# Exact and approximate limit behaviour of the Yule tree's cophenetic index

Krzysztof Bartoszek

March 27, 2017

## Abstract

In this work we study the limit distribution of an appropriately normalized cophenetic index of the pure birth tree on $n$ contemporary tips. We show that this normalized phylogenetic balance index is a submartingale that converges almost surely and in $L^2$. We link our work with studies on trees without branch lengths and show that in this case the limit distribution is a contraction type distribution, similar to the Quicksort limit distribution. In the continuous branch case we suggest approximations to the limit distribution. We propose heuristic methods of simulating from these distributions and it may be observed that these algorithms result in reasonable tails. Therefore, we postulate using quantiles of the derived distributions for hypothesis testing, whether an observed phylogenetic tree is consistent with the pure birth process. Simulating a sample by the proposed heuristics is rapid while exact simulation (simulating the tree and then calculating the index) is a time–consuming procedure.

Keywords : Contraction type distribution; Cophenetic index; Martingales; Phylogenetics; Significance testing

# 1   Introduction

Phylogenetic trees are now a standard when analyzing groups of species. They are inferred from molecular sequences by algorithms that often assume a Markov chain for mutations of the individual entries of the genetic sequence. Given a phylogenetic tree it is often of interest to quantify the rate(s) of

1

speciation and extinction for the studied species. To do this one commonly assumes a birth–death process with constant rates. However, what we seem to be lacking are formal statistical tests whether a given tree comes from a given branching process model. For example, is a tree consistent with a pure birth tree, i.e. a Yule tree? The reason for the apparent lack of widespread use of such tests (but see Blum and François, 2005) could be the lack of a commonly agreed on test statistic. This is as a tree is a complex object and there are multiple ways in which to summarize it in a single number.

One proposed way of summarizing a tree is through indices that quantify how balanced it is, i.e. how close is it to a fully symmetric tree. Two such indices have been with us for many years now: Sackin's (Sackin, 1972) and Colless' (Colless, 1982). Recently a new one was proposed—the cophenetic index (Mir et al., 2013). The two former have already been studied and here we focus on the latter. This work is inspired by private communication with evolutionary biologist Gabriel Yedid (current affiliation Nanjing Agricultural University, Nanjing, China) who posed the question of how to use the cophenetic index for significance testing of whether a given tree is consistent with the pure birth process. He noticed that simulated distributions of the index have much heavier tails than those of the normal and t distributions and hence, comparing centred and scaled cophenetic indices with the usual Gaussian or t quantiles is not appropriate for significance testing. It would lead to a higher false positive rate—rejecting the null hypothesis of no extinction when a tree was generated by a pure birth process.

Our aim here is to propose an approach for working analytically with the cophenetic index, especially to improve hypothesis tests for phylogenetic trees. We show that there is a relationship between the cophenetic index and the Quicksort algorithm. This suggests that the methods exploring (e.g. Fill and Janson, 2000, 2001; Janson, 2015) the limiting distribution of the Quicksort algorithm can be an inspiration for studying analytical properties of the cophenetic index.

The paper is organized as follows. In Section 2 we formally define the cophenetic index, derive an associated with it submartingale, show that it converges almost surely and in $L^2$ (Thm. 2.5), propose an elegant representation (Thm. 2.10) and very good approximation (Def. 2.12). Then, in Section 3 we study the second order properties of this decomposition and conjecture a Central Limit Theorem (CLT, Rem. 3.12). Afterwards in Section 4, we link our work with previous studies which considered trees without branch lengths. In this discrete setting we show that the limit law of the normalized

cophenetic index is a contraction type distribution. Based on this we propose alternative approximations to the limit law of the normalized (with branch lengths) cophenetic index. In Section 5 we describe heuristic algorithms to simulate from these limit laws, show simulated quantiles and discuss the usefulness of the various proposed approaches. We end the paper with Section 6 by describing alternative representations of the cophenetic index.

# 2  The cophenetic index

Mir et al. (2013) recently proposed a new balance index for phylogenetic trees.

**Definition 2.1 (Mir et al. (2013))** *For a given phylogenetic tree on $n$ tips and for each pair of tips $(i, j)$ let $\tilde{\phi}_{ij}$ be the number of branches from the root to the most recent common ancestor of tips $i$ and $j$. We then the define the discrete* cophenetic index *as*

$$\tilde{\Phi}^{(n)} = \sum_{1 \leq i < j \leq n} \tilde{\phi}_{ij}^{(n)}.$$

Mir et al. (2013) show that this index has a better resolution than the "traditional" ones. In particular the cophenetic index has a range of values of the order of $O(n^3)$ while Colless' and Sackin's ranges have an order of $O(n^2)$. Furthermore, unlike the other two previously mentioned, $\tilde{\Phi}^{(n)}$ makes mathematical sense also for not fully resolve (i.e. not binary) trees.

In this work we study phylogenetic trees with branch lengths and hence consider a variation of the cophenetic index.

**Definition 2.2** *For a given phylogenetic tree on $n$ tips and for each pair of tips $(i, j)$ let $\phi_{ij}$ be the time from the most recent common ancestor of tips $i$ and $j$ to the root/origin (depending on the tree model) of the tree. We then the define the* continuous *cophenetic index as*

$$\Phi^{(n)} = \sum_{1 \leq i < j \leq n} \phi_{ij}^{(n)}.$$

**Remark 2.3** *In the original setting, when the distance between two nodes was measured by counting branches, Mir et al. (2013) did not consider the*

3

*edge leading to the root. In our work here, where our prime concern is with trees with random branch lengths, we include the branch leading to the root. This is not a big difference, one just has to remember to add to each distance between nodes the same exponential with rate 1 random variable (see next paragraph for description of the tree's growth).*

We study the asymptotic distributional properties of $\Phi^{(n)}$ for the pure birth tree model using techniques from our previous papers on branching Brownian and Ornstein–Uhlenbeck processes (Bartoszek and Sagitov, 2015b; Bartoszek, 2014; Bartoszek and Sagitov, 2015a; Sagitov and Bartoszek, 2012). We assume that the speciation rate of the tree is $\lambda = 1$. The key property we will use is that in the pure–birth tree case the time between two speciation events, $k$ and $k+1$ (the first speciation event is at the root), is exponentially distributed with rate $k$, as the minimum of $k$ rate 1 exponential random variables. We furthermore, assume that the tree starts with a single species (the origin) that lives for $\exp(1)$ time and then splits (the root of the tree) into two species. We consider a conditioned on $n$ contemporary species tree. This conditioning translates into stopping the tree process just before the $n+1$ speciation event, i.e. the last interspeciation time is $\exp(n)$ distributed. We introduce the notation that $U^{(n)}$ is the height of the tree, $\tau^{(n)}$ is the time to coalescent of two randomly selected tip species and $T_k$ is the time between speciation events $k$ and $k+1$ (see Fig. 1 and Bartoszek and Sagitov, 2015b; Sagitov and Bartoszek, 2012).

**Theorem 2.4** *The cophenetic index is an increasing sequence of random variables, $\Phi^{(n+1)} > \Phi^{(n)}$ and has the recursive representation*

$$\Phi^{(n+1)} = \Phi^{(n)} + nU^{(n)} - \sum_{i=1}^{n} \xi_i^{(n)} \sum_{i \neq j}^{n} \tau_{ij}^{(n)}, \tag{1}$$

*where $\xi_i^{(n)}$ is an indicator random variable whether tip $i$ split at the $n$–th speciation event.*

PROOF From the definition we can see that

$$\Phi^{(n)} = \sum_{1 \leq i < j \leq n} \left( U^{(n)} - \tau_{ij}^{(n)} \right) = \binom{n}{2} \left( U^{(n)} - \mathrm{E}\left[\tau^{(n)}|\mathcal{Y}_n\right] \right),$$
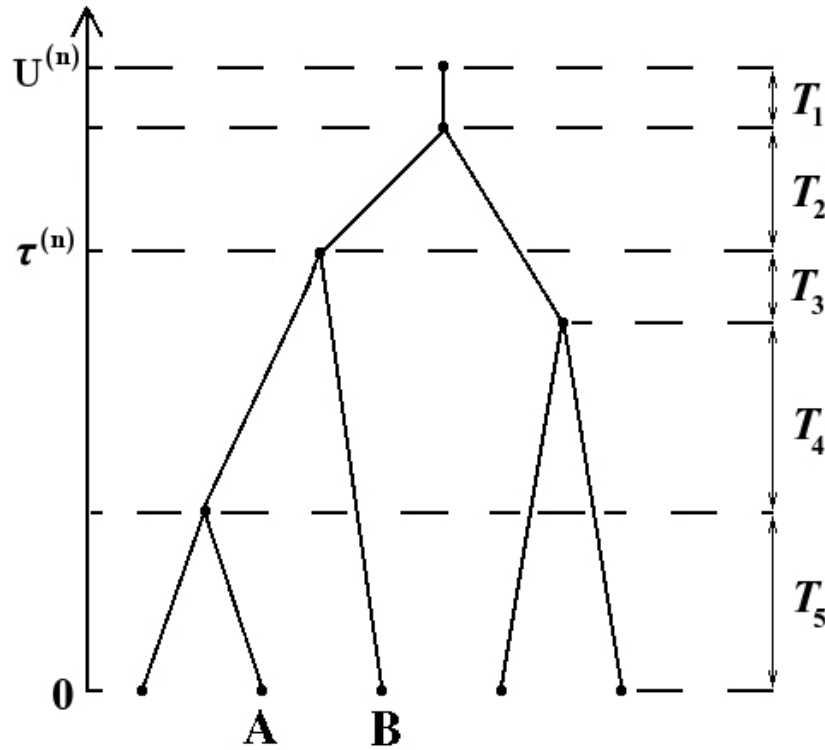
4

Figure 1: A pure–birth tree with the various time components marked on it. The between speciation times on this lineage are $T_1$, $T_2$, $T_3 + T_4$ and $T_5$. If we "randomly sample" the pair of extant species "A" and "B", then the two nodes coalesced at time $\tau^{(n)}$.

where $\tau_{ij}^{(n)}$ is the time to coalescent of tip species $i$ and $j$. We now develop a recursive representation for the cophenetic index. First notice that when a new speciation occurs all coalescent times are extended by $T_{n+1}$, i.e.

$$\sum_{1 \leq i < j \leq n+1} \tau_{ij}^{(n+1)} = \sum_{1 \leq i < j \leq n} \left( \tau_{ij}^{(n)} + T_{n+1} \right) + \sum_{i=1}^{n} \xi_i^{(n)} \sum_{i \neq j}^{n} \left( \tau_{ij}^{(n)} + T_{n+1} \right) + T_{n+1},$$

where the "lone" $T_{n+1}$ is the time to coalescent of the two descendants of the split tip. The vector $\left( \xi_1^{(n)}, \ldots, \xi_n^{(n)} \right)$ consists of $n - 1$ 0s and exactly one 1 (a categorical distribution with $n$ categories all with equal probability). For each $i$ the marginal probability that $\xi_i^{(n)}$ is 1 is $1/n$. We rewrite

5

$$\sum_{1 \leq i < j \leq n+1} \tau_{ij}^{(n+1)} = \binom{n+1}{2} T_{n+1} \sum_{1 \leq i < j \leq n} \tau_{ij}^{(n)} + \sum_{i=1}^{n} \xi_i^{(n)} \sum_{i \neq j}^{n} \tau_{ij}^{(n)}$$

and then obtain the recursive form

$$
\begin{aligned}
\Phi^{(n+1)} &= \binom{n+1}{2} U^{(n)} + \binom{n+1}{2} T_{n+1} - \binom{n+1}{2} T_{n+1} - \sum_{1 \leq i < j \leq n} \left( \tau_{ij}^{(n)} + T_{n+1} \right) \\
&\quad - \sum_{i=1}^{n} \xi_i^{(n)} \sum_{i \neq j}^{n} \tau_{ij}^{(n)} \\
&= \binom{n+1}{2} U^{(n)} - \sum_{1 \leq i < j \leq n} \left( \tau_{ij}^{(n)} + T_{n+1} \right) - \sum_{i=1}^{n} \xi_i^{(n)} \sum_{i \neq j}^{n} \tau_{ij}^{(n)} \\
&= \Phi^{(n)} + n U^{(n)} - \sum_{i=1}^{n} \xi_i^{(n)} \sum_{i \neq j}^{n} \tau_{ij}^{(n)}.
\end{aligned}
$$

Obviously, $\Phi^{(n+1)} > \Phi^{(n)}$.

$\square$

Let $\mathcal{Y}_n$ be the $\sigma$–algebra containing all the information on the Yule with $n$ tips tree. We introduce the notation

$$H_{n,m} := \sum_{k=1}^{n} 1/k^m.$$

We will now associate an almost surely and $L^2$ convergent submartingale with $\Phi^{(n)}$.

**Theorem 2.5** *Consider a scaled cophenetic index*

$$W_n = \binom{n}{2}^{-1} \Phi^{(n)}.$$

*$W_n$ is a positive submartingale that converges almost surely and in $L^2$ to a finite first and second moment random variable.*

PROOF Obviously

$$W_{n+1} = \frac{n-1}{n+1} W_n + \frac{2}{n+1} U^{(n)} - \binom{n+1}{2}^{-1} \sum_{i=1}^{n} \xi_i^{(n)} \sum_{i \neq j}^{n} \tau_{ij}^{(n)}$$

and

$$
\begin{aligned}
\mathrm{E}\left[W_{n+1}|\mathcal{Y}_n\right] &= \tfrac{n-1}{n+1}W_n + \binom{n+1}{2}^{-1}\left(nU^{(n)} - \tfrac{1}{n}\sum_{i=1}^{n}\sum_{i\neq j}^{n}\tau_{ij}^{(n)}\right) \\
&= \tfrac{n-1}{n+1}W_n + \binom{n+1}{2}^{-1}\tfrac{2}{n}\left(\tfrac{n^2}{2}U^{(n)} - \sum_{i<j}^{n}\tau_{ij}^{(n)}\right) \\
&= \tfrac{n-1}{n+1}W_n + \binom{n+1}{2}^{-1}\tfrac{2}{n}\left(\binom{n}{2}W_n + \tfrac{n}{2}U^{(n)}\right) \\
&= \left(\tfrac{n-1}{n+1} + \binom{n+1}{2}^{-1}\tfrac{2}{n}\binom{n}{2}\right)W_n + \binom{n+1}{2}^{-1}U^{(n)} \\
&= \tfrac{(n-1)(n+2)}{n(n+1)}W_n + \binom{n+1}{2}^{-1}U^{(n)} \\
&> W_n + \binom{n+1}{2}^{-1}U^{(n)} > W_n.
\end{aligned}
$$

Hence, $W_n$ is a positive submartingale with respect to $\mathcal{Y}_n$. Notice that

$$
\mathrm{E}\left[W_n^2\right] = \mathrm{E}\left[(U^{(n)} - \mathrm{E}\left[\tau^{(n)}|\mathcal{Y}_n\right])^2\right] \le \mathrm{E}\left[(U^{(n)} - \tau^{(n)})^2\right].
$$

Then using the general formula for the moment of $U^{(n)} - \tau^{(n)}$ (Appendix A, Bartoszek and Sagitov, 2015b) we see that

$$
\begin{aligned}
\mathrm{E}\left[(U^{(n)} - \tau^{(n)})^2\right] &= 2\tfrac{n+1}{n-1}\sum_{j=1}^{n-1}\tfrac{1}{(j+1)(j+2)}\left(H_{j,1}^2 + H_{j,2}\right) \\
&= 2\tfrac{n+1}{n-1}\left(\tfrac{n}{n+1}H_{n,2} - \tfrac{n}{n+1} - \tfrac{H_{n,2}}{n+1} + \sum_{j=1}^{n-1}\tfrac{H_{j,1}^2}{(j+1)(j+2)}\right) \nearrow \tfrac{2}{3}\pi^2.
\end{aligned}
$$

Hence, $\mathrm{E}\left[W_n\right]$ and $\mathrm{E}\left[W_n^2\right]$ are $O(1)$ and by the martingale convergence theorem $W_n$ converges almost surely and in $L^2$ to a finite first and second moment random variable.

$\square$

**Corollary 2.6** $W_n$ *has finite third moment and is* $L^3$ *convergent.*

PROOF Using the general formula for the moment of $U^{(n)} - \tau^{(n)}$ again we see

7

$$
\begin{aligned}
\mathrm{E}\left[(U^{(n)} - \mathrm{E}\left[\tau^{(n)}|\mathcal{Y}_n\right])^3\right] & \leq \mathrm{E}\left[(U^{(n)} - \tau^{(n)})^3\right] \\
& = 2\frac{n+1}{n-1}\sum_{j=1}^{n-1}\frac{1}{(j+1)(j+2)}\left(H_{j,1} + 3H_{j,1} + 3H_{j,2} + H_{j,3}\right) \\
& < 16\frac{n+1}{n-1}\sum_{j=1}^{n-1}\frac{H_{j,1}}{(j+1)(j+2)} \\
& = 16\frac{n+1}{n-1}\frac{n-H_{n,1}}{n+1} = 16\frac{n-H_{n,1}}{n-1} \nearrow 16.
\end{aligned}
$$

This implies that $\mathrm{E}\left[W_n^3\right] = O(1)$ and hence $L^3$ convergence and finiteness of the third moment.

$\square$

**Remark 2.7** *Notice that we (Appendix A, Bartoszek and Sagitov, 2015b) made a typo in the general formula for the cross moment of*

$$
\mathrm{E}\left[(U^{(n)} - \tau^{(n)})^m \tau^{(n)^r}\right].
$$

*The $(-1)^{m+r}$ should not be there, it will cancel with the $(-1)^{m+r}$ from the derivative of the Laplace transform.*

**Definition 2.8** *For $k = 1, \ldots, n-1$ let us define $1_k^{(n)}$ as the indicator random variable taking the value of 1 if a randomly sampled pair of species coalesced at the $k$–th (counting from the origin of the tree) speciation event.*

We know (e.g. Bartoszek and Sagitov, 2015b; Stadler, 2009; Steel and McKenzie, 2001) that

$$
\mathbb{P}(1_k^{(n)} = 1) = \mathrm{E}\left[1_k^{(n)}\right] = 2\frac{n+1}{n-1}\frac{1}{(k+1)(k+2)} \equiv \pi_{n,k}. \tag{2}
$$

**Definition 2.9** *For $i = 1, \ldots, n-1$ let us introduce the random variable*

$$
V_i^{(n)} := \frac{1}{i}\sum_{k=i}^{n-1}\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]. \tag{3}
$$

**Theorem 2.10** *$W_n$ can be represented as*

$$
W_n = \sum_{i=1}^{n-1}V_i^{(n)}Z_i, \tag{4}
$$

*where $Z_1, \ldots, Z_{n-1}$ are i.i.d. exponential with rate 1 random variables.*

8

PROOF We write $W_n$ as

$$
\begin{aligned}
W_n &= U^{(n)} - \mathrm{E}\left[\tau^{(n)}|\mathcal{Y}_n\right] = \mathrm{E}\left[U^{(n)} - \tau^{(n)}|\mathcal{Y}_n\right] = \mathrm{E}\left[\sum_{k=1}^{n-1} 1_k^{(n)} \sum_{i=1}^{k} T_i|\mathcal{Y}_n\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{n-1} T_i \sum_{k=i}^{n-1} 1_k^{(n)}|\mathcal{Y}_n\right] = \sum_{i=1}^{n-1} T_i \sum_{k=i}^{n-1} \mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right] \\
&= \sum_{i=1}^{n-1}\left(\frac{1}{i}\sum_{k=i}^{n-1} \mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]\right) Z_i = \sum_{i=1}^{n-1} V_i^{(n)} Z_i,
\end{aligned}
$$

where $Z_1, \ldots, Z_{n-1}$ are i.i.d. exponential with rate 1 random variables.

$\square$

**Remark 2.11** *We notice that we may equivalently rewrite*

$$
W_n = \sum_{k=1}^{n-1} \mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]\left(\sum_{i=1}^{k} T_i\right) = \sum_{k=1}^{n-1} \mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]\left(\sum_{i=1}^{k} \frac{1}{i} Z_i\right). \tag{5}
$$

**Definition 2.12** *Define the random variable $\overline{W}_n$ as*

$$
\overline{W}_n = \sum_{i=1}^{n-1} \mathrm{E}\left[V_i^{(n)}\right] Z_i, \tag{6}
$$

*where $Z_1, \ldots, Z_{n-1}$ are i.i.d. exponential with rate 1 random variables.*

**Remark 2.13** *Despite the apparent elegance, it is not visible how to derive a Central Limit Theorem (CLT) or limit statements concerning $W_n$ from the representations of Eqs. (4) or (5). Initially one could hope (based on "typical" results on limits for randomly weighted sums, e.g. Thm. 1 of Rosalsky and Sreehari, 1998) that $W_n$ could converge a.s. to a random variable that has the same limiting distribution as $\overline{W}_n$.*

*Similarly, as $((n+2)(n-1)/(n(n+1)) > 1$, we have that $\overline{W}_n$ is an $L^2$ bounded submartingale*

$$
\mathrm{E}\left[\overline{W}_{n+1}|\overline{W}_n\right] = \frac{(n+2)(n-1)}{n(n+1)}\overline{W}_n + \frac{2}{n^2(n+1)} > \overline{W}_n.
$$

*Hence $\overline{W}_n$ converges almost surely. Simulations presented in Fig. 2 can easily mislead one to believe in the equality of the limiting distributions of*

$W_n$ and $\overline{W}_n$. However, in Thm. 3.8 we can see that $\text{Var}\left[W_n\right]$ and $\text{Var}\left[\overline{W}_n\right]$ convergence to different limits. Therefore, $W_n$ and $\overline{W}_n$ cannot converge in distribution to the same limit.
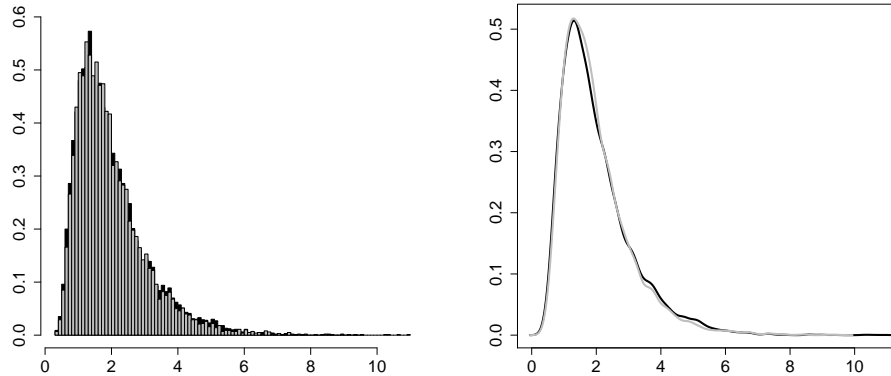


Figure 2: Left: the black is a histogram of values of $W_n$ from 10000 simulated Yule trees with $\lambda = 1$. The gray is a histogram of the approximation of $\overline{W}_n$ (independent exponentials with rate 1 variables were drawn). Right: the black curve is an estimate (via R's `density()` function) of $W_n$'s density and gray is an density estimate of $\overline{W}_n$'s density. The simulated sample of $W_n$ has mean 2, variance 1.214, skewness 1.609 and excess kurtosis 4.237 while the simulated sample of $\overline{W}_n$ has mean 1.973, variance 1.109, skewness 1.634 and excess kurtosis 4.159. It is obvious that $\text{E}\left[W_n\right] = \text{E}\left[\overline{W}_n\right]$, but we have shown that their variances differ (simulations agree with Thm. 3.8). Therefore, the differences in skewness and kurtosis could be true (despite being of the magnitude of differences in sample averages).

## 3  Second order properties

In this Section we prove a series of rather technical Lemmata and Theorems concerning the second order properties of $1_k^{(n)}$, $V_i^{(n)}$ and $W_n$. Even though we will not obtain any weak limit the derived properties do give insight on the delicate behaviour of $W_n$ and also show that no "simple" limit, e.g. Eq.

10

(6), is possible. To obtain our results we used Mathematica 9.0 for Linux x86 (64–bit) running on Ubuntu 12.04.5 LTS to evaluate the required sums in closed forms. The Mathematica code is available as an appendix to this paper.

**Lemma 3.1**

$$\mathrm{Var}\left[1_k^{(n)}\right] = 2\frac{n+1}{n-1}\frac{1}{(k+1)(k+2)}\left(1 - 2\frac{n+1}{n-1}\frac{1}{(k+1)(k+2)}\right) \quad (7)$$

PROOF

$$\begin{aligned}
\mathrm{Var}\left[1_k^{(n)}\right] &= \mathrm{E}\left[1_k^{(n)^2}\right] - \mathrm{E}\left[1_k^{(n)}\right]^2 = \pi_{n,k} - \pi_{n,k}^2 = \pi_{n,k}(1 - \pi_{n,k}) \\
&= 2\frac{n+1}{n-1}\frac{1}{(k+1)(k+2)}\left(1 - 2\frac{n+1}{n-1}\frac{1}{(k+1)(k+2)}\right).
\end{aligned}$$

$\square$

The following lemma is an obvious consequence of the definition of $1_k^{(n)}$.

**Lemma 3.2** *For $k_1 \neq k_2$*

$$\mathrm{Cov}\left[1_{k_1}^{(n)}, 1_{k_2}^{(n)}\right] = -\pi_{n,k_1}\pi_{n,k_2} = \frac{(-4)(n+1)^2}{(n-1)^2(k_1+1)(k_1+2)(k_2+1)(k_2+2)}. \quad (8)$$

**Lemma 3.3**

$$\mathrm{Var}\left[\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]\right] = 4\frac{n+1}{n(n-1)^2}\frac{(n-(k+1))(n(3k^2+5k-4)-(k^2-k-8))}{(k+1)^2(k+2)^2(k+3)(k+4)} \quad (9)$$

PROOF Obviously

$$\mathrm{Var}\left[\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]\right] = \mathrm{E}\left[\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]^2\right] - \mathrm{E}\left[\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]\right]^2.$$

We notice (as Bartoszek and Sagitov, 2015b; Bartoszek, 2016, in Lemmata 11 and 2 respectively) that we may write

$$\mathrm{E}\left[\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]^2\right] = \mathrm{E}\left[1_{k,1}^{(n)}1_{k,2}^{(n)}\right],$$

11

where $1_{k,1}^{(n)}$, $1_{k,2}^{(n)}$ are two independent copies of $1_k^{(n)}$, i.e. we sample a pair of tips twice and ask if both pairs coalesced at the $k$–th speciation event. There are three possibilities, we (i) drew the same pair, (ii) drew two pairs sharing a single node or (iii) drew two disjoint pairs. Event (i) occurs with probability $\binom{n}{2}^{-1}$, (ii) with probability $2(n-2)\binom{n}{2}^{-1}$ and (iii) with probability $\binom{n-2}{2}\binom{n}{2}^{-1}$. As a check notice that $1 + 2(n-2) + \binom{n-2}{2} = \binom{n}{2}$. In case (i) $1_{k,1}^{(n)} = 1_{k,2}^{(n)}$, hence writing informally

$$\mathrm{E}\left[1_{k,1}^{(n)}1_{k,2}^{(n)}|(i)\right] = \mathrm{E}\left[1_k^{(n)}\right] = \pi_{n,k}.$$
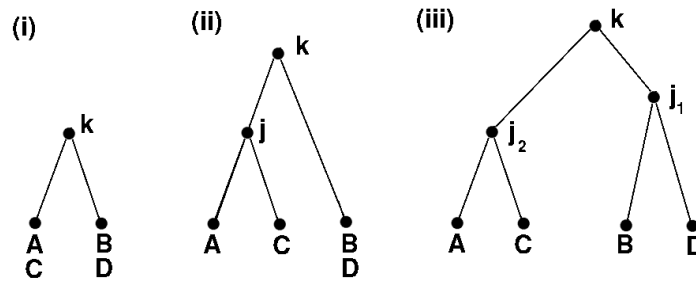


Figure 3: The three possible cases when drawing two random pairs of tip species that coalesce at the $k$–th speciation event. In the picture we "randomly draw" pairs $(A, B)$ and $(C, D)$.

To calculate cases (ii) and (iii) we visualize the situation in Fig. 3 and recall the proof of Bartoszek and Sagitov (2015b)'s Lemma 1. Using Mathematica we obtain

$$
\begin{aligned}
\mathrm{E}\left[1_{k,1}^{(n)}1_{k,2}^{(n)}|(ii)\right] &= \sum_{j=k+1}^{n-1}\left(1-\frac{3}{\binom{n}{2}}\right)\cdots\left(1-\frac{3}{\binom{j+2}{2}}\right)\frac{1}{\binom{j+1}{2}}\left(1-\frac{1}{\binom{j}{2}}\right)\cdots \\
&\quad \cdot\left(1-\frac{1}{\binom{k+2}{2}}\right)\frac{1}{\binom{k+1}{2}} \\
&= 4\frac{(n+1)}{(n-1)(n-2)}\frac{n-(k+1)}{(1+k)(2+k)(3+k)}.
\end{aligned}
$$

Similarly for case (iii)

$$
\begin{aligned}
\mathrm{E}\left[1_{k,1}^{(n)}1_{k,2}^{(n)}|(\mathrm{iii})\right] &= \sum_{j_2=k+2}^{n-1}\sum_{j_1=k+1}^{j_2+1}\left(1-\frac{6}{\binom{n}{2}}\right)\cdots\left(1-\frac{6}{\binom{j_2+2}{2}}\right)\frac{4}{\binom{j_2+1}{2}} \\
&\quad \cdot\left(1-\frac{3}{\binom{j_2}{2}}\right)\cdots\left(1-\frac{3}{\binom{j_1+2}{2}}\right)\frac{1}{\binom{j_1+1}{2}}\left(1-\frac{1}{\binom{j_1}{2}}\right)\cdots \\
&\quad \cdot\left(1-\frac{1}{\binom{k+2}{2}}\right)\frac{1}{\binom{k+1}{2}} \\
&= 16\frac{(n+1)}{(n-1)(n-2)(n-3)}\frac{(n-(k+1))(n-(k+2))}{(k+1)(k+2)(k+3)(k+4)}.
\end{aligned}
$$

We now put this together as

$$
\begin{aligned}
\mathrm{Var}\left[\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]\right] &= \binom{n}{2}^{-1}\pi_{n,k}+2(n-2)\binom{n}{2}^{-1}\mathrm{E}\left[1_{k,1}^{(n)}1_{k,2}^{(n)}|(\mathrm{ii})\right] \\
&\quad +\binom{n-2}{2}\binom{n}{2}^{-1}\mathrm{E}\left[1_{k,1}^{(n)}1_{k,2}^{(n)}|(\mathrm{iii})\right]-\pi_{n,k}^2
\end{aligned}
$$

we obtain (through Mathematica)

$$
\begin{aligned}
\mathrm{Var}\left[\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]\right] &= 4\frac{n+1}{n(n-1)^2}\frac{(n-(k+1))(n(3k^2+5k-4)-(k^2-k-8))}{(k+1)^2(k+2)^2(k+3)(k+4)} \\
&\to 4\frac{3k^2+5k-4}{(k+1)^2(k+2)^2(k+3)(k+4)}.
\end{aligned}
$$

$\square$

**Lemma 3.4** *For $k_1 < k_2$*

$$
\mathrm{Cov}\left[\mathrm{E}\left[1_{k_2}^{(n)}|\mathcal{Y}_n\right],\mathrm{E}\left[1_{k_1}^{(n)}|\mathcal{Y}_n\right]\right] = \frac{(-8)(n+1)}{n(n-1)^2}\frac{(3n-(k_2-2))(n-(k_2+1))}{(k_1+1)(k_1+2)(k_2+1)(k_2+2)(k_2+3)(k_2+4)}. \tag{10}
$$

PROOF Obviously

$$
\mathrm{Cov}\left[\mathrm{E}\left[1_{k_1}^{(n)}|\mathcal{Y}_n\right],\mathrm{E}\left[1_{k_2}^{(n)}|\mathcal{Y}_n\right]\right] = \mathrm{E}\left[\mathrm{E}\left[1_{k_1}^{(n)}|\mathcal{Y}_n\right]\mathrm{E}\left[1_{k_2}^{(n)}|\mathcal{Y}_n\right]\right]-\mathrm{E}\left[1_{k_1}\right]\mathrm{E}\left[1_{k_2}\right].
$$

We notice that

$$
\mathrm{E}\left[\mathrm{E}\left[1_{k_1}^{(n)}|\mathcal{Y}_n\right]\mathrm{E}\left[1_{k_2}^{(n)}|\mathcal{Y}_n\right]\right] = \mathrm{E}\left[1_{k_1}^{(n)}1_{k_2}^{(n)}\right],
$$

13

where $1_{k_1}^{(n)}$, $1_{k_2}^{(n)}$ are the indicator variables if two independently sampled pairs coalesced at speciation events $k_1 < k_2$ respectively. There are now two possibilities represented in Fig. 4 (notice that since $k_1 \neq k_2$ the counterpart of event (i) in Fig. 3 cannot take place). Event (ii) occurs with probability $4/(n+1)$ and (iii) with probability $(n-3)/(n+1)$. Event (iii) can be divided into three "subevents".
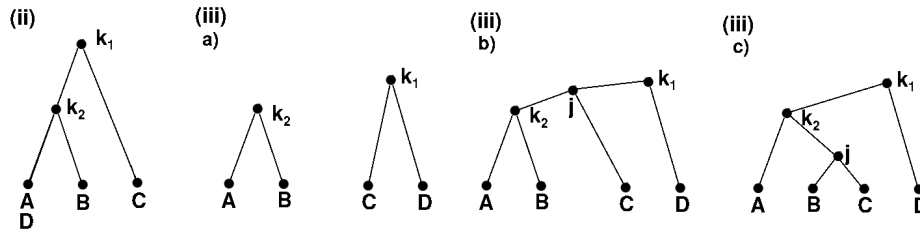


Figure 4: The possible cases when drawing two random pairs of tip species that coalesce at speciation events $k_1 < k_2$ respectively. In the picture we "randomly draw" pairs $(A, B)$ and $(C, D)$.

Again we recall the proof of Bartoszek and Sagitov (2015b)'s Lemma 1 and we write informally for (ii) using Mathematica

$$
\begin{aligned}
\mathrm{E}\left[1_{k_1}^{(n)} 1_{k_2}^{(n)} | (\mathrm{ii})\right] &= \left(1 - \tfrac{3}{\binom{n}{2}}\right) \ldots \left(1 - \tfrac{3}{\binom{k_2+2}{2}}\right) \tfrac{1}{\binom{k_2+1}{2}} \left(1 - \tfrac{1}{\binom{k_2}{2}}\right) \ldots \\
&\quad \cdot \left(1 - \tfrac{1}{\binom{k_1+2}{2}}\right) \tfrac{1}{\binom{k_1+1}{2}} \\
&= 4 \tfrac{(n+1)(n+2)}{(n-1)(n-2)} \tfrac{1}{(k_1+1)(k_1+2)(k_2+2)(k_2+3)}.
\end{aligned}
$$

In the same way for the subcases of (iii)

14

$$
\begin{aligned}
\mathrm{E}\left[1_{k_1}^{(n)}1_{k_2}^{(n)}|(\text{iii})\right] &= \left(1-\frac{6}{\binom{n}{2}}\right)\cdots\left(1-\frac{6}{\binom{k_2+2}{2}}\right)\frac{1}{\binom{k_2+1}{2}} \\
&\quad\cdot\left(1-\frac{3}{\binom{k_2}{2}}\right)\cdots\left(1-\frac{3}{\binom{k_1+2}{2}}\right)\frac{1}{\binom{k_1+1}{2}} \\
&\quad+\sum_{j=k_1+1}^{k_2-1}\left(1-\frac{6}{\binom{n}{2}}\right)\cdots\left(1-\frac{6}{\binom{k_2+2}{2}}\right)\frac{1}{\binom{k_2+1}{2}} \\
&\quad\cdot\left(1-\frac{3}{\binom{k_2}{2}}\right)\cdots\left(1-\frac{3}{\binom{j+2}{2}}\right)\frac{2}{\binom{j+1}{2}}\left(1-\frac{1}{\binom{j}{2}}\right)\cdots \\
&\quad\cdot\left(1-\frac{1}{\binom{k_1+2}{2}}\right)\frac{1}{\binom{k_1+1}{2}} \\
&\quad+\sum_{j=k_2+1}^{n-1}\left(1-\frac{6}{\binom{n}{2}}\right)\cdots\left(1-\frac{6}{\binom{j+2}{2}}\right)\frac{4}{\binom{j+1}{2}} \\
&\quad\cdot\left(1-\frac{3}{\binom{j}{2}}\right)\cdots\left(1-\frac{3}{\binom{k_2+2}{2}}\right)\frac{2}{\binom{k_2+1}{2}}\left(1-\frac{1}{\binom{k_2}{2}}\right)\cdots \\
&\quad\cdot\left(1-\frac{1}{\binom{k_1+2}{2}}\right)\frac{1}{\binom{k_1+1}{2}} \\
&= 4\frac{(n+2)(n+1)}{(n-1)(n-2)(n-3)}\frac{n(k_2+6)-5k_2-14}{(k_1+1)(k_1+2)(k_2+2)(k_2+3)(k_2+4)}.
\end{aligned}
$$

We now put this together as

$$
\begin{aligned}
\mathrm{Cov}\left[\mathrm{E}\left[1_{k_1}^{(n)}|\mathcal{Y}_n\right],\mathrm{E}\left[1_{k_2}^{(n)}|\mathcal{Y}_n\right]\right] &= 2(n-2)\binom{n}{2}^{-1}\mathrm{E}\left[1_{k_1}^{(n)}1_{k_2}^{(n)}|(\text{ii})\right] \\
&\quad+\binom{n-2}{2}\binom{n}{2}^{-1}\mathrm{E}\left[1_{k_1}^{(n)}1_{k_2}^{(n)}|(\text{iii})\right]-\pi_{n,k_1}\pi_{n,k_2}
\end{aligned}
$$

and we obtain

$$
\begin{aligned}
\mathrm{Cov}\left[\mathrm{E}\left[1_{k_1}^{(n)}|\mathcal{Y}_n\right],\mathrm{E}\left[1_{k_2}^{(n)}|\mathcal{Y}_n\right]\right] &= \frac{(-8)(n+1)}{n(n-1)^2}\frac{(3n-(k_2-2))(n-(k_2+1))}{(k_1+1)(k_1+2)(k_2+1)(k_2+2)(k_2+3)(k_2+4)} \\
&\to (-24)\frac{1}{(k_1+1)(k_1+2)(k_2+1)(k_2+2)(k_2+3)(k_2+4)}.
\end{aligned}
$$

$\square$

**Theorem 3.5**
$$
\mathrm{E}\left[V_i^{(n)}\right] = 2\frac{1}{n-1}\frac{n-i}{i(i+1)} \tag{11}
$$

15

PROOF We immediately have

$$
\begin{aligned}
\mathrm{E}\left[V_i^{(n)}\right] &= \frac{1}{i}\sum_{k=i}^{n-1}\mathrm{E}\left[\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]\right]\\
&= 2\frac{n+1}{n-1}\frac{1}{i}\sum_{k=i}^{n-1}\frac{1}{(k+1)(k+2)}\\
&= 2\frac{1}{n-1}\frac{n-i}{i(i+1)}\\
&\to \frac{2}{i(i+1)}.
\end{aligned}
$$

$\square$

**Theorem 3.6**

$$
\mathrm{Var}\left[V_i^{(n)}\right] = 4\frac{(n+1)}{n(n-1)^2}\frac{(n-i)(n-(i+1))(i-1)}{i^2(i+1)^2(i+2)(i+3)} \tag{12}
$$

PROOF We immediately may write using Lemmata 3.3, 3.4 and Mathematica

$$
\begin{aligned}
\mathrm{Var}\left[V_i^{(n)}\right] &= \frac{1}{i^2}\left(\sum_{k=i}^{n-1}\mathrm{Var}\left[\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]\right] + 2\sum_{i=k_1<k_2}^{n-1}\mathrm{Cov}\left[\mathrm{E}\left[1_{k_1}^{(n)}|\mathcal{Y}_n\right],\mathrm{E}\left[1_{k_2}^{(n)}|\mathcal{Y}_n\right]\right]\right)\\
&= \frac{4}{i^2}\left(\sum_{k=i}^{n-1}\frac{n+1}{n(n-1)^2}\frac{(n-(k+1))(n(3k^2+5k-4)-(k^2-k-8))}{(k+1)^2(k+2)^2(k+3)(k+4)}\right.\\
&\quad -4\sum_{i=k_1<k_2}^{n-1}\frac{(n+1)}{n(n-1)^2}\frac{(3n-(k_2-2))(n-(k_2+1))}{(k_1+1)(k_1+2)(k_2+1)(k_2+2)(k_2+3)(k_2+4)}\Bigg)\\
&= 4\frac{(n+1)}{n(n-1)^2}\frac{(n-i)(n-(i+1))(i-1)}{i^2(i+1)^2(i+2)(i+3)}\\
&\to 4\frac{(i-1)}{i^2(i+1)^2(i+2)(i+3)}.
\end{aligned}
$$

$\square$

**Theorem 3.7** *For $1 \le i_1 < i_2 \le n-1$ we have*

$$
\mathrm{Cov}\left[V_{i_1}^{(n)}, V_{i_2}^{(n)}\right] = 4\frac{(n+1)}{n(n-1)^2}\frac{(i_1-1)(n-i_2)(n-(i_2+1))}{i_1(i_1+1)i_2(i_2+1)(i_2+2)(i_2+3)}. \tag{13}
$$

PROOF Again using Lemmata 3.3, 3.4, Mathematica and the fact that $i_1 < i_2$

16

$$
\begin{aligned}
\mathrm{Cov}\left[V_{i_1}^{(n)}, V_{i_2}^{(n)}\right] &= \tfrac{1}{i_1 i_2}\left(\mathrm{Cov}\left[\sum_{k=i_1}^{n-1}\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right], \sum_{k=i_2}^{n-1}\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]\right]\right)\\
&= \tfrac{1}{i_1 i_2}\left(\mathrm{Var}\left[\sum_{k=i_2}^{n-1}\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]\right] + \mathrm{Cov}\left[\sum_{k=i_1}^{i_2-1}\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right], \sum_{k=i_2}^{n-1}\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]\right]\right)\\
&= \tfrac{1}{i_1 i_2}\left((i_2^2)\,\mathrm{Var}\left[V_{i_2}^{(n)}\right] + \sum_{k_1=i_1}^{i_2-1}\sum_{k_2=i_2}^{n-1}\mathrm{Cov}\left[\mathrm{E}\left[1_{k_1}^{(n)}|\mathcal{Y}_n\right], \sum_{k=i_2}^{n-1}\mathrm{E}\left[1_{k_2}^{(n)}|\mathcal{Y}_n\right]\right]\right)\\
&= 4\frac{(n+1)}{n(n-1)^2}\frac{(i_1-1)(n-i_2)(n-(i_2+1))}{i_1(i_1+1)i_2(i_2+1)(i_2+2)(i_2+3)}\\
&\to 4\frac{i_1-1}{i_1(i_1+1)i_2(i_2+1)(i_2+2)(i_2+3)}.
\end{aligned}
$$

$\square$

**Theorem 3.8**

$$
\begin{aligned}
\mathrm{Var}\left[\sum_{i=1}^{n-1}V_i^{(n)}\right] &= \tfrac{1}{54n^2(n-1)^2}\big(179n^4 + 588n^3 + 133n^2 - 432n\\
&\qquad -468 - 108n^2(n+1)(n+3)H_{n-1,2}\\
&\qquad -144nH_{n-1,1}\big) \to \tfrac{179}{54} - \tfrac{\pi^2}{3} \approx 1.347,\\
\mathrm{Var}\left[\sum_{i=1}^{n-1}V_i^{(n)}Z_i\right] &= \tfrac{1}{9n^2(n-1)^2}\big(12n^2(n^2 - 6n - 4)H_{n-1,2} - 9n^4\\
&\qquad +102n^3 + 51n^2 - 24nH_{n-1,1} - 72n - 72\big)\\
&\to \tfrac{2}{9}\pi^2 - 1 \approx 1.193,\\
\mathrm{Var}\left[\sum_{i=1}^{n-1}\mathrm{E}\left[V_i^{(n)}\right]Z_i\right] &= \tfrac{2}{3n^2(n-1)^2}\big((12H_{n-1,2} - 18)\,n^4 - 24n^3\\
&\qquad +12n^2(2n+1)H_{n-1,2} - 24n^2 + 24n + 12\big)\\
&\to \tfrac{4}{3}\pi^2 - 12 \approx 1.159,\\
\mathrm{Var}\left[\sum_{i=1}^{n-1}\left(V_i^{(n)} - \mathrm{E}\left[V_i^{(n)}\right]\right)Z_i\right] &= \tfrac{1}{9n^2(n-1)^2}\big(99n^4 + 174n^3 - 21n^2 - 144n\\
&\qquad -108 - 12n^2(n+1)(5n+7)H_{n-1,2}\\
&\qquad -24nH_{n-1,1}\big) \to 11 - \tfrac{10}{9}\pi^2 \approx 0.034.
\end{aligned}
$$

(14)

PROOF We use Mathematica to first calculate

$$
\begin{aligned}
\operatorname{Var}\left[\sum_{i=1}^{n-1} V_i^{(n)}\right] &= \sum_{i=1}^{n-1} \operatorname{Var}\left[V_i^{(n)}\right] + 2\sum_{1=i_1<i_2}^{n-1} \operatorname{Cov}\left[V_{i_1}^{(n)}, V_{i_2}^{(n)}\right] \\
&= \tfrac{1}{54n^2(n-1)^2}\left(179n^4 - 108n^2(n+1)(n+3)H_{n-1,2} + 588n^3\right. \\
&\quad \left. +133n^2 - 144nH_{n-1,1} - 432n - 468\right) \\
&\to \tfrac{179}{54} - \tfrac{\pi^2}{3} \approx 1.347.
\end{aligned}
$$

For the second we again use Mathematica and the fact that the $Z_i$s are i.i.d. $\exp(1)$.

$$
\begin{aligned}
\operatorname{Var}\left[\sum_{i=1}^{n-1} \mathrm{E}\left[V_i^{(n)}\right] Z_i\right] &= \sum_{i=1}^{n-1}\left(2\tfrac{1}{n-1}\tfrac{n-i}{i(i+1)}\right)^2 \\
&= \tfrac{2(12H_{n-1,2}-18)n^4 + 2\left(6n^2(2n+1)H_{n-1,2} - 12n^3 - 12n^2 + 12n + 6\right)}{3n^2(n-1)^2} \\
&\to \tfrac{4}{3}\pi^2 - 12 \approx 1.159.
\end{aligned}
$$

For the third equality we use Mathematica and the fact that for independent families $\{X\}$ and $\{Y\}$ of random variables we have

$$
\begin{aligned}
\operatorname{Var}\left[XY\right] &= \mathrm{E}\left[Y^2\right]\operatorname{Var}\left[X\right] + (\mathrm{E}\left[X\right])^2\operatorname{Var}\left[Y\right], \\
\operatorname{Cov}\left[X_1Y_1, X_2Y_2\right] &= \mathrm{E}\left[Y_1\right]\mathrm{E}\left[Y_2\right]\operatorname{Cov}\left[X_1, X_2\right] + \mathrm{E}\left[X_1\right]\mathrm{E}\left[X_2\right]\operatorname{Cov}\left[Y_1, Y_2\right].
\end{aligned}
$$

As the $Z_i$s are i.i.d. $\exp(1)$ we use Mathematica to obtain

$$
\begin{aligned}
\operatorname{Var}\left[\sum_{i=1}^{n-1} V_i^{(n)} Z_i\right] &= \sum_{i=1}^{n-1} \operatorname{Var}\left[V_i^{(n)} Z_i\right] + 2\sum_{1=i_1<i_2}^{n-1} \operatorname{Cov}\left[V_{i_1}^{(n)} Z_{i_1}, V_{i_2}^{(n)} Z_{i_2}\right] \\
&= 2\sum_{i=1}^{n-1} \operatorname{Var}\left[V_i^{(n)}\right] + \sum_{i=1}^{n-1}\left(\mathrm{E}\left[V_i^{(n)}\right]\right)^2 + 2\sum_{1=i_1<i_2}^{n-1} \operatorname{Cov}\left[V_{i_1}^{(n)}, V_{i_2}^{(n)}\right] \\
&= \tfrac{1}{9n^2(n-1)^2}\left(12n^2(n^2 - 6n - 4)H_{n-1,2} - 9n^4 + 102n^3\right. \\
&\quad \left. +51n^2 - 24nH_{n-1,1} - 72n - 72\right) \\
&\to \tfrac{1}{9}\left(2\pi^2 - 9\right) \approx 1.193.
\end{aligned}
$$

For the fourth equality we use the same properties and pair–wise independence of $Z_i$s.

$$\text{Var}\left[\sum_{i=1}^{n-1}\left(V_i^{(n)} - \text{E}\left[V_i^{(n)}\right]\right)Z_i\right] = \sum_{i=1}^{n-1}\text{Var}\left[V_i^{(n)}Z_i\right] + \sum_{i=1}^{n-1}\text{Var}\left[\text{E}\left[V_i^{(n)}\right]Z_i\right]$$

$$-2\sum_{1=i_1<i_2}^{n-1}\text{Cov}\left[V_{i_1}^{(n)}Z_{i_1}, \text{E}\left[V_{i_2}^{(n)}\right]Z_{i_2}\right]$$

$$= \sum_{i=1}^{n-1}\text{Var}\left[V_i^{(n)}Z_i\right] + \sum_{i=1}^{n-1}\left(\text{E}\left[V_i^{(n)}\right]\right)^2 - 2\sum_{i=1}^{n-1}\left(\text{E}\left[V_i^{(n)}\right]\right)^2$$

$$= \sum_{i=1}^{n-1}\text{Var}\left[V_i^{(n)}Z_i\right] - \sum_{i=1}^{n-1}\left(\text{E}\left[V_i^{(n)}\right]\right)^2$$

$$= \frac{1}{9n^2(n-1)^2}\left(99n^4 - 12n^2(n+1)(5n+7)H_{n-1,2} + 174n^3 - 21n^2\right.$$

$$\left. -24nH_{n-1,1} - 144n - 108\right) \to 11 - \frac{10}{9}\pi^2 \approx 0.034.$$

$\square$

It is worth noting that the above Lemmata and Theorems were confirmed by numerical evaluations of the formulae and comparing these to simulations performed to obtain Fig. 2. As a check also notice that, as implied by variance properties,

$$\text{Var}\left[\sum_{i=1}^{n-1}\text{E}\left[V_i^{(n)}\right]Z_i\right] + \text{Var}\left[\sum_{i=1}^{n-1}\left(V_i^{(n)} - \text{E}\left[V_i^{(n)}\right]\right)Z_i\right]$$

$$\to \frac{4}{3}\pi^2 - 12 + 11 - \frac{10}{9}\pi^2 = \frac{2}{9}\pi^2 - 1 \leftarrow \text{Var}\left[\sum_{i=1}^{n-1}V_i^{(n)}Z_i\right].$$

**Theorem 3.9**

$$\begin{aligned}
\text{E}\left[\Phi^{(n)}\right] &= \binom{n}{2}\frac{2(n-H_{n,1})}{n-1} \sim n^2 \\
\text{Var}\left[\Phi^{(n)}\right] &= \frac{\binom{n}{2}^2}{9n^2(n-1)^2}\left(12n^2(n^2-6n-4)H_{n-1,2} - 9n^4 + 102n^3\right. \\
&\qquad\left. +51n^2 - 24nH_{n-1,1} - 72n - 72\right) \\
&\sim \frac{1}{36}\left(2\pi^2 - 9\right)n^4.
\end{aligned} \qquad (15)$$

PROOF The proof of the expectation part is due to Mir et al. (2013); Sagitov and Bartoszek (2012) while the variance is a consequence of the previously derived lemmata and theorems in this Section.

$\square$

**Remark 3.10** *We recall that in Thm. 3.9 we include the branch leading to the root. If we would not, then we would have to decrease the expectation by $\binom{n}{2}$ and variance by $\binom{n}{2}^2$. This is due to each pair of tips "having" the $\exp(1)$ root edge included in the cophenetic distance between them.*

19

**Theorem 3.11**

$$\mathrm{E}\left[\mathrm{Var}\left[(n)^{-1}\sum_{i=2}^{n-2}\frac{V_i^{(n)}Z_i-\mathrm{E}\left[V_i^{(n)}\right]}{\sqrt{\mathrm{Var}\left[V_i^{(n)}\right]}}\,\middle|\,\{V_i^{(n)}\}\right]\right]\to 0.5. \qquad (16)$$

PROOF Using the limit for the variance of $V_i^{(n)}$ (Thm. 3.6) and the independence of the $Z_i$s we have

$$\mathrm{E}\left[\mathrm{Var}\left[\frac{1}{n}\sum_{i=2}^{n-2}\frac{V_i^{(n)}Z_i-\mathrm{E}\left[V_i^{(n)}\right]}{\sqrt{\mathrm{Var}\left[V_i^{(n)}\right]}}\,\middle|\,\{V_i^{(n)}\}\right]\right]\sim\frac{1}{4n^2}\sum_{i=2}^{n-2}\frac{i^2(i+1)^2(i+2)(i+3)}{(i-1)}\,\mathrm{E}\left[(V_i^{(n)})^2\right].$$

Now from Thms. 3.6 and 3.5 we have

$$\mathrm{E}\left[(V_i^{(n)})^2\right]=4\frac{n+1}{n(n-1)^2}\frac{(n-i)(n-(i+1))(i-1)}{i^2(i+1)^2(i+2)(i+3)}+\left(\frac{2}{n-1}\frac{n-i}{i(i+1)}\right)^2$$
$$=4\frac{1}{(n-1)^2}\frac{(n-i)^2}{i^2(i+1)^2}\left(\frac{n+1}{n(n-i)}\frac{(n-(i+1))(i-1)}{(i+2)(i+3)}+1\right)\to 4\frac{i+5}{i^2(i+1)(i+2)(i+3)}.$$

Plugging this in (and using Mathematica)

$$\frac{1}{4n^2}\sum_{i=2}^{n-2}\frac{i^2(i+1)^2(i+2)(i+3)}{(i-1)}\,\mathrm{E}\left[(V_i^{(n)})^2\right]\sim\frac{1}{4n^2}\sum_{i=2}^{n-2}\frac{i^2(i+1)^2(i+2)(i+3)4(i+5)}{(i-1)i^2(i+1)(i+2)(i+3)}$$
$$=n^{-2}\sum_{i=2}^{n-2}\frac{(i+1)(i+5)}{(i-1)}=n^{-2}\frac{1}{2}\left(n^2+11n+24H_{n,1}-42\right)\to 0.5.$$

$\square$

**Remark 3.12** *Simulations presented in Fig. 5 and Thm. 3.11 suggest a different possible CLT, namely*

$$(n)^{-1}\sum_{i=2}^{n-2}\frac{V_i^{(n)}Z_i-\mathrm{E}\left[V_i^{(n)}\right]}{\sqrt{\mathrm{Var}\left[V_i^{(n)}\right]}}\xrightarrow{\text{weakly}}\text{some distribution}(\text{mean}=0,\text{variance}=\frac{1}{2}).$$

$$(17)$$

*We sum over $i = 2, \ldots n - 2$ as $V_1^{(n)} = 1$ and $V_{n-1}^{(n)} = \binom{n}{2}^{-1}$ for all $n$. It would be tempting to take the distribution to be a normal one. However, we should be wary after Rem. 2.13 and Fig. 2 that for our rather delicate problem even very fine simulations can indicate incorrect weak limits. It remains to study the variance of the conditional variance in Eq. (16). It is not entirely clear if this variance of the conditional variance will converge to 0. Hence, it remains an open problem to investigate the conjecture of Eq. (17).*
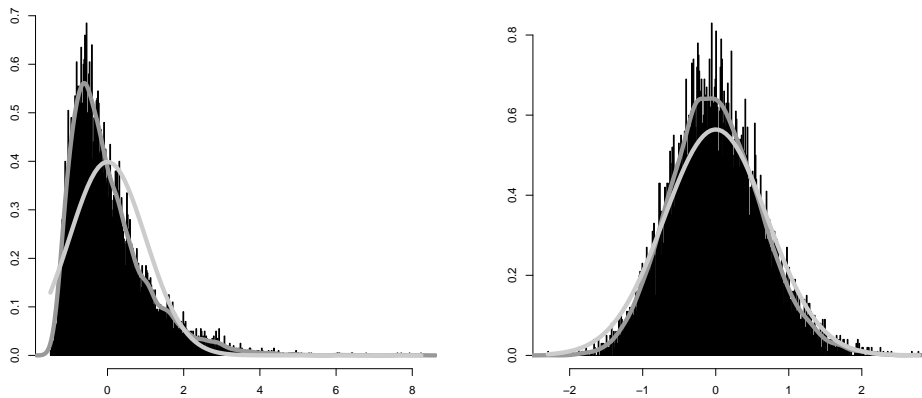


Figure 5: Histogram and density estimates of scaled and centred cophenetic indices for 10000 simulated 500 tip Yule trees with $\lambda = 1$. Left: the histogram of $(\Phi^{(n)} - \mathrm{E}\left[\Phi^{(n)}\right])/\sqrt{\mathrm{Var}\left[\Phi^{(n)}\right]}$. The darker curve is the density fitted by R's `density()` function, the light is the $\mathcal{N}(0, 1)$ density. Right: simulation of Eq. (17), the light curve is the $\mathcal{N}(0, 1/2)$ density, and the darker curve is R's `density()` from the simulated data. The sample variance of the simulated Eq. (17) values is 0.385 indicating that with $n = 500$ we still have a high variability or alternatively that the variance of the sample variance in Eq. (16) does not converge to 0.

# 4 Contraction type limit distribution

Even though the representations of Eqs. (4) and (5) are very elegant ones it is not obvious how to derive asymptotic properties of the process from it

(compare Section 3). We turn to considering the recursive representation proposed by Mir et al. (2013)

$$\tilde{\Phi}^{(n)} = \tilde{\Phi}^{(L_n)} + \tilde{\Phi}^{(R_n)} + \binom{L_n}{2} + \binom{R_n}{2} \tag{18}$$

where $L_n$ and $R_n$ are the number of left and right daughter tip descendants. We remind again that in their definition of the cophenetic index Mir et al. (2013) do not include the root tip (or root branch of length 1 in our formulation). Obviously $L_n + R_n = n$.

**Definition 4.1** *We naturally define the scaled discrete cophenetic index as*

$$\tilde{W}_n = \binom{n}{2}^{-1} \tilde{\Phi}^{(n)}. \tag{19}$$

**Theorem 4.2** $\tilde{W}_n$ *is a almost surely and $L^2$ convergent submartingale.*

PROOF The argumentation is analogous to the proof of Thm. 2.5 by using the recursion

$$\tilde{\Phi}^{(n+1)} = \tilde{\Phi}^{(n)} + \sum_{i=1}^{n} \xi_i^{(n)} \left( \sum_{i \neq j} \tilde{\phi}_{ij} + \Upsilon_i^{(n)} \right),$$

where $\Upsilon_i^{(n)}$ is the number of nodes on the path from the root (or appropriately origin) of the tree to tip $i$, (see also Bartoszek, 2014, esp. Fig. A.8). An alternative proof for almost sure convergence can be found in Section 6.1.
□

From Eq. (18) we will be able to deduce the form of the limit of the process. We recall (Mir et al., 2013; Cardona et al., 2013)

$$\begin{aligned}
\mathrm{E}\left[\tilde{\Phi}^{(n)}\right] &= \binom{n}{2}\left(\frac{4(n-H_{n,1})}{n-1} - 2\right) = 2\left(\mathrm{E}\left[\Phi^{(n)}\right] - \binom{n}{2}\right) \\
\mathrm{Var}\left[\tilde{\Phi}^{(n)}\right] &= \frac{1}{12}\left(n^4 - 10n^3 + 131n^2 - 2n\right) - 4n^2 H_{n,2} - 6n H_{n,1}.
\end{aligned} \tag{20}$$

Notice that if the root edge was included this would increase the expectation by $\binom{n}{2}$ and leave the variance unchanged (the distance between all pairs is increased by 1, hence the change is deterministic). Equation (18) is for the

discrete case (i.e. all branches of length 1). In the case with branch lengths we attempt to approximate the cophenetic index with the following contraction type law (NRE, no root edge)

$$\Phi_{NRE}^{(n)} = \Phi_{NRE}^{(L_n)} + \Phi_{NRE}^{(R_n)} + \binom{L_n}{2}T_{0.5} + \binom{R_n}{2}T'_{0.5} \tag{21}$$

where $T_{0.5}$, $T'_{0.5}$ are independent exponential random variables with rate 2 (we index with the mean to avoid confusion with $T_2$ the time between the second and third speciation event which is also exponential with rate 2). These are the branch lengths leading from the speciation point. The rationale behind the choice of distribution is that a randomly chosen internal branch of a conditioned Yule tree with rate 1 is exponentially distributed with rate 2 (Corr. 3.2 and Thm. 3.3 Stadler and Steel, 2012). This is of course an approximation, as we cannot expect that the law of every branch length should be the same as that of a random one. In fact we should expect this law to depend on $n$, i.e. the level of the recursion. However, as we shall see simulations indicate that approximating with the average law still could still yield acceptable heuristics, but not as good as the approximation by $\overline{W}_n$. We use the notation $\Phi_{NRE}^{(n)}$ to differentiate from $\Phi^{(n)}$ where the root branch is counted, i.e.

$$\Phi^{(n)} = \Phi_{NRE}^{(n)} + \binom{n}{2}T_1, \quad \text{where } T_1 \sim exp(1).$$

Define now

$$Y^{(n)} = n^{-2}\left(\Phi_{NRE}^{(n)} - \mathrm{E}\left[\Phi_{NRE}^{(n)}\right]\right) \quad \tilde{Y}^{(n)} = n^{-2}\left(\tilde{\Phi}^{(n)} - \mathrm{E}\left[\tilde{\Phi}^{(n)}\right]\right)$$

and using Eqs. (15) and (20) we obtain the following recursions

$$
\begin{aligned}
Y^{(n)} &= \left(\tfrac{L_n}{n}\right)^2 Y^{(L_n)} + \left(\tfrac{R_n}{n}\right)^2 Y^{(R_n)} + n^{-2}\binom{L_n}{2}T_{0.5} + n^{-2}\binom{R_n}{2}T'_{0.5} \\
&\quad + n^{-2}\left(\mathrm{E}\left[\Phi_{NRE}^{(L_n)}|L_n\right] + \mathrm{E}\left[\Phi_{NRE}^{(R_n)}|R_n\right] - \mathrm{E}\left[\Phi_{NRE}^{(n)}\right]\right)
\end{aligned}
$$

and

$$
\begin{aligned}
\tilde{Y}^{(n)} &= \left(\tfrac{L_n}{n}\right)^2 \tilde{Y}^{(L_n)} + \left(\tfrac{R_n}{n}\right)^2 \tilde{Y}^{(R_n)} + n^{-2}\binom{L_n}{2} + n^{-2}\binom{R_n}{2} \\
&\quad + n^{-2}\left(\mathrm{E}\left[\tilde{\Phi}^{(L_n)}|L_n\right] + \mathrm{E}\left[\tilde{\Phi}^{(R_n)}|R_n\right] - \mathrm{E}\left[\tilde{\Phi}^{(n)}\right]\right).
\end{aligned}
$$

23

The process $\tilde{Y}^{(n)}$ is related to the process $\tilde{W}_n$ as

$$\tilde{W}_n = 2(1 + n^{-1})\tilde{Y}^{(n)} + \binom{n}{2}^{-1} \mathrm{E}\left[\tilde{\Phi}^{(n)}\right].$$

In the continuous case we do not have an exact equality, we rather hope for

$$W_n \approx 2(1 + n^{-1})Y^{(n)} + \binom{n}{2}^{-1}\left(\mathrm{E}\left[\Phi^{(n)}\right] - \binom{n}{2}\right) + T_1$$

in some sense of approximation. Hence, knowledge of the asymptotic behaviour of $Y^{(\infty)}$, $\tilde{Y}^{(\infty)}$ will immediately give us information about $W^{(\infty)}$, $\tilde{W}^{(\infty)}$ in the obvious way

$$\begin{aligned} \tilde{W}^{(\infty)} &= 2\tilde{Y}^{(\infty)} + 2 \\ W^{(\infty)} &\approx 2Y^{(\infty)} + 1 + T_1. \end{aligned}$$

The processes $Y^{(n)}$, $\tilde{Y}^{(n)}$ look very similar to the scaled recursive representation of the Quicksort algorithm (e.g. Rösler, 1991). In fact, it is of interest that, just as in the present work, a martingale proof first showed convergence of Quicksort (Régnier, 1989), but then a recursive approach is required to show properties of the limit. The random variable $L_n/n \to \tau \sim \mathrm{Unif}[0,1]$ weakly and as weak convergence is preserved under continuous transformations (Thm. 18, p. 316 Grimmett and Stirzaker, 2009) we will have $(L_n/n)^2 \to \tau^2$ weakly. Therefore, we would expect the almost sure limits to satisfy the following equalities in distribution (remembering the asymptotic behaviour of the expectations)

$$Y^{(\infty)} \overset{\mathcal{D}}{=} \tau^2 Y'^{(\infty)} + (1 - \tau)^2 Y''^{(\infty)} + \frac{1}{2}\tau^2 T_{0.5} + \frac{1}{2}(1 - \tau)^2 T'_{0.5} - \tau(1 - \tau), \quad (22)$$

and

$$\tilde{Y}^{(\infty)} \overset{\mathcal{D}}{=} \tau^2 \tilde{Y}'^{(\infty)} + (1 - \tau)^2 \tilde{Y}''^{(\infty)} + \frac{1}{2} - 3\tau(1 - \tau) \qquad (23)$$

where $\tau$ is uniformly distributed on $[0,1]$, $Y^{(\infty)}$, $Y'^{(\infty)}$ and $Y''^{(\infty)}$ are identically distributed random variables, so are $\tilde{Y}^{(\infty)}$, $\tilde{Y}'^{(\infty)}$ and $\tilde{Y}''^{(\infty)}$, and $Y'^{(\infty)}$, $Y''^{(\infty)}$, $\tilde{Y}'^{(\infty)}$ and $\tilde{Y}''^{(\infty)}$ are independent. Following Rösler (1991)'s approach it turns out that the limiting distributions do satisfy the equalities of Eqs. (22) and (23).

24

Let $D$ be the space of distributions with zero first moment and finite second moment. We consider on $D$ the Wasserstein metric

$$d(F, G) = \inf_{X \sim F, Y \sim G} \|X - Y\|_{L^2}.$$

**Theorem 4.3 (cf. Thm.** 2.1**, Rösler (1991))** *Let $F \in D$ and assume that $Y, Y' \sim F$, $\tau \sim \mathrm{Unif}[0, 1]$, $T_{0.5}, T'_{0.5} \sim \exp(2)$ and $Y, Y', \tau, T, T'$ are all independent. Define transformations $S_1 : D \to D$, $S_2 : D \to D$ as*

$$S_1(F) = \tau^2 Y + (1 - \tau)^2 Y' + \frac{1}{2}\tau^2 T_{0.5} + \frac{1}{2}(1 - \tau)^2 T'_{0.5} - \tau(1 - \tau), \qquad (24)$$

*and*

$$S_2(F) = \tau^2 Y + (1 - \tau)^2 Y' + \frac{1}{2} - 3\tau(1 - \tau) \qquad (25)$$

*respectively. Both transformations $S_1$ and $S_2$ are contractions on $(D, d)$ and converge exponentially fast in the $d$–metric to the fixed points of $S_1$ and $S_2$ respectively.*

PROOF Let $C$ be some random variable. Then we can see that we can write both $S_1$ and $S_2$ in the form

$$S(F) = \tau^2 Y + (1 - \tau)^2 Y' + C,$$

where $S : D \to D$ and $C$ is some value which may depend on $\tau$ or other parameters/random variables. Now remembering that random variables distributed by laws from $D$ have 0 mean, it holds for $F, G \in D$, $X, X' \sim F$, $Y, Y' \sim G$

$$
\begin{aligned}
d^2(S(F), S(G)) &\leq \|\tau^2 Y + (1 - \tau)^2 Y' + C - \tau^2 X - (1 - \tau)^2 X' - C\|_{L^2}^2 \\
&= \|\tau^2(Y - X) + (1 - \tau)^2(Y' - X')\|_{L^2}^2 = \mathrm{E}\left[\tau^4\right]\mathrm{E}\left[(X - Y)^2\right] \\
&+ \mathrm{E}\left[(1 - \tau)^4\right]\mathrm{E}\left[(X' - Y')^2\right] = \tfrac{2}{5}\mathrm{E}\left[(X - Y)^2\right].
\end{aligned}
$$

Hence,

$$d(S(F), S(G)) \leq \sqrt{\frac{2}{5}}\, d(F, G).$$

25

Following Rösler (1991) the sequence $S^n(F)$ will be a Cauchy sequence in the $d$–metric as for $m \leq n$,

$$d(S^m(F), S^n(F)) \leq \sum_{j=m}^{n-1} d(S^j(F), S^{j+1}(F)) \leq \sum_{j=m}^{n-1} \left(\tfrac{2}{5}\right)^{j/2} d(F, S(F))$$
$$\leq \tfrac{5}{3} d(F, S(F)) \left(\tfrac{2}{5}\right)^{m/2}.$$

Hence, we have exponential convergence to some limit that must be a fixed point. We showed that the contraction is strict and hence, this fixed point is unique.

Hence, both $S_1$ and $S_2$ are contractions with a unique fixed point (but potentially different for the two maps). Both transformations converge exponentially fast to their fixed point.

$\square$

**Remark 4.4** *Compared to the Quicksort algorithm (Rösler, 1991) we can see that we have a $2/5$ instead of $2/3$ upper bound on the convergence rate. This speed–up should be expected as have $\tau^2$ and $(1 - \tau)^2$ instead of $\tau$ and $(1 - \tau)$.*

**Lemma 4.5** *Define for $i \in \{1, \ldots, n\}$*

$$\tilde{C}_n(i) = n^{-2} \left( \mathrm{E}\left[ \tilde{\Phi}^{(i)} \right] + \mathrm{E}\left[ \tilde{\Phi}^{(n-i)} \right] - \mathrm{E}\left[ \tilde{\Phi}^{(n)} \right] + \binom{n}{2} - i(n-i) \right)$$

*and for $x \in [0, 1]$*

$$\tilde{C}(x) = \frac{1}{2} - 3x(1-x)$$

*then*

$$\sup_{x \in [0,1]} |\tilde{C}_n(\lceil nx \rceil) - C(x)| \leq 2n^{-1} \ln n + O(n^{-1}).$$

PROOF Writing out

$$\begin{aligned}
\tilde{C}_n(i) \;=\;& n^{-2}\left(i^2 + i - 2iH_{i,1} + (n-i)^2 + (n-i) - 2(n-i)H_{n-i,1} - n^2\right.\\
& \left. -n + 2nH_{n,1} + \binom{n}{2} - i(n-i)\right)\\
\;=\;& n^{-2}\left(3i^2 - 3in + \tfrac{1}{2}n^2 + 2nH_{n,1} - \tfrac{1}{2}n - 2iH_{i,1} - 2(n-i)H_{n-i,1}\right)\\
\;<\;& \tfrac{1}{2} - 3\tfrac{i}{n}\left(1 - \tfrac{i}{n}\right) + 2n^{-1}\ln n
\end{aligned}$$

Therefore, assuming that $1 \le \lceil nx \rceil \le n-1$

$$\begin{aligned}
|\tilde{C}_n(\lceil nx \rceil) - \tilde{C}(x)| \le\;& 3\left|\tfrac{\lceil nx \rceil}{n}\left(1 - \tfrac{\lceil nx \rceil}{n}\right) - x(1-x)\right| + 2n^{-1}\ln n\\
\le\;& \sup_{|y-z|<1/n} |C(y) - C(z)| + 2n^{-1}\ln n \le \tfrac{6}{n} + 2n^{-1}\ln n + O(n^{-2}).
\end{aligned}$$

If $\lceil nx \rceil = n$, we notice that $x \in (1 - 1/n, 1]$ and directly obtain

$$|\tilde{C}_n(\lceil nx \rceil) - \tilde{C}(x)| \le 3|x(1-x)| + 2n^{-1}\ln n \le 2n^{-1}\ln n + \tfrac{3}{n}.$$

$\square$

**Lemma 4.6** *Define for $i \in \{1,\dots,n\}$, $T, T' \sim \exp(2)$*

$$C_n(i, T, T') = \frac{1}{n^2}\left(\mathrm{E}\left[\Phi_{NRE}^{(i)}\right] + \mathrm{E}\left[\Phi_{NRE}^{(n-i)}\right] - \mathrm{E}\left[\Phi_{NRE}^{(n)}\right] + \binom{i}{2}T + \binom{n-i}{2}T'\right)$$

*and for $x \in [0,1]$, $T, T' \sim \exp(2)$*

$$C(x, T, T') = \frac{1}{2}x^2 T + \frac{1}{2}(1-x)^2 T' - x(1-x)$$

*then*

$$\sup_{x \in [0,1]} |C_n(\lceil nx \rceil, T, T') - C(x, T, T')| \le n^{-1}\ln n + O(n^{-1}) + B_n,$$

*where $B_n$ is a positive random variable that converges to $0$ almost surely with expectation decaying as $O(n^{-1})$ and second moment as $O(n^{-2})$.*

27

PROOF Similarly, as in the proof of Lemma 4.5 we write out

$$
\begin{aligned}
C_n(i, T, T') &= n^{-2}\left(\binom{i}{2}T + \binom{n-i}{2}T' + \tfrac{1}{2}(i^2+i) - iH_{i,1} + \tfrac{1}{2}((n-i)^2 + (n-i)) \right. \\
&\quad \left. -(n-i)H_{n-i,1} - \tfrac{1}{2}(n^2-n) + nH_{n,1}\right) \\
&< \tfrac{1}{2}\left(\tfrac{i}{n}\right)^2 T + \tfrac{1}{2}\left(\tfrac{n-i}{n}\right)^2 T' - \tfrac{i}{n}\left(1 - \tfrac{i}{n}\right) + n^{-1}\ln n - \tfrac{1}{2}\left(\tfrac{i}{n^2}T + \tfrac{n-i}{n^2}T'\right).
\end{aligned}
$$

We denote $A_n = (1/2)\left(\frac{i}{n^2}T + \frac{n-i}{n^2}T'\right)$ and notice that it converges almost surely to 0 with $n$. Now, assuming that $1 \le \lceil nx \rceil \le n-1$

$$
\begin{aligned}
|C_n(\lceil nx \rceil) - C(x)| &\le \tfrac{1}{2}\left|\left(\tfrac{\lceil nx \rceil}{n}\right)^2 - x^2\right|T \\
&\quad + \tfrac{1}{2}\left|\left(1 - \tfrac{\lceil nx \rceil}{n}\right)^2 - (1-x)^2\right|T' + \left|\tfrac{\lceil nx \rceil}{n}\left(1 - \tfrac{\lceil nx \rceil}{n}\right) - x(1-x)\right| + n^{-1}\ln n + A_n \\
&< \sup_{|y-z|<1/n} \tfrac{1}{2}|y^2 - z^2|T + \sup_{|y-z|<1/n} \tfrac{1}{2}|y^2 - z^2|T' + \sup_{|y-z|<1/n} |y(1-y) + z(1-z)| \\
&\quad + n^{-1}\ln n + A_n \\
&\le (n^{-1} + O(n^{-2}))T + (n^{-1} + O(n^{-2}))T' + \tfrac{2}{n} + O(n^{-2}) + n^{-1}\ln n + A_n.
\end{aligned}
$$

If $\lceil nx \rceil = n$, we notice that $x \in (1 - 1/n, 1]$ and directly obtain

$$
|C_n(\lceil nx \rceil) - C(x)| \le \tfrac{1}{2}n^{-2}T + \tfrac{1}{2}n^{-2}T' + n^{-1} + n^{-1}\ln n + A_n.
$$

Therefore, if we now denote

$$
B_n = A_n + (n^{-1} + O(n^{-2}))T + (n^{-1} + O(n^{-2}))T'
$$

we obtain the statement of the Lemma.

$\square$

We now turn to showing that $Y^{(n)}$ and $\tilde{Y}^{(n)}$ converge in the Wasserstein $d-$metric to $Y^{(\infty)}$ and $\tilde{Y}^{(\infty)}$ whose laws are fixed points of $S_1$ and $S_2$ respectively.

**Definition 4.7** *Define the maps* $M_1, M_2 : \bigcup\limits_{n=1}^{\infty} D^n \to D$ *as*

$$
\begin{aligned}
M_1(G_1, \ldots, G_{n-1}) &= \mathcal{L}\left(\left(\tfrac{L_n}{n}\right)^2 Y^{(L_n)} + \left(\tfrac{n-L_n}{n}\right)^2 Y'^{(n-L_n)} + n^{-2}\binom{L_n}{1}T_{0.5}\right. \\
&\quad \left. + n^{-2}\binom{n-L_n}{n}T'_{0.5} + n^{-2}\left(\mathrm{E}\left[\Phi^{(L_n)}_{NRE}|L_n\right] + \mathrm{E}\left[\Phi^{(n-L_n)}_{NRE}|L_n\right] - \mathrm{E}\left[\Phi^{(n)}_{NRE}\right]\right)\right),
\end{aligned}
$$

$$
(26)
$$

*and*

$$M_2\left(G_1, \ldots, G_{n-1}\right) = \mathcal{L}\left(\left(\tfrac{L_n}{n}\right)^2 \tilde{Y}^{(L_n)} + \left(\tfrac{n-L_n}{n}\right)^2 \tilde{Y}'^{(n-L_n)} + n^{-2}\binom{L_n}{1}\right.$$
$$\left. + n^{-2}\binom{n-L_n}{n} + n^{-2}\left(\mathrm{E}\left[\tilde{\Phi}^{(L_n)}|L_n\right] + \mathrm{E}\left[\tilde{\Phi}^{(n-L_n)}|L_n\right] - \mathrm{E}\left[\tilde{\Phi}^{(n)}\right]\right)\right), \tag{27}$$

*where by $\mathcal{L}(X)$ we denote the law of the random variable $X$.*

For every $G \in D$ we are interested in two sequences

$$G_1^{(1)} = G, \quad G_2^{(1)} = M_1(G_1^{(1)}), \quad G_3^{(1)} = M_1(G_1^{(1)}, G_2^{(1)}), \quad \ldots \tag{28}$$

and

$$G_1^{(2)} = G, \quad G_2^{(2)} = M_2(G_1^{(2)}), \quad G_3^{(2)} = M_2(G_1^{(2)}, G_2^{(2)}), \quad \ldots \tag{29}$$

We obtain similarly to Rösler (1991)'s Thm. 3.1

**Theorem 4.8** *Let $G$ correspond to the point measure on $0$, i.e.*

$$G(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

*Then $G_n^{(1)}$ and $G_n^{(2)}$, as defined in Eqs. (28) and (29) converge in the Wasserstein $d$–metric to the unique fixed points of $S_1$ and $S_2$ respectively.*

PROOF We write the proof for $G_n^{(1)}$ as for $G_n^{(2)}$ it will be nearly identical. We take $Y^{(\infty)}$ and $Y'^{(\infty)}$ independent and distributed as $F$, the fixed point of $S_1$. Then, for $i = 1, \ldots, n-1$ we choose independent versions of $Y^{(i)}$ and $Y'^{(i)}$. We define

$$V_x = \sum_{i=1}^{n} 1_{(i-1)/n < x \leq i/n} Y^{(i-1)}$$

and

$$V'_x = \sum_{i=1}^{n} 1_{(i-1)/n < x \leq i/n} Y'^{(i-1)}.$$

Now, for independent random variables $\tau \sim \mathrm{Unif}[0,1]$, $T_{0.5}, T'_{0.5} \sim \exp(2)$

29

$$
\begin{aligned}
d^2(G_n^{(1)}, F) \leq \ & \mathrm{E}\left[\left(\left(\tfrac{\lceil n\tau\rceil-1}{n}\right)^2 V_\tau - \tau^2 Y^{(\infty)} + \left(\tfrac{n-\lceil n\tau\rceil-1}{n}\right)^2 V_\tau' - (1-\tau)^2 Y'^{(\infty)}\right.\right. \\
& \left.+ C_n(\lceil n\tau\rceil, T_{0.5}, T_{0.5}') - C(\tau, T_{0.5}, T_{0.5}'))^2\right] \\
\leq \ & 2\,\mathrm{E}\left[\left(\left(\tfrac{\lceil n\tau\rceil-1}{n}\right)^2 V_\tau - \tau^2 Y^{(\infty)}\right)^2\right] + 2\,\mathrm{E}\left[\left(\left(\tfrac{n-\lceil n\tau\rceil-1}{n}\right)^2 V_\tau' - (1-\tau)^2 Y'^{(\infty)}\right)^2\right] \\
& + 2\,\mathrm{E}\left[(C_n(\lceil n\tau\rceil, T_{0.5}, T_{0.5}') - C(\tau, T_{0.5}, T_{0.5}'))^2\right].
\end{aligned}
$$

We consider the first term of the right–hand side of the inequality.

$$
\begin{aligned}
\mathrm{E}\left[\left(\left(\tfrac{\lceil n\tau\rceil-1}{n}\right)^2 V_\tau - \tau^2 Y^{(\infty)}\right)^2\right] &= \mathrm{E}\left[\sum_{i=1}^n 1_{(i-1)/n < \tau \leq i/n}\left(\left(\tfrac{i-1}{n}\right)^2 Y^{(i-1)} - \tau^2 Y^{(\infty)}\right)^2\right] \\
&\leq \tfrac{1}{n}\sum_{i=1}^n \left(\tfrac{n-i}{n}\right)^4 \mathrm{E}\left[\left(Y'^{(n-i)} - Y'^{(\infty)}\right)^2\right] \to 0.
\end{aligned}
$$

The last convergence to 0 is a direct generalization of Rösler (1991)'s Prop. 3.3. Applying the same reasoning to the second term and invoking Lemma 4.6 we have

$$
\begin{aligned}
d^2(G_n^{(1)}, F) \ \leq \ & \tfrac{2}{n}\sum_{i=1}^n \left(\tfrac{i-1}{n}\right)^4 \mathrm{E}\left[\left(Y^{(i-1)} - Y^{(\infty)}\right)^2\right] + \tfrac{2}{n}\sum_{i=1}^n \left(\tfrac{i-1}{n}\right)^4 \mathrm{E}\left[\left(Y^{(i-1)} - Y^{(\infty)}\right)^2\right] \\
& + 2n^{-2}\ln^2 n + O(n^{-2}) + 2\,\mathrm{E}\left[B_n^2\right] \to 0.
\end{aligned}
$$

$\square$

**Remark 4.9** *One may directly obtain from the recursive representation that* $\mathrm{E}\left[Y^{(\infty)}\right] = E\tilde{Y}^{(\infty)} = 0$, $\mathrm{Var}\left[Y^{(\infty)}\right] = 1/16 = 0.0635$ *and* $\mathrm{Var}\left[\tilde{Y}^{(\infty)}\right] = 1/12$. *We can therefore, see that in the discrete case the variance agrees. However, in the continuous case we can see that it slightly differs*

$$
\mathrm{Var}\left[(W_n - T_1)/2\right] = \pi^2/18 - 0.5 \approx 0.048.
$$

**Remark 4.10** *One can of course calculate what the mean and variance of* $T_{0.5}$, $T_{0.5}'$ *should be so that* $\mathrm{E}\left[Y^{(\infty)}\right] = 0$ *and* $\mathrm{Var}\left[Y^{(\infty)}\right] = \mathrm{Var}\left[(W_n - T_1)/2\right]$. *We should have* $\mathrm{E}\left[T_{0.5}\right] = \mathrm{E}\left[T_{0.5}'\right] = 0.5$ *and* $\mathrm{Var}\left[T_{0.5}\right] = \mathrm{Var}\left[T_{0.5}'\right] = \pi^2/3 - 25/8$. *This, in particular, means that these branch lengths cannot be exponential. We therefore also experimented by drawing* $T_{0.5}$, $T_{0.5}'$ *from a gamma*

30

*distribution with rate equalling $1/(2(\pi^2/3 - 25/8))$ and shape equalling $\pi^2/6 - 25/16$. However, this increased the duration of the computations about $4.5$ times and did not result in any visible improvements in comparison to Table 1.*

# 5 Significance testing

One of the main motivations to undertake the study of the limiting behaviour of the cophenetic index is to construct a statistic based on it that will allow for testing if an observed phylogenetic tree is consistent with the pure birth process. Sackin's and Colless' indices have been studied with respect to their limiting distributions (Blum and François, 2005; Blum et al., 2006). In particular it was shown that the study of Sackin's index is equivalent to studying the Quicksort distribution (Blum and François, 2005). Alternatively McKenzie and Steel (2000) proposed to measure balance by counting cherries on the tree, this index after appropriate centring and scaling converges to the standard normal (McKenzie and Steel, 2000). However, the cophenetic index might be a better candidate as it has a higher resolution and can handle non–binary trees (Mir et al., 2013). The analysis done in this work clearly indicates that a normal approximation is not appropriate. The tail of the scaled cophenetic index is much heavier and using normal quantiles will result in wrong significance levels of a test.

Unfortunately an analytical form of the density of any scaled cophenetic index is not known so one will have to resort to simulations. Directly simulating a large number of pure birth trees can take an overly long time, measured in hours. In Tab. 1 we only report the time needed to simulate the sample of Yule trees, as we used a suboptimal $O(n^2)$ algorithm to obtain the cophenetic indices. The cophenetic index can be calculated in $O(n)$ time (Corr. 3 Mir et al., 2013). In order to speed up the computations we use the same set of simulated Yule trees for both the continuous and discrete cases. On the other hand, the suggestive (but wrong) approximations of Eq. (6) and contraction limiting distributions Eqs. (22) and (23) are significantly faster to simulate, see Tabs. 1 and 2. Therefore, these should methods should appeal to applied researchers who want to use this index for testing consistency with the pure–birth process.

Simulating from the approximate Eq. (6) is straightforward. One just draws $n - 1$ independent exponential 1 random variables. Simulating ran-

dom variables satisfying Eqs. (22) and (23) is more involved and probably an exact rejection algorithm can be developed (cf. Devroye et al., 2000). Here we choose simple, approximate but still effective, heuristics in order to demonstrate the usefulness of the approach for significance testing.

We describe the algorithms for simulating from a more general distribution, $F$, that satisfies

$$Y \overset{\mathcal{D}}{=} g_1(\tau)Y' + g_2(\tau)Y'' + C(\tau, \theta),$$

where $Y, Y', Y'' \sim F$, $Y', Y'', \tau, \theta$ are independent, $\tau \sim F_\tau$, $\theta \sim F_\theta$ is some random vector, $g_1, g_2 : \mathbb{R} \to \mathbb{R}$ and $C : \mathbb{R}^p \to \mathbb{R}$ for some appropriate $p$ that depends on $\theta$'s dimension. Of course for our case of $\Phi^{(n)}$, $\tilde{\Phi}^{(n)}$ we have $\tau \sim$ Unif$[0,1]$, $g_1(\tau) = \tau^2$, $g_2(\tau) = (1-\tau)^2$, $C(\tau, T, T') = \tau^2 T + (1-\tau)^2 T' - \tau(1-\tau)$ and $C(\tau) = 1/2 - 3\tau(1-\tau)$ for $\Phi^{(n)}$, $\tilde{\Phi}^{(n)}$ respectively. Of course, $T$, $T'$ are independent and exponential 2 distributed. If one considers also the root edge, then to the simulated random variable one needs to add $T_1 \sim \exp(1)$ when simulating $n^{-2}\Phi^{(n)}$ or appropriately 1 if one considers $n^{-2}\tilde{\Phi}^{(n)}$.

---

**Algorithm 1** Population approximation

---

1: Initiate population size $N$
2: Set $P[0, 1 : N] = Y_0$             ▷ Initial population
3: **for** $i = 1$ to $i_{max}$ **do**
4:     $f_{i-1} =$ `density`$(P[i-1,])$         ▷ density estimation by R
5:     **for** $j = 1$ to $N$ **do**
6:        Draw $\tau$ from $F_\tau$
7:        Draw $\theta$ from $F_\theta$
8:        Draw $Y_1, Y_2$ independently from $f_{i-1}$
9:        $P[i, j] = g_1(\tau)Y_1 + g_2(\tau)Y_2 + C(\tau, \theta)$
10:     **end for**
11: **end for**
12: **return** $P[i_{max},]$
13:           ▷ Add root branch (exp(1) or 1) if needed for each individual.

---

The recursion of Alg. 2 for a given realization of $\tau$ and $\theta$ random variables can be directly solved. In the case of $Y_0 = 0$ it will equal (in a directly implementable representation)

---

**Algorithm 2** Recursive approximation

1: **procedure** YRECURSION($n$, $Y_0$)
2:     **if** $n = 0$ **then**
3:         $Y_1 = Y_0$, $Y_2 = Y_0$
4:     **else if** $n = 1$ **and** $Y_0 = 0$ **then**
5:         Draw $\tau_1$, $\tau_2$ independently from $F_\tau$
6:         Draw $\theta_1$, $\theta_2$ independently from $F_\theta$
7:         $Y_1 = C(\tau_1, \theta_1)$
8:         $Y_2 = C(\tau_2, \theta_2)$
9:     **else**
10:         $Y_1 =$YRECURSION($n - 1$ , $Y_0$)
11:         $Y_2 =$YRECURSION($n - 1$ , $Y_0$)
12:     **end if**
13:     Draw $\tau$ from $F_\tau$
14:     Draw $\theta$ from $F_\theta$
15:     **return** $g_1(\tau)Y_1 + g_2(\tau)Y_2 + C(\tau, \theta)$,
16: **end procedure**
17: **return** YRECURSION($N$, $Y_0$)
18:                                         ▷ Add root branch (exp(1) or 1) if needed.

---

$$
Y^{(N)} = C(\tau_{0,0}, \theta_{0,0})
$$
$$
+ \sum_{j=1}^{N-1} \left( \sum_{i=0, i \text{ even}}^{2^j-1} \left( C(\tau_{j,i}, \theta_{j,i}) \cdot g_1(\tau_{0,0})Y_0 \cdot \prod_{k=1}^{j-1} \left( g_1(\tau_{k,i[0:(k-1)]})^{(i[k]+1)_2} \cdot g_2(\tau_{k,i[1:k]+1})^{i[k]} \right) \right) \right.
$$
$$
\left. + \sum_{i=0, i \text{ odd}}^{2^j-1} \left( C(\tau_{j,i}, \theta_{j,i}) \cdot g_2(\tau_{0,0})Y_0 \cdot \prod_{k=1}^{j-1} \left( g_1(\tau_{k,i[0:(k-1)]})^{(i[k]+1)_2} \cdot g_2(\tau_{k,i[1:k]+1})^{i[k]} \right) \right) \right),
$$

(30)

where $i[k]$ is the value (either 0 or 1) of the bit corresponding to $2^k$ in $i$'s binary representation and $i[0 : (k-1)]$ is the value obtained when only the $k$ youngest bits are taken from $i$'s binary representation. For different $i, j \in \{0, \ldots, N-1\}$ pairs, $\tau_{i,j} \sim F_\tau$ and $\theta_{i,j} \sim F_\theta$ are independent. The operation $(\cdot)_2$ means taking the value modulo 2. If $Y_0 \neq 0$, then the formula will be very similar except lengthier. However, in our case taking $Y_0 = 0$ is reasonable as the cophenetic index is 0 for trees with 0, 1 and 2 tips. The reader should remember that we do not have a leading root branch in the contraction setup, we correct for it later if needed.

33

In Tabs. 1 and 2 we also compare the quantiles from the different distributions. We can see that the approximation of $\overline{W}_n$ for $W_n$ is a very good one and can be used when one needs to work with the distribution of the cophenetic index with branch lengths. In the case of the discrete cophenetic index we have found an exact limit distribution which is a contraction–type distribution. Therefore, one can relatively quickly simulate a sample from it without the need to do lengthy simulations of the whole tree and then calculations of the cophenetic index. Unfortunately this contraction approach does not seem to give such good results in the Yule tree with branch lengths case. We used an approximation when constructing the contraction. Instead of taking the law of the length of two daughter branches, we took the law of an random internal branch. This induces a difference between the tails of the distributions that is clearly visible in the simulations. Even at the second moment level there is a large difference. We calculated (Thm. 3.8) that $\mathrm{Var}\left[W_n\right] \to 2\pi^2/9 - 1 \approx 1.193$, $\mathrm{Var}\left[\overline{W}_n\right] \to 4\pi^2/3 - 12 \approx 1.159$ while $\mathrm{Var}\left[2Y^{(n)} + T_1\right] = 1.25$. Therefore, the approximation by $\overline{W}_n$ seems better even at the second moment level. Generally if one cannot afford the time and memory to simulate a large sample of Yule tree, simulating $\overline{W}_n$ values seems a very attractive option, as the discrepancy between the two distributions seems very small.

In Fig. 6 we compare the histograms (and density estimates) of (scaled and centred) both continuous and discrete branches cophenetic indices and their respective contraction–type limit distributions. The histograms generally agree but we know from Tab. 1 that for $\Phi^{(n)}$ this is only an approximation. We simulated 10000 Yule trees and hence the 0.0005 upper and lower quantile estimates are very inaccurate. This is especially visible in Tab. 2. All the quantiles, except the 0.9995 one agree with the quantiles from the Yule tree simulation. We should expect this as for $\tilde{\Phi}^{(n)}$ we have shown an exact limit distribution.

It is also encouraging that all three proposed heuristics of simulating from the contraction–type limit give similar results even with as few as 10 iterations. Simulating with Alg. 1 is the fastest, however the results also seem to be the least accurate. This could certainly be due to the dependencies building up in the population after each generation. Simulating by Alg. 2 seems to give the best results (out of the three proposed methods) and is still very fast, especially compared to using Eq. (30). Hence, if the distribution of the cophenetic index (discrete branches) is required, then there is no need to

simulate Yule trees. The recursion–type limit distribution is an exact result and using Alg. 2 with even a recursion depth of $n = 10$ will yield a useful sample. Of course paying with time for a deeper recursion will give even more accuracy.
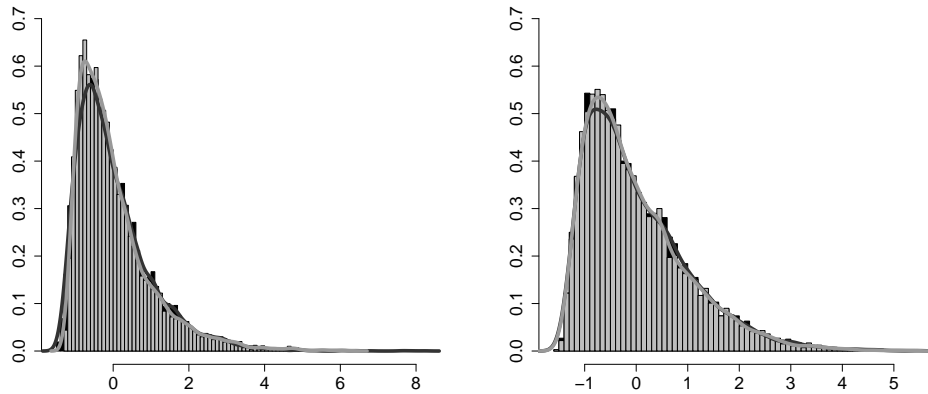


Figure 6: Histogram and density estimates of scaled (by theoretical standard deviation) and centred (by theoretical expectation) cophenetic indices (black) for 10000 simulated 500 tip Yule trees with $\lambda = 1$ and of simulation by Alg. 2 (gray), also scaled and centred to mean 0 and variance 1. Left: histogram of $\Phi^{(n)}$, right: $\tilde{\Phi}^{(n)}$. The curves are based on R's `density()` function.

# 6    Alternative descriptions

## 6.1    Difference process

Let us consider in detail the families of random variables $V_i^{(n)}$ and $\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]$. Obviously $V_i^{(n)}$ is $\binom{n}{2}i$ times the number of pairs that coalesced after the $i-1$ speciation event for a given Yule tree. Denote

$$A_i^{(n)} := iV_i^{(n)}.$$

As going from $n$ to $n + 1$ means a new speciation event and coalescent at this new $n$th event, then

35

$$\left(\sqrt{\mathrm{Var}\left[\Phi^{(n)}\right]}\right)^{-1}\left(\Phi^{(n)} - \mathrm{E}\left[\Phi^{(n)}\right]\right) \text{ limit approximation}$$

| | Yule | $\mathcal{N}(0,1)$ | $\overline{W}_N$ | $Y^{(N)}$ Alg. 1 | $Y^{(N)}$ Alg. 2 | $Y^{(N)}$ Eq. (30) |
|---|---|---|---|---|---|---|
| Running time | 4.458h | — | 15.877s | 1.661s | 8.617m | 46.018m |
| Average $(= 0)$ | 0.021 | 0 | $-0.003$ | 0.014 | 0.019 | $-0.001$ |
| Variance $(= 1)$ | 1.018 | 1 | 0.929 | 1.024 | 1.023 | 0.988 |
| Skewness | 1.609 | 0 | 1.634 | 1.836 | 1.778 | 1.865 |
| Excess kurtosis | 4.237 | 0 | 4.159 | 5.013 | 4.308 | 5.54 |
| $q(0.0005)$ | $-1.462$ | $-3.291$ | $-1.449$ | $-1.271$ | $-1.252$ | $-1.248$ |
| $q(0.001)$ | $-1.431$ | $-3.090$ | $-1.420$ | $-1.245$ | $-1.239$ | $-1.235$ |
| $q(0.005)$ | $-1.337$ | $-2.576$ | $-1.330$ | $-1.181$ | $-1.194$ | $-1.191$ |
| $q(0.01)$ | $-1.285$ | $-2.326$ | $-1.281$ | $-1.153$ | $-1.165$ | $-1.154$ |
| $q(0.025)$ | $-1.199$ | $-1.96$ | $-1.183$ | $-1.094$ | $-1.097$ | $-1.093$ |
| $q(0.05)$ | $-1.113$ | $-1.644$ | $-1.095$ | $-1.028$ | $-1.031$ | $-1.031$ |
| $q(0.95)$ | 1.995 | 1.644 | 1.855 | 2.004 | 2.052 | 1.962 |
| $q(0.975)$ | 2.658 | 1.96 | 2.482 | 2.687 | 2.773 | 2.61 |
| $q(0.99)$ | 3.326 | 2.326 | 3.366 | 3.593 | 3.574 | 3.587 |
| $q(0.995)$ | 3.935 | 2.576 | 4.056 | 4.359 | 4.393 | 4.211 |
| $q(0.999)$ | 5.342 | 3.090 | 5.161 | 5.943 | 5.576 | 5.580 |
| $q(0.9995)$ | 6.423 | 3.291 | 5.857 | 6.361 | 5.885 | 6.183 |

Table 1: Simulations based on 10000 draws of each random variable (population size for Alg. 1) i.e. columns, bar $\mathcal{N}(0,1)$. The rows $q(\alpha)$ correspond to the, simulated, bar $\mathcal{N}(0,1)$, quantiles i.e. for a random variable $X$, $P(X \leq q(\alpha)) = \alpha$. All simulations were done in R with the package TreeSim (Stadler, 2009, 2011) used to obtain the Yule trees. All results correspond to Yule trees with speciation rate $\lambda = 1$ and $n = 500$ tips. The Yule tree $\Phi^{(n)}$ values are centred and scaled by expectation and standard deviation from Eq. (15). $\overline{W}_n$ is centred by $\mathrm{E}[W_n]$ and scaled by $\sqrt{(2\pi^2 - 9)/9}$ (Thm. 3.8) $Y^{(\infty)}$ is centred by 1 (root branch), scaled by $\sqrt{1 + 1/16}$. $N = 10$ for Algs. 1, 2 and Eq. (30) is the number of generations and recursion depth of the respective algorithm. In Alg. 1 the initial population is set at 0 and also $Y_0 = 0$ for Alg. 2 and Eq. (30). The simulations were run in R 3.2.5 for Ubuntu 12.04.5 LTS on a 1.4GHz. AMD Opteron Proc. 6274

$$A_i^{(n+1)} \geq \binom{n+1}{2}^{-1}\left(\binom{n}{2}A_i^{(n)} + 1\right).$$

$$\left(\sqrt{\mathrm{Var}\left[\tilde{\Phi}^{(n)}\right]}\right)^{-1} (\tilde{\Phi}^{(n)} - \mathrm{E}\left[\tilde{\Phi}^{(n)}\right]) \text{ limit approximation}$$

|  | Yule | $\mathcal{N}(0,1)$ | $\tilde{Y}^{(N)}$ Alg. 1 | $\tilde{Y}^{(N)}$ Alg. 2 | $\tilde{Y}^{(N)}$ Eq. (30) |
|---|---|---|---|---|---|
| Running time | — | — | 1.384s | 3.141m | 38.089m |
| Average $(=0)$ | 0.011 | 0 | $-0.014$ | $-0.008$ | 0.003 |
| Variance $(=1)$ | 1.013 | 1 | 1.004 | 0.998 | 0.003 |
| Skewness | 1.206 | 0 | 1.289 | 1.279 | 1.003 |
| Excess kurtosis | 1.714 | 0 | 2.142 | 2.103 | 1.227 |
| $q(0.0005)$ | $-1.478$ | $-3.291$ | $-1.499$ | $-1.458$ | $-1.454$ |
| $q(0.001)$ | $-1.442$ | $-3.090$ | $-1.475$ | $-1.435$ | $-1.436$ |
| $q(0.005)$ | $-1.372$ | $-2.576$ | $-1.389$ | $-1.366$ | $-1.371$ |
| $q(0.01)$ | $-1.327$ | $-2.326$ | $-1.339$ | $-1.326$ | $-1.322$ |
| $q(0.025)$ | $-1.255$ | $-1.96$ | $-1.260$ | $-1.25$ | $-1.249$ |
| $q(0.05)$ | $-1.166$ | $-1.644$ | $-1.184$ | $-1.159$ | $-1.160$ |
| $q(0.95)$ | 1.961 | 1.644 | 1.913 | 1.933 | 1.951 |
| $q(0.975)$ | 2.496 | 1.96 | 2.535 | 2.463 | 2.472 |
| $q(0.99)$ | 3.167 | 2.326 | 3.261 | 3.152 | 3.213 |
| $q(0.995)$ | 3.649 | 2.576 | 3.813 | 3.684 | 3.600 |
| $q(0.999)$ | 4.642 | 3.090 | 4.682 | 4.899 | 4.627 |
| $q(0.9995)$ | 4.734 | 3.291 | 5.026 | 5.158 | 4.954 |

Table 2: Simulations based on 10000 draws of each random variable (population size for Alg. 1) i.e. columns, bar $\mathcal{N}(0,1)$. The rows $q(\alpha)$ correspond to the, simulated, bar $\mathcal{N}(0,1)$, quantiles i.e. for a random variable $X$, $P(X \leq q(\alpha)) = \alpha$. All simulations were done in R with the package TreeSim (Stadler, 2009, 2011) used to obtain the Yule trees. All results correspond to Yule trees with speciation rate $\lambda = 1$ and $n = 500$ tips. The Yule tree $\tilde{\Phi}^{(n)}$ values are centred and scaled by expectation and standard deviation from Eq. (20). $\tilde{Y}^{(\infty)}$ is scaled by $(2\sqrt{3})^{-1}$. $N = 10$ for Algs. 1, 2 and Eq. (30) is the number of generations and recursion depth of the respective algorithm. In Alg. 1 the initial population is distributed set at 0 and also $Y_0 = 0$ for Alg. 2 and Eq. (30). The simulations were run in R 3.2.5 for Ubuntu 12.04.5 LTS on a 1.4GHz. AMD Opteron Proc. 6274

We also know by previous calculations that

$$\mathrm{E}\left[A_i^{(n)}\right] = i\,\mathrm{E}\left[V_i^{(n)}\right] = 2(n-i)/((n-1)(i+1)) \to 2/(i+1).$$

37

Let $\binom{n+1}{2}\epsilon_i^{(n)}$ denote the number of newly introduced coalescent events after the $(i-1)$–one when we go from $n$ to $n+1$ species. Obviously $\epsilon_i^{(n)} > \binom{n+1}{2}$ Then, we may write

$$A_i^{(n+1)} = \binom{n+1}{2}^{-1}\binom{n}{2}A_i^{(n)} + \epsilon_i^{(n)}.$$

Now,

$$\mathrm{E}\left[\epsilon_i^{(n)}\right] = \mathrm{E}\left[A_i^{(n+1)}\right] - \binom{n+1}{2}^{-1}\binom{n}{2}\mathrm{E}\left[A_i^{(n)}\right] = \frac{2(n+1-i)}{n(i+1)} - \frac{n(n-1)}{n(n+1)}\frac{2(n-i)}{(n-1)(i+1)}$$
$$= \frac{2}{i+1}\left(\frac{n+1-i}{n} - \frac{n-i}{n+1}\right) = \frac{2}{i+1}\frac{(n-i+1)(n+1)-n(n-i)}{n(n+1)} = \frac{2}{i+1}\frac{n(n-i)+n+n-i+1-n(n-i)}{n(n+1)}$$
$$= \frac{2}{i+1}\frac{2n+1-i}{n(n+1)} \to 0.$$

Therefore, for every $i$, $\epsilon_i^{(n)} \to 0$ almost surely as it is a positive random variable whose expectation goes to 0. However, $A_i^{(n)}$ is bounded by 1, as it can be understood as the conditional (on tree) cumulative density function for the random variable $\kappa$—at which speciation event did a random pair of tips coalesce, i.e. for all $i = 1, \ldots, n-1$

$$P(\kappa \leq i - 1|\mathcal{Y}_n) = 1 - A_i^{(n)}.$$

Therefore, as $A_i^{(n)}$ is bounded by 1 and the difference process

$$A_i^{(n)} - A_i^{(n-1)} = \epsilon_i^{(n)}$$

goes almost surely to 0 we may conclude that $A_i^{(n)}$ converges almost surely to some random variable $A_i$. In particular, this implies the almost sure convergence of $V_i^{(n)}$ to a limiting random variable $V_i$. Furthermore, as $\mathrm{E}\left[\sum_{i=1}^{n-1} V_i^{(n)}\right]$ and $\mathrm{Var}\left[\sum_{i=1}^{n-1} V_i^{(n)}\right]$ are both $O(1)$ we may conclude that $\sum_{i=1}^{n-1} V_i^{(n)}$ also converges almost surely. This means that the discrete version (all $T_i = 1$, corresponding to $\tilde{\Phi}^{(n)}$) of the cophenetic index converges almost surely (compare with Thm. 4.2).

## 6.2 Polyá urn description

The cophenetic index both in the discrete and continuous version has the following Polyá urn description. We start with an urn filled with $n$ balls.

38

Each ball has a number painted on it, 0 initially. At each step we remove a pair of balls, say with numbers $x$ and $y$ and return a ball with the number $(x+1)(y+1)$ painted on it. We stop when there is only one ball, it will have value $\binom{n}{2}$. Denote $B_{k,i,n}$ as the value painted on the $k$–th ball in the $i$–th step when we initially started with $n$ balls. Then we can represent the cophenetic index as

$$\Phi^{(n)} = \sum_{i=1}^{n-1} \left( \sum_{k=1}^{i} B_{k,i,n} \right) T_i$$

or alternatively, in the discrete case

$$\tilde{\Phi}^{(n)} = \sum_{i=1}^{n-1} \sum_{k=1}^{i} B_{k,i,n}.$$

# Acknowledgments

# References

K. Bartoszek. Quantifying the effects of anagenetic and cladogenetic evolution. *Math. Biosci.*, 254:42–57, 2014.

K. Bartoszek. A central limit theorem for punctuated equilibrium. *ArXiv e-prints*, 2016.

K. Bartoszek and S. Sagitov. A consistent estimator of the evolutionary rate. *J. Theor. Biol.*, 371:69–78, 2015a.

K. Bartoszek and S. Sagitov. Phylogenetic confidence intervals for the optimal trait value. *J. App. Prob.*, 52:1115–1132, 2015b.

M. G. B. Blum and O. François. On statistical tests of phylogenetic tree imbalance: The Sackin and other indices revisited. *Math. Biosci.*, 195:141–153, 2005.

M. G. B. Blum, O. François, and S. Janson. The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *Ann. Appl. Probab.*, 16(4):2195–2214, 2006.

G. Cardona, A. Mir, and F. Rosselló. Exact formulas for the variance of several balance indices under the Yule model. *J. Math. Biol.*, 67:1833–1846, 2013.

D. H. Colless. Review of "Phylogenetics: the theory and practise of phylogenetic systematics". *Syst. Zool.*, 31:100–104, 1982.

L. Devroye, J. A. Fill, and R. Neininger. Perfect simulation from the Quicksort limit distribution. *Electronic Comm. Probab.*, 5(12):95–99, 2000.

J. A. Fill and S. Janson. Smoothness and decay properties of the limiting Quicksort density function. In D. Gardy and A. Mokkadem, editors, *Mathematics and Computer Science: Algorithms, Trees, Combinatorics and Probabilities*, Trends in Mathematics, pages 53–64. Birkhäuser, Basel, 2000.

J. A. Fill and S. Janson. Approximating the limiting Quicksort distribution. *Rand. Struct. Alg.*, 19(3-4):376–406, 2001.

G. Grimmett and D. Stirzaker. *Probability and Random Processes (Third Edition)*. Oxford University Press, Oxford, 2009.

S. Janson. On the tails of the limiting Quicksort distribution. *Electronic Comm. Probab.*, 81:1–7, 2015.

A. McKenzie and M. Steel. Distributions of cherries for two models of trees. *Math. Biosci.*, 164:81–92, 2000.

A. Mir, F. Rosselló, and L. Rotger. A new balance index for phylogenetic trees. *Math. Biosci.*, 241(1):125–136, 2013.

M. Régnier. A limiting distribution for Quicksort. *Theor. Inf. Applic.*, 23 (3):335–343, 1989.

A. Rosalsky and M. Sreehari. On the limiting behavior of randomly weighted partial sums. *Stat. & Prob. Lett.*, 40:403–410, 1998.

U. Rösler. A limit theorem for "Quicksort". *Theor. Inf. Applic.*, 25(1): 85–100, 1991.

M. J. Sackin. "Good" and "bad" phenograms. *Syst. Zool.*, 21:225–226, 1972.

S. Sagitov and K. Bartoszek. Interspecies correlation for neutrally evolving traits. *J. Theor. Biol.*, 309:11–19, 2012.

T. Stadler. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *J. Theor. Biol.*, 261(1):58–68, 2009.

T. Stadler. Simulating trees with a fixed number of extant species. *Syst. Biol.*, 60(5):676–684, 2011.

T. Stadler and M. Steel. Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. *J. Theor. Biol.*, 297:33–40, 2012.

M. Steel and A. McKenzie. Properties of phylogenetic trees generated by Yule–type speciation models. *Math. Biosci.*, 170:91–112, 2001.

# Appendix: Mathematica code for Section 3

```
(*
Mathematica code used to obtain the closed form formulae
of Section 3. Second order properties in K. Bartoszek
Exact and approximate limit behaviour of the Yule tree's
cophenetic index.
The script was run using Mathematica 9.0 for Linux x86
(64-bit) running on Ubuntu 12.04.5 LTS. It has to be noted
that Mathematica's output should be manually postprocessed
in order to have the formulae in terms of harmonic sums and
not derivatives of polygamma functions.
All the references in this script point to appropriate fragments
of the manuscript.
*)


(*
We choose the pairs in order, i.e. first the first pair to
coalesce then the second pair to coalesce.
*)
(* Compare with proof of Lemma 1 of Bartoszek and Sagitov (2015b) *)
FcoalProb[n_,k_,c_]=FullSimplify[Product[(1-c/((r*(r-1))/2))
,{r,k+2,n}]]

(* Def. 2.8, Eq. (2) *)
E1k[n_,k_]:=(2*(n+1)/((n-1)*(k+1)*(k+2)))
(* Lemma 3.1, Eq. (6) *)
Var1k[n_,k_]:=(E1k[n,k]-E1k[n,k]*E1k[n,k])
(* Lemma 3.2, Eq. (7) *)
Cov1k11k2[n_,k1_,k2_]:=(-E1k[n,k1]*E1k[n,k2])


(* Lemma 3.3, Eq. (8) *)
VarE1k[n_,k_]=(FullSimplify[
(1/(n*(n-1)/2))*(2*(n+1)/((n-1)*(k+1)*(k+2)))
+
```

42

```
(2*(n−2)/(n*(n−1)/2))*(Sum[FcoalProb[n,j,3]*(1/((j+1)*j/2))
*FcoalProb[j,k,1]*(1/((k+1)*k/2)),{j,k+1,n−1}])
+
((n−2)*(n−3)/2/(n*(n−1)/2))*(
Sum[Sum[FcoalProb[n,j1,6]*(4/((j1+1)*j1/2))
*FcoalProb[j1,j2,3]*(1/((j2+1)*j2/2))
*FcoalProb[j2,k,1]*(1/((k+1)*k/2)),{j1,j2+1,n−1}],{j2,k+1,n−2}]
)− (E1k[n,k])*(E1k[n,k])])

(* Lemma 3.4, Eq. (9) *)
CovE1k1E1k2[n_,k1_,k2_]=(FullSimplify[
(2*(n−2)/(n*(n−1)/2))*(FcoalProb[n,k2,3]*(1/((k2+1)*k2/2))
*FcoalProb[k2,k1,1]*(1/((k1+1)*k1/2)))
+
((n−2)*(n−3)/2/(n*(n−1)/2))*(FcoalProb[n,k2,6]
*(1/((k2+1)*k2/2))*FcoalProb[k2,k1,3]*(1/((k1+1)*k1/2))
+
Sum[FcoalProb[n,k2,6]*(1/((k2+1)*k2/2))*FcoalProb[k2,j,3]
*(2/((j+1)*j/2))*FcoalProb[j,k1,1]*(1/((k1+1)*k1/2)),{j,k1+1,k2−1}]
+
Sum[FcoalProb[n,j,6]*(4/((j+1)*j/2))*FcoalProb[j,k2,3]
*(1/((k2+1)*k2/2))*FcoalProb[k2,k1,1]*(1/((k1+1)*k1/2)),{j,k2+1,n−1}]
)−(E1k[n,k1])*(E1k[n,k2])])

(* Thm. 3.5, Eq. (10) *)
EVi[n_,i_]:=(FullSimplify[Sum[E1k[n,k],{k,i,n−1}]/i])

(* Thm. 3.6, Eq. (10) *)
VarVi[n_,i_]=(FullSimplify[(Sum[VarE1k[n,k],{k,i,n−1}]
+2*Sum[Sum[CovE1k1E1k2[n,k1,k2],{k2,k1+1,n−1}],{k1,i,n−1}])/(i*i)])


(* Thm. 3.7, Eq. (11) *)
CovVi1Vi2[n_,i1_,i2_]=(FullSimplify[(i2*i2*VarVi[n,i2]
+Sum[Sum[CovE1k1E1k2[n,k1,k2],{k2,i2,n−1}],{k1,i1,i2−1}])/(i1*i2)])

(* Thm. 3.8, formula 1 *)
EVi2[n_,i_]=(FullSimplify[VarVi[n,i]+(EVi[n,i]^2)])
```

43

```
VarSumVi[n_]=(FullSimplify[Sum[EVi2[n,i],{i,1,n−1}]
+2*Sum[Sum[CovVi1Vi2[n,i1,i2],{i2,i1+1,n−1}],{i1,1,n−2}]])

(* Thm. 3.8 Eq. (13), formula 2 *)
VarWn[n_]=(FullSimplify[2*Sum[VarVi[n,i],{i,1,n−1}]
+Sum[(EVi[n,i])^2,{i,1,n−1}]+2*Sum[Sum[CovVi1Vi2[n,i1,i2]
,{i2,i1+1,n−1}],{i1,1,n−1}]])

(* Thm. 3.8 Eq. (13), formula 3 *)
VarWnBar[n_]=(FullSimplify[Sum[(EVi[n,i])^2,{i,1,n−1}]])

(* Thm. 3.8 Eq. (13), formula 4 *)
VarWnCentre[n_]=(FullSimplify[2*Sum[VarVi[n,i],{i,1,n−1}]
+2*Sum[Sum[CovVi1Vi2[n,i1,i2],{i2,i1+1,n−1}],{i1,1,n−1}]])

(* Thm. 3.10 *)
FinalPart[n_]=(Sum[(i+1)*(i+5)/(i−1),{i,2,n−2}])
```