

# The conditioned reconstructed process

Tanja Gernhard

Department of Mathematics, Kombinatorische Geometrie (M9), TU München  
Boltzmannstr. 3, 85747 Garching, Germany

Phone +49 89 289 16882, Fax +49 89 289 16859, gernhard@ma.tum.de

## Abstract

We investigate a neutral model for speciation and extinction, the constant rate birth-death process. The process is conditioned to have  $n$  extant species today, we look at the tree distribution of the reconstructed trees— i.e. the trees without the extinct species. Whereas the tree shape distribution is well-known and actually the same as under the pure birth process, no analytic results for the speciation times were known. We provide the distribution for the speciation times and calculate the expectations analytically. This characterizes the reconstructed trees completely. We will show how the results can be used to date phylogenies.

**Keywords:** Phylogenetics, macroevolutionary models, birth-death process, reconstructed process.

## 1 Introduction

Phylogenetics is the science of reconstructing the evolutionary history of lineages (usually species). Besides providing data for systematics and for taxonomy, phylogenies are the pattern of past diversification and so can be analysed to infer past macroevolutionary process. The first common step is to compare the reconstructed trees with expectations from neutral models of diversification (Gould et al., 1977; Mooers and Heard, 1997; Nee et al., 1992; Raup et al., 1973). The simplest class of neutral model are entirely homogeneous, and assume that throughout time, whenever a speciation (or extinction) event occurs, each species is equally likely to be the one undergoing that event. Of course speciation is not just random – lineages will differ in their expected diversification rates for both intrinsic and extrinsic factors (Mooers et al., 2007). However, a neutral model is often used as a null model to analyze the data, with departures pointing the way to more sophisticated scenarios (Harvey et al., 1994).

We investigate the constant rate birth-death process (Feller, 1968; Kendall, 1948) as it is probably the most popular homogeneous model. A birth-death process is a stochastic process which starts with an initial species. A species gives birth to a new species after exponential (rate  $\lambda$ ) waiting times and dies after an exponential (rate

$\mu$ ) waiting time. Throughout this paper, we will have  $0 \leq \mu \leq \lambda$ . In the following, time 0 is today and  $t_{or}$  the origin of the tree, so time is increasing going into the past. Special cases of the birth-death process are the Yule model (Yule, 1924) where  $\mu = 0$  and the critical branching process (Aldous and Popovic, 2005; Popovic, 2004) where  $\mu = \lambda$ . When looking at phylogenies, we have a given number, say  $n$ , of extant taxa. We therefore condition the process to have  $n$  species today, we call that process the conditioned birth-death process (cBDP). The age of the tree, i.e. the time since origin of the birth-death process is  $t_{or}$ ; 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 Aldous and Popovic (2005); Popovic (2004). Note that a tree of age  $t_{or}$ , which evolved under a birth-death process includes extinct species, it is called the complete tree (Aldous and Popovic, 2005), see Figure 1, left. From the complete tree, delete the extinct lineages and suppress degree-two vertices. This is called the reconstructed tree shape, see Figure 1, right. Label its leaves uniformly at random (since each species evolves in the same way). The resulting tree is called the reconstructed tree (this follows the notation in Nee et al. (1994); it is also called a lineage tree (Aldous and Popovic, 2005)). Note that when reconstructing a phylogeny from (molecular) data, we see the reconstructed tree. Extinct lineages are only apparent when the fossil record is included.

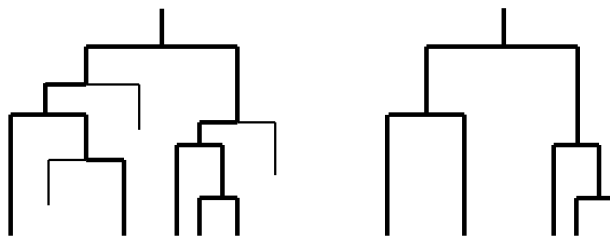


Figure 1: A complete tree (left) and its reconstructed tree shape (right).

In Nee et al. (1994), the reconstructed tree of a birth-death process after time  $t$  is discussed. In this paper we additionally condition on having  $n$  extant species, since this allows us to compare the model with phylogenies on  $n$  extant species. We will obtain the probability for each speciation event in a tree with  $n$  species. This has been done for the Yule model and the conditioned critical branching process (cCBP) in Gernhard (2008) (note that the conditioned critical branching process is the critical branching process conditioned on  $n$  extant species). For the general birth-death process, the joint probability for all speciation times and the shape as well as conditioning on the shape has been established in Thompson (1975); However, no individual probabilities have been established.

For establishing the individual probabilities, we introduce the point process representation for reconstructed trees (Section 2). This had been done for the critical branching process in Aldous and Popovic (2005); Popovic (2004). In Section 3, we calculate the probability distribution of the age of a given tree on  $n$  species, assuming a uniform prior on  $(0, \infty)$  for the age of the tree. This enables us to derive the density function for the time of the  $k$ -th speciation event in a tree with  $n$  extant

species (Section 4) and its expectation (Section 5) – assuming a uniform prior or conditioning on the age of the tree. In Section 6, we discuss some further properties of reconstructed trees. We will determine the point process when not conditioning the cBDP on the time of origin. Also, we describe the point process of the coalescent, the neutral model in population genetics. Further, we will discuss the backwards process of reconstructed trees. The backward process is the process of the coalescence of the extant species.

Knowing the time of the  $k$ -th speciation event in a reconstructed tree with  $n$  species allows us to calculate the time of a given vertex in the reconstructed tree (Gernhard et al., 2006; Gernhard, 2008). This becomes useful for dating phylogenies. If we are able to reconstruct the phylogeny of extant species, but do not obtain speciation times, we can use the expected time of a speciation event as an estimate for the speciation time. This estimate has been used for the undated vertices in the primate phylogeny (Vos, 2006), assuming the Yule model. Simulations were used for obtaining the expected speciation times. We provide analytic results assuming any constant rate birth-death model. The methods are implemented in python as part of our PhyloTree package and can be downloaded at <http://www-m9.ma.tum.de/twiki/pub/Allgemeines/TanjaGernhard/PhyloTree.zip>.

## 1.1 Basic definitions

Formally, a *reconstructed tree* (Nee et al., 1994) is a rooted, binary tree with unique leaf labels and ultrametric edge lengths assigned, i.e. the distance from any leaf to the root is the same, see Figure 2, left tree. We denote the set of interior vertices by  $\mathring{V}$ . A *ranked reconstructed tree* is a reconstructed tree without edge lengths but with a rank function defined on the interior vertices. A rank function (Semple and Steel, 2003) is a bijection from  $\mathring{V} \rightarrow \{1, 2, \dots, |\mathring{V}|\}$  where the ranks are increasing on any path from the root to the leaves. Note that a ranked reconstructed tree is also called a ranked phylogenetic tree in the literature (Semple and Steel, 2003). A rank function induces an order on  $\mathring{V}$  which can be interpreted as the order of speciation events. A *(ranked) reconstructed tree shape* is a (ranked) reconstructed tree without leaf labels. A *(ranked) oriented tree* is a (ranked) reconstructed tree without leaf labels but where we distinguish between the two daughter edges of the interior vertices, w.l.o.g. label them  $l$  and  $r$ , see Figure 2, middle tree. Note that a (ranked) oriented tree has  $n!$  possible labelings. We introduce the oriented tree to make the proofs clearer and the statements easier.

**Remark 1.1.** The cBDP induces a (ranked) reconstructed tree in the following way. Consider the complete tree which evolved under the cBDP. We delete the extinct lineages and label the  $n$  leaves uniformly at random with  $\{1, 2, \dots, n\}$  to obtain the reconstructed tree (there are  $n!2^{-k}$  possible labelings, where  $k$  is the number of cherries in the reconstructed tree shape).

The interior vertices shall be ordered according to the time of speciation, this defines the rank function. To make the reconstructed tree oriented, for each interior vertex, we label the two daughter lineages with  $l$  and  $r$  uniformly at random, there are  $2^{n-1}$  possibilities. We then ignore the leaf labelings (note that each labeling of

the oriented tree is equally likely, since each labeling of the reconstructed tree was equally likely).

On the other hand, if we know the distribution on (ranked) oriented trees induced by the cBDP, we obtain the distribution on (ranked) reconstructed trees in the following way. We choose a labeling of the leaves with  $\{1, 2, \dots, n\}$  uniformly at random from the  $n!$  possible labelings. We then ignore the orientation. This gives us back the distribution on (ranked) reconstructed tree. Therefore, it is sufficient to determine the distribution on (ranked) oriented trees in order to determine the distribution on (ranked) reconstructed trees. Overall, let  $\tau_r$  be a reconstructed tree, and let  $\tau_o$  be a oriented tree which was induced by  $\tau_r$ . Then  $\mathbb{P}[\tau_r] = \mathbb{P}[\tau_o]2^{n-1}/n!$ , since an oriented tree has  $n!$  possible labelings and for the  $n - 1$  interior vertices, we have the distinction between the  $l$  and  $r$  daughter branches.

## 2 The point process

In this section, we provide the density for the time of a speciation event in the reconstructed tree given  $n$  species today and the time of origin being at time  $t_{or}$  in the past. We do that using a point process representation. The following point process has first been considered in connection with trees in Aldous and Popovic (2005); Popovic (2004).

**Definition 2.1.** A point process for  $n$  points and of age  $t_{or}$  is defined as follows. Draw the  $n$  points on the horizontal axis at  $1, 2, \dots, n$ . Now pick  $n - 1$  points to be at the location  $(i + 1/2, s_i)$ ,  $i = 1, 2, \dots, (n - 1)$ ;  $0 < s_i < t_{or}$ .

**Lemma 2.2.** *We have a bijection between oriented trees of age  $t_{or}$  and the point process of age  $t_{or}$ .*

*Proof.* Draw the given oriented tree from the top to the bottom. At each speciation event, choose the branch with label  $r$  to be on the right and with label  $l$  on the left. On a horizontal axis, the leaves are located at position  $1, 2, \dots, n$ . The speciation events are at the location  $(i + 1/2, s_i)$ ,  $i = 1, 2, \dots, (n - 1)$ ;  $0 < s_i < t_{or}$ . These are the  $n - 1$  points of the point process. The mapping to the point process is obviously injective and surjective, i.e. bijective.  $\square$

For completeness, we give the mapping from the point process to the oriented trees. Consider a realization of the point process. Connect the most recent speciation event with the two neighboring leaves. This 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 oriented tree. An example of the point process is given in Figure 2.

The following theorem has been proven for the Yule process in Edwards (1970). Thompson (1975) derives the theorem for the constant rate birth-death process from results in Harding (1971). Since the theorem will be crucial for further results in the paper, we give a direct proof.

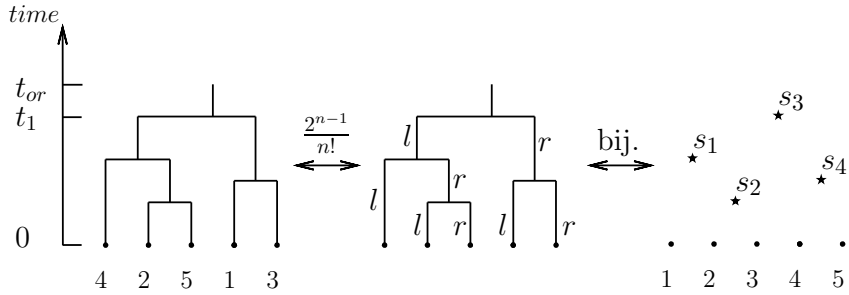


Figure 2: Reconstructed tree, oriented tree and the corresponding point process. The time of origin of the process is  $t_{or}$ , the time of the most recent common ancestor is  $t_1$ .

**Theorem 2.3.** *Each ranked oriented tree on  $n$  leaves induced by the constant rate birth and death process has equal probability. Note that this is true with or without conditioning on the time since origin.*

*Proof.* We first determine the probability of a ranked oriented tree with leaf labels. Consider the process backwards. We have  $n$  species today. We pick two species uniformly at random, which coalesce first. The one new branch gets label  $l$  and the other new branch gets label  $r$ . Overall there are  $n(n-1)$  possible choices for the first coalescent event, each one being equally likely. The two chosen leaves are replaced with their common ancestor. We proceed in this way until all species are connected. There are  $n!(n-1)!$  possible scenarios for the coalescent, each one being equally likely, i.e. a ranked oriented tree with leaf labels has probability  $\frac{1}{n!(n-1)!}$ . Each of the  $n!$  possible labelings are equally likely, therefore the probability of a ranked oriented tree is  $\frac{1}{(n-1)!}$ . This is the uniform distribution on ranked oriented trees of size  $n$ .  $\square$

**Corollary 2.4.** *Each permutation of the  $n-1$  speciation points  $s_1, \dots, s_{n-1}$  in the point process of the birth-death process has equal probability.*

*Proof.* We have a bijection between the oriented trees and the point process (Lemma 2.2). Choosing the  $n-1$  speciation points  $s_1, \dots, s_{n-1}$  induces a ranked oriented tree, let the probability of that tree be  $p$ . Now permute the  $n-1$  speciation points arbitrary. This induces a different ranked oriented tree. Since we have a uniform distribution on ranked oriented trees (Theorem 2.3), the probability of the new tree is again  $p$ . So each permutation is equally likely.  $\square$

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 Kendall (1949), we know

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

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

In Thompson (1975), page 56, the density  $g$  of the ordered speciation times,  $x_2 > x_3 > \dots > x_{n-1}$ , given  $n$  and  $x_1 = t_1$  is derived,

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

This joint density is used in Yang and Rannala (1997) 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  $x_2, x_3, \dots, x_n$  are the order statistic of say  $s_2, s_3, \dots, s_{n-1}$ . Each permutation of the  $n-2$  random variables  $s_2, s_3, \dots, s_{n-1}$  has equal probability (Corollary 2.4), and therefore the density  $f$  of the speciation times is,

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

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

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

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 (2) w.r.t.  $s_i$ ,

$$F(s_i | t_1 = t) = \frac{1 - e^{-(\lambda-\mu)s_i}}{\lambda - \mu e^{-(\lambda-\mu)s_i}} \frac{\lambda - \mu e^{-(\lambda-\mu)t}}{1 - e^{-(\lambda-\mu)t}}. \quad (3)$$

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— maybe first seeming surprisingly— property that

$$f(s_i | t_1 = t) = f(s_i | t_{or} = t). \quad (4)$$

The following argument verifies Equation (4). 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

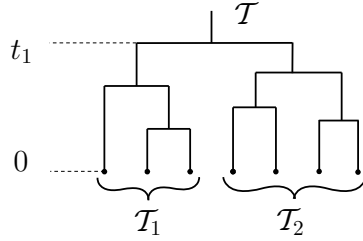


Figure 3: Reconstructed tree  $\mathcal{T}$  with daughter trees  $\mathcal{T}_1, \mathcal{T}_2$ . We have  $mrca(\mathcal{T}) = origin(\mathcal{T}_1) = origin(\mathcal{T}_2) = t_1$ .

species. The speciation times in  $\mathcal{T}_n, \mathcal{T}_m$  occurred according to Equation (2). 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. With Remark 1.1, this establishes the following theorem.

**Theorem 2.5.** *The speciation times  $s_1, \dots, s_{n-1}$  in a oriented tree (reconstructed tree) with  $n$  species conditioned on the age of the tree are i.i.d. The speciation times  $s_2, \dots, s_{n-1}$  in a oriented tree (reconstructed 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,*

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

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

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.

## 2.1 Special models

### 2.1.1 The Yule model

For the special case of a pure birth process, i.e.  $\mu = 0$ , which is the Yule model, Equation (2) 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 Nee (2001)– he conditioned on the time since the  $mrca$  though.

### 2.1.2 The conditioned critical branching process

In a cCBP, we have  $\lambda = \mu$ . As  $\mu \rightarrow \lambda$ , we get in the limit using Equation (2), (3) and (4), and the property  $e^{-\epsilon} \sim 1 - \epsilon$  for  $\epsilon \rightarrow 0$ ,

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

This has already been established in a different way for  $\lambda = 1$  in Aldous and Popovic (2005); Popovic (2004).

## 3 The time of origin

Suppose nothing is known about  $t$ , the time of origin of a tree. As in Aldous and Popovic (2005); Popovic (2004), we then assume a uniform prior on  $(0, \infty)$ , i.e. a tree is equally likely to origin at any point in time. Note that the prior does not integrate to 1. For any constant function, the integral is  $\infty$ . Therefore the prior is not a density. Such a prior is called improper; a discussion and justification is found e.g. in Berger (1980). Assuming the uniform prior, we will establish the density for  $t$  given  $n$  extant species. From Equation (1), we have the probability of  $n$  extant species given the time of origin is  $t$ ,

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

In order to derive the density for  $t$  given  $n$ , we need the following lemma.

**Lemma 3.1.** *Let  $\mathbb{P}_{or}[n|t]$  be the probability that a tree has  $n$  extant species given the time of origin  $t$ . We have*

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

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

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

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 3.2.** *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*

$$q_{or}(t|n) = n\lambda^n(\lambda - \mu)^2 \frac{(1 - e^{-(\lambda-\mu)t})^{n-1} e^{-(\lambda-\mu)t}}{(\lambda - \mu e^{-(\lambda-\mu)t})^{n+1}}. \quad (5)$$

*Proof.* With Bayes' law, we have

$$\begin{aligned} q_{or}(t|n) &= \frac{\mathbb{P}_{or}[n|t]q_{or}(t)}{\mathbb{P}_{or}[n]} = \frac{\mathbb{P}_{or}[n|t]q_{or}(t)}{\int_0^\infty \mathbb{P}_{or}[n, t]dt} \\ &= \frac{\mathbb{P}_{or}[n|t]q_{or}(t)}{\int_0^\infty \mathbb{P}_{or}[n|t]q_{or}(t)dt} = \frac{\mathbb{P}_{or}[n|t]}{\int_0^\infty \mathbb{P}_{or}[n|t]dt} \\ &\stackrel{(3.1)}{=} \lambda n \mathbb{P}_{or}[n|t] \\ &= n\lambda^n(\lambda - \mu)^2 \frac{(1 - e^{-(\lambda-\mu)t})^{n-1} e^{-(\lambda-\mu)t}}{(\lambda - \mu e^{-(\lambda-\mu)t})^{n+1}}. \end{aligned}$$

□

**Corollary 3.3.** *The distribution for the time of origin given  $n$  species today is*

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

*Proof.* We have  $\lim_{t \rightarrow \infty} Q_{or}(t|n) = 1$ . Differentiation of  $Q_{or}(t|n)$  w.r.t.  $t$  yields  $\frac{d}{dt}Q_{or}(t|n) = n\lambda^n(\lambda - \mu)^2 \frac{(1 - e^{-(\lambda-\mu)t})^{n-1} e^{-(\lambda-\mu)t}}{(\lambda - \mu e^{-(\lambda-\mu)t})^{n+1}} = q_{or}(t|n)$  which completes the proof. □

## 4 The time of speciation events

In this section, we calculate the density for the time of the  $k$ -th speciation event given we have  $n$  species today. Knowing that the distribution on ranked reconstructed trees is uniform (Theorem 2.3), this characterizes the reconstructed trees completely. These results allow us to calculate the density for the time of a given vertex in a reconstructed tree (Gernhard et al., 2006; Gernhard, 2008).

### 4.1 Known age of the tree

Let  $\mathcal{A}_{n,t}^k$  be the time of the  $k$ -th speciation event in a reconstructed tree  $\mathcal{A}$  with  $n$  extant species and age  $t$ . The  $n - 1$  speciation events in  $\mathcal{A}$  are i.i.d. and have the density function  $f(s|t)$ , see Theorem 2.5. The density of  $\mathcal{A}_{n,t}^k$  is therefore the  $(n - k)$ -th order statistic, which is (see e.g. Dehling and Haupt (2003), Theorem 9.17),

$$f_{\mathcal{A}_{n,t}^k}(s) = (n - k) \binom{n - 1}{n - k} F(s|t)^{n-k-1} (1 - F(s|t))^{k-1} f(s|t) \quad (6)$$

for  $s \leq t$  and  $f_{\mathcal{A}_{n,t}^k}(s) = 0$  else. The distribution function of  $\mathcal{A}_{n,t}^k$  is

$$F_{\mathcal{A}_{n,t}^k}(s) = \sum_{i=0}^{k-1} \binom{n-1}{i} F(s|t)^{n-i-1} (1 - F(s|t))^i \quad (7)$$

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

## 4.2 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 4.1.** *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  $f_{\mathcal{A}_{n,t}^k}$  for the time of the  $k$ -th speciation event (Section 4.1). With our uniform prior, the time of origin has density function  $q_{or}(t|n)$ . The density  $f_{\mathcal{A}_n^k}(s)$  is therefore

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

which establishes the theorem.  $\square$

**Remark 4.2.** Under the Yule model, i.e. setting  $\mu = 0$  and  $\lambda$  arbitrary in Theorem 4.1, 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 Gernhard (2008) for  $\lambda = 1$  in a different way.

In the cCBP, the birth rate equals the death rate,  $\lambda = \mu$ . Taking the limit  $\mu \rightarrow \lambda$  in  $f_{\mathcal{A}_n^k}$ , we obtain from Theorem 4.1 using the property  $e^{-\epsilon} \sim 1 - \epsilon$ ,

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

which has been established in Gernhard (2008) for  $\lambda = 1$  in a direct way.

## 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 for dating trees uses the analytic results. Higher moments are calculated numerically.

### 5.1 Known age of the tree

**Theorem 5.1.** *The expectation of  $\mathcal{A}_{n,t}^k$  is, for  $0 < \mu < \lambda$ ,*

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

where

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

and

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

For  $\mu = 0$ , we have,

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

For  $\mu = \lambda$ , we have

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

The proof is found in the appendix.

## 5.2 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 and for all moments under the cCBP (with the setting  $\lambda = 1$ ) is given in Gernhard (2008),

$$\mathbb{E}^{Yule}[\mathcal{A}_n^k] = \sum_{i=k+1}^n \frac{1}{i}, \quad (8)$$

$$\mathbb{E}^{Yule}[(\mathcal{A}_n^k)^2] = \sum_{i=k+1}^n \frac{1}{i^2} + \sum_{i=k+1}^n \sum_{j=k+1}^n \frac{1}{ij}, \quad (9)$$

$$\mathbb{E}^{cCBP}[(\mathcal{A}_n^k)^m] = \begin{cases} \frac{\binom{n-k+m-1}{m}}{\binom{n}{m}} & \text{if } k \geq m, \\ \infty & \text{else.} \end{cases} \quad (10)$$

For general  $\lambda, \mu$ , we have the following analytic expression for the expectation (the proof is found in the appendix).

**Theorem 5.2.** For  $0 < \mu < \lambda$ , the moments of  $\mathcal{A}_n^k$  are ( $\rho := \mu/\lambda$ ),

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

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  $\rho$ . Different  $\lambda$  just scale time by  $1/\lambda$ .

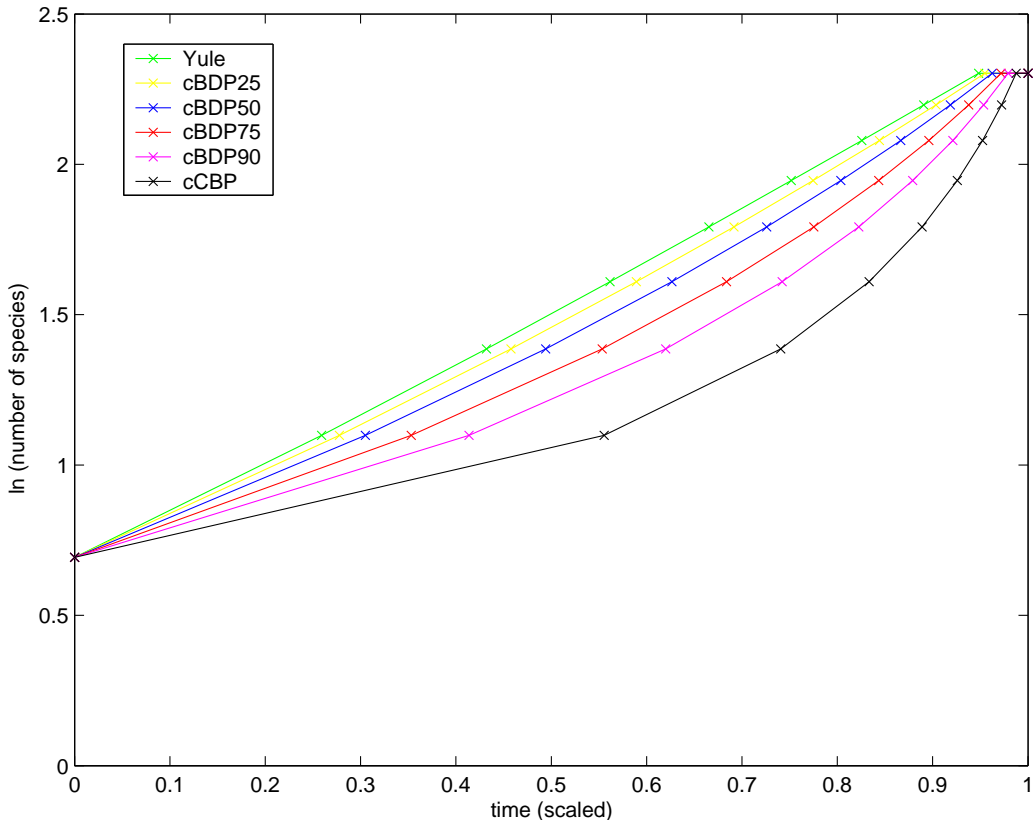


Figure 4: Lineage through time plot for  $n = 10$  species. Time is scaled such that the *mrca* is at 0 and today is 1. We have  $\rho = 0, 0.25, 0.5, 0.75, 0.9, 1$  from top to bottom. Note that for varying  $\lambda$ , time is scaled by  $1/\lambda$  (compared to  $\lambda = 1$ ). That means, since we scale time, the plots are the same for any  $\lambda$ .

Knowing the expected time of the  $k$ -th speciation event allows us to draw a lineage-through-time (LTT) plot (Nee et al., 1994) analytically. In a LTT plot, the time vs. the number of species at that time (on a logarithmic scale) is drawn. These plots are frequently used for comparing the data with a model. Commonly, the LTT plots for different models are obtained via simulations. Since we know the expected time of the speciation events in our model analytically, we can plot the LTT plot analytically, see Figure 4.

## 6 Properties of the speciation times

### 6.1 More on the point process of cBDP

In Section 2, we showed that a reconstructed tree of a cBDP 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 2.5 we obtain the density function for  $x = (x_1, \dots, x_{n-1})$ , the order statistic of the speciation times, conditioned on the time of origin,  $t$ ,

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

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

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

If the  $n-1$  speciation points would be 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 2.4), 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 2.5 (without using Theorem 2.5, we could also show the independence by calculating the density of  $x$  conditioning on  $n, x_1$  as  $f(x|n, x_1) = \frac{f(x|n)}{f_{\mathcal{A}_n^1}(x_1)} = \frac{f(x|n)}{f_{\mathcal{A}_n^1}(x_1)}$ ). For  $\mu = 0$ , i.e. for the Yule model, we can establish the same result,

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

i.e. the  $s_i$  are not independent. Conditioning on  $x_1$ , we again have independent points, as stated in Theorem 2.5.

**Remark 6.1.** 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,

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

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

## 6.2 The point process of the coalescent

The coalescent is the standard neutral model for population genetics. 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 coalescent is distributed exponential (rate  $\binom{n}{2}\lambda$ ) where  $\lambda$  is the rate of coalescent. We will show that this process – even though it is very similar to the Yule process – does not have a point process representation with i.i.d. coalescent points.

Let  $x = (x_1, x_2, \dots, x_{n-1})$  be the order statistic of the coalescent times (with  $x_1 > x_2 > \dots > x_{n-1}$ ). Note that  $x_i - x_{i+1}$  is distributed exponential 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 would be 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  $i$  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 shape is equally likely (see Aldous (2001) or argue as in Theorem 2.3), therefore each permutation of the  $s_i$  has the same probability. That means that the  $s_i$  are identical distributed– but not independent.

## 6.3 Backwards process of a cBDP

In the birth-death process, extant 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 describe the backward process, i.e. determine the waiting time until the extant species coalesce in the reconstructed tree.

**Theorem 6.2.** *Under the conditioned birth death process, a pair of species coalesces according to density function  $f(s|t)$  from Theorem 2.5.*

*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 2.5. The time  $s_i$  has the distribution with density function  $f(s|t)$  from Theorem 2.5.  $\square$

In a reconstructed tree with  $n$  species, the time to the last speciation event, i.e. the time between the  $(n-1)$ -st speciation event and today is  $\mathcal{A}_{n,t}^{n-1}$ ,  $\mathcal{A}_n^{n-1}$ . The time between the  $k$ -th and the  $(k+1)$ -st 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  $\tau$  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  $\tau$  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  $\tau$  is  $\binom{n-3}{k-1}(1-F(\tau))^{k-1}$ . The probability that the remaining  $n-k-2$  speciation points happen after  $\tau-s$  is  $F(\tau-s)^{n-k-2}$ . Overall,

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

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  $\tau$  and  $\tau-s$ ,

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

Note that  $\mathcal{A}_{n,t}^{k-1} - \mathcal{A}_{n,t}^k$  is the time until a coalescent event for  $k$  species in the reconstructed tree. We found analytic solutions for the above integrals under the Yule model (see next Section). For  $\mu \neq 0$ , the densities can be derived with numerical integration in the PhyloTree package. If assuming a uniform prior for the time of origin, we additionally need to integrate the above densities over  $t$ , weighted by  $q_{or}(t|n)$ .

## 6.4 Backwards process of the Yule model

### 6.4.1 Known tree age

For the Yule model, we can calculate the time between any two speciation events in the reconstructed tree analytically. The time between the  $k$ -th speciation event and the  $l$ -th speciation event,  $l > k$ , given the time between the  $n$ -th speciation event (today) and the origin of the tree is  $t$  has been calculated in Gernhard et al. (2007) for  $\lambda = 1$  which yields for general  $\lambda$  to

$$f_{\mathcal{A}_{n,t}^k - \mathcal{A}_{n,t}^l}(s) = \lambda \sum_{i=0}^{k-1} \sum_{j=0}^{n-l-1} B_{i,j} e^{\lambda(n-l)s} \frac{(e^{\lambda s} - 1)^{l-k-1}}{(e^{\lambda t} - 1)^{n-1}} (e^{\lambda(n-k+i)(t-s)} - e^{\lambda j(t-s)})$$

with  $B_{i,j} = k(k+1) \binom{l}{k+1} \binom{n-1}{l} \binom{k-1}{i} \binom{n-l-1}{j} \frac{(-1)^{n+k-l-i-j}}{n-k+i-j}$ . Note that  $\mathcal{A}_{n,t}^{k-1} - \mathcal{A}_{n,t}^k$  is the time until a coalescent event of  $k$  extant species.

Analogous results are not straightforward to obtain in a process with extinction. However, the expectation can be calculated straightforward for the cBDP,  $\mathbb{E}[\mathcal{A}_{n,t}^k - \mathcal{A}_{n,t}^l] = \mathbb{E}[\mathcal{A}_{n,t}^k] - \mathbb{E}[\mathcal{A}_{n,t}^l]$ .

### 6.4.2 Unknown tree age

We assume a 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 see  $n$  species today.

For the pure birth process, this is equivalent to start growing a tree and wait until the tree has  $n$  species. After the  $(k-1)$ -th speciation event, we always have an exponential (rate  $k$ ) waiting time. Therefore, also the coalescent has an exponential (rate  $k$ ) waiting time. This is not true in a process with extinction.

It remains to consider the time between the  $(n-1)$ -st speciation event and today. Note that today is not defined as the  $n$ -th speciation event, but whenever we look at the process. The density for the time between the  $(n-1)$ -st speciation event and today is

$$f_{\mathcal{A}_n^{n-1}}(s) = n\lambda e^{-\lambda ns}$$

which is the exponential (rate  $n$ ) distribution. Therefore, looking at the tree today is equivalent to looking at the tree at the time of the  $n$ -th speciation event. This observation has been discussed in Hartmann et al. (2008).

In Gernhard et al. (2006), the time between the  $k$ -th and the  $l$ -th speciation event,  $l > k$ , is established for  $\lambda = 1$  which yields for general  $\lambda$  to

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

For the general birth-death process, the waiting times cannot be calculated straightforward, since the time to the  $n$ -th speciation event differs from the time until today (note that we might have the  $n$ -th speciation event before today – i.e. the  $n$ -th speciation event is followed by extinction). However, obtaining the expectation for  $\mathcal{A}_n^{k,l}$  is straightforward,  $\mathbb{E}[\mathcal{A}_n^l - \mathcal{A}_n^k] = \mathbb{E}[\mathcal{A}_n^l] - \mathbb{E}[\mathcal{A}_n^k]$ .

## 7 Applications

Knowing the density and expectation of the  $k$ -th speciation time given we have  $n$  species today, we can obtain the density and expectation for the time of each interior node of a given tree. This can be used for dating phylogenies, if only the shape is inferred— missing dates in phylogenies could be due to supertree methods, morphological data or absence of a molecular clock. In earlier work (Gernhard et al., 2006; Gernhard, 2008), we gave the method and computer programs for dating phylogenetic trees. So far though, we only knew the speciation times for the Yule model and the cCBP. With the results in this paper, we can date a phylogeny

assuming any constant rate birth-death model. The methods are implemented in our PhyloTree package for python.

The point process representation is useful for simulating reconstructed trees on  $n$  taxa. If we have no extinction, we can simulate until obtaining  $n$  species. More precise, we stop at the  $n + 1$ -st speciation event, since that is the same as stopping today (Section 6.4.2). However, with extinction, simulations are tricky – we could return to  $n$  species again and again. With the point process, it is easy to sample trees on  $n$  species. First sample a time of origin according to the density  $q_{or}(t|n)$  in Equation (5). Then sample  $n - 1$  speciation times according to the density  $f(s|t)$  in Theorem 2.5.

If we want to simulate reconstructed trees on  $n$  taxa and age  $t$ , direct simulation of the process seems almost impossible – we simulate until  $t$  and only keep the realization if we see  $n$  species. We will throw away a lot of realizations (always if we do not see  $n$  species), therefore the time amount until we have a reasonable size of samples is huge. With the point process, we simply sample  $n - 1$  speciation times according to the density  $f(s|t)$  in Theorem 2.5. These sampling methods and more general sampling methods will be discussed in detail in Hartmann et al. (2008).

## 8 Results and Outlook

The  $n - 1$  speciation points in the point process representation are i.i.d. if conditioning on the time of origin or the most recent common ancestor. As discussed, this allows us to calculate the speciation times in a reconstructed phylogeny. So far, we calculate the speciation time with only conditioning on the shape of the phylogeny. With the point process, one might be able to condition on the shape as well as on some known dates in the phylogeny. This would be valuable for dating supertrees, since some speciation times are usually known.

In the application section, we showed that simulations of reconstructed trees become easy using the point process. This becomes useful for comparing the model with the data on aspects where no analytical results are known.

## 9 Acknowledgements

The author thanks Mike Steel, Arne Mooers, Daniel Ford, Dirk Metzler, Anusch Taraz and the anonymous reviewer for very helpful comments and discussions. 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

Aldous, D. and Popovic, L. (2005). A critical branching process model for biodiversity. *Adv. in Appl. Probab.*, 37(4):1094–1115.

- Aldous, D. J. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statist. Sci.*, 16(1):23–34.
- Berger, J. O. (1980). *Statistical decision theory: foundations, concepts, and methods*. Springer-Verlag, New York. Springer Series in Statistics.
- Dehling, H. and Haupt, B. (2003). *Einfuehrung in die Wahrscheinlichkeitstheorie und Statistik*. Springer.
- Edwards, A. W. F. (1970). Estimation of the branch points of a branching diffusion process. (With discussion.). *J. Roy. Statist. Soc. Ser. B*, 32:155–174.
- Feller, W. (1968). *An introduction to probability theory and its applications. Vol. I*. Third edition. John Wiley & Sons Inc., New York.
- Gernhard, T. (2008). New analytic results for speciation times in neutral models. *To appear in Bull. Math. Biol.*
- Gernhard, T., Ford, D., Vos, R., and Steel, M. (2006). Estimating the relative order of speciation or coalescence events on a given phylogeny. *Evolutionary Bioinformatics Online*, 2:309–317.
- Gernhard, T., Hartmann, K., and Steel, M. (2007). Stochastic properties of generalised yule models, with biodiversity applications. *Submitted*.
- Gould, S. J., Raup, D. M., Sepkowski, J. J., Schopf, T. J. M., and Simberloff, D. S. (1977). The shape of evolution: a comparison of real and random clades. *Paleobiology*, 3:23–40.
- Harding, E. F. (1971). The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Appl. Probability*, 3:44–77.
- Hartmann, K., Gernhard, T., and Wong, D. (2008). Sampling trees from evolutionary models. *Submitted*.
- Harvey, P. H., May, R. M., and Nee, S. (1994). Phylogenies without fossils. *Evolution*, 48:523–529.
- Kendall, D. G. (1948). On the generalized “birth-and-death” process. *Ann. Math. Statist.*, 19(1):1–15.
- Kendall, D. G. (1949). Stochastic processes and population growth. *J. Roy. Statist. Soc. Ser. B.*, 11:230–264.
- Mooers, A. O., Harmon, L. J., Blum, M. G. B., Wong, D. H. J., and Heard, S. B. (2007). Some models of phylogenetic tree shape. *Reconstructing Evolution: new mathematical and computational advances*, pages 149–170.
- Mooers, A. O. and Heard, S. B. (1997). Inferring evolutionary process from phylogenetic tree shape. *The quarterly review of Biology*, 72(1):31–54.

- Nee, S. C. (2001). Inferring speciation rates from phylogenies. *Evolution*, 55(4):661–668.
- Nee, S. C., May, R. M., and Harvey, P. (1994). The reconstructed evolutionary process. *Philos. Trans. Roy. Soc. London Ser. B*, 344:305–311.
- Nee, S. C., Mooers, A. O., and Harvey, P. (1992). Tempo and mode of evolution revealed from molecular phylogenies. *Proc. Natl. Acad. Sci. USA*, 89:8322–8326.
- Popovic, L. (2004). Asymptotic genealogy of a critical branching process. *Ann. Appl. Probab.*, 14(4):2120–2148.
- Raup, D. M., Gould, S. J., Schopf, T. J. M., and Simberloff, D. S. (1973). Stochastic models of phylogeny and the evolution of diversity. *Journal of Geology*, 81:449–452.
- Semple, C. and Steel, M. (2003). *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford.
- Thompson, E. A. (1975). *Human evolutionary trees*. Cambridge University Press.
- Vos, R. A. (2006). A new dated supertree of the primates. *PhD thesis*.
- Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: A markov chain monte carlo method. *Mol. Biol. Evol.*, 17(7):717–724.
- Yule, G. U. (1924). A mathematical theory of evolution: based on the conclusions of Dr. J.C. Willis. *Philos. Trans. Roy. Soc. London Ser. B*, 213:21–87.

## A Proofs

*Proof of Theorem 5.1.* Under the Yule model, i.e.  $\mu = 0$ , the expectation of  $\mathcal{A}_{n,t}^k$  has been calculated in Gernhard et al. (2007) for  $\lambda = 1$ ,

$$\mathbb{E}[\mathcal{A}_{n,t}^k(\lambda = 1)] = \sum_{i=0}^{n-k-1} \sum_{j=0}^{k-1} \frac{k \binom{n-1}{k} \binom{n-k-1}{i} \binom{k-1}{j} (-1)^{i+j}}{(k+i-j)^2} \times \\ (1 - e^{-t})^{1-n} (e^{-jt} - ((k+i-j)t + 1)e^{-(k+i)t}).$$

For general  $\lambda$ , since

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

we have

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

Substituting  $x = \lambda s$  yields

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

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

$$\begin{aligned}\mathbb{E}[\mathcal{A}_{n,t}^k] &= \int_0^t s f_{\mathcal{A}_{n,t}^k}(s) ds = [s F_{\mathcal{A}_{n,t}^k}(s)]_0^t - \int_0^t F_{\mathcal{A}_{n,t}^k}(s) ds \\ &= t - \sum_{i=0}^{k-1} \sum_{j=0}^i \binom{n-1}{i} \binom{i}{j} (-1)^{i+j} \int_0^t F(s|t)^{n-j-1} ds \quad (11) \\ &= t - \sum_{i=0}^{k-1} \sum_{j=0}^i \binom{n-1}{i} \binom{i}{j} (-1)^{i+j} \left( \frac{\lambda - \mu e^{-(\lambda-\mu)t}}{1 - e^{-(\lambda-\mu)t}} \right)^{n-j-1} \\ &\quad \int_0^t \left( \frac{1 - e^{-(\lambda-\mu)s}}{\lambda - \mu e^{-(\lambda-\mu)s}} \right)^{n-j-1} ds \\ &= t - \sum_{i=0}^{k-1} \sum_{j=0}^i \sum_{l=0}^{n-j-1} \binom{n-1}{i} \binom{i}{j} \binom{n-j-1}{l} (-1)^{i+j+l} \\ &\quad \left( \frac{\lambda - \mu e^{-(\lambda-\mu)t}}{1 - e^{-(\lambda-\mu)t}} \right)^{n-j-1} \int_0^t \frac{e^{-(\lambda-\mu)ls}}{(\lambda - \mu e^{-(\lambda-\mu)s})^{n-j-1}} ds.\end{aligned}$$

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

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

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

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

So overall,

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

For  $\mu = \lambda$ , we obtain from Equation (11),

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

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

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

Overall, this is

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

□

*Proof of Theorem 5.2.* For  $\mu = 0$  and for  $\mu = \lambda$ , the expectation is established with Remark 4.2 and Equations (8) and (10). For  $\mu \neq 0$  and  $\mu \neq \lambda$  we have with Theorem 4.1,

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

Set

$$C_1 := (k+1) \binom{n}{k+1} \lambda^{n-k} (\lambda - \mu)^{k+2},$$

$$f(s) := e^{-(\lambda-\mu)(k+1)s} \frac{(1 - e^{-(\lambda-\mu)s})^{n-k-1}}{(\lambda - \mu e^{-(\lambda-\mu)s})^{n+1}}.$$

Therefore,

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

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

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

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

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  $\lim_{s \rightarrow \infty} F(s)s = 0$  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

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

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.  $\square$