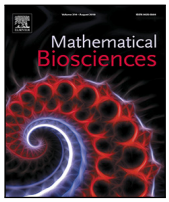
Contents lists available at ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs



Highlights

Squaring within the Colless index yields a better balance index

Mathematical Biosciences xxx (xxxx) xxx

Krzysztof Bartoszek, Tomás M. Coronado, Arnau Mir, Francesc Rosselló*

- We define a modification of the classical Colless index for bifurcating trees.
- Our Q-Colless index classifies as most balanced exactly the maximally balanced trees.
- Our Q-Colless index has a larger resolutive power than the Colless index.
- We give closed formulas for its first moments under the Yule and uniform models.

Graphical abstract and Research highlights will be displayed in online search result lists, the online contents list and the online article, but will not appear in the article PDF file or print unless it is mentioned in the journal specific style requirement. They are displayed in the proof pdf for review purpose only.

Please cite this article as: K. Bartoszek, T.M. Coronado, A. Mir et al., Squaring within the Colless index yields a better balance index, Mathematical Biosciences (2020) 108503, https://doi.org/10.1016/j.mbs.2020.108503.

MBS: 108503

ARTICLE IN PRESS

Mathematical Biosciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

Original Research Article

Squaring within the Colless index yields a better balance index

Krzysztof Bartoszek [a], Tomás M. Coronado [b,c], Arnau Mir [b,c], Francesc Rosselló [b,c,*]

[a] Department of Computer and Information Science, Linköping University, 581 83 Linköping, Sweden
[b] Department of Mathematics and Computer Science, University of the Balearic Islands, E-07122 Palma, Spain
[c] Balearic Islands Health Research Institute (IdISBa), E-07010 Palma, Spain

ARTICLE INFO

Keywords:
Phylogenetic tree
Balance index
Colless index
Yule model
Uniform model

ABSTRACT

The Colless index for bifurcating phylogenetic trees, introduced by Colless (1982), is defined as the sum, over all internal nodes v of the tree, of the absolute value of the difference of the sizes of the clades defined by the children of v. It is one of the most popular phylogenetic balance indices, because, in addition to measuring the balance of a tree in a very simple and intuitive way, it turns out to be one of the most powerful and discriminating phylogenetic shape indices. But it has some drawbacks. On the one hand, although its minimum value is reached at the so-called maximally balanced trees, it is almost always reached also at trees that are not maximally balanced. On the other hand, its definition as a sum of absolute values of differences makes it difficult to study analytically its distribution under probabilistic models of bifurcating phylogenetic trees. In this paper we show that if we replace in its definition the absolute values of the differences of clade sizes by the squares of these differences, all these drawbacks are overcome and the resulting index is still more powerful and discriminating than the original Colless index.

1. Introduction

Evolutionary biology is concerned, among other major things, about understanding what forces influence speciation and extinction processes, and how they affect macroevolution [1]. In order to do so, there has been a natural interest in the development of techniques and measures whose goal is to assess the imprint of these forces in what has become the standard representation of joint evolutionary histories of groups of species: phylogenetic trees [2–4]. There are two aspects of a phylogenetic tree that can expose such an imprint: its branch lengths – determined by the timing of speciation events – and its *shape*, or *topology* —which, in turn, is determined by the differences in the diversification rates among clades [5, Chap. 33]. But, as it turns out, the accurate reconstruction of branch lengths associating, to a given phylogenetic tree, a robust timeline is not straightforward [6] while, on the other hand, phylogenetic reconstruction methods over the same empirical data tend to agree on the topology of the reconstructed tree [7–9]. Therefore, the shape of phylogenetic trees has become the focus of most of the studies performed on this topic, be it *via* the definition of indices quantifying topological features – see, for instance, [3,10,11] and the references on balance indices given below – or the frequency distribution of small rooted subtrees [12–15].

In his 1922 paper, Yule [16] first observed that taxonomic trees have a tendency towards asymmetry, with most clades being small and only a few of them large at every taxonomic level. Thus, *balance*, understood as the propensity of the children of any given node to have the same number of descendant leaves, has become the most popular topological measure used to describe the topology of a phylogenetic tree. Therefore, *per negationem*, the *imbalance* of a phylogenetic tree gives a measure of the tendency of diversification events to occur mostly along specific lineages [11,17]. Several such measures have been proposed, in order to quantify the balance (or, in many cases, the imbalance) of a phylogenetic tree, and they are referred to in the literature as *balance indices*. For example, see [10–12,18–24] and the section "Measures of overall asymmetry" in [5] (pp. 562–563).

For instance, these indices have then been thoroughly used in order to test the validity evolutionary models [3,21,25–29]; to assess possible biases in the distribution of shapes that are obtained through different phylogenetic tree reconstruction methods [30–34]; to compare tree shapes [35–37]; as a tool to discriminate between input parameters in phylogenetic tree simulations [38,39]; or simply to describe phylogenies existing in the literature [40–43].

Introduced in [18], the *Colless index* has become one of the most popular balance indices in the literature. Given a bifurcating tree T, it is defined as the sum, over all internal nodes v in T, of the absolute value of the difference between the numbers of descendant leaves of the pair of children of v (even so, there exists a recent extension to multifurcating trees, see [23]). Its popularity springs from several

* Corresponding author.
E-mail addresses: krzysztof.bartoszek@liu.se, krzbar@protonmail.ch (K. Bartoszek), t.martinez@uib.eu (T.M. Coronado), arnau.mir@uib.eu (A. Mir), cesc.rossello@uib.es (F. Rosselló).

https://doi.org/10.1016/j.mbs.2020.108503
Received 29 July 2020; Received in revised form 28 October 2020; Accepted 28 October 2020
Available online xxxx
0025-5564/

sources. First of all, its antiquity: it is one of the first balance indices found in the literature, dating back to 1982. Secondly, the way it measures the "global imbalance" by adding the "local imbalances" of each internal node in T is fairly intuitive. Finally, it has been classified as one of the most powerful tree shape statistics in goodness-of-fit tests of probabilistic models of phylogenetic trees [21,44,45], as well as one of the most shape-discriminant balance indices [46].

Due to this popularity, the statistical properties of the Colless index under several probabilistic models have been thoroughly studied [47–50] as well as its maximum [23] and minimum [51] values. The characterization of this last value, as well as that of the trees attaining it, apart from recent turns out to be rather complex and fails to shed light on the intuitive concept of balance. Indeed, other balance indices, such as the total cophenetic index [22] and the rooted quartet index rQI [19] classify as "most balanced" trees only those that are maximally balanced, in the sense that the imbalance of each internal node is either 0 or 1. Even though these trees are effectively considered to be "most balanced" by the Colless index, they are seldom the only ones being so considered.

In this manuscript, we introduce a modification of the Colless index that offers some benefits over the original definition, consisting in squaring the difference of the number of descendant leaves to each child of an internal node instead of considering its absolute value. On the one hand, we have been able to compute both its expected value and its variance under the Yule and uniform probabilistic models for phylogenetic trees. In contrast, notice that the expected value of the Colless index under the uniform model is still unknown in the literature. On the other hand, its maximum and minimum values are attained exactly at the caterpillars and the maximally balanced trees, respectively, and the proofs of these results are rather easy—more so when compared to those concerning the Colless index. Furthermore, it proves to be less prone to have ties between different trees than any other balance index in the literature is, as well as more shape-discriminant than any of the balance indices tested in [46] are.

Before leaving the Introduction, we want to note that, even though the Colless index, as well as other indices, was invented for its application to the description and analysis of phylogenetic trees, it is a shape index, i.e. one whose value does not depend on the specific labels associated to the leaves of the tree, but on its underlying topological features. Thus, in the rest of this manuscript we will restrict ourselves to unlabelled trees.

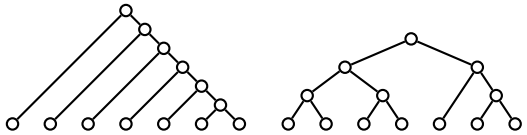
2. Preliminaries

2.1. Trees

In this paper, by a *tree* T we always mean a *bifurcating rooted tree*, that is, a directed tree with one, and only one, node of in-degree 0 (called the *root* of the tree) and all its nodes of out-degree either 0 (the *leaves*, forming the set $L(T)$) or 2 (the *internal nodes*, forming the set $V_{int}(T)$). For every $n \geqslant 1$, we denote by $\mathcal{T}_n^*$ the set of (isomorphism classes of) trees with n leaves.

Let T be a tree. If there exists an edge from a node u to a node v in T, we say that v is a *child* of u and that u is the *parent* of v. Notice that, since T is bifurcating, all internal nodes of T have exactly two children. In addition, if there exists a path from a node u to a node v in T, we say that v is a *descendant* of u. For every node v of T, we denote by $\kappa_T(v)$ the number of its descendant leaves. If $n \geqslant 2$, the *maximal pending subtrees* of T are the pair of subtrees rooted at the children of its root. We shall denote the fact that T_1 and T_2 are the maximal pending subtrees of T by writing $T = T_1 \star T_2$. This notation is commutative, that is $T_1 \star T_2 = T_2 \star T_1$.

For every $n \geqslant 1$, the *comb* with n leaves, K_n, is the unique tree in $\mathcal{T}_n$ all whose internal nodes have different numbers of descendant leaves; cf. Fig. 1(a).



(a) K_7 (b) B_7

Fig. 1. (a) The comb K_7 with 7 leaves; (b) The maximally balanced tree B_7 with 7 leaves.

2.2. The Colless index and the maximally balanced trees

Given a tree T and an internal node $v \in V_{int}(T)$ with children v_1 and v_2, the *balance value* of v is $bal_T(v) = |\kappa_T(v_1) - \kappa_T(v_2)|$. The *Colless index* [18] of a tree $T \in \mathcal{T}_n$ is the sum of the balance values of its internal nodes:

$$C(T) = \sum_{v \in V_{int}(T)} bal_T(v).$$

An internal node v is *balanced* when $bal_T(v) \leqslant 1$, i.e. when its two children have $\lceil \kappa_T(v)/2 \rceil$ and $\lfloor \kappa_T(v)/2 \rfloor$ descendant leaves, respectively. A tree is *maximally balanced* if all its internal nodes are balanced (cf. Fig. 1(b)). Recursively, a bifurcating tree is maximally balanced if its root is balanced and its two maximal pending subtrees are maximally balanced. This easily implies that, for every $n \in \mathbb{N}$, there exists a unique maximally balanced tree with n leaves, which we denote by B_n.

The maximum Colless index in $\mathcal{T}_n^*$ is reached exactly at the *comb* K_n. The fact that $C(K_n)$ is maximum was already hinted at by Colless in [18], but to our knowledge a formal proof that $C(K_n) > C(T)$ for every $T \in \mathcal{T}_n^* \setminus \{K_n\}$ was not provided until [23, Lem. 1]. As to the minimum Colless index in $\mathcal{T}_n^*$, it is proved in [51, Thm. 1] that it is achieved at the maximally balanced tree B_n, although (unlike the situation with the maximum Colless index) for almost every $n \in \mathbb{N}_{\geqslant 1}$ there exist other trees in $\mathcal{T}_n^*$ with minimum Colless index (see [51, Cor. 7]). If we write $n = \sum_{j=1}^{\ell} 2^{m_j}$, with $\ell \geqslant 1$ and $m_1, \ldots, m_\ell \in \mathbb{N}$ such that $m_1 > \cdots > m_\ell$, then

$$C(B_n) = \sum_{j=2}^{\ell} 2^{m_j}(m_1 - m_j - 2(j-2)). \tag{1}$$

For a proof, see Thm. 2 in [51].

2.3. Phylogenetic trees

A *phylogenetic tree* on a set X is a (rooted and bifurcating) tree with its leaves bijectively labelled by the elements of X. We shall denote by $\mathcal{T}_X$ the space of (isomorphism classes of) phylogenetic trees on X. When the specific set of labels X is irrelevant and only its cardinality $|X| = n$ matters, we shall identify X with the set $\{1, \ldots, n\}$, we shall write $\mathcal{T}_n$ instead of $\mathcal{T}_X$, and we shall call the members of this set *phylogenetic trees with n leaves*.

A *probabilistic model of phylogenetic trees* P_n, $n \geqslant 1$, is a family of probability mappings $P_n : \mathcal{T}_n \to [0,1]$, each one sending each phylogenetic tree in $\mathcal{T}_n$ to its probability under this model.

The two most popular probabilistic models of phylogenetic trees are the *Yule*, or *Equal-Rate Markov*, *model* [16,52] and the *uniform*, or *Proportional to Distinguishable Arrangements*, *model* [53,54]. The *Yule model* produces bifurcating phylogenetic trees on $[n]$ through the following stochastic process: starting with a single node, at every step a leaf is chosen randomly and uniformly and it is replaced by a pair of sister leaves; when the desired number n of leaves is reached, the labels are assigned randomly and uniformly to these leaves. The probability $P_{Y,n}(T)$ of each $T \in \mathcal{T}_n$ under this model is the probability of being obtained through this process. As to the *uniform model*, it assigns the same probability to all trees $T \in \mathcal{T}_n$, which is then $P_{U,n} = 1/(2n-3)!!$. For more information on these two models, see [55, §3.2].

3. Main theoretical results

The *Quadratic Colless index*, *Q-Colless index* for short, of a bifurcating tree T is the sum of the squared balance values of its internal nodes:

$$QC(T) = \sum_{v \in V_{int}(T)} bal_T(v)^2 = \sum_{v \in V_{int}(T)} (\kappa_T(v_1) - \kappa_T(v_2))^2,$$

where v_1 and v_2 denote the children of each $v \in V_{int}(T)$.

For instance, the trees depicted in Fig. 1 have Q-Colless indices $QC(K_7) = 55$ and $QC(B_7) = 2$. As we shall see, these are the maximum and minimum values of QC on $\mathcal{T}_7$.

It is straightforward to check that the Q-Colless index satisfies the following recurrence; cf. [56] for the corresponding recurrence for the "classical" Colless index.

Lemma 1. *For every $T \in \mathcal{T}_n^*$, with $n \geqslant 2$, if $T = T_k \star T'_{n-k}$, with $T_k \in \mathcal{T}_k^*$ and $T'_{n-k} \in \mathcal{T}_{n-k}^*$, then*

$$QC(T) = QC(T_k) + QC(T'_{n-k}) + (n-2k)^2.$$

The Colless index and the Q-Colless index satisfy the following relation.

Lemma 2. *For every $T \in \mathcal{T}_n^*$, $QC(T) \geqslant C(T)$ and the equality holds if and only if T is maximally balanced.*

Proof. By definition,

$$QC(T) = \sum_{u \in V_{int}(T)} \mathrm{bal}_T(u)^2 \geqslant \sum_{u \in V_{int}(T)} \mathrm{bal}_T(u) = C(T)$$

because $\mathrm{bal}_T(u) \in \mathbb{N}$ for all $u \in V_{int}(T)$. This inequality is an equality if, and only if, each $\mathrm{bal}_T(u)$ is either 0 or 1, and, by definition, this only happens in the maximally balanced trees. □

3.1. Extremal values

In this subsection we prove that, according to the Q-Colless index, the most balanced trees are exactly the maximally balanced trees and the most unbalanced trees are exactly the combs.

Theorem 3. *The minimum of the Q-Colless index on $\mathcal{T}_n^*$ is always reached at the maximally balanced tree B_n, and only at this tree. Moreover, $QC(B_n) = C(B_n)$ and hence this minimum value is given by Eq. (1).*

Proof. Let $T \in \mathcal{T}_n^*$. By [51, Thm. 1], we know that $C(T) \geqslant C(B_n)$. Therefore, by Lemma 2,

$$QC(T) \geqslant C(T) \geqslant C(B_n) = QC(B_n)$$

and therefore $QC(B_n)$ is minimum on $\mathcal{T}_n^*$. Furthermore, the first inequality is strict if $T \neq B_n$, and therefore $QC(T) > QC(B_n)$ if $T \neq B_n$. □

Theorem 4. *The maximum of the Q-Colless index on $\mathcal{T}_n^*$ is always reached at the comb K_n, and only at this tree, and it is equal to.*

$$QC(K_n) = \binom{n}{3} + \binom{n-1}{3}.$$

Proof. The formula for $QC(K_n)$ comes from the fact that the balance values of the internal nodes of K_n are $\{0, 1, \ldots, n-2\}$ and therefore

$$QC(K_n) = \sum_{i=1}^{n-2} i^2 = \frac{(n-1)(n-2)(2n-3)}{6} = \binom{n}{3} + \binom{n-1}{3}.$$

We prove now the maximality assertion in the statement by induction on the number n of leaves. For $n \in \{1, 2, 3\}$, the assertion is obviously true because in these cases $\mathcal{T}_n^*$ consists of a single tree. Assume now that $n \geqslant 4$ and that, for every $m < n$, $QC(K_m) > QC(T_m)$ for every $T_m \in \mathcal{T}_m^* \setminus \{K_m\}$. Let $T \in \mathcal{T}_n^*$ and let T_{n_1} and T_{n-n_1} be its two maximal pending subtrees, with $T_{n_1} \in \mathcal{T}_{n_1}^*$ and $T_{n-n_1} \in \mathcal{T}_{n-n_1}^*$ and, say, $n_1 \leqslant n/2$. In this way, by Lemma 1,

$$QC(T) = QC(T_{n_1}) + QC(T_{n-n_1}) + (n-2n_1)^2.$$

We want to prove that $QC(K_n) \geqslant QC(T)$ and that the equality holds only when $T = K_n = K_1 \star K_{n-1}$. Since, by induction, $QC(K_{n_1}) \geqslant QC(T_{n_1})$ and $QC(K_{n-n_1}) \geqslant QC(T_{n-n_1})$ and the corresponding equalities hold only when $T_{n_1} = K_{n_1}$ and $T_{n-n_1} = K_{n-n_1}$, it is enough to prove that

$$QC(K_n) \geqslant QC(K_{n_1}) + QC(K_{n-n_1}) + (n-2n_1)^2,$$

i.e., that

$$\frac{(n-1)(n-2)(2n-3)}{6} \geqslant \frac{(n_1-1)(n_1-2)(2n_1-3)}{6} + \frac{(n-n_1-1)(n-n_1-2)(2n-2n_1-3)}{6} + (n-2n_1)^2,$$

for every $1 \leqslant n_1 \leqslant n/2$, and that the equality holds only when $n_1 = 1$.

Consider now the function $\kappa : [1, n/2] \to \mathbb{R}$, defined as

$$\kappa(x) = \frac{1}{6}\Big((x-1)(x-2)(2x-3) + (n-x-1)(n-x-2)(2n-2x-3) + 6(n-2x)^2\Big)$$
$$= (n+1)x^2 - n(n+1)x + \frac{1}{6}\left(2n^3 - 3n^2 + 13n - 12\right)$$

The graph of this function is a convex parabola with vertex at $x = n/2$. Therefore, the maximum value of κ on the interval $[1, n/2]$ is reached at $x = 1$, which is exactly what we wanted to prove. □

By [51, Cor. 5],

$$QC(B_n) = C(B_n) < \min\{n/2, 2^{\lceil \log_2(n) \rceil}/3\}$$

and therefore the range of values of QC on $\mathcal{T}_n^*$ goes from below this bound to $\binom{n}{3} + \binom{n-1}{3}$ and hence its width grows in $n^3/3$, one order of magnitude larger than the range of the Colless index.

3.2. Statistics under the uniform and the Yule model

Let QC_n be the random variable that chooses a phylogenetic tree $T \in \mathcal{T}_n$ and computes $QC(T)$.

Theorem 5. *For every $n \geqslant 1$:*

(a) The expected value of QC_n under the uniform model is

$$E_U(QC_n) = \binom{n+1}{2} \cdot \frac{(2n-2)!!}{(2n-3)!!} - n(2n-1).$$

(b) The variance of QC_n under the uniform model is

$$\sigma_U^2(QC_n) = \frac{2}{15}(2n-1)(7n^2+9n-1)\binom{n+1}{2} - \frac{1}{8}(5n^2+n+2)\binom{n+1}{2}\frac{(2n-2)!!}{(2n-3)!!} - \binom{n+1}{2}^2\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2.$$

Regarding the Yule model, we have the following result. In it, H_n and $H_n^{(2)}$ denote, respectively, the nth *harmonic number* and *second order harmonic number*:

$$H_n = \sum_{i=1}^{n} \frac{1}{i}, \quad H_n^{(2)} = \sum_{i=1}^{n} \frac{1}{i^2}.$$

Theorem 6. *For every $n \geqslant 1$:*

(a) The expected value of QC_n under the Yule model is

$$E_Y(QC_n) = n(n+1) - 2nH_n.$$

(b) The variance of QC_n under the Yule model is

$$\sigma_Y^2(QC_n) = \frac{1}{3}n\left(n^3 - 8n^2 + 50n - 1 - 30H_n - 12nH_n^{(2)}\right).$$

We prove these theorems in the Appendix at the end of the paper.

Using Stirling's approximation for large factorials it is easy to prove that

$$\frac{(2n-2)!!}{(2n-3)!!} \sim \sqrt{\pi n}$$

(see, for instance, [57, Rem. 2]). Moreover, it is known (see, for instance, [58]) that

$$H_n \sim \ln(n), \quad H_n^{(2)} \sim \frac{\pi^2}{6}.$$

Then, from the last two theorems we obtain the following limit behaviours:

$$\begin{aligned} E_U(QC_n) &\sim \frac{\sqrt{\pi}}{2}n^{5/2} & \sigma_U(QC_n) &\sim \sqrt{\frac{14}{15}}n^{5/2} \\ E_Y(QC_n) &\sim n^2 & \sigma_Y(QC_n) &\sim \frac{1}{\sqrt{3}}n^2 \end{aligned}$$

So, both under the Yule and the uniform models, the Q-Colless index satisfies that the expected value and the standard deviation grow with n in the same order. This is in contrast with the Colless index, for which it only happens under the uniform model (see [47] for details):

$$\begin{aligned} E_U(C_n) &\sim \sqrt{\pi}n^{3/2} & \sigma_U(C_n) &\sim \sqrt{\frac{10-3\pi}{3}}n^{3/2} \\ E_Y(C_n) &\sim n\log(n) & \sigma_Y(C_n) &\sim \sqrt{\frac{18-6\log(2)-\pi^2}{6}}n. \end{aligned}$$

3.3. Limit distribution under the Yule distribution

Let us now consider the following Yule-normalized version of the random variable QC_n:

$$Y_n = \frac{QC_n - E_Y(QC_n)}{n^2},$$

where notice that the denominator n^2 is the order of growth of $\sigma_Y(QC_n)$.

The limit distribution of this random variable under the Yule model satisfies the following theorem.

Theorem 7. *As $n \to \infty$, the distribution under the Yule model of Y_n tends to the distribution of a random variable Y satisfying the following equality in distribution:*

$$Y \overset{D}{=} \tau^2 Y' + (1-\tau)^2 Y'' + (1 + 6\tau^2 - 6\tau),$$

where $\tau \sim \mathrm{Unif}[0,1]$ and Y', Y'' are independent and distributed according to the same law as Y.

We also postpone the proof of this theorem to the Appendix at the end of the paper. It is interesting to compare the formula obtained in this theorem with the formula of the limit distribution of the corresponding normalization of the Colless index,

$$Z_n = \frac{C_n - E_Y(C_n)}{n}.$$

Blum, François and Janson proved in [47] that as $n \to \infty$, the distribution under the Yule model of Z_n tends to the distribution of a random variable Z such that

$$Z \overset{D}{=} \tau Z' + (1-\tau)Z'' + \tau\log(\tau) + (1-\tau)\log(1-\tau) + 1 - 2\min(\tau, 1-\tau), \quad (2)$$

where $\tau \sim \mathrm{Unif}[0,1]$ and Z', Z'' are independent and distributed according to the same law as Z. So, the independent term of the recurrent equation for the limit distribution is much simpler for the Q-Colless index than for the Colless index.

We have not carried out here a similar study for the uniform model, because in this case the random variable L_n that chooses a tree in $\mathcal{T}_n$, and then chooses a maximal pending subtree of it and counts the number of leaves of the latter (cf. A.3) has a more complicated behaviour than in the Yule case (cf. the $\beta \leqslant -1$ case in [59, Lemma 3]). Similarly, the weak convergence results obtained in [60] cannot be carried over to this model (see Remark 4.2 in *loc. cit.*).

Table 1
Range of values of the Colless index C, the Sackin index S, the total cophenetic index Φ and the Q-Colless index QC; recall that a function $f(n)$ is in $\Theta(n^k)$ when there exist constants $0 < c_1 \leqslant c_2$ such that, for large enough values of n, $c_1 \cdot n^k \leqslant f(n) \leqslant c_2 \cdot n^k$.

Index	Minimum	Maximum
C	$\Theta(n)$	$\binom{n-1}{2}$
S	$\Theta(n\log(n))$	$\binom{n+1}{2} - 1$
Φ	$\Theta(n^2)$	$\binom{n}{3}$
QC	$\Theta(n)$	$\binom{n}{3} + \binom{n-1}{3}$

4. Numerical results

4.1. Discriminative power of QC

Since the range of values of the Q-Colless index on $\mathcal{T}_n^*$ is wider than those of the Colless index C, the Sackin index S [11,24] or the total cophenetic index Φ (see Table 1), our intuition told us that the probability of two trees with the same number of leaves having the same Q-Colless index would be smaller than for these other balance indices. To simplify the language, when a balance index I takes the same value on two trees in the same space $\mathcal{T}_n^*$, we call it a *tie*. Of course, since for $n \geqslant 12$ the range of possible QC values is narrower than the number of trees in $\mathcal{T}_n^*$ (see [5, Table 3.3] for the cardinality of $\mathcal{T}_n^*$ for small values of n), the pigeonhole principle implies that the Q-Colless index cannot avoid ties for large numbers of leaves.

To check the discriminative power of QC with respect to C, S, and Φ, we have computed the probability of tie $p_n(I)$ for these four balance indices I and for number of leaves of n between $n = 4$ and $n = 20$.

More concretely, first of all, for every balance index $I = C, S, \Phi, QC$ and for $n = 4, \ldots, 20$, we have considered all pairs of different trees (T_1, T_2) in $\mathcal{T}_n^* \times \mathcal{T}_n^*$ and we have calculated the number n_I of such pairs of trees such that $I(T_1) = I(T_2)$. Finally, we have computed the probability $p_n(I)$ as $p_n(I) = \frac{n_I}{\binom{|\mathcal{T}_n^*|}{2}}$, where $|\mathcal{T}_n^*|$ is the cardinal of the set $\mathcal{T}_n^*$.

The results obtained are shown in Fig. 2. The Q-Colless balance index is the balance index with the least probability of a tie.

In relation with this last point, another way to assess the discriminating skill of an index is to evaluate its power to distinguish between dissimilar trees, and compare it with that of other shape indices. In their paper [46], the authors (whom we thank for their support with the software provided in the article) develop a new resolution function to evaluate the power of tree shape statistics when it comes to discriminate between dissimilar trees (based on the Laplacian matrix of the tree, which allows for less spatial and time complexity in the operations), and then test it together with the usual resolution function based on the NNI metric. Therefore, they are able to rank some balance indices according to their power in discriminating all possible phylogenetic trees on the same number of leaves.

We have performed the same experiment on the same data (which was provided along with [46]). It turns out that the QC performs better than all the other tested indices do, including the Saless index [46], a linear combination of the Sackin and Colless indices which was introduced in the same article and performed best when tested under the NNI metric— although not with the resolution function proposed in the article, under which it was the Colless index that performed better. We present here the two tables (Tables 2 and 3), the first of them computing the score under the NNI distance (bigger values represent more power), and the second one under their proposed resolution function (lower values represent more power).

Table 2
Scaled resolution scores for shape indices on the NNI distance matrix for different numbers of leaves n. The value of the resolution is between 0 and 1. Higher values represent more discriminating power.

n	Colless	Sackin	Variance	I_2	B_1	B_2	Saless	Q-Colless
5	1	1	1	1	1	1	1	1
6	0.8157	0.8510	0.8144	0.7611	0.7546	0.8705	0.8315	0.8709
7	0.9251	0.9303	0.9023	0.8844	0.8649	0.9254	0.9297	0.9360
8	0.9255	0.9122	0.8753	0.8612	0.8326	0.9113	0.9235	0.9218
9	0.9184	0.9208	0.8826	0.8539	0.8324	0.907	0.9224	0.9302
10	0.941	0.9380	0.8985	0.8545	0.8326	0.9085	0.9426	0.9475
11	0.9531	0.9514	0.9102	0.8552	0.8375	0.9132	0.9551	0.9604
12	0.9533	0.9523	0.9086	0.8504	0.8311	0.9045	0.9556	0.9632
13	0.9541	0.9542	0.9078	0.8416	0.8247	0.8992	0.9567	0.9657
14	0.9552	0.9548	0.9070	0.8374	0.82	0.8902	0.9575	0.967
15	0.9546	0.9544	0.9049	0.8298	0.813	0.8826	0.9569	0.9674
16	0.9543	0.9541	0.9034	0.8265	0.8089	0.8743	0.9564	0.9677
17	0.9534	0.9534	0.9006	0.8199	0.8024	0.8678	0.9555	0.9679

Table 3
Scaled resolution scores for shape indices on the resolution function presented in [46]. The value of the resolution is between 0 and 1. Lower values represent more discriminating power.

n	Colless	Sackin	Variance	I_2	B_1	B_2	Q-Colless
7	0.0984	0.0937	0.1082	0.1115	0.1178	0.0989	0.0948
8	0.0808	0.0955	0.111	0.0893	0.1164	0.0965	0.0941
9	0.0507	0.0566	0.0662	0.068	0.0797	0.0653	0.0558
10	0.0327	0.0379	0.0471	0.0535	0.0629	0.0451	0.0357
11	0.0222	0.0255	0.0326	0.0458	0.0511	0.0348	0.0236
12	0.0183	0.0217	0.0282	0.0429	0.0473	0.0304	0.0194
13	0.016	0.0185	0.0238	0.0413	0.0441	0.0283	0.0163
14	0.0147	0.0170	0.0217	0.04	0.0421	0.0265	0.0147
15	0.0137	0.0157	0.0197	0.039	0.0404	0.0256	0.0134
16	0.013	0.0148	0.0184	0.038	0.0389	0.0247	0.0126
17	0.0123	0.014	0.017	0.037	0.0375	0.0238	0.0118
18	0.0117	0.0132	0.016	0.0358	0.0361	0.0229	0.0111
19	0.0112	0.0127	0.015	0.0347	0.0349	0.0222	0.0105
20	0.0107	0.012	0.0141	0.0339	0.0338	0.0217	0.01
21	0.0102	0.0114	0.0133	0.0329	0.0327	0.0209	0.01

4.2. Limit behaviour of QC

We can use Theorem 7 to simulate directly the behaviour of QC in the limit, using Algorithm 3 in [61]. We have simulated 10000 values of Y using this algorithm, with recursion depth 10. It has to be noted that the drawn number will come from the limiting distribution only when the recursion depth is infinite. We have also generated 10000 independent phylogenetic trees in $\mathcal{T}_{1000}$ under the Yule model, and calculated their Yule normalized Q-Colless indices

$$Y(T) = \frac{QC(T) - E_Y(QC_n)}{n^2}.$$

The corresponding histograms are shown in Fig. 3. As it can be seen, they are indistinguishable. Performing a t-test on the two simulated samples yields a p-value of 0.2315, and an F-test (R's `var.test()`), to compare the sample variances, yields a p-value of 0.199. We can also notice that both sample variances are close to $1/3$, as they should be according to Theorem 6.

In Fig. 3 we also compare the (Yule-normalized) Q-Colless index limit distribution with the limit distribution of the (Yule-normalized) Colless index simulated by means of Eq. (2). We can see that the Colless index is more symmetric and also has wider support than the Q-Colless index. This in turn implies that a statistical test based on the Q-Colless index should have better power at detecting deviations from the Yule model, in line with the conclusions concerning its discriminating abilities given in the previous subsection. Interestingly, the histogram of the limit of the normalized Q-Colless index exhibits a similar skewness as the limit of the normalized total cophenetic index (cf. Fig. 2 in [61], although that figure is not directly comparable with our Fig. 3 as there the simulated points were also normalized by the leading constant of the variance).

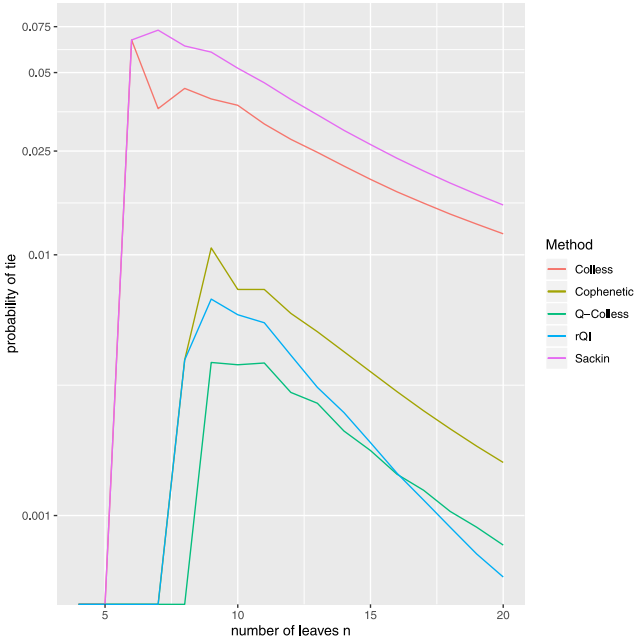


Fig. 2. Probability of tie using the Colless, Total cophenetic, Quadratic Colless, Rooted Quartet (rQI), and Sackin balance indices as function of the trees' number of leaves n, for $n = 4, \ldots, 20$..

5. Conclusions

The Colless index [18] is one of the oldest and most popular balance indices appearing in the literature. Its number of cites more than doubles that of the second most cited balance index in Google Scholar, the Sackin index. Nevertheless, it presents some drawbacks related to the difficult characterization of the trees that achieve its minimum value – which clashes with the intuition that only the maximally balanced trees should be considered the most balanced bifurcating trees – and the fact that its moments under one of the most widely used probabilistic models for bifurcating phylogenetic trees, the uniform model, are still unknown.

In this paper we have presented an alternative to the Colless index that captures both its intuitive definition and its statistical benefits. In the first part of this manuscript we have proved that its extremal values are attained exactly by the trees that are usually considered to be the "most" and "least" balanced family of bifurcating trees, respectively. This contrasts vividly with the Colless and Sackin indices, whose minimum value, although being always reached by the maximally balanced trees, is seldom attained only by it; although the Colless index was defined in 1982 [18], these characterizations have been only very recently found [51,62]. We have thus shown that the range of values of the Quadratic Colless index, $O(n^3)$, is bigger than that of the original Colless index, $O(n^2)$, on pair with that of the total cophenetic index.

Then, we have proceeded to the computation of both the expected value and the variance under the Yule and the Uniform models of the Q-Colless index. We want to remark to the reader that the expected value and the variance of the Colless index in its original definition are, under the uniform model, still unknown. So, in this regard the Quadratic Colless index presents an improvement over the original measure of balance. We have also obtained a recurrent equation for the limit distribution of the Q-Colless index under the Yule model.

Finally, we have empirically shown that it possesses more discriminatory power than the original Colless index does by, firstly, computing the probability of producing a tie between a pair of trees for numbers of leaves up to 20, and, secondly, testing it under the metrics provided in [46]. In both cases, it has systematically been one of the best performing measures, being often superior to the Colless and Sackin

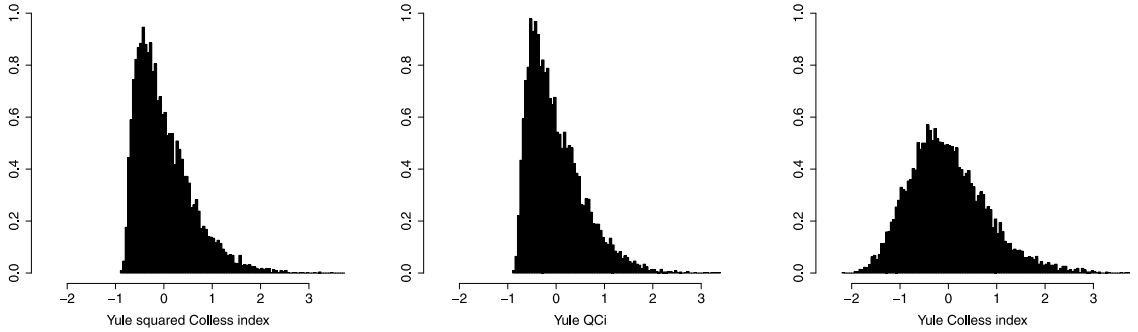


Fig. 3. Left: histogram of 10,000 normalized Q-Colless indices of trees in $\mathcal{T}_{1000}$ independently generated under the Yule model (sample mean 0.008, sample variance 0.333). Centre: histogram 10,000 simulations from the limit distribution under the Yule model (Eq. (12)) of the normalized Q-Colless index on $\mathcal{T}_{1000}$ (sample mean −0.002, sample variance 0.325). Right: histogram 10,000 simulations from the limit distribution under the Yule model (Eq. (12)) (Eq. (2)) of the normalized Colless index on $\mathcal{T}_{1000}$ (sample mean −0.012, sample variance 0.654). The simulations were done in R 3.6.1 running on an openSUSE Leap 42.3 operating system. Algorithm 3 in [61] was used for the centre and right panel with a recursion depth of 10.

indices. All scripts and data used in these computations are available at the GitHub repository associated to this paper (https://github.com/biocom-uib/QColless).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

TMC, AM and FR were partially supported by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund through project PGC2018-096956-B-C43 (MINECO/FEDER). KB was partially supported by the Swedish Research Council (Vetenskapsrådet) grant no. 2017–04951.

Appendix

A.1. Proof of Theorem 6

The following lemma summarizes Lemma 16 in [22] and Lemma 2 in [48].

Lemma 8. *Let $I : \bigcup_{n\geqslant 1}\mathcal{T}_n \to \mathbb{R}$ be a mapping satisfying the following two conditions:*

- *It is invariant under phylogenetic tree isomorphisms and relabellings of leaves.*
- *There exists a symmetric mapping $f_I : \mathbb{N}_{n\geqslant 1}\times\mathbb{N}_{n\geqslant 1} \to \mathbb{R}$ such that, for every pair of phylogenetic trees T, T' on disjoint sets of taxa X, X', respectively,*

$$I(T \star T') = I(T) + I(T') + f_I(|X|, |X'|).$$

For every $n \geqslant 1$, let I_n and I_n^2 be the random variables that choose a tree $T \in \mathcal{T}_n$ and compute $I(T)$ and $I(T)^2$, respectively. Then, for every $n \geqslant 2$, their expected values under the Yule model are:

$$E_Y(I_n) = \frac{1}{n-1}\sum_{k=1}^{n-1}\big(2E_Y(I_k) + f_I(k, n-k)\big)$$

$$E_Y(I_n^2) = \frac{1}{n-1}\sum_{k=1}^{n-1}\Big(2E_Y(I_k^2) + 4f_I(k,n-k)E_Y(I_k) + 2E_Y(I_k)E_Y(I_{n-k}) + f_I(k,n-k)^2\Big).$$

Claim 1. *For every $n \geqslant 1$, the expected value of QC_n under the Yule model is*

$$E_Y(QC_n) = n(n+1) - 2nH_n.$$

Proof. By Lemma 8(a),

$$\begin{aligned}
E_Y(QC_n) &= \frac{2}{n-1}\sum_{k=1}^{n-1}E_Y(QC_k) + \frac{1}{n-1}\sum_{k=1}^{n-1}(n-2k)^2\\
&= \frac{2}{n-1}\sum_{k=1}^{n-1}E_Y(QC_k) + \frac{1}{3}n(n-2)\\
&= \frac{2}{n-1}E_Y(QC_{n-1}) + \frac{n-2}{n-1}\Big(\frac{2}{n-2}\sum_{k=1}^{n-2}E_Y(QC_k)\Big) + \frac{1}{3}n(n-2)\\
&= \frac{2}{n-1}E_Y(QC_{n-1}) + \frac{n-2}{n-1}\Big(E_Y(QC_{n-1}) - \frac{1}{3}(n-1)(n-3)\Big)\\
&\quad + \frac{1}{3}n(n-2)\\
&= \frac{n}{n-1}E_Y(QC_{n-1}) + n - 2
\end{aligned}$$

Dividing this equation by n and setting $X_n = E_Y(QC_n)/n$, we obtain the equation

$$X_n = X_{n-1} + 1 - \frac{2}{n}$$

whose solution with initial condition $X_1 = E_Y(QC_1) = 0$ is

$$X_n = \sum_{k=2}^{n}\Big(1 - \frac{2}{k}\Big) = n + 1 - 2H_n$$

and hence, finally,

$$E_Y(QC_n) = nX_n = n(n+1) - 2nH_n. \quad \square$$

Claim 2. *For every $n \geqslant 1$, the variance of QC_n under the Yule model is*

$$\sigma_Y^2(QC_n) = \frac{1}{3}n\big(n^3 - 8n^2 + 50n - 1 - 30H_n - 12nH_n^{(2)}\big).$$

Proof. We shall compute the variance $\sigma_Y^2(QC_n)$ by means of the identity

$$\sigma_Y^2(QC_n) = E_Y(QC_n^2) - E_Y(QC_n)^2 \tag{3}$$

where the value of $E_Y(QC_n)$ is given by Claim 1. What remains is to compute $E_Y(QC_n^2)$. Now, by Lemma 8(b),

$$\begin{aligned}
E_Y(QC_n^2) &= \frac{1}{n-1}\sum_{k=1}^{n-1}\Big(2E_Y(QC_k^2) + (n-2k)^4\\
&\quad + 4(n-2k)^2E_Y(QC_k) + 2E_Y(QC_k)E_Y(QC_{n-k})\Big)\\
&= \frac{2}{n-1}\sum_{k=1}^{n-1}E_Y(QC_k^2) + \frac{1}{n-1}\sum_{k=1}^{n-1}(n-2k)^4
\end{aligned}$$

$$+\frac{4}{n-1}\sum_{k=1}^{n-1}(n-2k)^2k(k+1-2H_k)$$
$$+\frac{2}{n-1}\sum_{k=1}^{n-1}k(n-k)(k+1-2H_k)(n-k+1-2H_{n-k})$$

Let us denote by T_n the independent term in this equation, so that this equation can be written as

$$\begin{aligned}E_Y(QC_n^2)&=\frac{2}{n-1}\sum_{k=1}^{n-1}E_Y(QC_k^2)+T_n\\&=\frac{2}{n-1}E_Y(QC_{n-1}^2)+\frac{n-2}{n-1}\cdot\frac{2}{n-2}\sum_{k=1}^{n-2}E_Y(QC_k^2)+T_n\\&=\frac{2}{n-1}E_Y(QC_{n-1}^2)+\frac{n-2}{n-1}(E_Y(QC_{n-1}^2)-T_{n-1})+T_n\\&=\frac{n}{n-1}E_Y(QC_{n-1}^2)+T_n-\frac{n-2}{n-1}T_{n-1}\end{aligned}$$

Dividing this equation by n and setting $Y_n=E_Y(QC_n^2)/n$, we obtain the equation

$$Y_n=Y_{n-1}+\frac{1}{n}\Big(T_n-\frac{n-2}{n-1}T_{n-1}\Big). \tag{4}$$

We want to compute now the independent term in this equation as an explicit expression in n. To do that, we first compute the three sums that form T_n. On the one hand,

$$\frac{1}{n-1}\sum_{k=1}^{n-1}(n-2k)^4=\frac{1}{15}n(n-2)(3n^2-6n-4). \tag{5}$$

On the other hand,

$$\begin{aligned}&\frac{4}{n-1}\sum_{k=1}^{n-1}(n-2k)^2k(k+1-2H_k)\\&=\frac{4}{n-1}\Bigg(\sum_{k=1}^{n-1}(n-2k)^2k(k+1)-2(n-2)^2\sum_{k=1}^{n-1}kH_k\\&\qquad+16(n-3)\sum_{k=1}^{n-1}\binom{k}{2}H_k-48\sum_{k=1}^{n-1}\binom{k}{3}H_k\Bigg)\\&=\frac{4}{n-1}\Bigg(\frac{1}{15}(n-1)n(n+1)(2n^2-5n+2)-2(n-2)^2\binom{n}{2}\Big(H_n-\frac{1}{2}\Big)\\&\qquad+16(n-3)\binom{n}{3}\Big(H_n-\frac{1}{3}\Big)-48\binom{n}{4}\Big(H_n-\frac{1}{4}\Big)\Bigg)\\&=\frac{2}{45}n(n-2)(12n^2+16n+9)-\frac{4}{3}n^2(n-2)H_n\end{aligned} \tag{6}$$

using, in the second last equality above, that

$$\sum_{k=1}^{n-1}\binom{k}{m}H_k=\binom{n}{m+1}\Big(H_n-\frac{1}{m+1}\Big); \tag{7}$$

see Eq. (6.70) in [58].

As to the third sum,

$$\begin{aligned}&\frac{2}{n-1}\sum_{k=1}^{n-1}k(n-k)(k+1-2H_k)(n-k+1-2H_{n-k})\\&=\frac{2}{n-1}\Bigg[\sum_{k=1}^{n-1}k(k+1)(n-k)(n-k+1)\\&\qquad-2\sum_{k=1}^{n-1}k(n-k)(n-k+1)H_k-2\sum_{k=1}^{n-1}k(n-k)(k+1)H_{n-k}\\&\qquad+4\sum_{k=1}^{n-1}k(n-k)H_kH_{n-k}\Bigg]\\&=\frac{2}{n-1}\Bigg[\sum_{k=1}^{n-1}k(k+1)(n-k)(n-k+1)-4\sum_{k=1}^{n-1}k(n-k)(n-k+1)H_k\\&\qquad+4n\sum_{k=1}^{n-1}kH_kH_{n-k}-4\sum_{k=1}^{n-1}k^2H_kH_{n-k}\Bigg]\\&=\frac{2}{n-1}\Bigg[\sum_{k=1}^{n-1}k(k+1)(n-k)(n-k+1)\\&\qquad-4\sum_{k=1}^{n-1}\Big(6\binom{k}{3}-4(n-1)\binom{k}{2}+n(n-1)k\Big)H_k\\&\qquad+4n\sum_{k=1}^{n-1}kH_kH_{n-k}-4\sum_{k=1}^{n-1}k^2H_kH_{n-k}\Bigg]\\&=\frac{2}{n-1}\Bigg[4\binom{n+3}{5}-24\binom{n}{4}\Big(H_n-\frac{1}{4}\Big)\\&\qquad+16(n-1)\binom{n}{3}\Big(H_n-\frac{1}{3}\Big)-4n(n-1)\binom{n}{2}\Big(H_n-\frac{1}{2}\Big)\\&\qquad+4n\binom{n+1}{2}(H_{n+1}^2-H_{n+1}^{(2)}-2H_{n+1}+2)\\&\qquad-\frac{4}{3}\binom{n+1}{2}\Big((2n+1)(H_{n+1}^2-H_{n+1}^{(2)})\\&\qquad-\frac{13n+5}{3}H_{n+1}+\frac{71n+37}{18}\Big)\Bigg]\\&=\frac{1}{270}n(18n^3+303n^2+1163n+98)-\frac{2}{9}n(n+1)(3n+16)H_n\\&\qquad+\frac{4}{3}n(n+1)(H_{n+1}^2-H_{n+1}^{(2)})\end{aligned} \tag{8}$$

using, in the second last equality above, Eq. (7) and the identities

$$\sum_{k=1}^{n-1}kH_kH_{n-k}=\binom{n+1}{2}(H_{n+1}^2-H_{n+1}^{(2)}-2H_{n+1}+2)$$
$$\sum_{k=1}^{n-1}k^2H_kH_{n-k}=\frac{n(n+1)}{6}\Big[(2n+1)(H_{n+1}^2-H_{n+1}^{(2)})-\frac{13n+5}{3}H_{n+1}+\frac{71n+37}{18}\Big]$$

proved in [63].

So,

$$\begin{aligned}T_n&=\frac{1}{15}n(n-2)(3n^2-6n-4)\\&\qquad+\frac{2}{45}n(n-2)(12n^2+16n+9)-\frac{4}{3}n^2(n-2)H_n\\&\qquad+\frac{1}{270}n(18n^3+303n^2+1163n+98)-\frac{2}{9}n(n+1)(3n+16)H_n\\&\qquad+\frac{4}{3}n(n+1)(H_{n+1}^2-H_{n+1}^{(2)})\\&=\frac{1}{270}n(216n^3-9n^2+1031n+26)-\frac{2}{9}n(9n^2+7n+16)H_n\\&\qquad+\frac{4}{3}n(n+1)(H_{n+1}^2-H_{n+1}^{(2)})\end{aligned}$$

and, hence, the independent term in Eq. (4) is

$$\begin{aligned}&\frac{1}{n}\Big(T_n-\frac{n-2}{n-1}T_{n-1}\Big)\\&=\frac{1}{n}\Bigg[\frac{1}{270}n(216n^3-9n^2+1031n+26)-\frac{2}{9}n(9n^2+7n+16)H_n\\&\qquad+\frac{4}{3}n(n+1)(H_{n+1}^2-H_{n+1}^{(2)})\\&\qquad-\frac{n-2}{n-1}\Bigg(\frac{1}{270}(n-1)(216(n-1)^3-9(n-1)^2+1031(n-1)+26)\\&\qquad-\frac{2}{9}(n-1)(9(n-1)^2+7(n-1)+16)H_{n-1}\\&\qquad+\frac{4}{3}(n-1)n(H_n^2-H_n^{(2)})\Bigg)\Bigg]\\&=\frac{1}{n}\Bigg[\frac{1}{270}n(216n^3-9n^2+1031n+26)\\&\qquad-\frac{2}{9}n(9n^2+7n+16)H_{n-1}-\frac{2}{9}(9n^2+7n+16)\\&\qquad+\frac{4}{3}n(n+1)(H_n^2-H_n^{(2)})+\frac{8}{3}nH_{n-1}+\frac{8}{3}\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{270}(n-2)(216n^3 - 657n^2 + 1697n - 1230) \\
&+ \frac{2}{9}(n-2)(9n^2 - 11n + 18)H_{n-1} \\
&- \frac{4}{3}(n-2)n(H_n^2 - H_n^{(2)})\Big] \\
&= \frac{1}{n}\Big(\frac{1}{3}(12n^3 - 28n^2 + 47n - 30) - 8(n^2 - n + 1)H_{n-1} \\
&\quad + 4n(H_n^2 - H_n^{(2)})\Big) \\
&= 4n^2 - \frac{28}{3}n + \frac{47}{3} - \frac{10}{n} - 8(n-1)H_{n-1} - \frac{8H_{n-1}}{n} + 4H_n^2 - 4H_n^{(2)}
\end{aligned}
$$

The solution of Eq. (4) with initial condition $Y_1 = E_Y(QC_1^2) = 0$ is

$$
\begin{aligned}
Y_n &= \sum_{k=2}^{n} \frac{1}{k}\Big(T_k - \frac{k-2}{k-1}T_{k-1}\Big) \\
&= \sum_{k=2}^{n}\Big(4k^2 - \frac{28}{3}k + \frac{47}{3} - \frac{10}{k} - 8(k-1)H_{k-1} - \frac{8H_{k-1}}{k} + 4H_k^2 - 4H_k^{(2)}\Big) \\
&= \sum_{k=1}^{n-1}\Bigg(4(k+1)^2 - \frac{28}{3}(k+1) + \frac{47}{3} - \frac{10}{k+1} \\
&\quad - 8kH_k - \frac{8H_k}{k+1} + 4H_{k+1}^2 - 4H_{k+1}^{(2)}\Bigg) \\
&\overset{(*)}{=} \frac{1}{3}(4n^3 - 8n^2 + 35n - 31) - 10(H_n - 1) \\
&\quad - 8\binom{n}{2}\Big(H_n - \frac{1}{2}\Big) - 4(H_n^2 - H_n^{(2)}) \\
&\quad + 4\big((n+1)H_n^2 - (2n+1)H_n + 2n - 1\big) \\
&\quad - 4\big((n+1)H_n^{(2)} - H_n - 1\big) \\
&= \frac{1}{3}(4n^3 - 2n^2 + 53n - 1) - 2(2n^2 + 2n + 5)H_n + 4n(H_n^2 - H_n^{(2)})
\end{aligned}
$$

where, in the second last equality (marked with (*)) we have used Eq. (7) and the identities

$$\sum_{k=1}^{n-1} \frac{H_k}{k+1} = \frac{1}{2}(H_n^2 - H_n^{(2)})$$

(cf. Eqn. (6.71) in [58]) and

$$\sum_{k=1}^{n-1} H_k^2 = nH_n^2 - (2n+1)H_n + 2n$$

$$\sum_{k=1}^{n-1} H_k^{(2)} = nH_n^{(2)} - H_n$$

(see [64, §1.2.7]).

Therefore, finally

$$
\begin{aligned}
E_Y(QC_n^2) &= nY_n \\
&= \frac{n}{3}(4n^3 - 2n^2 + 53n - 1) - 2n(2n^2 + 2n + 5)H_n + 4n^2(H_n^2 - H_n^{(2)})
\end{aligned}
$$

and

$$
\begin{aligned}
\sigma_Y^2(QC_n) &= E_Y(QC_n^2) - E_Y(QC_n)^2 \\
&= \frac{1}{3}n\big(n^3 - 8n^2 + 50n - 1 - 30H_n - 12nH_n^{(2)}\big)
\end{aligned}
$$

as we claimed. □

A.2. Proof of Theorem 5

To simplify the notations, for every $n \geqslant 2$ and for every $1 \leqslant k \leqslant n-1$, set

$$C_{k,n-k} := \frac{1}{2}\binom{n}{k}\frac{(2k-3)!!(2(n-k)-3)!!}{(2n-3)!!}.$$

The proof of the following lemma is identical to the proof of Lemma 8 given in the references provided in the previous subsection, simply replacing the probabilities under the Yule model by probabilities under the uniform model. We leave the details to the reader.

Lemma 9. *Let $I : \bigcup_{n\geqslant 1}\mathcal{T}_n \to \mathbb{R}$ be a mapping satisfying the same conditions as in the statement of Lemma 8 and, for every $n \geqslant 1$, let I_n and I_n^2 be the random variables that choose a tree $T \in \mathcal{T}_n$ and compute $I(T)$ and $I(T)^2$, respectively. Then, for every $n \geqslant 2$, their expected values under the uniform model are:*

$$E_U(I_n) = \sum_{k=1}^{n-1} C_{k,n-k}\big(2E_U(I_k) + f_I(k,n-k)\big)$$

$$
\begin{aligned}
E_U(I_n^2) = \sum_{k=1}^{n-1} C_{k,n-k}\Big(&2E_U(I_k^2) + f_I(k,n-k)^2 \\
&+4f_I(k,n-k)E_U(I_k) + 2E_U(I_k)E_U(I_{n-k}).\Big)
\end{aligned}
$$

In the proofs provided in this subsection we shall use the following technical lemmas. They are proved in the Section SN-4 of the Supplementary Material of [57]; Lemma 12 is Proposition 6 in that paper.

Lemma 10. *For every $n \geqslant 2$:*

(a) $$\sum_{k=1}^{n-1} C_{k,n-k} = 1$$

(b) For every $m \geqslant 1$,

$$\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{m} = \frac{1}{2}\binom{n}{m}\Big(1 - \frac{m-1}{n-1}\cdot\frac{(2m-3)!!}{(2m-2)!!}\cdot\frac{(2n-2)!!}{(2n-3)!!}\Big).$$

Lemma 11. *For every $n \geqslant 2$,*

(a) $$\sum_{k=1}^{n-1} C_{k,n-k}\cdot\frac{(2k-2)!!}{(2k-3)!!} = \frac{1}{2}\cdot\frac{(2n-2)!!}{(2n-3)!!} + \frac{1}{4}\big(2H_{2n-2} - H_{n-1} - 2\big).$$

(b) For every $m \geqslant 1$,

$$\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{m}\frac{(2k-2)!!}{(2k-3)!!} = \frac{1}{2}\binom{n}{m}\Big(\frac{(2n-2)!!}{(2n-3)!!} - \frac{(2m-2)!!}{(2m-3)!!}\Big).$$

Lemma 12. *The solution X_n of the equation*

$$X_n = 2\sum_{k=1}^{n-1} C_{k,n-k}X_k + \sum_{l=1}^{r} a_l\binom{n}{l} + \frac{(2n-2)!!}{(2n-3)!!}\sum_{l=1}^{s} b_l\binom{n}{l}$$

with given initial condition X_1 is

$$X_n = \sum_{l=1}^{s+1}\widehat{a}_l\binom{n}{l} + \frac{(2n-2)!!}{(2n-3)!!}\sum_{l=1}^{r}\widehat{b}_l\binom{n}{l}$$

with

$$
\begin{aligned}
\widehat{a}_1 &= X_1 - a_1 \\
\widehat{a}_l &= \frac{l\cdot(2l-2)!!}{(2l-3)!!}\Big(\frac{b_l}{l} + \frac{b_{l-1}}{l-1}\Big), \quad l = 2,\dots,s \\
\widehat{a}_{s+1} &= \frac{(s+1)\cdot(2s)!!}{s\cdot(2s-1)!!}\cdot b_s \\
\widehat{b}_l &= \frac{(2l-3)!!}{(2l-2)!!}\cdot a_l, \quad l = 1,\dots,r
\end{aligned}
$$

Claim 3. *For every $n \geqslant 1$, the expected value of QC_n under the uniform model is*

$$E_U(QC_n) = \binom{n+1}{2}\cdot\frac{(2n-2)!!}{(2n-3)!!} - n(2n-1).$$

Proof. By Lemma 9(a),

$$
\begin{aligned}
E_U(QC_n) &= 2\sum_{k=1}^{n-1} C_{k,n-k}E_U(QC_k) + \sum_{k=1}^{n-1} C_{k,n-k}(n-2k)^2 \\
&= 2\sum_{k=1}^{n-1} C_{k,n-k}E_U(QC_k) + n^2\sum_{k=1}^{n-1} C_{k,n-k} - 4(n-1)\sum_{k=1}^{n-1} C_{k,n-k}k
\end{aligned}
$$

$$
\begin{aligned}
&+8\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{2}\\
&=2\sum_{k=1}^{n-1} C_{k,n-k}E_U(QC_k)+n^2-2n(n-1)\\
&\quad+4\binom{n}{2}\Big(1-\frac{1}{2(n-1)}\cdot\frac{(2n-2)!!}{(2n-3)!!}\Big)\\
&=2\sum_{k=1}^{n-1} C_{k,n-k}E_U(QC_k)+2\binom{n}{2}+n-n\cdot\frac{(2n-2)!!}{(2n-3)!!}
\end{aligned}
$$

where in the second last equality we have used Lemma 10. Therefore, by Lemma 12 and using that $E_U(QC_1)=0$, we have that

$$
\begin{aligned}
E_U(QC_n)&=\Big(\binom{n}{2}+n\Big)\frac{(2n-2)!!}{(2n-3)!!}-\Big(4\binom{n}{2}+n\Big)\\
&=\binom{n+1}{2}\cdot\frac{(2n-2)!!}{(2n-3)!!}-n(2n-1)
\end{aligned}
$$

as we claimed. □

Claim 4. *For every $n\geqslant 1$, the variance of QC_n under the uniform model is*

$$
\begin{aligned}
\sigma_U^2(QC_n)&=\frac{2}{15}(2n-1)(7n^2+9n-1)\binom{n+1}{2}\\
&\quad-\frac{1}{8}(5n^2+n+2)\binom{n+1}{2}\frac{(2n-2)!!}{(2n-3)!!}-\binom{n+1}{2}^2\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2
\end{aligned}
$$

Proof. To simplify the notations, we shall denote $(2n-2)!!/(2n-3)!!$ by α_n. We shall compute the variance $\sigma_U^2(QC_n)$ by means of the identity

$$\sigma_U^2(QC_n)=E_U(QC_n^2)-E_U(QC_n)^2 \tag{9}$$

where the value of $E_U(QC_n)$ is given by Claim 3. Now, we must compute $E_U(QC_n^2)$. By Lemma 9(b),

$$
\begin{aligned}
E_U(QC_n^2)&=\sum_{k=1}^{n-1} C_{k,n-k}\Big(2E_U(QC_k^2)+(n-2k)^4\\
&\qquad+4(n-2k)^2E_U(QC_k)+2E_U(QC_k)E_U(QC_{n-k})\Big)\\
&=2\sum_{k=1}^{n-1} C_{k,n-k}E_U(QC_k^2)\\
&\quad+\sum_{k=1}^{n-1} C_{k,n-k}\Big[(n-2k)^4+4(n-2k)^2\Big(\binom{k+1}{2}\alpha_k-k(2k-1)\Big)\\
&\qquad+2\Big(\binom{k+1}{2}\alpha_k-k(2k-1)\Big)\\
&\qquad\cdot\Big(\binom{n-k+1}{2}\alpha_{n-k}-(n-k)(2(n-k)-1)\Big)\Big]\\
&=2\sum_{k=1}^{n-1} C_{k,n-k}E_U(QC_k^2)\\
&\quad+\sum_{k=1}^{n-1} C_{k,n-k}\Big((n-2k)^4-4(n-2k)^2k(2k-1)\\
&\qquad+2k(2k-1)(n-k)(2(n-k)-1)\Big)\\
&\quad+\sum_{k=1}^{n-1} C_{k,n-k}\Big[4(n-2k)^2\binom{k+1}{2}\alpha_k-2\binom{n-k+1}{2}k(2k-1)\alpha_{n-k}\\
&\qquad-2\binom{k+1}{2}(n-k)(2(n-k)-1)\alpha_k\Big]\\
&\quad+2\sum_{k=1}^{n-1} C_{k,n-k}\binom{k+1}{2}\binom{n-k+1}{2}\alpha_k\alpha_{n-k}\\
&=2\sum_{k=1}^{n-1} C_{k,n-k}E_U(QC_k^2)\\
&\quad-\sum_{k=1}^{n-1} C_{k,n-k}\Big(8k^4+16(n-1)k^3-2(12n^2-6n-1)k^2\\
&\qquad-(2n-8n^3)k-n^4\Big)\\
&\quad+\sum_{k=1}^{n-1} C_{k,n-k}\Big[4(n-2k)^2\binom{k+1}{2}-4\binom{k+1}{2}(n-k)(2(n-k)-1)\Big]\alpha_k\\
&\quad+2\sum_{k=1}^{n-1} C_{k,n-k}\binom{k+1}{2}\binom{n-k+1}{2}\alpha_k\alpha_{n-k}\\
&=2\sum_{k=1}^{n-1} C_{k,n-k}E_U(QC_k^2)\\
&\quad-\sum_{k=1}^{n-1} C_{k,n-k}\Big[192\binom{k}{4}+96(n+2)\binom{k}{3}-4(12n^2-30n-5)\binom{k}{2}\\
&\qquad+(8n^3-24n^2+26n-6)k-n^4\Big]\\
&\quad+\sum_{k=1}^{n-1} C_{k,n-k}\Big[96\binom{k}{4}+156\binom{k}{3}-4(n^2-n-16)\binom{k}{2}\\
&\qquad-4(n^2-n-1)k\Big]\alpha_k\\
&\quad+2\sum_{k=1}^{n-1} C_{k,n-k}\binom{k+1}{2}\binom{n-k+1}{2}\alpha_k\alpha_{n-k}.
\end{aligned}
\tag{10}
$$

Let us compute the independent term in this equation. The first sum can be computed using Lemma 10:

$$
\begin{aligned}
&\sum_{k=1}^{n-1} C_{k,n-k}\Big(192\binom{k}{4}+96(n+2)\binom{k}{3}-4(12n^2-30n-5)\binom{k}{2}\\
&\qquad+(8n^3-24n^2+26n-6)k-n^4\Big)\\
&=96\binom{n}{4}\Big(1-\frac{3}{n-1}\cdot\frac{5!!}{6!!}\cdot\alpha_n\Big)\\
&\quad+48(n+2)\binom{n}{3}\Big(1-\frac{2}{n-1}\cdot\frac{3!!}{4!!}\cdot\alpha_n\Big)\\
&\quad-2(12n^2-30n-5)\binom{n}{2}\Big(1-\frac{1}{2(n-1)}\cdot\alpha_n\Big)\\
&\quad+\frac{1}{2}(8n^3-24n^2+26n-6)n-n^4\\
&=(3n-2)n^3-\frac{n(15n^2-15n+4)}{4}\cdot\alpha_n.
\end{aligned}
$$

The second sum in this independent term can be computed using Lemma 11:

$$
\begin{aligned}
&\sum_{k=1}^{n-1} C_{k,n-k}\Big[96\binom{k}{4}+156\binom{k}{3}-4(n^2-n-16)\binom{k}{2}-4(n^2-n-1)k\Big]\alpha_k\\
&=48\binom{n}{4}\Big(\alpha_n-\frac{6!!}{5!!}\Big)+78\binom{n}{3}\Big(\alpha_n-\frac{4!!}{3!!}\Big)\\
&\quad-2(n^2-n-16)\binom{n}{2}(\alpha_n-2)-2(n^2-n-1)n(\alpha_n-1)\\
&=n^3(n+1)\alpha_n-\frac{2n(33n^3-13n^2-12n+7)}{15}.
\end{aligned}
$$

Finally, the third sum in the independent term of this equation can be computed as follows:

$$
\begin{aligned}
&2\sum_{k=1}^{n-1} C_{k,n-k}\binom{k+1}{2}\binom{n-k+1}{2}\frac{(2k-2)!!}{(2k-3)!!}\frac{(2n-2k-2)!!}{(2n-2k-3)!!}\\
&=\sum_{k=1}^{n-1}\frac{n!(2k-3)!!(2(n-k)-3)!!k(k+1)(n-k)(n-k+1)2^{k-1}(k-1)!2^{n-k-1}(n-k-1)!}{k!(n-k)!(2n-3)!!2^2(2k-3)!!(2(n-k)-3)!!}\\
&=\frac{n!2^{n-4}}{(2n-3)!!}\sum_{k=1}^{n-1}(k+1)(n-k+1)\\
&=\frac{n!2^{n-3}(n-1)(n+1)(n+6)}{(2n-3)!!6}=\frac{n+6}{8}\cdot\binom{n+1}{3}\cdot\alpha_n.
\end{aligned}
$$

So, the independent term of Eq. (10) is

$$\frac{n(15n^2-15n+4)}{4}\cdot\alpha_n-(3n-2)n^3$$

$$+ n^3(n+1)\alpha_n - \frac{2n(33n^3 - 13n^2 - 12n + 7)}{15} + \frac{n+6}{8} \cdot \binom{n+1}{3} \cdot \alpha_n = \frac{n(49n^3 + 234n^2 - 181n + 42)}{48} \cdot \alpha_n - \frac{n(111n^3 - 56n^2 - 24n + 14)}{15} = \left(3n + 36\binom{n}{2} + 66\binom{n}{3} + \frac{49}{2}\binom{n}{4}\right)\alpha_n - 3n - 78\binom{n}{2} - 244\binom{n}{3} - \frac{888}{5}\binom{n}{4}$$

and, hence, Eq. (10) simplifies to

$$E_U(QC_n^2) = 2\sum_{k=1}^{n-1} C_{k,n-k} E_U(QC_k^2) - 3n - 78\binom{n}{2} - 244\binom{n}{3} - \frac{888}{5}\binom{n}{4} + \left(3n + 36\binom{n}{2} + 66\binom{n}{3} + \frac{49}{2}\binom{n}{4}\right)\alpha_n.$$

This equation can be solved using Lemma 12 and the fact that $E_U(QC_1^2) = 0$. Its solution is

$$E_U(QC_n^2) = 3n + 84\binom{n}{2} + 320\binom{n}{3} + 360\binom{n}{4} + 112\binom{n}{5} - \left(3n + 39\binom{n}{2} + \frac{183}{2}\binom{n}{3} + \frac{111}{2}\binom{n}{4}\right)\alpha_n = \frac{n}{15}(14n^4 + 85n^3 - 60n^2 + 5n + 1) - \frac{n}{16}(37n^3 + 22n^2 - 13n + 2)\alpha_n.$$

Finally,

$$\sigma_U^2(QC_n) = E_U(QC_n^2) - E_U(QC_n)^2 = \frac{2}{15}(2n-1)(7n^2 + 9n - 1)\binom{n+1}{2} - \frac{1}{8}(5n^2 + n + 2)\binom{n+1}{2}\frac{(2n-2)!!}{(2n-3)!!} - \binom{n+1}{2}^2\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2$$

as we claimed. □

A.3. Proof of Theorem 7

Recall from Lemma 1 that the Q–Colless index has the following recursive representation for a tree $T_n = T_k * T'_{n-k} \in \mathcal{T}_n$:

$$QC(T_n) = QC(T_k) + QC(T'_{n-k}) + (n-2k)^2.$$

Recall moreover from Theorem 5 that, under the Yule model,

$$E_Y(QC_n) = n(n+1) - 2nH_n$$

and

$$\sigma_Y^2(QC_n) = \frac{1}{3}n\left(n^3 - 8n^2 + 50n - 1 - 30H_n - 12nH_n^{(2)}\right) \sim n^4.$$

Let us now consider the following normalized version of the Q-Colless index on $\mathcal{T}_n$:

$$Y(T) = \frac{QC(T) - E_Y(QC_n)}{n^2} = \frac{QC(T) - (n(n+1) - 2nH_n)}{n^2}.$$

Then, if $T = T_k * T'_{n-k}$,

$$Y(T) = n^{-2}\left(QC(T_k) + QC(T'_{n-k}) + (n-2k)^2 - (n(n+1) - 2nH_n)\right) = n^{-2}\left[k^2\left(\frac{QC(T_k) - (k(k+1) - 2kH_k)}{k^2} + \frac{k(k+1) - 2kH_k}{k^2}\right) + (n-k)^2\left(\frac{QC(T'_{n-k}) - ((n-k)(n-k+1) - 2(n-k)H_{n-k})}{(n-k)^2} + \frac{(n-k)(n-k+1) - 2(n-k)H_{n-k}}{(n-k)^2}\right) + (n-2k)^2 - (n(n+1) - 2nH_n)\right]$$

$$= n^{-2}(k^2 Y(T_k) + (n-k)^2 Y(T'_{n-k}) + k(k+1) - 2kH_k + (n-k)(n-k+1) - 2(n-k)H_{n-k} + (n-2k)^2 - (n(n+1) - 2nH_n)$$

$$= n^{-2}(k^2 Y(T_k) + (n-k)^2 Y(T'_{n-k}) + n^2 + 6k^2 - 6nk + 2nH_n - 2kH_k - 2(n-k)H_{n-k})$$

$$= n^{-2}(k^2 Y(T_k) + (n-k)^2 Y(T'_{n-k}) + A_n(k)) \quad (11)$$

where

$$A_n(k) = n^2 + 6k^2 - 6nk + 2nH_n - 2kH_k - 2(n-k)H_{n-k}.$$

Let us denote now by Y_n the random variable that chooses a tree $T \in \mathcal{T}_n$ and computes $Y(T)$. Moreover, and as it is usual in this context (see, e.g., [47]) let L_n denote the random variable that chooses a tree $T \in \mathcal{T}_n$ and one of its maximal pending subtrees of T and counts its number of leaves. Recall that, under the Yule model, L_n is uniformly distributed on $\{1, \ldots, n-1\}$. When we translate Eq. (11) in terms of these random variables, we obtain

$$Y_n = n^{-2}\left(L_n^2 Y_{L_n} + (n - L_n)^2 Y_{n-L_n} + A_n(L_n)\right).$$

We will now look at the limit in distribution of Y_n under the Yule model as $n \to \infty$. If one simplifies the proof of Thm. 3.1 in [60] to the Yule case (which corresponds to Aldous's β-model with $\beta = 0$ [59]), or, alternatively, follows the logic behind the proof Thm. 3.1 in [61] (which is a minor modification of the proof of Thm. 2.1 in [65]), using the function A_n that gives the independent term in Eq. (11), and noticing that $E_Y(Y_n^2)$ is uniformly bounded and that

$$E_Y(A_n(L_n)) = E_Y(n^2 + 6L_n(L_n - n) + 2nH_n - 2L_nH_{L_n} - 2(n - L_n)H_{n-L_n}) = n^2 + \frac{6}{n-1}\sum_{k=1}^{n-1} k(k-n) + 2nH_n - \frac{2}{n-1}\sum_{k=1}^{n-1} kH_k - \frac{2}{n-1}\sum_{k=1}^{n-1} kH_k = n^2 - n(n+1) + 2nH_n - n(2H_n - 1) = 0$$

(for the value of $\sum_{k=1}^{n-1} kH_k$, see Eqn. (6.68) in [58]), one obtains that $Y_n \xrightarrow{D} Y$, for a random variable Y satisfying the equality in distribution

$$Y \overset{D}{=} \tau^2 Y' + (1-\tau)^2 Y'' + (1 + 6\tau^2 - 6\tau), \quad (12)$$

where $\tau \sim \mathrm{Unif}[0,1]$ and Y', Y'' are independent, distributed according to the same law as Y.

References

[1] D.J. Futuyma (Ed.), Evolution, Science and Society: Evolutionary Biology and the National Research Agenda, The State University of New Jersey, 1999.
[2] T. Kubo, Y. Iwasa, Inferring the rates of branching and extinction from molecular phylogenies, Evolution 49 (1995) 694–704.
[3] A.O. Mooers, S.B. Heard, Inferring evolutionary process from phylogenetic tree shape, Q. Rev. Biol. 72 (1997) 31–54.
[4] M. Stich, S. Manrubia, Topological properties of phylogenetic trees in evolutionary models, Eur. Phys. J. B 70 (2009) 583–592.
[5] J. Felsenstein, Inferring Phylogenies, Sinauer Associates, 2004.
[6] A. Drummond, S.Y.W. Ho, et al., Relaxed phylogenetics and dating with confidence, PLoS Biol. 4 (2006) e88.
[7] A. Brower, E. Rindal, Reality check: A reply to smith, Cladistics 29 (2013) 464–465.
[8] D. Hillis, J. Bull, M. White, et al., Experimental phylogenetics: Generation of a known phylogeny, Science 255 (1992) 589–592.
[9] E. Rindal, A. Brower, Do model-based phylogenetic analyses perform better than parsimony? A test with empirical data, Cladistics 27 (2011) 331–334.
[10] G. Fusco, Q.C. Cronk, A new method for evaluating the shape of large phylogenies, J. Theoret. Biol. 175 (1995) 235–243.
[11] K. Shao, R. Sokal, Tree balance, Syst. Zool. 39 (1990) 266–276.
[12] A. McKenzie, M. Steel, Distributions of cherries for two models of trees, Math. Biosci. 164 (2000) 81–92.
[13] H.M. Savage, The shape of evolution: Systematic tree topology, Biol. J. Linnean Soc. 20 (1983) 225–244.
[14] J. Slowinski, Probabilities of n-trees under two models: A demonstration that asymmetrical interior nodes are not improbable, Syst. Zool. 39 (1990) 89–94.

[15] T. Wu, K. Choi, On joint subtree distributions under two evolutionary models, Theor. Popul. Biol. 108 (2015) 13–23.
[16] G.U. Yule, A mathematical theory of evolution based on the conclusions of Dr J.C. Willis, Philos. Trans. R. Soc. London B 213 (1924) 21–87.
[17] M.I. Nelson, E.C. Holmes, The evolution of epidemic influenza, Nature Rev. Genet. 8 (2007) 196–205.
[18] D. Colless, Review of Phylogenetics: the theory and practice of phylogenetic systematics, Syst. Zool. 31 (1982) 100–104.
[19] T.M. Coronado, A. Mir, F. Rosselló, G. Valiente, A balance index for phylogenetic trees based on rooted quartets, J. Math. Biol. 79 (2019) 1105–1148.
[20] M. Fischer, V. Liebscher, On the balance of unrooted trees, 2015, arXiv preprint arXiv:1510.07882.
[21] M. Kirkpatrick, M. Slatkin, Searching for evolutionary patterns in the shape of a phylogenetic tree, Evolution 47 (1993) 1171–1181.
[22] A. Mir, F. Rosselló, L. Rotger, A new balance index for phylogenetic trees, Math. Biosci. 241 (2013) 125–136.
[23] A. Mir, F. Rosselló, L. Rotger, Sound Colless-like balance indices for multifurcating trees, PLoS ONE 13 (2018) e0203401.
[24] M. Sackin, Good and bad phenograms, Syst. Zool. 21 (1972) 225–226.
[25] D. Aldous, Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today, Statist. Sci. 16 (2001) 23–34.
[26] M. Blum, O. François, On statistical tests of phylogenetic tree imbalance: The Sackin and other indices revisited, Math. Biosci. 195 (2005) 141–153.
[27] S. Duchene, R. Bouckaert, et al., Phylodynamic model adequacy using posterior predictive simulations, Syst. Biol. 68 (2018) 358–364.
[28] A. Purvis, Using interspecies phylogenies to test macroevolutionary hypotheses, in: New Uses for New Phylogenies, Oxford University Press, 1996, pp. 153–168.
[29] G. Verboom, F. Boucher, D. Ackerly, et al., Species selection regime and phylogenetic tree shape, Syst. Biol. 69 (2020) 774–794.
[30] D. Colless, Relative symmetry of cladograms and phenograms: An experimental study, Syst. Biol. 44 (1995) 102–108.
[31] J. Farris, M. Källersjö, Asymmetry and explanations, Cladistics 14 (1998) 159–166.
[32] T. Holton, M. Wilkinson, D. Pisani, The shape of modern tree reconstruction methods, Syst. Biol. 63 (2014) 436–441.
[33] E. Sober, Experimental tests of phylogenetic inference methods, Syst. Biol. 42 (1993) 85–89.
[34] E. Stam, Does imbalance in phylogenies reflect only bias?, Evolution 56 (2002) 1292–1295.
[35] M. Avino, T.N. Garway, et al., Tree shape-based approaches for the comparative study of cophylogeny, Ecol. Evol. 9 (2019) 6756–6771.
[36] P. Goloboff, J. Arias, C. Szumik, Comparing tree shapes: beyond symmetry, Zool. Scripta 46 (2017) 637–648.
[37] H. Kayondo, S. Mwalili, J. Mango, Inferring multi-type birth-death parameters for a structured host population with application to HIV epidemic in Africa, Comput. Mol. Biosci. 9 (2019) 108–131.
[38] A.F. Poon, Phylodynamic inference with kernel ABC and its application to HIV epidemiology, Mol. Biol. Evol. 32 (2015) 2483–2495.
[39] E. Saulnier, S. Alizon, O. Gascuel, Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study, PLoS Comput. Biol. 13 (2017) e1005416.
[40] L. Chalmandrier, C. Albouy, et al., Comparing spatial diversification and meta-population models in the Indo-Australian Archipelago, Royal Soc. Open Sci. 5 (2018) 171366.
[41] T. Cunha, G. G. Giribet, A congruent topology for deep gastropod relationships, Proc. R. Soc. B 286 (2019) 20182776.
[42] C. Metzig, O. Ratmann, D. Bezemer, C. Colijn, Phylogenies from dynamic networks, PLoS Comput. Biol. 15 (2019) e1006761.
[43] A. Purvis, S. Fritz, J. Rodríguez, et al., The shape of mammalian phylogeny: Patterns, processes and scales, Phil. Trans. R. Soc. B 366 (2011) 2462–2477.
[44] P. Agapow, A. Purvis, Power of eight tree shape statistics to detect nonrandom diversification: A comparison by simulation of two models of cladogenesis, Syst. Biol. 51 (2002) 866–872.
[45] F. Matsen, A geometric approach to tree shape statistics, Syst. Biol. 55 (2006) 652–661.
[46] M. Hayati, B. Shadgar, L. Chindelevitch, A new resolution function to evaluate tree shape statistics, PLoS One 14 (2019) e0224197.
[47] M. Blum, O. François, S. Janson, The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance, Ann. Appl. Probab. 16 (2006) 2195–2214.
[48] G. Cardona, A. Mir, F. Rosselló, Exact formulas for the variance of several balance indices under the Yule model, J. Math. Biol. 67 (2013) 1833–1846.
[49] D. Ford, Probabilities on Cladograms: Introduction to the Alpha Model (Ph.D. thesis), Stanford University, 2005, arXiv preprint arXiv:math/0511246.
[50] S.B. Heard, Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees, Evolution 46 (1992) 1818–1826.
[51] T.M. Coronado, M. Fischer, L. Herbst, F. Rosselló, K. Wicke, On the minimum value of the Colless index and the bifurcating trees that achieve it, J. Math. Biol. 80 (2020) 1993–2054.
[52] E. Harding, The probabilities of rooted tree-shapes generated by random bifurcation, Adv. Appl. Probab. 3 (1971) 44–77.
[53] L.L. Cavalli-Sforza, A. Edwards, Phylogenetic analysis: Models and estimation procedures, Evolution 21 (1967) 550–570.
[54] D.E. Rosen, Vicariant patterns and historical explanation in biogeography, Syst. Biol. 27 (1978) 159–188.
[55] M. Steel, Phylogeny: Discrete and random processes in evolution, SIAM (2016).
[56] J.S. Rogers, Response of Colless's tree imbalance to number of terminal taxa, Syst. Biol. 42 (1993) 102–105.
[57] T.M. Coronado, A. Mir, F. Rosselló, L. Rotger, On Sackin's original proposal: The variance of the leaves' depths as a phylogenetic balance index, BMC Bioinformatics 21 (2020) 154.
[58] R. Graham, D. Knuth, O. Patashnik, Concrete Mathematics, second ed., Addison-Wesley, 1994.
[59] D. Aldous, Probability distributions on cladograms, in: D. Aldous, R. Pemantle (Eds.), Random Discrete Structures, Springer–Verlag, 1996, pp. 1–18.
[60] K. Bartoszek, Limit distribution of the quartet balance index for Aldous's $\beta \geqslant 0$–model, Appl. Math. 47 (2020) 29–44.
[61] K. Bartoszek, Exact and approximate limit behaviour of the Yule tree's cophenetic index, Math. Biosci. 303 (2018) 26–45.
[62] M. Fischer, Extremal values of the Sackin balance index for rooted binary trees, 2018, arXiv preprint arXiv:1801.10418.
[63] C. Wei, D. Gong, Q. Wang, Chu-Vandermonde convolution and harmonic number identities Chu–Vandermonde convolution and harmonic number identities, Integral Transforms Spec. Funct. 24 (2013) 324–330.
[64] D. Knuth, The Art of Computer Programming, third ed., in: Vol. 1: Fundamental Algorithms, Addison-Wesley, 1997.
[65] U. Rösler, A limit theorem for quicksort, Inform. Theor. Appl. 25 (1991) 85–100.