PLOS ONE

RESEARCH ARTICLE

Sound Colless-like balance indices for multifurcating trees

Arnau Mir[1,3☯], **Lucía Rotger**[2☯], **Francesc Rosselló**[1,3☯*]

1 Dept. of Mathematics and Computer Science, University of the Balearic Islands, E-07122 Palma, Spain, **2** Dept. of Mathematics and Computing, University of La Rioja, E-26004 Logroño, Spain, **3** Balearic Islands Health Research Institute (IdISBa), E-07010 Palma, Spain

☯ These authors contributed equally to this work.
* cesc.rossello@uib.es

Citation: Mir A, Rotger L, Rosselló F (2018) Sound Colless-like balance indices for multifurcating trees. PLoS ONE 13(9): e0203401. https://doi.org/10.1371/journal.pone.0203401

Editor: Ulrich Melcher, Oklahoma State University, UNITED STATES

Received: May 3, 2018

Accepted: August 20, 2018

Published: September 25, 2018

Data Availability Statement: All data used in this paper are available at the companion GitHub page https://github.com/LuciaRotger/CollessLike. The balance indices computed in this paper are also available as a dataset in the R package "CollesLike." The phylogenetic trees which have been used in the numerical experiments (and which amount to 4 Gb) are available from the authors on request.

Funding: This research has been partially supported by the Obra Social la Caixa through the "Programa Pont La Caixa per a grups de recerca de la UIB" and by the Spanish Ministry of Economy and Competitiveness and European Re- gional

Abstract

The Colless index is one of the most popular and natural balance indices for bifurcating phylogenetic trees, but it makes no sense for multifurcating trees. In this paper we propose a family of *Colless-like* balance indices $\mathfrak{C}_{D,f}$ that generalize the Colless index to multifurcating phylogenetic trees. Each $\mathfrak{C}_{D,f}$ is determined by the choice of a dissimilarity D and a weight function $f : \mathbb{N} \to \mathbb{R}_{\geq 0}$. A balance index is *sound* when the most balanced phylogenetic trees according to it are exactly the fully symmetric ones. Unfortunately, not every Colless-like balance index is sound in this sense. We prove then that taking $f(n) = \ln(n + e)$ or $f(n) = e^n$ as weight functions, the resulting index $\mathfrak{C}_{D,f}$ is sound for every dissimilarity D. Next, for each one of these two functions f and for three popular dissimilarities D (the variance, the standard deviation, and the mean deviation from the median), we find the most unbalanced phylogenetic trees according to $\mathfrak{C}_{D,f}$ with any given number n of leaves. The results show that the growth pace of the function f influences the notion of "balance" measured by the indices it defines. Finally, we introduce our R package "CollessLike," which, among other functionalities, allows the computation of Colless-like indices of trees and their comparison to their distribution under Chen-Ford-Winkel's α-γ-model for multifurcating phylogenetic trees. As an application, we show that the trees in TreeBASE do not seem to follow either the uniform model for multifurcating trees or the α-γ-model, for any values of α and γ.

Introduction

Since the early 1970s, the shapes of phylogenetic trees have been used to test hypothesis about the evolutive forces underlying their assembly [1]. The most used topological feature of phylogenetic trees in this regard is their symmetry, which captures the symmetry of the evolutionary histories described by them. The symmetry of a tree is usually measured through its *balance* (see [2], pp. 559–560), the tendency of the children of any given node to have the same number of descendant leaves. Several *balance indices* have been proposed so far to quantify the balance of a phylogenetic tree. The two most popular ones are the *Colless index* [3], whose definition we recall below and that only works for bifurcating trees, and the *Sackin index* [4–6], which is

Development Fund through project DPI2015-67082-P (MINECO/FEDER) (FR). There was no additional external funding received for this study.

Competing interests: The authors have declared that no competing interests exist.

defined as the sum of the depths of the leaves in the tree and can be used on multifurcating trees. Other balance indices for bifurcating trees introduced so far include the variance of the depths of the leaves [4, 5], the sum of the reciprocals of the orders of the rooted subtrees [6], and the number of cherries [7]. As for balance indices for multifurcating trees, two recent additions are the total cophenetic index [8] and the quartet index [9]; for more proposals, see the section "Measures of overall asymmetry" in Felsenstein's book [2] (pp. 562–563). This abundance of balance indices is partly motivated by Shao and Sokal's advice on using more than one such index to quantify tree balance: see [6], p. 1990.

The *Colless index* $C(T)$ of a bifurcating phylogenetic tree T is defined as the sum of the balance values of its internal nodes, where by the *balance value* of an internal node we mean the absolute value of the difference between the number of descendant leaves of its pair of children. In this way, the Colless index of a bifurcating tree measures the average balance value of its internal nodes, and therefore it quantifies in a very intuitive way its balance. In particular, $C(T) = 0$ if, and only if, T is a fully symmetric bifurcating tree with 2^m leaves, for some m.

Unfortunately, the Colless index can only be used as defined on bifurcating trees. A natural generalization to multifurcating trees would be to define the balance value of a node as some measure of the dissimilarity of the numbers of descendant leaves of its children, like for instance their standard deviation, and then to add up all these balance values. But this definition has a drawback: this sum can be 0 on non-symmetric multifurcating trees, and hence the resulting index need not capture the symmetry of a tree in a sound way. For an example of this misbehavior, consider the tree depicted in Fig 1: each one of its nodes has all its children with the same number of descendant leaves and therefore the balance value of each node is 0 independently on the dissimilarity used to define it, but the tree is not symmetric. Replacing the number of descendant leaves by the number of descendant nodes, which in a bifurcating tree is simply twice the number of descendant leaves minus 1, does not overcome this drawback: again, all children of each node in the tree depicted in Fig 1 have the same number of descendant nodes.

In this paper we solve this problem by taking a suitable function $f : \mathbb{N} \to \mathbb{R}_{\geq 0}$ and then replacing in this schema the number of descendant leaves or the number of descendant nodes of a node by the *f-size* of the subtree rooted at the node, defined as the sum of the images under f of the out-degrees of the nodes in the subtree. Then, we define the *balance value* (relative to such a function f and a dissimilarity D) of an internal node in a phylogenetic tree as the value of D applied to the f-sizes of the subtrees rooted at the children of the node. Finally, we define the *Colless-like index* $\mathfrak{C}_{D,f}$ of a phylogenetic tree as the sum of the balance values relative to f and D of its internal nodes.

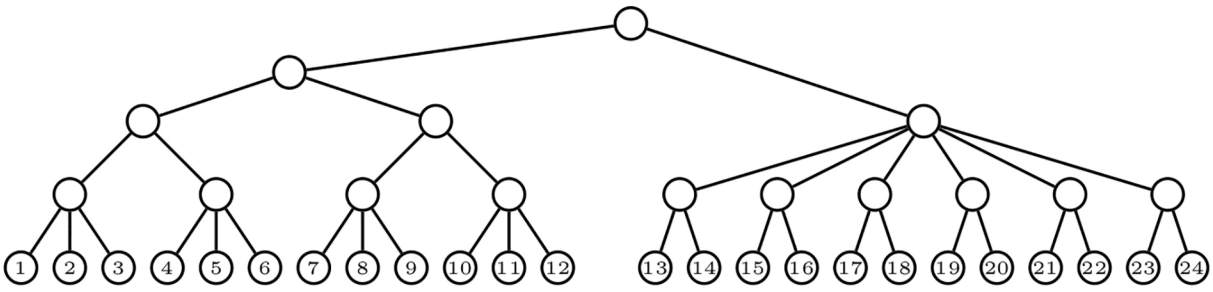

Fig 1. Each node in this asymmetric tree has all its children with the same number of descendant leaves as well as with the same number of descendant nodes.

https://doi.org/10.1371/journal.pone.0203401.g001

The advantage of such a general definition is that there exist functions f such that, for every dissimilarity D, the resulting index $\mathfrak{C}_{D,f}$ satisfies that $\mathfrak{C}_{D,f}(T) = 0$ if, and only if, T is *fully symmetric*, in the sense that, for every internal node v, the subtrees rooted at the children of v have all the same shape. Two such functions turn out to be $f(n) = \ln(n + e)$ and $f(n) = e^n$.

The different growth pace of these two functions make them quantify the trees' balance in different ways. We show it by finding the trees that are maximally unbalanced according to $\mathfrak{C}_{D,f}$, that is, the trees with largest $\mathfrak{C}_{D,f}$ value, when f is one of these two functions and D is the variance, the standard deviation, and the mean deviation from the median. We show that the choice of the dissimilarity D does not cause any major difference in the maximally unbalanced trees relative to $\mathfrak{C}_{D,f}$ for a fixed f, but that changing the function f implies completely different maximally unbalanced trees.

We have written an R package called *CollessLike*, available at the CRAN, that, among other functionalities, computes Colless-like indices and simulates their distribution under the α-γ-model for multifurcating trees [10]. We have used the functions in this package to perform two experiments on the TreeBASE phylogenetic database [11]. First, we have compared the behavior of the Colless-like index obtained by taking $f(n) = \ln(n + e)$ and as dissimilarity D the mean deviation from the median, MDM, with two other balance indices for multifurcating trees: the Sackin index and the total cophenetic index. Next, we have used this Colless-like index to contrast the goodness of fit of the trees in TreeBASE to the uniform distribution for multifurcating trees and to the α-γ-model.

Materials

Notations and conventions

Throughout this paper, by a *tree* we always mean a rooted, finite tree without out-degree 1 nodes. As usual, we consider such a tree to be a directed graph, with its arcs pointing away from the root. Given a tree T, we shall denote its sets of nodes, of *internal* (that is, non-leaf) nodes, and of arcs by $V(T)$, $V_{int}(T)$, and $E(T)$, respectively, and the out-degree of a node $v \in V(T)$ by $\deg(v)$. A tree T is *bifurcating* when $\deg(v) = 2$ for every $v \in V_{int}(T)$. Whenever we want to emphasize the fact that a tree need not be bifurcating, we shall call it *multifurcating*. The *depth* of a node in a tree T is the length (i.e., the number of arcs) of the directed path from the root to it, and the *depth* of T is the largest depth of any of its leaves. We shall always make the abuse of language of saying that two isomorphic trees are equal, and hence we shall always identify any tree with its isomorphism class. We shall denote by $\mathcal{T}^*_n$ the set of (isomorphism classes of) trees with n leaves, and by $\mathcal{T}^*$ the union $\cup_{n\geq 1}\mathcal{T}^*_n$.

A *phylogenetic tree* on a (non-empty, finite) set X of *labels* is a tree with its leaves bijectively labelled in the set X. We shall always identify every leaf in a phylogenetic tree T on X with its label, and in particular we shall denote its set of leaves by X. Two phylogenetic trees T_1, T_2 on X are *isomorphic* when there exists an isomorphism of directed graphs between them that preserves the labelling of the leaves. We shall also make always the abuse of language of considering two isomorphic phylogenetic trees as equal. Given a set of labels X, we shall denote by $\mathcal{T}_X$ the set of (isomorphism classes of) phylogenetic trees on X, and we shall denote by $\mathcal{T}_n$, for every $n \geq 1$, the set $\mathcal{T}_{\{1,2,\ldots,n\}}$. Notice that if $|X| = n$, then any bijection $X \leftrightarrow \{1, 2, \ldots, n\}$ induces a bijection $\mathcal{T}_X \leftrightarrow \mathcal{T}_n$. Moreover, if $|X| = n$, there is a forgetful mapping $\pi_X : \mathcal{T}_X \to \mathcal{T}^*_n$ that sends every phylogenetic tree to the corresponding unlabeled tree, which we shall call its *shape*.

No closed formula is known for the numbers $|\mathcal{T}^*_n|$ or $|\mathcal{T}_n|$. Felsenstein gives in Chapter 3 in [2] an easy recurrence to compute $|\mathcal{T}_n|$ and describes how to obtain such a recurrence for

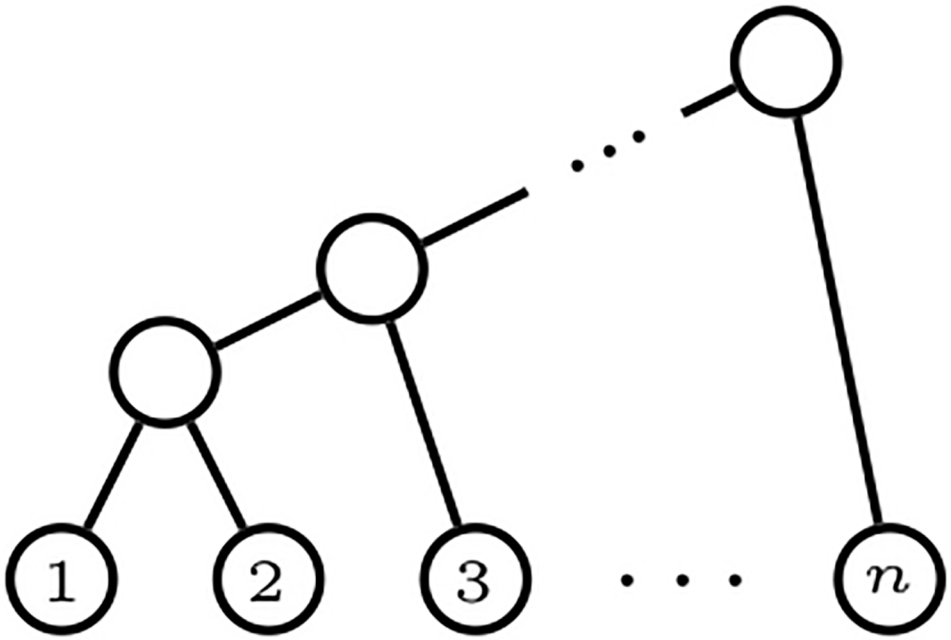

Fig 2. A comb K_n with n leaves.

https://doi.org/10.1371/journal.pone.0203401.g002

$|\mathcal{T}_n^*|$; an explicit algorithm to compute the latter is provided in [12]. These numbers $(|\mathcal{T}_n|)_n$ and $(|\mathcal{T}_n^*|)_n$ form sequences A000311 and A000669, respectively, in Sloane's *On-Line Encyclopedia of Integer Sequences* [13], where more information about them can be found.

A *comb* is a bifurcating phylogenetic tree with all its internal nodes having a leaf child: see Fig 2. We shall generically denote every comb in $\mathcal{T}_n$, as well as their shape in $\mathcal{T}_n^*$, by K_n. A *star* is a phylogenetic tree of depth 1: see Fig 3. For consistency with later notations, we shall denote the star in $\mathcal{T}_n$, and its shape in $\mathcal{T}_n^*$, by FS_n.

Let $T_1, \ldots, T_k$ be phylogenetic trees on pairwise disjoint sets of labels $X_1, \ldots, X_k$, respectively. The phylogenetic tree $T_1 \star \cdots \star T_k$ on $X_1 \cup \cdots \cup X_k$ is obtained by adding to the disjoint union of $T_1, \ldots, T_k$ a new node r and new arcs from r to the root of each T_i. In this way, the

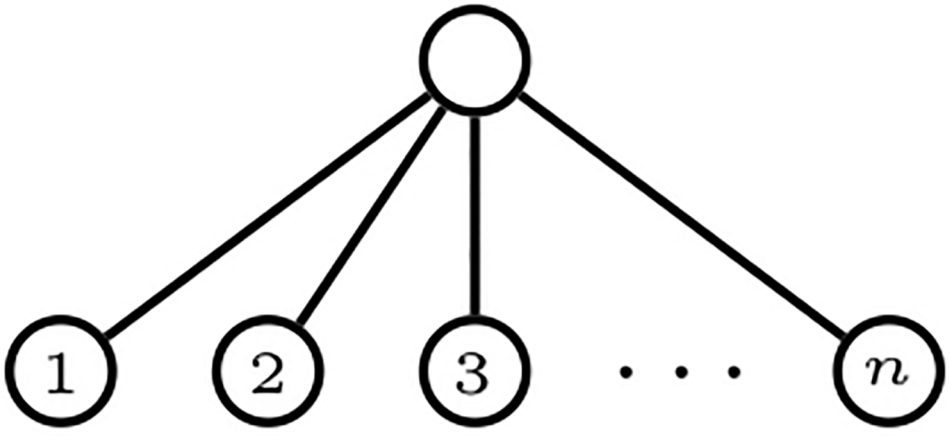

Fig 3. A star FS_n with n leaves.

https://doi.org/10.1371/journal.pone.0203401.g003

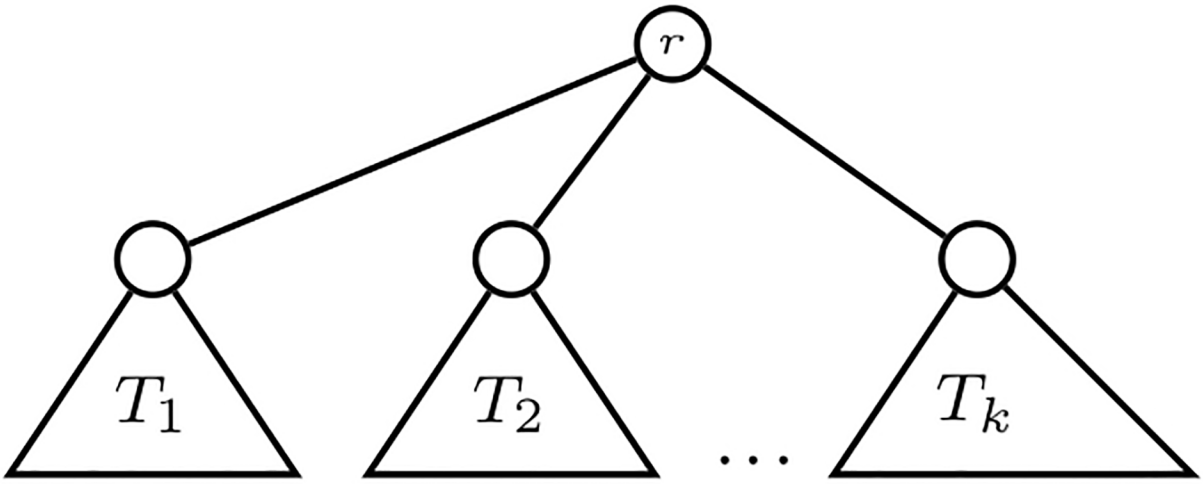

Fig 4. The (phylogenetic) tree $T_1 \star \cdots \star T_k$.

https://doi.org/10.1371/journal.pone.0203401.g004

trees $T_1, \ldots, T_k$ become the subtrees of $T_1 \star \cdots \star T_k$ rooted at the children of its root r; cf. Fig 4. A similar construction produces a tree $T_1 \star \cdots \star T_k$ from a set of (unlabeled) trees $T_1, \ldots, T_k$.

Given a node v in a tree T, we shall denote by T_v the subtree of T rooted at v and by κ_v its number of descendant leaves, that is, the number of leaves of T_v. An internal node v of a tree T is *symmetric* when, if $v_1, \ldots, v_k$ are its children, the trees $T_{v_1}, \ldots, T_{v_k}$ are isomorphic. A tree T is *fully symmetric* when all its internal nodes are symmetric, and a phylogenetic tree is *fully symmetric* when its shape is so.

Given a number n of leaves, there may exist several fully symmetric trees with n leaves. For instance, there are three fully symmetric trees with 6 leaves, depicted in Fig 5. In fact, every fully symmetric tree with n leaves is characterized by an ordered factorization $n_1 \ldots n_k$ of n, with $n_1, \ldots, n_k \geq 2$. More specifically, for every $k \geq 1$ and $(n_1, \ldots, n_k) \in \mathbb{N}^k$ with $n_1, \ldots, n_k \geq 2$, let $FS_{n_1,\ldots,n_k}$ be the tree defined, up to isomorphism, recursively as follows:

- FS_{n_1} is the star with n_1 leaves.
- If $k \geq 2$, $FS_{n_1,\ldots,n_k}$ is a tree whose root has n_1 children, and the subtrees at each one of these children are (isomorphic to) $FS_{n_2,\ldots,n_k}$.

Every $FS_{n_1,\ldots,n_k}$ is fully symmetric, and every fully symmetric tree is isomorphic to some $FS_{n_1,\ldots,n_k}$. Therefore, for every n, the number of fully symmetric trees with n leaves is equal to the number $H(n)$ of ordered factorizations of n (sequence A074206 in Sloane's *On-Line Encyclopedia of Integer Sequences* [13]).

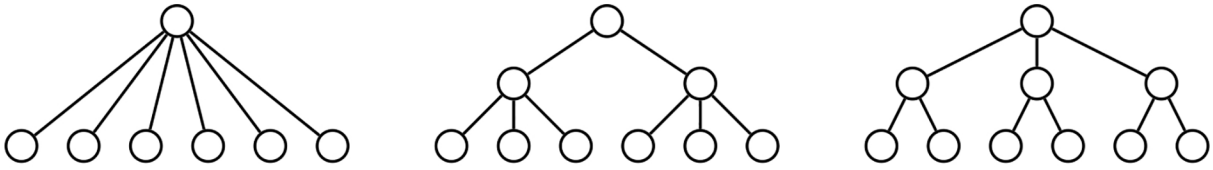

Fig 5. Three fully symmetric trees with 6 leaves: From left to right, FS_6, $FS_{2,3}$ and $FS_{3,2}$.

https://doi.org/10.1371/journal.pone.0203401.g005

The Colless index

The *Colless index* $C(T)$ of a bifurcating tree T with n leaves is defined as follows [3]: if, for every $v \in V_{int}(T)$, we denote by v_1 and v_2 its two children and by κ_{v_1} and κ_{v_2} their respective numbers of descendant leaves, then

$$C(T) = \sum_{v \in V_{int}(T)} |\kappa_{v_1} - \kappa_{v_2}|.$$

The Colless index of a phylogenetic tree is simply defined as the Colless index of its shape.

It is well-known that the maximum Colless index on the set of bifurcating trees with n leaves is reached at the comb K_n, and it is

$$C(K_n) = \binom{n-1}{2}$$

(see, for instance, [14]). In fact, this maximum is only reached at the comb. Since we have not been able to find an explicit reference for this last result in the literature and we shall make use of it later, we provide a proof here.

Lemma 1. *For every bifurcating tree T with n leaves, if $T \neq K_n$, then $C(T) < C(K_n)$.*

Proof. Let T a bifurcating tree with n leaves different from the comb K_n. Let x be an internal node of smallest depth in it without any leaf child, and let $T_1 \star T_2$ and $T_3 \star T_4$ be the subtrees rooted at its children (see Fig 6); for every $i = 1, 2, 3, 4$, let t_i be the number of leaves of T_i. Assume, without any loss of generality, that $t_1 \leq t_2$ and $t_1 + t_2 \leq t_3 + t_4$. Let then T' be the tree obtained by pruning T_2 from T and regrafting it to the other arc starting in x (see again Fig 6).

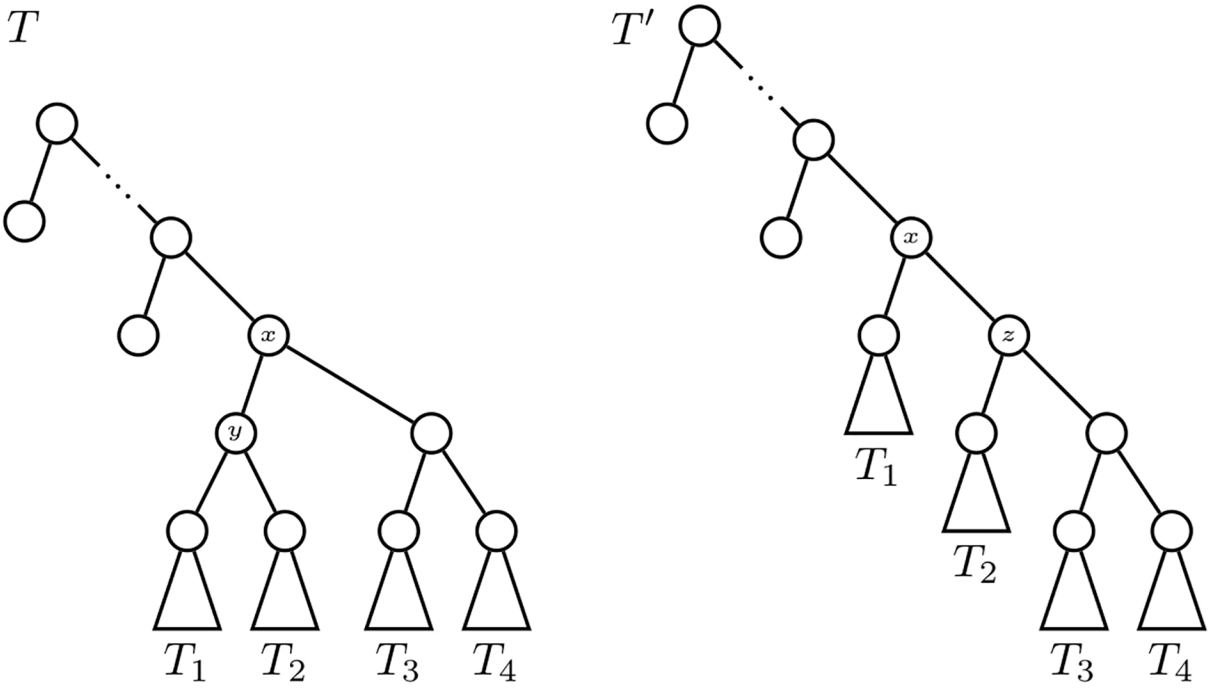

Fig 6. The trees T and T' in the proof of Lemma 1.

https://doi.org/10.1371/journal.pone.0203401.g006

Then $C(T') > C(T)$. Indeed, the only nodes whose children change their numbers of descendant leaves from T to T' are (cf. Fig 6): the node x; the parent y of the roots of T_1 and T_2 in T, which is removed in T'; and the parent z of the root of T_2 in T', which does not exist in T. Therefore,

$$\begin{aligned} & C(T') - C(T) \\ & \quad = |t_3 + t_4 - t_2| + [t_3 + t_4 + t_2 - t_1| - |t_2 - t_1| - |t_3 + t_4 - t_2 - t_1| \\ & \quad = t_3 + t_4 - t_2 + t_3 + t_4 + t_2 - t_1 - t_2 + t_1 - t_3 - t_4 + t_2 + t_1 \\ & \quad = t_1 + t_3 + t_4 > 0. \end{aligned}$$

So, this procedure takes a bifurcating tree with n leaves $T \neq K_n$ and produces a new bifurcating tree T' with the same number n of leaves and strictly larger Colless index. Since the number of bifurcating trees with n leaves is finite, the Colless index cannot increase indefinitely, which means that if we iterate this procedure, we must eventually stop at a comb K_n. And since the Colless index strictly increases at each iteration, we conclude that if $T \neq K_n$, then $C(T) < C(K_n)$.

Methods

Colless-like indices

Let $f : \mathbb{N} \to \mathbb{R}_{\geq 0}$ be a function that sends each natural number to a positive real number. The *f-size* of a tree $T \in \mathcal{T}^*$ is defined as

$$\delta_f(T) = \sum_{v \in V(T)} f(\deg(v)).$$

If $T \in \mathcal{T}_X$, for some set of labels X, then $\delta_f(T)$ is defined as $\delta_f(\pi_X(T))$.

Therefore, $\delta_f(T)$ is the sum of the degrees of all nodes in T, with these degrees weighted by means of the function f. Examples of f-sizes include:

- The *number of leaves*, κ, which is obtained by taking $f(0) = 1$ and $f(n) = 0$ if $n > 0$.
- The *order* (the number of nodes), τ, which corresponds to $f(n) = 1$ for every $n \in \mathbb{N}$.
- The usual *size* (the number of arcs), θ, which corresponds to $f(n) = n$ for every $n \in \mathbb{N}$.

Notice that δ_f satisfies the following recursion:

$$\delta_f(T_1 \star \cdots \star T_k) = \delta_f(T_1) + \cdots + \delta_f(T_k) + f(k).$$

Table A in S2 File gives the abstract values of $\delta_f(T)$ for every $T \in \mathcal{T}_n^*$ with $n = 2, 3, 4, 5$.

Example 2. *If T is a bifurcating tree with n leaves, and hence with n − 1 internal nodes, all of them of out-degree 2, then*

$$\delta_f(T) = (f(0) + f(2))n - f(2).$$

Example 3. *For every fully symmetric tree* $FS_{n_1,\ldots,n_k}$,

$$\delta_f(FS_{n_1,\ldots,n_k}) = n_1 \cdots n_k \cdot f(0) + n_1 \cdots n_{k-1} \cdot f(n_k) + \cdots + n_1 \cdot f(n_2) + f(n_1).$$

Now let

$$\mathbb{R}^+ = \bigcup_{k\geq 1} \mathbb{R}^k = \{(x_1, \ldots, x_k) \mid k \geq 1, x_1, \ldots, x_k \in \mathbb{R}\}$$

be the set of all non-empty finite-length sequences of real numbers. A *dissimilarity* on $\mathbb{R}^+$ is any mapping $D : \mathbb{R}^+ \to \mathbb{R}_{\geq 0}$ satisfying the following conditions: for every $(x_1, \ldots, x_k) \in \mathbb{R}^+$,

- $D(x_1, \ldots, x_k) = D(x_{\sigma(1)}, \ldots, x_{\sigma(k)})$, for every permutation σ of $\{1, \ldots, k\}$;
- $D(x_1, \ldots, x_k) = 0$ if, and only if, $x_1 = \cdots = x_k$.

The dissimilarities that we shall explicitly use in this paper are the *mean deviation from the median*,

$$\mathrm{MDM}(x_1, \ldots, x_k) = \frac{1}{k}\sum_{i=1}^{k} |x_i - Median(x_1, \ldots, x_k)|,$$

the (*sample*) *variance*,

$$\mathrm{var}(x_1, \ldots, x_k) = \frac{1}{k-1}\sum_{i=1}^{k} \big(x_i - Mean(x_1, \ldots, x_k)\big)^2,$$

and the (*sample*) *standard deviation*,

$$sd(x_1, \ldots, x_k) = +\sqrt{\mathrm{var}(x_1, \ldots, x_k)}.$$

Let D be a dissimilarity on $\mathbb{R}^+$, $f : \mathbb{N} \to \mathbb{R}_{\geq 0}$ a function, and δ_f the corresponding f-size, and let $T \in \mathcal{T}^*$. For every internal node v in T, with children $v_1, \ldots, v_k$, the (D, f)-*balance value* of v is

$$bal_{D,f}(v) = D(\delta_f(T_{v_1}), \ldots, \delta_f(T_{v_k})).$$

So, $bal_{D,f}(v)$ measures, through D, the spread of the f-sizes of the subtrees rooted at the children of v. In particular, $bal_{D,f}(v) = 0$ if, and only if, $\delta_f(T_{v_1}) = \cdots = \delta_f(T_{v_k})$.

Definition 4. *Let D be a dissimilarity on $\mathbb{R}^+$ and $f : \mathbb{N} \to \mathbb{R}_{\geq 0}$ a function. For every $T \in \mathcal{T}^*$, its* Colless-like index relative to D and f, $\mathfrak{C}_{D,f}(T)$, *is the sum of the (D, f)-balance values of the internal nodes of T*:

$$\mathfrak{C}_{D,f}(T) = \sum_{v \in V_{int}(T)} bal_{D,f}(v).$$

If $T \in \mathcal{T}_X$, for some set of labels X, then $\mathfrak{C}_{D,f}(T)$ is defined as $\mathfrak{C}_{D,f}(\pi_X(T))$.

Example 5. *If we take $D = \mathrm{MDM}$ and f the constant mapping 1, so that $\delta_f = \tau$, the usual order of a tree, then*

$$\begin{aligned}\mathfrak{C}_{\mathrm{MDM},\tau}(T) &= \sum_{v \in V_{int}(T)} \mathrm{MDM}(\tau_{v_1}, \ldots, \tau_{v_{\deg(v)}}) \\ &= \sum_{v \in V_{int}(T)} \frac{1}{\deg(v)} \sum_{i=1}^{\deg(v)} |\tau_{v_1} - Median(\tau_{v_1}, \ldots, \tau_{v_{\deg(v)}})|,\end{aligned}$$

where, for every $v \in V_{int}(T)$, $v_1, \ldots, v_{\deg(v)}$ denote its children and $\tau_{v_1}, \ldots, \tau_{v_{\deg(v)}}$ their numbers of descendant nodes.

Notice that $\mathfrak{C}_{D,f}$ gets larger as the f-sizes of the subtrees rooted at siblings get more different, and therefore it behaves as a balance index for trees, in the same way as, for instance, the Colless index for bifurcating trees: the smaller the value of $\mathfrak{C}_{D,f}(T)$, the more balanced is T relative to the f-size δ_f.

It is clear that $\mathfrak{C}_{D,f}$ satisfies the following recursion:

$$\mathfrak{C}_{D,f}(T_1 \star \cdots \star T_k) = \mathfrak{C}_{D,f}(T_1) + \cdots + \mathfrak{C}_{D,f}(T_k) + D(\delta_f(T_1), \ldots, \delta_f(T_k)).$$

Therefore these Colless-like indices are *recursive tree shape statistics* in the sense of [15], relative to the f-size δ_f. Table A in S2 File also gives the abstract values of $\mathfrak{C}_{D,f}(T)$, for D = MDM, var, and *sd*, and for every $T \in \mathcal{T}_n^*$ with $n = 2, 3, 4, 5$.

The next result shows that, if we take D = MDM or D = *sd*, then any index $\mathfrak{C}_{D,f}$ restricted to only bifurcating trees defines, up to a constant factor, the usual Colless index.

Proposition 6. *Let T be a bifurcating tree with n leaves and $f : \mathbb{N} \to \mathbb{R}_{\geq 0}$ any function. Then,*

$$\mathfrak{C}_{\mathrm{MDM},f}(T) = \frac{f(0)+f(2)}{2} \cdot C(T), \qquad \mathfrak{C}_{sd,f}(T) = \frac{f(0)+f(2)}{\sqrt{2}} \cdot C(T).$$

Proof. Notice that, for every $x, y \in \mathbb{R}$, $\mathrm{MDM}(x,y) = \frac{1}{2}|x-y|$ and $sd(x,y) = \frac{1}{\sqrt{2}}|x-y|$. We shall prove the statement for MDM; the proof for *sd* is identical, replacing the 2 in the denominator by $\sqrt{2}$. For every internal node v in a bifurcating tree T, if v_1 and v_2 denote its children,

$$\begin{aligned} bal_{MDM,f}(v) &= \frac{1}{2}|\delta_f(T_{v_1}) - \delta_f(T_{v_2})| \\ &= \frac{1}{2}|((f(0)+f(2))\kappa_{v_1} - f(2)) - ((f(0)+f(2))\kappa_{v_2} - f(2))| \\ &\quad \text{(by Example 2)} \\ &= \frac{f(0)+f(2)}{2} \cdot |\kappa_{v_1} - \kappa_{v_2}| \end{aligned}$$

and therefore

$$\begin{aligned} \mathfrak{C}_{MDM,f}(T) &= \sum_{v \in V_{int}(T)} bal_{MDM,f}(v) = \frac{f(0)+f(2)}{2} \cdot \sum_{v \in V_{int}(T)} |\kappa_{v_1} - \kappa_{v_2}| \\ &= \frac{f(0)+f(2)}{2} \cdot C(T), \end{aligned}$$

as we claimed.

If we define the *quadratic Colless index* of a bifurcating tree T as

$$C^{(2)}(T) = \sum_{v \in V_{int}(T)} (\kappa_{v_1} - \kappa_{v_2})^2$$

(where, for every $v \in V_{int}(T)$, v_1, v_2 denote its children), then, given that $\mathrm{var}(x,y) = \frac{1}{2}(x-y)^2$, a similar argument proves the following result.

Proposition 7. *Let T be a bifurcating tree with n leaves and $f : \mathbb{N} \to \mathbb{R}_{\geq 0}$ any function. Then,*

$$\mathfrak{C}_{\mathrm{var},f}(T) = \frac{(f(0)+f(2))^2}{2} \cdot C^{(2)}(T).$$

As for the cost of computing Colless-like indices, we have the following result.

Proposition 8. *If the cost of computing $D(x_1, \ldots, x_k)$ is in $O(k)$ and the cost of computing each $f(k)$ is at most in $O(k)$, then, for every $T \in \mathcal{T}_n^*$, the cost of computing $\mathfrak{C}_{D,f}(T)$ is in $O(n)$.*

Proof. Assume that every $f(k)$ is computed in time at most $O(k)$. For every $k \geq 2$, let m_k be the number of internal nodes in T of out-degree k. Since the sizes $\delta_f(v)$ are additive, in the sense that if v has children $v_1, \ldots, v_k$, then $\delta_f(v) = \sum_{i=1}^{k} \delta_f(v_i) + f(k)$, we can compute the whole vector $(\delta_f(v))_{v \in V(T)}$ in time $O(n + \sum_{k \geq 2} m_k \cdot k) = O(n)$ by traversing the tree in post-order.

Assume now that $D(x_1, \ldots, x_k)$ can be computed in time $O(k)$. Then, for every internal node v of out-degree k, $bal_{D,f}(v) = D(\delta_f(T_{v_1}), \ldots, \delta_f(T_{v_k}))$ can be computed in time $O(k)$, by simply reading the k f-sizes of its children (which are already computed) and applying D to them. This shows that the whole vector $(bal_{D,f}(v))_{v \in V(T)}$ can be computed again in time $O(\sum_{k \geq 2} m_k \cdot k) = O(n)$. Finally, we compute $\mathfrak{C}_{D,f}(T)$ by adding up the entries of $(bal_{D,f}(v))_{v \in V(T)}$, which still can be done in time $O(n)$.

The dissimilarities mentioned previously in this subsection can be computed in a number of sums and multiplications that is linear in the length of the input vector, and the specific functions f that we shall consider in the next subsection, basically exponentials and logarithms, can be approximated to any desired precision in constant time by using addition and look-up tables [16].

Sound Colless-like indices

It is clear that for every dissimilarity D, for every function $f : \mathbb{N} \to \mathbb{R}_{\geq 0}$ and for every fully symmetric tree $FS_{n_1,\ldots,n_k}$, $\mathfrak{C}_{D,f}(FS_{n_1,\ldots,n_k}) = 0$ because $bal_{D,f}(v) = 0$ for every $v \in V_{int}(FS_{n_1,\ldots,n_k})$. We shall say that a Colless-like index $\mathfrak{C}_{D,f}$ is *sound* when the converse implication is true.

Definition 9. *A Colless-like index $\mathfrak{C}_{D,f}$ is sound when, for every $T \in \mathcal{T}^*$, $\mathfrak{C}_{D,f}(T) = 0$ if, and only if, T is fully symmetric.*

In other words, $\mathfrak{C}_{D,f}$ is sound when, according to it, the most balanced trees are exactly the fully symmetric trees.

The Colless index C and its quadratic version $C^{(2)}$ are sound for *bifurcating* trees. Unfortunately, this is not always so for Colless-like indices for multifurcating trees. It is not so even for direct generalizations of C or $C^{(2)}$. For instance, $\mathfrak{C}_{\mathrm{MDM},\kappa}$, $\mathfrak{C}_{sd,\kappa}$ and $\mathfrak{C}_{\mathrm{var},\kappa}$, where κ denotes the number of leaves, are not sound; neither are $\mathfrak{C}_{\mathrm{MDM},\tau}$, $\mathfrak{C}_{sd,\tau}$ and $\mathfrak{C}_{\mathrm{var},\tau}$, where τ denotes the number of nodes; and they are not sound even when replacing τ by θ, the usual size, which is simply $\tau - 1$. For example, the tree T in Fig 1 is not fully symmetric, but $\mathfrak{C}_{\mathrm{MDM},\kappa}(T) = \mathfrak{C}_{\mathrm{var},\kappa}(T) = \mathfrak{C}_{sd,\kappa}(T) = \mathfrak{C}_{\mathrm{MDM},\tau}(T) = \mathfrak{C}_{\mathrm{var},\tau}(T) = \mathfrak{C}_{sd,\tau}(T) = 0$.

The following lemma shows that the soundness of $\mathfrak{C}_{D,f}(T) = 0$ does not depend on D, but only on f.

Lemma 10. *$\mathfrak{C}_{D,f}$ is sound if, and only if, $\delta_f(T_1) \neq \delta_f(T_2)$ for every pair of different fully symmetric trees T_1, T_2.*

Proof. For the "only if" implication: if there exist two different (i.e., non isomorphic) fully symmetric trees T_1, T_2 such that $\delta_f(T_1) = \delta_f(T_2)$, then the tree $T = T_1 \star T_2$ is not fully symmetric, but

$$\mathfrak{C}_{D,f}(T) = \mathfrak{C}_{D,f}(T_1) + \mathfrak{C}_{D,f}(T_2) + D(\delta_f(T_1), \delta_f(T_2)) = 0.$$

Conversely, assume that, for every pair of fully symmetric trees T_1, T_2, if $\delta_f(T_1) = \delta_f(T_2)$ then $T_1 = T_2$. We shall prove by complete induction on n that if T is a tree with n leaves such that

$\mathfrak{C}_{D,f}(T) = 0$, then T is fully symmetric. If T has only one leaf, it is clearly fully symmetric. Now, assume that $n > 1$ and hence that T has depth at least 1. Let $T_1, \ldots, T_k$, $k \geq 2$, be its subtrees rooted at the children of its root, so that $T = T_1 \star \cdots \star T_k$. Then,

$$0 = \mathfrak{C}_{D,f}(T) = \sum_{i=1}^{k} \mathfrak{C}_{D,f}(T_i) + D(\delta_f(T_1), \ldots, \delta_f(T_k))$$

implies, on the one hand, that $\mathfrak{C}_{D,f}(T_1) = \cdots = \mathfrak{C}_{D,f}(T_k) = 0$, and hence, by induction, that $T_1, \ldots, T_k$ are fully symmetric, and, on the other hand, that $D(\delta_f(T_1), \ldots, \delta_f(T_k)) = 0$, and hence that $\delta_f(T_1) = \cdots = \delta_f(T_k)$, which, by assumption, implies that $T_1 = \cdots = T_k$: in summary, T is fully symmetric.

The following problem now arises:

Problem. *To find functions* $f : \mathbb{N} \to \mathbb{R}_{\geq 0}$ *such that* $\mathfrak{C}_{D,f}$ *is sound.*

Unfortunately, many natural functions f do not define sound Colless-like indices, as the following examples show.

Example 11. *If* $f(n) = an^2 + bn + c$, *for any* a, b, c, *then* $\mathfrak{C}_{D,f}$ *is not sound, because, for example,* $\delta_f(FS_{2,2,2,7}) = \delta_f(FS_{14,4}) = 420a + 70b + 71c$.

Example 12. *If* $f(n) = n^d$, *for any* $d \geq 0$, *then* $\mathfrak{C}_{D,f}$ *is not sound. Indeed, for every* $d \geq 3$ (*the case when* $d \leq 2$ *is a particular case of the last example*), *take*

- $k = 2^d + 1$ *and* $l = 2$;
- $n_i = 2^{(d-1)^i d^{k-i-1}}$ *for* $i = 1, \ldots, k-1$;
- $n_k = 2$;
- $m_1 = 2^{(d-1)d^{k-2}+1}$;
- $m_2 = 2^{((d-1)^2(d^{k-2}-(d-1)^{k-2})+d-1)/d}$; *notice that this exponent is an integer number, because k is odd and therefore d divides* $(d-1)^k + 1$.

Then

$$n_1 \ldots n_{i-1} \cdot n_i^d = n_1^d$$

and hence, on the one hand,

$$\begin{aligned} n_1^d + \cdots + n_1 \cdots n_{k-2} \cdot n_{k-1}^d &= (k-1)n_1^d = 2^d \cdot 2^{(d-1)d^{k-1}} \\ &= (2^{1+(d-1)d^{k-2}})^d = m_1^d, \end{aligned}$$

and, on the other hand,

$$\begin{aligned} n_1 \cdots n_{k-1} \cdot n_k^d &= n_1^{\frac{\left(1-\left(\frac{d-1}{d}\right)^{k-1}\right)}{\left(1-\frac{d-1}{d}\right)}} \cdot n_k^d = n_1^{\frac{d^{k-1}-(d-1)^{k-1}}{d^{k-2}}} n_k^d \\ &= 2^{(d-1)(d^{k-1}-(d-1)^{k-1})+d} = m_1 m_2^d. \end{aligned}$$

Therefore, $\delta_{n^d}(FS_{n_1,\ldots,n_k}) = \delta_{n^d}(FS_{m_1,m_2})$.

Of course, for any given d there may exist "smaller" counterexamples: for instance, $\delta_{n^3}(FS_{2,10,4}) = \delta_{n^3}(FS_{6,8}) = 3288$ *and* $\delta_{n^4}(FS_{2,6,2,3}) = \delta_{n^4}(FS_{8,3}) = 4744$.

Example 13. *If* $f(n) = \log_a(n)$ *(for some* $a > 1$*) when* $n > 0$*, and* $f(0) = 0$*, then* $\mathfrak{C}_{D,f}$ *is not sound: for instance,* $\delta_f(FS_{2,2}) = \delta_f(FS_8) = \log_a(8)$*. In a similar way, if* $f(n) = \log_a(n+1)$ *(for some* $a > 1$*), then* $\mathfrak{C}_{D,f}$ *is not sound, either: for instance,* $\delta_f(FS_{2,3,3}) = \delta_f(FS_{5,7}) = \log_a(196608)$.

On the positive side, we shall show now two functions that define sound indices. The following lemmas will be useful to prove it.

Lemma 14. *For every* $k, l \geq 1$ and $n_1, n_2, \ldots, n_k, m_1, m_2, \ldots, m_l \geq 2$, *if* $\delta_f(FS_{n_1,n_2,\ldots,n_k}) = \delta_f(FS_{m_1,m_2,\ldots,m_l})$, $n_1 \cdot n_2 \cdots n_k = m_1 \cdot m_2 \cdots m_l$, *and* $n_k = m_l$, *then* $\delta_f(FS_{n_1,\ldots,n_{k-1}}) = \delta_f(FS_{m_1,\ldots,m_{l-1}})$.

Proof. *If* $n_1 \cdots n_k = m_1 \cdots m_l$ and $n_k = m_l$, then $n_1 \cdots n_{k-1} = m_1 \cdots m_{l-1}$. If, moreover, $\delta_f(FS_{n_1,n_2,\ldots,n_k}) = \delta_f(FS_{m_1,m_2,\ldots,m_l})$, that is,

$$\begin{aligned} &n_1 \cdots n_k f(0) + n_1 \cdots n_{k-1} f(n_k) + n_1 \cdots n_{k-2} f(n_{k-1}) + \cdots + f(n_1) \\ &\quad = m_1 \cdots m_l f(0) + m_1 \cdots m_{l-1} f(m_l) + m_1 \cdots m_{l-2} f(m_{l-1}) + \cdots + f(m_1), \end{aligned}$$

then

$$\begin{aligned} &n_1 \cdots n_{k-2} f(n_{k-1}) + \cdots + n_1 f(n_2) + f(n_1) \\ &\quad = m_1 \cdots m_{l-2} f(m_{l-1}) + \cdots + m_1 f(m_2) + f(m_1) \end{aligned}$$

and hence

$$\begin{aligned} &\delta_f(FS_{n_1,n_2,\ldots,n_{k-1}}) \\ &\quad = n_1 \cdots n_{k-1} f(0) + n_1 \cdots n_{k-2} f(n_{k-1}) + \cdots + n_1 f(n_2) + f(n_1) \\ &\quad = m_1 \cdots m_{l-1} f(0) + m_1 \cdots m_{l-2} f(m_{l-1}) + \cdots + m_1 f(m_2) + f(m_1) \\ &\quad = \delta_f(FS_{m_1,\ldots,m_{l-1}}) \end{aligned}$$

as claimed.

Lemma 15. *If* $n_1, \ldots, n_k \geq 2$, *then*

$$1 + n_1 + n_1 n_2 + \cdots + n_1 \cdots n_{k-1} < n_1 \cdots n_k.$$

Proof. By induction on k. If $k = 1$, the statement says that $1 < n_1$, which is true by assumption. Assume now that the statement is true for any $n_1, \ldots, n_k \geq 2$, and let $n_{k+1} \geq 2$. Then,

$$\begin{aligned} 1 + n_1 + n_1 n_2 + \cdots + n_1 \cdots n_{k-1} + n_1 \cdots n_k &< n_1 \cdots n_k + n_1 \cdots n_k \\ &= 2 n_1 \cdots n_k \leq n_1 \cdots n_k \cdot n_{k+1}. \end{aligned}$$

Proposition 16. *If* $f(n) = e^n$, *then* $\mathfrak{C}_{D,f}$ *is sound*.

Proof. Assume that there exist two non-isomorphic fully symmetric trees $FS_{n_1,\ldots,n_k}$ and $FS_{m_1,\ldots,m_l}$ such that

$$\delta_{e^n}(FS_{n_1,\ldots,n_k}) = \delta_{e^n}(FS_{m_1,\ldots,m_l}),$$

that is, such that

$$\begin{aligned} &n_1 \cdots n_k + n_1 \cdots n_{k-1} e^{n_k} + \cdots + n_1 e^{n_2} + e^{n_1} \\ &\quad = m_1 \cdots m_l + m_1 \cdots m_{l-1} e^{m_l} + \cdots + m_1 e^{m_2} + e^{m_1}. \end{aligned}$$

Assume that l is the smallest depth of a fully symmetric tree with e^n-size equal to the e^n-size of another fully symmetric tree non-isomorphic to it.

Since e is transcendental, the equality above implies the equality of polynomials in $\mathbb{Z}[x]$

$$n_1 \cdots n_k + n_1 \cdots n_{k-1} x^{n_k} + \cdots + n_1 x^{n_2} + x^{n_1} = m_1 \cdots m_l + m_1 \cdots m_{l-1} x^{m_l} + \cdots + m_1 x^{m_2} + x^{m_1}.$$

If $l = 1$, the right-hand side polynomial is simply $m_1 + x^{m_1}$ and then the equality of polynomials implies that $k = 1$ and $n_1 = m_1$, which contradicts the assumption that $FS_{n_1,\ldots,n_k} \neq FS_{m_1,\ldots,m_l}$. Now assume that $l \geq 2$. This equality of polynomials implies the equality of their independent terms: $n_1 \cdots n_k = m_1 \cdots m_l$. On the other hand, the non-zeroth power of x with the largest coefficient in the left-hand side polynomial is x^{n_k} (because all coefficients are non-negative, and, by Lemma 15, $n_1 \cdots n_{k-1}$ alone is larger than the sum $n_1 \cdots n_{k-2} + \cdots + n_1 + 1$ of all other coefficients of non-zeroth powers of x) and, by the same reason, the non-zeroth power of x with the largest coefficient in the right-hand side polynomial is x^{m_l}. The equality of polynomials implies then that $n_k = m_l$ and hence, by Lemma 14, that $\delta_{e^n}(FS_{n_1,\ldots,n_{k-1}}) = \delta_{e^n}(FS_{m_1,\ldots,m_{l-1}})$, against the assumption on l. We reach thus a contradiction that implies that there does not exist any pair of non-isomorphic fully symmetric trees with the same e^n-size. By Lemma 10, this implies that $\mathfrak{C}_{D,e^n}$ is sound.

The same argument shows that $\mathfrak{C}_{D,f}$ is sound for every exponential function $f(n) = r^n$ with base r a transcendental real number. However, if r is not transcendental, then $\mathfrak{C}_{D,r^n}$ need not be sound. For instance, $\delta_{2^n}(FS_{2,3}) = \delta_{2^n}(FS_{3,2}) = 26$ and $\delta_{\sqrt{2}^n}(FS_{8,10}) = \delta_{\sqrt{2}^n}(FS_{12,8}) = 352$.

Proposition 17. *If* $f(n) = \ln(n + e)$, *then* $\mathfrak{C}_{D,f}$ *is sound*.

Proof. The argument is similar to that of the previous proof. Let $f(n) = \ln(n + e)$ and assume that there exist two non-isomorphic fully symmetric trees $FS_{n_1,\ldots,n_k}$ and $FS_{m_1,\ldots,m_l}$ such that $\delta_f(FS_{n_1,\ldots,n_k}) = \delta_f(FS_{m_1,\ldots,m_l})$, that is, such that

$$n_1 \cdots n_k + n_1 \cdots n_{k-1} \ln(n_k + e) + \cdots + \ln(n_1 + e) = m_1 \cdots m_l + m_1 \cdots m_{l-1} \ln(m_l + e) + \cdots + \ln(m_1 + e).$$

Assume that l is the smallest depth of a fully symmetric tree with f-size equal to the f-size of a fully symmetric tree non-isomorphic to it.

Applying the exponential function to both sides of the equality above, we obtain

$$e^{n_1 \cdots n_k}(n_k + e)^{n_1 \cdots n_{k-1}} \cdots (n_2 + e)^{n_1}(n_1 + e) = e^{m_1 \cdots m_l}(m_l + e)^{m_1 \cdots m_{l-1}} \cdots (m_2 + e)^{m_1}(m_1 + e).$$

Since e is transcendental, this implies the equality of polynomials in $\mathbb{Z}[x]$

$$x^{n_1 \cdots n_k}(n_k + x)^{n_1 \cdots n_{k-1}} \cdots (n_2 + x)^{n_1}(n_1 + x) = x^{m_1 \cdots m_l}(m_l + x)^{m_1 \cdots m_{l-1}} \cdots (m_2 + x)^{m_1}(m_1 + x),$$

which, since $n_1, \ldots, n_k, m_1, \ldots, m_l \geq 2$, on its turn implies the equalities

$$x^{n_1 \cdots n_k} = x^{m_1 \cdots m_l}, \text{ i.e., } n_1 \cdots n_k = m_1 \cdots m_l,$$

$$(x + n_k)^{n_1 \cdots n_{k-1}} \cdots (x + n_2)^{n_1}(x + n_1) = (x + m_l)^{m_1 \cdots m_{l-1}} \cdots (x + m_2)^{m_1}(x + m_1).$$

If $l = 1$, the right-hand side polynomial in the second equality is simply $x + m_1$ and then this equality of polynomials implies that $k = 1$ and $n_1 = m_1$, which contradicts the assumption that $FS_{n_1,\dots,n_k} \neq FS_{m_1,\dots,m_l}$. Now assume that $l \geq 2$. From the first equality we know that $n_1 \cdots n_k = m_1 \cdots m_l$. But, the root of the left-hand side polynomial in the second equality with largest multiplicity is $-n_k$ (because, by Lemma 15, $n_1 \cdots n_{k-1}$ alone is greater than the degree of $(x + n_{k-1})^{n_1 \cdots n_{k-2}} \cdots (x + n_2)^{n_1}(x + n_1)$) and, similarly, the root of the right-hand side polynomial in the second equality with largest multiplicity is $-m_l$. Therefore, the equality of both polynomials implies that $n_k = m_l$ and hence, by Lemma 14, $\delta_f(FS_{n_1,\dots,n_{k-1}}) = \delta_f(FS_{m_1,\dots,m_{l-1}})$, against the assumption on l. As in the previous proof, this contradiction implies that $\mathfrak{C}_{D,f}$ is sound.

The same argument proves that, for every transcendental number $r > 1$, the function $f(n) = \log_r(n + r)$ defines sound indices $\mathfrak{C}_{D,f}$. However, if r is not transcendental, then such a $\mathfrak{C}_{D,f}$ need not be sound. For instance, $\delta_{\log_2(n+2)}(FS_{9,6}) = \delta_{\log_2(n+2)}(FS_{20,2}) = 81 + \log_2(11)$.

Summarizing, each one of the functions $f(n) = \ln(n + e)$ and $f(n) = e^n$ defines, for every dissimilarity D, a Colless-like index $\mathfrak{C}_{D,f}$ that reaches its minimum value on each $\mathcal{T}_n^*$, 0, at exactly the fully symmetric trees.

Results

Maximally unbalanced trees

The next results give the maximum values of $\mathfrak{C}_{D,f}$ on $\mathcal{T}_n^*$ when D = MDM, var or *sd* and $f(n) = \ln(n+ e)$ or $f(n) = e^n$. These maxima define the range of each $\mathfrak{C}_{D,f}$ on $\mathcal{T}_n^*$, and, dividing by them, we can define normalized Colless-like indices that can be used to compare the balance of trees with different numbers of leaves.

We begin with the function $f(n) = \ln(n + e)$, which is covered by the following theorem.

Theorem 18. *Let f be a function* $\mathbb{N} \to \mathbb{R}_{\geq 0}$ *such that* $0 < f(k) < f(k-1) + f(2)$, *for every* $k \geq 3$. *Then, for every* $n \geq 2$, *the indices* $\mathfrak{C}_{\mathrm{MDM},f}$, $\mathfrak{C}_{sd,f}$ *and* $\mathfrak{C}_{\mathrm{var},f}$ *reach their maximum values on* $\mathcal{T}_n^*$ *exactly at the comb* K_n. *These maximum values are, respectively,*

$$\mathfrak{C}_{\mathrm{MDM},\delta_f}(K_n) = \frac{f(0) + f(2)}{4}(n-1)(n-2),$$

$$\mathfrak{C}_{sd,\delta_f}(K_n) = \frac{f(0) + f(2)}{2\sqrt{2}}(n-1)(n-2),$$

$$\mathfrak{C}_{\mathrm{var},\delta_f}(K_n) = \frac{(f(0) + f(2))^2}{12}(n-1)(n-2)(2n-3).$$

The proof of this theorem is very long, and we devote to it the first three sections in S1 File, one section for each dissimilarity.

It is straightforward to check that the function $f(n) = \ln(n + e)$ satisfies the hypothesis of Theorem 18 (for the inequality $f(k) \leq f(k-1) + f(2)$, notice that $\ln(k + e) \leq \ln(k + e - 1) + \ln(2)$ if, and only if, $k + e \leq 2(k + e - 1)$, and this last inequality is true for every $k \in \mathbb{N}$). Therefore, $\mathfrak{C}_{\mathrm{MDM},\ln(n+e)}$, $\mathfrak{C}_{\mathrm{var},\ln(n+e)}$, and $\mathfrak{C}_{sd,\ln(n+e)}$ take their maximum values on $\mathcal{T}_n^*$ at the comb K_n. In other words, the combs are the most unbalanced trees according to these indices. Table B in S2 File gives the values of $\mathfrak{C}_{\mathrm{MDM},\ln(n+e)}$, $\mathfrak{C}_{\mathrm{var},\ln(n+e)}$, and $\mathfrak{C}_{sd,\ln(n+e)}$ on $\mathcal{T}_n^*$, for $n = 2, 3, 4, 5$, and the positions of the different trees in each $\mathcal{T}_n^*$ according to the increasing order of the corresponding index.

And for $f(n) = e^n$, we have the following result. We have also moved its proof to the S1 File.

Theorem 19. *For every* $n \geq 2$:

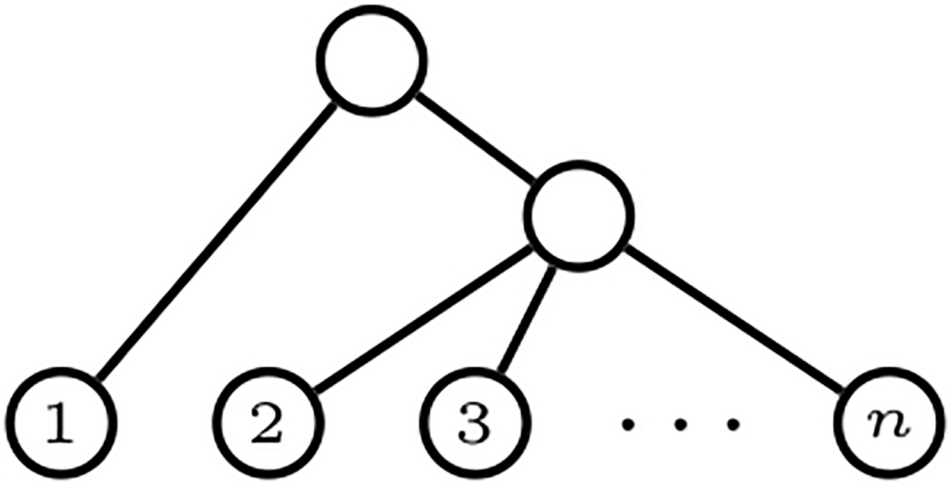

Fig 7. The tree $FS_1 \star FS_{n-1}$.

https://doi.org/10.1371/journal.pone.0203401.g007

(*a*) *If* $n \neq 4$, *then both* $\mathfrak{C}_{\mathrm{MDM},e^n}$ *and* $\mathfrak{C}_{sd,e^n}$ *reach their maximum on* $\mathcal{T}_n^*$ *exactly at the tree* $FS_1 \star FS_{n-1}$ (*see* Fig 7), *and these maximum values are*

$$\mathfrak{C}_{\mathrm{MDM},e^n}(FS_1 \star FS_{n-1}) = \frac{1}{2}(e^{n-1} + n - 2),$$
$$\mathfrak{C}_{sd,e^n}(FS_1 \star FS_{n-1}) = \frac{1}{\sqrt{2}}(e^{n-1} + n - 2).$$

(*b*) *Both* $\mathfrak{C}_{\mathrm{MDM},e^n}$ *and* $\mathfrak{C}_{sd,e^n}$ *reach their maximum on* $\mathcal{T}_4^*$ *exactly at the comb* K_4, *and these maximum values are*

$$\mathfrak{C}_{\mathrm{MDM},e^n}(K_4) = \frac{3}{2}(e^2 + 1),$$
$$\mathfrak{C}_{sd,e^n}(K_4) = \frac{3}{\sqrt{2}}(e^2 + 1).$$

(*c*) $\mathfrak{C}_{\mathrm{var},e^n}$ *always reaches its maximum on* $\mathcal{T}_n^*$ *exactly at the tree* $FS_1 \star FS_{n-1}$, *and the maximum value is*

$$\mathfrak{C}_{\mathrm{var},e^n}(FS_1 \star FS_{n-1}) = \frac{1}{2}(e^{n-1} + n - 2)^2.$$

So, according to $\mathfrak{C}_{\mathrm{MDM},e^n}$, $\mathfrak{C}_{\mathrm{var},e^n}$, and $\mathfrak{C}_{sd,e^n}$, the trees of the form $FS_1 \star FS_{n-1}$ are the most unbalanced (except for $n = 4$ and D = MDM or *sd*, in which case the most unbalanced tree is the comb). Table B in S2 File also gives the values of these indices on $\mathcal{T}_n^*$, for $n = 2, 3, 4, 5$, and the positions of the different trees in each $\mathcal{T}_n^*$ according to the increasing order of the corresponding index.

The R package "CollessLike"

We have written an R package called *CollessLike*, available at the CRAN (https://cran.r-project.org/web/packages/CollessLike/index.html), that computes the Colless-like indices and their

normalized version, as well as several other balance indices, and simulates the distribution of these indices on $\mathcal{T}_n$ under the α-γ-model [10]. This package contains functions that:

- Compute the following balance indices for multifurcating trees: the Sackin index S [4–6], the total cophenetic index Φ [8], and the Colles-like index $\mathfrak{C}_{D,f}$ for several predefined dissimilarities D and functions f as well as for any user-defined ones.
  Our functions also compute the normalized versions (obtained by subtracting their minimum value and dividing by the width of their range, so that they take values in [0, 1]) of S, Φ and the Colless-like indices $\mathfrak{C}_{D,f}$ for which we have computed the range in Theorems 18 and 19. Recall from the aforementioned references that, for every $n \geq 2$:
  - the range of S on $\mathcal{T}_n^*$ is $S(FS_n) = n$ to $S(K_n) = \frac{1}{2}(n+2)(n-1)$
  - the range of Φ on $\mathcal{T}_n^*$ is $\Phi(FS_n) = 0$ to $\Phi(K_n) = \binom{n}{3}$

  Therefore, for every $T \in \mathcal{T}_n$, the normalized Sackin and total cophenetic index are, respectively,

$$S_{norm}(T) = \frac{S(T) - n}{\frac{1}{2}(n+2)(n-1) - n}, \quad \Phi_{norm}(T) = \frac{\Phi(T)}{\binom{n}{3}},$$

  while, for instance, the normalized version of $\mathfrak{C}_{\mathrm{MDM},\ln(n+e)}$ is

$$\mathfrak{C}_{\mathrm{MDM},\ln(n+e),norm}(T) = \frac{\mathfrak{C}(T)}{\frac{1+\ln(e+2)}{4}(n-1)(n-2)}.$$

- Given two natural numbers n and N, produce a random sample of N values of a balance index S, Φ, or $\mathfrak{C}_{D,f}$ on trees in $\mathcal{T}_n$ generated following an α-γ-model: the parameters N, n, α, γ (with $0 \leq \gamma \leq \alpha \leq 1$) can be set by the user.
  Due to the computational cost of this function, we have stored the values of S, Φ, and $\mathfrak{C}_{\mathrm{MDM},\ln(n+e)}$ (denoted henceforth simply by $\mathfrak{C}$) on the random samples of $N = 5000$ trees in each $\mathcal{T}_n$ (for every $n = 3, \ldots, 50$ and for every $\alpha, \gamma \in \{0, 0.1, 0.2, \ldots, 0.9, 1\}$ with $\gamma \leq \alpha$) generated in the study reported in the next subsection. In this way, if the user is interested in this range of numbers of leaves and this range of parameters, he or she can study the distribution of the corresponding balance index efficiently and quickly.
- Given a tree $T \in \mathcal{T}_n$, estimate the percentile $q_{T,n,\alpha,\gamma}$ of its balance index S, Φ, or $\mathfrak{C}_{D,f}$ with respect to the distribution of this index on $\mathcal{T}_n$ under some α-γ-model. If n, α, γ are among those mentioned in the previous item, for the sake of efficiency this function uses the database of computed indices to simulate the distribution of the balance index on $\mathcal{T}_n$ under this α-γ-model.

For instance, the unlabeled tree $T \in \mathcal{T}_8^*$ in Fig 8 is the shape of a phylogenetic tree randomly generated under the α-γ-model with $\alpha = 0.7$ and $\gamma = 0.4$ (using `set.seed(1000)` for reproducibility). The values of its balance indices are given in the figure's caption.

Fig 9 displays the estimation of the density function of the balance indices $\mathfrak{C}$, S, and Φ under the α-γ-model with $\alpha = 0.7$ and $\gamma = 0.4$ on $\mathcal{T}_8$, obtained using the 5000 random trees gathered in our database. Moreover, the estimated percentiles of the balance indices of the tree of Fig 8 are also shown in the figure.

Fig 10 shows a percentile plot of $\mathfrak{C}$, S, and Φ under the α-γ-model for $\alpha = 0.7$ and $\gamma = 0.4$ on $\mathcal{T}_8$. The percentiles of the tree of Fig 8 are given by the area to the left of the vertical lines.

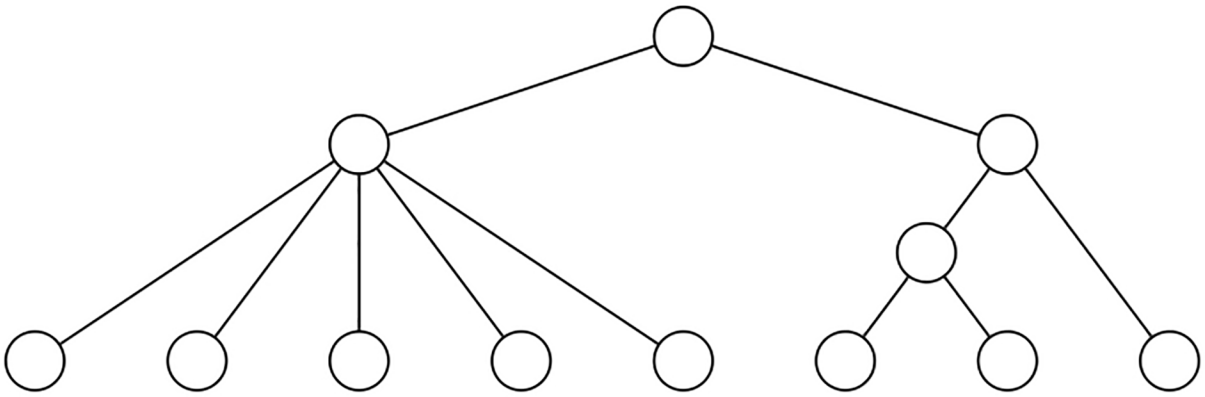

Fig 8. A tree with 8 leaves randomly generated under the α-γ-model with $\alpha = 0.7$ and $\gamma = 0.4$. Its indices are $\mathfrak{C}(T) = 1.746$, $S(T) = 18$, and $\Phi(T) = 14$, and its normalized indices are $\mathfrak{C}_{norm}(T) = 0.06518$, $S_{norm}(T) = 0.3704$, and $\Phi_{norm}(T) = 0.25$.

https://doi.org/10.1371/journal.pone.0203401.g008

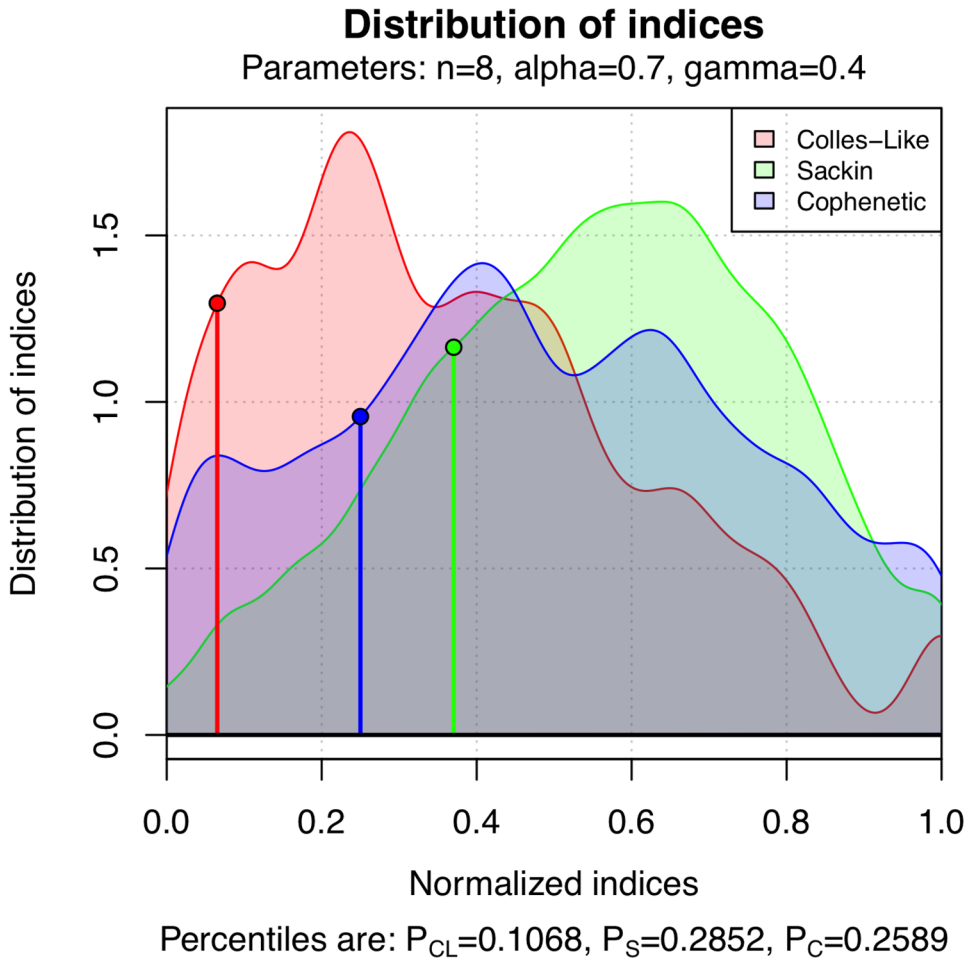

Fig 9. The estimated density function of the distribution of $\mathfrak{C}$, S and Φ on $\mathcal{T}_8$ under the α-γ-model with $\alpha = 0.7$ and $\gamma = 0.4$. The percentiles of the tree in Fig 8 are also represented.

https://doi.org/10.1371/journal.pone.0203401.g009

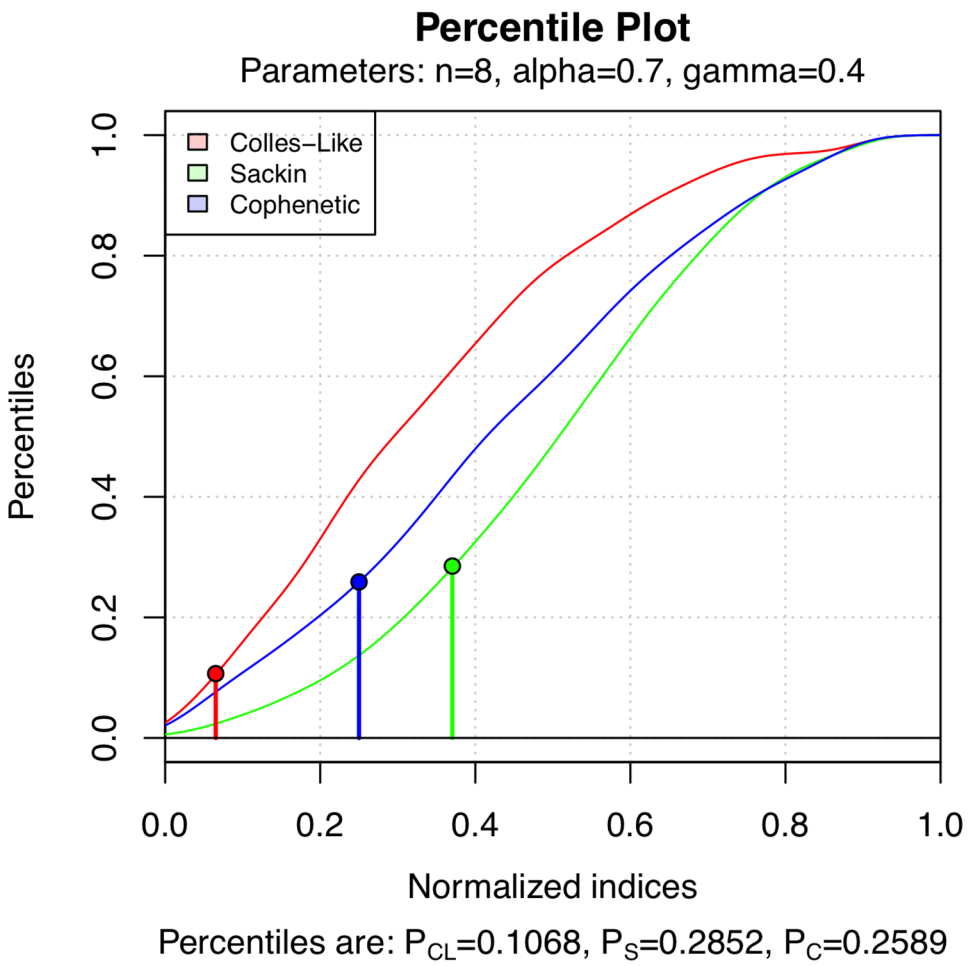

Fig 10. Percentile plot of the distribution of $\mathfrak{C}$, S and Φ on $\mathcal{T}_8$ under the α-γ-model with $\alpha = 0.7$ and $\gamma = 0.4$. The percentiles of the tree of Fig 8 are also highlighted.

https://doi.org/10.1371/journal.pone.0203401.g010

A special case of the α-γ-model, corresponding to the case $\alpha = \gamma$, is Ford's α-model for bifurcating phylogenetic trees [17]. This model includes as special cases the Yule, or Equal-Rate Markov, model [18, 19] and the uniform, or Proportional to Distinguishable Arrangements, model [20, 21]. So, this package allows also to study this model. For example, the unlabeled tree in Fig 11 has been generated (with `set.seed(1000)`) using $n = 8$ and $\alpha = \gamma = 0.5$, which corresponds to the uniform model. The figure also depicts the estimation of the density functions and of the percentile plots of $\mathfrak{C}$, S, and Φ on $\mathcal{T}_8$ under this model, as well as the percentile values of the tree.

Experimental results on TreeBASE

To assess the performance of $\mathfrak{C}_{\mathrm{MDM},\ln(n+\mathrm{e})}$, which we abbreviate by $\mathfrak{C}$, we downloaded (December 13-14, 2015) all phylogenetic trees in the TreeBASE database [11] using the function `search_treebase()` of the R package `treebase` [22]. We obtained 13,008 trees, from which 80 had format problems that prevented R from reading them, so we restricted ourselves to the remaining 12,928 trees. To simplify the language, we shall still refer to this slightly

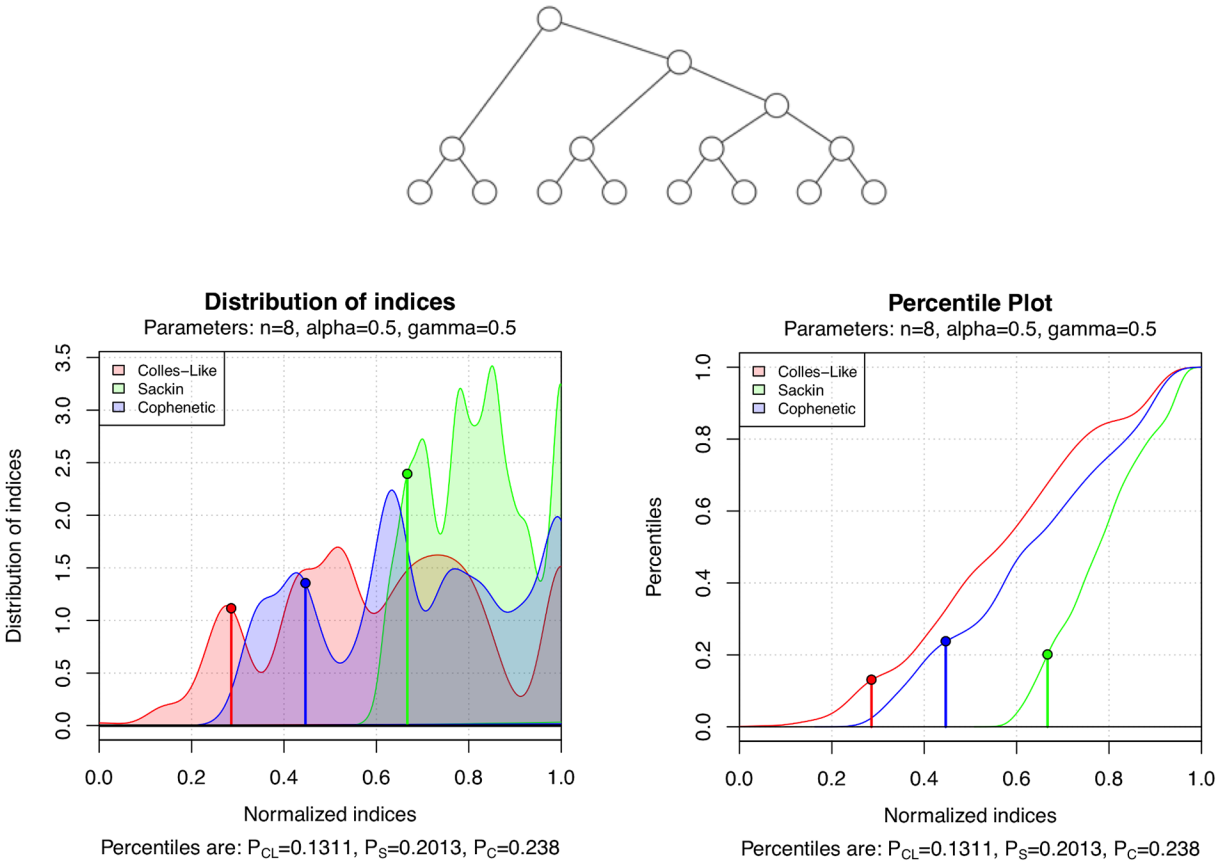

Fig 11. A bifurcating tree randomly generated under the uniform model, the estimated density function of the distribution of the three balance indices on T_8 under the uniform model, and their percentile plot.

https://doi.org/10.1371/journal.pone.0203401.g011

smaller subset of phylogenetic trees as "all trees in TreeBASE". Only 4,814 among these 12,928 trees in TreeBASE are bifurcating.

Then, for every phylogenetic tree T in this set, we have computed its Colless-like index $\mathfrak{C}(T)$, its Sackin index $S(T)$, and its total cophenetic index $\Phi(T)$. We have compared the results in the ways we show next (all analysis have been performed with R [23]).

Behavior as functions of the number of leaves. For every number of leaves n, we have computed the mean and the variance of $\mathfrak{C}$, S and Φ on all trees with n leaves in TreeBASE. Then, we have computed the regression of these values as a function of n.

For the means, the best fits have been:

- *Colless-like index*: $\overline{\mathfrak{C}} \approx 0.5351 \cdot n^{1.5848}$, with a coefficient of determination of $R^2 = 0.9869$ and a p-value for the exponent $p < 2 \cdot 10^{-16}$.
- *Sackin index*: $\overline{S} \approx 1.4512 \cdot n^{1.4359}$, with a coefficient of determination of $R^2 = 0.9953$ and a p-value for the exponent $p < 2 \cdot 10^{-16}$.
- *Total cophenetic index*: $\overline{\Phi} \approx 0.1894 \cdot n^{2.5478}$, with a coefficient of determination of $R^2 = 0.9945$ and a p-value for the exponent $p < 2 \cdot 10^{-16}$.

Fig 12 depicts these mean values of $\mathfrak{C}$ (left), S (center), and Φ (right) as functions of n.

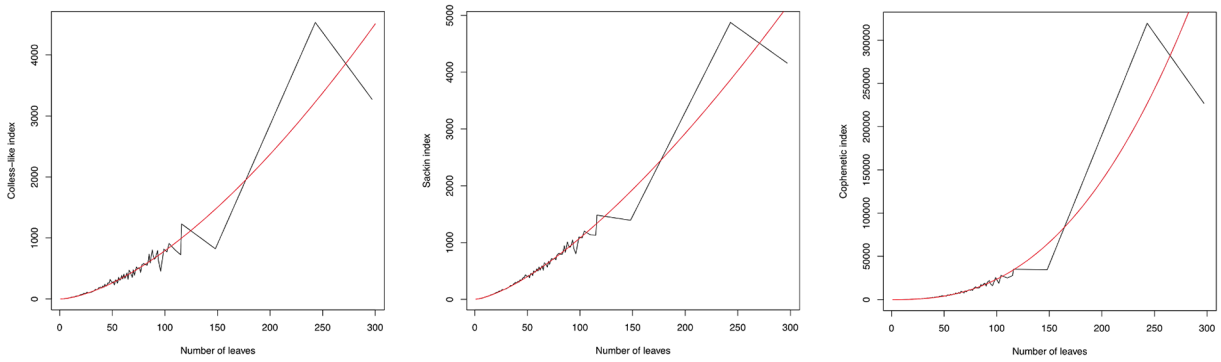

Fig 12. Growth of the mean value of $\mathfrak{C}$ (left), S (center), and Φ (right) in TreeBASE, as functions of the trees' numbers of leaves n.

https://doi.org/10.1371/journal.pone.0203401.g012

Thus, S and $\mathfrak{C}$ have similar mean growth rates, while Φ has a mean growth rate one order higher in magnitude. This difference vanishes if we normalize the indices by their range width, which is $O(n^2)$ for $\mathfrak{C}$ and S, and $O(n^3)$ for Φ:

$$\overline{\mathfrak{C}_{norm}} \approx 0.8389 \cdot n^{-0.4152}$$

$$\overline{S_{norm}} \approx 2.9024 \cdot n^{-0.5641}$$

$$\overline{\Phi_{norm}} \approx 1.1364 \cdot n^{-0.4522}$$

As for the behavior of the variances, the best fits are the following:

- *Colless index*: $\mathrm{var}(\mathfrak{C}) \approx 0.07599 \cdot n^{3.12831}$, with a coefficient of determination of $R^2 = 0.962$ and a p-value for the exponent $p < 2 \cdot 10^{-16}$.
- *Sackin index*: $\mathrm{Var}(S) \approx 0.03182 \cdot n^{3.22441}$, with a coefficient of determination of $R^2 = 0.9575$ and a p-value for the exponent $p < 2 \cdot 10^{-16}$.
- *Total cophenetic index*: $\mathrm{Var}(\Phi) \approx 0.0041 \cdot n^{5.2075}$, with a coefficient of determination of $R^2 = 0.9812$ and a p-value for the exponent $p < 2 \cdot 10^{-16}$.

The results are in the same line as before, with the variances of $\mathfrak{C}$ and S having similar growth rates, and the variance of Φ having a growth rate two orders of magnitude higher. This difference vanishes again when we normalize the indices:

$$\mathrm{var}(\mathfrak{C}_{norm}) \approx 0.18677 \cdot n^{-0.87169}$$

$$\mathrm{var}(S_{norm}) \approx 0.12728 \cdot n^{-0.77559}$$

$$\mathrm{var}(\Phi_{norm}) \approx 0.1476 \cdot n^{-0.7925}$$

So, in summary, $\mathfrak{C}$ has, on TreeBASE and relative to the range of values, a slightly larger mean growth rate and a slightly smaller variance growth rate than the other two indices.

Numbers of ties. The number of ties (that is, of pairs of different trees with the same index value) of a balance index is an interesting measure of quality, because the smaller its frequency of ties, the bigger its ability to rank the balance of any pair of different trees. Although, in our opinion, this ability need not always be an advantage: for instance, neither Φ nor S take the same, minimum, value on all different fully symmetric trees with the same numbers of leaves (for example, $S(FS_6) = 6$ but $S(FS_{2,3}) = S(FS_{3,2}) = 12$; and $\Phi(FS_6) = 0$, but $\Phi(FS_{3,2}) = 3$

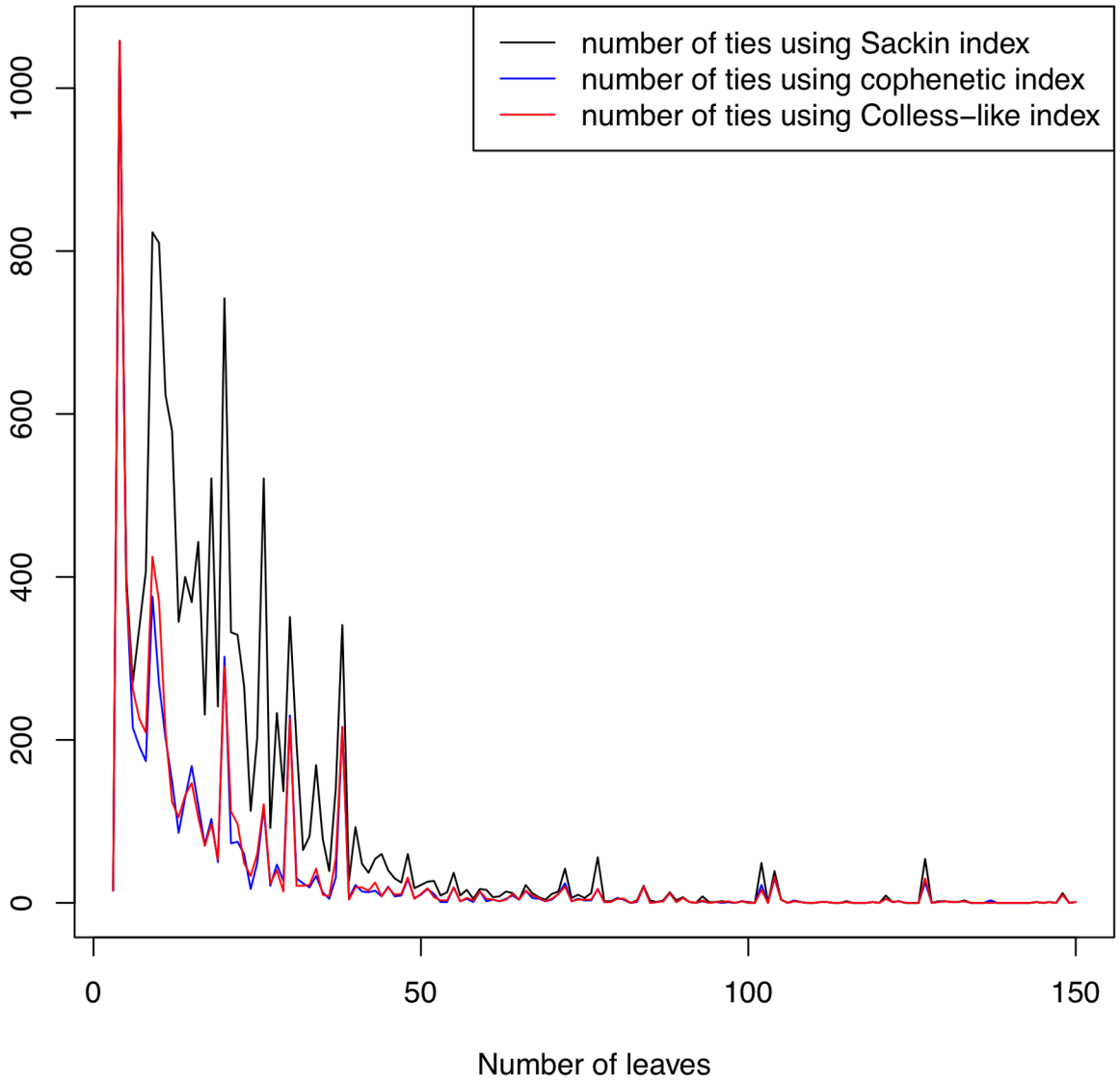

Fig 13. Numbers of ties of $\mathfrak{C}$, S, and Φ in TreeBASE, as functions of the trees' numbers of leaves n.

https://doi.org/10.1371/journal.pone.0203401.g013

and $\Phi(FS_{2,3}) = 6$; cf. Fig 5), while $\mathfrak{C}$ applied to any fully symmetric tree is always 0. In this case, we believe that these ties are fair.

Anyway, for every number of leaves n and for every one of all three indices under scrutiny, we have computed the numbers of pairs of trees with n leaves in TreeBASE having the same value of the corresponding index (in the case of $\mathfrak{C}$, up to 16 decimal digits). Fig 13 plots the frequencies of ties of $\mathfrak{C}$, S and Φ as functions of n. As it can be seen in this graphic, $\mathfrak{C}$ and Φ have a similar number of ties, and consistently less ties than S.

Spearman's rank correlation. In order to measure whether all three indices sort the trees according to their balance in the same way or not, we have computed the Spearman's rank correlation coefficient [24] of the indices on all trees in TreeBASE, as well as grouping them by their number of leaves n.

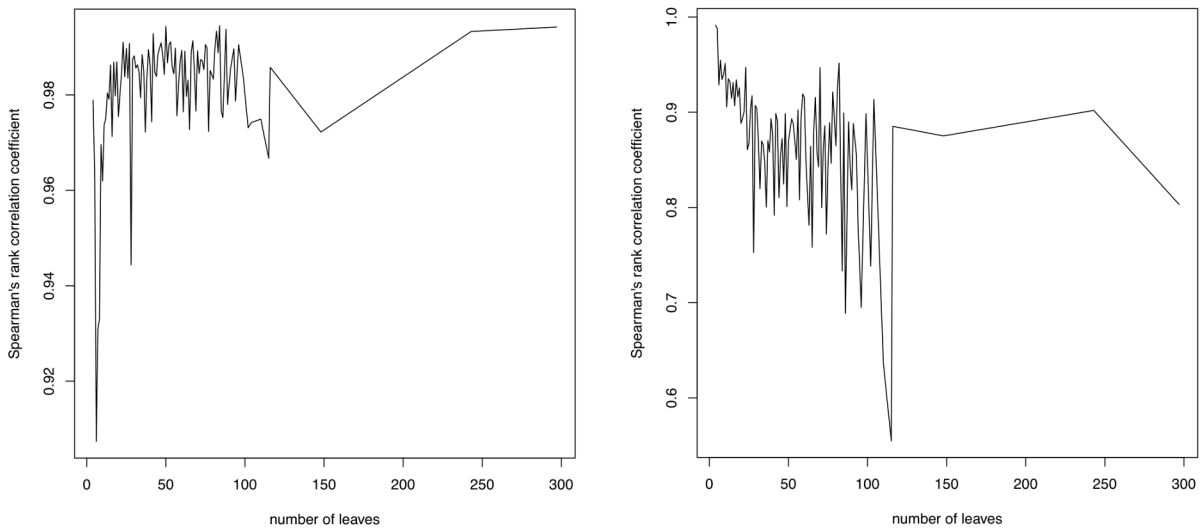

Fig 14. Spearman's rank correlation coefficient of $\mathfrak{C}$ and S (left) and of $\mathfrak{C}$ and Φ (right) in TreeBASE, as functions of the trees' numbers of leaves n.

https://doi.org/10.1371/journal.pone.0203401.g014

The global Spearman's rank correlation coefficient of $\mathfrak{C}$ and S is 0.9765, and that of $\mathfrak{C}$ and Φ is 0.9619. The graphics in Fig 14 plot these coefficients as functions of n. As it can be seen, Spearman's rank correlation coefficient for $\mathfrak{C}$ and S grows with n, approaching to 1, while the coefficient for $\mathfrak{C}$ and Φ shows a decreasing tendency with n.

Does TreeBASE fit the uniform model or the alpha-gamma model?

In this subsection, we test whether the distribution of the Colless-like index of the phylogenetic trees in TreeBASE agrees with its theoretical distribution under either the uniform model for multifurcating phylogenetic trees [25] or the α-γ-model [10] for some parameters α, γ. To do it, we use the normalized version $\mathfrak{C}_{norm}$ of $\mathfrak{C}$, which can be used simultaneously on trees with different numbers of leaves.

To estimate the theoretical distribution of this index under the two aforementioned theoretical models, for every $n = 3, \ldots, 50$ we have generated, on the one hand, 10,000 random phylogenetic trees in $\mathcal{T}_n$ under the uniform model using the algorithm described in [25], and, on the other hand, 5000 random phylogenetic trees in $\mathcal{T}_n$ under the α-γ-model for every pair of parameters $(\alpha, \gamma) \in \{0, 0.1, 0.2, \ldots, 0.9, 1\}^2$ with $\gamma \leq \alpha$. We have computed the value of $\mathfrak{C}_{norm}$ on all these trees, and we have used the distribution of these values as an estimation of the corresponding theoretical distribution. To test whether the distribution of the normalized Colless-like index on TreeBASE (or on some subset of it: see below) fits one of these theoretical distributions, we have performed two non-parametric statistical tests on the observed set of indices of TreeBASE and the corresponding simulated set of indices: Pearson's chi-squared test and the Kolmogorov-Smirnov test, using bootstrapping techniques in the latter to avoid problems with ties.

As a first approach, we have performed these tests on the whole set of trees in TreeBASE. The p-values obtained in all tests, be it for the uniform model or for any considered pair (α, γ), have turned out to be negligible. Then, we conclude confidently that the distribution of the normalized Colless-like index on TreeBASE does not fit either the uniform model or any α-γ-model when we round α, γ to one decimal place. For instance, Fig 15 displays the distribution

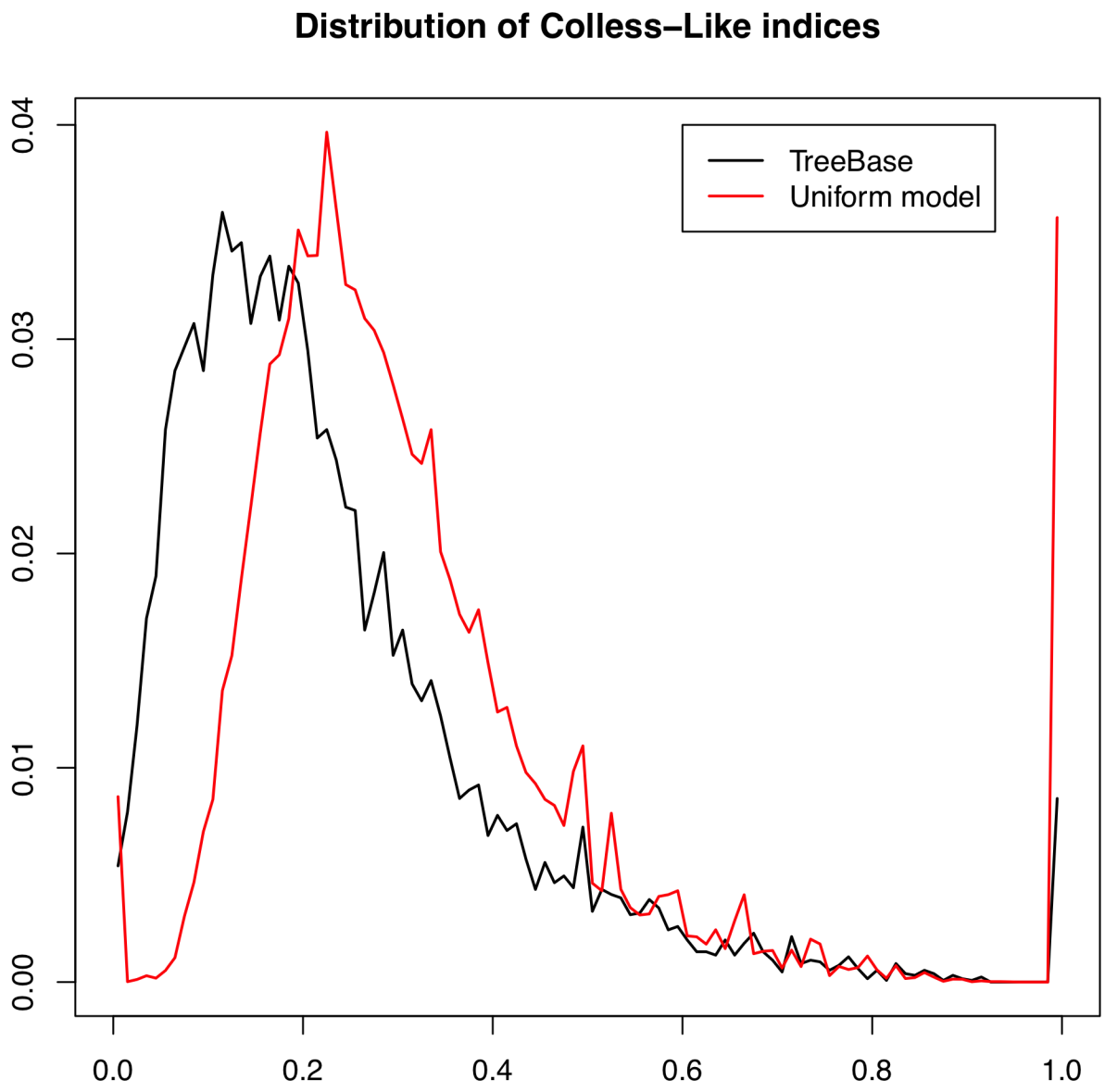

Fig 15. The distribution of $\mathfrak{C}_{norm}$ on all trees in TreeBASE (black line) and its estimated theoretical distribution under the uniform model (red line).

https://doi.org/10.1371/journal.pone.0203401.g015

of $\mathfrak{C}_{norm}$ on TreeBASE and its estimated theoretical distribution under the uniform model. As it can be seen, these distributions are quite different, which confirms the conclusion of the statistical test.

Fig 16 displays the distribution of $\mathfrak{C}_{norm}$ for all trees in TreeBASE and its estimated theoretical distribution under the α-γ-model for the pair of parameters α, γ that gave the largest p-values in the goodness of fit tests, which are $\alpha = 0.7$ and $\gamma = 0.4$. Although graphically both distributions are quite similar, the p-values of the Pearson chi-squared test and of the Kolmogorov-Smirnov test are virtually zero. One might think that the high "peaks" of the theoretical distribution near 0 and 1 could have influenced the outcome of these statistical tests. For this

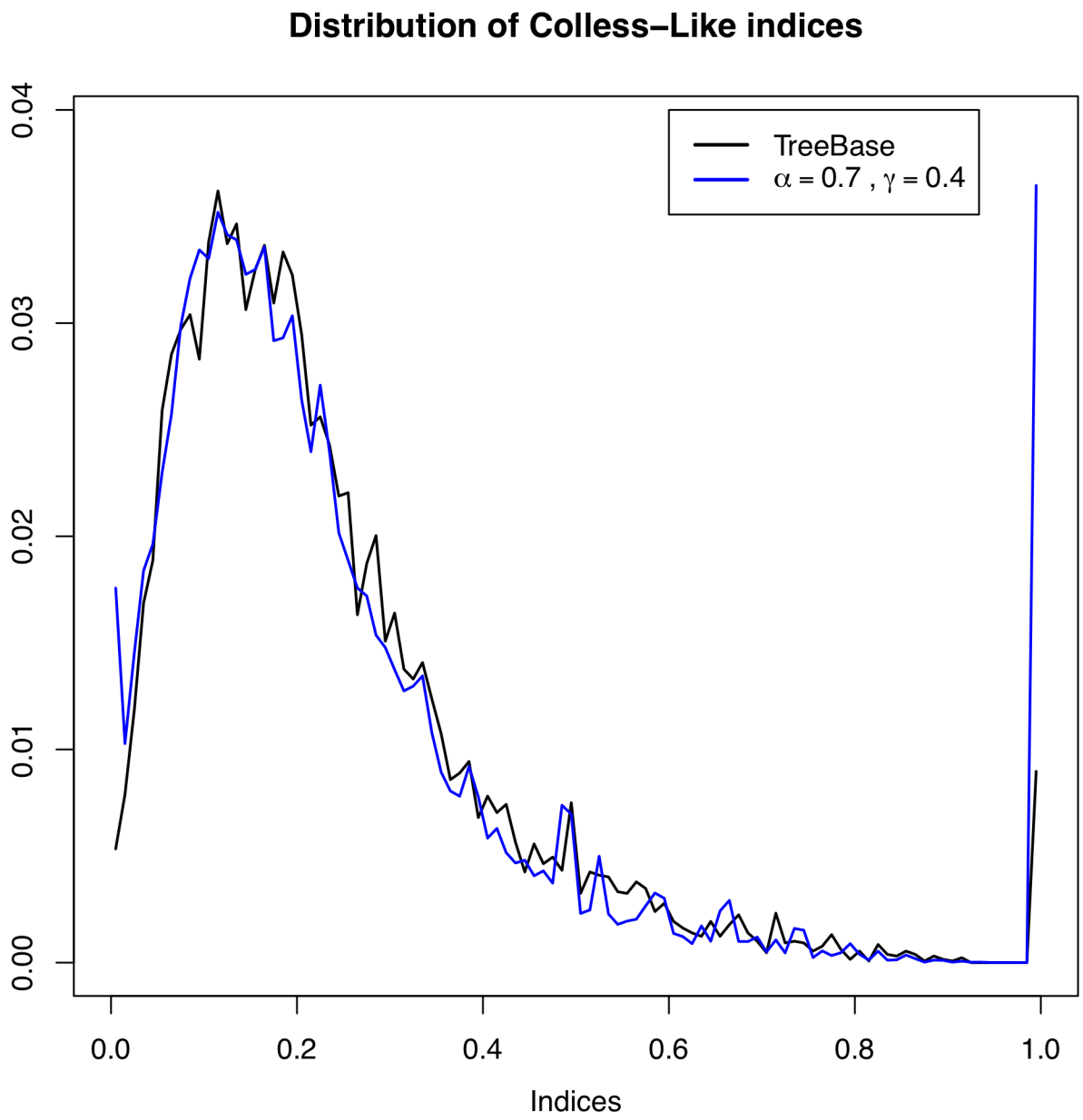

Fig 16. The distribution of $\mathfrak{C}_{norm}$ on all trees in TreeBASE (black line) and its estimated theoretical distribution under the α-γ-model with $\alpha = 0.7$ and $\gamma = 0.4$ (blue line).

https://doi.org/10.1371/journal.pone.0203401.g016

reason, we have repeated them without taking into account these "extreme" values, and the results have been the same.

Since TreeBASE gathers phylogenetic trees of different types and from different sources, we have also considered subsets of it defined by means of attributes. More specifically, besides the whole TreeBASE as explained above, we have also considered the following subsets of it:

- All trees in TreeBASE up to repetitions: we have removed 513 repeated trees (which represent about a 4% of the total).

- All trees with their `kind` attribute equal to "Species". This `kind` attribute can take three values: "Barcode tree", "Gene Tree" and "Species Tree".
- All trees with their `kind` attribute equal to "Species" and their `type` attribute equal to "Consensus". This `type` attribute can take two values: "Consensus" and "Single".
- All trees with their `kind` attribute equal to "Species" and their `type` attribute equal to "Single".

We have repeated the study explained above for these four subsets of TreeBASE, comparing the distribution of the normalized Colless-like indices of their trees with the estimated theoretical distributions by means of goodness-of-fit tests, and the results have been the same, that is, all p-values have also turned out to be negligible. Our conclusion is, then, that neither the whole TreeBASE nor any of these four subsets of it seem to fit either the uniform model or some α-γ-model.

Conclusions

In this paper we have introduced a family of *Colless-like* balance indices $\mathfrak{C}_{D,f}$, which depend on a dissimilarity D and a function $f : \mathbb{N} \to \mathbb{R}_{\geq 0}$, that generalize the Colless index to multifurcating phylogenetic trees. We have proved that every combination of a dissimilarity D and a function either $f(n) = \ln(n + e)$ or $f(n) = e^n$, defines a Colless-like index that is *sound* in the sense that the maximally balanced trees according to it are exactly the fully symmetric ones. But, the growth of the function f determines strongly which are the most unbalanced trees according to $\mathfrak{C}_{D,f}$, and hence it has influence on the very notion of "balance" measured by the index.

In our opinion, choosing $\ln(n + e)$ instead of e^n seems a more sensible decision, because, on the one hand, the most unbalanced trees according to the former are the expected ones—the combs—and, on the other hand, we have encountered several hard numeric problems when working with the extremely large figures that appear when using e^n-sizes on trees with internal nodes of high degree. With respect to the choice of the dissimilarity D, MDM and *sd* define indices that are proportional to the Colless index when applied to bifurcating trees. From these two options, we recommend to use MDM because it only involves linear operations, and hence it has less numerical precision problems than *sd*, that uses a square root of a sum of squares. This is the reason we have stuck to $\mathfrak{C}_{\mathrm{MDM},\ln(n+e)}$ in the numerical experiments reported in the Results section.

To end this paper, we would like to call the reader's attention on the problem posed in the subsection "Sound Colless-like indices:" to find functions f such that $\mathfrak{C}_{D,f}$ is sound. Our conjecture is that there is no function $f : \mathbb{N} \to \mathbb{N}$ taking values in the set of natural numbers that satisfies this property.

Supporting information

S1 File. Proofs of Theorems 18 and 19. The file provides the detailed proofs of Theorems 18 and 19.
(PDF)

S2 File. Tables. The file provides two tables, quoted in the main text, with the values of several Colless-like indices on $\mathcal{T}_n^*$ for $n = 2, 3, 4, 5$.
(PDF)

Acknowledgments

This research has been partially supported by the *Obra Social la Caixa* through the "Programa Pont La Caixa per a grups de recerca de la UIB" and by the Spanish Ministry of Economy and Competitiveness and European Regional Development Fund through project DPI2015-67082-P (MINECO/FEDER). There was no additional external funding received for this study. We thank K. Bartoszek and J. Miró-Juliá for several useful suggestions and A. Saldaña Plomer for making available to us his Java script that generates random phylogenetic trees with a fixed number of leaves with uniform distribution.

Author Contributions

Conceptualization: Arnau Mir, Lucía Rotger, Francesc Rosselló.

Data curation: Lucía Rotger.

Formal analysis: Arnau Mir, Lucía Rotger, Francesc Rosselló.

Funding acquisition: Francesc Rosselló.

Investigation: Arnau Mir, Lucía Rotger, Francesc Rosselló.

Methodology: Arnau Mir, Lucía Rotger, Francesc Rosselló.

Software: Lucía Rotger.

Writing – original draft: Francesc Rosselló.

Writing – review & editing: Arnau Mir, Lucía Rotger, Francesc Rosselló.

References

1. Mooers A, Heard S. Inferring evolutionary process from phylogenetic tree shape. Q Rev Biol 1997; 72: 31–54. https://doi.org/10.1086/419657
2. Felsenstein J. Inferring Phylogenies. Sinauer Associates Inc.; 2004.
3. Colless D. Review of "Phylogenetics: the theory and practice of phylogenetic systematics". Sys Zool 1982; 31: 100–104. https://doi.org/10.2307/2413420
4. Kirkpatrick M, Slatkin M. Searching for evolutionary patterns in the shape of a phylogenetic tree. Evolution; 1993; 47: 1171–1181. https://doi.org/10.2307/2409983
5. Sackin M. "Good" and "bad" phenograms. Sys Zool 1972; 21: 225–226. https://doi.org/10.2307/2412292
6. Shao K, Sokal R. Tree balance. Sys Zool 1990; 39: 226–276. https://doi.org/10.2307/2992186
7. McKenzie A. Distributions of cherries for two models of trees. Math Biosci 2000; 164: 81–92. https://doi.org/10.1016/S0025-5564(99)00060-7 PMID: 10704639
8. Mir A, Rosselló F, Rotger L. A new balance index for phylogenetic trees. Math Biosci 2013; 241: 125–136. https://doi.org/10.1016/j.mbs.2012.10.005 PMID: 23142312
9. Coronado T, Mir A, Rosselló F, Valiente G. A balance index for phylogenetic trees based on quartets. J Math Biol, forthcoming 2018. Available from: arXiv:1803.01651.
10. Chen B, Ford D, Winkel M. A new family of Markov branching trees: the alpha-gamma model. Electron J Probab 2009; 14: 400–430. https://doi.org/10.1214/EJP.v14-616
11. Sanderson M, Donoghue M, Piel W, Eriksson T. TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. Am J Botany 1994; 81: 183. Available from: https://treebase.org.
12. Xiang Y, Zhu Z, Li Y. Enumerating unlabeled and root labeled trees for causal model acquisition. In: Advances in Artificial Intelligence: Springer; 2009; 158–170.
13. The On-Line Encyclopedia of Integer Sequences; 2010. Available from: http://oeis.org/.
14. Rogers J. Response of Colless's tree imbalance to number of terminal taxa. Sys Biol 1993; 42: 102–105. https://doi.org/10.1093/sysbio/42.1.102

15. Matsen F. Optimization Over a Class of Tree Shape Statistics. IEEE/ACM Trans Comput Biol Bioinform 2007; 4: 506–512. https://doi.org/10.1109/tcbb.2007.1020 PMID: 17666770
16. Yan X, Tang T, Deng Y, Du J, Yang X. Evaluation of transcendental functions on Imagine architecture. In: International Conference on Parallel Processing 2007: IEEE Press; 2007; 53–53.
17. Ford D. Probabilities on cladograms: Introduction to the alpha model. 2005. Preprint. Available from: arXiv:math/0511246.
18. Harding E. The probabilities of rooted tree-shapes generated by random bifurcation. Adv Appl Prob 1971; 3: 44–77. https://doi.org/10.2307/1426329
19. Yule G. A mathematical theory of evolution based on the conclusions of Dr J. C. Willis. Phil Trans Royal Soc (London) Series B 1924; 213: 21–87. https://doi.org/10.1098/rstb.1925.0002
20. Cavalli-Sforza L, Edwards A. Phylogenetic analysis. Models and estimation procedures. Am J Hum Genet 1967; 19: 233–257. PMID: 6026583
21. Pinelis I. Evolutionary models of phylogenetic trees. Roy Soc Lond Proc Ser Biol Sci 2003; 270: 1425–1431. https://doi.org/10.1098/rspb.2003.2374
22. Boettiger C, Temple Lang D. Treebase: an R package for discovery, access and manipulation of online phylogenies. Methods Ecol Evol 2012; 3: 1060–1066. https://doi.org/10.1111/j.2041-210X.2012.00247.x
23. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2008.
24. Spearman C. The proof and measurement of association between two things. Am J Psychol 1904; 15: 72–101. https://doi.org/10.2307/1412159
25. Oden N, Shao K. An algorithm to equiprobably generate all directed trees with *k* labeled terminal nodes and unlabeled interior nodes. Bull Math Biol 1984; 46: 379–387 PMID: 6547358