

New analytic results for speciation times in neutral models

Tanja Gernhard

Department of Mathematics, Kombinatorische Geometrie (M9), TU München

Bolzmannstr. 3, 85747 Garching, Germany

Phone +49 89 289 16882, Fax +49 89 289 16859

gernhard@ma.tum.de

Keywords: phylogenetics, Yule model, critical branching process, reconstructed process.

Abstract

In this paper we investigate the standard Yule model, and a recently studied model of speciation and extinction, the ‘critical branching process’. We develop an analytic way— as opposed to the common simulation approach— for calculating the speciation times in a reconstructed phylogenetic tree. Simple expressions for the density and the moments of the speciation times are obtained.

Methods for dating a speciation event become valuable, if for the reconstructed phylogenetic trees, no time scale is available. A missing time scale could be due to supertree methods, morphological data or molecular data which violates the molecular clock. Our analytic approach is in particular useful for the model with extinction, since simulations of birth-death processes which are conditioned on obtaining n extant species today are quite delicate. Further, simulations are very time consuming for big n under both models.

1 Introduction

In phylogenetics, a major task is to reconstruct the evolutionary history for some extant species. With the reconstructed trees, the aim is to understand evolution better. A central question is if the species evolved under a neutral model. A neutral model assumes that throughout time, whenever a speciation or extinction event occurs, each species is equally likely to be the one undergoing that event. That is, if a phylogeny evolved under a neutral model, all species behave alike. Of course speciation is not always just random – some lineages can speciate faster than others due to differing selective pressures and environmental factors. However, for rejecting a neutral model for some data set, we need to know properties of the neutral models

(which are not present in the data set). 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 properties of the neutral model. In this paper, we will calculate the speciation times under a neutral model in a given tree shape.

The most popular neutral model is the so-called Yule model [2, 6, 25] which G.U. Yule introduced in 1924. Under the Yule model, no extinction occurs and each species has an exponential (rate λ) lifetime. 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 [4, 18, 23], even though extinction clearly occurred in the considered phylogenies.

Recently, a critical branching process as a neutral model for speciation was introduced [1, 17]. In the critical branching process, each species has an exponential (rate λ) lifetime during which it produces offspring according to a Poisson (rate λ) process. After a species lifetime, it goes extinct (without further offspring). We condition the process to have n extant - present day - species; this is the conditioned critical branching process (cCBP). Conditioning on n extant species is crucial for data analysis, since a tree which we obtain from data has a fixed number of species. A lineage tree is the smallest subtree of the cCBP containing the n extant species, see Figure 1. Note that in other work, a lineage tree is also called a reconstructed process [16].

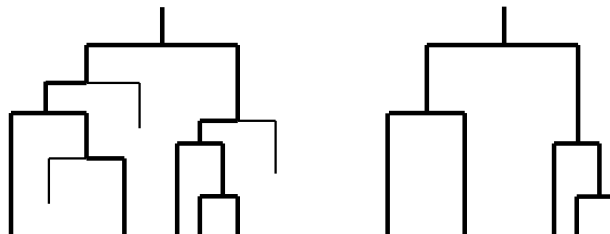


Figure 1: A complete tree (left) and its lineage tree (right)

Both the Yule model and the critical branching process are special cases of a birth-death process. In a birth-death process, we again start with one individual. The extant individuals act independently from each other, and each individual has a birth rate λ and a death rate μ . So the Yule model is a birth-death process with $\mu = 0$ and the critical branching process is a birth-death process with $\lambda = \mu$. Note that for all birth-death processes where $\mu < \lambda$, the number of species is increasing exponentially fast. Only for $\mu = \lambda$, we do not have an exponential increase, which is biological more reasonable [2].

In Nee et al. [16], lineage trees which are obtained from birth-death processes are discussed. The authors investigate the lineage tree after a time t . Our approach differs, since we investigate the lineage tree with n extant species. The joint prob-

ability for the shape of the lineage tree and the times of all speciation events has been considered in Rannala and Yang [19] for the general birth-death process and in Edwards [6] for the Yule model. In Yang and Rannala [24], the joint probability for the times of all speciation events in lineage trees has been considered – disregarding the shape. They all did not obtain the marginal probabilities for the times of speciation events given the tree shape.

We are interested in the time of the successive speciation events. For the two special cases of birth-death models, the Yule model and the cCBP, we calculate the time for the k -th speciation event, $k = 1, \dots, n - 1$ (Section 3 and 2). For the Yule model, the density for the time of any speciation event (without an order) is established in Nee [15] – conditioned that the time since the most recent common ancestor of the extant species is t ; for the cCBP, the density for the time of any speciation event is established in Popovic [17] – conditioned such that the time since the origin of the tree is t . However, often the time of origin is unknown. We will assume that the time of origin of the tree is distributed uniformly at random and only condition on obtaining n species today, as it was done in Aldous and Popovic [1]. Note that in this paper, we will first discuss the cCBP and then the Yule model. The reason is, that the method used for the cCBP in Section 2 could be used for a general birth death process (and in particular for the Yule model as well). However, since the Yule model is a pure birth process, we can use an alternative and easier method which only works for pure-birth processes (Section 3).

Knowing the time of the k -th speciation event, we can calculate the time of any interior node in a given tree (Section 4). This becomes very useful for supertrees. In supertree reconstruction, 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 [18] and the mammal phylogeny [4], 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 Purvis [18]. 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 [23].

In Vos [23], an undated speciation event is estimated via the expectation of the time of that vertex. The expectation is obtained via simulations. Our approach in Section 4 calculates the distribution and the first two moments analytically for the Yule model and the cCBP – in the cCBP we can even determine all moments. 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 [23].

The algorithms for calculating the time of the speciation events in a given tree were implemented in Python as part of our PhyloTree Package, which can be downloaded: <http://www-m9.ma.tum.de/homepages/gernhard/PhyloTree.zip>. In an earlier paper [9], we have already determined the expectation for the Yule model (in a more complicated way though). No expressions were given for the distribution and variance under the Yule model and any value under the cCBP model. We will

show in Section 4 a surprising connection between the cCBP and the coalescent – a popular null model in population genetics. The two models induce the same tree shapes and expected speciation times. Only the higher moments for the speciation times differ.

If the reconstructed tree has dates associated for internal nodes, we can use the expected dates to accept or reject a null model for the reconstructed phylogeny. A common approach is to compare the reconstructed dates to the lineage-through-time (LTT) plots under a specified model [16]. Conditioning on n extant species, we will provide an analytic way to obtain the expected LTT plots.

2 The cCBP model

The cCBP model has first been introduced in Popovic [17]. Let a random binary tree \mathcal{T} be generated according to the cCBP model. This stochastic process operates 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. In Popovic [17], time is scaled such that $\lambda = 1$, so a time unit represents the expected lifetime of a species.

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 time between two Poisson events is exponentially (rate λ) distributed, an individual is equally likely to die or to speciate. This model is a neutral model for speciation and extinction. At any point in time, each species is equally likely to die or speciate next. The described process with conditioning on having n extant individuals today is called the cCBP.

The asymptotic behavior of the cCBP for large n has been analyzed [1, 17]. The authors mention that the model is of biological significance mainly for small n . They calculate the distribution and expected time of the most recent common ancestor for the extant species. We extend this idea and calculate the distribution and all moments for each ancestor of the extant species.

2.1 Basic properties about the cCBP model

For our calculations in Section 2.2, we need the following properties of the cCBP model, which have been established in Aldous and Popovic [1], Popovic [17].

Write $\mathcal{T}_{n,t}$ for a tree that has evolved under the cCBP model conditioned to have n species today, the value $t > 0$ is the time of origin and today is time 0 (so the time parameter decreases from origin until now). However, the time of origin, t , is often not known. We assume a uniform prior on $(0, \infty)$ for the time of origin of a tree as has been done in Aldous and Popovic [1]. Note that the 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 Berger [3]. Define the random variable t_n , the ‘time of origin in a tree with n extant species’. The density function $q_n(\tau)$ of the random variable t_n , assuming the uniform prior for the time of origin t , is [1, 20]

$$q_n(\tau) = \frac{n\tau^{n-1}}{(1+\tau)^{n+1}}. \quad (1)$$

Let \mathcal{T}_n denote the random tree generated under the cCBP model conditioned to have n extant species. The time of origin of \mathcal{T}_n has density $q_n(\tau)$ assuming the uniform prior. As stated above, $\mathcal{T}_{n,t}$ is the random tree generated under the cCBP model conditioned to have n extant species and age t . The trees $\mathcal{T}_{n,t}$ and \mathcal{T}_n are called complete trees. The lineage tree $\mathcal{A}_{n,t}$ is the smallest subtree of $\mathcal{T}_{n,t}$ which contains all divergence times for the extant species. The lineage tree \mathcal{A}_n is defined analogously using \mathcal{T}_n .

Generally, a binary tree on n extant species can be described by $n - 1$ points as described in the following. Draw the tree from the top to the bottom. At each speciation event, choose one branch to be on the right and one on the left. On a horizontal axis, the leaves are located at $1, 2, \dots, n$. The speciation events are at the location $(j + 1/2, s_j)$, $j = 1, 2, \dots, (n - 1)$; $s_j > 0$. In Popovic [17], it is shown that the times s_j of the speciation events in a lineage tree under the cCBP model are independent and identically distributed with density function

$$f_t(s) = \begin{cases} (1 + 1/t)(1 + s)^{-2} & \text{if } 0 < s < t, \\ 0 & \text{else,} \end{cases} \quad (2)$$

where $s > 0$. This yields the distribution function

$$F_t(s) = \begin{cases} \int_0^s f_t(u) du = \frac{t+1}{t} \frac{s}{1+s} & \text{if } 0 < s < t, \\ 1 & \text{else.} \end{cases} \quad (3)$$

The function $F_t(s)$ is the probability that a speciation event occurred between s and today.

Aldous and Popovic [1] calculate the density and expectation for the random variable ‘most recent common ancestor of the extant species’, \mathcal{T}_n^{mrca} ,

$$f_{\mathcal{T}_n^{mrca}}(s) = \frac{n(n-1)s^{n-2}}{(1+s)^{n+1}}, \quad (4)$$

$$\mathbb{E}[\mathcal{T}_n^{mrca}] = n - 1. \quad (5)$$

Apart from that, the authors focus on asymptotic results. As stated in Aldous and Popovic [1], the model is mainly valuable in biology for small n , so exact expressions are of interest when applying the cCBP model in biology.

2.2 Our calculations for the cCBP model

Let \mathcal{A}_n^k be the time to the k -th speciation event in the lineage tree \mathcal{A}_n , $k = 1, \dots, (n-1)$. Note that $k = 1$ is the most recent common ancestor. We will obtain simple expressions for the density, expectation and the higher moments of \mathcal{A}_n^k .

Theorem 2.1. *The density of \mathcal{A}_n^k is*

$$f_{\mathcal{A}_n^k}(s) = (k+1) \binom{n}{k+1} \frac{s^{n-k-1}}{(s+1)^{n+1}}. \quad (6)$$

Proof. Let $f_{\mathcal{A}_n^k}(s, t)$ be the density function and $F_{\mathcal{A}_n^k}(s, t)$ be the distribution function of $\mathcal{A}_{n,t}^k$. The $n-1$ speciation points are i.i.d., with the density function given in Equation (2). Therefore, $\mathcal{A}_{n,t}^k$ is the $(n-k)$ -th order statistic, its density and distribution is (see e.g. Dehling and Haupt [5], Theorem 9.17),

$$\begin{aligned} f_{\mathcal{A}_n^k}(s, t) &= (n-k) \binom{n-1}{n-k} F_t(s)^{n-k-1} (1-F_t(s))^{k-1} f_t(s), \\ F_{\mathcal{A}_n^k}(s, t) &= \sum_{i=0}^{k-1} \binom{n-1}{i} F_t(s)^{n-i-1} (1-F_t(s))^i. \end{aligned}$$

With our uniform prior, the time of origin has density function $q_n(t)$. The density of \mathcal{A}_n^k is therefore

$$f_{\mathcal{A}_n^k}(s) = \frac{d}{ds} F_{\mathcal{A}_n^k}(s) = \frac{d}{ds} \int_0^\infty F_{\mathcal{A}_n^k}(s, t) q_n(t) dt. \quad (7)$$

We may change the order of differentiation and integration (Lemma A.2). Therefore,

$$f_{\mathcal{A}_n^k}(s) = \int_0^\infty \frac{d}{ds} F_{\mathcal{A}_n^k}(s, t) q_n(t) dt = \int_s^\infty f_{\mathcal{A}_n^k}(s, t) q_n(t) dt$$

With Equation (14) from Lemma A.2, we get

$$\begin{aligned} f_{\mathcal{A}_n^k}(s) &= nk \binom{n-1}{k} \frac{s^{n-k-1}}{(1+s)^n} \int_s^\infty \frac{(t-s)^{k-1}}{(1+t)^{k+1}} dt \\ &= nk \binom{n-1}{k} \frac{s^{n-k-1}}{(1+s)^n} \left[\frac{1}{k(1+s)} \left(\frac{t-s}{t+1} \right)^k \right]_s^\infty \\ &= (k+1) \binom{n}{k+1} \frac{s^{n-k-1}}{(1+s)^{n+1}} \end{aligned}$$

which establishes the theorem. □

Corollary 2.2. *The expectation and variance of \mathcal{A}_n^k are*

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

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

$$\mathbb{E}[(\mathcal{A}_n^k)^m] = \begin{cases} \frac{\binom{n-k+m-1}{m}}{\binom{k}{m}} & \text{if } k \geq m, \\ \infty & \text{else.} \end{cases} \quad (8)$$

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

Proof. For calculating the moments, we need the following result which can be found in Lebedew [13],

$$\int_0^\infty \frac{s^a}{(1+s)^b} ds = \begin{cases} \frac{1}{(b-a-1)\binom{b-1}{a}} & \text{if } b > a + 1, \\ \infty & \text{else,} \end{cases} \quad (9)$$

where $a, b \in \mathbb{N}_0$. With Theorem 2.1, 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 (9) and therefore $\mathbb{E}[(\mathcal{A}_n^k)^m] = \infty$. For $k \geq m$, we obtain with Equation (9),

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

which completes the proof of Corollary 2.2. \square

Corollary 2.3. *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 Corollary 2.2, we obtain for $k \geq m$

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

Multiplying those expectations yields to the formula for $\mathbb{E}[(\mathcal{A}_n^k)^m]$ in Corollary 2.2. \square

3 The Yule model

Under the Yule model [25], each species has an exponential (rate λ) lifetime; each extant species is equally likely to speciate next. We will set $\lambda = 1$ in the following.

Again, we want to calculate the density and the moments for the time of the k -th speciation event in a tree on n species.

For obtaining the density of the time of the k -th speciation event, \mathcal{A}_n^k , under the Yule model, we could take the same approach as for the cCBP. For a tree on n extant species which evolved under the Yule model, the $n - 1$ speciation events are i.i.d., with the density and distribution function [15]

$$f_t(s) = \frac{e^{-s}}{1 - e^{-t}}, \quad F_t(s) = \frac{1 - e^{-s}}{1 - e^{-t}}.$$

Under the Yule model, the density function of \mathcal{A}_n^k can therefore be established with the same approach as under the cCBP. However, it is difficult to obtain a simple expression for the moments from the density function analytically. Therefore, we chose an alternative way for obtaining the density function, which yields on its way to nice expressions for the first and second moment without integrating the density function. This approach can be applied whenever we have to add exponentially distributed random variables.

Let X_k be the random variable ‘time between $(k - 1)$ -st speciation event and k -th speciation event’ under the Yule model – without conditioning on any value, in particular without conditioning on obtaining n species today.

We assume an uniform prior for the time of origin of a 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 have n species today.

This is equivalent to growing a tree and waiting until the tree has n species. After the $(k - 1)$ -th speciation event, we always have an exponential (rate k) waiting time. 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 X_k . After the $(k - 1)$ -th speciation event, we have k extant species with independent exponential (rate 1) distributed lifetimes S_1, S_2, \dots, S_k . The random variable X_k is distributed as follows,

$$\mathbb{P}[X_k \geq t] = \mathbb{P}[S_1, S_2, \dots, S_k \geq t] = \prod_{j=1}^k \mathbb{P}[S_j \geq t] = e^{-kt}.$$

The density function $f_{X_k}(t)$ is therefore

$$f_{X_k}(t) = \frac{d}{dt}(1 - \mathbb{P}[X_k \geq t]) = ke^{-kt}$$

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

$$\mathbb{E}[X_k] = \frac{1}{k}, \quad \text{Var}[X_k] = \frac{1}{k^2}.$$

Let \mathcal{A}_n^k be – as under the cCBP – the random variable ‘time of k -th speciation event’. Time is again 0 today and increases going back to the past. Therefore, $\mathcal{A}_n^k = \sum_{i=k+1}^n X_i$ and

$$\mathbb{E}[\mathcal{A}_n^k] = \mathbb{E}\left[\sum_{i=k+1}^n X_i\right] = \sum_{i=k+1}^n \frac{1}{i}. \quad (10)$$

For the variance, we get, since the X_i are independent,

$$\text{Var}[\mathcal{A}_n^k] = \text{Var}\left[\sum_{i=k+1}^n X_i\right] = \sum_{i=k+1}^n \frac{1}{i^2}.$$

For the second moment, we get

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

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

Calculating density functions under the Yule model

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 Ma and Liu [14], a formula for the convolution of n exponential distributed random variables is established. Note that $Y_i^j := \sum_{k=i}^j X_k$ is a convolution of $j - i + 1$ exponential distributed random variables. From the general formula for the convolution in Ma and Liu [14], we obtain for our setting

$$f_{Y_i^j}(s) = i \cdot (i + 1) \cdot \dots \cdot j e^{-js} \varphi_{j-i+1}(s) \quad (11)$$

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 of $f_{Y_i^j}(s)$ in our setting.

Lemma 3.1. *With the notation above, we have*

$$\varphi_n(s) = \frac{1}{(n-1)!} (e^s - 1)^{n-1}. \quad (12)$$

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. *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 (11) and (12) we obtain

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

□

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

Theorem 3.3. *The density of \mathcal{A}_n^k is*

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

4 Applications

4.1 Calculating the distribution of the time of a speciation event

So far, we calculated the time of the k -th speciation event. However, in a given phylogenetic tree, we are interested in the time of any interior vertex. The expected time of a vertex can be used as an estimate for the time of an undated speciation event in a supertree. In Vos [23], the author uses the expectation for estimating the undated vertices in the primate supertree. The expectation is obtained via simulations. We will now show how to calculate the distribution and the first two moments analytically for the Yule model and the cCBP – in the cCBP we will even determine all moments. The expected time for a vertex under the Yule model is calculated in Gernhard et al. [9], however the distribution and variance are not calculated there, nor any values for the cCBP.

For calculating the time of a speciation event in a phylogenetic tree, we need the concept of a rank function r [21]. A rank function on a phylogenetic tree is a bijection from the set of interior vertices \mathring{V} to $\{1, 2, \dots, |\mathring{V}|\}$ with the property that the ranks are increasing on any path from the root to a leaf. We call a phylogenetic tree with a rank function a ranked phylogenetic tree. In the following, we want to calculate $p_u := (\mathbb{P}[r(u) = i])_{i=1, \dots, n-1}$ where $\mathbb{P}[r(u) = i]$ is the probability that vertex u has rank i – assuming that each rank function is equally likely. In Gernhard [8], a formula for calculating p_u is given. Label the vertices on the path from the vertex u to the root ρ with $u = x_1, x_2, \dots, x_n = \rho$. Define λ_j as the number of leaves below x_j minus 1. Further, recall that the 1-norm $|\cdot|_1$ is defined for a vector $x = (x_1, \dots, x_n)$ as $|x|_1 = \sum_{i=1}^n |x_i|$. We obtain from Gernhard [8] that

$$p_u = \frac{M_{n-1} M_{n-2} \dots M_1 e_1}{|M_{n-1} M_{n-2} \dots M_1 e_1|_1} \quad (13)$$

where $e_1 = (1, 0, 0, \dots, 0)^T$ and the matrix M_k 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 Gernhard [8] calculates p_u according to Equation (13).

A neutral model always induces a uniform distribution on the ranked phylogenetic lineage trees on n species [2]. Therefore, we can calculate $\mathbb{P}[r(u) = i]$ for a tree which evolved under the cCBP model or under the Yule model with the algorithm RANKPROB.

Let v be a interior node of a phylogenetic tree and let \mathcal{A}_v be the random variable ‘time of the speciation event v ’. The distribution, expectation, higher moments and variance of \mathcal{A}_v are obviously

$$\begin{aligned} f_{\mathcal{A}_v}(s) &= \sum_{i=1}^{n-1} f_{\mathcal{A}_v|r(v)=i}(s) \mathbb{P}[r(v) = i] = \sum_{i=1}^{n-1} f_{\mathcal{A}_n^i}(s) \mathbb{P}[r(v) = i], \\ \mathbb{E}[(\mathcal{A}_v)^m] &= \sum_{i=1}^{n-1} \mathbb{E}[(\mathcal{A}_v)^m | r(v) = i] \mathbb{P}[r(v) = i] = \sum_{i=1}^{n-1} \mathbb{E}[(\mathcal{A}_n^i)^m] \mathbb{P}[r(v) = i], \\ \text{Var}[\mathcal{A}_v] &= \mathbb{E}[(\mathcal{A}_v)^2] - (\mathbb{E}[\mathcal{A}_v])^2. \end{aligned}$$

For an edge $e = (u, v)$, the length of an edge is the time between speciation event u and speciation event v . For the random variable \mathcal{A}_e , ‘length of edge e ’, we have the expected edge length

$$\mathbb{E}[\mathcal{A}_e] = \mathbb{E}[\mathcal{A}_v - \mathcal{A}_u] = \mathbb{E}[\mathcal{A}_v] - \mathbb{E}[\mathcal{A}_u] = \sum_{i=1}^{n-1} \mathbb{E}[\mathcal{A}_n^i] (\mathbb{P}[r(v) = i] - \mathbb{P}[r(u) = i]).$$

For the Yule model, the expected edge length had already been established in Gernhard [8].

Comparing the neutral models

As mentioned above, all neutral models induce the same distribution on tree shapes. In Equation (10) we established $\mathbb{E}_{Yule}[\mathcal{A}_n^i] = \sum_{k=i+1}^n \frac{1}{k}$. Since

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

with γ being the Euler constant, we have

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

as $n \rightarrow \infty$ and therefore, for fixed i , $\sum_{k=i+1}^n \frac{1}{k} = \ln n + O(1)$. Asymptotically, this is $\sum_{k=i+1}^n \frac{1}{k} \sim \ln n$. From Corollary 2.2, we have $\mathbb{E}_{cCBP}[\mathcal{A}_n^i] = \frac{n-i}{i} \sim \frac{n}{i}$. So

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

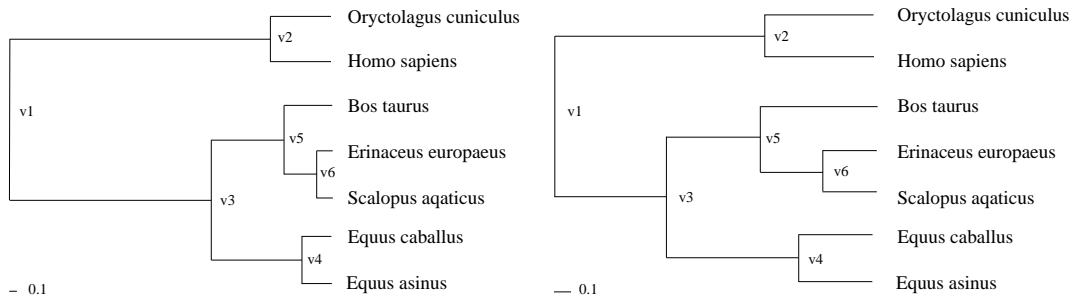


Figure 2: Given the tree shape, we obtained the displayed expected edge lengths under the cCBP (left) and Yule model (right).

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

Remark 4.1. We will show a surprising connection between the cCBP and the coalescent – a popular neutral model in population genetics. Under the coalescent setting [10, 11, 12] the random variable \mathcal{A}_n^k , ‘waiting time between the k -th and the $(k - 1)$ -th coalescent event’ is exponential ($\lambda \binom{k}{2}$) distributed where λ is the rate of the coalescent. We set $\lambda = 1$ for the calculations. For the first and second moment, Kingman [10, 11, 12] established

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

So in expectation, the coalescent with rate 1 is equivalent to the cCBP with rate $\lambda = \frac{2}{n}$. The ranked trees under the coalescent are distributed uniformly at random [2] – so the distribution is the same as under the cCBP or the Yule model. 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

$$\mathbb{E}_{Coal}[(\mathcal{A}_n^i)^2] = \sum_{k=i+1}^n \frac{1}{\binom{k}{2}^2} + \left(\frac{2(n-i)}{ni} \right)^2,$$

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

Calculating the distribution and the moments for \mathcal{A}_v (under the cCBP and the Yule model) with the expressions given above has been implemented in Python as part of our PhyloTree Package, which can be downloaded: <http://www-m9.ma.tum.de/homepages/gernhard/PhyloTree.zip>.

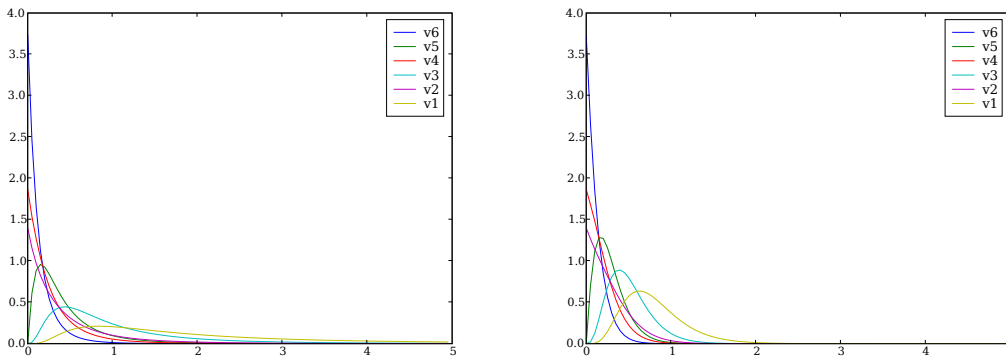


Figure 3: Plot of the density functions for the time of each interior vertex under the cCBP (left) and Yule model (right). Since $f_{\mathcal{A}_n^k}(0) = 0$ for $k < n-1$, but $f_{\mathcal{A}_n^{n-1}}(0) \neq 0$ for $k = n-1$, for the vertices v with $\mathbb{P}[r(v) = n-1] > 0$, we obtain $f_{\mathcal{A}_v}(0) \neq 0$.

To illustrate the method, consider the following example of a tree found in Figure 2, available on TreeBase [22]. The reconstructed tree has no time scale. 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 node the density function for the time of speciation under the cCBP and under the Yule model, see Figure 3. Further, we calculated the expected speciation times, see Figure 2. 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.

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

4.2 Lineage-through-time plots

If the reconstructed tree has a time scale, we can compare the speciation time in the reconstructed tree with the estimated time under the Yule model and under the cCBP. This should indicate whether much extinction occurred during the evolution of that particular phylogeny.

Lineage-through-time (LTT) plots are frequently used for that purpose [16]. In a LTT plot, the time is plotted vs. the number of species. We can either consider the LTT plot for the complete tree or for the lineage tree. In the following, we

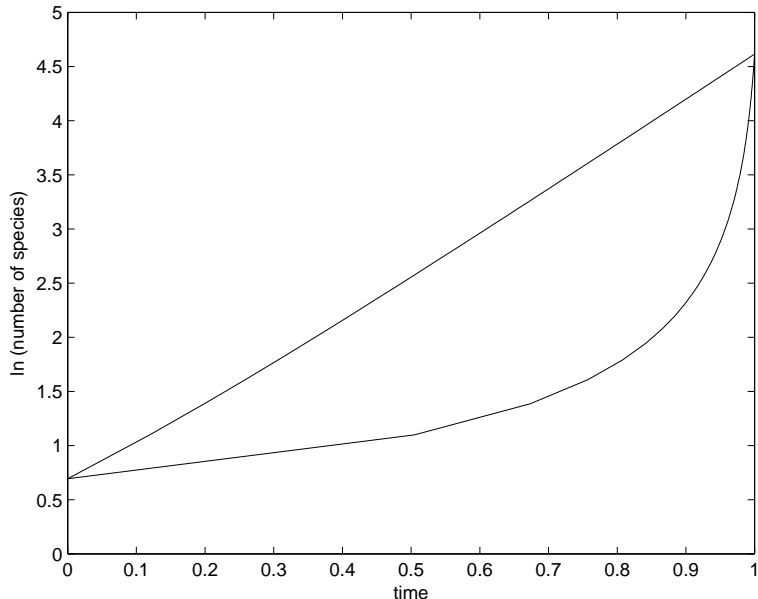


Figure 4: Expected lineage-through-time plot for a lineage tree on $n = 100$ extant species under the Yule model (upper line) and the cCBP model (lower line).

only consider the lineage tree. We condition on having n extant species today. The number of species is displayed on a logarithmic scale. Time is scaled such that the most recent ancestor of the extant species speciates at time 0 and today is time 1. Between the time of the k th speciation event and the time of the $k + 1$ st speciation event, we interpolate with a straight line.

It is of interest to look at the LTT plots for different neutral models in order to compare them with the data [16]. Obtaining the LTT plots is commonly done 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 analytical approach is of special interest.

For the lineage tree, the expected time for having k species is $\mathbb{E}[\mathcal{A}_n^k]$, therefore plotting $\mathbb{E}[\mathcal{A}_n^k]$ vs. k with the appropriate scaling yields the expected LTT plot, see Figure 4. This can be done analytically for the Yule model and the cCBP with the results obtained above.

5 Results and Outlook

We obtained analytic results for the distribution and moments of the time of speciation events in a reconstructed phylogeny under neutral models. The existing methods for estimating the undated vertices in a supertree via the expected time – which relied on simulations – can be done analytically now. In particular, we can calculate the expected speciation time for a model which includes extinction. This is quite

useful, since simulating trees with n extant species today is quite tricky – n extant species can reoccur again and again, so one has to simulate forever.

It is of interest to test if the real times of speciation are in accordance with the cCBP or the Yule model. With the provided algorithms, such a test could be done at a broad scale.

The presented methods can also be applied for calculating analytically the expected loss in biodiversity when some species become extinct. However this is not discussed further here. Another issue worth following up is deriving the time of a speciation event under a general birth-death process which is conditioned to have n extant species.

6 Acknowledgements

The author thanks Mike Steel and Anusch Taraz for very helpful discussions, Dennis Wong for suggesting to consider lineage-through-time plots and the two anonymous reviewers for very helpful comments. Financial support by the Deutsche Forschungsgemeinschaft through the graduate program “Angewandte Algorithmische Mathematik” at the Munich University of Technology and by the Allan Wilson Center through a summer studentship is gratefully acknowledged.

References

- [1] D. Aldous and L. Popovic. A critical branching process model for biodiversity. *Adv. in Appl. Probab.*, 37(4):1094–1115, 2005. ISSN 0001-8678.
- [2] D. J. Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statist. Sci.*, 16(1):23–34, 2001. ISSN 0883-4237.
- [3] J. O. Berger. *Statistical decision theory: foundations, concepts, and methods*. Springer-Verlag, New York, 1980. ISBN 0-387-90471-9. Springer Series in Statistics.
- [4] 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.
- [5] H. Dehling and B. Haupt. *Einfuehrung in die Wahrscheinlichkeitstheorie und Statistik*. Springer, 2003.
- [6] 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. ISSN 0035-9246.
- [7] O. Forster. *Analysis 3*. Friedr. Vieweg & Sohn, 1981.
- [8] T. Gernhard. Stochastic models of speciation events in phylogenetic trees. *Diplom thesis, Technical University of Munich*, 2006.

- [9] 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.
- [10] J. F. C. Kingman. On the genealogy of large populations. *J. Appl. Probab.*, 19A:27–43, 1982.
- [11] J. F. C. Kingman. The coalescent. *Stochastic Processes and Their Applications*, 13:235–248, 1982.
- [12] J. F. C. Kingman. Exchangeability and the evolution of large populations. *Exchangeability in Probability and Statistics*, pages 97–112, 1982.
- [13] N. N. Lebedew. *Spezielle Funktionen und ihre Anwendung*. B.I.-Wissenschaftsverlag, 1973.
- [14] 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. ISSN 0096-3003.
- [15] S. C. Nee. Inferring speciation rates from phylogenies. *Evolution*, 55(4):661–668, 2001.
- [16] S. C. Nee, R. M. May, and P. Harvey. The reconstructed evolutionary process. *Philos. Trans. Roy. Soc. London Ser. B*, 344:305–311, 1994.
- [17] L. Popovic. Asymptotic genealogy of a critical branching process. *Ann. Appl. Probab.*, 14(4):2120–2148, 2004. ISSN 1050-5164.
- [18] A. Purvis. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London, Series B*, 348:405–421, 1995.
- [19] B. Rannala and Z. Yang. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.*, 43:304–311, 1996.
- [20] W. Reed and B. Hughes. On the size distribution of live genera. *Journal of Theoretical Biology*, 213(1):125–135, 2002.
- [21] C. Semple and M. Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003. ISBN 0-19-850942-1.
- [22] 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.
- [23] R. A. Vos. A new dated supertree of the primates. *PhD thesis*, 2006.
- [24] 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.

- [25] 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.

A Appendix

For obtaining the density function of \mathcal{A}_n^k , we need the following theorem which can be found in Forster [7], Theorem 11.2.

Theorem A.1. *Let $\rho(s, t)$ be a function on $\mathbb{R}^2 \rightarrow \mathbb{R}$ with the following properties. For fixed s , $\rho(s, t)$ is integrable w.r.t. t . For fixed t , $\frac{\partial}{\partial s}\rho(s, t)$ exists. Further, there exists a integrable function $\phi(t)$ on $\mathbb{R} \rightarrow \mathbb{R}_+ \cup \{\infty\}$ with $|\frac{\partial}{\partial s}\rho(s, t)| \leq \phi(t)$ for all s, t . Then*

$$\frac{d}{ds} \int_0^\infty \rho(s, t) dt = \int_0^\infty \frac{\partial}{\partial s} \rho(s, t) dt.$$

Lemma A.2. *Let $F_{\mathcal{A}_n^k}(s, t)$ be the distribution function of $\mathcal{A}_{n,t}^k$ and let $q_n(t)$ be the function defined in Equation (1). Then,*

$$\frac{d}{ds} \int_0^\infty F_{\mathcal{A}_n^k}(s, t) q_n(t) dt = \int_0^\infty \frac{\partial}{\partial s} F_{\mathcal{A}_n^k}(s, t) q_n(t) dt.$$

Proof. We will show that all requirements from Theorem A.1 are fulfilled. Let $f_{\mathcal{A}_n^k}(s, t)$ be the density function of \mathcal{A}_n^k . Define the function $\rho(s, t) := F_{\mathcal{A}_n^k}(s, t) q_n(t)$. Note that $F_{\mathcal{A}_n^k}(s, t)$ is continuous since $F_t(s)$ is continuous (in s and t). Further $q_n(t)$ is continuous, therefore $\rho(s, t)$ is continuous. Thus $\rho(s, t)$ is integrable w.r.t. t .

The function $\rho(s, t)$ is differentiable w.r.t. s , $\frac{\partial}{\partial s}\rho(s, t) = f_{\mathcal{A}_n^k}(s, t) q_n(t) \geq 0$. For $s > t$, we have $\frac{\partial}{\partial s}\rho(s, t) = 0$ and for $s \leq t$,

$$\begin{aligned} 0 &\leq \frac{\partial}{\partial s} \rho(s, t) = f_{\mathcal{A}_n^k}(s, t) q_n(t) \\ &= (n-k) \binom{n-1}{n-k} \left(\frac{t+1}{t} \frac{s}{1+s} \right)^{n-k-1} \times \\ &\quad \left(1 - \frac{t+1}{t} \frac{s}{1+s} \right)^{k-1} (1+1/t)(1+s)^{-2} \frac{nt^{n-1}}{(1+t)^{n+1}} \\ &= nk \binom{n-1}{k} \frac{s^{n-k-1} (t-s)^{k-1}}{(1+s)^n (1+t)^{k+1}} \\ &\leq nk \binom{n-1}{k} \frac{(1+t)^{k-1}}{(1+t)^{k+1}} := \phi(t) \end{aligned} \tag{14}$$

Since $\phi(t)$ is continuous, it is integrable. So all requirements for Theorem A.1 are fulfilled, and we may change the order of differentiation and integration which establishes the lemma. \square