f_{\mathcal{A}_n^{k+1}}(s). \end{aligned}$$

With more non-sampled taxa, we can proceed recursively:

Theorem 3.7.5. *Let $\mathcal{A}_{n,m}^k$ be the random variable "k-th speciation event in an oriented tree with m uniformly sampled species, out of n species overall". We have,*

$$f_{\mathcal{A}_{n,m-1}^k}(s) = \left(1 - \frac{(k+1)k}{m(m-1)}\right) f_{\mathcal{A}_{n,m}^k}(s) + \frac{(k+1)k}{m(m-1)} f_{\mathcal{A}_{n,m}^{k+1}}(s).$$

for $1 \leq k \leq m-2$ and $2 \leq m \leq n$, and $\mathcal{A}_{n,n}^k := \mathcal{A}_n^k$.

Corollary 3.7.6. *The expectation of $\mathcal{A}^k_{n,m-1}$ is,*

$$\mathbb{E}[\mathcal{A}^k_{n,m-1}] = \left(1 - \frac{(k+1)k}{m(m-1)}\right)\mathbb{E}[\mathcal{A}^k_{n,m}] + \frac{(k+1)k}{m(m-1)}\mathbb{E}[\mathcal{A}^{k+1}_{n,m}].$$

3.7.3 Constant rate birth-death processes

With the results of the last section, we can calculate the expected speciation times under random taxon sampling for the cBDP. We will calculate the time of the k-th speciation event assuming a uniform prior for the time of origin; further we obtain the horizontal LTT plot. The calculations for a fixed age of the tree are analogous.

The expectation of $\mathcal{A}^k_n$ for the cBDP is stated in Theorem 3.4.16. With Corollary 3.7.6, we can calculate the expectation of $\mathcal{A}^k_{n,m}$. Note that for the cBDP, $\mathbb{E}[\mathcal{A}^k_n]$ is a function of $\rho := \mu/\lambda$ multiplied by $\frac{1}{\lambda}$ (Theorem 3.4.16) . So when scaling time, $\mathbb{E}[\mathcal{A}^k_n]$ only depends on $\rho = \mu/\lambda$.

In Figure 3.7, the expected LTT plots for $m = 10$ and different values for n, ρ are shown. We scale time such that the *mrca*, i.e. the first speciation event, is at time zero and the present is at time one. In Figure 3.8, we plot the expected age of the *mrca* – this visualizes how much scaling we do for the different parameter combinations n, ρ.

Remark 3.7.7. The cCBP is a cBDP with $\lambda = \mu$. After scaling time, the expected k-th speciation time in a tree with m leaves is invariant when sampling from bigger trees:

$$\begin{aligned}
\mathbb{E}[\mathcal{A}^k_{n,n-1}] &= \left(1 - \frac{(k+1)k}{n(n-1)}\right)\mathbb{E}[\mathcal{A}^k_n] + \frac{(k+1)k}{n(n-1)}\mathbb{E}[\mathcal{A}^{k+1}_n] \\
&\overset{\text{Thm. 3.4.16}}{=} \left(1 - \frac{(k+1)k}{n(n-1)}\right)\frac{n-k}{\lambda k} + \frac{(k+1)k}{n(n-1)}\frac{n-k-1}{\lambda(k+1)} \\
&= \frac{n(n-1-k)}{\lambda(n-1)k} = \frac{n}{n-1}\mathbb{E}[\mathcal{A}^k_{n-1}].
\end{aligned}$$

Note that the invariance does not hold for a general cBDP as displayed in Figure 3.7. Further, it does not hold if conditioning the cCBP on an age t as displayed in Figure 3.9. Note that for obtaining this figure, we used the analytic results in Theorem 3.4.13, so again no simulations were involved.

Further, we derive a closed form equation for the expectation of $\mathcal{A}^k_{n,m}$ under the cCBP.

Theorem 3.7.8. *Consider the cCBP. We have*

$$\mathbb{E}[\mathcal{A}^k_{n,m}] = \frac{n}{m}\mathbb{E}[\mathcal{A}^k_m] = \frac{n(m-k)}{\lambda m k}. \tag{3.27}$$

i.e. for m, k fixed, the expectation increases linear with n.

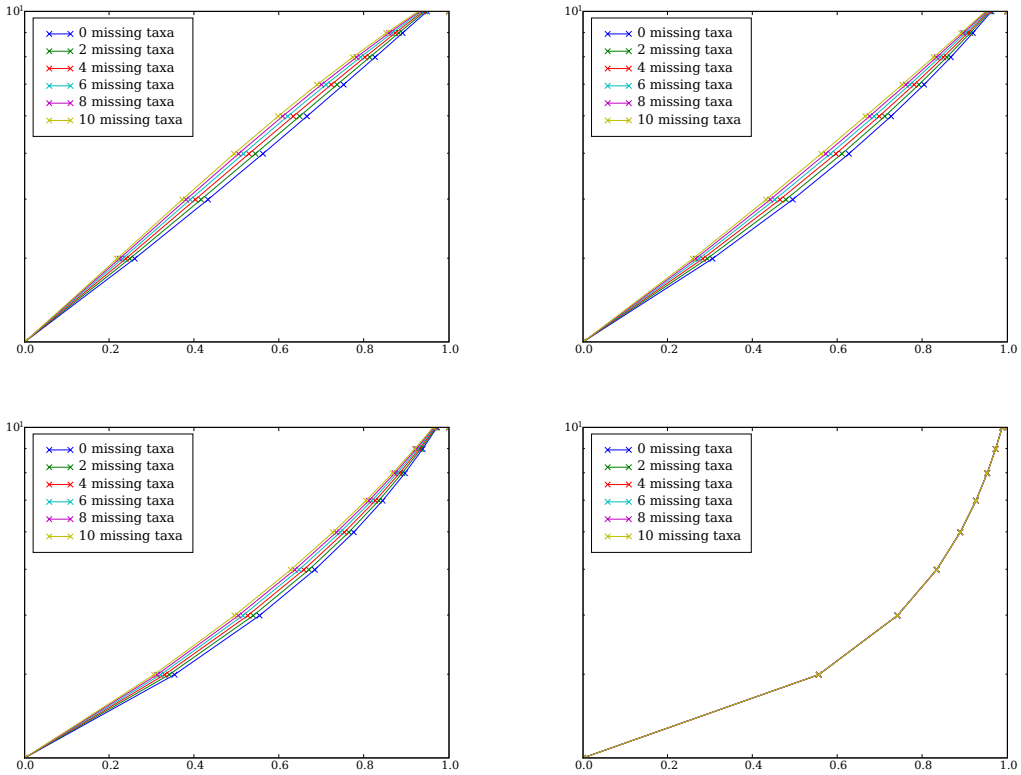

Figure 3.7: Expected speciation times for $\rho = 0, 1/2, 3/4, 1$ (left to right and top to bottom) and $m = 10$ species today (time is on x-axis, number of species is on y-axis). We sample leaves from bigger trees as labeled in the figure. For each parameter combination, time is scaled such that the *mrca* is at time zero and today is one.

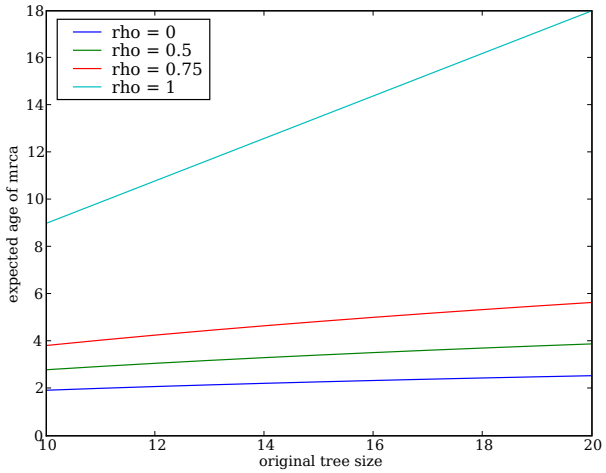

Figure 3.8: Expected age of the *mrca* in a tree with 10 leaves for different values of ρ and original tree sizes variing between 10 and 20. Note that under the cCBP we have a linear curve.

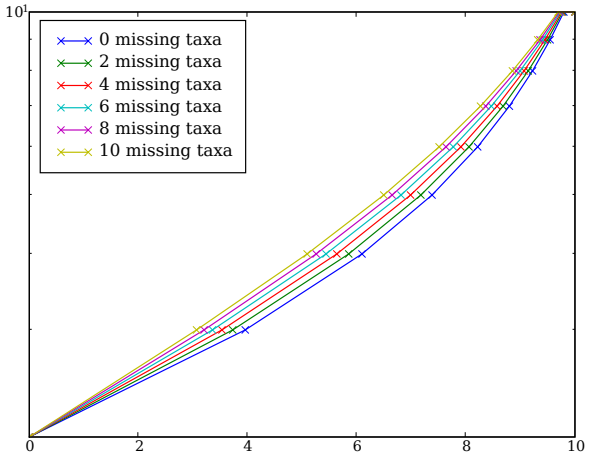

Figure 3.9: Expected speciation times for $\mu = \lambda = 1$ conditioning on the age of the tree being $t = 10$ and $m = 10$ species today (time is on x-axis, number of species is on y-axis). We sample leaves from bigger trees as labeled in the figure.

Proof. We prove the theorem by induction on m. Equation (3.27) holds for $m = n$, since $\mathcal{A}^k_{m,m} = \mathcal{A}^k_m$ by definition and $\mathbb{E}[\mathcal{A}^k_m] = \frac{m-k}{\lambda k}$ (Theorem 3.4.16).

Assume Equation (3.27) holds for all M with $n \geq M \geq m$. Then the expectation of $\mathcal{A}^k_{n,m-1}$ is with Corollary 3.7.6,

$$\begin{aligned}
\mathbb{E}[\mathcal{A}^k_{n,m-1}] &= \left(1 - \frac{(k+1)k}{m(m-1)}\right)\mathbb{E}[\mathcal{A}^k_{n,m}] + \frac{(k+1)k}{m(m-1)}\mathbb{E}[\mathcal{A}^{k+1}_{n,m}] \\
&= \left(1 - \frac{(k+1)k}{m(m-1)}\right)\frac{n(m-k)}{\lambda mk} + \frac{(k+1)k}{m(m-1)}\frac{n(m-k-1)}{\lambda m(k+1)} \\
&= \frac{n}{m-1}\frac{m-k-1}{\lambda k} = \frac{n}{m-1}\mathbb{E}[\mathcal{A}^k_{m-1}].
\end{aligned}$$

□

There have been methods proposed to estimate the birth and death parameter, λ and μ, for a phylogeny from the information contained in the LTT plot [71]; the method assumes complete taxon sampling. As we see in Figure 3.7, the LTT plot becomes "less convex" under incomplete taxon sampling. Further, the LTT plot becomes "less convex" for decreasing μ. Therefore, the death parameter μ is underestimated if taxon sampling is incomplete. In particular, a cBDP with a positive death rate and many non-sampled taxa looks like a cBDP with $\mu = 0$ and complete taxon sampling.

Remark 3.7.9. The expected coalescent times under the coalescent model [53, 51, 52] equal the expected speciation times under the cCBP (Section 3.5.2). Therefore, Figure 3.7 (bottom right) and Figure 3.8 (straight line) show the expected behavior of the coalescent.

3.8 Vertical LTT plots

In the previous sections, we calculated the time of the k-th speciation event. Another approach for obtaining LTT plots is to calculate the number of species at a fixed time. In Section 3.8.1, this is done for trees of known age, in Section 3.8.2, a uniform prior is assumed. We assume complete taxon sampling.

3.8.1 LTT plots for trees of known age

In a reconstructed oriented tree of age t, we calculate the density and expectation of the number of species at time σt after origin, $M_{\sigma,t}$ ($\sigma \in [0,1]$). We condition $M_{\sigma,t}$ on $M_{1,t} = n$, i.e. having n species today.

Theorem 3.8.1. *In a reconstructed oriented tree of age t with n extant species, the probability that at time σt after origin ($\sigma \in [0,1]$), we have exactly m species is,*

$$\mathbb{P}[M_{\sigma,t} = m | M_{1,t} = n] = \begin{cases} \binom{n-1}{m-1}\frac{f(\sigma,t,\rho,\delta)^{m-1}}{(1+f(\sigma,t,\rho,\delta))^{n-1}} & \text{if } m \leq n \\ 0 & \text{else} \end{cases}$$

with $f(\sigma,t,\rho,\delta)=(1-\rho)\frac{(1-e^{-\sigma\delta t})e^{-(1-\sigma)\delta t}}{(1-e^{-(1-\sigma)\delta t})(1-\rho e^{-\delta t})}$ *and* $\rho=\mu/\lambda, \delta=\mu-\lambda$.

Proof. We will need the following functions as defined in [71]:

$$P(t) := \frac{\lambda-\mu}{\lambda-\mu e^{-(\lambda-\mu)t}}, \tag{3.28}$$

$$u(t) := \lambda\frac{1-e^{-(\lambda-\mu)t}}{\lambda-\mu e^{-(\lambda-\mu)t}}. \tag{3.29}$$

Since we are considering reconstructed trees, we obviouly have $\mathbb{P}[M_{\sigma,t}=m|M_{1,t}=n]=0$ if $m>n$. For $m\leq n$, we have with Bayes' law,

$$\mathbb{P}[M_{\sigma,t}=m|M_{1,t}=n]=\mathbb{P}[M_{1,t}=n|M_{\sigma,t}=m]\frac{\mathbb{P}[M_{\sigma,t}=m]}{\mathbb{P}[M_{1,t}=n]}. \tag{3.30}$$

The probability that a lineage at time σt after origin has m descendants today (i.e. after time t) is [46]

$$\mathbb{P}[M_{1,(1-\sigma)t}=m]=(1-u((1-\sigma)t))u((1-\sigma)t)^{m-1}.$$

Therefore, with N being the normalizing constant, and $\mathbf{e}=(1,1,\ldots,1)^T$, we get

$$\begin{aligned}
\mathbb{P}[M_{1,t}=n|M_{\sigma,t}=m] &= \frac{1}{N}\sum_{\substack{i\in\mathbb{N}^m\\ i^T\mathbf{e}=n}}\prod_{k=1}^{m}\mathbb{P}[M_{1,(1-\sigma)t}=i_k]\\
&= \frac{1}{N}\sum_{\substack{i\in\mathbb{N}^m\\ i^T\mathbf{e}=n}}\prod_{k=1}^{m}(1-u((1-\sigma)t))u((1-\sigma)t)^{i_k-1}\\
&= \frac{1}{N}\sum_{\substack{i\in\mathbb{N}^m\\ i^T\mathbf{e}=n}}(1-u((1-\sigma)t))^m u((1-\sigma)t)^{n-m}\\
&= \frac{1}{N}|\{i\in\mathbb{N}^m : i^T\mathbf{e}=n\}|(1-u((1-\sigma)t))^m u((1-\sigma)t)^{n-m}.
\end{aligned}$$

We determine $|\{i\in\mathbb{N}^m : i^T\mathbf{e}=n\}|$. Since we have $i_k\geq 1, k=1,\ldots,m$, determining $|\{i\in\mathbb{N}^m : i^T\mathbf{e}=n\}|$ is equivalent to counting in how many ways we can distribute $n-m$ ones to m components. Distributing the $n-m$ ones to m components is equivalent to drawing $n-m$ times from a urn with m different balls and returning the ball to the urn after each draw. There are $\binom{n-m+m-1}{n-m}=\binom{n-1}{m-1}$ different outcomes. So $|\{i\in\mathbb{N}^m : i^T\mathbf{e}=n\}|=\binom{n-1}{m-1}$. Therefore,

$$\mathbb{P}[M_{1,t}=n|M_{\sigma,t}=m]=\frac{1}{N}\binom{n-1}{m-1}(1-u((1-\sigma)t))^m u((1-\sigma)t)^{n-m}.$$

In [71], the authors establish (Equation (9) and (3)),

$$\begin{aligned}
\mathbb{P}[M_{\sigma,t}=m] &= \left(1-u(\sigma t)\frac{P(t)}{P(\sigma t)}\right)\left(u(\sigma t)\frac{P(t)}{P(\sigma t)}\right)^{m-1},\\
\mathbb{P}[M_{1,t}=n] &= P(t)(1-u(t))u(t)^{n-1}.
\end{aligned}$$

Plugging these equations into Equation (3.30) yields

$$\begin{aligned}
&\mathbb{P}[M_{\sigma,t} = m | M_{1,t} = n] \\
= \;& \frac{1}{N}\binom{n-1}{m-1}(1-u((1-\sigma)t))^m u((1-\sigma)t)^{n-m} \frac{\left(1-u(\sigma t)\frac{P(t)}{P(\sigma t)}\right)\left(u(\sigma t)\frac{P(t)}{P(\sigma t)}\right)^{m-1}}{P(t)(1-u(t))u(t)^{n-1}} \\
= \;& \frac{1}{N}\binom{n-1}{m-1}\left(u(\sigma t)\frac{1-u((1-\sigma)t)}{u((1-\sigma)t)}\frac{P(t)}{P(\sigma t)}\right)^{m-1} \\
& u((1-\sigma)t)^{n-1}(1-u((1-\sigma)t))\frac{\left(1-u(\sigma t)\frac{P(t)}{P(\sigma t)}\right)}{P(t)(1-u(t))u(t)^{n-1}} \\
= \;& \frac{1}{N}\binom{n-1}{m-1}\left(u(\sigma t)\frac{1-u((1-\sigma)t)}{u((1-\sigma)t)}\frac{P(t)}{P(\sigma t)}\right)^{m-1} g_{\sigma,t,n}
\end{aligned}$$

where $g_{\sigma,t,n} = u((1-\sigma)t)^{n-1}(1-u((1-\sigma)t))\frac{\left(1-u(\sigma t)\frac{P(t)}{P(\sigma t)}\right)}{P(t)(1-u(t))u(t)^{n-1}}$. In the following, we determine N. Since probabilities add up to 1, we have $\sum_{m=1}^{n}\mathbb{P}[M_{\sigma,t} = m | M_{1,t} = n] = 1$. We have with the binomial theorem,

$$N = N\sum_{m=1}^{n}\mathbb{P}[M_{\sigma,t} = m | M_{1,t} = n] = \left(1+u(\sigma t)\frac{1-u((1-\sigma)t)}{u((1-\sigma)t)}\frac{P(t)}{P(\sigma t)}\right)^{n-1} g_{\sigma,t,n}.$$

Therefore,

$$\mathbb{P}[M_{\sigma,t} = m | M_{1,t} = n] = \binom{n-1}{m-1}\frac{\left(u(\sigma t)\frac{1-u((1-\sigma)t)}{u((1-\sigma)t)}\frac{P(t)}{P(\sigma t)}\right)^{m-1}}{\left(1+u(\sigma t)\frac{1-u((1-\sigma)t)}{u((1-\sigma)t)}\frac{P(t)}{P(\sigma t)}\right)^{n-1}}.$$

We evaluate

$$u(\sigma t)\frac{1-u((1-\sigma)t)}{u((1-\sigma)t)}\frac{P(t)}{P(\sigma t)} \;=\; (\lambda-\mu)\frac{(1-e^{-(\lambda-\mu)\sigma t})e^{-(\lambda-\mu)((1-\sigma)t)}}{(1-e^{-(\lambda-\mu)((1-\sigma)t)})(\lambda-\mu e^{-(\lambda-\mu)t})}$$

with $P(t)$ and $u(t)$ from Equation (3.28) and (3.29). We define,

$$f(\sigma,t,\rho,\delta) := u(\sigma t)\frac{1-u((1-\sigma)t)}{u((1-\sigma)t)}\frac{P(t)}{P(\sigma t)} = (1-\rho)\frac{(1-e^{-\sigma\delta t})e^{-(1-\sigma)\delta t}}{(1-e^{-(1-\sigma)\delta t})(1-\rho e^{-\delta t})}.$$

Therefore,

$$\mathbb{P}[M_{\sigma,t} = m | M_{1,t} = n] = \binom{n-1}{m-1}\frac{f(\sigma,t,\rho,\delta)^{m-1}}{(1+f(\sigma,t,\rho,\delta))^{n-1}}$$

which establishes the theorem. □

Remark 3.8.2. Note that $f(\sigma,t,\rho,\delta) = f(\sigma,\delta t,\rho,1)$. Therefore, the conditional distribution $\mathbb{P}[M_{\sigma,t} = m | M_{1,t} = n]$ with parameters ρ,δ is the same as $\mathbb{P}[M_{\sigma,\delta t} = m | M_{1,\delta t} = n]$ with parameters $\rho, 1$.

Corollary 3.8.3. *The expectation of $M_{\sigma,t}$ given $M_{1,t} = n$ is,*

$$\mathbb{E}[M_{\sigma,t}|M_{1,t} = n] = \frac{1 + nf(\sigma, t, \rho, \delta)}{1 + f(\sigma, t, \rho, \delta)}.$$

Proof. From Theorem 3.8.1, we get

$$\begin{aligned}
\mathbb{E}[M_{\sigma,t}|M_{1,t} = n] &= \sum_{m=1}^{n} m\mathbb{P}[M_{\sigma,t} = m|M_{1,t} = n] \\
&= \frac{1}{(1 + f(\sigma, t, \rho, \delta))^{n-1}} \sum_{m=0}^{n-1} (m+1) \binom{n-1}{m} f(\sigma, t, \rho, \delta)^m \\
&= \frac{1}{(1 + f(\sigma, t, \rho, \delta))^{n-1}} \left[(1 + f(\sigma, t, \rho, \delta))^{n-1} + \sum_{m=1}^{n-1} m \binom{n-1}{m} f(\sigma, t, \rho, \delta)^m \right] \\
&= 1 + \frac{(n-1)f(\sigma, t, \rho, \delta)}{(1 + f(\sigma, t, \rho, \delta))^{n-1}} \sum_{m=1}^{n-1} \binom{n-2}{m-1} f(\sigma, t, \rho, \delta)^{m-1} \\
&= 1 + \frac{(n-1)f(\sigma, t, \rho, \delta)(1 + f(\sigma, t, \rho, \delta))^{n-2}}{(1 + f(\sigma, t, \rho, \delta))^{n-1}} \\
&= \frac{1 + nf(\sigma, t, \rho, \delta)}{1 + f(\sigma, t, \rho, \delta)}
\end{aligned}$$

which establishes the corollary. □

Note that for a fixed n, the conditional expectation $\mathbb{E}[M_{\sigma,t}|M_{1,t} = n]$ only depends on ρ and δt. For $\rho = 0, 1/4, 1/2, 3/4, 1$, $t = 10$ and varying values of δ, we calculated the expectation, see Figure 3.10. The graph looks quite unfamiliar for an LTT plot of a reconstructed tree since we have concave curves, and the Yule model is – for large λ – "more convex" than models with extinction.

This behavior is due to conditioning on both t and n: Consider the curves for arbitrary λ and $\mu = 0$. We condition on the age t of the tree. If λ is very large, i.e. the process (if not conditioned on n) will have more than n lineages at time t with high probability, then the most likely trees with n species are the trees where nothing happens at the beginning, and later we have speciation. If many speciation events occur at the beginning, we would later allow all species to only speciate rarely, since we want to end up with n species. This is very unlikely though, since λ is big. If at the beginning, the one lineage does not speciate, and after a while, we have "normal" speciation, this is much more likely, since we only force the first lineages to behave abnormally. This yields a "very convex" LTT plot.

In the case of λ being small compared to t, we need the early lineages to speciate a lot. Then the later lineages can behave quite "normal" in order to end up with n lineages today. This yields "very concave" LTT plots.

The following result had already been established with a different approach in [71].

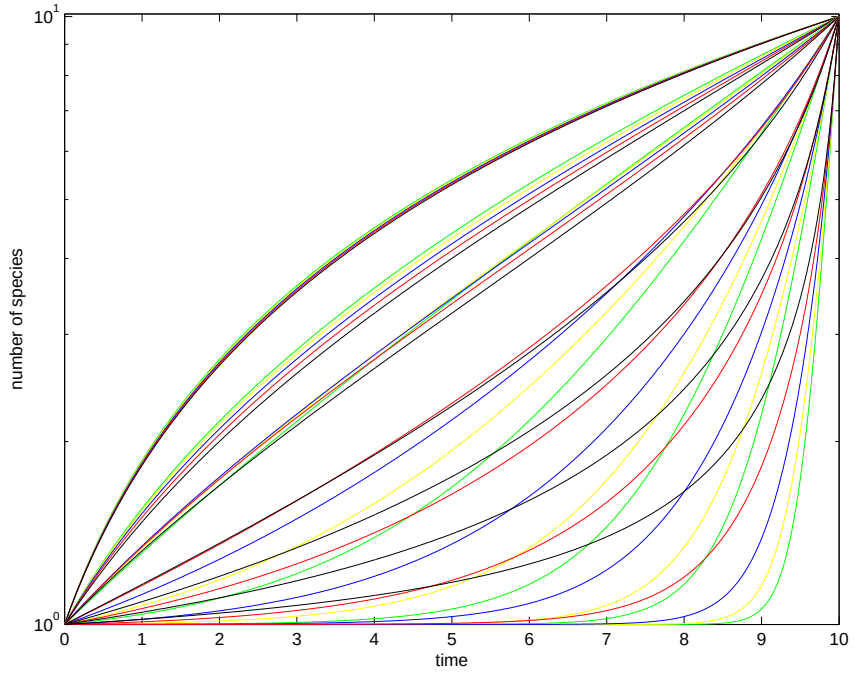

Figure 3.10: Expected number of species given we have $n = 10$ species at time $t = 10$ (the present) for $\lambda = 5, 2, 1, 0.5, 0.2, 0.1, 0.01$, from bottom to top. The different colours correspond to, green: $\rho = 0$, yellow: $\rho = 1/4$, blue: $\rho = 1/2$, red: $\rho = 3/4$, black: $\rho = 1$.

Corollary 3.8.4. *The expected number of species at time σt after origin conditioned on the process surviving until t is*

$$\mathbb{E}[M_{\sigma,t}|M_{1,t} > 0] = e^{(\lambda-\mu)\sigma t}\frac{\lambda - \mu e^{-(\lambda-\mu)t}}{\lambda - \mu e^{-(\lambda-\mu)(1-\sigma t)}}.$$

Proof. From [46], we have $\mathbb{P}[M_{1,t} = n|M_{1,t} > 0] = (1 - u(t))u(t)^{n-1}$. We can write the expectation as,

$$\begin{aligned}
\mathbb{E}[M_{\sigma,t}|M_{1,t} > 0] &= \sum_{n=1}^{\infty} \mathbb{E}[M_{\sigma,t}|M_{1,t} = n]\mathbb{P}[M_{1,t} = n|M_{1,t} > 0] \\
&= \sum_{n=1}^{\infty} \frac{1 + nf(\sigma,t,\rho,\delta)}{1 + f(\sigma,t,\rho,\delta)}\lambda^{n-1}(\lambda-\mu)\frac{e^{-(\lambda-\mu)t}(1 - e^{-(\lambda-\mu)t})^{n-1}}{(\lambda - \mu e^{-(\lambda-\mu)t})^n} \\
&= \frac{(\lambda-\mu)e^{-(\lambda-\mu)t}}{(1 + f(\sigma,t,\rho,\delta))(\lambda - \mu e^{-(\lambda-\mu)t})} \sum_{n=0}^{\infty} \left(\frac{\lambda(1 - e^{-(\lambda-\mu)t})}{\lambda - \mu e^{-(\lambda-\mu)t}}\right)^n \\
&\quad + \frac{(\lambda-\mu)e^{-(\lambda-\mu)t}f(\sigma,t,\rho,\delta)}{(1 + f(\sigma,t,\rho,\delta))(\lambda - \mu e^{-(\lambda-\mu)t})} \sum_{n=1}^{\infty} n\left(\frac{\lambda(1 - e^{-(\lambda-\mu)t})}{\lambda - \mu e^{-(\lambda-\mu)t}}\right)^{n-1}.
\end{aligned}$$

Evaluating the geometric series yields,

$$\begin{aligned}\mathbb{E}[M_{\sigma,t}|M_{1,t}>0] &= \frac{(\lambda-\mu)e^{-(\lambda-\mu)t}}{(1+f(\sigma,t,\rho,\delta))(\lambda-\mu e^{-(\lambda-\mu)t})}\frac{1}{1-\frac{\lambda(1-e^{-(\lambda-\mu)t})}{\lambda-\mu e^{-(\lambda-\mu)t}}} \\ &\quad + \frac{(\lambda-\mu)e^{-(\lambda-\mu)t}f(\sigma,t,\rho,\delta)}{(1+f(\sigma,t,\rho,\delta))(\lambda-\mu e^{-(\lambda-\mu)t})}\frac{(\lambda-\mu e^{-(\lambda-\mu)t})^2}{\lambda(\lambda-\mu)^2e^{-(\lambda-\mu)t}} \\ &\quad \times\frac{d}{dt}\left(\frac{1}{1-\frac{\lambda(1-e^{-(\lambda-\mu)t})}{\lambda-\mu e^{-(\lambda-\mu)t}}}\right) \\ &= \frac{1}{1+f(\sigma,t,\rho,\delta)} + \frac{f(\sigma,t,\rho,\delta)(\lambda-\mu e^{-(\lambda-\mu)t})}{(1+f(\sigma,t,\rho,\delta))(\lambda-\mu)e^{-(\lambda-\mu)t}}.\end{aligned}$$

The corollary follows by plugging the definition of $f(\sigma,t,\rho,\delta)$ into the derived expression. □

Conditioning on the most recent common ancestor

So far, we conditioned on the time of origin of our tree. In other situations, we might know the time of the *mrca* of the extant species opposed to the time of origin.

Let $M^{mrca}_{\sigma,t}$ be the random variable "number of lineages in a reconstructed oriented tree at time σt after the *mrca* given the time since the *mrca* is t".

Corollary 3.8.5. *For $M^{mrca}_{\sigma,t}$, we have the following conditional density,*

$$\mathbb{P}[M^{mrca}_{\sigma,t}=m|M^{mrca}_{1,t}=n] = \frac{1}{n-1}\frac{f(\sigma,t,\rho,\delta)^{m-2}}{(1+f(\sigma,t,\rho,\delta)^{n-2}}\sum_{k=1}^{n-1}\sum_{l=1}^{k}\binom{k-1}{l-1}\binom{n-k-1}{m-l-1}.$$

for $m\leq n$ and $\mathbb{P}[M^{mrca}_{\sigma,t}=m|M^{mrca}_{1,t}=n]=0$ otherwise.

Proof. Label the two daughter trees of the reconstructed oriented tree $\mathcal{T}$ by $\mathcal{T}_1,\mathcal{T}_2$, these two trees originate at the time of the *mrca* of $\mathcal{T}$, and together they have n leaves. The random variable "number of leaves of tree $\mathcal{T}_i$ (having age t) at time σt" is $M^{\mathcal{T}_i}_{\sigma,t}$, $i\in\{1,2\}$. The probability that $\mathcal{T}_1$ has k leaves ($k=1,2,\dots n-1$) is $\frac{1}{n-1}$ (Lemma 3.7.2). Therefore,

$$\begin{aligned}&\mathbb{P}[M^{mrca}_{\sigma,t}=m|M^{mrca}_{1,t}=n] \\ =\ & \frac{1}{n-1}\sum_{k=1}^{n-1}\mathbb{P}[M^{mrca}_{\sigma,t}=m|M^{\mathcal{T}_1}_{1,t}=k,M^{\mathcal{T}_2}_{1,t}=n-k] \\ =\ & \frac{1}{n-1}\sum_{k=1}^{n-1}\sum_{l=1}^{k}\mathbb{P}[M^{\mathcal{T}_1}_{\sigma,t}=l,M^{\mathcal{T}_2}_{\sigma,t}=m-l|M^{\mathcal{T}_1}_{1,t}=k,M^{\mathcal{T}_2}_{1,t}=n-k] \\ =\ & \frac{1}{n-1}\sum_{k=1}^{n-1}\sum_{l=1}^{k}\mathbb{P}[M^{\mathcal{T}_1}_{\sigma,t}=l|M^{\mathcal{T}_1}_{1,t}=k]\mathbb{P}[M^{\mathcal{T}_2}_{\sigma,t}=m-l|M^{\mathcal{T}_2}_{1,t}=n-k] \\ =\ & \frac{1}{n-1}\frac{f(\sigma,t,\rho,\delta)^{m-2}}{(1+f(\sigma,t,\rho,\delta)^{n-2}}\sum_{k=1}^{n-1}\sum_{l=1}^{k}\binom{k-1}{l-1}\binom{n-k-1}{m-l-1}\end{aligned}$$

which establishes the theorem. □

Corollary 3.8.6. *For $M_{\sigma,t}^{mrca}$, we have the following conditional expectation,*

$$\mathbb{E}[M_{\sigma,t}^{mrca}|M_{1,t}^{mrca}=n] \;=\; \frac{2+nf(\sigma,t,\rho,\delta)}{1+f(\sigma,t,\rho,\delta)}.$$

Proof. Label the two daughter trees of the reconstructed oriented tree $\mathcal{T}$ by $\mathcal{T}_1, \mathcal{T}_2$, these two trees originate at the time of the *mrca* of $\mathcal{T}$, and together they have n leaves. The random variable "number of leaves of tree $\mathcal{T}_i$ (having age t) at time σt" is $M_{\sigma,t}^{\mathcal{T}_i}$, $i \in \{1,2\}$. Since the probability of $\mathcal{T}_1$ having k leaves ($k = 1, 2, \ldots n-1$) is $\frac{1}{n-1}$ (Lemma 3.7.2), we have

$$\begin{aligned}
\mathbb{E}[M_{\sigma,t}^{mrca}|M_{1,t}^{mrca}=n] &= \frac{1}{n-1}\sum_{k=1}^{n-1}[\mathbb{E}[M_{\sigma,t}^{\mathcal{T}_1}|M_{1,t}^{\mathcal{T}_1}=k]+\mathbb{E}[M_{\sigma,t}^{\mathcal{T}_2}|M_{1,t}^{\mathcal{T}_2}=n-k]] \\
&= \frac{1}{n-1}\sum_{k=1}^{n-1}\frac{2+nf(\sigma,t,\rho,\delta)}{1+f(\sigma,t,\rho,\delta)} \\
&= \frac{2+nf(\sigma,t,\rho,\delta)}{1+f(\sigma,t,\rho,\delta)}
\end{aligned}$$

which completes the proof. □

3.8.2 LTT plots for trees of unknown age

So far, we assumed that the time since origin is known to be t. We then calculated the expected number of species for each point in time between the origin and today. We will now assume a uniform prior on $(0,\infty)$ for the time of origin. Let M_σ be the random variable "number of lineages in a reconstructed oriented tree when the fraction σ of the time between the origin and the present has passed". We obtain:

Remark 3.8.7. The probability of having m lineages in the reconstructed oriented tree after the fraction $\sigma \in [0,1]$ of time between the origin and the present has passed, given n species at the present, is:

$$\mathbb{P}[M_\sigma = m|M_1 = n] = \int_0^\infty \mathbb{P}[M_{\sigma,t} = m|M_{1,t} = n]q_{or}(t|n)dt.$$

We did not find an analytic expression for this integral.

Theorem 3.8.8. *The expectation of M_σ given $M_1 = n$ is*

$$\mathbb{E}[M_\sigma|M_1=n] = \begin{cases} 1+n(n-1)\sum_{k=0}^{n-2}\binom{n-2}{k}\frac{(-1)^k}{k+2}\frac{\sigma}{k+2-\sigma} & \text{if } \mu=0, \\ 1+n(n-1)\sigma\int_0^\infty \frac{t^{n-1}}{(1+(1-\sigma)t)(1+t)^{n+1}}dt & \text{if } \mu=\lambda, \\ 1+n(n-1)(1-\rho)^2\int_0^\infty \frac{e^{-(2-\sigma)t}-e^{-2t}}{1-\rho e^{-(1-\sigma)t}}\frac{(1-e^{-t})^{n-2}}{(1-\rho e^{-t})^{n+1}}dt & \text{else.} \end{cases}$$

Proof. We have,

$$\begin{aligned}
& \mathbb{E}[M_\sigma|M_1=n] \\
= & \int_0^\infty \mathbb{E}[M_{\sigma,t}|M_{1,t}=n]q_{or}(t|n)dt \\
= & \int_0^\infty \left(1+\frac{(n-1)f(\sigma,t,\rho,\delta)}{1+f(\sigma,t,\rho,\delta)}\right)\left(n\lambda^n(\lambda-\mu)^2\frac{(1-e^{-(\lambda-\mu)t})^{n-1}e^{-(\lambda-\mu)t}}{(\lambda-\mu e^{-(\lambda-\mu)t})^{n+1}}\right)dt \\
= & 1+n(n-1)\lambda^n(\lambda-\mu)^3 \\
& \times\int_0^\infty \frac{e^{-(\lambda-\mu)(2-\sigma)t}-e^{-(\lambda-\mu)2t}}{\lambda-\mu e^{-(\lambda-\mu)(1-\sigma)t}}\frac{(1-e^{-(\lambda-\mu)t})^{n-2}}{(\lambda-\mu e^{-(\lambda-\mu)t})^{n+1}}dt \qquad (3.31) \\
\overset{\lambda\neq\mu}{=} & 1+n(n-1)\lambda^n(\lambda-\mu)^2\int_0^\infty \frac{e^{-(2-\sigma)t}-e^{-2t}}{\lambda-\mu e^{-(1-\sigma)t}}\frac{(1-e^{-t})^{n-2}}{(\lambda-\mu e^{-t})^{n+1}}dt \\
\overset{\rho:=\mu/\lambda}{=} & 1+n(n-1)(1-\rho)^2\int_0^\infty \frac{e^{-(2-\sigma)t}-e^{-2t}}{1-\rho e^{-(1-\sigma)t}}\frac{(1-e^{-t})^{n-2}}{(1-\rho e^{-t})^{n+1}}dt \qquad (3.32)
\end{aligned}$$

Note that the expectation only depends on $\rho=\mu/\lambda$. In general, we could not find an analytic solution for the integral. However, for the Yule model, $\mu=0$, we can evaluate the integral. From Equation (3.32), we get

$$\begin{aligned}
\mathbb{E}_{Yule}[M_\sigma|M_1=n] &= 1+n(n-1)\int_0^\infty (e^{-(2-\sigma)t}-e^{-2t})(1-e^{-t})^{n-2}dt \\
&= 1+n(n-1)\sum_{k=0}^{n-2}\binom{n-2}{k}(-1)^k\int_0^\infty (e^{-(k+2-\sigma)t}-e^{-(k+2)t})dt \\
&= 1+n(n-1)\sum_{k=0}^{n-2}\binom{n-2}{k}(-1)^k\left[-\frac{1}{k+2-\sigma}e^{-(k+2-\sigma)t}+\frac{1}{k+2}e^{-(k+2)t}\right]_0^\infty \\
&= 1+n(n-1)\sum_{k=0}^{n-2}\binom{n-2}{k}\frac{(-1)^k}{k+2}\frac{\sigma}{k+2-\sigma}.
\end{aligned}$$

For the cCBP, i.e. $\lambda=\mu$, we observe with the property $e^{-\epsilon}\sim 1-\epsilon$ for $\epsilon\to 0$, from Equation (3.31),

$$\begin{aligned}
& \mathbb{E}_{CBP}[M_\sigma|M_1=n] \\
= & \lim_{\mu\to\lambda}\left(1+n(n-1)\int_0^\infty \frac{\lambda^n(\lambda-\mu)^3((\lambda-\mu)\sigma t)((\lambda-\mu)t)^{n-2}}{(\lambda-\mu(1-(\lambda-\mu)(1-\sigma)t))(\lambda-\mu(1-(\lambda-\mu)t))^{n+1}}\right) \\
= & 1+n(n-1)\int_0^\infty \frac{\lambda^n\sigma t^{n-1}}{(1+\lambda(1-\sigma)t)(1+\lambda t)^{n+1}}dt \\
= & 1+n(n-1)\sigma\int_0^\infty \frac{t^{n-1}}{(1+(1-\sigma)t)(1+t)^{n+1}}dt.
\end{aligned}$$

This establishes the theorem. □

Note that the expectation of M_σ only depends on ρ (i.e. is independent of δ). In particular, the expectation for $\mu=0$ and $\mu=\lambda$ is independent of λ. The expectation

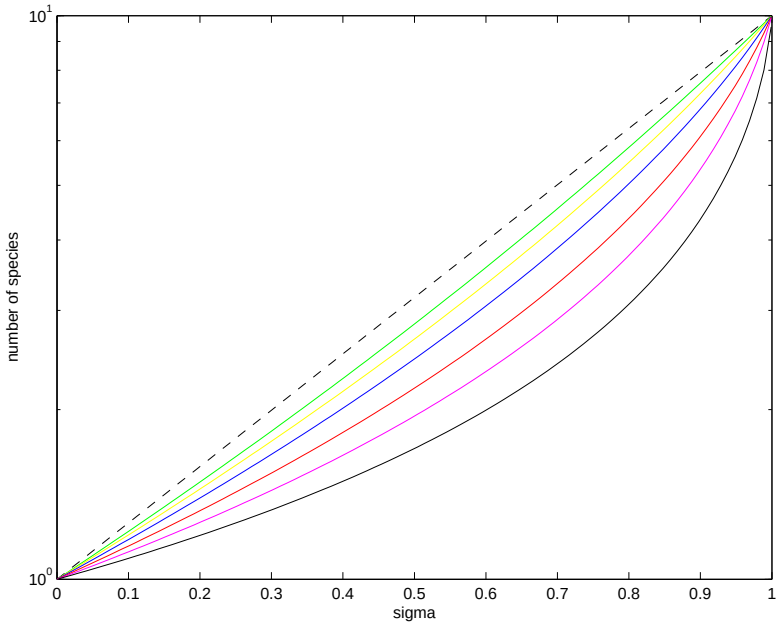

Figure 3.11: Expected number of species given we have $n = 10$ species today. According to Theorem 3.8.8, the expectation only depends on $\rho = \mu/\lambda$, we calculated $\rho = 0, 1/4, 1/2, 3/4, 9/10, 1$ (from top to bottom). The dashed black line is the straight line. Note that the curve for the Yule model is "more convex" than the straight line.

of M_σ is plotted for different values of ρ, see Figure 3.11. The numerical integration was done with the Matlab ode45 tool.

3.9 Estimating divergence times in partially dated phylogenies

We estimate the speciation times in a given phylogeny by the expected speciation time of a vertex (Section 3.1.1). This estimate does not have a bias opposed to the method in [75]. However, in reconstructed phylogenies, we often have some speciation events dated. The approach in Section 3.1.1 does not use the information of the dated vertices for estimating the other dates, the approach only uses the tree shape information.

Assuming the cBDP, we calculate in this section the density and expectation for the time of any undated interior vertex in a given oriented tree $\mathcal{T}$, conditioning on the dates $t_{v_1} > \ldots > t_{v_m}$ for the dated vertices $v_1, \ldots, v_m$ ($v_i \in \mathring{V}$, $i = 1, \ldots, m$). Let v_0 be the origin of the tree and v_{m+1} an arbitrary leaf.

For an undated vertex v in an oriented tree $\mathcal{T}$ on n species, the density of the time t_v of vertex v conditioning on the times $t_{v_1} > \ldots > t_{v_m}$ of vertices $v_1, \ldots, v_m$ can be calculated as follows. Let $r(\mathcal{T})$ be the set of rank functions on $\mathcal{T}$, and for a

given rank function r, $r(v)$ is the rank of vertex v. Let x_k, $k = 1, \ldots, n-1$ be the time of speciation event with rank k. Let t_{or} be the time of origin and $x_n = 0$ be today. We have for the density of t_v, conditioned on the speciation times $t_{v_1}, \ldots, t_{v_m}$, the time of origin t_{or} and the oriented tree $\mathcal{T}$,

$$\begin{aligned} f_{t_v}(s|t_{v_1}, \ldots, t_{v_m}, t_{or}, \mathcal{T}) &= \sum_{r \in r(\mathcal{T})} f_{t_v}(s, r|t_{v_1}, \ldots, t_{v_m}, t_{or}, \mathcal{T}) \\ &= \sum_{r \in r(\mathcal{T})} f_{t_v}(s|r, t_{v_1}, \ldots, t_{v_m}, t_{or}, \mathcal{T}) f(r|t_{v_1}, \ldots, t_{v_m}, t_{or}, \mathcal{T}). \end{aligned} \tag{3.33}$$

Note that, since under the cBDP the ranked tree is independent of the time of the k-th speciation event [95],

$$\begin{aligned} f(r|t_{v_1}, \ldots, t_{v_m}, t_{or}, \mathcal{T}) &= \frac{f(r, t_{v_1}, \ldots, t_{v_m}|t_{or}, \mathcal{T})}{\sum_{\tilde{r} \in r(\mathcal{T})} f(\tilde{r}, t_{v_1}, \ldots, t_{v_m}|t_{or}, \mathcal{T})} \\ &= \frac{f(x_{r(v_1)}, \ldots, x_{r(v_m)}|t_{or}, \mathcal{T}) f(r|t_{or}, \mathcal{T})}{\sum_{\tilde{r} \in r(\mathcal{T})} f(x_{\tilde{r}(v_1)}, \ldots, x_{\tilde{r}(v_m)}|t_{or}, \mathcal{T}) f(\tilde{r}|t_{or}, \mathcal{T})} \\ &= \frac{f(x_{r(v_1)}, \ldots, x_{r(v_m)}|t_{or}, n)}{\sum_{\tilde{r} \in r(\mathcal{T})} f(x_{\tilde{r}(v_1)}, \ldots, x_{\tilde{r}(v_m)}|t_{or}, n)} \end{aligned} \tag{3.34}$$

where n is the number of species in $\mathcal{T}$. Note that $f(r|t_{or}, \mathcal{T})$ is the uniform distribution and therefore cancels out. We calculate $f(x_{r(v_1)}, \ldots, x_{r(v_m)}|t_{or}, n)$ as,

$$f(x_{r(v_1)}, \ldots x_{r(v_m)}|t_{or}, n) = \int f(x_1, \ldots x_{n-1}|t_{or}, n) dx_{r_1} \ldots dx_{r_{n-m-1}} \tag{3.35}$$

where $\{x_{r_1}, x_{r_2}, \ldots, x_{r_{n-m-1}}\} := \{x_i : i \in \{1, \ldots, n-1\} \setminus \{r(v_1), \ldots, r(v_m)\}\}$; we integrate over all possible values of $x_{r_1}, x_{r_2}, \ldots, x_{r_{n-m-1}}$. The density $f(x_1, x_2, \ldots x_{n-1}|t_{or}, n)$ is calculated via

$$f(x_1, x_2, \ldots x_{n-1}|t_{or}, n) = (n-1)! \prod_{i=1}^{n-1} f(x_i|t_{or})$$

with $f(x_i|t_{or})$ given in Theorem 3.4.4.

Next, we derive the density $f_{t_v}(s|r, t_{v_1}, \ldots, t_{v_m}, t_{or}, \mathcal{T})$. Since the cBDP before the k-th speciation event is independent from the process after the k-th speciation event (more precise, the process only depends on the time of the k-th speciation event), we have,

$$f_{t_v}(s|r, t_{v_1}, \ldots, t_{v_m}, t_{or}, \mathcal{T}) = f_{x_{r(v)}}(s|x_{r(v_1)}, \ldots, x_{r(v_m)}, t_{or}) = f_{x_{r(v)}}(s|x_{r(v_a)}, x_{r(v_{a+1})}, t_{or})$$

where $a \in \{0, 1, \ldots, m+1\}$ is determined such that $r(v_a) < r(t_v) < r(v_{a+1})$. Set $r(v_a) =: k, r(v_{a+1}) =: j, r(v) =: i$. Note that between the k-th and j-th bifurcation

event, we have $j-k-1$ bifurcation events which follow the point process introduced in Section 3.4.1; the point process is conditioned to have the $j-k-1$ points appearing in the interval (x_j, x_k). The $j-k-1$ speciation times between x_k and x_j conditioned on t_{or} are i.i.d. as established in Section 3.4.1, and we obtain for the density with $f(s|t_{or})$ from Theorem 3.4.4,

$$g(s|t_{or}, x_k, x_j) = \frac{f(s|t_{or})}{\int_{x_j}^{x_k} f(s|t_{or})} = \frac{f(s|t_{or})}{F(x_k|t_{or}) - F(x_j|t_{or})}$$

for $x_k > s > x_j$ and $g(s|t_{or}, x_k, x_j) = 0$ else. The distribution of g is

$$G(s|t_{or}, x_k, x_j) = \frac{F(s|t_{or}) - F(x_j|t_{or})}{F(x_k|t_{or}) - F(x_j|t_{or})}.$$

The density of x_i ($i \in \{k+1, \ldots, j-1\}$), conditioned on x_j, x_k and t_{or} is the $(j-i)$-th order statistic of $j-k-1$ i.i.d. random variables with density $g(s|t_{or}, x_j, x_k)$, this is

$$\begin{aligned} f_{x_i}(s|t_{or}, x_k, x_j) &= (j-i)\binom{j-k-1}{j-i} G(s|t_{or}, x_k, x_j)^{j-i-1} \times \\ &\quad (1 - G(s|t_{or}, x_k, x_j))^{i-k-1} g(s|t_{or}, t_k, t_j). \end{aligned} \tag{3.36}$$

With Equations (3.33), (3.34), (3.35), and (3.36) we established the density for the time of an undated speciation event v, t_v, in a partially dated phylogeny of age t_{or},

$$\begin{aligned} f_{t_v}(s|t_{v_1}, \ldots, t_{v_m}, t_{or}, \mathcal{T}) &= \sum_{r \in r(\mathcal{T})} (r(v_{a+1}) - r(v)) \binom{r(v_{a+1}) - r(v_a) - 1}{r(v_{a+1}) - r(v)} \times \\ &\quad G(s|t_{or}, x_{r(v_a)}, x_{r(v_{a+1})})^{r(v_{a+1}) - r(v) - 1} \times \\ &\quad (1 - G(s|t_{or}, x_{r(v_a)}, x_{r(v_{a+1})}))^{r(v) - r(v_a) - 1} g(s|t_{or}, t_{r(v_a)}, t_{r(v_{a+1})}) \times \\ &\quad \frac{\int f(x_1, x_2, \ldots x_{n-1}|t_{or}, n) dx_{r_1} \ldots dx_{r_{n-m-1}}}{\sum_{\tilde{r} \in r(\mathcal{T})} \int f(x_1, x_2, \ldots x_{n-1}|t_{or}, n) dx_{\tilde{r}_1} \ldots dx_{\tilde{r}_{n-m-1}}} \end{aligned} \tag{3.37}$$

where for each rank function r, we choose $a \in \{0, \ldots, m+1\}$ such that $r(v_a) < r(v) < r(v_{a+1})$. If t_{or} is not known, we have to integrate Equation (3.37) over t_{or} weighted by the prior distribution $q_{or}(t_{or}|n)$ (Equation 3.17) to obtain $f_{t_v}(s|t_{v_1}, \ldots, t_{v_m}, \mathcal{T})$.

The expectation of t_v can be obtained numerically with Equation (3.37). The expectation is an estimate for the speciation time of v. The estimate uses all the information given in the data. We do not use any estimated vertices to date another vertex, and therefore do not get a bias.

Chapter 4

Neutrality test on ranked trees

In this chapter we present a new statistical test to investigate the timing of branching events in phylogenetic trees. Our method explicitly considers the relative timing of diversification events between sister clades; as such it is complementary to existing methods using LTT plots which consider diversification in aggregate. The method looks for evidence of diversification happening in lineage-specific *bursts*, or the opposite, where diversification between two clades happens in an unusually regular fashion. The null models of our statistical test are the CAL and the UR models. For calculating the p-values, we use the analytic results for the CAL and UR class of models derived in the previous chapters. We apply the new statistic to several data sets: first, we show that the evolution of the Hepatitis C virus appears to proceed in a lineage-specific bursting fashion. Second, we analyze a large tree of ants, demonstrating that a period of elevated diversification rates does not appear to have occurred in a bursting manner. Last, we test the phylogeny of the genus *Dina* for lineage-specific bursting. We detect less balanced trees than under a CAL model, but there is no evidence for lineage-specific bursts.

4.1 Motivation

Understanding the tempo and mode of diversification is one of the major challenges of evolutionary biology. Phylogenetic trees with timing information are powerful tools for answering questions about tempo and mode. Such trees were once available only in situations with a rich fossil record, where the timing information might have come from radiocarbon dating or stratigraphic information (layering of rocks). However, modern techniques of phylogenetic analysis are capable of reconstructing not only the shape of phylogenetic trees, but can also reconstruct information about the timing of diversification events even when limited or no fossil evidence is available. This can be done in one of a number of ways. One can first test if a molecular clock is appropriate (see [21] p. 323), then reconstruct under the assumption of a molecular clock. One can reconstruct a tree with branch lengths using any method and then apply rate smoothing [84]. One may also choose from the variety of *relaxed clock* methods which allow the rate of substitution to vary within the tree [29, 40, 15].

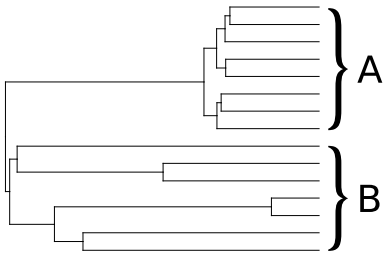

Figure 4.1: Example for lineage-specific bursting diversification.

Of course, the accuracy of any these techniques depends on a correct choice of model and a strong phylogenetic signal along with perhaps some fossil calibration points.

Phylogenetic trees with timing information can then be used to make inferences about the forces guiding the evolution of the taxa. For example, the paper of [66] notes that there was a period of high diversification rate in ant lineages during the rise of angiosperms (flowering plants). Another paper by [32] uses the deviation of four groups of lizards from the pure-birth model of diversification to make inferences about their evolutionary radiations.

Given the number of methods available for reconstructing phylogenetic trees with diversification timing information, and the interest in investigating temporal properties of those trees, the number of direct methods available to investigate timing information on phylogenetic trees is surprisingly small. The most popular ways of investigating timing in phylogenetic trees are LTT plots and the associated γ statistic. LTT plots have time t on the x axis and simply show the number of lineages which were present in the phylogenetic tree at time t on the y axis (Section 3.6). The γ statistic is computed based on the periods during which the LTT plot stays constant (called the *inter-node intervals*); the γ for a pure-birth diversification process will have a standard normal distribution. Broadly speaking $\gamma < 0$ implies that diversification rates were high early in history, while $\gamma > 0$ implies that most diversification has happened more recently [76].

However, much more information is available in a phylogenetic tree with diversification timing information than can be summarized in a LTT plot or a derivative statistic. Consider the tree in Figure 4.1, with two sets of sister taxa, A and B. The extant taxa in B had a period of relatively high diversification rate early in evolutionary history, during which time the lineage leading to A is in a period of stasis. Then lineage A experiences a burst of diversification, and the taxa in B do not experience any lineage-splitting events during this time. We will call the sort of diversification seen in Figure 4.1 *lineage-specific bursting* (LSB) diversification.

The lineage-specific bursting diversification seen in Figure 4.1 would not be apparent in an LTT plot. Indeed, LTT plots take the timing information out of the context of the phylogenetic tree from which they are derived, and thus ignore information about how the timings relate to the shape of the tree. This context can be crucial, as we now argue.

One would like to be able to say if, for example, the pattern seen in Figure 4.1 arose simply *by chance.* In order to do so, we need two things: first, a convenient way to summarize the timing information, and second, a set of null models which define what we mean with "by chance." For a given internal node, we summarize the timing information at that node by writing down the order of diversification events by clade. For instance, we associate with the root node of Figure 4.1 the sequence $s = BBBBBAAAAAAAB$ which we call a *shuffle*, as defined in Section 2.1.

Now that we have summarized the timing information at the root node as a shuffle s, we would like to think about if s arose "by chance", i.e. if no shuffle is favored over another shuffle. This is equivalent to each shuffle being equally likely. The UR models introduced in Section 2.3.3 induce the uniform distribution on shuffles. Recall that the UR models are precisely the set of pure-birth ET models which induce a uniform distribution on shuffles.

The uniform distribution in this setting is what one would get by throwing the A's and B's of the shuffle into a bag and drawing them out one by one uniformly. Thus it seems reasonably unlikely that the shuffle s would arise by chance, having first a long run of B's then a long run of A's.

Note that the pure-birth CAL models are a subset of the UR models. Under a CAL model, each species evolves in the same way. We will assume the UR as well as the CAL model as null models for our statistic. Departure of a CAL model can be interpreted as different clades evolving with different rates. Departure of the UR model means that speciation events are clustered in an unusual fashion. Given a tree shape, some shuffles are favored over other shuffles.

We can attach a p-value to a shuffle by using the *runs distribution.* The number of *runs* is simply the number of sequences of the same letter: in the example above, there is a run of B's, then a run of A's, then another B. That totals three runs. Under the uniform distribution, the probability of seeing a given number of runs in this setting is known from classical statistics, and can be calculated via Equation (4.1). The probability of seeing 3 runs with 6 A's and 7 B's is about 0.00641, and the probability of seeing 2 runs is about 0.00117. We can interpret the sum of these two probabilities, 0.00758, as the significance level of the LSB diversification seen in Figure 4.1. Being below the 1% significance level, we can interpret this shuffle as being quite significant; thus if the tree in Figure 4.1 came from data, the observed lineage-specific diversification might require some explanation. Please note that for simplicity this example only considers the root shuffle; however the main body of the chapter is dedicated to investigating all shuffles simultaneously.

Above we defined the clustering of the same letter in the shuffle as "lineage-specific bursting (LSB) diversification," but would like to note that clustering of the same letter might also have other reasons than different relative diversification rates. For example, assume in the macroevolutionary case that there are two lineages descending from the root; the left lineage has a moderate speciation rate and no extinction, whereas the right lineage has a very high speciation and extinction rate. Because of the high extinction rates, the right lineage will have most of its internal nodes at a time close to the present day. This phenomenon has been called "the pull

of the present" [69].

However, we do not think that such biases will in general pose a problem for the following reasons. Although the "pull of the present" is a phenomenon which generally comes with extinction, it will only pose a problem for our method when extinction rates differ significantly between pairs of sister clades. Also, because we are taking the sum of the numbers of runs across all internal nodes of the tree it is not enough to have just a single pair of sister clades with significantly differing extinction rates: there must be a general pattern of such extinction rate differences across the tree. Although this is certainly possible, we will call cases with number of runs significantly fewer than expectation "lineage-specific bursting speciation."

We further note that, since the uniform ranked oriented tree distribution is invariant under random (uniform) taxon sampling (Theorem 3.7.1), the shuffle p-value is not biased by random taxon sampling. However, like any method based on phylogenetic trees, the shuffle p-value is subject to biases introduced by non-uniform taxon sampling. It is not hard to devise a sampling scheme which would bias the results. For example, say we have two subpopulations descending from a single internal node by a process that induces the uniform distribution on shuffles. On side A sampling is done uniformly, whereas on side B similar lineages are unlikely to be part of the sampling. Such a scheme would bias the surviving internal nodes in B to be further back in the past, resulting in a non-uniform distribution on shuffles.

In Section 4.2, we provide analytic tools to compare diversification rates between lineages. In doing so, we hope to provide a complementary perspective to that provided by LTT plots and associated statistics. In particular, our method can detect LSB diversification. One might expect LSB diversification if a lineage diversifies to fill variants of a single niche, or if a key innovation appears which makes further diversifications more likely. By comparing the results of our analysis to results using LLT plots, we may be able to tease apart causes of diversification rate changes – are they lineage-specific or due to global events? The assumed null models are the UR models and the CAL models. Under the UR models, each ranking given the oriented tree is equally likely. Under the CAL model, each ranked oriented tree is equally likely.

In Section 4.3, we apply our statistic on various data. Our first example application uses Hepatitis C virus (HCV) data, and shows that trees from this data demonstrate a limited but significant amount of LSB diversification. This analysis may imply a note of caution for researchers using coalescent methods to analyze HCV data. Our second application is to the ant data of [66] and [67], the lineages of which do not appear to demonstrate significant LSB diversification, despite some other interesting characteristics of their history. For the phylogeny of the genus *Dina* in an ancient lake, we do not observe a significant departure from the UR models, i.e. we do not observe evidence for LSB diversification.

4.2 Tests for bursting diversification based on shuffles

In this section we describe a way of testing for deviation from the uniform distribution on tree shuffles, and thus test for deviation from the neutral models, the UR models. We emphasize that this can go significantly beyond testing the coalescent/Yule model, which is typically considered to be the definition of neutrality. Indeed, rejection of the uniform distribution on shuffles rejects all of the UR models simultaneously, and the coalescent/Yule model is only one model in this class. We note further that although the focus of this section is to consider all of the shuffles of a ranked tree at once, one can also consider a shuffle at a particular node as described in the last section.

There are several useful tools available to test whether a shuffle is likely to have come from the uniform distribution on shuffles. In fact, a number of tests in the statistics literature have been developed for testing equality of distributions which actually implement a test of deviation from the uniform distribution for shuffles. These tests work as follows: assume we are given two sets of samples $\{\ell_i\}_{i=1,\dots,m}$ and $\{r_j\}_{j=1,\dots,n}$ and would like to test the hypothesis that they are draws from the same distribution. To test, combine the draws and put the samples in increasing order (assume that all draws are distinct). This clearly gives a shuffle on symbols ℓ and r. If the draws are from identical distributions then the induced distribution on shuffles will be uniform; if on the other hand symbols cluster together in the shuffle, there is some evidence that the draws are from unequal distributions.

One can then test deviation from the uniform distribution on shuffles in one of several ways. One way is to count the number of *runs*. As described in Section 4.1, a run is simply a sequence within the shuffle using only one symbol; the shuffle $\ell\ell rrrr\ell$ has three runs. Let $X_{m,n}$ denote the number of runs under the uniform distribution on shuffles on m symbols of one type and n of another. The distribution of $X_{m,n}$ is classical (see, e.g. [38]):

$$\begin{aligned} \mathbb{P}\{X_{m,n} = 2k+1\} &= \frac{\binom{m-1}{k}\binom{n-1}{k-1} + \binom{m-1}{k-1}\binom{n-1}{k}}{\binom{m+n}{m}}, \\ \mathbb{P}\{X_{m,n} = 2k\} &= \frac{2\binom{m-1}{k-1}\binom{n-1}{k-1}}{\binom{m+n}{m}}. \end{aligned} \tag{4.1}$$

Asymptotic results for the mean and variance are also known:

$$\mathbb{E}[X_{m,n}] = \mu_{m,n} = 2\frac{mn}{m+n} + 1, \qquad \mathrm{Var}[X_{m,n}] = \frac{(\mu_{m,n}-1)(\mu_{m,n}-2)}{m+n-1}.$$

The usual application of the runs test makes a shuffle from the two draws as described above, calculates the number of runs in the shuffle, and then uses the above-calculated probabilities to test deviation from the uniform distribution on shuffles. In the present case, we can use an analogous process to investigate tree

shuffles as introduced in Section 2.1. A tree shuffle simply assigns a shuffle of appropriate size to each internal node of the tree; these shuffles are distributed uniformly under the UR models. Using runs we can test whether a single shuffle is drawn from the uniform distribution, but some method is needed to combine this information across the internal nodes of the tree.

For a ranked tree shape τ, we chose to combine our data from each vertex by simply summing the number of runs across all of the shuffles in the corresponding tree shuffle. Let $\mathcal{R}(\tau)$ denote this number. The distribution of $\mathcal{R}(\tau)$ (under the assumption that each shuffle is equally likely) can be calculated recursively as shown in the next several paragraphs. Note that under the UR model, each rank function conditioned on an oriented tree is equally likely. With Corollary 2.2.4, this yields each rank function on a given tree shape being equally likely under the UR model. Therefore we may consider tree shapes rather than oriented trees, which reduces the time complexity of the calculations.

There are two cases to consider. First, one may condition on the observed tree shape and calculate the neutral distribution of $\mathcal{R}(\tau)$ in that setting. A second option is to test deviation from a neutral model which gives the uniform distribution on ranked oriented trees. This is a stronger statement than saying that a given model induces the uniform distribution on shuffles conditioned on the tree shape.

We first condition on the observed tree. Uniform shuffles conditioned on the tree shape are obtained in the UR class of models. In the case of pure-birth models, the class of models which induce a uniform distribution on shuffles is exactly the set of UR models. For a tree with one leaf, we have $\mathbb{P}\{\mathcal{R}(\tau) = 0\} = 1$. For a tree with two leaves, we also have $\mathbb{P}\{\mathcal{R}(\tau) = 0\} = 1$ (the two daughter subtrees have no internal nodes).

For a tree shape τ with uniform random ranking, composed of two ranked subtrees L and R with m and n leaves, respectively, we have:

$$\mathbb{P}\{\mathcal{R}(\tau) = k\} = \sum_{i=0}^{k} \mathbb{P}\{X_{m-1,n-1} = i\} \sum_{j=0}^{k-i} \mathbb{P}\{\mathcal{R}(L) = j\}\mathbb{P}\{\mathcal{R}(R) = k-i-j\}. \quad (4.2)$$

It is shown in the next section, that this distribution can be calculated on a tree with n leaves in time $O(n^3 \log^2 n)$. Thus it is practical to obtain a p-value for $\mathcal{R}(\tau)$ analytically.

Now we take the second approach, assuming we want to test a model such that each ranked oriented tree is equally likely. This includes the CAL models, and in the case of pure-birth models, this is exactly the set of the CAL models (Proposition 2.3.3). Let $\mathcal{R}(n)$ be the random variable "number of runs of an oriented tree with n leaves" where the tree is drawn from the uniform distribution on ranked oriented trees. The distribution of $\mathcal{R}(n)$ can again be obtained recursively. Note that for a uniform ranked oriented tree on n leaves, the probability that the left daughter tree has size r and the right daughter tree has size $n-r$ is $1/(n-1)$ for all r (Lemma

3.7.2). Thus

$$\mathbb{P}\{\mathcal{R}(n)=k\} = \frac{1}{n-1}\sum_{r=1}^{n-1}\sum_{i=1}^{k}\mathbb{P}\{X_{r-1,n-r-1}=i\}\times \sum_{j=0}^{k-i}\mathbb{P}\{\mathcal{R}(r)=j\}\mathbb{P}\{\mathcal{R}(n-r)=k-i-j\}. \tag{4.3}$$

The complexity for recursively calculating the distribution of runs for trees with n leaves is $O(n^4 \log^2 n)$, by an argument analogous to that for Equation (4.2).

Note that there are a number of alternative ways to "sums of runs" for testing deviation from the uniform distribution on shuffles. First, we have made one choice – namely, summation – concerning how the statistics for each shuffle are combined. One certainly could use an alternative method, potentially including weights. Second, there are other statistics such as Mann-Whitney-Wilcoxon which could be used in place of the runs statistic. The advantage of summation is that it results in simple formulae. Further, through summation the runs statistic has a recursive formulation analogue to the Colless statistic. For a ranked tree τ with the daughter trees L, R, the number of runs in τ, $\mathcal{R}(\tau)$ is,

$$\mathcal{R}(\tau) = \mathcal{R}(L) + \mathcal{R}(R) + \mathcal{R}(\rho),$$

where $\mathcal{R}(\rho)$ is the number of runs in the root shuffle of τ. The advantage of the runs statistic is that it is easy to interpret. We have not tested any alternate formulations.

Computing quantiles of shuffles is available in the Cass package. One of the main features is the ability to calculate the quantile of the runs statistic assuming a uniform distribution on rankings for a set of input trees. The quantiles can be calculated conditioned on a given tree shape, or under the assumption of a uniform distribution on ranked trees. For a collection of trees (e.g. a sample from the Bayesian posterior), the individual quantiles can be averaged. In addition to the calculation of the runs statistic and the quantile for the whole tree, the package can calculate the runs statistic and quantile for each interior vertex of a tree. This feature may be useful for biologists looking for signals of a key innovation.

4.2.1 Complexity of computing the runs distribution

Here we provide a proof of the time-complexity bound for the computation of the runs distribution $\mathcal{R}(\tau)$ (i.e. conditioning on a given tree shape). This distribution may be computed easily for certain tree shapes, such as the comb tree. However, here we provide a bound which holds for all tree shapes. This bound makes use of a bound on the number of runs in a ranked tree shape.

Let $r(n)$ denote the maximum number of runs for a ranked tree shape with n leaves. Thus $r(1) = r(2) = 0$, $r(3) = 1$ and $r(4) = 2$. Let $I_{i=n/2}$ be 1 if $i = n/2$ and 0 otherwise. For a tree with at least 2 leaves, if the first branch point has i leaves on one side and $n - i$ leaves on the other, with $i \leq n - i$, then the number of runs at

this vertex may be up to $2(i-1)+1-I_{i=n/2}$ (note that we have an $(i-1,n-i-1)$ shuffle at this vertex). This maximum is obtained by a shuffle which interleaves the elements from each set, one from each side for as long as possible, starting with the largest side.

Thus, $r(n)$ satisfies the following recurrence: $r(1)=r(2)=0$ and for $n\geq 2$:

$$r(n)=\max_{1\leq i\leq n/2}\left(2i-1-I_{i=n/2}+r(i)+r(n-i)\right)$$

Proposition 4.2.1. *For all integers $n\geq 1$, $r(n)\leq n\log_2 n$.*

Proof. The statement is true for $n=1$. Suppose that the statement is true for all $k<n$. Then,

$$\begin{aligned} r(n) &= \max_{1\leq i\leq n/2}\left(2i-1-I_{i=n/2}+r(i)+r(n-i)\right) \\ &\leq \max_{1\leq i\leq n/2}\left(2i-1+i\log_2 i+(n-i)\log_2(n-i)\right). \end{aligned}$$

Note that $2i-1$, $i\log_2 i$ and $(n-i)\log_2(n-i)$ are all convex functions of i so their sum is convex also. Thus, the maximum of $2i-1+i\log_2 i+(n-i)\log_2(n-i)$ occurs at an extreme value. Setting $i=1$ gives $1+0+(n-1)\log_2(n-1)$, while setting $i=\frac{n}{2}$ gives $2\frac{n}{2}-1+2\frac{n}{2}\log_2\frac{n}{2}=n(\log_2 2+\log_2\frac{n}{2})-1=n\log_2 n-1$. Both of these values are less than $n\log_2 n$ and so $r(n)$ must be at most $n\log_2 n$. The result follows for all $n\geq 1$ by induction. □

We now proceed to bound the complexity of computing the distribution of runs for a tree. For a tree shape τ with 1 or 2 leaves, the number of runs is always 0.

Let τ be a tree shape with $n\geq 3$ leaves; we assume a uniform distribution on tree shuffles. Let L and R be the two randomly ranked subtees of τ, with a and b leaves respectively.

Equation (4.2) may be rewritten as follows:

$$\begin{aligned} \mathbb{P}\{\mathcal{R}(\tau)=k\} &= \sum_{i=0}^{A_1}\mathbb{P}\{X_{a-1,b-1}=i\}\sum_{j=0}^{A_2}\mathbb{P}\{\mathcal{R}(L)=j\}\mathbb{P}\{\mathcal{R}(R)=k-i-j\} \\ &= \sum_{i=1}^{A_1}\mathbb{P}\{X_{a-1,b-1}=i\}\mathbb{P}\{\mathcal{R}(L)+\mathcal{R}(R)=k-i\} \qquad (4.4) \end{aligned}$$

where $A_1=\min(k,n)$ and $A_2=\min(k-i,r(a))$. Note that $a+b=n\geq 3$ implies $X_{a,b}\geq 1$ and $\mathcal{R}(\tau)\geq 1$.

Since $\mathcal{R}(\tau)$ is supported on (i.e. zero outside of) $k=1,\ldots,\lfloor n\log_2 n\rfloor$, the cost of computing its distribution with this formula is $(\lfloor n\log_2 n\rfloor)(2n-1)$ arithmetic operations plus the cost of computing $\mathbb{P}\{X_{a-1,b-1}=i\}$ for $i=1,\ldots,n$ and $\mathbb{P}\{\mathcal{R}(L)+\mathcal{R}(R)=x\}$ for $x=0,\ldots,r(n)-1\leq n\log_2 n-1$.

For these fixed a and b, the values of $\mathbb{P}\{X_{a,b}=i\}$ can be calculated using Equation (4.1) in constant time (at most $5*2+4=14$ arithmetic operations each) with a linear overhead as follows. The binomial coefficients $\binom{a}{k}$ for $a\leq b$ and $k\leq b$ in

Equation (4.1) may be calculated with at most two arithmetic operations from the factorials, $j!$ for $1 \leq j \leq n$, which may in turn be pre-calculated in linear time ($n-1$ multiplications). Thus, calculating $\mathbb{P}\{X_{a,b} = i\}$ for $i = 1, \ldots, n$ takes at most $14n$ arithmetic operations, with a one-time overhead of $n - 1$.

The distribution of $\mathbb{P}\{\mathcal{R}(L)+\mathcal{R}(R) = x\}$ is supported on $x = 0, \ldots, \lfloor n \log_2 n \rfloor - 1$. It may be computed by repeated application of the formula

$$\mathbb{P}\{\mathcal{R}(L) + \mathcal{R}(R) = x\} = \sum_{j=0}^{\lfloor (n-1)\log_2(n-1) \rfloor} \mathbb{P}\{\mathcal{R}(L) = j\}\mathbb{P}\{\mathcal{R}(R) = x - j\}$$

as long as the distributions of $\mathcal{R}(L)$ and $\mathcal{R}(R)$ are known. This computation requires at most $n \log_2 n\,(2(n-1)\log_2(n-1) + 1)$ arithmetic operations: at most $(n-1)\log_2(n-1) + 1$ multiplications and $(n-1)\log_2(n-1)$ additions for each of $n \log_2 n$ values of x. Note that the distribution of $\mathbb{P}\{\mathcal{R}(L)\}$ is supported by $j = 0, \ldots, \lfloor (n-1)\log_2(n-1) \rfloor$, since L has at most $n - 1$ leaves.

So, if the distribution of $\mathcal{R}(L)$ and $\mathcal{R}(R)$ are known, the distribution of $\mathcal{R}(\tau)$ may be calculated in at most

$$(\lfloor n \log_2 n \rfloor)(2n - 1) + 14n + n \log_2 n\,(2(n-1)\log_2(n-1) + 1)$$

arithmetic operations. This is at most

$$2n^2 \log_2 n + 2n^2 \log_2^2 n + 14n$$

for all $n \geq 3$. Since $\mathcal{R}(\tau)$ is 0 for $n = 1, 2$ the time to calculate it is 0.

This procedure may be applied recursively, computing the distribution of runs of all subtrees before finally computing the run distribution of τ. Since there are $n - 1$ internal vertices and each has at most n leaves below it, the total number of arithmetic operations required is at most $n(2n^2 \log_2 n + 2n^2 \log_2^2 n + 14n + 1)$ (including the overhead for pre-computing $j!$). This is $O(n^3 \log_2^2 n)$.

4.2.2 Shuffles in the Bayesian setting

In our work up to now, we have assumed that the correct tree and diversification timing information is known. This assumption is not realistic for a number of datasets. For example, below we apply our methodology to a sample of Hepatitis C viruses, which probably do not have enough sequence divergence to perfectly reconstruct a tree with timing information.

One way of working with such datasets is to take a Bayesian approach, where rather than a single tree one gets a posterior distribution on trees. For each single tree, one can compute the p-value of the total runs statistic, either conditioning on the topology or assuming a uniform distribution on ranked oriented trees. We then simply take the average of the p-values thus computed for each tree. The average of p-values in this case is a simple type of posterior predictive p-value [63, 79]. As such, it is not exactly uniformly distributed under the neutral model as a proper p-value should be, although the average does share many of the characteristics of a classical p-value.

4.2.3 Neutrality in population genetics

The coalescent is the standard model in population genetics. When introducing the coalescent in Section 3.5 we implicitly assumed a constant population size through time. However, the population size might vary in a lot of cases. The runs statistic can be used to test the coalescent in the presence of ancestral population size variation since this process is still a CAL model – each individual evolves in the same way. Tests of neutrality in the presence of historical population size variation are of particular recent importance, as new coalescent-based methods are in use to infer population size history in a Bayesian framework [17, 73]. If these methods are to be used on a given set of sequences it is important to test the central assumption of the methods, namely that the sequences have a genealogy which can be accurately described using the coalescent with arbitrary population size history.

Unfortunately, classical statistics such as the D statistics of [94] and [24] confound ancestral population size changes and non-neutral evolution. One solution to this problem is to investigate the Bayesian posterior on phylogenetic trees for evidence of non-neutral evolution rather than using the sequence information directly. Since the coalescent with arbitrary population size history is a CAL model and thus will induce the uniform distribution on ranked oriented trees, we can apply the runs statistic; thus by rejecting the CAL class we reject a general coalescent model. We will apply this fact below in the example application to Hepatitis C data.

4.2.4 Generalization for non-binary trees

Polytomies (i.e. non-binary splits) are common in reconstructed phylogenetic trees. Some polytomies are certainly due to a lack of information to resolve the splits, however it has been argued that molecular and species level polytomies actually exist [41, 88]. The methodology described in this chapter can be extended to trees with *hard* polytomies, i.e. cases of multiple divergence which are essentially simultaneous in evolutionary time.

The new ingredient needed is the *multiple runs distribution,* i.e. the analog of (4.1) for shuffles on more than two symbols. This is described in [12]. Using these distributions, the probability of a shuffle consisting of symbols from the k daughter trees can be found for a shuffle at a non-bifurcating split v.

4.3 Example applications

In this section we describe three distinct applications of the runs statistic. First, we apply the methods to E1 gene data for the Hepatitis C virus (HCV). This data set shows some limited – though consistent – lineage-specific bursting diversification, showing that neither CAL nor the more general UR models accurately describe the sort of evolution observed. An analysis not conditioning on tree shape clearly rejects any CAL model, such as the coalescent with varying population size. The second application is to phylogenetic trees for ants, whose timing information was recon-

structed through fossils and the r8s [84] rates smoothing program. These ant trees do not show any evidence of lineage-specific bursting evolution, despite some interesting history in terms of diversification rates. Last, we investigate the phylogeny of the genus *Dina*. There is no evidence of lineage-specific bursting. However, the tree shape is less balanced than under a CAL model.

4.3.1 Hepatitis C Virus

Our HCV data comes from two independent studies: one in China [58], and one in Egypt [78]. The HCV alignments were retrieved from the LANL HCV database [54] via PubMed article ID numbers. The Chinese dataset contained samples from 132 infected individuals, and the Egyptian dataset had samples from 71 individuals. We randomly partitioned the taxa from the Chinese dataset into three sets of 44 taxa each and used the corresponding sub-alignments as distinct data sets. The Egyptian data was similarly split into two sub-alignments of size 37 and 36. This partitioning was done in order to have a larger number of similar datasets from which we could investigate the dynamics of HCV evolution, and to demonstrate that non-neutral evolution can be seen even with a moderate number of taxa.

In order to avoid confounding temporal information with molecular rate variation, we applied the relaxed clock model of [15] as implemented in the BEASTv1.4 suite of computer programs [16]. We chose uncorrelated lognormally distributed local clocks, the HKY model, and four categories of gamma rate parameters in the gamma + invariant sites model of sequence evolution. We used both the constant population size and exponential growth coalescent priors. All other parameters were left as default; the corresponding BEAST XML input files are available from the authors upon request.

In each case the MCMC chain was run for 10 million generations, and convergence to stationarity was checked with the BEAST program Tracer. For each model parameter, the minimum effective sample size was at least 164, with most being significantly greater. The coefficient of variation of the relaxed clocks in the analysis had a minimum of 0.336 and an average of 0.491, indicating a significant deviation from a strict clock for this data set. The first 10% of the run was removed and 100 trees were taken from the tree log file, equally spaced along the run of the MCMC chain. We interpret these trees as being independent samples from the posterior. As a check, the analysis was run with an empty alignment and no consistent deviation from the uniform distribution on shuffles was detected (results not shown).

We have displayed the results in Table 4.1. In the columns labeled "UR" we show the quantile of the number of runs conditioning on tree shape, calculated as in Equation (4.2). As can be seen, the results are substantially below one half, with the maximum being 0.308. Although this is not exceptionally strong lineage-specific bursting behavior, it does so consistently across five samples from two independent studies. Thus we feel confident in saying that the evolution of HCV displays lineage-specific bursting behavior. It might also be noted that these results were gained despite the fact that the coalescent was used as a prior. That is, if any bias could

Data set	Const. UR	Exp. UR	Const. CAL	Exp. CAL
China set 1	0.232	0.254	0.0415	0.0601
China set 2	0.191	0.17	0.041	0.0349
China set 3	0.239	0.259	0.0287	0.0265
Egypt set 1	0.261	0.299	0.045	0.0624
Egypt set 2	0.308	0.256	0.0242	0.0188

Table 4.1: Expected quantiles of the number of runs in the posterior for a Bayesian analysis as described in the text. Each row represents one dataset. "Const." means the BEAST analysis with a constant population coalescent prior, and "Exp." denotes analysis with an exponentially increasing population size coalescent prior. The "UR" label means that we analyze conditional on the tree shape, which gives us thus the quantile for any neutral model inducing the uniform distribution on shuffles. The UR models are precisely the class of pure-birth models with this property. "CAL" denotes the runs quantile under the assumption of the uniform distribution on ranked oriented trees, as would be the case for any CAL model, such as the coalescent with arbitrary population size history. As described in the text, the "UR" columns show that some limited lineage-specific bursting is seen, and the "CAL" column rejects the coalescent with arbitrary population size history.

be expected in the Bayesian analysis, it would be towards a coalescent prior and a uniform distribution on shuffles, thus we believe our results form an upper bound for the actual statistics of the HCV lineages.

We have displayed a graphical representation of the results for the second Chinese data set in Figure 4.2. Each horizontal bar represents one of the 100 ranked trees from the posterior. One side of the bar gives the number of runs in the ranked tree T, and the other side gives the expected number of runs for a neutral (i.e. CAL or UR) tree of the same unranked tree shape as T. If T has more runs than the expectation, the bar is colored gray; if fewer it is colored black. In both the cases of constant population size and exponentially increasing population size coalescent prior for BEAST, it can be seen that there are fewer runs than the expectation, meaning that it appears that the HCV data under investigation may have had periodic bursts of diversification in its past.

Now we apply our techniques as a statistical test for the coalescent with ancestral population size variation as described above. This is topical: we note that the [78] HCV data was analyzed by [73] as an example application of a reversible-jump Bayesian MCMC algorithm for estimating demographic history of the virus. In doing so they made an implicit assumption of neutrality because their method [and other such methods [17]] are based on the coalescent. They did not test this neutrality assumption as no methods were available at the time to test for neutral evolution in the presence of ancestral population size changes.

Our method can do so. Specifically, we compare the number of runs to the distribution for an arbitrary CAL model, as in Equation (4.3). By the results in the right half of Table 4.1, one can see that the data does not follow a coalescent

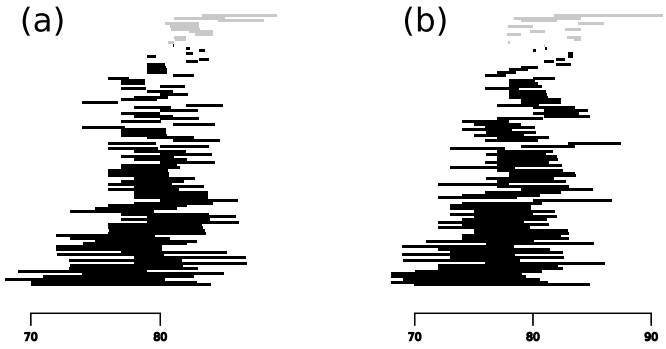

Figure 4.2: A visualization of the number of runs in a posterior sample of trees for an alignment of Hepatitis C sequence data of [78]. Black bars represent fewer runs than neutral, and gray the opposite. As described in the text, the width of the bars represents the amount of divergence from a broad class of neutral models. Said simply, each black bar represents a tree in the posterior which displays evidence of lineage-specific bursting diversification, and the longer the bar, the more bursting the tree. The bars are sorted vertically by increasing size (with sign.) Figure (a) shows the results when the tree prior in BEAST was taken to be coalescent with constant population size. Figure (b) shows the corresponding results with the exponentially increasing population size prior.

model with arbitrary population size history. This implies a significant model misspecification in the [73] paper; it would be interesting to know how this would impact the historical population size estimates in their paper.

4.3.2 Ants

For the second application we investigated two different trees of ant taxa. The first tree is that of [66], showing the diversification of the major ant lineages. The timing information in this tree is quite remarkable, in that the corresponding LTT plot shows a substantial increase in diversification rate during the Late Cretaceous to Early Eocene, which corresponds to the rise of angiosperms (flowering plants). Given the tools at our disposal, one might wonder if this increase in diversification rates affected all lineages equally, or if it occurred in lineage-specific bursts. The second ant tree we investigated was that of *Pheidole*, a "hyperdiverse" ant genus. Pheidole is almost certainly monophyletic, and yet comprises about 9.5% of the ant species in the world, according to latest estimates [67]. Moreau has recently reconstructed a phylogeny of this genus which we have analyzed along with the tree of the ant lineages in general. Both trees were reconstructed via maximum likelihood, then made ultrametric using the penalized likelihood method of the r8s rates smoothing program [84].

In Figure 4.3 we show a plot of the internal nodes of each tree. The x coordinate in the plot is the number of taxa below an internal node, and the y axis is the

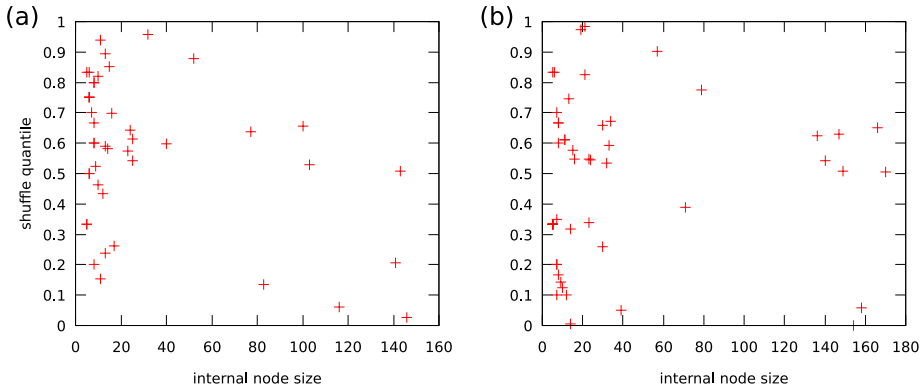

Figure 4.3: The distribution of the runs statistic for the internal nodes in two trees of ant taxa. Each point in each plot represents an internal node in the corresponding tree; the x axis gives the number of taxa below the internal node and the y axis gives the quantile of that internal node in terms of the runs statistic. Figure (a) is the tree of [66], and Figure (b) is a tree of Pheidole. These two trees do not appear to consistently show either lineage-specific bursting or refractory diversification.

quantile of the number of runs in the shuffle statistic. As can be seen, there is no clear correlation between number of taxa below an internal node and the shuffle statistic, and at no stage does diversification appear to be consistently bursting or refractory in a lineage-specific sense. We can also compute the quantile of the total number of runs across the tree: for tree (a) this is about 0.9052 and for tree (b) this is about 0.6718. Thus for these two ant trees we do not see any significant evidence of lineage-specific bursting or refractory diversification. This analysis forms an interesting counterpoint to the LTT results for the ants, which shows an overall increase in diversification in rate during the Late Cretaceous to Early Eocene across the entire tree.

4.3.3 The genus *Dina*

In order to reduce sampling artifacts in speciation studies, some workers utilize relatively small isolated systems with a high number of endemic species such as oceanic islands, dessert spring systems or ancient lakes systems (ancient lakes are lakes with an age of more than $100,000$ years). Particularly ancient lakes constitute a prime system for studying evolutionary processes as they often harbor species rich, monophyletic groups of endemic species (so called ancient lake species flocks) that are presumably little affected by surrounding biota.

We will infer and analyze the phylogeny of the genus *Dina* Blanchard 1892. The species of the genus *Dina* were collected in the ancient Lake Ohrid. The genus *Dina* belongs to the family of leech. A Bayesian posterior sample for 14 species is generated in MrBayes under the assumption of the molecular clock and a K2P model of sequence evolution. We then take the 1000 best trees of the Bayesian sample for

further analysis. For details of the collected sample and the inference method refer to our paper [96].

In order to test for variation in speciation rates over time, we used two independent approaches. First we tested for lineage specific bursting, i.e., whether subsets of taxa were likely to speciate faster than others. Then we investigated whether, over the whole tree, there are changes in speciation rates over time by testing for deviation from a BDP.

Lineage specific bursting

The average quantile for the run statistic (over the 1000 best trees in the posterior) is 0.6204, i.e. we clearly do not reject the neutral model in favour of lineage specific bursting.

Global speciation rate changes

We did not detect lineage-specific bursting. However, there could be a global change in the speciation rate which we test in the following via LTT plots and the γ statistic. As the individual trees have different node depths, we scaled all trees such that the most recent common ancestor of the *Dina* species flock always resides at time 0 and the leaves of the tree at time 1. Then we averaged over the number of lineages at each point in time between zero and one, see Figure 4.4.

The results were compared to a BDP. Maximum likelihood estimates of μ and λ for the Bayesian trees showed that the death rate is almost zero, so we can focus on a pure-birth BDP, i.e. a Yule model. We calculated the maximum likelihood of parameter λ for each tree, the average over all estimated λ is $\lambda_{ML} = 87.139$ (standard deviation: 5.126).

First we obtained a visualization of the speciation times via LTT plots under the Yule model. We took two approaches. We calculated the expected LTT plot conditioned on observing 14 species today. We used a uniform prior for the time of origin. This plot is obtained with Theorem 3.8.8, see Figure 4.4.

Further, we calculated the expected LTT plot under a Yule model with parameter λ_{ML} evolving for a time 1 after the *mrca* (not conditioning on observing 14 species). For obtaining the LTT plot, we applied the following theory for Yule processes. The probability that a single lineage has i lineages at time t later has a geometric distribution (see e.g. [71]),

$$p_t(i) = e^{-\lambda t}(1 - e^{-\lambda t})^{i-1}$$

After divergence of the *mrca*, we have two independently evolving lineages, each having a geometric distribution for the number of descendants at time t. The convolution of two independent geometric distributions with parameter p is a negative binomial distribution with parameters $p, 2$. Therefore, the probability of obtaining i lineages at time t after the *mrca* is,

$$q_t(i) = (i-1)e^{-2\lambda t}(1 - e^{-\lambda t})^{i-2}.$$

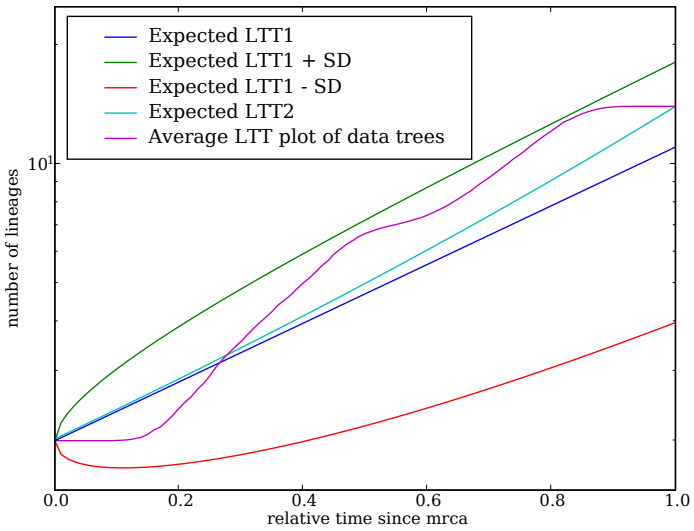

Figure 4.4: Average LTT plot for the posterior trees of the *Dina* samples. Additionally, we calculate the expected LTT plots of the Yule model. Expected LTT1 is the expected LTT plot of the Yule model assuming a speciation rate λ_{ML}. Expected LTT2 is the expected LTT plot of the Yule model conditioning on n species today (see Theorem 3.8.8).

The expectation is $2e^{\lambda t}$ and its variance is $2(1-e^{-\lambda t})e^{2\lambda t}$.

This allows us to plot the expected LTT plot with its standard deviation for a Yule process evolving for a time 1 after the *mrca*, see Figure 4.4.

To test if the LTT plot of the data is significantly different from a purely stochastic process, in our case the Yule model, we calculate the γ statistic [76] for all trees in the Bayesian posterior sample and averaged over all γ values. Under the Yule model, we expect $\gamma = 0$. A γ-value smaller than zero means that speciation occurred closer to the root than under the Yule model. A γ-value bigger than zero means that speciation occurred closer to the leaves (which is the case for a BDP with positive extinction rate).

The average γ value for the thousand trees is $\gamma = -1,642$. This shows again that we can neglect a positive death rate – for a positive death rate, we should obtain positive γ-values. Under the Yule model, we expect $\gamma = 0$ and we can reject the hypothesis of a Yule process with a 5% error, if $\gamma < -1.645$. So for our sample, the null hypothesis, a constant speciation model, can not be rejected.

4.4 Discussion of the new statistic

The method in this chapter was conceived for the macroevolutionary case, in order to find historical evolutionary patterns requiring explanation. However, it is also

quite applicable in the microevolutionary case, where it can test neutrality in the presence of historical population size variation. This is particularly relevant as methods are becoming available to describe historical population size under a coalescent assumption.

We emphasize that our methodology can go substantially beyond testing for deviation from the BDP, which are usually the entire class of *neutral* models considered. Indeed, because *any* UR model induces the uniform distribution on shuffles, deviation from this distribution is evidence to reject any model in the UR class. Such a conclusion is much stronger than deviation from a BDP, which is only one of the CAL models.

However, sometimes one may wish to test only a more restricted set of models, such as only the CAL models (which include the coalescent with arbitrary population size history) and not the more general UR models. By testing a more restricted class of models, a particular dataset will be more likely to fall outside the chosen class. For example, in the application of our methods to the Hepatitis C data above, the data consistently shows evidence of not coming from an UR model, although the corresponding quantile is in the 0.17 to 0.31 range. However, if one tests for conformity to the CAL class (again, including the coalescent with arbitrary population size history) one obtains rejection at the 5% level. Note that rejecting the class of CAL (resp. UR) models for a dataset where extinction might be present, we even reject a broader class of models – all models which induce a uniform distribution on ranked oriented trees (resp. rankings given the tree shape). However, classifying these models is tricky, see Section 2.3.

We recall that our method uses "relative" timing information rather than actual branch lengths. In many ways this is an advantage. In a microevolutionary setting this means that the corresponding tests are invariant to changes in ancestral population size, and thus our test for neutrality is not "fooled" by ancestral population size variation. In a macroevolutionary setting the statistics are robust to branch length estimation error over long time scales. Such estimations are known to be difficult [50]. We note further that from a modeling perspective it is possible to specify a probability distribution on ranked phylogenetic trees without specifying a particular distribution on branch lengths. This flexibility means that it may be possible to reject many models at once as described above.

Nevertheless it may be useful at some future stage to combine tree shape and continuous branch length information, rather than the discretized version considered here. However, quantifying the shape of such objects appears to be challenging, as the relevant geometry is quite intricate [6, 68]. In contrast, by discretization to ranked trees we obtain a purely combinatorial object.

Chapter 5

Samples of trees via simulations

A wide range of evolutionary models for speciation have been developed, some of the neutral models were discussed in the previous chapters. These models can be used to test evolutionary hypotheses (as done in Chapter 4) and provide comparisons with phylogenetic trees constructed from real data (e.g. via LTT plots, see Section 3.6). To carry out these tests and comparisons it is often necessary to sample – or simulate – trees from the evolutionary models, since no analytic results are available for most models. Even for simple models, simulations might be necessary. For example only little is known about the distribution of the γ statistic [76] under a cBDP.

For most models sampling trees appears to be a relatively easy exercise. If the aim is to produce trees of a given age this is indeed true. However in many circumstances it is preferable to sample trees with a given number of species. There are numerous ways to produce trees with a given number of species from an evolutionary model, however many seemingly intuitive approaches sample trees from unexpected and unrealistic distributions. This introduces some potential pitfalls, a problem that is exacerbated by the fact that there is no easy method for testing whether the sampling approach is correct.

Some simple approaches for sampling trees with a given number of species are in common usage. We show that these approaches are appropriate for the widely used Yule and coalescent models but there are some fundamental problems applying these approaches to other evolutionary models. We provide alternative sampling approaches that are theoretically sound and easy to apply.

We investigate the importance of using our correct sampling approach over established methods. This is achieved by comparing samples of trees produced by the different sampling approaches for given models. Existing sampling approaches introduce a strong bias in the age of a tree and a less pronounced bias in the relative timing of the speciation events. For the considered models, existing approaches introduce a negligible bias in the tree shape distribution and in an incomplete taxon sampling scenario. We identify attributes of other models that will result in existing sampling approaches producing more biased samples.

The methods we present are not the fastest or most sophisticated, however in our opinion they are the easiest to implement and applicable to the broadest pos-

sible range of models. Klaas Hartmann implemented the algorithms in the Perl Bio::Phylo package, where they can easily be applied to any suitable evolutionary model. For those users unfamiliar with Perl we have also made them available using a stand-alone GUI TreeSample. These tools are freely available [33]. Lastly we note that although we present our work in the context of evolutionary models of species diversification, our methods can be applied to other scenarios where birth-death processes are modeled, for example gene trees [72, 44, 30].

5.1 Sampling methods

Our aim is to produce a correct sample from the tree probability distribution induced by an evolutionary model. The first problem is that this tree probability distribution is ill defined for most evolutionary models. Under most models trees evolve perpetually – trees of all ages are possible and the expected age of the tree (the time between the root and the leaves) is infinite. To obtain a probability distribution it is therefore necessary to condition on some aspect of the tree; the number of species or the age of the tree are arguably the two most common and useful choices.

Conditioning on the age of a tree is appropriate if we wish to compare a model with trees of known age or if we want to test methods on simulated trees of a given age. It is relatively easy to sample trees of a given age from an evolutionary model. The tree is simply evolved according to the model until it has reached the desired age. This process is repeated until a sufficient number of trees have been sampled.

Conditioning on the number of species, n, in a tree may be of more interest for real applications. The age of a reconstructed tree may only be known with limited accuracy, however the number of species in the (reconstructed) tree is fixed. Consequently it may be more appropriate to use samples from an evolutionary model with a fixed number of species (we also consider incomplete taxon sampling). Sampling from the tree distribution conditional on the number of species, $p(\mathcal{T}|n)$, is the basis of this chapter.

Throughout this chapter we assume a uniform prior on the age of the tree as introduced in Section 1.1. Consider a large number of simulation runs that begin at a uniformly distributed time before the present. Trees obtained by selecting only those simulations that have n species at the present are a sample from $p(\mathcal{T}|n)$. This is a convenient way of interpreting the distribution but is not a practical sampling approach as the simulation starting time is taken from an ill defined distribution (between an infinite time in the past and the present). A given model (and its parameters) will induce a distribution on the age of the tree given its size as explained in Section 1.1, Equation (1.1). All our knowledge about the age of a tree is encapsulated in the model and the chosen parameter values; the uniform prior on the tree age represents the fact that we have no further knowledge about the tree age outside of these parameters.

5.1.1 Current approaches

One simple sampling approach (which we refer to as *SSA*) for sampling trees with n species has seen wide usage. With this approach a tree is evolved under the model until it has $n+1$ species and the last speciation event is disregarded. This approach produces trees conditional on the next speciation event occurring immediately after the end of the tree, which as we show here is generally not the same distribution as $p(\mathcal{T}|n)$. It is difficult to justify this approach as it produces a sample of trees equivalent to what we would expect if all 'real' trees were observed immediately prior to a speciation event.

PhyloGen [77] is a freely available tree sampler that has been used in a number of studies (for example [39, 86, 97, 99]). It permits users to sample trees from the BDP and episodic speciation models. These trees are conditioned on the age of the tree or the number of species, n. Conditioning on n in *PhyloGen* simply terminates a tree after it first reaches n species. Trees sampled with *PhyloGen* are younger than expected for our definition of $p(\mathcal{T}|n)$ and the pendant edges are shorter than expected – in fact the species produced by the last speciation event have zero length edges. If the last speciation event is removed (creating a tree with $n-1$ species) sampling trees with *PhyloGen* is equivalent to *SSA* with $n-1$ species. Due to this similarity throughout the remainder of this chapter we only consider *SSA*. There are three main possible problems with *SSA* and *PhyloGen*:

Problem 1. As has already been noted the pendant edge lengths produced by *SSA* and *PhyloGen* have what appears to be extreme values. With *PhyloGen* the pendant edges are as short as possible and with *SSA* they seem too long (this will be discussed in more detail later).

Problem 2. SSA and *PhyloGen* stop evolving the tree during (or just after) the first period of time where the tree has n species. For models with extinction the number of species will fluctuate up and down so there may be many periods during which the tree has n leaves. For such models *SSA* and *PhyloGen* will result in younger trees than expected.

Problem 3. A final concern with *SSA* and *PhyloGen* is that each model simulation run makes the same contribution to the final sample – one single tree. However, from our definition of $p(\mathcal{T}|n)$ the probability of observing a given simulation depends on the duration for which the simulated tree had n species – for example, if this duration is short it is unlikely that the simulated tree will be observed whilst it has n species.

5.1.2 Pure-birth memoryless models

We begin by considering pure-birth memoryless models – models that do not explicitly include extinction (pure-birth) and where future evolution depends only on the number of extant species (memoryless). This class of models is of particular interest as an approach similar to *SSA* can be used to correctly sample phylogenetic trees from them. Furthermore this class of models includes the most widely used speciation model – the Yule model (Section 3.2) – and the most widely used null model in population genetics – the coalescent (Section 3.5).

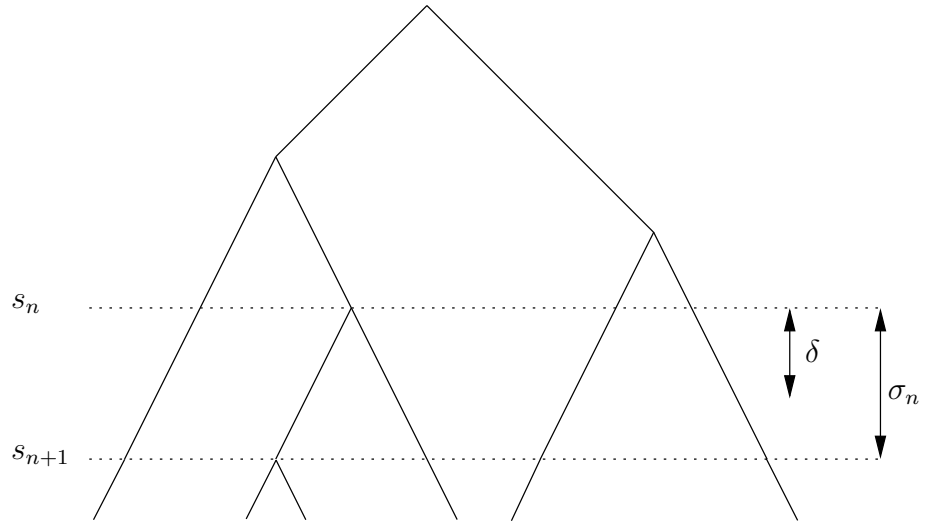

Figure 5.1: Some of the notation used throughout this chapter is illustrated in this figure where $n = 5$. $\mathcal{T}$ is the simulated tree until the point in time when the tree first has greater than n species. This point in time is the speciation event creating the $(n+1)$-th species, s_{n+1}. The duration for which a simulated tree has n species is denoted by σ_n, this is the time between the creation of the n-th species (s_n) and the $(n+1)$-th species (s_{n+1}). The time for which an observed tree has n species is necessarily less than σ_n and is denoted by δ.

Recall that under the Yule model each species has the same probability of speciating per unit time and this speciation rate is constant over time. Consequently the time between speciation events is exponentially distributed with parameter $m\lambda$, where m is the number of species that are extant and λ is the rate of speciation. The coalescent is derived from population genetics principles but is essentially the same as the Yule model with one exception – the time between coalescent events is exponentially distributed with parameter $\binom{m}{2}\lambda$ where λ encodes the population size. In the following we will use 'speciation' for both speciation and coalescent events.

In this section we show that although *SSA* is generally inappropriate for pure-birth memoryless models it is actually a correct approach for the Yule model and the coalescent. As these models are pure-birth models there will only be one period during which n species exist, so *Problem 2* does not apply. This leaves *Problems 1* and *3* which we will show cancel each other out under the Yule model and the coalescent. We speculate that the suitability of *SSA* for sampling from the most widely used null models has led to its application to other models for which it is unsuitable.

An important aspect of memoryless models is that the evolution after the speciation event that created the n-th species (s_n) is completely independent of the evolution that occurred up to that point. Consequently it is possible to simulate trees from these models in two separate stages. Firstly, using the model, a tree is simulated to the speciation event that created the n-th species (denoted by s_n; see

Figure 5.1). A length δ is then added onto the pendant edges to produce the final tree. Due to the independence of these two processes, *Problems 1* and *3* do not effect the simulation to s_n and are addressed entirely by an appropriate choice of δ. This raises the question from what probability density, $h(\delta)$, the additional time δ should be sampled.

We begin by noting that any pure-birth memoryless model can be uniquely defined by the probability densities of the intervals between speciation events. We denote the time between the speciation event that created the n-th and the $(n+1)$-th species by σ_n (the time between s_n and s_{n+1}) and its probability density by $g_n(\sigma_n)$.

Note that *SSA* makes the assumption that

$$h(\delta) = g_n(\delta).$$

This effectively produces a tree with n species conditional on the next speciation event occurring immediately – clearly not what was intended.

A seemingly better (but still generally incorrect) approach would be to simulate the tree until s_{n+1} and randomly terminate the tree between s_n and s_{n+1} (since all trees between these two events should be equally likely). This addresses *Problem 1* and gives us:

$$\begin{aligned} h(\delta) &= \int_{\delta}^{\infty} h(\delta|\sigma_n) g_n(\sigma_n) d\sigma_n \\ &= \int_{\delta}^{\infty} \frac{g_n(\sigma_n)}{\sigma_n} d\sigma_n \end{aligned}$$

However this does not take into account the variable contribution to the $p(\mathcal{T}|n)$ that different values of σ_n should make (*Problem 3.*).

From the definition of $p(\mathcal{T}|n)$ the contribution from a simulated tree with a given σ_n should be proportional to σ_n, therefore the correct distribution from which to sample δ is:

$$\begin{aligned} h(\delta) &\propto \int_{\delta}^{\infty} \sigma_n h(\delta|\sigma_n) g_n(\sigma_n) d\sigma_n \\ &= \int_{\delta}^{\infty} g_n(\sigma_n) d\sigma_n \end{aligned} \tag{5.1}$$

Thus the following will produce correct samples from $p(\mathcal{T}|n)$ for any pure-birth memoryless model:

Pure-birth memoryless sampling approach (*PBMSA*)

1. Simulate a tree terminating at s_n
2. Add a distance, δ, to the pendant edges using the correct $h(\delta)$ from Equation (5.1)
3. Repeat from step 1 until all samples are obtained

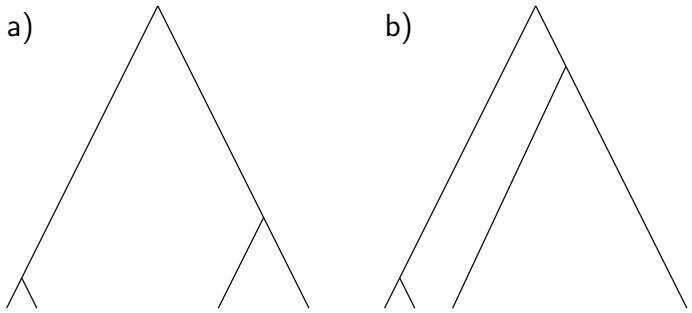

Figure 5.2: Consider an evolutionary model where 'young' species have a higher chance of speciating. Under this model the tree in Panel a is expected to have four species for a shorter duration than the tree in Panel b. The tree in Panel b should therefore make a greater contribution to our sample if we want to sample trees from the model conditional on them having four leaves.

For *SSA* to be appropriate we require $h(\delta) = g_n(\delta)$. Inspection of Equation (5.1) reveals that this requirement is met if $g_n(\sigma_n)$ is an exponential distribution. Furthermore as the model is memoryless the parameter may depend only on the number of species that are extant. These conditions are clearly satisfied by the Yule model, the coalescent and the related Moran [65] and Hey models [37].

PBMSA is appropriate for any model where the time between speciation events depends only on the number of extant species, however the Yule model and the coalescent are the only widely used models that fit this category. *PBMSA* is inappropriate for models with explicit extinction events and models with a memory. Explicit extinction events will result in a simulated tree that may have n species for several intervals – *PBMSA* would only sample from the first of these intervals resulting in a tree that is younger than expected.

Many models feature a memory, this may be in the form of hereditary speciation rates (e.g. [36]) or a dependence of speciation rates on the absolute age of a tree or a species (e.g. [9]). *PBMSA* cannot sample from such models as the evolution before and after s_n is not independent and different simulations to s_n should make different contributions to the final sample. Consider a model where young species are much more likely to speciate than their older counterparts. Figure 5.2 shows two simulated trees to s_n where $n = 4$. In Figure 5.2a there are four young species, in Figure 5.2b there are only two young species (those produced at s_n). Consequently the tree in Figure 5.2a is expected to have n species for a shorter time than the tree in Figure 5.2b and by the definition of $p(\mathcal{T}|n)$, the tree in Figure 5.2a should give a smaller contribution to that density than the tree in Figure 5.2b. Consequently it is necessary to take different numbers of samples from each of the evolutionary histories and *PBMSA* cannot be used.

5.1.3 A general sampling approach

We now introduce a general sampling method that works for a broad class of models that can include both speciation and extinction events. Our sampling approach simulates a tree, $\mathcal{T}$, until it is highly unlikely that the tree will return to n species. This will occur either when all species are extinct or when there has been sufficient speciation such that the number of extinctions required to return to n species are highly improbable.

The only restriction on the class of models from which our algorithm can sample is that we must be able to guarantee that each simulation 'run' will eventually terminate. The efficiency of the algorithm depends on the time that is required until a simulation terminates. An example of a model to which this algorithm can not be applied is one where the number of species perpetually fluctuates over a range including n.

Determining how unlikely a tree is to return to n species depends on the model. Throughout the remainder of this section we assume that we can determine a critical number of species, n^*, from which it is unlikely that extinctions will bring the number of species back to n. A simulation therefore ends when the number of species reaches 0 or n^*.

For some models the termination condition may be much more complicated, consider a model with evolving speciation and extinction rates – an appropriate termination condition will depend both on the number of species and on the speciation and extinction rates.

A simulation run will have k periods during which n species were extant, we denote the length of each of these periods by $\phi_i, i = 1, \ldots, k$. As previously discussed, the probability of observing a simulated tree whilst it has n species is directly proportional to the duration for which n species existed: $\Phi = \sum_{i=1}^{k} \phi_i$. The value Φ will vary between simulations so each simulation should make a different contribution to the final sample – a simulated tree where n species existed for a short period of time should make a lower contribution to the sample than a simulated tree where n species existed for a longer period.

The question remains how to decide on the number of samples to take from a given simulated tree, this should be proportional to Φ. To take this into account we introduce a sampling rate, r, such that we will take $r\Phi$ sampled trees from a given simulated tree. As we can only take whole samples of trees, for each simulated tree $r\Phi$ will be randomly rounded: If $r\Phi$ is between integers k and $k+1$, it is rounded down with probability $r\Phi - k$ and up with probability $1 - (r\Phi - k)$. This ensures that the randomly rounded $r\Phi$ has an expected value of $r\Phi$.

If the sampling rate is too low many simulations will be required for each sampled tree and the process will be very inefficient. If it is too high many sampled trees may be derived from a single simulated tree and these sampled trees will have a higher degree of correlation than expected for random samples. Ideally r should be determined experimentally (by simulations) such that it is as high as possible whilst ensuring that many simulated trees produce trees toward the final sample.

Lastly we introduce $S_i(\mathcal{T})$ as the set of trees that can be obtained by truncating

a simulated tree during the i-th interval during which it had n species. Combining these elements we have the following sampling approach:

General sampling approach (*GSA*)

1. Determine a suitable sampling rate, r, and a critical number of species, n^*
2. Simulate a tree, $\mathcal{T}$, until n^* species or extinction is reached
3. Find the number of trees to sample from $\mathcal{T}$: $r\Phi = \sum_{i=1}^{k} r\phi_i$
4. Randomly round $r\Phi$
5. For each sample required:
   (a) Randomly choose an interval, i, according to the weights ϕ_i
   (b) Sample a tree uniformly at random from $S_i(\mathcal{T})$
6. Repeat from step 2 until the required number of samples has been obtained

5.1.4 Extension of *GSA* to incomplete taxon sampling

Most n species trees based on real data will be a subsample of the m species contained in the true underlying tree, such that $m - n$ species are missing. This problem is referred to as incomplete taxon sampling (see e.g. [102]) and may be due to several reasons including inability to sample the species or a species being 'undiscovered'. If the number of species that are missing in a tree is substantial, incomplete taxon sampling should be included explicitly. A common approach is to sample trees with m species and randomly remove $m - n$ species, thus producing an n species tree as desired. For example, if only 75% of species are being sampled and we wish to sample a tree with 30 species, we would generate a tree with 40 species and remove 10 species uniformly at random. The problem with this approach is that we will generally only have an estimate of the number of missing species (25% in our example), hence we should consider a range of possible missing numbers of species. For instance in the previous example the true tree may have somewhere between, say, 35 and 50 species. We extend *GSA* to explicitly take into account incomplete taxon sampling. This extension of *GSA* requires either an estimate of the probability, s, of any given species being sampled or alternatively the probability distribution of the size of the true tree m, given the number of sampled species n, $p(m|n)$. Without one of these quantities our method cannot be applied and indeed, it is difficult to see how to proceed otherwise. Our method also assumes that sampled species are uniformly at random distributed through the tree. It is relatively straightforward to relax this last assumption, although we do not present any details here. One instance where this would be necessary is if the probability of sampling any two species is positively correlated to their proximity in the phylogenetic tree (as might be the case if whole clades are likely to be missed, or thoroughly sampled).

Given the sampling probability s, for a given real tree size, m, the number of sampled species, n, will be distributed according to a binomial distribution:

$$p(n|m) = \binom{m}{n} s^n (1-s)^{m-n}. \tag{5.2}$$

However the number of sampled species, n, is the size of the final tree and is what we wish to condition on, thus Bayes' Law gives us:

$$p(m|n) \propto p(n|m)p(m), \tag{5.3}$$

where $p(m)$ is the probability of a tree having m leaves and $p(n|m)$ is the probability of sampling n of those leaves. For $m \geq n$ it is always possible to obtain n leaves from a tree with m leaves, however the probability of this occurring decreases with m, such that $p(n|m)$ becomes small enough to make $p(m|n)$ negligible. This permits us to restrict the range of m that must be examined to $n \leq m \leq m^*$ where m^* is a limit that needs to be established. If we assume that $p(m)$ does not increase with m, an appropriate condition to solve for m^* is:

$$p(n|m^*+1) \leq \sum_{m=n}^{m^*} \frac{p(n|m)}{N}, \tag{5.4}$$

where N is the number of trees which are being sampled. This condition ensures that the first value of m being excluded is expected to contribute less than one tree to the final sample. If $p(m)$ increases with m extra analysis will be required to find an appropriate m^* (eg. using simulation studies).

Given a particular simulated tree we have $p(m) \propto \Phi_m$ (the duration for which a simulated tree had m species), hence substitution in Equation (5.3) gives:

$$p(m|n) \propto \Phi_m \binom{m}{n} s^n (1-s)^{m-n}, \tag{5.5}$$

which is readily normalised to give $p(m|n)$. The expected contribution to the sample from a given simulated tree consists of the expected contribution for each value of m:

$$r \sum_{m=n}^{m^*} \Phi_m p(m|n). \tag{5.6}$$

When a tree is simulated, the expected contribution to the sample is found and a sample of the corresponding size is taken. This process is repeated until the sample has the desired size.

GSA with incomplete taxon sampling

1. Find m^* analytically or by simulation / investigation (eg. Equation (5.4))

2. Simulate a tree, $\mathcal{T}$, until m^* species are reached or all species become extinct

3. If $p(m|n)$ is not given, calculate $p(m|n)$ for all m for this simulated tree (use Equation (5.5))

4. Find the expected number of samples to take from $\mathcal{T}$ (Equation (5.6))

5. Randomly round the expected number of samples

6. For each sample:

   (a) Randomly choose the original tree size, $\hat{m}$, according to $p(m|n)$

   (b) Uniformly at random choose a time when $\mathcal{T}$ had $\hat{m}$ species

   (c) Randomly delete $\hat{m} - n$ species

7. Repeat from step 2 until all samples have been obtained

5.1.5 Efficient sampling from the BDP

In this section we present an efficient algorithm for sampling trees with n species from the cBDP via inverse transform sampling. The method we propose relies on representing an oriented tree as a point process, as introduced in Section 3.4.1. In that section, we showed that the times s_i of the point process are independent and identically distributed. For $\lambda > \mu$, we have, from Theorem 3.4.4, the distribution function for a point in the point process:

$$F(s|t,\lambda,\mu,n) = \frac{1-e^{-(\lambda-\mu)s}}{\lambda-\mu e^{-(\lambda-\mu)s}} \frac{\lambda-\mu e^{-(\lambda-\mu)t}}{1-e^{-(\lambda-\mu)t}}$$

where t is the time of origin of the tree. The inverse of $F(s|t,\lambda,\mu,n)$ is:

$$F^{-1}(s|t,\lambda,\mu,n) = \frac{1}{\lambda-\mu} \ln\left(\frac{\lambda-\mu e^{-(\lambda-\mu)t} - \mu(1-e^{-(\lambda-\mu)t})s}{\lambda-\mu e^{-(\lambda-\mu)t} - \lambda(1-e^{-(\lambda-\mu)t})s}\right).$$

Further, for our sampling approach, we need the probability density of the time of origin of the tree, t, conditional on having n species at the present (assuming a uniform prior for the age of the tree). This distribution is derived in Theorem 3.4.7 for $\lambda > \mu$:

$$Q(t|\lambda,\mu,n) = \left(\frac{\lambda(1-e^{-(\lambda-\mu)t})}{\lambda-\mu e^{-(\lambda-\mu)t}}\right)^n.$$

The inverse of Q is

$$Q^{-1}(t|\lambda,\mu,n) = \frac{1}{\lambda-\mu} \ln\left(\frac{1-\frac{\mu}{\lambda}t^{1/n}}{1-t^{1/n}}\right).$$

For $\lambda = \mu$, the functions $F(s|t, \lambda, \lambda, n)$ and $Q(t|\lambda, \lambda, n)$ are the limiting functions of $F(s|t, \lambda, \mu, n)$ and $Q(t|\lambda, \mu, n)$ as $\mu \to \lambda$. In Section 3.4, we obtained,

$$\begin{aligned} F(s|t, \lambda, \lambda, n) &= \frac{s}{1+\lambda s}\frac{1+\lambda t}{t} \\ F^{-1}(s|t, \lambda, \lambda, n) &= \frac{st}{1+\lambda t(1-s)} \\ Q(t|\lambda, \lambda, n) &= \left(\frac{\lambda t}{1+\lambda t}\right)^n \\ Q^{-1}(t|\lambda, \lambda, n) &= \frac{1}{\lambda(s^{-1/n}-1)} \end{aligned}$$

Combining these probability distributions and the point process representation we obtain the following algorithm:

Constant rate birth-death approach (*BDA*)

1. Sample $r_0, \ldots, r_{n-1}$ uniformly at random from $[0, 1]$
2. Calculate the age of the tree, $t = Q^{-1}(r_0|\lambda, \mu, n)$
3. Calculate the $n-1$ branching times, $s_i = F^{-1}(r_i|t, \lambda, \mu, n), i = 1, \ldots, n-1$
4. Construct the sampled tree from the point process representation
5. Repeat from step 1 until all samples have been obtained

The advantage of this method over *GSA* is that it is unnecessary to determine n^* and r. The disadvantage of this method is that it gives no information about extinct lineages (regardless of the value of μ). If this information is required, *GSA* must be used for sampling from the BDP. Finally note that a sample from the Yule model can be obtained by setting $\mu = 0$.

5.1.6 Other sampling approaches

We have presented two main sampling approaches – *PBMSA* and *GSA*. *PBMSA* applies only to a limited class of evolutionary models that includes the Yule model and the coalescent (for which *PBMSA* becomes equivalent to *SSA*). *GSA* applies to a much wider class of models including some for which *SSA* has been used inappropriately. Application of *GSA* to a given model is relatively straightforward regardless of the model's complexity. However the generality of this approach makes it a mathematically unsatisfying and relatively inefficient process (from a computational perspective).

For some models it may be possible to derive the probability density for the time of individual speciation events explicitly. This has been done for the cBDP, trees can be sampled from the model via *BDA* – this is the most efficient way to sample trees from a cBDP of which we are aware.

For other models it may be possible to obtain the joint density of the speciation times. For example this was all that was known for the cBDP [100], prior to the result discussed in Section 5.1.5. In such cases an MCMC approach can be used to sample trees from this density.

5.2 Comparison of the sampling approaches

We have shown that *SSA* is only appropriate for models without extinction where the time between speciation events is exponentially distributed with a rate parameter that depends only on the number of species that are extant. The two most popular models – the Yule and coalescent – satisfy these conditions and it is appropriate to sample from them using *SSA*. We speculate that the simplicity of *SSA* combined with its correctness for the two most popular models has resulted in its inappropriate application to other models.

Existing approaches (such as *SSA*) are conceptually and computationally simpler than those introduced in this chapter, they have also been applied to many situations in existing studies for which they are inappropriate. It is therefore of great importance to consider how significantly the samples produced by the approaches differ. In situations where the difference is minimal it may be appropriate to use the simpler existing approaches to produce an approximate sample, if the difference is great it will be necessary to use more complicated approaches such as those presented here.

Lastly we note that for the remainder of this chapter we will disregard the events before the *mrca*. We define the *mrca age* of a sampled tree as the distance between the *mrca* and the leaves. This corresponds to realistic situations where it is often difficult to determine the evolutionary history before the *mrca*.

5.2.1 Speciation times under the cBDP

We begin by comparing *SSA* and *GSA* using a cBDP model. A cBDP includes two parameters – the speciation rate and the extinction rate – for our analysis it is sufficient to consider the ratio of these, hence we set the speciation rate to one. If the extinction rate is zero the model is equivalent to the Yule model. By increasing the extinction rate from zero to one the model becomes increasingly different from the Yule model and *SSA* should become increasingly inappropriate.

Figure 5.3 shows the expected *mrca* age of the tree as a function of the extinction rate for samples of five thousand trees produced by both sampling algorithms. When the extinction rate is zero the model is equivalent to the Yule model and the two approaches provide the same sample of speciation times. As the extinction rate is increased, the *mrca* age of the trees sampled by *GSA* also increases as this effectively reduces the net speciation rate, resulting in older trees.

We have shown that the absolute *mrca* age of the tree differs for the two sampling approaches, however in some situations the relative timing of the speciation events may be all that matters. To investigate this feature we consider LTT plots

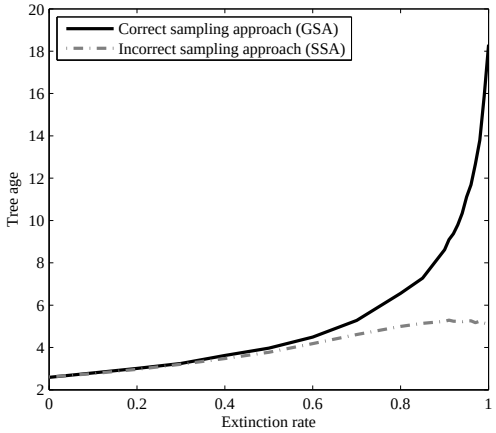

Figure 5.3: This figure shows the expected *mrca* age for twenty-species trees sampled from the BDP as a function of the extinction risk. The speciation rate was set to one and five thousand trees were sampled for each extinction rate using *SSA* (dotted line) and *GSA* (solid line). The *mrca* age of the trees sample by *GSA* increases as the extinction rate increases – this is because the net speciation rate is effectively reduced. *SSA* only considers the first time period during which n species existed, hence trees sampled using *SSA* do not exhibit the same *mrca* age increase.

(introduced in Section 3.6). Figure 5.4 shows the expectation of the LTT plot for an extinction rate of 0.95 from a sample of five thousand trees produced using the two algorithms. There is a clear difference between *GSA* and *SSA*.

The slope near the origin of a log transformed LTT plot can be used to give an estimate of the net speciation rate. In Figure 5.4 we consider the difference between this slope for the two methods, as a function of the extinction rate. Interestingly around an extinction rate of 0.9 the bias switches from negative to positive.

Extinction rates have been estimated to be around 0.9 of the speciation rate [61, 80]. At this value the two sampling approaches differ significantly in the estimated *mrca* age of the tree. For the relative timing of speciation events the result is not as clear, the severity (and direction) of the bias depends strongly on the extinction rate.

5.2.2 Tree shapes

The shape or topology of a tree is the structure obtained by disregarding the timing of speciation events. All memoryless models are CAL models and produce trees with the same tree shape distribution at all times – the uniform distribution on ranked oriented trees, see Proposition 2.3.1. The reason for this is that there is nothing to differentiate between species, hence, regardless of the model, each species is always equally likely to be the one that undergoes the next speciation or extinction event. Since *SSA* does not distinguish between species it correctly samples the tree shape

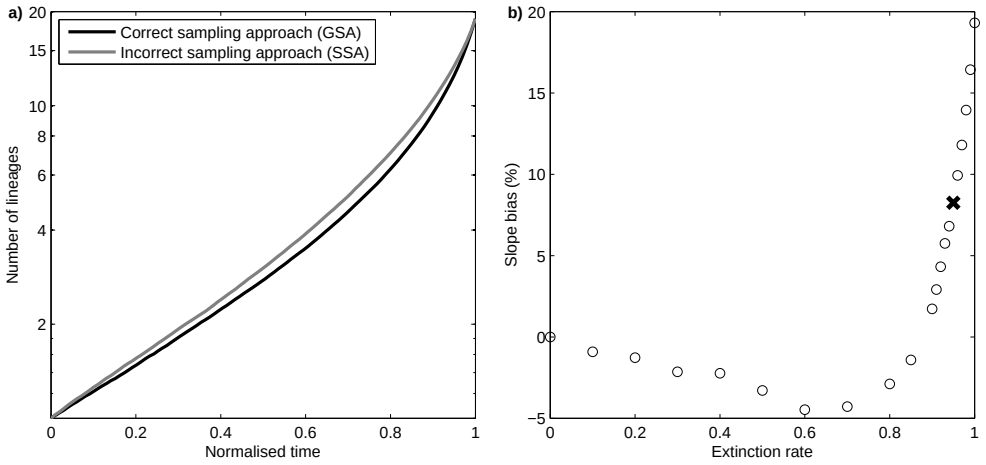

Figure 5.4: Panel a: An expected LTT plot is shown here for five thousand, twenty-species trees sampled from a BDP using both *SSA* and *GSA*. The speciation rate was set to one and the extinction rate to 0.95. The trees have been rescaled to have *mrca* age one – this removes the effect seen in Figure 5.3 and permits us to explore the relative speciation times of both samples. Panel b: The initial slope in Panel a gives an estimate of the net speciation rate. Here we depict the percentage deviation of the slope obtained using *SSA* to that obtained with *GSA* for different extinction rates. The point corresponding to Panel a is marked.

distribution for memoryless models (with or without extinction).

SSA may incorrectly sample the tree shape distribution from models that feature a memory. For pure-birth models, the mechanism behind this would require a correlation between the shape of a tree and the duration for which n species exist. This correlation is not explicit in any common models of which we are aware, but may exist implicitly; the strength of the correlation will determine the suitability of *SSA* to sample from a given model. We investigated two of the more common models with a memory [36, 8] and found minimal bias in the tree shape distribution produced by *SSA*.

For other models *SSA* may introduce a more serious bias in the tree shape distribution. One of the most obvious cases is a model with extinction where the tree shape distribution changes over time – as we have seen *SSA* produces trees that are too young, hence the tree shape distribution would be sampled too early.

5.2.3 Incomplete taxon sampling

The most common approach for incomplete taxon sampling samples a tree containing the expected true number of species, $\hat{m}$, and then randomly deletes $n - \hat{m}$ of these species. We applied this common approach and our developed approach in Section 5.1.4 to the cBDP and found that the sampled trees differed negligibly for these two approaches. There are two main issues with the common approach, in this section

we illustrate why each issue results in only a negligible bias:

Issue 1: Consider the cBDP, Figure 5.5 top panal shows how the expected *mrca* age of a ten-species tree suffering from incomplete taxon sampling increases as a function of the true tree size. It is important to note that this is near-linear; in Section 3.7.3 we showed that for the cBDP the relationship is linear when the extinction rate is one, and becomes slightly non-linear as the extinction rate is decreased. If this relationship were perfectly linear, for arbitrary $\hat{m} \geq n$, simulating a tree with $\hat{m}$ species and then deleting $\hat{m} - n$ species gives a correct sample up to scaling time linearly (i.e. determining the *mrca* age of the tree). For the cBDP the deviation from linearity seems sufficiently small to be irrelevant for most purposes.

Issue 2: Given a probability s of sampling each species, a naive method for calculating the expected number of species would be $\hat{m} = n/s$. In Figure 5.5 bottom panal we show the distribution of the true tree size as calculated using Equation (5.5) for $s = 0.7$, due to the asymmetry of this distribution, its expecation exceeds n/s. In this example the difference between these expectations is about 0.5, this will result in a small bias towards younger trees.

For the cBDP, the bias introduced by using a simplistic incomplete sampling method is insignificant in contrast with uncertainty regarding the true number of species. For other models it may be necessary to use the approach outlined above. This will particularly be the case for models that exhibit a strong non-linearity in the expected *mrca* age curve shown in Figure 5.5.

5.3 Concluding comments on sampling trees

When exploring evolutionary models, analytic results are preferable to simulation studies because of the smaller computational burden and greater insight they provide. However analytic results may be difficult to obtain and simulation studies may answer questions more quickly – once a result has been confirmed by simulation studies an analytic approach can be pursued with extra confidence.

Simulation studies have an inherent danger – it is extremely easy to simulate trees using a given model, however understanding what distribution these trees come from can be difficult. This makes it easy to proceed with a (possibly incorrect) method and therefore sample of trees. This is particularly problematic with more complicated evolutionary models where seemingly intuitive methods of simulating trees (such as *SSA*) often sample from undesirable and unrealistic probability distributions.

We have shown that a commonly used sampling approach is appropriate for two of the most common evolutionary models – the Yule model and the coalescent. However this approach is inappropriate for many other models to which it has been applied. For the cBDP, *SSA* produces a strong bias in the *mrca* age of the tree and the relative timing of speciation events. It does not produce a bias in the tree shape distribution. Further, for the cBDP, the common approach for incorporating incomplete taxon sampling seems adequate for most applications. More complex models with certain characteristics as discussed in this chapter may result in stronger biases of any of these attributes of a sampled tree.

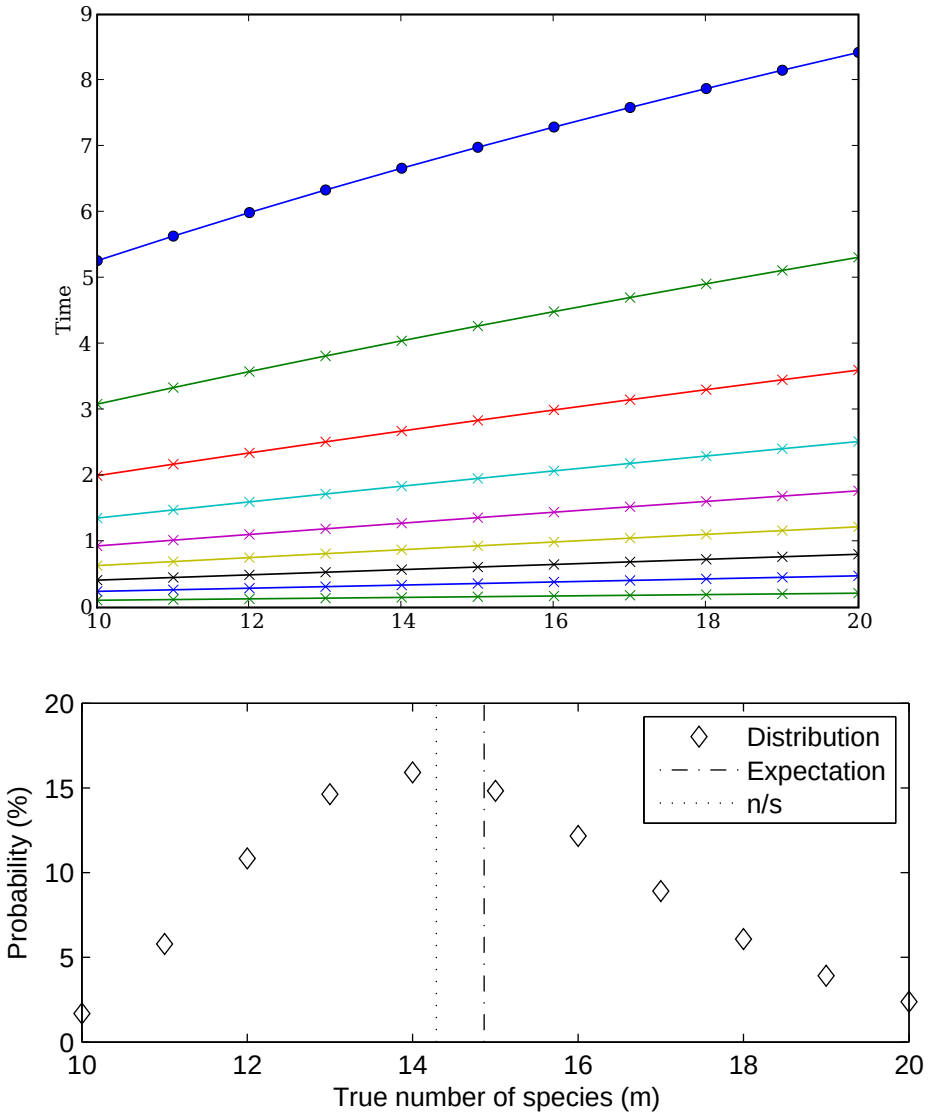

Figure 5.5: Top panel: The circles show the expected *mrca* age of a ten-species cBDP tree that has been sampled by constructing an m species tree and deleting $m - 10$ species. The speciation rate was set to 1 and the extinction rate to 0.9, the plot is drawn analytically with the formulae in Section 3.7.3. The crosses show the expected time of the speciation events in the same situation. The bottom panel shows the probability distribution of the true tree size, m, as calculated from Equation (5.5) for a sampling probability of $s = 0.7$. Also depicted are the expectation of this distribution (about 14.8) and a simple estimate of this, n/s (about 14.3). The plot is drawn via simulating a sample of five thousand trees for each value of m.

We suggest that for some of the studies that have sampled trees using *PhyloGen* and *SSA*, it would have been more appropriate to use our presented methods. It should be noted that in the cited studies the sampled trees were only one part of a complicated process (eg. to generate a data set for testing a tree construction method) and it is unlikely that the results would have been significantly effected by the chosen sampling method. For studies explicitly comparing speciation times in trees or sampling from more complicated models the distinction between these distributions will become crucial.

Our simulation package TreeSample [33] has built in support for the Yule model and the general cBDP model and is extendable to permit sampling from additional models.

Chapter 6

Reticulate evolution

Reticulation events are evolutionary events in which species pass on genetic material to co-existing lineages. Evolution with reticulation events, "reticulate evolution", cannot be displayed by a tree; binary networks are used to represent reticulate evolution. The common reconstruction methods for inferring these "reticulation networks" do not infer timing information. However, there are some restrictions for the dating of the vertices in a network without timing information: Reticulation events are assumed to happen instantaneously, whereas the time between successive speciation events is strictly positive. A network which can be dated with these constraints has a temporal labeling. Networks which do not have a temporal labeling can be modified via adding new taxa – species which we did not sample or which are extinct – to have a temporal labeling, see Figure 6.1. We show that determining the minimal number of taxa to add such that the network has a temporal labeling is NP-complete. This minimal number is a lower bound for the number of non-sampled taxa or extinct species in a (correctly) reconstructed network.

6.1 Introduction

The most common way to think of evolution is tree-like: Species evolve and pass genetic information to descendant species. However, for various organisms, reticulate evolution is a common process: genetic material is passed to co-existing lineages. We can have different reasons for reticulate evolution. In this chapter we consider the two major reticulation scenarios, hybridization and horizontal gene transfer (HGT). In the case of hybridization, two ancestor species combine their DNA and form a new species. This process is commonly found in plants and fish. In the case of HGT, one species contributes some DNA to another lineage, which especially occurs in bacteria. Directed acyclic graphs are used to model reticulate evolution, we call the graph displaying reticulate evolution a reticulation network. There are several algorithms around to infer a reticulation network for some given species, see for example the review article [62]. However, the reconstructed networks may not fulfill the necessary biological condition: We can assign a time to all nodes (temporal labeling) such that all species exist for a strictly positive time, whereas reticulation

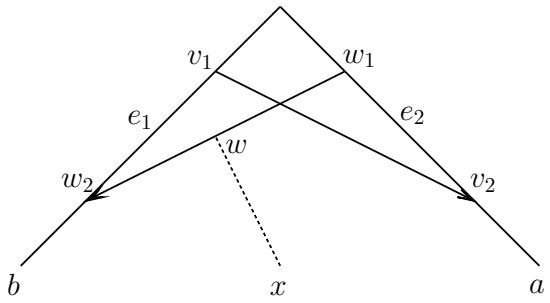

Figure 6.1: Network on species a and b with two reticulation events (v_1, v_2) and (w_1, w_2) at time $t_v := t(v_1) = t(v_2)$ and $t_w := t(w_1) = t(w_2)$. The network does not have a valid dating, since edge e_1 requires t_v to be before t_w and edge e_2 requires t_w to be before t_v. When inserting an additional species x, the resulting vertex has a temporal labeling: first we observe reticulation edge (v_1, v_2), then reticulation edge (w, w_2).

events happen instantaneously, see Figure 6.1.

If we have incomplete taxon sampling or extinct species not displayed, this condition can be violated in a (correctly) reconstructed network. When adding the missing taxa, the network again has a temporal labeling [4]. For example, the network in Figure 6.1 on species a, b, x has a temporal labeling, but the network restricted to species a, b has no temporal labeling. A natural question arises: Assume we have reconstructed a reticulation network correctly. What is the minimum number of taxa we do not display? We call this problem AddTaxa. We will show that AddTaxa is NP-complete via a reduction from FeedbackVertexSet. However we also show that AddTaxa is fixed parameter tractable.

Algorithms on reticulation networks like calculating the parsimony score of a reticulation network [42] implicitly assume a temporal labeling. In order to run these algorithms on any reconstructed network, we need to modify the reconstructed networks such that they have a temporal labeling. Adding the minimal number of (possibly) non-sampled taxa seems a reasonable way for the modification.

In Section 6.2, we define reticulation networks and temporal labelings formally and discuss their properties. In Section 6.3, we establish a reduction from FeedbackVertexSet to AddTaxa for HGT networks. We will see that HGT networks can be considered as special hybridization networks. Since FeedbackVertexSet is NP-hard, also AddTaxa for hybridization or HGT networks is NP-hard. We will show how to use FeedbackVertexSet algorithms for solving AddTaxa for HGT or hybridization networks in Section 6.4.

6.2 Modeling reticulate evolution

Directed acyclic graphs are commonly used to model reticulate evolution.

Definition 6.2.1. A directed connected acyclic graph is called a *general reticulation network* $\mathcal{N} = (V_\mathcal{N}, E_\mathcal{N})$, if its vertices belong to one of the following groups:

- *root* (outdegree 2, indegree 0)
- *leaves* (indegree 1, outdegree 0)
- *tree vertices* (indegree 1, outdegree 2)
- *reticulation vertices* (indegree 2, outdegree 1)

We will write $u < v$ for $u, v \in V_\mathcal{N}$ if there is a directed path from u to v, i.e. u is an ancestor of v, v is a descendant of u. The edge (u, v) is called parent edge of v.Further, u is a direct ancestor of v, and v a direct descendant of u.

A reticulation vertex due to hybridization is also called a *hybridization vertex*. We call the two edges pointing to a hybridization vertex a *reticulation or hybridization edge*. A reticulation vertex due to HGT is also called a *HGT vertex*. We call the edge pointing to an HGT vertex which contributes some DNA to the other lineage a *reticulation or HGT edge*. In the displayed networks, an arrow corresponds to a reticulation edge.

Note that for a hybridization vertex, both parent edges are hybridization edges, whereas for an HGT vertex, only one parent edge is an HGT edge. The remaining edges are called tree edges.

We will assign dates to all interior vertices, modeling the time between speciation events. In doing so, we will see that it is sufficient to consider recombination networks (which will be defined below) instead of general recombination networks. We formalize the idea of assigning dates to the vertices:

Definition 6.2.2. A *labeling* of a network is any map from the set of interior vertices of the network to the real numbers. The real number assigned to a vertex by a labeling is called the vertex label. A *temporal labeling* is a *labeling* with the property that the vertices adjacent to a reticulation edge have the same labels and all other vertex labels are strictly increasing on any path from the root to the leaves.

In a general reticulation network, we could encounter (i) a species x undergoing two hybridization events at the same time and then immediately becoming extinct, see Figure 6.2, left. This scenario seems biologically not plausible. There must be the non-sampled species x' or x'', we add one of them.

Further, we could encounter (ii) the scenario that a hybrid x passes on DNA and goes extinct in the very moment of being created, see Figure 6.2, right. This scenario seems not plausible either. The non-sample species x' must exist, and thus we add it.

By adding the species x' or x'' in the cases (i) and (ii), we can consider reticulation networks which are specific general reticulation networks without the described anomalies (i) or (ii):

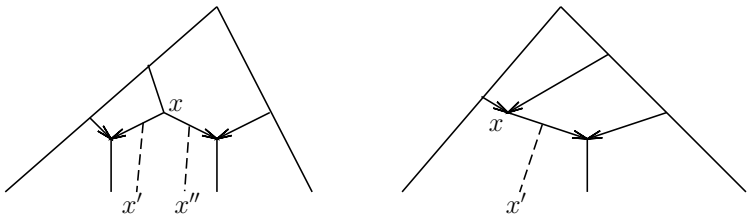

Figure 6.2: Reason for only considering reticulation networks and not general reticulation networks. In the left case, x would pass DNA to two different new species at the very same time and then go instinct which is not observed in biology. In the right case, x would pass on DNA in the moment of creation and go extinct which is not observed either. Therefore we add x' (or x'') and consider reticulation networks instead of general reticulation networks.

Definition 6.2.3. A *reticulation network* is a general reticulation network where at least one tree edge is descending from any interior vertex. A *hybridization / HGT network* is a reticulation network where all reticulation events are due to hybridization / HGT. Figure 6.3 shows a hybridization network and Figure 6.4 shows an HGT network. A reticulation network $\mathcal{N}$ is a *temporal network*, if it has a temporal labeling.

If a reticulation network does not have a temporal labeling, we identify the reticulation vertices which are problematic:

Definition 6.2.4. Let (v, v_1) be a reticulation edge in the reticulation network $\mathcal{N}$. Let V_v be the subset of $V_\mathcal{N}$ with $w \in V_v$ iff $v < w$ but $\neg(v_1 \leq w)$. We call V_v the critical area of v. The vertex v is called *critical vertex*, if V_v contains at least one reticulation vertex. If v is critical, the reticulation event v_1 is called a *critical reticulation vertex* or *critical reticulation event*.

Consider a hybridization network. We will show that if the critical reticulation vertices and the critical vertices have a labeling such that the vertex labels increase on any path from the root to a leaf (constant on reticulation edges and strictly increasing otherwise), the whole network has a temporal labeling. We will see later that the temporal labeling of an HGT network can be described in a hybridization network setting, therefore the next lemma also holds for HGT networks.

Lemma 6.2.5. *Consider a hybridization network. Suppose we have a labeling for all critical reticulation events and all critical vertices, such that the vertex labels are increasing on any path from the root to a leaf (constant on reticulation edges and strictly increasing otherwise). Then $\mathcal{N}$ has a temporal labeling.*

Proof. W.l.o.g. let the labels of the critical vertices be bigger than zero. First, we show that we can find a labeling for the non-critical reticulation events such that the labelings are increasing on any path from the root to a leaf (constant on reticulation

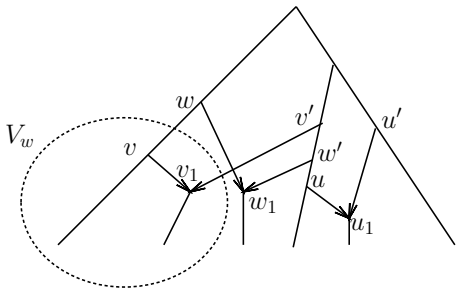

Figure 6.3: Network modeling hybridization. The vertices v', w, w' are critical vertices, i.e. the hybridization events v_1, w_1 are critical hybridization events. The critical area of w is indicated by the dotted circle. Note that the network does not have a temporal labeling.

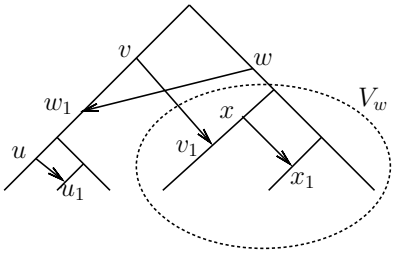

Figure 6.4: Network modeling horizontal gene transfer. The vertices v, w, x are critical vertices, i.e. the HGT events v_1, w_1, x_1 are critical HGT events. The critical area of w is indicated by the dotted circle. Note that the network does not have a temporal labeling.

edges and strictly increasing otherwise). Let the reticulation vertex v_1 be without a label, with the parent vertices v and v'. The vertices v and v' are tree vertices (by definition of a hybridization network – each vertex has at least one tree edge descending, i.e. the indegree of v and v' is one) and have only tree vertices in V_v and $V_{v'}$ (since v_1 is a non-critical reticulation vertex). If v_1 has descendants with a label, the label of the oldest descendant is denoted by t_d. If there is no labeled descendant, set $t_d = \infty$. If v_1 has a labeled ancestor, the label of the youngest labeled ancestor is denoted by t_a. If there is no labeled ancestor, set $t_a = 0$. Label the vertices v, v_1, v' with $t_v \in (t_a, t_d)$. In that way we assure $t_d - t_a > 0$ throughout the whole construction of the labeling.

Once this labeling is done for all reticulation events, we only have tree vertices left to label. Start at the root, call it v_r. Set the label of the root to 0. Define $V_r := \emptyset$. Set the label of the leaves to t_l where t_l is bigger than any existing label. The following algorithm dates all tree vertices.

1. Consider $V_{\mathcal{T}}$, the largest connected subset of tree vertices which is connected

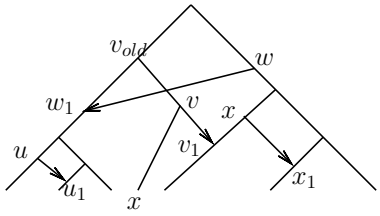

Figure 6.5: The result of applying the operation $add(\mathcal{N}, v)$ to the network $\mathcal{N}$ shown in Figure 6.4. The new network has a temporal labeling.

to v_r. Let the label of v_r be t_a. Add the set of direct descendants of $V_\mathcal{T}$ to V_r. Let t_d be the smallest label of the direct descendant of $V_\mathcal{T}$. Assign labels with values in (t_a, t_d) to the vertices in $V_\mathcal{T}$ such that the labels are increasing on any path from the root to the leaves.

2. If $V_r = \emptyset$, stop the algorithm. The whole network is labeled. If $V_r \neq \emptyset$, take any node v_r from V_r, delete it from V_r. Proceed with (1).

□

Note that we can modify any network $\mathcal{N}$, such that only non-critical vertices are in the network, with the following operation:

Definition 6.2.6. Given an edge $e = (v, v_1)$ in a reticulation network $\mathcal{N}$. Delete e. Rename vertex v to v_{old} and add a new vertex v. Add an edge (v_{old}, v) and an edge (v, v_1). Add a taxa x below v. This operation is called $add(\mathcal{N}, v)$, for an illustration see Figure 6.5. Consider a set of vertices $\mathcal{V}$ in $\mathcal{N}$. The result of applying $add(\mathcal{N}, v)$ for all $v \in \mathcal{V}$ to $\mathcal{N}$ is called $add(\mathcal{N}, \mathcal{V})$.

In a given network $\mathcal{N}$, we can eliminate all critical vertices in the following way: For a reticulation edge (v, v_1) where v is critical, do the operation $add(\mathcal{N}, v)$. The new vertex v is non-critical, since the descendant x is just a leave. In a hybridization network without critical vertices, we can always find a temporal labeling according to Lemma 6.2.5. So we can modify any hybridization network $\mathcal{N}$ by adding new taxa, i.e. new leaves, such that it has a temporal labeling. This motivates the following definition.

Definition 6.2.7. AddTaxa
Instance: Reticulation Network $\mathcal{N} = (V, E)$, positive integer K.
Question: Is it possible to obtain a network with a temporal labeling by adding k new taxa to $\mathcal{N}$, $k \leq K$?

Remark 6.2.8. A HGT network as an input for AddTaxa can be seen as a special case of a hybridization network. Let $e = (v, v_1)$ be an HGT edge in an HGT network. Let w be the second direct ancestor of v_1. Do the operation $add(\mathcal{N}, w)$. Note that the new w and v_1 can always occur at the same time since w has as a directed

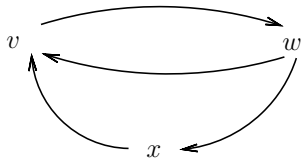

Figure 6.6: The critical graph $\mathcal{C}_\mathcal{N}$ of the network $\mathcal{N}$ shown in Figure 6.4.

descendant (besides v_1) only a leaf x. So if we want to know whether an HGT network has a temporal labeling, we add a taxa to each HGT event as described here. We can then check if this modified network has a temporal labeling under hybridization. If the HGT network does not have a temporal labeling, the number of taxa we have to add is the same as in the modified hybridization network, since v_1 and w can always occur instantaneously (v_1 again being an HGT vertex, and w being the direct ancestor where a new taxa was attached).

It is possible to check in polynomial time whether a hybridization network has a temporal labeling [4]. So the problem AddTaxa is in NP.

We will show that AddTaxa is actually NP-complete. To do so, we prove that AddTaxa is NP-complete for HGT networks. Since the HGT networks can be considered as a special class of hybridization networks (Remark 6.2.8), we establish that AddTaxa is NP-complete for hybridization networks as well.

6.3 Proof: AddTaxa is NP-complete

6.3.1 The critical graph

Given an HGT network, we want to add taxa such that the critical reticulation events and critical vertices have a labeling where labels are increasing on any path from the root to the leaves (constant on reticulation edges and strictly increasing otherwise). Then the whole network has a temporal labeling by Lemma 6.2.5. Note that adding taxa to non-critical reticulation events does not influence the existence of a temporal labeling. We will introduce the *critical graph* for a network, in order to prove NP-completeness of AddTaxa. For the HGT network $\mathcal{N}$, construct the critical graph $\mathcal{C}_\mathcal{N}$ as follows.

- The vertex set $V_{\mathcal{C}_\mathcal{N}}$ is the set of critical vertices of $\mathcal{N}$.
- The edge set $E_{\mathcal{C}_\mathcal{N}}$ is defined as follows. $(v, u) \in E_{\mathcal{C}_\mathcal{N}}$ iff $(v < u)$ or $(v < u_1)$ in $\mathcal{N}$ (where u_1 is the reticulation vertex adjacent to u). Note that the edges of $\mathcal{C}_\mathcal{N}$ are directed.

We define an operation on $\mathcal{C}_\mathcal{N}$:

Definition 6.3.1. Delete vertex v and all its incident edges in $\mathcal{C}_\mathcal{N}$. This operation is called $delete(\mathcal{C}_\mathcal{N}, v)$. Consider a set of vertices $\mathcal{V}$ in $\mathcal{C}_\mathcal{N}$. The result of applying $delete(\mathcal{C}_\mathcal{N}, v)$ for all $v \in \mathcal{V}$ to $\mathcal{C}_\mathcal{N}$ is called $delete(\mathcal{C}_\mathcal{N}, \mathcal{V})$.

Theorem 6.3.2. *For a vertex $v \in \mathcal{C}_\mathcal{N}$, we have*

$$\mathcal{C}_{add(\mathcal{N},v)} = delete(\mathcal{C}_\mathcal{N}, v).$$

Proof. We add a taxa x between v and v_1 by $add(\mathcal{N}, v)$. The new parent v of v_1 has just the new taxa x as a descendant. In particular, v has no HGT vertex as a descendant in V_v, so v is non-critical. So the vertex set of $\mathcal{C}_{add(\mathcal{N},v)}$ is the vertex set of $\mathcal{C}_\mathcal{N}$ excluding v. The edges between the remaining vertices are the same as in $E_{\mathcal{C}_\mathcal{N}}$. So $\mathcal{C}_{add(\mathcal{N},v)} = delete(\mathcal{C}_\mathcal{N}, v)$. □

Lemma 6.3.3. *Let $G = (V, E)$ be a finite acyclic graph. Then G has a vertex with indegree 0.*

Proof. Let $G = (V, E)$ be a finite acyclic graph with n vertices. Assume G contains no vertex with indegree 0. Choose any vertex $u_1 \in V$. This vertex is an endpoint of the edge (u_2, u_1) since its indegree is non-zero. Vertex u_2 has non-zero indegree, so there exists an edge (u_3, u_2). Proceed in that way until the n vertices are chosen or a cycle is closed. Suppose the path $u_n, u_{n-1}, \ldots, u_1$ contains no cycle. Since u_n has non-zero indegree, we have an edge (u_i, u_n) which closes a cycle. This contradicts our assumption. □

Theorem 6.3.4.

$$\mathcal{C}_\mathcal{N} \text{ is acyclic} \Leftrightarrow \mathcal{N} \text{ has a temporal labeling.}$$

Proof. $' \Rightarrow'$ Since $\mathcal{C}_\mathcal{N}$ is acyclic, it has a vertex v with indegree 0 by Lemma 6.3.3. In $\mathcal{N}$, the edge (v, v_1) is the reticulation edge attached to v. Note that no vertex in $\mathcal{C}_\mathcal{N}$ is ancestor of v or v_1 in $\mathcal{N}$ since v has indegree 0. Let the labels of v_1, v be $t(v_1) = t(v) = 1$. Delete v in $\mathcal{C}_\mathcal{N}$. The remaining graph again has a vertex with indegree 0, the label of the corresponding reticulation event shall be 2. Continue until we end up with the empty graph. Now label the remaining vertices in $\mathcal{N}$. This is straightforward; only tree nodes and non-critical HGT events are left. These vertices can be labeled such that we have a temporal labeling according to Remark 6.2.8 and Lemma 6.2.5.
$' \Leftarrow'$ Assume $\mathcal{C}_\mathcal{N}$ has a cycle $u, v, \ldots, u$. Let u_1, v_1 be the reticulation vertices adjacent to u, v. Because of the cycle, u is an ancestor of v or v_1, and also v is an ancestor of u or u_1. This violates the assumption of the existence of a temporal labeling. □

Corollary 6.3.5. *Let $\mathcal{V}$ be a subset of vertices of $\mathcal{C}_\mathcal{N}$. Then*

$$add(\mathcal{N}, \mathcal{V}) \text{ has temporal labeling} \Leftrightarrow delete(\mathcal{C}_\mathcal{N}, \mathcal{V}) \text{ is acyclic}$$

Proof. With Theorem 6.3.2, we have $\mathcal{C}_{add(\mathcal{N},\mathcal{V})} = delete(\mathcal{C}_\mathcal{N}, \mathcal{V})$. Theorem 6.3.4 establishes the corollary. □

Figure 6.7: Network $\mathcal{N}_0$. Note that $\mathcal{C}_{\mathcal{N}_0} = G_0$.

The following problem can be reduced to AddTaxa as we will show below.

Definition 6.3.6. FeedbackVertexSet
Instance: Directed connected graph $G = (V, E)$, positive integer $K \leq |V|$.
Question: Is there a subset $V' \subseteq V$ with $|V'| \leq K$ such that $G \setminus V'$ is acyclic?

FeedbackVertexSet is NP-complete [45]. It is even NP-complete for input graphs of indegree at most two and outdegree at most two (degree-two graphs), we call that problem FeedbackVertexSet2.

With Corollary 6.3.5, we have for an HGT network $\mathcal{N}$,

$$\text{AddTaxa}(\mathcal{N}, K) = \text{FeedbackVertexSet}(\mathcal{C}_{\mathcal{N}}, K),$$

i.e. solving AddTaxa for $\mathcal{N}$ is equivalent to solving FeedbackVertexSet for the critical graph $\mathcal{C}_{\mathcal{N}}$.

In the following, we reduce FeedbackVertexSet2 to AddTaxa – we show that we can convert (in polynomial time) any input graph G for FeedbackVertexSet2 to an HGT network $\mathcal{N}$ with $\mathcal{C}_{\mathcal{N}} = G$. Since $\text{AddTaxa}(\mathcal{N}, K)$ and $\text{FeedbackVertexSet2}(\mathcal{C}_{\mathcal{N}}, K)$ return the same answer, and since we know that FeedbackVertexSet2 is NP-hard, AddTaxa is also NP-hard.

6.3.2 The reduction

Let G be any degree-two graph with n vertices. We want to convert G into an HGT network $\mathcal{N}$, such that $\mathcal{C}_{\mathcal{N}} = G$. We will show that this conversion is always possible by giving a polynomial algorithm for the conversion. This will establish the NP-completeness of AddTaxa, since we can solve FeedbackVertexSet2 with an algorithm for AddTaxa.

We will show by induction on k that for any degree-two graph with k edges, G_k, there is a network $\mathcal{N}_k$ with $\mathcal{C}_{\mathcal{N}_k} = G_k$. For $k = 0$, we have $\mathcal{C}_{\mathcal{N}_0} = G_0$ for the network $\mathcal{N}_0$ as shown in Figure 6.7. Assume that for for all $m < k$, we have for arbitrary G_m a network $\mathcal{N}_m$ such that $\mathcal{C}_{\mathcal{N}_m} = G_m$.

For arbitrary G_k, we will now construct a network $\mathcal{N}_k$ such that $\mathcal{C}_{\mathcal{N}_k} = G_k$. Delete an arbitrary edge (u, v) from G_k to obtain the graph G_{k-1}. By our induction

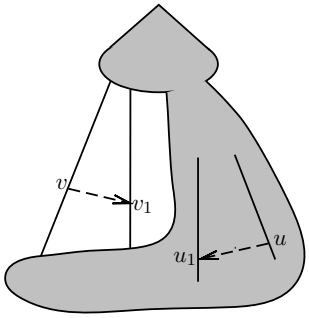

Figure 6.8: Network $\mathcal{N}_{k-1}$. In the gray area, we have some arbitrary network structure with reticulation edge (u, u_1). Note that u is not an ancestor of v or v_1.

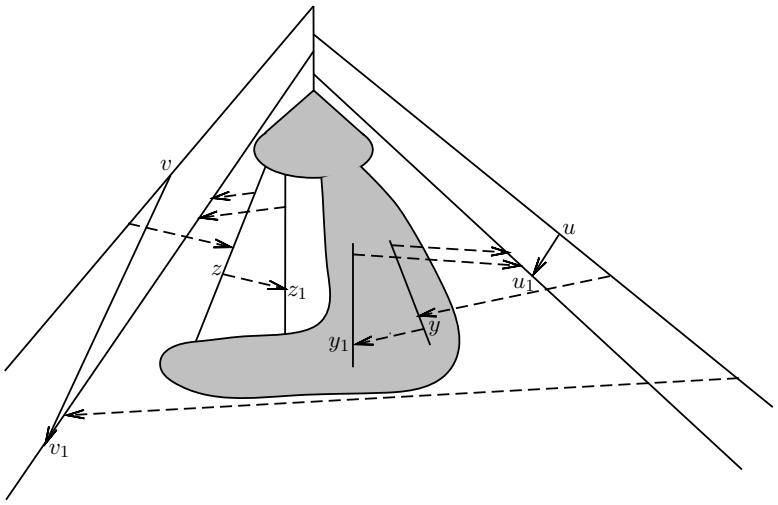

Figure 6.9: The network $\mathcal{N}_{k-1}$ in Figure 6.8 is altered to a network $\mathcal{N}_k$ shown in this figure. The gray area remains unchanged. We rename u, u_1 to y, y_1 and v, v_1 to z, z_1. The critical graph of $\mathcal{N}_k$ is the critical graph of $\mathcal{N}_{k-1}$ with the additional edge (u, v). Note that each dashed HGT edge e in $\mathcal{N}_k$ has a new taxa attached with $add(\mathcal{N}_k, e)$. We did not display the additional taxa for more clarity in the figure.

assumption we have a network $\mathcal{N}_{k-1}$ with $\mathcal{C}_{\mathcal{N}_{k-1}} = G_{k-1}$. Now we alter $\mathcal{N}_{k-1}$ to $\mathcal{N}_k$ such that we have $\mathcal{C}_{\mathcal{N}_k} = G_k$.

Since there is no edge (u, v) in G_{k-1}, the network $\mathcal{N}_{k-1}$ has a structure as displayed in Figure 6.8. In the gray area, we have some network structure. The importance is, that vertex u is not an ancestor of v or v_1.

We convert $\mathcal{N}_{k-1}$ to the network $\mathcal{N}_k$ as shown in Figure 6.9. Note that each dashed HGT edge e does have a taxa added as well ($add(\mathcal{N}_k, e)$). So the HGT event is non-critical. We did not display the added taxa for more clarity of the figure. The gray area has the same network structure as the gray area in Figure 6.8. Vertices u, u_1, v, v_1 are renamed to y, y_1, z, z_1. It is easy to check that any ancestor of v (resp. u) in $\mathcal{N}_{k-1}$ remains an ancestor of v or v_1 (resp. u or u_1) in $\mathcal{N}_k$. Further any descendant of v (resp. u) in $\mathcal{N}_{k-1}$ remains a descendant of v (resp. u) in $\mathcal{N}_k$. All other relationships between critical vertices remain unchanged as well since we do not change the gray area. As we intended, u is now an ancestor of v, and overall $\mathcal{C}_{\mathcal{N}_k} = G_k$.

So for any degree-two graph G, we can iteratively obtain a network $\mathcal{N}$ with $\mathcal{C}_{\mathcal{N}} = G$. Since we add 13 new taxa and 9 new reticulation events to the network in each step, the network grows linear with the number of edges in G. Therefore the conversion from the degree-two graph to an HGT network is polynomial. This shows that every instance of FeedbackVertexSet2 can be reduced to AddTaxa, therefore we have proven,

Theorem 6.3.7. *The problem* AddTaxa *is NP-complete.*

Note that AddTaxa is NP-complete for HGT networks as well as hybridization networks, since HGT networks can be considered as a special case of hybridization networks (Remark 6.2.8).

6.4 Algorithms for solving AddTaxa

6.4.1 HGT and hybridization networks

For HGT networks, the previous section provides an algorithm for solving AddTaxa: We construct the critical graph and use an algorithm for FeedbackVertexSet (see Section 6.4.2). For hybridization networks, we can do the same approach. First, we define the critical graph for a hybridization network $\mathcal{N}$:

- The vertex set $V_{\mathcal{C}_\mathcal{N}}$ is the set of critical vertices of $\mathcal{N}$.
- The edge set $E_{\mathcal{C}_\mathcal{N}}$ is defined as follows. $(v, u) \in E_{\mathcal{C}_\mathcal{N}}$ iff $(v < u)$ or $(v < u_1)$ or $(v < u')$ in $\mathcal{N}$ (where u_1 is the reticulation vertex with direct ancestors u, u').

It is straightforward to see that Theorem 6.3.2 holds for the critical graph of hybridization networks as well. Theorem 6.3.4 is also valid for hybridization networks: Indeed, consider the hybridization vertex v_1 with parents v, v'. Note that if v and v' are in the critical graph, then $(u, v) \in E_{\mathcal{C}_\mathcal{N}} \Leftrightarrow (u, v') \in E_{\mathcal{C}_\mathcal{N}}$ by definition

of the critical graph. Therefore vertex v having indegree 0 is equivalent to vertex v' having indegree 0. For assigning the labels, proceed as in the proof of Theorem 6.3.4: Choose a critical vertex v with indegree 0. Label $t(v) = t(v') = t(v_1) = 1$. Delete v, v' from $\mathcal{C}_\mathcal{N}$. Choose the next vertex with indegree 0 and so on.

Since Theorem 6.3.2 and 6.3.4 hold for hybridization networks, Corollary 6.3.5 also holds for hybridization networks. Therefore solving AddTaxa for hybridization or HGT networks can be done by solving FeedbackVertexSet on the critical graph.

6.4.2 Algorithms for solving FeedbackVertexSet

FeedbackVertexSet and FeedbackVertexSet2 are NP-complete. In [10] it is shown that FeedbackVertexSet for directed graphs is fixed parameter tractable. The authors provide an algorithm which solves FeedbackVertexSet(G, k) in $O(4^k k! n^{O(1)})$ where n is the number of vertices in G.

No polynomial time approximation algorithms with constant ratio have been found. There is a polynomial time approximation algorithm with a ratio of $O(\log k \log\log k)$ [19] where k is the size of the minimum feedback vertex set.

FeedbackVertexSet2 is reducible to AddTaxa, therefore AddTaxa is NP-hard. Furter, AddTaxa is reducible to FeedbackVertexSet, therefore AddTaxa is fixed parameter tractable. The size of our critical graph is determined by the number of critical vertices which is less than the number of reticulation events and this is usually much less than the size of the network. Therefore, for biological reticulation networks, the critical graph will be reasonable "small".

6.5 Summary

We showed that determining the minimal number of taxa to add to a hybridization or HGT network, such that the network has a temporal labeling (AddTaxa) is NP-complete. However, the critical graph of a reticulation network will be reasonable "small" for most biological instances, and therefore even brute-force algorithms might be feasible. Furthermore, AddTaxa is fixed parameter tractable.

Chapter 7

Outlook

The analytic results in the thesis have been applied to improve a variety of methods as explained in the different chapters. Concluding the thesis, I would like to point out the – in my opinion – next most important steps to generalize the various results and methods and to obtain further conclusions about evolution, in particular about the process of speciation and extinction.

We exhaustively discussed models for speciation and extinction, these models induce the "species tree distribution". Furthermore, we briefly discussed the coalescent as a model for the evolution within a population, the coalescent induces the "gene tree distribution". Note that when reconstructing phylogenies from genes, we actually reconstruct a gene tree which is evolving on a species tree. It is commonly assumed that the gene tree and the species tree are equal, therefore we reconstructed the species tree. However, this does not have to be the case. Recently, surprising and maybe unexpected results have been established for gene trees evolving on species trees [13]. I am very curious to see whether and how the distributions discussed in this thesis change, when considering gene trees evolving on species trees, rather than considering only species trees as done in the thesis. In particular, it will be interesting to understand the distribution of coalescent gene trees evolving on cBDP species trees – note that the cBDP and the coalescent on their own both induce a uniform distribution on ranked, oriented trees as discussed in the thesis. Another challenge is to establish the distribution of the Colless statistic and the runs statistic when gene trees are evolving on species trees.

In this thesis, we established the prior distribution for the cBDP which is implemented in the Bayesian inference program Beast. Since most clades are not fully sampled, an obvious next goal is to derive a prior which incooporates random taxon sampling.

Supertrees, like the primate tree of Rutger Vos and Arne Mooers, can now be dated efficiently without introducing a bias. It will be exciting seeing this method applied to other undated phylogenies, currently Jonathan Davies is using Cass for dating his Carnivora supertree. As soon as having an understanding of more complex models for speciation than the cBDP, establishing a method for dating phylogenies under such complex models should be addressed.

We derived LTT plots for the cBDP analytically. For the corresponding γ statistic, we only have limited analytic knowledge. In order to avoid unnecessary simulations, it will be useful to derive analytic results for the γ statistic as well.

Phylogenies can be tested for lineage-specific bursting with the runs statistic. We applied the statistic to example applications. It would be of great value to apply the statistic to the known dated phylogenies to investigate how frequent lineage-specific bursting appears. With our package Cass this is possible as soon as enough data is available. For the Colless statistic, such an analysis had been done on published tree shapes, with the result that data trees are less balanced than expected under the neutral models. Since the Colless statistic considers tree shapes, but our analysis requires a ranking with the shape, we cannot apply the runs statistic to the data used for the Colless analysis.

Note that the runs statistic is not only useful for phylogenies – samples of populations can be tested for neutrality with the runs statistic. As a great advantage over other tests, this test is not fooled by population size variation. Therefore, previous work assuming the coalescent for estimating the population size history should be checked again by applying the runs statistic to the data. This will indicate whether the coalescent is an appropriate assumption for the considered data.

Bibliography

[1] D. Aldous. Probability distributions on cladograms. In D. Aldous and R. Pemantle, editors, *Random Discrete Structures*, pages 1–18. Springer, Berlin, 1995.

[2] D. Aldous and L. Popovic. A critical branching process model for biodiversity. *Adv. in Appl. Probab.*, 37(4):1094–1115, 2005.

[3] D. J. Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statist. Sci.*, 16(1):23–34, 2001.

[4] M. Baroni, C. Semple, and M. Steel. Hybrids in real time. *Systematic Biology*, 44(1):46–56, 2006.

[5] J. O. Berger. *Statistical decision theory: foundations, concepts, and methods*. Springer-Verlag, New York, 1980. Springer Series in Statistics.

[6] L. J. Billera, S. P. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Adv. Appl. Math.*, 27(4):733–767, November 2001.

[7] O. R. Bininda-Emonds, M. Cardillo, K. E. Jones, R. D. MacPhee, R. M. Beck, R. Grenyer, S. A. Price, R. A. Vos, J. L. Gittleman, and A. Purvis. The delayed rise of present-day mammals. *Nature*, 446(7135):507–12, 2007.

[8] M. Blum and O. Francois. Which random processes describe the Tree of Life? A large-scale study of phylogenetic tree imbalance. *Systematic Biology*, 55(4):685–691, 2006. Comparison of the beta splitting model to pandit and tree base with good fits.

[9] K. M. A. Chan and B. R. Moore. Accounting for mode of speciation increases power and realism of tests of phylogenetic asymmetry. *American Naturalist*, 153(3):332–346, 1999.

[10] J. Chen, Y. Liu, and S. Lu. A fixed-parameter algorithm for the directed feedback vertex set problem. *Proceedings of the fourtieth annual ACM symposium on Theory of computing*, pages 177–186, 2008.

[11] D. H. Colless. Phylogenetics: the theory and practice of phylogenetic systematics. *Syst. Zool.*, 31(1):100–104, 1982.

[12] F. N. David and D. E. Barton. *Combinatorial chance*. Hafner Publishing Co., New York, 1962.

[13] J. H. Degnan and N. A. Rosenberg. Discordance of species trees with their most likely gene trees. *PLoS Genet*, 2(5), 2006.

[14] H. Dehling and B. Haupt. *Einfuehrung in die Wahrscheinlichkeitstheorie und Statistik*. Springer, 2003.

[15] A. Drummond, S. Ho, M. Phillips, and A. Rambaut. Relaxed phylogenetics and dating with confidence. *PLoS Biol.*, 4:e88, May 2006.

[16] A. Drummond and A. Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, 7:214, 2007.

[17] A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.*, 22(5):1185–1192, May 2005.

[18] A. W. F. Edwards. Estimation of the branch points of a branching diffusion process. (With discussion.). *J. Roy. Statist. Soc. Ser. B*, 32:155–174, 1970.

[19] G. Even. Approximating Minimum Feedback Sets and Multicuts in Directed Graphs. *Algorithmica*, 20(2):151–174, 1998.

[20] W. Feller. *An introduction to probability theory and its applications. Vol. I.* Third edition. John Wiley & Sons Inc., New York, 1968.

[21] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, Massachusetts, 2004.

[22] D. Ford, E. Matsen, and T. Stadler. A method for investigating relative timing information on phylogenetic trees. *submitted*, 2008.

[23] D. J. Ford. Probabilities on cladograms: introduction to the alpha model. *Manuscript*, 2005.

[24] Y. X. Fu and W. H. Li. Statistical tests of neutrality of mutations. *Genetics*, 133(3):693–709, Mar 1993.

[25] T. Gernhard. The conditioned reconstructed process. *J. Theo. Biol.*, 253(4):769–778, 2008.

[26] T. Gernhard. New analytic results for speciation times in neutral models. *Bull. Math. Biol.*, 70(4):1082–1097, 2008.

[27] T. Gernhard, D. Ford, R. Vos, and M. Steel. Estimating the relative order of speciation or coalescence events on a given phylogeny. *Evolutionary Bioinformatics Online*, 2:309–317, 2006.

[28] T. Gernhard, K. Hartmann, and M. Steel. Stochastic properties of generalised yule models, with biodiversity applications. *J. Math. Biol.*, 57:713–735, 2008.

[29] J. Gillespie. The molecular clock may be an episodic clock. *PNAS*, 81:8009–8013, Dec 1984.

[30] M. Hahn, T. De Bie, J. Stajich, C. Nguyen, and N. Cristianini. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research*, 15(8):1153, 2005.

[31] E. F. Harding. The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Appl. Probability*, 3:44–77, 1971.

[32] L. Harmon, J. Schulte, A. Larson, and J. Losos. Tempo and mode of evolutionary radiation in iguanian lizards. *Science*, 301:961–964, Aug 2003.

[33] K. Hartmann. TreeSample. http://www.klaashartmann.com/treesample/, 2007.

[34] K. Hartmann, T. Stadler, and D. Wong. Sampling trees from evolutionary models. *Submitted*, 2008.

[35] P. H. Harvey, R. M. May, and S. Nee. Phylogenies without fossils. *Evolution*, 48:523–529, 1994.

[36] S. B. Heard. Patterns in phylogenetic tree balance with variable and evolving speciation rates. *Evolution*, 50(6):2141–2148, 1996.

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

[38] R. V. Hogg and A. Craig. *Introduction to Mathematical Statistics (5th Edition)*. Prentice Hall, December 1994.

[39] M. Hohl and M. A. Ragan. Is multiple-sequence alignment required for accurate inference of phylogeny? *Systematic Biology*, 56(2):206–221, 2007.

[40] J. Huelsenbeck, B. Larget, and D. Swofford. A compound poisson process for relaxing the molecular clock. *Genetics*, 154:1879–1892, Apr 2000.

[41] T. R. Jackman, A. Larson, K. de Queiroz, and J. B. Losos. Phylogenetic relationships and tempo of early diversication in anolis lizards. *Syst. Biol.*, 48(2):254–285, 1999.

[42] G. Jin, L. Nakhleh, S. Snir, and T. Tuller. Efficient parsimony-based methods for phylogenetic network reconstruction. *Bioinformatics*, 23(2):e123, 2007.

[43] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. *Mammalian Protein Metabolism*, pages 21–132, 1969.

[44] G. P. Karev, Y. I. Wolf, and E. V. Koonin. Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? *Bioinformatics*, 19:1889–1900, 2003.

[45] R. M. Karp. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

[46] D. G. Kendall. On some modes of population growth leading to r. a. fisher's logarithmic series distribution. *Biometrika*, 35(1/2):6–15, 1948.

[47] D. G. Kendall. On the generalized "birth-and-death" process. *Ann. Math. Statist.*, 19(1):1–15, 1948.

[48] D. G. Kendall. Stochastic processes and population growth. *J. Roy. Statist. Soc. Ser. B.*, 11:230–264, 1949.

[49] M. Kimura. A simple model for estimatin evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

[50] M. Kimura. Estimation of evolutionary distances between homologous nucleotide sequences. *PNAS*, 78:454–458, Jan 1981.

[51] J. F. C. Kingman. The coalescent. *Stochastic Processes and Their Applications*, 13:235–248, 1982.

[52] J. F. C. Kingman. Exchangeability and the evolution of large populations. *Exchangeability in Probability and Statistics*, pages 97–112, 1982.

[53] J. F. C. Kingman. On the genealogy of large populations. *J. Appl. Probab.*, 19A:27–43, 1982.

[54] C. Kuiken, K. Yusim, L. Boykin, and R. Richardson. The Los Alamos hepatitis C sequence database. *Bioinformatics*, 21(3):379–384, Feb 2005.

[55] C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20:86–93, 1984.

[56] N. N. Lebedew. *Spezielle Funktionen und ihre Anwendung.* B.I.-Wissenschaftsverlag, 1973.

[57] J. Losos and F. Adler. Stumped by trees— a generalized null model for patterns of organismal diversity. *Am. Nat.*, 145(3):329–342, 1995.

[58] L. Lu, T. Nakano, Y. He, Y. Fu, C. H. Hagedorn, and B. H. Robertson. Hepatitis C virus genotype distribution in China: predominance of closely related subtype 1b isolates and existence of new genotype 6 variants. *J. Med. Virol.*, 75(4):538–549, Apr 2005.

[59] N.-Y. Ma and F. Liu. A novel analytical scheme to compute the n-fold convolution of exponential-sum distribution functions. *Appl. Math. Comput.*, 158(1):225–235, 2004.

[60] W. Maddison. Estimating a Binary Character's Effect on Speciation and Extinction. *Systematic Biology*, 56(5):701–710, 2007.

[61] S. Magallón and M. Sanderson. Absolute diversification rates in agiosperm clades. *Evolution*, 55(9):1762–1780, 2001.

[62] V. Makarenkov, D. Kevorkov, and P. Legendre. Phylogenetic Network Construction Approaches. *Bioinformatics*, 2006.

[63] X.-L. Meng. Posterior predictive p-values. *Ann. Statist.*, 22(3):1142–1160, 1994.

[64] A. Mooers, L. J. Harmon, M. G. B. Blum, D. H. J. Wong, and S. Heard. Some models of phylogenetic tree shape. In O. Gascuel and M. Steel, editors, *Reconstructing Evolution: new mathematical and computational advances*, pages 149–170. Oxford University Press, Oxford, 2007.

[65] P. Moran. A General Theory of the Distribution of Gene Frequencies. I. Overlapping Generations. *Proceedings of the Royal Society of London. Series B, Biological Sciences (1934-1990)*, 149(934):102–112, 1958.

[66] C. Moreau, C. Bell, R. Vila, S. Archibald, and N. Pierce. Phylogeny of the ants: diversification in the age of angiosperms. *Science*, 312:101–104, Apr 2006.

[67] C. S. Moreau. Unraveling the evolutionary history of the hyperdiverse ant genus *Pheidole*. *submitted*, 2008.

[68] V. Moulton and M. Steel. Peeling phylogenetic 'oranges'. *Adv. Appl. Math.*, 33(4):710–727, 2004.

[69] S. Nee, E. C. Holmes, R. M. May, and P. H. Harvey. Extinction rates can be estimated from molecular phylogenies. *Philosophical Transactions: Biological Sciences*, 344(1307):77–82, 1994.

[70] S. C. Nee. Infering speciation rates from phylogenies. *Evolution*, 55(4):661–668, 2001.

[71] S. C. Nee, R. M. May, and P. Harvey. The reconstructed evolutionary process. *Philos. Trans. Roy. Soc. London Ser. B*, 344:305–311, 1994.

[72] T. Oakley, B. Ostman, and A. Wilson. Repression and loss of gene expression outpaces activation and gain in recently duplicated fly genes. *Proceedings of the National Academy of Sciences*, 103(31):11637, 2006.

[73] R. Opgen-Rhein, L. Fahrmeir, and K. Strimmer. Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. *BMC Evol. Biol.*, 5(1):6, 2005.

[74] L. Popovic. Asymptotic genealogy of a critical branching process. *Ann. Appl. Probab.*, 14(4):2120–2148, 2004.

[75] A. Purvis. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London, Series B*, 348:405–421, 1995.

[76] O. G. Pybus and P. H. Harvey. Testing macro-evolutionary models using incomplete molecular phylogenies. *Proc. R. Soc. London*, 267(1459):2267–2272, 2000.

[77] A. Rambaut. PhyloGen: Phylogenetic tree simulator package. *Department of Zoology, University of Oxford*, 2002.

[78] S. C. Ray, R. R. Arthur, A. Carella, J. Bukh, and D. L. Thomas. Genetic epidemiology of hepatitis C virus throughout Egypt. *J. Infect. Dis.*, 182(3):698–707, Sep 2000.

[79] R. Ree. Detecting the historical signature of key innovations using stochastic models of character evolution and cladogenesis. *Evolution*, 59:257–265, Feb 2005.

[80] R. Ricklefs. Global diversification rates of passerine birds. *Proceedings- Royal Society of London. Biological sciences*, 270(1530):2285–2291, 2003.

[81] J. S. Rogers. Central moments and probability distributions of three measures of phylogenetic tree imbalance trees, from Yule to today. *Syst. Biol.*, 45(1):99–110, 1996.

[82] D. Rosen. Vicariant patterns and historical explanation in biogeography. *Systematic Zoology*, 27(20):159–188, 1978.

[83] K. Rosen. *Handbook of Discrete and Combinatorial Mathematics*. CRC Press, 2000.

[84] M. Sanderson. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19:301–302, Jan 2003.

[85] C. Semple and M. Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.

[86] A. Shaw, C. Cox, B. Goffinet, W. Buck, and S. Boles. Phylogenetic evidence of a rapid radiation of pleurocarpous mosses (bryophyta). *Evolution*, 57(10):2226–2241, 2003.

[87] J. Slowinski. Probabilities on n-trees under two models: a demonstration that asymmetrical interior nodes are not improbable. *Systematic Zoology*, 39:89–94, 1990.

[88] J. B. Slowinski. Molecular polytomies. *Mol. Phylogenet. Evol.*, 19(1):114–120, 2001.

[89] T. Stadler. Cass. http://www.tb.ethz.ch/people/tstadler, 2006–2008.

[90] T. Stadler. Lineages-through-time plots of neutral models for speciation. *Mathematical Biosciences*, 216:163–171, 2008.

[91] M. J. Stanhope, V. G. Waddell, O. Madsen, W. de Jong, S. B. Hedges, G. C. Cleven, D. Kao, and M. S. Springer. Molecular evidence for multiple origins of insectivora and for a new order of endemic african insectivore mammals. *Proc. Natl. Acad. Sci. USA*, 95:9967–9972, 1998.

[92] M. Steel and A. McKenzie. Properties of phylogenetic trees generated by Yule-type speciation models. *Math. Biosci.*, 170(1):91–112, 2001.

[93] H. Stöcker. *Taschenbuch mathematischer Formeln und moderner Verfahren*. Verlag Harri Deutsch, Frankfurt am Main, 2003.

[94] F. Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, Nov 1989.

[95] E. A. Thompson. *Human evolutionary trees*. Cambridge University Press, 1975.

[96] S. Trajanovski, C. Albrecht, R. Schultheiß, T. Gernhard, M. Benke, and T. Wilke. Testing the temporal framework of speciation in an ancient lake species flock: the genus dina (hirudinea: Erpobdellidae) in lake ohrid. *submitted to Hydrobiologia - Special Issue*, 2008.

[97] C. Venditti, A. Meade, and M. Pagel. Detecting the node-density artifact in phylogeny reconstruction. *Systematic Biology*, 55(4):637–643, 2006.

[98] R. A. Vos. A new dated supertree of the primates. *PhD thesis*, 2006.

[99] J. T. Weir. Divergent timing and patterns of species accumulation in lowland and highland neotropical birds. *Evolution*, 60(4):842–855, 2006.

[100] Z. Yang and B. Rannala. Bayesian phylogenetic inference using DNA sequences: A markov chain monte carlo method. *Mol. Biol. Evol.*, 17(7):717–724, 1997.

[101] 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.

[102] D. J. Zwickl and D. M. Hillis. Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology*, 51(4):588–598, 2002.