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

Krzysztof Bartoszek

April 30, 2018

Abstract

In this work we study the limit distribution of an appropriately normalized cophenetic index of the pure–birth tree conditioned 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 propose a way based on the 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. We conduct a power study to investigate how well the cophenetic indices detect deviations from the Yule tree and apply the methodology to empirical phylogenies.

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 at the individual positions of the genetic sequence (e.g. Ewens and Grant, 2005; Felsenstein, 2004; Yang, 2006). Given a phylogenetic tree it is often of interest to quantify the rate(s) of speciation and extinction for the studied species. To do this one commonly assumes a birth–death process with constant rates. However, the development of formal statistical tests whether a given tree comes from a given branching process model is an open area of research (see the still relevant "Work remaining" part at the end of Ch. 33 in Felsenstein, 2004). 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). Alternatively, McKenzie and Steel (2000) proposed to measure balance by counting cherries on the tree and they showed that after appropriate centring and scaling, this index converges to the standard normal distribution (for examples of other indices see Ch. 33 in Felsenstein, 2004).

Recently, a new balance index was proposed—the cophenetic index (Mir et al., 2013). The work here is inspired by private communication with evolutionary biologist Gabriel Yedid (current affiliation Nanjing Agricultural University, Nanjing, China) concerning the usage of 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, i.e. how to recognize if the tree is out of the "Yule zone" (Yang et al., 2017). 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 (for trees with and without branch lengths) and present the most important results of the manuscript. We define an associated submartingale that converges almost surely and in L^2 (Thm. 2.4), propose an elegant representation (Thm. 2.7) and a very promising approximation (Def. 2.8). Afterwards in Section 3, we show that in the discrete setting 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 4 we describe heuristic algorithms to simulate from these limit laws, show simulated quantiles, explore the power of the cophenetic index to recognize deviations from the Yule tree (comparing with Sackin's and Colless' indices' powers), and apply the indices to example empirical data. In Section 5 we prove the claims presented in Section 2 alongside other supporting results. Then, in Section 6 we study the second order properties of this decomposition and conjecture a Central Limit Theorem (CLT, Rem. 6.10). We end the paper with Section 7 by describing alternative representations of the cophenetic index.

2 The cophenetic index and summary of main results

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 for trees that are not fully resolved (i.e. not binary).

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 ϕ_{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 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 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 (*$\exp(1)$*—parametrization by the rate) random variable (see Section 5 for description of the tree's growth).*

The results of the present manuscript are built around a scaled version of the cophenetic index which is an almost surely and L^2 convergent submartingale. We first introduce some notation. Let $\mathcal{Y}_n$ be the σ–algebra containing all the information on the Yule with n tips tree and define

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

Below we present the main results concerning the cophenetic index, leaving the proofs and supporting theorems for Section 5.

Theorem 2.4 *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.

Definition 2.5 *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

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

Definition 2.6 *For $i = 1, \ldots, n-1$ let us introduce the random variable*

$$V_i^{(n)} := \frac{1}{i}\sum_{k=i}^{n-1} \mathrm{E}\left[1_k^{(n)} | \mathcal{Y}_n\right]. \tag{2}$$

Theorem 2.7 *W_n can be represented as*

$$W_n = \sum_{i=1}^{n-1} V_i^{(n)} Z_i, \tag{3}$$

where $Z_1, \ldots, Z_{n-1}$ are i.i.d. $\exp(1)$ *random variables.*

Definition 2.8 *Define the random variable $\overline{W}_n$ as*

$$\overline{W}_n = \sum_{i=1}^{n-1} \mathrm{E}\left[V_i^{(n)}\right] Z_i, \tag{4}$$

where $Z_1, \ldots, Z_{n-1}$ are i.i.d. $\exp(1)$ *random variables.*

Remark 2.9 *Despite the apparent elegance, it is not straightforward to derive a Central Limit Theorem (CLT) or limit statements concerning W_n from the representation of Eq.* (3)*. 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 in the proof of Thm. 2.4 in Section 5, because $((n+2)(n-1)/(n(n+1)) > 1$, 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. Figure 1 can easily mislead one to believe in the equality of the limiting distributions of W_n and $\overline{W}_n$. However,

in Thm. 6.8 we can see that $\mathrm{Var}\left[W_n\right]$ *and* $\mathrm{Var}\left[\overline{W}_n\right]$ *convergence to different limits. Therefore,* W_n *and* $\overline{W}_n$ *cannot converge in distribution to the same limit. However, as we shall see in Section 4,* $\overline{W}_n$ *provides a reasonable approximation (and importantly extremely cheap, in terms of computational time and memory) to* W_n *in the sense of their distributions.*

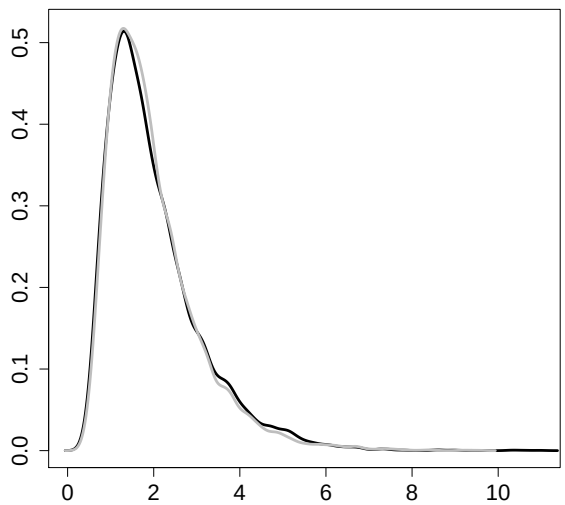

Figure 1: The curves are density estimates, via R's (R Core Team, 2013) `density()` function, of W_n's density (black) and $\overline{W}_n$'s density (gray). They are based on simulated values of W_n from 10000 simulated 500–tips Yule trees with $\lambda = 1$. To obtain a sample from $\overline{W}_n$, independent exp(1) random variables were drawn. 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 $\mathrm{E}\left[W_n\right] = \mathrm{E}\left[\overline{W}_n\right]$, but we have shown that their variances differ (simulations agree with Thm. 6.8).

Definition 2.10 *We naturally define the scaled discrete cophenetic index as*

$$\tilde{W}_n = \binom{n}{2}^{-1} \tilde{\Phi}^{(n)}. \tag{5}$$

Theorem 2.11 *$\tilde{W}_n$ is an almost surely and L^2 convergent submartingale.*

The applied reader will be most interested in how the results here can be practically used. As written already in the Introduction balance indices are often used to provide a single–number summary of the tree's shape. Such statistics can be then used e.g. to test if the tree is consistent with some null model (here the Yule model). Naturally, there has been extensive work on using different balance indices for significance testing (e.g. Agapow and Purvis, 2002; Blum and François, 2005; Yang et al., 2017). However previous works nearly always worked with indices that only considered the topology and often obtained the rejection regions through direct simulations.

Unfortunately, looking only at the tree's topology will not allow for distinguishing between some models. In particular (as seen in Tab. 3) there is no difference (from the topological indices perspective) between a Yule tree, a constant rate birth–death tree and a coalescent tree. Hence, a temporal index that also takes into account the branch lengths should be used (as indicated in the "Work remaining" section at the end of Ch. 33 in Felsenstein, 2004). A statistic based on $\Phi^{(n)}$ performs significantly better (but in these cases still leaves a lot to be desired). However, $\Phi^{(n)}$ shows it true usefulness when employed to distinguish a biased speciation (Blum and François, 2005) from a Yule model. Blum and François (2005) indicated that there is a regime where topological indices fail completely. Table 3 shows that in this setup (and also certain others) the temporal index in superior in recognizing the deviation from the Yule tree.

Directly simulating a tree from a null model (Yule here) and then calculating the index will of course give a sample from the correct null distribution. However, this approach is costly both in terms of time and memory. Therefore, if theoretical results that provide equivalent, asymptotic or approximate representations of the index's law are available they could speed up any study by orders of magnitude. In fact this is clearly visible in Tab. 1, calculating the cophenetic index directly from a sample of simulated pure–birth trees is over 170 times slower than considering $\overline{W}_n$. Even more dramatically one can obtain a sample from an approximation to the equivalent representation of the asymptotic distribution of $\tilde{\phi}^{(n)}$ (after normalization) nearly 3000 times faster than directly sampling the discrete cophenetic index.

In Alg. 1 we describe how the presented here approach can be used for significance testing. Then, in Section 4 we discuss in detail the required computational procedures, present simulation results concerning the power of

the tests and apply the tests to empirical data. Preceding this computational Section is a characterization of the limit distribution of (normalized) $\tilde{\Phi}^{(n)}$ and another proposal to approximate the limit of (normalized) $\Phi^{(n)}$. This section justifies the described simulation algorithms in Section 4.

Algorithm 1 Significance testing

input: A phylogenetic tree $\mathcal{T}^{(n)}$ with n tips and significance level α
output: A decision if the null hypothesis of $\mathcal{T}^{(n)}$ coming from a pure–birth process is rejected (**TRUE**) or not (**FALSE**)
Correct, when necessary, the tree for the speciation rate, by multiplying all branch lengths by $\hat{\lambda}$, if cophenetic index with branch lengths is used. ▷ See Section 4.2.
Calculate Φ, $\mathcal{T}^{(n)}$'s cophenetic index ▷ Exactly which version is used, $\Phi^{(n)}$, $\tilde{\Phi}^{(n)}$, $\Phi^{(n)}_{NRE}$, $\tilde{\Phi}^{(n)}_{NRE}$, depends on the particular tree, if it has branch lengths or root edge
Standardize Φ as $X = (\Phi - \mathrm{E}_{Yule}[\Phi])/\sqrt{\mathrm{Var}_{Yule}[\Phi]}$ ▷ $\mathrm{E}_{Yule}[\Phi]$ and $\mathrm{Var}_{Yule}[\Phi]$ depend which version of the cophenetic index is considered. In Thm. 2.12 all the possibilities are presented.
Obtain the quantiles $q_{Yule}(\alpha/2)$, $q_{Yule}(1-\alpha/2)$ (if test is two–sided), $q_{Yule}(\alpha)$ (left–tailed), $q_{Yule}(1-\alpha)$ (right–tailed) of X under the Yule model, i.e. $P(X \leq q_{Yule}(\alpha)) = \alpha$. ▷ Exactly how to obtain the quantiles is a matter of which version of the cophenetic index is used and computational resources (see Section 4).
if X is inside rejection region **then return TRUE**
else
    return FALSE
end if

Theorem 2.12 *A random variable with subscript NRE (no root–edge) indicates that this random variable comes from a tree lacking the edge leading to the root.*

$$
\begin{aligned}
\mathrm{E}\left[\Phi^{(n)}\right] &= \binom{n}{2}\frac{2(n-H_{n,1})}{n-1} \sim n^2 \\
\mathrm{E}\left[\Phi^{(n)}_{NRE}\right] &= \binom{n}{2}\left(\frac{2(n-H_{n,1})}{n-1}-1\right) \sim \tfrac{1}{2}n^2 \\
\mathrm{Var}\left[\Phi^{(n)}\right] &= \frac{\binom{n}{2}^2}{9n^2(n-1)^2}\left(12n^2(n^2-6n-4)H_{n-1,2}-9n^4+102n^3\right. \\
&\quad \left.+51n^2-24nH_{n-1,1}-72n-72\right) \\
&\sim \tfrac{1}{36}\left(2\pi^2-9\right)n^4 \\
\mathrm{Var}\left[\Phi^{(n)}_{NRE}\right] &= \frac{\binom{n}{2}^2}{9n^2(n-1)^2}\left(12n^2(n^2-6n-4)H_{n-1,2}-9n^4+102n^3\right. \\
&\quad \left.+51n^2-24nH_{n-1,1}-72n-72\right)-\binom{n}{2}^2 \\
&\sim \tfrac{1}{36}\left(2\pi^2-18\right)n^4
\end{aligned}
\tag{6}
$$

$$
\begin{aligned}
\mathrm{E}\left[\tilde{\Phi}^{(n)}\right] &= \binom{n}{2}\left(\frac{4(n-H_{n,1})}{n-1}-1\right) \sim 3n^2/2 \\
\mathrm{E}\left[\tilde{\Phi}^{(n)}_{NRE}\right] &= \binom{n}{2}\left(\frac{4(n-H_{n,1})}{n-1}-2\right) \sim n^2 \\
\mathrm{Var}\left[\tilde{\Phi}^{(n)}\right] &= \tfrac{1}{12}\left(n^4-10n^3+131n^2-2n\right)-4n^2H_{n,2}-6nH_{n,1} \sim n^4/12 \\
\mathrm{Var}\left[\tilde{\Phi}^{(n)}_{NRE}\right] &= \tfrac{1}{12}\left(n^4-10n^3+131n^2-2n\right)-4n^2H_{n,2}-6nH_{n,1} \sim n^4/12
\end{aligned}
\tag{7}
$$

Proof The proof of the expectation part is due to Mir et al. (2013); Sagitov and Bartoszek (2012). The variance of $\tilde{\Phi}^{(n)}$ is due to Cardona et al. (2013); Mir et al. (2013). The variance of $\Phi^{(n)}$ is a consequence of the lemmata and theorems presented in Section 6. When the root edge is not included, then we have to decrease the expectation by $\binom{n}{2}$. This is due to each pair of tips "having" the root edge included in the cophenetic distance between them. In the case of branch lengths, the expectation of the root edge, distributed as exp(1), is one. Without a root edge for the same reason the variance of $\Phi^{(n)}$ has to be decreased by $\binom{n}{2}^2$. In the discrete case the root edge has a deterministic length of 1 and hence no effect on the variance.

□

3 Contraction–type limit distribution

Even though the representation of Eq. (3) is a very elegant one, it is not obvious how to derive asymptotic properties of the process from it (compare

Section 6). We turn to considering the recursive representation proposed by Mir et al. (2013)

$$\tilde{\Phi}^{(n)}_{NRE} = \tilde{\Phi}^{(L_n)}_{NRE} + \tilde{\Phi}^{(R_n)}_{NRE} + \binom{L_n}{2} + \binom{R_n}{2}, \tag{8}$$

where L_n and R_n are the number of left and right daughter tip descendants. Obviously $L_n + R_n = n$.

From Eq. (8) we will be able to deduce the form of the limit of the process. In the case with branch lengths we attempt to approximate the cophenetic index with the following contraction–type law

$$\Phi^{(n)}_{NRE} = \Phi^{(L_n)}_{NRE} + \Phi^{(R_n)}_{NRE} + \binom{L_n}{2} T_{0.5} + \binom{R_n}{2} T'_{0.5}, \tag{9}$$

where $T_{0.5}$, $T'_{0.5}$ are independent exp(2) random variables (we index with the mean to avoid confusion with T_2, Section 5, the time between the second and third speciation event which is also exp(2) distributed). 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 exp(2) distributed (Cor. 3.2 and Thm. 3.3 Stadler and Steel, 2012). This is of course an approximation, as we cannot expect that the laws of the branch lengths with the depth of the recursion should become indistinguishable from the law of the average branch. In fact, we should expect that the law of Eq. (9) has to depend on n, i.e. the level of the recursion. For larger n the branches have distributions concentrated on smaller values, e.g. compare the randomly sampled root adjacent branch length law (Thm. 5.1 Stadler and Steel, 2012) with the law of the average branch length.

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 $\tilde{\Phi}^{(n)}_{NRE}$, $\Phi^{(n)}_{NRE}$ to differentiate from $\tilde{\Phi}^{(n)}$, $\Phi^{(n)}$ where the root branch is included, i.e.

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

Define now

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

and using Eqs. (6) and (7) we obtain the following recursions

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

and

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

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

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

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

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

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)

$$Y^{(\infty)} \stackrel{\mathcal{D}}{=} \tau^2 Y'^{(\infty)} + (1-\tau)^2 Y''^{(\infty)} + \frac{1}{2}\tau^2 T_{0.5} + \frac{1}{2}(1-\tau)^2 T'_{0.5} - \tau(1-\tau), \quad (10)$$

and

$$\tilde{Y}^{(\infty)} \stackrel{\mathcal{D}}{=} \tau^2 \tilde{Y}'^{(\infty)} + (1-\tau)^2 \tilde{Y}''^{(\infty)} + \frac{1}{2} - 3\tau(1-\tau) \quad (11)$$

where τ 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. (10) and (11).

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 3.1 *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', τ, T, T' are all independent. Define transformations $S_1 : D \to D$, $S_2 : D \to D$ as*

$$S_1(F) = \tau^2 Y + (1-\tau)^2 Y' + \frac{1}{2}\tau^2 T_{0.5} + \frac{1}{2}(1-\tau)^2 T'_{0.5} - \tau(1-\tau), \quad (12)$$

and

$$S_2(F) = \tau^2 Y + (1-\tau)^2 Y' + \frac{1}{2} - 3\tau(1-\tau) \quad (13)$$

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

Remark 3.2 *The proof of Thm. 3.1 is the same as Rösler (1991)'s proof of his Thm.* 2.1. *However, compared to the Quicksort algorithm (Rösler, 1991)*

we will have a $\sqrt{2/5}$ *upper bound on the rate of decay instead of* $\sqrt{2/3}$*. This speed–up should be expected as we have* τ^2 *and* $(1-\tau)^2$ *instead of* τ *and* $(1-\tau)$*. Thm. 3.1 can also be seen as a consequence of Rösler (1992)'s more general Thms.* 3 *and* 4*. The rate of convergence is also a consequence of the general contraction lemma (Lemma 1, Rösler and Rüschendorf, 2001).*

Now, using Lemmata 7.1, 7.2 (their proofs in 7.2 differ only in detail from the proof of Prop. 3.2 in Rösler, 1991) and arguing in the same way as Rösler (1991) did in his Section 3, especially his proof of his Thm. 3.1 we obtain 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. A minor point should be made. Here, we will have $(i/n)^4$ instead of $(i/n)^2$ in a counterpart of Rösler (1991)'s Prop 3.3.

Remark 3.3 *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*

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

Remark 3.4 *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 exponentially distributed. We therefore, also experimented by drawing* $T_{0.5}$*,* $T'_{0.5}$ *from a gamma distribution with rate equalling* $1/(2(\pi^2/3 - 25/8))$ *and shape equalling* $\pi^2/6 - 25/16$*. However, this significantly increased the duration of the computations but did not result in any visible improvements in comparison to Tab. 1.*

4 Significance testing

4.1 Obtaining the quantiles

Algorithm 1 requires knowledge of the quantiles of the underlying distribution in order to define the rejection region. Unfortunately, an analytical form of

the density of any scaled cophenetic index is not known so one will have to resort to some sort of simulations to obtain the critical values. Directly simulating a large number of pure–birth trees can take an overly long time, measured in minutes (on a modern machine with a large amount of memory, or hours on an older one). Fortunately, the cophenetic index can be calculated in $O(n)$ time (Cor. 3 Mir et al., 2013) and such a tree–traversing algorithm was employed to obtain $\Phi^{(n)}$ and $\tilde{\Phi}^{(n)}$. On the other hand, the suggestive (but wrong) approximations of Eq. (4) and contraction limiting distributions Eqs. (10) and (11) are significantly faster to simulate, see Tab. 1.

Simulating from the approximate Eq. (4) is straightforward. One simply draws $n-1$ independent exp(1) random variables. Simulating random variables satisfying Eqs. (10) and (11) is more involved and it may be possible to develop an exact rejection algorithm (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 now describe algorithms (Algs. 2 and 3) for simulating from a more general distribution, F, that satisfies

$$Y \stackrel{\mathcal{D}}{=} g_1(\tau)Y' + g_2(\tau)Y'' + C(\tau,\theta), \tag{14}$$

where $Y, Y', Y'' \sim F$, Y', Y'', τ, θ 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 θ's dimension. Of course in our case here we have $\tau \sim \mathrm{Unif}[0,1]$, $g_1(\tau) = \tau^2$, $g_2(\tau) = (1-\tau)^2$,

$$C(\tau, T, T') = \tau^2 T/2 + (1-\tau)^2 T'/2 - \tau(1-\tau)$$

and

$$C(\tau) = 1/2 - 3\tau(1-\tau)$$

for $\Phi^{(n)}$, $\tilde{\Phi}^{(n)}$ respectively. Of course, T, T' are independent and exp(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 $n^{-2}\tilde{\Phi}^{(n)}$.

The recursion of Alg. 3 for a given realization of τ and θ random variables can be directly solved. However, from numerical experiments implementing Alg. 3 iteratively seemed computationally ineffective.

In Tab. 1 we report on the simulations from the different distribution. For each distribution we draw a sample of size 10000 and repeat this 100

Algorithm 2 Population approximation

```
Initiate population size N
Set P[0, 1 : N] = Y_0                                   ▷ Initial population
for i = 1 to i_max do
    f_{i-1} = density(P[i − 1, ])                       ▷ density estimation by R
    for j = 1 to N do
        Draw τ from F_τ
        Draw θ from F_θ
        Draw Y_1, Y_2 independently from f_{i-1}
        P[i, j] = g_1(τ)Y_1 + g_2(τ)Y_2 + C(τ, θ)
    end for
end for
return P[i_max, ]
        ▷ Add root branch (exp(1) or 1) if needed for each individual.
```

Algorithm 3 Recursive approximation

```
procedure YRECURSION(n, Y_0)
    if n = 0 then
        Y_1 = Y_0, Y_2 = Y_0
    else if n = 1 and Y_0 = 0 then
        Draw τ_1, τ_2 independently from F_τ
        Draw θ_1, θ_2 independently from F_θ
        Y_1 = C(τ_1, θ_1)
        Y_2 = C(τ_2, θ_2)
    else
        Y_1 = YRECURSION(n − 1 , Y_0)
        Y_2 = YRECURSION(n − 1 , Y_0)
    end if
    Draw τ from F_τ
    Draw θ from F_θ
    return g_1(τ)Y_1 + g_2(τ)Y_2 + C(τ, θ),
end procedure
return YRECURSION(N, Y_0)
                                ▷ Add root branch (exp(1) or 1) if needed.
```

times. We compare the quantiles from the different distributions. We can see that the approximation of $\overline{W}_n$ for W_n is a 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 difference. We calculated (Thm. 6.8) 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 already 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 an attractive option, as the discrepancy between the two distributions seems small.

In Fig. 2 we compare the density estimates of (scaled and centred) both continuous and discrete branches cophenetic indices and their respective contraction–type limit distributions. The density estimates generally agree but we know from Tab. 1 that for $\Phi^{(n)}$ this is only an approximation. We simulated 10000 Yule trees and hence we report only the quantiles between 2.5% and 97.5%. Quantiles further out in the tails seemed less accurate and hence are not included in the table. Similarly, we can see less correspondence between the different estimates of kurtosis. This statistic relies on fourth moments and hence is more sensitive to the tails. On the other hand we can see much greater Monte Carlo error for the kurtosis in all simulations, including the setup where the values are extracted directly from Yule trees. The values for $\tilde{\Phi}^{(n)}$ seem more similar to values from the Yule tree. We should expect this as here we have shown an exact limit distribution.

An overall assessment of the quantiles is given by the root–mean–square–error (RMSE) row in Tab. 1. We consider the quantiles at the $\alpha_1 = 0.001$, $\alpha_2 = 0.005$, $\alpha_3 = 0.01$, $\alpha_4 = 0.025$, $\alpha_5 = 0.05$, $\alpha_6 = 0.95$, $\alpha_7 = 0.975$, $\alpha_8 = 0.99$, $\alpha_9 = 0.995$, $\alpha_{10} = 0.999$ levels. The RMSE is defined as

$$\mathrm{RMSE} = \left(\frac{1}{100} \sum_{j=1}^{100} \left(\sum_{i=1}^{5} (\alpha_i - \alpha_{i-1}) \left(\hat{q}_j(\alpha_i) - q(\alpha_i)\right)^2 + \sum_{i=6}^{10} (\alpha_{i+1} - \alpha_i) \left(\hat{q}_j(\alpha_i) - q(\alpha_i)\right)^2 \right) \cdot (0.1)^{-1} \right)^{\frac{1}{2}} \tag{15}$$

with dummy levels $\alpha_0 = 0$ and $\alpha_{11} = 1$. The $(0.1)^{-1}$ normalizes the whole mean–square–error. We only look at the error at the tails, so we correct by the fraction of the distributions' support that we consider. As a proxy for the true quantiles we take the pooled values (as explained in Tab. 1) from the "Yule columns". The j index runs over the 100 repeats of the simulations.

The RMSE, when using $\overline{W}_N$, seems to be on the level of the RMSE of the "direct simulations". $Y^{(N)}$ has an error of about twice the size (both simulation methods). Looking at $\tilde{Y}^{(N)}$ one can see that the RMSE is exactly on the level of the "Yule column's" RMSE. This is even though we used a recursion of level 10, while an exact match of distributions should take place in the limit (infinite depth recursion). However, the rapid, exponential convergence of the contraction seems to make any differences invisible, already at this recursion level.

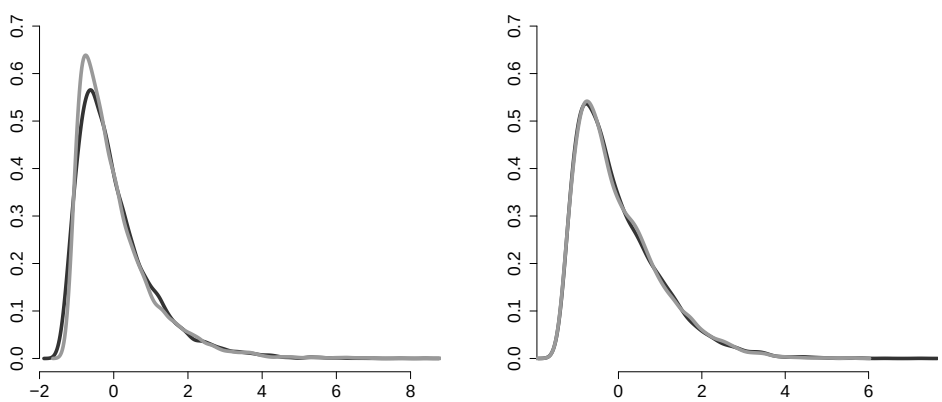

Figure 2: Density estimates of scaled (by theoretical standard deviation) and centred (by theoretical expectation) cophenetic indices (black) from 10000 simulated 500 tip Yule trees with $\lambda = 1$ and of simulation by Alg. 3 (gray), also scaled and centred to mean 0 and variance 1. Left: density estimates for $\Phi^{(n)}$, right: $\tilde{\Phi}^{(n)}$. The curves are calculate by R's `density()` function.

	$\left(\sqrt{\mathrm{Var}\left[\Phi^{(n)}\right]}\right)^{-1}(\Phi^{(n)} - \mathrm{E}\left[\Phi^{(n)}\right])$ limit approximation					$\left(\sqrt{\mathrm{Var}\left[\tilde{\Phi}^{(n)}\right]}\right)^{-1}(\tilde{\Phi}^{(n)} - \mathrm{E}\left[\tilde{\Phi}^{(n)}\right])$ limit ap		
	Yule	$\mathcal{N}(0,1)$	$\overline{W}_N$	$Y^{(N)}$ Alg. 2	$Y^{(N)}$ Alg. 3	Yule	$\mathcal{N}(0,1)$	$\tilde{Y}^{(N)}$ Alg. 2
Run time	690.918s	—	3.905s	0.318s	110.021s	698.269s	—	0.233s
Avg. (= 0)	$-0.023, 0.029$	0	$-0.025, 0.024$	$-0.024, 0.026$	$-0.019, 0.026$	$-0.02, 0.025$	0	$-0.033, 0.032$
	0.002	0	-0.001	0.006	0	0	0	-0.001
Var. (= 1)	$0.946, 1.074$	1	$0.921, 1.025$	$0.928, 1.072$	$0.932, 1.061$	$0.939, 1.038$	1	$0.953, 1.087$
	1.003	1	0.97	1.014	1	1	1	1.012
Skew.	$1.480, 1.834$	0	$1.487, 1.917$	$1.67, 2.124$	$1.62, 2.197$	$1.138, 1.368$	0	$1.163, 1.368$
	1.643	0	1.68	1.97	1.858	1.245	0	1.25
Ex. kurt.	$3.123, 7.222$	0	$3.148, 6,753$	$3.690, 8.31$	$3.5, 9.428$	$1.374, 2.853$	0	$1.392, 2.707$
	4.639	0	4.575	6.377	5.435	1.95	0	1.989
$q(0.025)$	$-1.235, -1.194$	-1.96	$-1.206, -1.174$	$1.115, 1.087$	$-1.114, -1.091$	$-1.257, -1.226$	-1.96	$-1.276, -1.239$
	-1.215	-1.96	-1.19	-1.1	-1.101	-1.245	-1.96	-1.257
$q(0.05)$	$-1.15, -1.104$	-1.644	$-1.115, -1.085$	$-1.048, -1.023$	$-1.047, -1.024$	$-1.18, -1.146$	-1.644	$-1.184, -1.154$
	-1.123	-1.644	-1.1	-1.033	-1.036	-1.162	-1.644	-1.17
$q(0.95)$	$1.861, 2.07$	1.644	$1.82, 2.013$	$1.863, 2.081$	$1.873, 2.066$	$1.844, 2.024$	1.644	$1.883, 2.066$
	1.946	1.644	1.914	1.942	1.969	1.949	1.644	1.958
$q(0.975)$	$2.436, 2.735$	1.96	$2.436, 2.732$	$2.434, 2.823$	$2.536, 2.792$	$2.328, 2.607$	1.96	$2.383, 2.645$
	2.587	1.96	2.549	2.642	2.634	2.486	1.96	2.5
RMSE	0.053	0.762	0.062	0.11	0.108	0.040	0.663	0.048

Table 1: Simulations based on 100 independent repeats of 10000 independent draws of each random variable (population size for Alg. 2) i.e. columns, bar $\mathcal{N}(0,1)$. The value on the left is the minimum observed from the 100 repeats, on the right the maximum and in the line below from pooling all repeats together. The running times are averages of 100 independent repeats with 10000 draws each. The abbreviations in the row names are for average (Avg.), variance (Var.), skewness (Skew.) excess kurtosis (Ex. kurt.) and root–mean–square–error (RMSE). 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 with speciation rate $\lambda = 1$, $n = 500$ tips and a root edge. The Yule tree $\Phi^{(n)}$, $\tilde{\Phi}^{(n)}$ values are centred and scaled by expectation and standard deviation from Eqs. (6) and (7). Other centrings and scalings are summarized in Tab. 2. $N = 10$ for Algs. 2 and 3 is the number of generations and recursion depth of the respective algorithm. In Alg. 2 the initial population is set at 0 and also $Y_0 = 0$ for Alg. 3. The simulations were run in R 3.4.2 for openSUSE 42.3 (x86_64) on a 3.50GHz. Intel®Xeon®CPU E5–1620 v4. The calculation of the RMSE is described in the text next to Eq. (15).

	$\overline{W}_N$	$Y^{(N)}$	$\tilde{Y}^{(N)}$
Centring (μ)	$2(n - H_{n,1})/(n-1)$	1	1
Scaling (σ)	$\sqrt{(2\pi^2 - 9)/9}$	$\sqrt{1 + 1/16}$	$(\sqrt{12})^{-1}$

	$\overline{W}_N^{NRE}$	$Y_{NRE}^{(N)}$	$\tilde{Y}_{NRE}^{(N)}$
Centring (μ)	$(n + 1 - 2H_{n,1})/(n-1)$	0	0
Scaling (σ)	$\sqrt{(2\pi^2 - 18)/9}$	1/4	$(\sqrt{12})^{-1}$

Table 2: Centrings and scalings applied to obtain the entries in Tab. 1. For a random variable X by its centred and scaled version we mean $(X - \mu)/\sigma$. These centrings and scalings are required to obtain mean zero, variance 1 versions of the random variables, i.e. so that they have the same location and scale as the z–transformed cophenetic index. In case of $\overline{W}_N$ we take the asymptotic scaling (Thm. 6.8) to be comparable with $Y^{(N)}$. For the convenience of the reader we also provide corresponding centrings and scalings in the no root–edge setup (not considered in Tab. 1).

4.2 Power of the tests

For a given test statistic to be useful one also needs to know its power, the ability to reject the null hypothesis (here Yule tree) when a given alternative one is true. For example, balance indices based only on topology like Sackin's, Colless' or $\tilde{\Phi}^{(n)}$ cannot be expected to differentiate between any trees that are generated by different constant rate birth–death processes or by the coalescent. The rationale behind this is that the topologies induced by the n contemporary species (i.e. we forget about lineages leading to extinct ones) are stochastically indistinguishable no matter what the death or birth rate is (Thm. 2.3, Cor. 2.4 Gernhard, 2008). Similarly, regarding the coalescent at the bottom of their p. 93 Steel and McKenzie (2001) write "..., one has the coalescent model [1,18,19]. In this model one starts with n objects, then picks two at random to coalesce, giving $n - 1$ objects. This process is repeated until there is only a single object left. If this process is reversed, starting with one object to give n objects, then it is equivalent to the Yule model. Note that in the coalescent model there is commonly a probability distribution for the times of coalescences, but in the Yule model we ignore this element." To differentiate between such trees one needs to take into consideration the branch lengths. Here we compare the power of

the Sackin's, Colless', $\Phi^{(n)}$ and $\tilde{\Phi}^{(n)}$ indices at the 5% significance level.

The null hypothesis is always that the tree is generated by a pure–birth process with rate $\lambda = 1$. The alternative ones are birth–death processes ($\lambda = 1$, death rate $\mu = 0.25$, and 0.5 using the TreeSim package), coalescent process (ape's `rcoal()` function Paradis et al., 2004) and the biased speciation model for

$$p \in \{0.05, 0.1, 0.125, 0.15, 0.18, 0.2, 0.25, 0.4, 0.5\}.$$

We also simulate a pure–birth process to check if the significance level is met. All trees were simulated with an exp(1) root edge.

The so–called biased speciation model with parameter p is the tree growth model as described by Blum and François (2005). In their words, "Assume that the speciation rate of a specific lineage is equal to r ($0 \leq r \leq 1$). When a species with speciation rate r splits, one of its descendant species is given the rate pr and the other is given the speciation rate $(1-p)r$ where p is fixed for the entire tree. These rates are effective until the daughter species themselves speciate. Values of p close to 0 or 1 yield very imbalanced trees while values around 0.5 lead to over–balanced phylogenies." We simulated such trees with in–house R code.

The quantiles of Sackin's and Colless' indices were obtained using Alg. 3. It is known (Eqs. 2 and 3 Blum and François, 2005; Blum et al., 2006) that after normalization (centring by expectation and dividing by n) in the limit they satisfy a contraction–type distribution of the form of Eq. (14), i.e.

$$Y \stackrel{\mathcal{D}}{=} \tau Y' + (1-\tau)Y'' + C(\tau)$$

for $\tau \sim \text{Unif}[0,1]$. The function $C(\tau)$ takes the form

$$C(\tau) = 2\tau \log \tau + 2(1-\tau)\log(1-\tau) + 1$$

in Sackin's case and

$$C(\tau) = \tau \log \tau + (1-\tau)\log(1-\tau) + 1 - 2\min(\tau, 1-\tau)$$

in Colless' case. It particular, studying the limit of Sackin's index is equivalent to studying the Quicksort distribution (Blum and François, 2005). We can immediately see the main qualitative difference, the limit of the normalized cophenetic index has the square in τ^2, $(1-\tau)^2$ in the "recursion part" while Sackin's and Colless' have τ, $(1-\tau)$.

Using 10000 repeats of Alg. 3 with recursion depth 10 we obtained the following sets of quantiles $q(0.025) = -0.983$, $q(0.95) = 1.189$, $q(0.975) = 1.493$, and $q(0.025) = -1.354$, $q(0.95) = 1.494$, $q(0.975) = 1.868$ respectively for the normalized Sackin's and Colless' indices.

Under each model we simulated 10000 trees conditioned on 500 contemporary tips. We then checked if the tree was outside the 95% "Yule zone" (Yang et al., 2017) by the procedure described in Alg. 1. We calculated the normalized Sackin's, Colless', discrete and continuous cophenetic indices (normalizations from Thm. 2.12). The functions `sackin.phylo()` and `colless.phylo()` of the phyloTop (Kendall et al., 2016) R package were used while the two cophenetic indices were calculated using a linear time in–house R implementation based on traversing the tree (Cor. 3 Mir et al., 2013). Two tests were considered, a two–sided one and a right–tailed one. For the discrete cophenetic index the quantiles from the simulation by Alg. 3 were considered, for the continuous those from $\overline{W}_N$ (Tab. 1). The power is then estimated as the fraction of times the null hypothesis was rejected and represented in Tab. 3 by the corresponding Type II error rates. For the Yule tree simulation we can see that the significance level is met. All simulated trees are independent of the trees used to obtain the values in Tab. 1 and quantiles of Sackin's and Colless' indices. Hence, they offer a validation of the rejection regions. We summarize the power study in Tab. 3.

As indicated in Alg. 1 one should first "correct" the tree for the speciation rate, when using the cophenetic index with branch lengths. The distributional results derived here on the cophenetic indices are for a unit speciation rate Yule tree. For a mathematical perspective this is not a significant restriction. If one has a pure–birth tree generated by a process with speciation rate $\lambda \neq 1$, then multiplying all branch lengths by λ will make the tree equivalent to one with unit rate. Hence, all the results presented here are general up to a multiplicative constant. However, from an applied perspective the situation cannot be treated so lightly. For example, if we used the cophenetic index with branch lengths from a Yule tree with a very large speciation rate, then we would expect a significant deviation. However, unless one is interested in deviations from the unit speciation rate Yule tree, this would not be useful. Hence, one needs to correct for this effect. If the tree did come from a Yule process, then an estimate, $\hat{\lambda}$, of the speciation rate by maximum likelihood can be obtained. For example, in the work here we used ape's `yule()` function. Then, one multiplies all branch lengths in the tree by $\hat{\lambda}$ and calculates the cophenetic index for this transformed tree. It is

Model	Sackin's		Colless'		$\tilde{\Phi}$		Φ					
	$>$	2	$>$	2	$>$	2	$>$	$>^c$	2	2^c	mean($\hat{\lambda}$)	variance($\hat{\lambda}$)
Yule	0.952	0.952	0.955	0.955	0.953	0.952	0.949	0.95	0.944	0.944	1	0.002
Coalescent	0.955	0.954	0.956	0.959	0.952	0.955	0.936	0	0.881	0	37.836	42.06
birth–death $\mu = 0.25$	0.952	0.953	0.956	0.955	0.948	0.952	0.853	0.874	0.903	0.91	0.87	0.002
birth–death $\mu = 0.5$	0.95	0.95	0.952	0.955	0.951	0.953	0.635	0.729	0.739	0.808	0.722	0.001
biased speciation $p = 0.05$	0	0	0	0	0	0	0	0.98	0	0.542	0.004	$4.213 \cdot 10^{-8}$
biased speciation $p = 0.1$	0	0	0	0	0	0	0	0.982	0	0.521	0.004	$4.241 \cdot 10^{-8}$
biased speciation $p = 0.125$	0	0	0	0	0.016	0.431	0	0.981	0	0.522	0.004	$4.211 \cdot 10^{-8}$
biased speciation $p = 0.18$	1	1	0.497	0.959	1	1	0	0.981	0	0.524	0.004	$4.191 \cdot 10^{-8}$
biased speciation $p = 0.2$	1	1	1	1	1	1	0	0.982	0	0.508	0.004	$4.222 \cdot 10^{-8}$
biased speciation $p = 0.25$	1	0.834	1	1	1	1	0	0.98	0	0.515	0.004	$4.243 \cdot 10^{-8}$
biased speciation $p = 0.4$	1	0	1	0	1	0	0	0.982	0.001	0.51	0.004	$4.218 \cdot 10^{-8}$
biased speciation $p = 0.5$	1	0	1	0	1	0	0	0.983	0.002	0.509	0.004	$4.335 \cdot 10^{-8}$

Table 3: Power, presented as Type II error rates, of the various indices to detect deviations from the Yule tree for various alternative models at the 5% significance level. In the first row the trees were simulates under the Yule (i.e. we present the Type I error rate) so this is a confirmation of correct significance level. Each probability is the fraction of 10000 independently simulated trees that were accepted as Yule trees by the various tests. Columns with "$>$" label indicate right–tailed test and with label "2" the two–sided test. The critical regions for the cophenetic indices were taken from the pooled estimates in Tab. 1. The superscript c indicates tests, where the trees were corrected for the speciation rate through multiplying all branch lengths with $\hat{\lambda}$. The mean and variance over all trees of $\hat{\lambda}$, as obtained through ape's `yule()` function is reported. Each tree's branches were scaled by its particular $\hat{\lambda}$ estimate.

important to point out that $\hat{\lambda}$ is only an estimate and hence a random variable. The effects of the this source of randomness on the limit distribution deserve a separate, detailed study. Balance indices that do not use branch lengths do not suffer from this issue but on the other hand miss another aspect of the tree—proportions between branch lengths that are non–Yule like.

The power analysis presented in Tab. 3 generally agrees with intuition and the power analysis done by Blum and François (2005). The first row shows that for all tests and statistics the 5% significance level is approximately kept. Then, in the next three rows (coalescent and birth–death process) all topology based indices fail completely (the power is at the significance level). This is completely unsurprising as the after one removes all speciation events (with lineages) leading to extinct species from a birth–death tree, the remaining tree is topologically equivalent to a pure birth tree. The same is true for the coalescent model, its topology is identical in law to the Yule tree's one. The cophenetic index with branch lengths has a high Type II error rate but is still better, than the topological indices. However, when one "corrects for λ" this index manages to nearly (2 trees were not rejected by the two–sided test) perfectly reject the coalescent model trees.

Power for the biased speciation model follows the same pattern as Blum and François (2005) observed. When imbalance is evident, $p \leq 0.125$, all (λ uncorrected) tests were nearly perfect (two–sided discrete cophenetic is an exception). However, the λ correction significantly worsened the ability of $\Phi^{(n)}$ to detect deviations. As imbalance decreased so did the power of the topological indices. For overbalanced trees one–sided tests failed, two sided worked (just as Blum and François, 2005, observed). The cophenetic index with branch lengths (without correction), that does not consider only the topology, was able to successfully reject the pure–birth tree for all p (with only minimal Type II error for $p \geq 0.4$ in the two–sided test case). Interestingly, $\Phi^{(n)}$'s (both corrected and uncorrected) power seems invariant with respect to p. These results are especially promising as $\Phi^{(n)}$ seems to be an index that functions significantly better in the difficult, $0.18 \leq p \leq 0.25$, regime, even after correcting.

At this stage we can point out that a normal approximation to the cophenetic indices' limit distribution is not appropriate. When doing the above power study we observed that when using the quantiles of the standard normal distribution the right–tailed test based on $\Phi^{(n)}$ rejects 6.81% of Yule trees, based on $\tilde{\Phi}^{(n)}$ rejects 7.03% of Yule trees, two–sided $\Phi^{(n)}$ test rejects

4.87% of Yule trees and two–sided $\tilde{\Phi}^{(n)}$ test rejects 4.66% of Yule trees. The Type I error rates of the two–sided tests are within the observed Monte Carlo errors (in Tab. 3) but the right–tailed tests' Type I error are evidently inflated. This confirms that the right tail of the scaled cophenetic index is much heavier than normal.

In short the power study indicates that the cophenetic index with branch lengths should be considered as an option to detect deviations from the Yule tree. This is because it is able to use information from two sources—the topology and time (a needed direction of development, as indicated in Ch. 33 of Felsenstein, 2004). Actually, this is evident in the decomposition of Eq. (3). The $V_i^{(n)}$s describe the topology and the Z_is branch lengths. With more information a more powerful testing procedure is possible. Deviations that are not topologically visible, e.g. biased speciation in the $0.18 \leq p \leq 0.25$ regimes, are now detectable. To use $\Phi^{(n)}$ one should correct for the effects of the speciation rate, as otherwise one merely detects deviations from the unit rate Yule tree. This correction is a mixed blessing. It can help or hinder detection.

4.3 Examples with empirical phylogenies

It is naturally interesting to ask how do the indices behave for phylogenies estimated from sequence data. Comparing a database of phylogenies, like TreeBase (`http://www.treebase.org`), with yet another index's distribution under the Yule model should not be expected to yield interesting results. The Yule model has been indicated as inadequate to describe the collection of TreeBase's trees (e.g. Blum and François, 2007). Therefore, we choose a particular study that estimated a tree and also reported a collection of posterior trees. Sosa et al. (2016a) is a recent work, providing all trees from BEAST's (Drummond and Rambaut, 2007) output, well suited for such a purpose. Sosa et al. (2016a) estimate the evolutionary relationships between a set of 109 tree ferns species. They report a posterior set of 22498 phylogenies (Sosa et al., 2016b).

In Tab. 4 we look what percentage of the trees from the posterior was accepted as being consistent with the Yule tree by the various tests and indices. It can be seen that the discrete cophenetic index has a high acceptance rate. The continuous one, which also takes into account branch lengths did not accept a single tree. However, this is lost when one corrects for the speciation rate (first ape's `multi2di()` was used to make the trees binary ones). Most

tests and indices rejected the Yule tree for the maximum likelihood estimate of the phylogeny with some exceptions. The two–sided discrete cophenetic index test did not reject the null hypothesis of the pure–birth tree. Also after correcting for the speciation rate (estimated at $\hat{\lambda} = 0.023$), neither test based on the continuous cophenetic index rejected the Yule tree. Therefore, one should conclude (based on the "topological balances") that the Yule tree null hypothesis can be rejected for this clade of plants.

Sackin's		Colless'		$\tilde{\Phi}$		Φ			
$>$	Sackin's 2	$>$	2	$>$	2	$>$	$>^c$	2	2^c
0.03	0.109	0.019	0.062	0.385	0.559	0	0.974	0	0.995

Table 4: Percentage of trees from Sosa et al. (2016b)'s set of posterior trees accepted as Yule trees by the various tests and indices. Columns with "$>$" label indicate right–tailed test and with label "2" the two–sided test. The critical regions for the cophenetic indices were taken from the pooled estimates in Tab. 1. The superscript c indicates tests, where the trees were corrected for the speciation rate through multiplying all branch lengths with $\hat{\lambda}$. Ape's `yule()` function returned an average over all trees estimate of λ of 0.023 with variance $2.988 \cdot 10^{-6}$. Each tree's branches were scaled by its particular $\hat{\lambda}$ estimate. Each tree was first transformed by ape's `multi2di()` into a binary one.

We also followed Blum and François (2005) in looking at Yusim et al. (2001)'s phylogeny of the human immunodeficiency virus type 1 (HIV–1) group M gene sequences, available in the ape R package. The phylogeny consists of 193 tips and Blum and François (2005) could not reject the null hypothesis of the pure–birth tree (using Sackin's index amongst others). After pruning the tree to keep "only the old internal branches that corresponded to the 30 oldest ancestors" they were able to reject the Yule tree. They conclude that the "results probably indicate a change in the evolutionary rate during the evolution which had more impact on cladogenesis during the early expansion of the virus." Repeating their experiment we find that only the two versions of the cophenetic index point to a deviation but only in the two–sided test (see Tab. 5). Based only on the $\tilde{\Phi}^{(n)}$'s test and that it conflicts with the conclusions of Sackin's and Colless' one should not draw any conclusions. However, as $\Phi^{(n)}$'s test indicates a deviation, we can be inclined to

reject the null hypothesis of the Yule tree. This is further strengthened by the fact that the significance remains after the correction for λ. Even though the topology as a whole seems consistent with the pure–birth tree the branch lengths are not. The fact that only the two–sided test rejected the Yule tree indicates that the HIV phylogeny is over–balanced in comparison to a pure–birth tree. In fact, in the biased speciation model tree over–balance is observed for values of p close to 0.5 (Blum and François, 2005). Such trees have a declining speciation rate as they grow and hence this supports Blum and François (2005)'s aforementioned explanation.

Sackin's	Colless'	$\tilde{\Phi}$	Φ	Φ^c	$\hat{\lambda}$
$0.823^{-,-}$	$0.993^{-,-}$	$-1.689^{-,*}$	$-1.765^{-,*}$	$-1.602^{-,*}$	9.313

Table 5: Values of the normalized indices for Yusim et al. (2001)'s HIV–1 phylogeny. Above each index is an indication if the index deviates at the 5% significance level from the Yule tree, dash insignificant, asterisk significant. The first symbol concerns the right–tailed test, the second the two–sided test. The superscripted Φ^c is calculated from the tree corrected for the speciation rate by multiplying all branch lengths by $\hat{\lambda}$.

5 Almost sure behaviour of the cophenetic index

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, 2014; Bartoszek and Sagitov, 2015a,b; 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 $\exp(k)$ distributed, as the minimum of k $\exp(1)$ 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. 3 and Bartoszek and Sagitov, 2015b; Sagitov and Bartoszek, 2012).

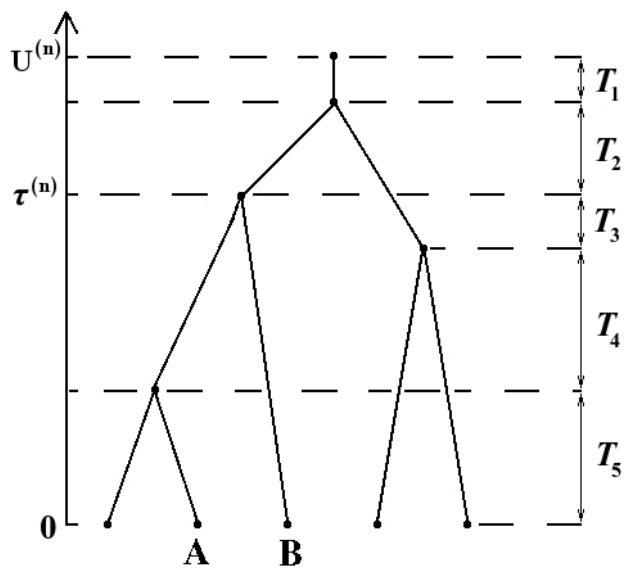

Figure 3: A pure–birth tree with the various time components marked on it. The between speciation times on this lineage are T_1, T_2, $T_3 + T_4$ and T_5. If we "randomly sample" the pair of extant species "A" and "B", then the two nodes coalesced at time $\tau^{(n)}$.

Theorem 5.1 *The cophenetic index is an increasing sequence of random variables, $\Phi^{(n+1)} > \Phi^{(n)}$ and has the recursive representation*

$$\Phi^{(n+1)} = \Phi^{(n)} + nU^{(n)} - \sum_{i=1}^{n} \xi_i^{(n)} \sum_{i \neq j}^{n} \tau_{ij}^{(n)}, \tag{16}$$

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

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

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.

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

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

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

and then obtain the recursive form

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

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

□

Proof *of Theorem 2.4.* 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

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

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

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

Then, using the general formula for the moments 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.

□

Corollary 5.2 *W_n has finite third moment and is L^3 convergent.*

Proof We first recall the W_n is positive. Using the general formula for the moments of $U^{(n)} - \tau^{(n)}$ again we see

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

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

□

Remark 5.3 *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.

Proof *of Theorem 2.7.* We write W_n as

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

where $Z_1, \ldots, Z_{n-1}$ are i.i.d. exp(1) random variables.

□

Remark 5.4 *We notice that we may equivalently rewrite*

$$
W_n = \sum_{k=1}^{n-1}\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]\left(\sum_{i=1}^{k} T_i\right) = \sum_{k=1}^{n-1}\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]\left(\sum_{i=1}^{k}\frac{1}{i}Z_i\right). \tag{17}
$$

The above and Eq. (3) *are very elegant representations of the cophenetic index with branch lengths. They explicitly describe the way the cophenetic index is constructed from a given tree.*

Proof *of Theorem 2.11.* The argumentation is analogous to the proof of Thm. 2.4 by using the recursion

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

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 7.1.

□

6 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. (4), 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 6.1

$$\mathrm{Var}\left[1_k^{(n)}\right] = 2\frac{n+1}{n-1}\frac{1}{(k+1)(k+2)}\left(1 - 2\frac{n+1}{n-1}\frac{1}{(k+1)(k+2)}\right) \tag{18}$$

Proof

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

□

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

Lemma 6.2 *For $k_1 \neq k_2$*

$$\mathrm{Cov}\left[1_{k_1}^{(n)}, 1_{k_2}^{(n)}\right] = -\pi_{n,k_1}\pi_{n,k_2} = \frac{(-4)(n+1)^2}{(n-1)^2(k_1+1)(k_1+2)(k_2+1)(k_2+2)}. \tag{19}$$

Lemma 6.3

$$\mathrm{Var}\left[\mathrm{E}\left[1_k^{(n)}|\mathcal{Y}_n\right]\right] = 4\tfrac{n+1}{n(n-1)^2}\tfrac{(n-(k+1))(n(3k^2+5k-4)-(k^2-k-8))}{(k+1)^2(k+2)^2(k+3)(k+4)}. \tag{20}$$

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

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

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

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

To calculate cases (ii) and (iii) we visualize the situation in Fig. 4 and recall the proof of Bartoszek and Sagitov (2015b)'s Lemma 1. Using Mathematica we obtain

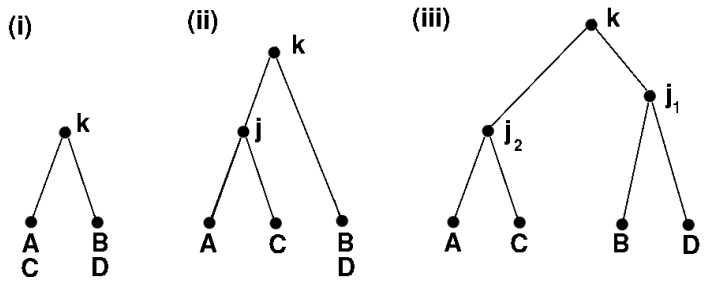

Figure 4: The three possible cases when drawing two random pairs of tip species that coalesce at the k–th speciation event. In the picture we "randomly draw" pairs (A, B) and (C, D).

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

Similarly for case (iii)

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

We now put this together as

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

and we obtain (through Mathematica)

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

□

Lemma 6.4 *For $k_1 < k_2$*

$$\mathrm{Cov}\left[\mathrm{E}\left[1_{k_2}^{(n)}|\mathcal{Y}_n\right], \mathrm{E}\left[1_{k_1}^{(n)}|\mathcal{Y}_n\right]\right] = \tfrac{(-8)(n+1)}{n(n-1)^2}\tfrac{(3n-(k_2-2))(n-(k_2+1))}{(k_1+1)(k_1+2)(k_2+1)(k_2+2)(k_2+3)(k_2+4)}. \tag{21}$$

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

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

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. 5 (notice that since $k_1 \neq k_2$ the counterpart of event (i) in Fig. 4 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".

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

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

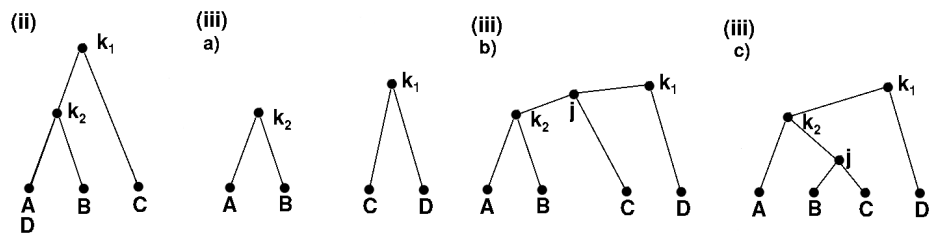

Figure 5: The possible cases when drawing two random pairs of tip species that coalesce at speciation events $k_1 < k_2$ respectively. In the picture we "randomly draw" pairs (A, B) and (C, D).

In the same way for the subcases of (iii)

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

We now put this together as

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

and we obtain

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

□

Theorem 6.5

$$\mathrm{E}\left[V_i^{(n)}\right] = 2\frac{1}{n-1}\frac{n-i}{i(i+1)} \tag{22}$$

Proof We immediately have

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

□

Theorem 6.6

$$\mathrm{Var}\left[V_i^{(n)}\right] = 4\frac{(n+1)}{n(n-1)^2}\frac{(n-i)(n-(i+1))(i-1)}{i^2(i+1)^2(i+2)(i+3)} \tag{23}$$

Proof We immediately may write using Lemmata 6.3, 6.4 and Mathematica

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

□

Theorem 6.7 *For $1 \le i_1 < i_2 \le n-1$ we have*

$$\mathrm{Cov}\left[V_{i_1}^{(n)}, V_{i_2}^{(n)}\right] = 4\frac{(n+1)}{n(n-1)^2}\frac{(i_1-1)(n-i_2)(n-(i_2+1))}{i_1(i_1+1)i_2(i_2+1)(i_2+2)(i_2+3)}. \tag{24}$$

Proof Again using Lemmata 6.3, 6.4, Mathematica and the fact that $i_1 < i_2$

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

□

Theorem 6.8

$$\begin{aligned}
\mathrm{Var}\left[\sum_{i=1}^{n-1} V_i^{(n)}\right] &= \tfrac{1}{54n^2(n-1)^2}\left(179n^4 + 588n^3 + 133n^2 - 432n\right.\\
&\quad -468 - 108n^2(n+1)(n+3)H_{n-1,2}\\
&\quad \left.-144nH_{n-1,1}\right) \to \tfrac{179}{54} - \tfrac{\pi^2}{3} \approx 1.347,\\
\mathrm{Var}\left[\sum_{i=1}^{n-1} V_i^{(n)} Z_i\right] &= \tfrac{1}{9n^2(n-1)^2}\left(12n^2(n^2-6n-4)H_{n-1,2} - 9n^4\right.\\
&\quad \left.+102n^3 + 51n^2 - 24nH_{n-1,1} - 72n - 72\right)\\
&\to \tfrac{2}{9}\pi^2 - 1 \approx 1.193,\\
\mathrm{Var}\left[\sum_{i=1}^{n-1}\mathrm{E}\left[V_i^{(n)}\right] Z_i\right] &= \tfrac{2}{3n^2(n-1)^2}\left(\left(12H_{n-1,2} - 18\right)n^4 - 24n^3\right.\\
&\quad \left.+12n^2(2n+1)H_{n-1,2} - 24n^2 + 24n + 12\right)\\
&\to \tfrac{4}{3}\pi^2 - 12 \approx 1.159,\\
\mathrm{Var}\left[\sum_{i=1}^{n-1}\left(V_i^{(n)} - \mathrm{E}\left[V_i^{(n)}\right]\right) Z_i\right] &= \tfrac{1}{9n^2(n-1)^2}\left(99n^4 + 174n^3 - 21n^2 - 144n\right.\\
&\quad -108 - 12n^2(n+1)(5n+7)H_{n-1,2}\\
&\quad \left.-24nH_{n-1,1}\right) \to 11 - \tfrac{10}{9}\pi^2 \approx 0.034.
\end{aligned} \tag{25}$$

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_is are i.i.d. exp(1).

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

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

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

As the Z_is are i.i.d. exp(1) we use Mathematica to obtain

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

For the fourth equality we use the same properties and pair–wise independence of Z_is.

$$\mathrm{Var}\left[\sum_{i=1}^{n-1}\left(V_i^{(n)}-\mathrm{E}\left[V_i^{(n)}\right]\right)Z_i\right]=\sum_{i=1}^{n-1}\mathrm{Var}\left[V_i^{(n)}Z_i\right]+\sum_{i=1}^{n-1}\mathrm{Var}\left[\mathrm{E}\left[V_i^{(n)}\right]Z_i\right]$$
$$-2\sum_{1=i_1<i_2}^{n-1}\mathrm{Cov}\left[V_{i_1}^{(n)}Z_{i_1},\mathrm{E}\left[V_{i_2}^{(n)}\right]Z_{i_2}\right]$$
$$=\sum_{i=1}^{n-1}\mathrm{Var}\left[V_i^{(n)}Z_i\right]+\sum_{i=1}^{n-1}\left(\mathrm{E}\left[V_i^{(n)}\right]\right)^2-2\sum_{i=1}^{n-1}\left(\mathrm{E}\left[V_i^{(n)}\right]\right)^2$$
$$=\sum_{i=1}^{n-1}\mathrm{Var}\left[V_i^{(n)}Z_i\right]-\sum_{i=1}^{n-1}\left(\mathrm{E}\left[V_i^{(n)}\right]\right)^2$$
$$=\tfrac{1}{9n^2(n-1)^2}\left(99n^4-12n^2(n+1)(5n+7)H_{n-1,2}+174n^3-21n^2\right.$$
$$\left.-24nH_{n-1,1}-144n-108\right)\to 11-\tfrac{10}{9}\pi^2\approx 0.034.$$

□

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. 1. As a check also notice that, as implied by variance properties,

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

Theorem 6.9

$$\mathrm{E}\left[\mathrm{Var}\left[(n)^{-1}\sum_{i=2}^{n-2}\frac{V_i^{(n)}Z_i-\mathrm{E}\left[V_i^{(n)}\right]}{\sqrt{\mathrm{Var}\left[V_i^{(n)}\right]}}\middle|\{V_i^{(n)}\}\right]\right]\to 0.5. \tag{26}$$

Proof Using the limit for the variance of $V_i^{(n)}$ (Thm. 6.6) and the independence of the Z_is 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. 6.6 and 6.5 we have

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

Plugging this in (and using Mathematica)

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

□

Remark 6.10 *Simulations presented in Fig. 6 and Thm. 6.9 suggest a different possible CLT, namely*

$$(n)^{-1}\sum_{i=2}^{n-2}\frac{V_i^{(n)}Z_i - \mathrm{E}\left[V_i^{(n)}\right]}{\sqrt{\mathrm{Var}\left[V_i^{(n)}\right]}} \overset{\text{weakly}}{\longrightarrow} \text{some distribution}(\text{mean} = 0, \text{variance} = \frac{1}{2}). \tag{27}$$

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. 2.9 and Fig. 1 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. (26)*. 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.* (27)*.*

7 Alternative descriptions

7.1 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

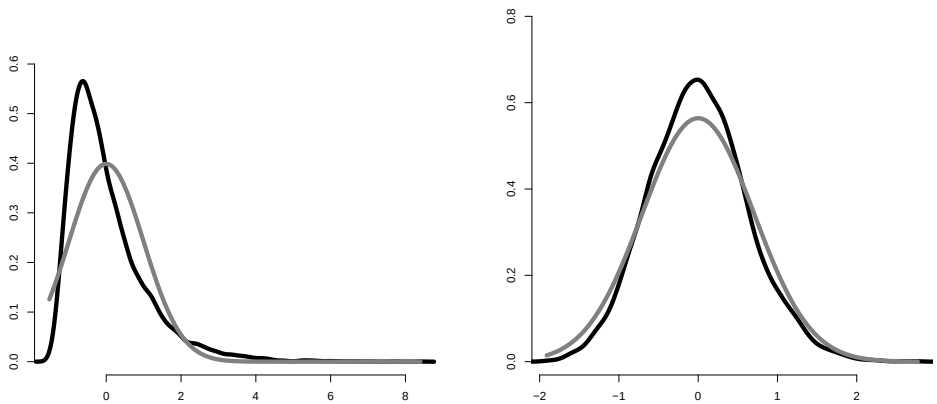

Figure 6: Density estimates of scaled and centred cophenetic indices for 10000 simulated 500 tip Yule trees with $\lambda = 1$. Left: density estimate of $(\Phi^{(n)} - \mathrm{E}\left[\Phi^{(n)}\right])/\sqrt{\mathrm{Var}\left[\Phi^{(n)}\right]}$. The black curve is the density fitted to simulated data by R's `density()` function, the gray is the $\mathcal{N}(0,1)$ density. Right: simulation of Eq. (27), the gray curve is the $\mathcal{N}(0,1/2)$ density, and the black curve is the density fitted to simulated data by R's `density()` function. The sample variance of the simulated Eq. (27) 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. (26) does not converge to 0.

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

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

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

We also know by previous calculations that

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

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,

$$\begin{aligned}
&\mathrm{E}\left[\epsilon_i^{(n)}\right] = \mathrm{E}\left[A_i^{(n+1)}\right] - \tbinom{n+1}{2}^{-1}\tbinom{n}{2}\,\mathrm{E}\left[A_i^{(n)}\right] = \tfrac{2(n+1-i)}{n(i+1)} - \tfrac{n(n-1)}{n(n+1)}\tfrac{2(n-i)}{(n-1)(i+1)} \\
&= \tfrac{2}{i+1}\left(\tfrac{n+1-i}{n} - \tfrac{n-i}{n+1}\right) = \tfrac{2}{i+1}\tfrac{(n-i+1)(n+1)-n(n-i)}{n(n+1)} = \tfrac{2}{i+1}\tfrac{n(n-i)+n+n-i+1-n(n-i)}{n(n+1)} \\
&= \tfrac{2}{i+1}\tfrac{2n+1-i}{n(n+1)} \to 0.
\end{aligned}$$

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 in terms of the conditional (on tree) cumulative distribution function for the random variable κ—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. 2.11).

7.2 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. 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

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

Acknowledgments

I was supported by the Knut and Alice Wallenberg Foundation and am now by the Swedish Research Council (Vetenskapsrådet) grant no. 2017–04951. I am grateful to the Barcelona Graduate School of Mathematics (BGSMath) for sponsoring the Workshop on Algebraical and Combinatorial Phylogenetics which significantly contributed to the development of my work. I would like to thank the whole Computational Biology and Bioinformatics Research Group of the Balearic Islands University for hosting me on multiple occasions, many discussions and suggestions on phylogenetic indices. My visits to the Balearic Islands University were partially supported by the the G S Magnuson Foundation of the Royal Swedish Academy of Sciences (grants no. MG2015–0055, MG2017–0066) and The Foundation for Scientific Research and Education in Mathematics (SVeFUM). I would like to ac-

knowledge Gabriel Yedid for numerous discussions on the distribution of the cophenetic index and sharing his cophenetic index simulation R code. I am grateful to Cecilia Holmgren and Svante Janson for pointing me to the works on contraction–type distributions and many discussions. I would furthermore like to acknowledge Wojciech Bartoszek, Sergey Bobkov, Joachim Domsta, Serik Sagitov, Mike Steel for helpful comments and discussions related to this work. I am indebted to two anonymous reviewers, an anonymous editor and Haochi Kiang for careful reading of an earlier version of the manuscript and comments significantly improving it.

References

P.–M. Agapow and A. Purvis. Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. *Syst. Biol.*, 51(6):866–872, 2002.

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 and O. François. Which random processes describe the Tree of Life? A large–scale study of phylogenetic tree imbalance. *Syst. Biol*, 55 (4):685–691, 2007.

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.

A. J. Drummond and A. Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, 7:214, 2007.

W. Ewens and G. Grant. *Statistical Methods in Bioinformatics: An Introduction*. Springer, New York, 2005.

J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates Inc., Sundarland, U.S.A., 2004.

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.

T. Gernhard. The conditioned reconstructed process. *J. Theor. Biol.*, 253: 769–778, 2008.

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.

M. L. Kendall, M. Boyd, and C. Colijn. phyloTop, 2016. `https://cran.r-project.org/web/packages/phyloTop/index.html`.

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.

E. Paradis, J. Claude, and K. Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290, 2004.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org`.

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.

U. Rösler. A fixed point theorem for distributions. *Stoch. Proc. Applic.*, 42: 195–214, 1992.

U. Rösler and L. Rüschendorf. The contraction method for recursive algorithms. *Algorithmica*, 29:3–33, 2001.

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.

V. Sosa, J. F. Ornelas, S. Ramírez-Barahona, and E. Gándara. Historical reconstruction of climatic and elevation preferences and the evolution of cloud forest–adapted tree ferns in Mesoamerica. *PeerJ*, 4:e2696, 2016a.

V. Sosa, J. F. Ornelas, S. Ramírez-Barahona, and E. Gándara. Data from: Historical reconstruction of climatic and elevation preferences and the evolution of cloud forest–adapted tree ferns in Mesoamerica. *Dryad Digital Repository*, 2016b. `https://doi.org/10.5061/dryad.709t8`.

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.

G.–D. Yang, P.–M. Agapow, and G. Yedid. The tree balance signature of mass extinction is erased by continued evolution in clades of constrained size with trait–dependent speciation. *PLoS ONE*, 12(6):e0179553, 2017.

Z. Yang. *Computational Molecular Evolution.* Oxford Series in Ecology and Evolution. Oxford University Press, Oxford, 2006.

K. Yusim, M. Peeters, O. G. Phybus, T. Bhattacharya, E. Delaporte, C. Mulanga, M. Muldoon, J. Theiler, and B. Korber. Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immunedeficiency syndrome epidemic and human immunodeficiency virus evolution. *Philos. Trans. Roy. Soc. Lond. B*, 356:855–866, 2001.

Appendix A: Mathematica code for Section 6

```
(*
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. 2.3, Eq. (1) *)
E1k[n_,k_]:=(2*(n+1)/((n-1)*(k+1)*(k+2)))
(* Lemma 6.1, Eq. (18) *)
Var1k[n_,k_]:=(E1k[n,k]-E1k[n,k]*E1k[n,k])
(* Lemma 6.2, Eq. (19) *)
Cov1k11k2[n_,k1_,k2_]:=(-E1k[n,k1]*E1k[n,k2])

(* Lemma 6.3, Eq. (20) *)
VarE1k[n_,k_]=(FullSimplify[
(1/(n*(n-1)/2))*(2*(n+1)/((n-1)*(k+1)*(k+2)))
+
(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 6.4, Eq. (21) *)
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. 6.1, Eq. (22) *)
EVi[n_,i_]:=(FullSimplify[Sum[E1k[n,k],{k,i,n-1}]/i])

(* Thm. 6.2, Eq. (23) *)
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. 6.3, Eq. (24) *)
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. 6.4, formula 1 *)
EVi2[n_,i_]=(FullSimplify[VarVi[n,i]+(EVi[n,i]^2)])
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. 6.4 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. 6.4 Eq. (13), formula 3 *)
VarWnBar[n_]=(FullSimplify[Sum[(EVi[n,i])^2,{i,1,n-1}]])

(* Thm. 6.4 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. 6.5 *)
FinalPart[n_]=(Sum[(i+1)*(i+5)/(i-1),{i,2,n-2}])
```

Appendix B: Counterparts of Rösler (1991)'s Prop. 3.2 for the cophenetic index

Lemma 7.1 *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 $\tilde{C}(x) = 0.5 - 3x(1-x)$ *for* $x \in [0,1]$*, then*

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

Proof Writing out

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

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

$$|\tilde{C}_n(\lceil nx\rceil) - \tilde{C}(x)| \leq 3|\tfrac{\lceil nx\rceil}{n}(1-\tfrac{\lceil nx\rceil}{n}) - x(1-x)| + 2n^{-1}\ln n$$
$$\leq \sup_{|y-z|<1/n} |\tilde{C}(y) - \tilde{C}(z)| + 2n^{-1}\ln n \leq \tfrac{6}{n} + 2n^{-1}\ln n + O(n^{-2}).$$

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

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

□

Lemma 7.2 *Define for $i \in \{1,\ldots,n\}$, $T, T' \sim \exp(2)$*

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

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

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

then

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

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 $O(n^{-2})$.

Proof Similarly, as in the proof of Lemma 7.1 we write out

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

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 $1 \leq \lceil nx\rceil \leq n-1$

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

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.

□