Systematic Biology (2022), **0**, 0, *pp.* 1–37
doi:10.1093/sysbio/LemantEtAl˙WithFigures

Robust, Universal Tree Balance Indices

Jeanne Lemant[1,2,3], Cécile Le Sueur[1], Veselin Manojlović[4], and Robert Noble[1,4,*]

[1] *Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland*
[2] *Current address: Swiss Tropical and Public Health Institute, Allschwil, Switzerland*
[3] *Current address: University of Basel, Basel, Switzerland*
[4] *Department of Mathematics, City, University of London, London, UK*

**robert.noble@city.ac.uk*

Abstract

Balance indices that quantify the symmetry of branching events and the compactness of trees are widely used to compare evolutionary processes or tree-generating algorithms. Yet existing indices are not defined for all rooted trees, are unreliable for comparing trees with different numbers of leaves, and are sensitive to the presence or absence of rare types. The contributions of this article are twofold. First, we define a new class of robust, universal tree balance indices. These indices take a form similar to Colless' index but can account for population sizes, are defined for trees with any degree distribution, and enable meaningful comparison of trees with different numbers of leaves. Second, we show that for bifurcating and all other full m-ary cladograms (in which every internal node has the same out-degree), one such Colless-like index is equivalent to the normalised reciprocal of Sackin's index. Hence we both unify and generalise the two most popular existing tree balance indices. Our indices are intrinsically normalised and can be computed in linear time. We conclude that these more widely applicable indices have potential to supersede those in current use.

Key words: tree balance, Sackin index, Colless index, cancer, species tree, clone tree

Tree balance indices – most notably those credited to Sackin (1972) and Colless

(1982) – are widely used to describe speciation processes, compare cladograms, and assert the correctness of tree reconstruction methods (Shao and Sokal, 1990; Mooers and Heard, 1997; Fischer et al., 2021). Existing tree balance indices have several important flaws. First, they cannot be applied to any tree in which any node has only one descendant. Second, existing indices are unreliable for comparing trees with different numbers of leaves. Third, because they do not account for population sizes, these indices are sensitive to the omission or inclusion of rare types. The latter issue is, for example, a problem in oncology (Chkhaidze et al., 2019; Scott et al., 2020), where methods for determining and classifying evolutionary modes have clinical value (Maley et al., 2017; Davis et al., 2017).

Here we develop a new class of robust, universal tree balance indices. Our definitions not only extend the tree balance concept and open up new applications but also unify the two main approaches to quantifying balance as proposed by Sackin and Colless. We describe several general advantages of our indices compared to those in current use.

Materials and Methods

Rooted trees

We consider exclusively rooted trees in which all edges are oriented away from the root (which will be topmost in our figures). This orientation defines a natural order on the tree, from top to bottom: edges descend from the root to the other *internal nodes* and finally to the terminal nodes or *leaves*. The *out-degree* of a node i, written $d^+(i)$, is the number of direct descendants, ignoring any subtrees in which all nodes have zero size. Internal nodes have out-degree at least one, whereas leaves have out-degree zero. If all internal nodes have out-degree 1 then the tree is called *linear*. If all internal nodes have out-degree $m > 1$ then the tree is a *full m-ary tree*, and if $m = 2$ then it is also called *bifurcating* (such as Figs. 1a and 1b).

Some other tree topologies have particular names. A *caterpillar tree* (Fig. 1a) is a

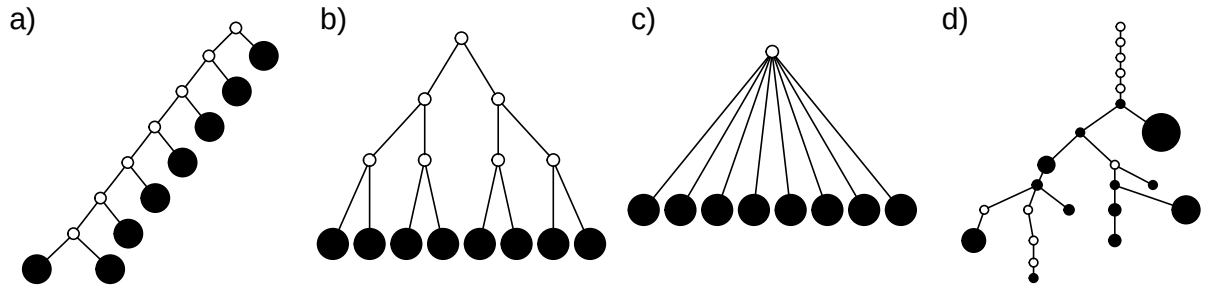

Fig. 1. Contrasting trees. **a**: Caterpillar tree with $I_S = 35$, $I_{S,norm} = 1$, $I_C = 21$, $I_{C,norm} = 1$, $I_\Phi = 56$, $I_{\Phi,norm} = 1$. **b**: Fully symmetric bifurcating tree with $I_S = 24$, $I_{S,norm} \approx 0.59$, $I_C = I_{C,norm} = 0$, $I_\Phi = 16$, $I_{\Phi,norm} \approx 0.29$. **c**: Star tree with $I_S = 8$, $I_{S,norm} = 0$, I_C and $I_{C,norm}$ undefined, $I_\Phi = I_{\Phi,norm} = 0$. **d**: Clone tree of the lung tumour CRUK0065 in the TRACERx cohort (Jamal-Hanjani et al., 2017). In the clone tree, nodes represented by empty circles correspond to extinct clones, and the diameters of other nodes are proportional to the corresponding clone population sizes.

bifurcating tree in which every internal node except one has exactly one leaf. A *fully symmetric* tree (Fig. 1b) is such that every internal node with the same depth has the same degree or, equivalently, for each internal node i all the subtrees rooted at i are identical. A *star tree* (Fig. 1c) is a tree whose leaves are all attached to the root, which is the only internal node.

Node sizes, tree magnitudes, and leafy trees

Although our definitions can be applied in other contexts, we will assume that nodes correspond to biological taxa or clones, and on this basis we assign non-negative *node sizes*. If we know (or care) only whether each type is extant or extinct – as is typical in taxonomy – then we assign size zero to every node representing an extinct type, and size one otherwise. If nodes represent clones with known population sizes – as is often the case in studies of cancer and microbial evolution – then each node size is equal to the population size of the corresponding clone. The *magnitude* of a tree or subtree is then defined as the sum of its node sizes (we use magnitude here because a tree's size is conventionally defined as its number of nodes). We define a *leafy tree* as a rooted tree in which all internal nodes have size zero.

Cladograms, taxon trees and clone trees

Tree types can also be defined in terms of what they represent. Following Podani (2013), we distinguish between two representations used in systematic biology.

We define a *cladogram* as a rooted tree in which internal nodes represent hypothetical extinct ancestors, leaves represent extant biological taxa, and edges represent evolutionary relationships. This is equivalent to the synchronous cladogram definition of Podani (2013). Every cladogram is by definition a leafy tree, with magnitude equal to its number of leaves. A common conception is that only bifurcating cladograms can be considered fully resolved. However, the linear two-node cladogram is appropriate for representing serial anagenesis (in which each descendant replaces its ancestor), while budding (in which an ancestor produces a descendant and remains extant) can give rise to cladogram nodes with out-degree greater than two (Podani, 2013). Hence there is no restriction on cladogram node degrees. An extant ancestor is represented in a cladogram by a leaf stemming from the internal ancestor node, in which case, as Podani notes, "an ancestor is identical to an extant taxon connected directly to it".

Alternatively, extant or known ancestors may be represented uniquely by internal nodes (like in a genealogy with overlapping generations). Such diagrams are known to organismal biologists as species trees or taxon trees, and to oncologists as clone trees. We define a *taxon tree* as a rooted tree in which all nodes represent biological taxa, and edges represent ancestor-descendant relationships. Similarly, a clone tree is defined as a rooted tree in which each node represents a clone (a set of cells that share alterations of interest due to common descent), and edges represent the chronology of alterations. Both taxon tree and clone tree fit the achronous tree definition of Podani (2013). Clone tree nodes can have any out-degree, including $d^+ = 1$, and each node – including internal nodes – can be associated with a non-negative size, as illustrated in Figure 1d.

When nodes are associated with sizes, the addition of subtrees comprising even vanishingly small nodes can change leaves into internal nodes and so substantially change

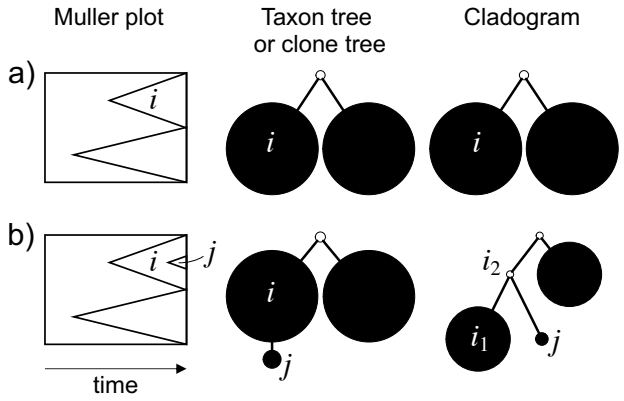

Fig. 2. Muller plots (left column), taxon or clone trees (middle column), and cladograms (right column) representing evolution by splitting only (a) and both splitting and budding (b). In a Muller plot, polygons represent proportional subpopulation sizes (vertical axis) over time (horizontal axis), and each descendant is shown emerging from its parent polygon. In the trees, nodes represented by empty circles correspond to extinct types.

the value of existing tree balance indices. This behaviour is unsatisfactory because relatively small nodes typically represent either newly-created types that have yet to experience evolutionary forces or types on the verge of extinction, and in either case convey negligible information about the mode of evolution. Data sets may also omit rare types due to sampling error or because genetic sequencing methods have imperfect sensitivity (Turajlic et al., 2018).

The change due to the addition of terminal nodes is greater when the tree is a cladogram rather than a taxon or clone tree. For example, when a three-node, two-leaf tree (Fig. 2a) is augmented by adding a node j to a leaf i (Fig. 2b), the three original nodes retain their positions in the clone tree (middle column of Figure 2), but in the cladogram (right column) node i becomes two nodes (i_1 and i_2), the larger of which is now further from the root (see Podani (2013) for further illustrations of this difference). As the size of the new node j is continuously reduced to zero, the clone tree changes continuously, whereas the cladogram undergoes an abrupt change of topology when the size of node j reaches zero. We conclude that the taxon tree or clone tree representation is more robust than the cladogram representation in the general case in which nodes are associated with sizes and ancestors can be extant. Also an index that accounts for non-zero internal node

sizes can be made more robust than one that does not. Accordingly, we will define indices for the more general domain of clone trees, and then obtain results for cladograms as a special case.

Existing tree balance indices

The most widely used tree balance indices are in fact imbalance indices, such that more balanced trees are assigned smaller values. These indices were introduced to study cladograms; they take no account of node size, and, even after applying standard normalisations, they are appropriate only for comparing trees with equal numbers of leaves. The most popular are Sackin's index and Colless' index.

Sackin's index.– Let T be a tree with set of leaves $L(T)$. For a leaf $l \in L(T)$, let ν_l be the number of internal nodes between l and the root, which is included in the count. Then the index credited to Sackin (1972) is

$$I_S(T) = \sum_{l \in L(T)} \nu_l.$$

For two bifurcating trees on the same number of leaves, a less balanced tree has higher values of ν as the tree is in a sense less compact (compare trees **a** and **b** in Figure 1).

Since the value tends to increase with the number of nodes, Shao and Sokal (1990) proposed normalising I_S with respect to trees on $n > 2$ leaves by subtracting its minimum possible value for such trees and then dividing by the difference between the maximum and minimum possible values. The minimal I_S is reached on the star tree, such as tree **c** in Figure 1, and hence $\min_n(I_S) = n$. The maximum is attained on the caterpillar tree, such as tree **a**:

$$\max_n(I_S) = n - 1 + \sum_{\nu=1}^{n-1} \nu = n - 1 + n(n-1)/2 = (n-1)(n+2)/2.$$

The normalised index is then

$$I_{S,norm}(T) = \frac{I_S(T) - n}{(n+2)(n-1)/2 - n}.$$

This normalised index is not very satisfactory as a balance index because it fails to capture an intuitive notion of balance. For example, it is not obvious why fully symmetric tree **b** should be considered less balanced than star tree **c** in Figure 1, yet its $I_{S,norm}$ value is much larger. To address this issue, Shao and Sokal (1990) further suggested normalising I_S relative to its extremal values among trees with the same number of internal nodes as well as the same number of leaves. But even then the index remains unreliable for comparing trees with different numbers of leaves. For example, the index is 1 for every caterpillar tree, yet long caterpillar trees are intuitively less balanced than short ones. The conventional I_S normalisations are not defined for trees containing linear parts. Moreover, since I_S doesn't account for node size, it is sensitive to the addition or removal of subtrees comprising relatively small nodes.

Colless' index.– For an internal node i of a bifurcating tree T, define n_{i_1} as the number of leaves of the left branch of the subtree rooted at i, and n_{i_2} as the number of leaves of the right branch. Then the index defined by Colless (1982) is

$$I_C(T) = \sum_{i \in \widetilde{V}(T)} |n_{i_1} - n_{i_2}|,$$

where $\widetilde{V}(T)$ is the set of all internal nodes of T. The index can be normalised for the set of trees on $n > 2$ leaves by dividing by its maximal value, $\binom{n-1}{2}$, which is reached on the caterpillar tree (as in Figure 1a).

Because Colless' index cannot be applied to multifurcating trees, Mir et al. (2018) recently introduced a family of Colless-like balance indices, including I_C as a special case. Each of these indices $C_{D,f}$ is determined by a weight function f, which assigns a size to each subtree as a function of its out-degree, and a dissimilarity function D. By definition of D, Colless-like indices are zero if and only if each internal node divides its descendants into subtrees of equal size. But since these indices are normalised by dividing by the maximal value for trees on the same number of leaves, they are unreliable for comparing trees with different numbers of leaves. In common with Sackin's index, the total cophenetic index I_Φ

(Mir et al., 2013) (see Appendix), and other existing indices (surveyed by Fischer et al. (2021)), the Colless-like indices so far defined do not account for node sizes and can be applied only to trees in which all nodes have out-degree greater than one.

Desirable properties of a universal, robust tree balance index

Our aim is to derive a tree balance index J that is useful for classifying and comparing rooted trees that can have any distributions of node degrees and node sizes. Here we specify four desirable properties that such an index should have. The first two axioms relate to extrema. We will call an index *universal* if it is defined for trees with any degree distribution and obeys these first two axioms. An index that conforms to the other three axioms – which are relevant only when nodes can have arbitrary sizes – will be called *robust*.

We will begin by introducing some additional notation (see also Table 1). For a tree T, we will use $V(T)$ to denote the set of all nodes of T, which we will abbreviate to V when the identity of the tree is unambiguous. Let $f(v) \geqslant 0$ denote the size of node v. Then T_i denotes the subtree rooted at node i (that is, the subtree that contains node i and all its descendants); S_i is the magnitude of T_i; and S_i^* is the magnitude of T_i excluding its root:

$$S_i \coloneqq \sum_{v \in V(T_i)} f(v); \quad S_i^* \coloneqq \sum_{\substack{v \in V(T_i) \\ v \neq i}} f(v) = S_i - f(i).$$

We will use $\widetilde{V}(T)$ or simply $\widetilde{V}$ to denote the set of all internal nodes such that $\{i \in \widetilde{V}\} \coloneqq \{i \in V : S_i^* > 0\}$.

Conventionally, a tree is considered maximally balanced only if every internal node splits its descendants into subtrees on the same number of leaves (Shao and Sokal, 1990). We generalise this concept by requiring that every internal node splits its descendants into at least two subtrees of equal magnitude, as in Figure 3a. We call this the *equal splits* property, and we make it a necessary and sufficient condition for maximal balance.

Properties of a node i	
$d^+(i)$	Out-degree
$C(i)$	Set of children
$\nu(i)$	Depth
$f(i)$	Size
T_i	Subtree rooted at i
n_i	Number of leaves of T_i
S_i	Magnitude of T_i (sum of node sizes)
S_i^*	Magnitude of T_i excluding its root
g_i	Importance factor
p_{ij}	S_j/S_i^*, where $j \in C(i)$
W_i	Balance score
W_i^q	Balance score based on ${}^qH$
h_i	Non-root dominance factor
Sets of nodes	
V	All nodes
$\widetilde{V}$	Internal nodes i such that $S_i^* > 0$
L	Leaves
Entropies and tree balance indices	
${}^qH$	Generalised entropy with parameter q
${}^1H_b$	Shannon entropy with base b
I_S	Sackin's index
I_C	Colless' index
I_Φ	Total cophenetic index
$C_{D,f}$	Colless-like index
$I_{S,gen}$	Generalised Sackin's index
$I_{C,gen}$	Generalised Colless' index
J^q	Tree balance index based on ${}^qH$
J_S	Normalised inverse Sackin index
J^{1c}	A conservative tree balance index

Table 1. Notation used throughout this paper.

Axiom 0.1 (Maximum value) $J(T) \leqslant 1$ for all trees T, and $J(T) = 1$ if and only if T has equal splits.

Another convention is that trees with relatively many internal nodes are considered highly imbalanced. According to this convention, linear trees (that is, trees in which every node i has $d^+(i) \leqslant 1$, as in Figure 3b) should be considered even less balanced than caterpillar trees. Also, given that balance implies branching, the most imbalanced split is

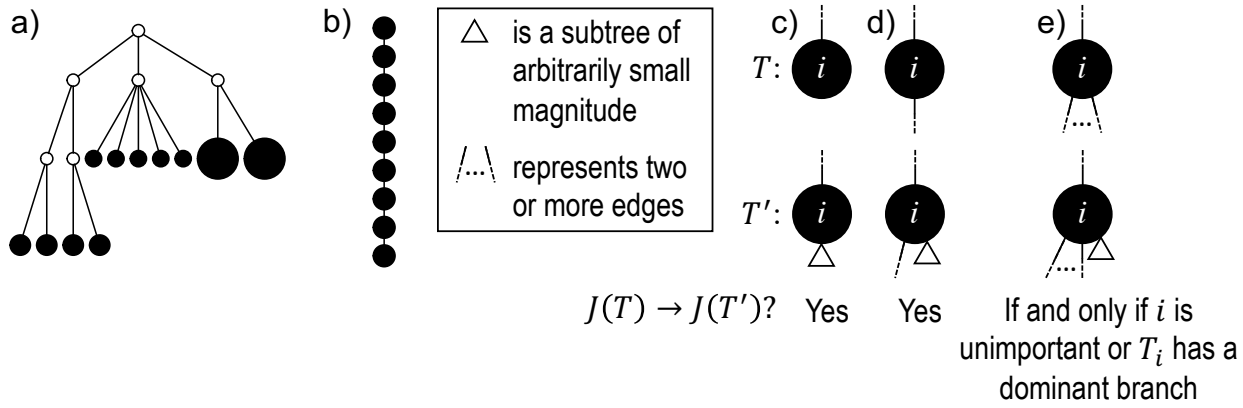

Fig. 3. **a**: A tree in which each internal node has null size and splits its descendants into subtrees of equal magnitude, and hence $J = 1$. This tree can be considered balanced only according to an index that accounts for node size. **b**: A linear tree, for which $J = 0$. **c-e**: A robust, universal tree balance index J is insensitive to the addition of a subtree of arbitrarily small magnitude if it is added to a leaf (a) or a non-root node with out-degree 1 (b), but not necessarily if the subtree is added to a non-root node with greater out-degree (c).

one that assigns all descendants to one branch and none to any other branches. Hence our second desirable property:

Axiom 0.2 (Minimum value) $J(T) \geqslant 0$ for all trees T, and $J(T) = 0$ if and only if T is a linear tree.

Our third desirable property ensures that our index is insensitive to the properties of nodes that have relatively few descendants.

Axiom 0.3 (Insensitivity) Let T be a tree and l be one of its leaves. If we create a new tree T' from T by adding a subtree with finitely many nodes rooted at l then $J(T') \to J(T)$ as $S_l^* / \sum_{j \in \widetilde{V}(T')} S_j^* \to 0$.

Our fourth axiom ensures that a linear section of a tree is regarded as a maximally unequal split.

Axiom 0.4 (Linear limit) Let T be a tree and $i \in \widetilde{V}(T)$ with $d^+(i) = 1$. Let i_1 be the unique child of i. If we create a new tree T' from T by adding additional subtrees with finitely many nodes rooted at i then $J(T') \to J(T)$ as $S_{i_1}/S_i^* \to 1$.

Lastly, we require continuity with respect to varying node size:

Axiom 0.5 (Continuity) Suppose we create a new tree T' by selecting a node of tree T and changing the node's size from x to x'. Then $J(T') \to J(T)$ as $x' \to x$.

Alternative axioms are considered in the Appendix.

Sensitivity to changes in out-degree of non-root nodes

By design, our definition of a robust tree balance index does not require insensitivity to the addition or removal of rare types in all cases. To see why, suppose we transform a tree T into T' by adding one or more subtrees of arbitrarily small magnitude, attached to a non-root node $i \in V(T)$. As illustrated in Figure 3c-e, there are three topologically distinct cases to consider. If i is a leaf of T (Fig. 3c) or $d^+(i) = 1$ in T (Fig. 3d) then $J(T') \to J(T)$ due to Axiom 0.3 or Axiom 0.4. In the first case, i is an *unimportant* node, which we define to mean that $S_i^* / \sum_{j \in \widetilde{V}} S_j^* \to 0$. In the second case, if i is not an unimportant node in T then T_i must have a *dominant branch*, meaning that i has a child i_1 such that $S_{i_1}/S_i^* \to 0$. The third case, when $d^+(i) \geqslant 2$ in T (Fig. 3e), is more complicated. If i is an unimportant node in T then $J(T') \to J(T)$ as $S_i^* / \sum_{j \in \widetilde{V}} S_j^* \to 0$ in T', by Axiom 0.3. If T_i in T has a dominant branch T_{i_1} in T then $J(T') \to J(T)$ as $S_{i_1}/S_i^* \to 1$ in T', by Axiom 0.4. But if neither of those conditions hold then our axioms do not specify the size of the effect on J.

Although we could modify Axiom 0.4 so that J is always insensitive to the addition of relatively low-magnitude subtrees – thus increasing the index's robustness – we argue that this would undermine its utility as a tree balance index. The balance of a node can be conventionally defined as the extent to which it splits its descendants into multiple subtrees of equal magnitude. By this definition, the attachment of a new, relatively low-magnitude subtree to a perfectly balanced node will create imbalance even as – in fact especially as – the magnitude of this new subtree, relative to the magnitude of the node's pre-existing descendants, approaches zero. Therefore it is desirable for a tree balance index

to be sensitive to certain changes of node degree, such that in the third scenario considered above, $J(T') \to J(T)$ if and only if i is an unimportant node or T_i has a dominant branch (Fig. 3e).

Results

General definition of universal, robust tree balance indices

Our general definition depends on two continuous functions of subtree magnitudes:

- An *importance* factor $g : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ with $g(x) \to 0$ as $x \to 0$;
- A *balance score* W that assigns $W_i \in [0, 1]$ to each internal node i such that $W_i = 0$ if and only if $d^+(i) = 1$, and $W_i = 1$ if and only if i splits its descendants into at least two equal-magnitude subtrees.

To allow us to define W more rigorously, let $\mathcal{S}$ denote the set of vectors with positive components that sum to unity:

$$\mathcal{S} \coloneqq \cup_{k \geqslant 1} \{(x_1, \ldots, x_k) | x_1, \ldots, x_k > 0, x_1 + \ldots + x_k = 1\}.$$

Then $W \colon \mathcal{S} \to [0, 1]$ is such that, for all $(x_1, \ldots, x_k) \in \mathcal{S}$:

- (Associativity) For every permutation π, $W(x_1, \ldots, x_k) = W(x_{\pi(1)}, \ldots, x_{\pi(k)})$;
- (Maximum value) $W(x_1, \ldots, x_k) = 1$ if and only if $k > 1$ and $x_1 = \ldots = x_k$;
- (Minimum value) $W = 0$ if and only if $\max(x_1, \ldots, x_k) = 1$;
- (Continuity) W is a continuous function with respect to each of its arguments.

We then define a balance index in terms of subtree magnitudes as

$$J \coloneqq \frac{1}{\sum_{k \in \widetilde{V}} g_k} \sum_{i \in \widetilde{V}} g_i W_i, \tag{0.1}$$

where $W_i = W(S_{i_1}/S_i^*, \ldots, S_{i_p}/S_i^*)$, $g_i = g(S_i^* / \sum_{j \in \widetilde{V}} S_j^*)$, and $i_1, \ldots, i_p$ are the children of node i (see Table 1 for a recap of notation). A short proof that this type of index satisfies

our five axioms for robustness and universality (Axioms 0.1-0.5) is presented in the Appendix.

The balance score W in Equation 0.1 measures the extent to which an internal node splits its descendants into equal-magnitude subtrees. The importance factor g assigns more weight to nodes that are the roots of large subtrees. In biological terms, this means giving more weight to types that have more descendants. Sackin's and Colless' indices similarly assign more weight to nodes that have more descendant leaves or are closer to the root. Mooers and Heard (1997) have argued that it is reasonable to put more weight on nodes deeper within the tree because "those nodes are the most informative, as the subclades they define are older and therefore sample longer periods of evolutionary time."

A specific index based on the Shannon entropy

In defining a specific index, we start by opting for the simplest importance factor function: $g(x) = x$. The role of the balance score function W is to quantify the extent to which a set of objects (specifically subtrees) have equal magnitude. A well-known index that satisfies the necessary conditions is the normalised Shannon entropy.

Assume a population is partitioned into $n \in \mathbb{N}$ types, with each type i accounting for a proportion p_i. Then the Shannon entropy with base b is defined as ${}^1H_b \coloneqq -\sum_{i=1}^{n} p_i \log_b p_i$. If all types have equal frequencies $p_i = 1/n$ then ${}^1H_b = \log_b n$. If the types have unequal sizes then ${}^1H_b < \log_b n$. And if the abundance is mostly concentrated on one type j, such that $p_j \to 1$, then ${}^1H_b \to 0$.

Let $C(i)$ denote the set of children (immediate descendants) of a node i, and for $j \in C(i)$ let $p_{ij} \coloneqq S_j/S_i^*$ denote the relative magnitude of subtree T_j compared to all subtrees attached to i.

A balance score based on the normalised Shannon entropy is then

$$W_i^1 = \sum_{j \in C(i)} W_{ij}^1, \qquad \text{with } W_{ij}^1 = \begin{cases} -p_{ij} \log_{d^+(i)} p_{ij} & \text{if } p_{ij} > 0 \text{ and } d^+(i) \geqslant 2, \\ 0 \quad \text{otherwise.} \end{cases} \tag{0.2}$$

For every internal node i, the number of frequencies p_{ij} is equal to $d^+(i)$, and if all these frequencies are equal then $-\sum_{i=1}^{n} p_{ij} \log_b p_{ij} = \log_b d^+(i)$, for any base b. Changing the base of the logarithm from b to $d^+(i)$ is equivalent to dividing the sum by $\log_b d^+(i)$, which implies that $-\sum_{i=1}^{n} p_{ij} \log_{d^+(i)} p_{ij} = 1$ when all the p_{ij} are equal. From aforementioned properties of the Shannon entropy, it then follows that $W_i^1 \in [0, 1]$, with $W_i^1 = 0$ if and only if $d^+(i) = 1$, and $W_i^1 = 1$ if and only if i splits its descendants into at least two equal-magnitude subtrees. Therefore the following specific balance index satisfies our robustness and universality axioms:

$$J^1 \coloneqq \frac{1}{\sum_{k \in \widetilde{V}} S_k^*} \sum_{i \in \widetilde{V}} S_i^* W_i^1. \tag{0.3}$$

The calculation of J^1 is illustrated in Figure 4a.

The definition simplifies when we restrict the domain to the set of multifurcating leafy trees in which all leaves have equal size f_0. This includes cladograms in which internal nodes represent extinct ancestors and leaves correspond to equally important extant types. For all internal nodes i in such trees, $S_i^* = S_i = f_0 n_i$, where n_i is the number of leaves of the subtree rooted at node i. The general definition of Equation 0.1 can then be expressed in terms of node balance scores and leaf counts:

$$J = \frac{1}{\sum_{k \in \widetilde{V}} n_k} \sum_{i \in \widetilde{V}} n_i W_i, \tag{0.4}$$

and the specific definition of Equation 0.3 becomes

$$J^1 = \frac{-1}{\sum_{k \in \widetilde{V}} n_k} \sum_{i \in \widetilde{V}} \sum_{j \in C(i)} n_j \log_{d^+(i)} \frac{n_j}{n_i}. \tag{0.5}$$

For example, Figure 4b shows the J^1 values of all leafy trees on six equally sized leaves without linear parts. Unlike Sackin's and Colless' indices, J^1 does not consider the caterpillar tree the least balanced of these trees.

There are of course many alternative options for W. For example, Colless' index can be generalised to define a robust, though not universal, tree balance index on the domain of bifurcating trees (see Appendix). Since the Shannon entropy belongs to families of

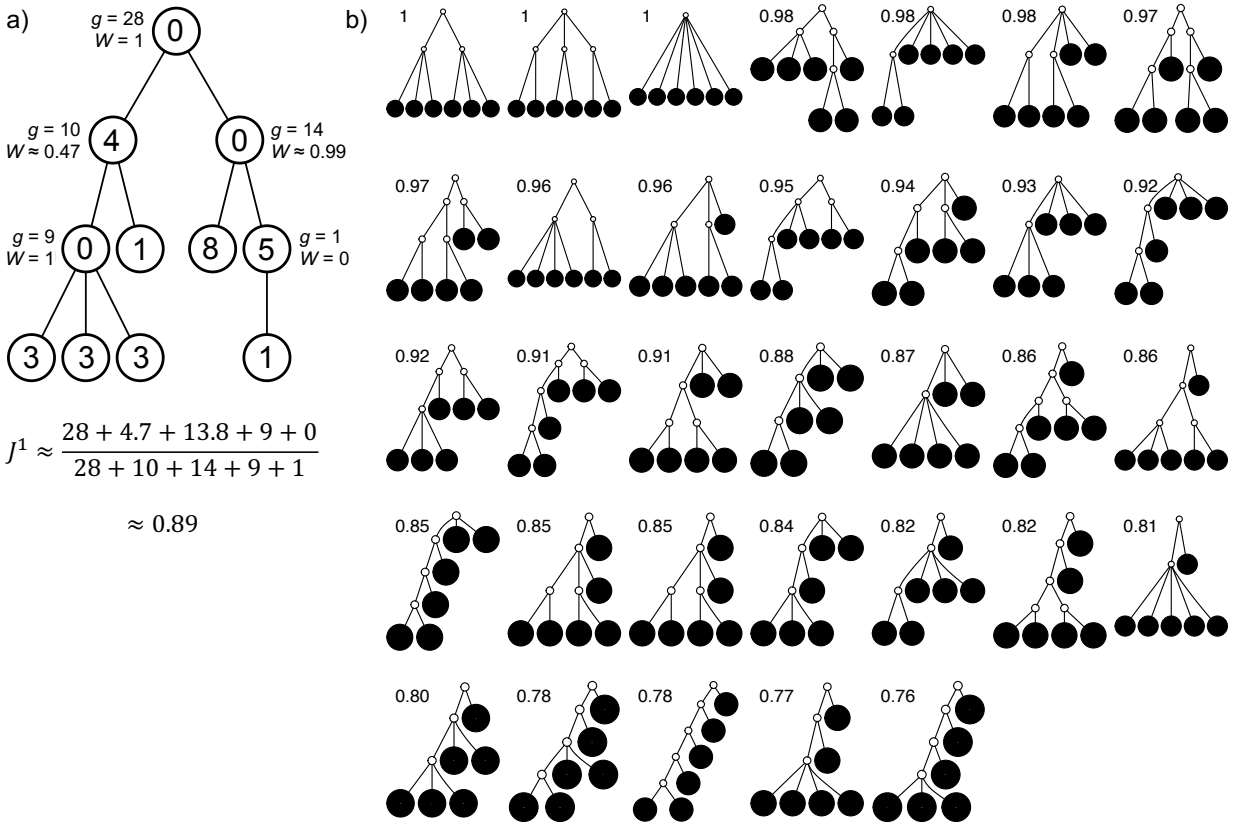

Fig. 4. **a**: An example calculation of J^1. Numbers shown inside nodes are the node sizes. **b**: All multifurcating leafy trees on six leaves without linear parts and with equally sized leaves, sorted and labelled by J^1 value.

generalised entropies (Chao et al., 2014; Rényi, 1961) parameterised by $q > 0$, the above reasoning can be generalised to define a balance score W^q, and hence a robust, universal balance index J^q, for every $q > 0$ (see Appendix). Other candidates for W include one minus the variance of the proportional subtree magnitudes, or one minus the mean deviation from the median (Mir et al., 2018). We prefer W^1 mostly because, as we shall show, it is the only function for which Equation 0.4 is a generalisation of the normalised inverse Sackin index.

Relationship with Colless' index

Like Colless' index and Colless-like indices as previously defined, our new family of tree balance indices is based on the intuitive idea of assigning a value to each internal

node, summing these values, and then normalising the sum. A Colless-like index in the sense of Mir et al. (2018) depends on a function $f : \mathbb{N} \to \mathbb{R}_{\geqslant 0}$, which assigns node sizes, and a dissimilarity score $D : \mathcal{R} \to \mathbb{R}_{\geqslant 0}$, where $\mathcal{R}$ is the set of non-null real vectors. Before normalisation, such an index has the form

$$C_{D,f} = \sum_{i \in \widetilde{V}} D(\delta_f(T_{i_1}), \ldots, \delta_f(T_{i_k})),$$

where $\{i_1, \ldots, i_k\}$ are the children of node i. The function δ_f assigns a size to each subtree by summing the node sizes: $\delta_f(T) = \sum_{j \in V(T)} f(d^+(j))$. Neglecting the initial normalising factor, our general definition (Equation 0.1) has a similar form and can be considered Colless-like in only a slightly broader sense. Our definition nevertheless differs in two important ways.

First, whereas the unbounded dissimilarity index D measures both node imbalance and importance, and is undefined for nodes with out-degree one, we split these two roles into a normalised balance score W and an unbounded importance factor g, and we assign a W value (specifically zero) to nodes with out-degree one. This difference enables us to extend the balance index definition to trees with any degree distribution. It also makes it easy to normalise our indices for any tree, simply by dividing by the sum of the importance factors. Furthermore, our normalisation is universal, rather than being based on comparison with other trees with the same number of leaves. For example, our J^q indices judge long caterpillar trees less balanced than short ones (Fig. 5a), whereas Sackin's index, Colless' index, and the total cophenetic index consider all caterpillar trees on more than two leaves equally imbalanced.

Second, instead of assigning a size to each node as a function of its out-degree, we associate a node's size with the size of the biological population it represents. This ensures that our indices can be made reliably robust by including population size data.

Relationship with Sackin's index

The sum $\sum_{k\in\widetilde{V}} n_k$ is just another way of expressing Sackin's index (summing over internal nodes instead of leaves). Therefore J in Equation 0.4 is essentially a weighted Sackin index (with each term in the sum weighted by the balance score W) divided by the unweighted Sackin index. In the special, important case of full m-ary leafy trees (including full m-ary cladograms), the weighted sum in J^1 (Equation 0.5) simplifies yet further. Let $\mathcal{T}^*_{n,m}$ denote the set of all trees on n leaves such that all internal nodes have the same out-degree $m > 1$, every internal node has null size, and all leaf sizes are equal. Then we obtain a remarkably simple relationship between J^1 and Sackin's index:

Proposition 0.6 Let T be a tree on n leaves with $d^+(i) = m > 1$ and $f(i) = 0$ for every internal node i. Then

$$J^1(T) = \frac{{}^1H_m(T)S(T)}{I_{S,gen}(T)},$$

where ${}^1H_m(T)$ is the Shannon entropy (base m) of the proportional node sizes, $S(T)$ is the magnitude of T, and $I_{S,gen}(T) \coloneqq \sum_{i\in\widetilde{V}(T)} S_i^*$. If additionally all leaves of T have the same size (so $T \in \mathcal{T}^*_{n,m}$) then

$$J^1(T) = \frac{\min_{n,m} I_S}{I_S(T)} = \frac{n \log_m n}{I_S(T)}, \tag{0.6}$$

where $\min_{n,m} I_S$ is the minimum I_S value of trees in $\mathcal{T}^*_{n,m}$.

The above result is somewhat surprising as it unifies our Colless-like index, which can be viewed as a weighted average of internal node balance scores, and Sackin's index, which is the sum of all leaf depths. A short proof of Proposition 0.6 is presented in the Appendix. The converse result, which is also proved in the Appendix, justifies our choice of W^1 instead of alternative balance score functions:

Proposition 0.7 Let J be a tree balance index such that

$$J(T) = \frac{1}{\sum_{k\in\widetilde{V}} n_k} \sum_{i\in\widetilde{V}} n_i W\left(\frac{n_{i_1}}{n_i}, \ldots, \frac{n_{i_{p(i)}}}{n_i}\right),$$

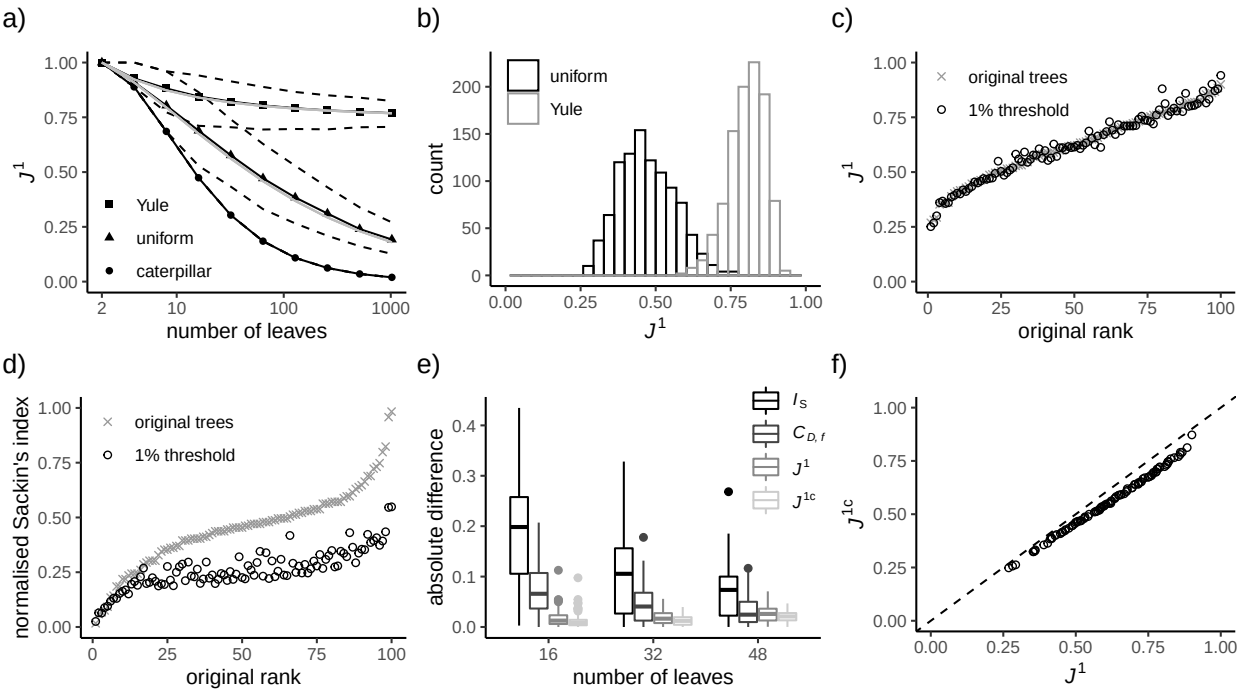

Fig. 5. **a**: J_1 values for caterpillar trees and random trees generated from the Yule and uniform models (1,000 trees per data point). All internal nodes have null size and all leaves have equal size. Solid black curves are the means; dashed curves are the 5th and 95th percentiles; and grey curves are $n \log_2 n$ divided by the corresponding expectation of I_S (where n is the number of leaves). **b**: J^1 distributions for random trees on 64 leaves generated from the Yule and uniform models (1,000 trees per model). **c**: J^1 values for 100 random trees on 16 leaves, before and after applying a 1% sensitivity threshold. These random trees were generated from the alpha-gamma model with $\alpha \sim \text{Unif}(0, 1)$ and $\gamma \sim \text{Unif}(0, \alpha)$. **d**: $I_{S,norm}$ values for the same set of random trees. **e**: Absolute change in normalised index values due to applying a 1% sensitivity threshold. Results are based on 100 random trees for each number of leaves, generated as in panels c and d. $C_{D,f}$ here is the Colless-like index with $f(n) = \ln(n + e)$ and D is the mean deviation from the median, as recommended by Mir et al. (2018). **f**: Values of J^{1c} versus J^1 for random multifurcating trees on 16 leaves, with node sizes drawn from a continuous uniform distribution. The dashed reference line has slope 1.

where $i_1, \ldots, i_{p(i)}$ are the children of node i, and W is a balance score satisfying the conditions stated before Equation 0.1. Suppose that for all trees $T \in \mathcal{T}^*_{n,m}$, $J(T) = n \log_m n / I_S(T)$. Then $W = W^1$.

The right-hand side of Equation 0.6 incidentally provides an alternative way of normalising Sackin's index on full m-ary leafy trees, including the bifurcating cladograms on which the index was originally defined. This normalised inverse Sackin index, which we can define as $J_S := n \log_m n / I_S$, provides a more satisfactory way of comparing trees that differ in their node degrees or leaf counts. $J_S = 1$ if and only if the tree has minimal depth

given m, which is equivalent to being fully symmetric, and so J_S is a *sound* tree balance index in the sense defined by Mir et al. (2018) (see Appendix for a proof). For $m > 1$, we have $J_S > 0$ but $\min J_S \to 0$ as $n \to \infty$, which makes sense because trees with more leaves can be made less balanced. In particular, when T is a caterpillar tree on $n \geqslant 2$ leaves,

$$J_S(T) = \frac{2n \log_2 n}{(n-1)(n+2)},$$

as illustrated in Figure 5a. The definition of J_S can be naturally extended to the case $m \leqslant 1$ by setting $J_S(T) \coloneqq 0$ if T is linear or has only one node. From this point of view, J^1 (a Colless-like index) is a generalisation of J_S (the normalised reciprocal of Sackin's index) to the domain of trees with arbitrary degree distributions and arbitrary node sizes.

Distributions under the Yule and uniform models

An immediate corollary of Proposition 0.6 is that J^1 can be used to test whether a set of full m-ary cladograms is consistent with a particular tree-generating model, with exactly the same sensitivity as Sackin's index. For example, Figures 5a and 5b show J^1 distributions for random bifurcating trees in $\mathcal{T}^*_{n,2}$ generated from the Yule and uniform models. These two distributions have insignificant overlap when the trees have at least a few dozen leaves.

Kirkpatrick and Slatkin (1993) showed that the expectation of I_S for the Yule model is

$$\mathbb{E}_{Yule}(I_S) = 2n \sum_{i=2}^{n} \frac{1}{i} = 2n \ln n + (2\gamma - 2)n + o(n),$$

where γ is Euler's constant and n is the number of leaves. Mir et al. (2013) have shown that the expectation of I_S for the uniform model is

$$\mathbb{E}_{Unif}(I_S) = n \left(\frac{(2n-2)!!}{(2n-3)!!} - 1 \right) = n \left(\frac{(2n-2)(2n-4)\dots(4)(2)}{(2n-3)(2n-5)\dots(3)(1)} - 1 \right),$$

which approaches $\sqrt{\pi} n^{3/2}$ as the number of leaves n approaches infinity (Blum et al., 2006; King and Rosenberg, 2021). Consistent with Proposition 0.6, we find that for random trees in $\mathcal{T}^*_{n,2}$ generated by either the Yule or the uniform model, a good approximation to the J^1

mean is $n \log_2 n$ divided by the corresponding expectation of I_S (grey curves in Fig. 5a). As $n \to \infty$, these approximations approach $1/(2 \ln 2) \approx 0.72$ and zero for the Yule and uniform models, respectively.

Robustness when applied to random trees

To test the robustness of J^1, we generated random multifurcating trees with node sizes drawn from a continuous uniform distribution, and then compared J^1 values for these trees before and after applying a 1% sensitivity threshold. In the latter case, whenever the combined frequency of a clone and its descendants was below 1%, we merged the corresponding subtree with the clone's parent, to simulate imperfect detection of rare types. As expected, the J^1 values for the two sets of trees were highly similar, with a median absolute difference of only 0.01 for trees that initially had 16 leaves (Fig. 5c). In contrast, the median absolute difference in the normalised Sackin's index for the same two sets of trees (after resolving any linear parts in the manner of Figure 2) was 0.20 (Fig. 5d), confirming that J^1 is much more robust to the omission of rare types.

As the number of leaves per tree increases, indices such as Sackin's index and the Colless-like index recommended by Mir et al. (2018) become more robust to the removal of rare types (Fig. 5e). Like J^1, these previously defined indices give more weight to nodes nearer the root. In larger trees, the nodes near the root tend to have large numbers of descendant leaves. It follows that removing a random sample of nodes from near the tips of the tree is likely to have only a modest effect on balance, as the tree's core structure is preserved. In our results, this effect outweighs an increase in the proportion of nodes removed (a median of 7%, 19% and 24% of nodes were removed from trees that originally had 16, 32 and 48 leaves, respectively, by applying the 1% sensitivity threshold). Therefore the robustness benefit of J^1 is more pronounced in trees with fewer leaves.

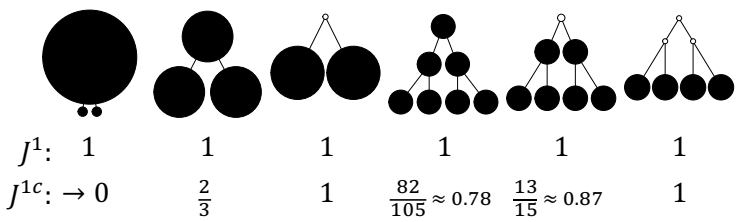

Fig. 6. Example values of J^1 versus the conservative tree balance index J^{1c}. The latter index takes account of the size of each internal node, relative to the sum of its descendant node sizes.

Comparison with a conservative tree balance index

We additionally investigated the robustness of an alternative new tree balance index J^{1c}, defined as

$$J^{1c} := \frac{1}{\sum_{k \in \widetilde{V}} S_k^*} \sum_{i \in \widetilde{V}} S_i^* \frac{S_i^*}{S_i} W_i^1.$$

J^{1c} – which we denoted J^1 in a previous paper (Noble et al., 2021) – conforms to an alternative set of axioms that define what we call a *conservative* tree balance index. This index is maximal not for all trees with equal splits, but only for leafy trees with equal splits (see Appendix for details).

An advantage of J^{1c} is that, unlike J^1, it is always insensitive to adding relatively low-magnitude subtrees to the root of the tree. Nevertheless, as the number of nodes increases, the difference between J^1 and J^{1c} rapidly diminishes, unless the root node is disproportionately large (Fig. 6). For example, when J^1 and J^{1c} are applied to random multifurcating trees on 16 leaves, with node sizes drawn from a continuous uniform distribution, the linear correlation between the two indices is 0.998 (J^{1c} is approximately 10% smaller than J^1 in this case; Fig. 5f). Accordingly, we find that J^{1c} is only slightly more robust than J^1 to the removal of rare types when applied to reasonably large random trees (Fig. 5e). For most practical purposes, we see no strong reason to favour J^{1c} over the simpler index J^1.

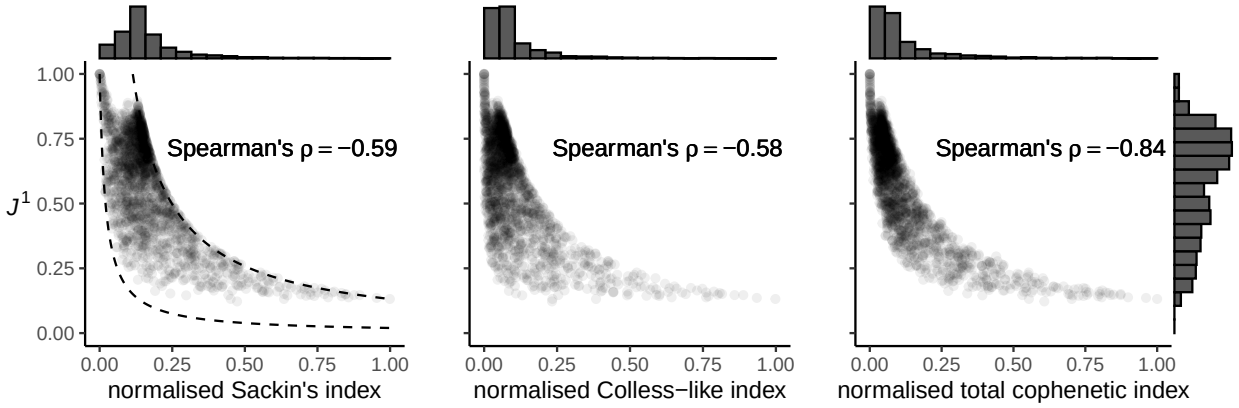

Fig. 7. Scatter plots of J^1 versus normalised Sackin's, Colless-like, and total cophenetic indices for 2,000 random multifurcating leafy trees with 100 equally sized leaves. Histograms in the margins show the marginal distributions. Dashed reference curves in the first panel are obtained by substituting $I_{S,norm}$ into Equation 0.6 with $n = 100$ and $m = 2$ (upper curve) or $m = 100$ (lower curve). We use the Colless-like index with $f(n) = \ln(n + e)$ and D the mean deviation from the median, as recommended by Mir et al. (2018). Normalisation of each index other than J^1 depends only on the number of leaves and so does not affect correlations. Trees were generated from the alpha-gamma model with $\alpha \sim \text{Unif}(0, 1)$ and $\gamma \sim \text{Unif}(0, \alpha)$.

Resolution power

Mir et al. (2013) have argued that a useful tree balance index should have good resolution power, meaning a low probability of assigning the same value to two trees with the same number of leaves, chosen uniformly at random. Proposition 0.6 implies that, when applied to full m-ary leafy trees with equally sized leaves, J^1 has the same resolution power as Sackin's index.

Correlations with preexisting indices

To compare J^1 to Sackin's index, a Colless-like index, and the total cophenetic index (defined in the Appendix) on a diverse set of trees, we generated 2,000 random multifurcating leafy trees on 100 equally sized leaves using the alpha-gamma model (Chen et al., 2009) via the R package *CollessLike* (Mir et al., 2018). As shown in Figure 7, our new balance index correlates negatively with the previously defined imbalance indices on this set of random trees, indicating that it captures a similar notion of balance. The strongest correlation is between J^1 and the total cophenetic index (Spearman's $\rho = -0.84$

for all trees, and $\rho = -0.97$ for trees with mean out-degree greater than 3). The marginal histograms in Figure 7 additionally show that more than 85% of these random trees have balance values less than 0.25 according to the previously defined indices, whereas J^1 values are more evenly distributed between zero and one, with mean and median approximately equal to 0.6.

Sensitivity to certain changes in node degree

As explained in Methods, we consider it desirable for tree balance indices to be sensitive to certain changes in node degree. In J^1 this sensitivity arises because, in the calculation of the node balance score, the node out-degree features as the base of the logarithm. For example, consider a star tree T with $l > 1$ leaves each of size $f_0 > 0$. Suppose we add to the root another $n - l$ leaves, each of size $x > 0$. If $x = f_0$ then $J^1(T) = 1$ since all the leaves have the same size. Otherwise

$$J^1(T) = -\left[l \frac{f_0}{lf_0 + (n-l)x} \log_n \left(\frac{f_0}{lf_0 + (n-l)x} \right) + (n-l) \frac{x}{lf_0 + (n-l)x} \log_n \left(\frac{x}{lf_0 + (n-l)x} \right) \right].$$

As x decreases from f_0 towards zero, $J^1(T)$ decreases monotonically to account for the growing loss of balance. And as $x \to 0$, so $J^1(T) \to \log_n l$. If we then remove these vanishingly small leaves, the value of $J^1(T)$ will jump from $\log_n l$ back to 1 because the remaining leaves are of equal size. The sensitivity of J^1 to such changes in node degree is thus a straightforward consequence of the conventional notion of node balance. The size of the jump in J^1 is at most $1 - \log_3 2 \approx 0.37$, and it approaches zero as $l/n \to 1$ (that is, when the new nodes are relatively few). The analyses shown in Figures 5e and 5f show that such discontinuities do not compromise the overall robustness of J^1 to the removal of rare types.

Implementation and algorithmic complexity

Assuming the identity of the root is known, our new indices can be computed from an adjacency matrix in $\mathcal{O}(N)$ time, where N is the number of nodes (or the number of edges plus one). Subtree magnitudes are computed via depth-first search, which takes linear time, and the computation of the balance index takes at most $\sum_{i=1}^{N} |\mathrm{Adj}(i)| = N - 1$ steps, where $\mathrm{Adj}(i)$ is the adjacency list of node i. Efficient R code for calculating J^q is shared in an online repository (Noble and Lemant, 2021).

Discussion

Here we have defined a new class of tree balance index that unifies, generalises, and in various ways improves upon previous definitions. Even when restricted to the tree types on which pre-existing indices are defined, our indices enable more meaningful comparison of trees with different degree distributions or different numbers of leaves. Due to these advantages, our indices have potential to supersede those in current use.

Our indices also enable important new applications. A challenge in comparing simulated phylogenies and trees inferred from data is that the former are exact, whereas the latter are often incomplete (Scott et al., 2020). In oncology, for example, it has been shown that whether or not a rare tumour clone is detected depends on both methodology and chance (Turajlic et al., 2018). Our balance indices largely solve this problem as they are insensitive to the omission of rare types, as demonstrated briefly here and more comprehensively in a companion paper (Noble et al., 2021).

Because of its unique relationship with Sackin's index, we especially recommend J^1 – a weighted average of the normalised entropies of the internal nodes – as defined in general by Equation 0.3 and more simply for cladograms by Equation 0.5. Given that Sackin's index has been well studied, it is convenient that J^1 inherits some of the properties of that index when applied to full m-ary cladograms, including its relatively

high sensitivity in distinguishing between alternative tree-generating models (Kirkpatrick and Slatkin, 1993; Agapow and Purvis, 2002). Within our framework, Sackin's index is seen not as a general balance index but rather as a normalising factor, which works as a balance index only in the special case of full m-ary leafy trees (for which the numerator of J^1 is independent of tree topology).

Proposition 0.6 implies that determining the precise moments of J^1 for a model that generates full m-ary leafy trees is equivalent to determining the moments of the reciprocal of Sackin's index. Figure 7 suggests that J^1 has interesting relationships with other indices such as the total cophenetic index. These are promising areas for further investigation.

Funding

This work was supported by the National Cancer Institute at the National Institutes of Health (grant number U54CA217376) to RN and VM. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Acknowledgements

We thank Laura Keller, Lisa Lamberti, Niko Beerenwinkel, Francesco Marass, Jack Kuipers and Katharina Jahn for helpful conversations, and János Podani for advice on terminology.

Author contributions

RN conceived the project. JL and RN developed the balance indices with helpful input from CLS. JL and RN obtained mathematical results with the assistance of VM. RN wrote the paper based on a chapter of JL's master's thesis. All authors have read and approved this manuscript.

References

Paul Michael Agapow and Andy Purvis. Power of eight tree shape statistics to detect nonrandom diversification: A comparison by simulation of two models of cladogenesis. *Systematic Biology*, 51(6):866–872, 2002.

Michael G. B. Blum, Olivier François, and Svante Janson. The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *The Annals of Applied Probability*, 16(4):2195–2214, 2006.

Anne Chao, Chun-Huo Chiu, and Lou Jost. Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics*, 45(1): 297–324, 2014.

Bo Chen, Daniel Ford, and Matthias Winkel. A new family of Markov branching trees: The alpha-gamma model. *Electronic Journal of Probability*, 14:400–430, 2009.

Ketevan Chkhaidze, Timon Heide, Benjamin Werner, Marc J. Williams, Weini Huang, Giulio Caravagna, Trevor A. Graham, and Andrea Sottoriva. Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. *PLOS Computational Biology*, 15(7):e1007243, 2019.

Donald H. Colless. Review of phylogenetics: the theory and practice of phylogenetic systematics. *Systematic Zoology*, 31(1):100–104, 1982.

Alexander Davis, Ruli Gao, and Nicholas Navin. Tumor evolution: Linear, branching, neutral or punctuated? *Biochimica et Biophysica Acta - Reviews on Cancer*, 1867(2): 151–161, 2017.

Mareike Fischer, Lina Herbst, Sophie Kersting, Luise Kühn, and Kristina Wicke. Tree balance indices: a comprehensive survey. *arXiv*, 2021.

Mariam Jamal-Hanjani, Gareth A. Wilson, Nicholas McGranahan, Nicolai J. Birkbak, Thomas B.K. Watkins, Selvaraju Veeriah, Seema Shafi, Diana H. Johnson, Richard Mitter, Rachel Rosenthal, et al. Tracking the evolution of non–small-cell lung cancer. *New England Journal of Medicine*, 376(22):2109–2121, 2017.

Matthew C. King and Noah A. Rosenberg. A simple derivation of the mean of the Sackin index of tree balance under the uniform model on rooted binary labeled trees. *Mathematical Biosciences*, 342(June):108688, 2021.

Mark Kirkpatrick and Montgomery Slatkin. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution*, 47(4):1171–1181, 1993.

Carlo C. Maley, Athena Aktipis, Trevor A. Graham, Andrea Sottoriva, Amy M. Boddy, Michalina Janiszewska, Ariosto S. Silva, Marco Gerlinger, Yinyin Yuan, Kenneth J. Pienta, Karen S. Anderson, Robert Gatenby, Charles Swanton, David Posada, Chung-I Wu, Joshua D. Schiffman, E. Shelley Hwang, Kornelia Polyak, Alexander R. A. Anderson, Joel S. Brown, Mel Greaves, and Darryl Shibata. Classifying the evolutionary and ecological features of neoplasms. *Nature Reviews Cancer*, 17(10):605–619, 2017.

Arnau Mir, Francesc Rosselló, et al. A new balance index for phylogenetic trees. *Mathematical biosciences*, 241(1):125–136, 2013.

Arnau Mir, Lucía Rotger, and Francesc Rosselló. Sound Colless-like balance indices for multifurcating trees. *PloS one*, 13(9), 2018.

Arne O. Mooers and Stephen B. Heard. Inferring Evolutionary Process from Phylogenetic Tree Shape. *The Quarterly Review of Biology*, 72(1):31–54, 1997.

Robert Noble and Jeanne Lemant. *RUtreebalance: Robust, universal tree balance indices*, 2021. URL `https://zenodo.org/badge/latestdoi/399934945`.

Robert Noble, Dominik Burri, Cécile Le Sueur, Jeanne Lemant, Yannick Viossat,

Jakob Nikolas Kather, and Niko Beerenwinkel. Spatial structure governs the mode of tumour evolution. *Nature Ecology and Evolution*, 2021.

János Podani. Tree thinking, time and topology: comments on the interpretation of tree diagrams in evolutionary/phylogenetic systematics. *Cladistics*, 29(3):315–327, 2013.

Alfréd Rényi. On measures of entropy and information. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:547–561, 1961.

M.J. Sackin. "Good" and "bad" phenograms. *Systematic Biology*, 21(2):225–226, 1972.

Jacob G Scott, Philip K Maini, Alexander RA A Anderson, and Alexander G Fletcher. Inferring Tumor Proliferative Organization from Phylogenetic Tree Measures in a Computational Model. *Systematic Biology*, 69(4):623–637, 2020.

Kwang-Tsao Shao and Robert R Sokal. Tree Balance. *Systematic Zoology*, 39(3):266, 1990.

Samra Turajlic, Hang Xu, Kevin Litchfield, Andrew Rowan, Stuart Horswell, Tim Chambers, Tim O'Brien, Jose I. Lopez, Thomas B.K. Watkins, David Nicol, Mark Stares, Ben Challacombe, Steve Hazell, Ashish Chandra, Thomas J. Mitchell, Lewis Au, Claudia Eichler-Jonsson, Faiz Jabbar, Aspasia Soultati, Simon Chowdhury, Sarah Rudman, Joanna Lynch, Archana Fernando, Gordon Stamp, Emma Nye, Aengus Stewart, Wei Xing, Jonathan C. Smith, Mickael Escudero, Adam Huffman, Nik Matthews, Greg Elgar, Ben Phillimore, Marta Costa, Sharmin Begum, Sophia Ward, Max Salm, Stefan Boeing, Rosalie Fisher, Lavinia Spain, Carolina Navas, Eva Grönroos, Sebastijan Hobor, Sarkhara Sharma, Ismaeel Aurangzeb, Sharanpreet Lall, Alexander Polson, Mary Varia, Catherine Horsfield, Nicos Fotiadis, Lisa Pickering, Roland F. Schwarz, Bruno Silva, Javier Herrero, Nick M. Luscombe, Mariam Jamal-Hanjani, Rachel Rosenthal, Nicolai J. Birkbak, Gareth A. Wilson, Orsolya Pipek, Dezso Ribli, Marcin Krzystanek, Istvan Csabai, Zoltan Szallasi, Martin Gore, Nicholas McGranahan, Peter Van Loo, Peter Campbell, James Larkin, and Charles Swanton. Deterministic

Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. *Cell*, 173(3):595–610.e11, 2018.

APPENDIX

Definition of the total cophenetic index

The cophenetic value $\phi(k,l)$ of a pair of leaves (k,l) is the depth of their lowest common ancestor. The total cophenetic index (Mir et al., 2013) is then the sum of the cophenetic values over all pairs of leaves:

$$I_\Phi(T) = \sum_{N-n+1 \leqslant k < l \leqslant n} \phi(k,l),$$

where N is the number of nodes and n the number of leaves. As in Sackin's index, the principle is that an unbalanced tree stretches more than a balanced tree. Being explicitly defined for all multifurcating trees, the total cophenetic index permits meaningful comparison of any two multifurcating trees on the same number of leaves.

For trees on $n > 2$ leaves, the minimum of the total cophenetic index is reached on the star tree, with $\min_n(I_\Phi) = 0$. The maximum is attained on the caterpillar tree:

$$\begin{aligned}\max_n(I_\Phi) = \sum_{k=2}^{n-1}\sum_{l=1}^{k-1} m = \sum_{k=2}^{n-1} \frac{1}{2}k(k-1) &= \frac{1}{2}\left(\frac{(n-1)n(2n-1)}{6} - \frac{n(n-1)}{2}\right) \\ &= \frac{n(n-1)(n-2)}{6} = \binom{n}{3}.\end{aligned}$$

Hence a normalised version of the total cophenetic index is $I_{\Phi,norm}(T) = I_\Phi(T)/\binom{n}{3}$. This normalised imbalance index is not minimal for all fully symmetric trees. For example, the cophenetic value of the two leftmost leaves of the fully symmetric tree in Figure 1b is two, and so both the unnormalised and normalised cophenetic indices of this tree will be nonzero.

Conservative tree balance indices

Our axioms permit J to change discontinuously when we add rare types to the root. This is because Axioms 0.3 and 0.4 consider the addition of subtrees that have vanishingly small magnitude relative to other subtrees excluding their roots, whereas the relative size of the root of the entire tree is immaterial. For example, consider a two-node linear tree T in which the non-root node has size δ, relative to the size of the root. Then $J(T) = 0$ by Axiom 0.4. But if we add another child to the root of T, also of relative size δ, then the J value of the new tree will be 1 (by Axiom 0.1), even as $\delta \to 0$. To make our index robust in such cases, we can add another axiom:

Axiom A.8 (Root limit) Let T be a tree with root r. Then $J(T) \to 0$ as $S_r^*/S_r \to 1$.

But this new axiom conflicts with Axiom 0.1, which we must then modify, such that equal splits are no longer sufficient for maximal balance:

Axiom A.9 (Alternative maximum value) $J(T) \leqslant 1$ for all trees T, and $J(T) = 1$ only if T has equal splits. Furthermore, if T has equal splits and is a leafy tree then $J(T) = 1$.

We will call a tree balance index *conservative* if it conforms to these two alternative axioms in addition to Axioms 0.2, 0.3, 0.4 and 0.5. This name is appropriate because Axiom A.8 implies that a tree will be considered imbalanced unless there is strong evidence to the contrary (in the form of a relatively small root node). Every conservative index is both universal and robust.

One way to define a class of conservative indices is to add to Equation 0.1 a *non-root dominance* factor $h : \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to (0, 1]$ with $h(x_1, x_2) \to 0$ as $x_1/x_2 \to 0$, and $h(x_1, x_2) = 1$ if and only if $x_1 = x_2$. We then obtain

$$J \coloneqq \frac{1}{\sum_{k \in \widetilde{V}} g_k} \sum_{i \in \widetilde{V}} g_i h_i W_i,$$

with $h_i = h(S_i^*, S_i)$. The role of h is to quantify the extent to which a node should be

considered a leaf (which doesn't contribute to the index's value) as opposed to an internal node (which does). Adding this factor has no effect on the balance values assigned to leafy trees, including cladograms, because if an internal node i has zero size then $h_i = 1$. Setting $h(x_1, x_2) = x_1/x_2$, we can modify Equation 0.3 to obtain the specific conservative index

$$J^{1c} \coloneqq \frac{1}{\sum_{k\in\widetilde{V}} S_k^*} \sum_{i\in\widetilde{V}} S_i^* \frac{S_i^*}{S_i} W_i^1.$$

We previously used J^1 instead of J^{1c} to denote the above index (Noble et al., 2021).

Alternative axioms proposed by Fischer et al. (2021)

Shortly after we posted a preprint version of the current article, Fischer et al. (2021) posted a preprint in which they proposed two alternative axioms for non-robust, non-universal tree balance indices, such as Sackin's and Colless' indices. In these axioms, $\mathcal{BT}_n^*$ denotes the set of rooted bifurcating trees with n leaves, $\mathcal{T}_n^*$ is the set of all rooted trees with n leaves such that $d^+(i) > 1$ for all internal nodes i, and the tree balance index is denoted t.

Axiom A.10 (Fischer et al. minimum value) The caterpillar tree with n leaves is the unique tree minimising t on $\mathcal{T}_n^*$ (if t is defined on multifurcating trees) or on $\mathcal{BT}_n^*$ (if t is defined only on bifurcating trees) for all $n \geqslant 1$.

Axiom A.11 (Fischer et al. maximum value) The fully symmetric bifurcating tree with n leaves is the unique tree maximising t on $\mathcal{BT}_n^*$ for all $n = 2^h$ with $h \in \mathbb{N}_{\geqslant 0}$.

These axioms can be compared with our axioms if we consider only leafy trees in which all leaves have equal size (such as cladograms). Axiom A.11 is then just a special case of our more general Axiom 0.1, because the fully symmetric bifurcating tree with n leaves is the only tree in $\mathcal{BT}_n^*$ that has equal splits. But Axiom A.10 is not necessarily consistent with our Axiom 0.2. In particular, as shown in Figure 4b, our index J^1 does not comply with Axiom A.10 in the case of multifurcating leafy trees. We can resolve this

incompatibility with the following simplification:

Axiom A.12 (Alternative Fischer et al. minimum value) The caterpillar tree with n leaves is the unique tree minimising t on $\mathcal{BT}_n^*$ for all $n \geqslant 1$ (whether or not t is defined on multifurcating trees).

J^1 is consistent with Axiom A.12 because, when we consider only bifurcating leafy trees in which all leaves have equal size, J^1 is equal to J_S (by Proposition 0.6), which is inversely proportional to I_S by definition, and the caterpillar tree is the unique bifurcating tree that maximises I_S (Fischer et al., 2021). Although Axiom 0.1 does not necessarily imply Axiom A.12, it is reasonable to expect useful universal tree balance indices to satisfy both conditions.

Proof that the index of Equation 0.1 satisfies our five axioms

Proof. Axiom 0.1 (Maximum value): We have $J \leqslant 1$ since g and W lie between zero and one by definition. Also if any internal node j of tree T doesn't split its descendants into at least two equal-magnitude subtrees then $W_j < 1$ by definition and so

$$\sum_{i \in \widetilde{V}} g_i W_i < \sum_{i \in \widetilde{V}} g_i \implies J(T) < 1.$$

Now let T be a tree such that every internal node splits its descendants into at least two equal-magnitude subtrees. Then $W_i = 1$ for all $i \in \widetilde{V}$ by definition. Hence

$$J(T) = \frac{1}{\sum_{k \in \widetilde{V}} g_k} \sum_{i \in \widetilde{V}} g_i = 1.$$

Axiom 0.2 (Minimum value): We have $J \geqslant 0$ since g and W are always non-negative by definition. Also if T is a linear tree then $W_i = 0$ for all $i \in \widetilde{V}$ by definition, and hence $J(T) = 0$. Conversely, if some internal node j has $d^+(j) > 1$ then $W_j > 0$ by definition and, because g_j must be positive by definition, we must have $J(T) > 0$.

Axiom 0.3 (Insensitivity): Adding a subtree to a leaf l changes the tree balance value via the contributions of two sets of nodes: the internal nodes of T_l (including l), and

all other internal nodes. For each internal node $i \in \widetilde{V}(T_l)$, as $S_l^* / \sum_{j \in \widetilde{V}(T')} S_j^* \to 0$ so also $S_i^* / \sum_{j \in \widetilde{V}(T')} S_j^* \to 0$ (because $S_i^* \leqslant S_l^*$), which implies $g_i \to 0$ by definition, and hence all such contributions approach zero. The contribution of all other internal nodes also approaches zero because g and W are continuous by definition.

Axiom 0.4 (Linear limit): Let $i \in \widetilde{V}(T)$ with $d^+(i) = 1$. Without loss of generality, let i_1 denote the original child of i, and $i_2, \ldots, i_p$ denote the newly added children of i. Adding subtrees to i changes the tree balance value via the contributions of the newly added nodes and of node i. As $S_{i_1}/S_i^* \to 1$, so $S_{i_k}/S_i^* \to 0$ for all $k \in \{2, \ldots, p\}$. This implies that $S_{i_k} / \sum_{j \in \widetilde{V}(T')} S_j^* \to 0$ and hence $g_{i_k} \to 0$ by definition for all $k \in \{2, \ldots, p\}$. Therefore the first contribution approaches zero. Also as $S_{i_1}/S_i^* \to 1$, we have $\max(S_{i_1}/S_i^*, \ldots, S_{i_p}/S_i^*) \to 1$, and so $W_i \to 0$ by definition. Therefore the second contribution also approaches zero.

Axiom 0.5 (Continuity): The continuity of J follows immediately from the continuity of g and W. □

New generalisations of Sackin's and Colless' indices

The number of distinct subtrees that contain a given leaf l is equal to its number of ancestors, which is the same as ν_l, the depth of l. Hence Sackin's index is equivalent to the sum of the leaf counts of the subtrees rooted at each internal node. By extension, we can define a new, more general form of Sackin's index that accounts for node sizes:

$$I_{S,gen}(T) \coloneqq \sum_{i \in \widetilde{V}(T)} S_i^*,$$

where S_i^* is the magnitude of the subtree rooted at node i, excluding the root. In the special case of leafy trees in which all leaves have size one, we recover $I_{S,gen} = I_S$. This new index is not very useful for assessing tree balance because it increases with the total tree magnitude, but in our framework it performs an important role as a normalising factor.

If we let S_{i_1} denote the magnitude of the left branch of the subtree rooted at i, and

S_{i_2} denote the magnitude of the right branch, then we can generalise Colless' index to account for node sizes in bifurcating trees:

$$I_{C,gen}(T) \coloneqq \sum_{i \in \widetilde{V}(T)} |S_{i_1} - S_{i_2}| = \sum_{i \in \widetilde{V}(T)} S_i^* |p_{i_1} - p_{i_2}|,$$

where $p_{i_j} = S_{i_1}/S_i^*$. This definition reduces to I_C in the case of leafy trees in which all leaves have size one. The right-hand expression above clarifies that the contribution of each node to Colless' index is the product of the node's importance (that is, its number of descendants) and its balance (the degree to which the node splits its descendants into two equal-magnitude subtrees). We further see that $I_{C,gen}(T) \leqslant I_{S,gen}(T)$ for all trees T (because $|p_{i_1} - p_{i_2}| \leqslant 1$ for all i_1, i_2), which suggests the normalisation

$$I_{C,gen,norm} \coloneqq \frac{I_{C,gen}}{I_{S,gen}} = \frac{1}{\sum_{k \in \widetilde{V}} S_k^*} \sum_{i \in \widetilde{V}(T)} S_i^* |p_{i_1} - p_{i_2}|.$$

This new generalisation of Colless' index is more robust than the conventional form, in the sense that its value is insensitive to the addition or removal of relatively small nodes. $I_{C,gen,norm}$ also enables meaningful comparison of trees with different numbers of leaves. But the problem remains that $I_{C,gen,norm}$ applies only to bifurcating trees.

Other balance indices based on generalised entropies

As defined by Chao et al. (2014), generalised entropies for $q \geqslant 0, q \neq 1$ are

$${}^qH \coloneqq \frac{1}{q-1}\left(1 - \sum_{i=1}^{P} p_i^q\right).$$

Parameter q determines the sensitivity to the type frequencies. ${}^0H$ is simply the richness (minus 1) of the population, which corresponds to ignoring the frequencies and just counting the types. For $0 < q < 1$, rare types are given more weight than implied by their proportion, whereas for $q > 1$ abundant types matter more. ${}^2H$ is the Gini-Simpson coefficient. In the limit $q \to 1$ we recover the Shannon entropy ${}^1H_e$.

For $q > 0$, ${}^qH$ attains its maximum value if and only if all types have equal

frequency $p_i = 1/m$:

$$\max({}^qH) = \frac{1}{q-1}\left(1 - \frac{1}{m^{q-1}}\right) = \frac{m^{q-1}-1}{m^{q-1}(q-1)}.$$

We can therefore define a normalised balance score W_i^q for $q > 0, q \neq 1$ and $i \in \widetilde{V}$:

$$W_i^q \coloneqq \begin{cases} \dfrac{d^+(i)^{q-1}}{d^+(i)^{q-1}-1}\left(1 - \sum\limits_{j\in C(i)} p_{ij}^q\right) & \text{if } d^+(i) \geqslant 2 \\ 0 \quad \text{otherwise.} \end{cases}$$

Similarly, one can define W_i^q for $q > 0, q \neq 1$ based on the entropy defined by Rényi (1961):

$$W_i^q \coloneqq \begin{cases} \dfrac{1}{(1-q)\log d^+(i)} \log\left(\sum\limits_{j\in C(i)} p_{ij}^q\right) & \text{if } d^+(i) \geqslant 2 \\ 0 \quad \text{otherwise.} \end{cases}$$

In either case, a balance index J^q satisfying our axioms is

$$J^q \coloneqq \frac{1}{\sum_{k\in\widetilde{V}} S_k^*} \sum_{i\in\widetilde{V}} S_i^* W_i^q,$$

for any $q > 0$. And in either case, $J^q \to J^1$ as $q \to 1$.

Proof of Proposition 0.6

Proof. By definition of J^1, if T is a tree on n leaves with $d^+(i) = m > 1$ and $f(i) = 0$ for every internal node i then

$$J^1(T) = \frac{-1}{\sum_{k\in\widetilde{V}} S_k} \sum_{i\in\widetilde{V}} \sum_{j\in C(i)} S_j \log_m \frac{S_j}{S_i}.$$

The sum of subtree magnitudes over the set of all internal nodes is equal to the sum of ν_l multiplied by leaf size over the set of all leaves:

$$I_{S,gen} \coloneqq \sum_{k\in\widetilde{V}} S_k = \sum_{k\in L} \nu_k f(k).$$

Summing first over the internal nodes and then over their children gives the same result:

$$\sum_{i\in\widetilde{V}} \sum_{j\in C(i)} S_j = \sum_{i\in\widetilde{V}} S_i = \sum_{i\in L} \nu_i f(i) = \sum_{i\in L} f(i) \sum_{j=1}^{\nu_i} 1.$$

Let $a(i,j)$ denote the ancestor of node i at distance j, with $a(i,0) = i$ and $a(i,\nu_i) = r$ (the root) for all i. Then by extension,

$$\sum_{i\in\widetilde{V}} \sum_{j\in C(i)} S_j\theta(S_i,S_j) = \sum_{i\in L} f(i) \sum_{j=1}^{\nu_i} \theta(S_{a(i,j)}, S_{a(i,j-1)}),$$

for any function θ. In particular, we have

$$\sum_{i\in\widetilde{V}} \sum_{j\in C(i)} S_j \log_m \frac{S_j}{S_i} = \sum_{i\in L} f(i) \sum_{j=1}^{\nu_i} \log_m \frac{S_{a(i,j-1)}}{S_{a(i,j)}}.$$

Substituting this result into the expression for J^1 we find

$$\begin{aligned} J^1(T) &= \frac{-1}{\sum_{k\in\widetilde{V}} S_k} \sum_{i\in L} \sum_{j=1}^{\nu_i} f(i) \log_m \frac{S_{a(i,j-1)}}{S_{a(i,j)}} \\ &= \frac{-1}{\sum_{k\in\widetilde{V}} S_k} \sum_{i\in L} f(i) \sum_{j=1}^{\nu_i} (\log_m S_{a(i,j-1)} - \log_m S_{a(i,j)}). \end{aligned}$$

The right-hand sum is a telescoping series that collapses to give

$$J^1(T) = \frac{-1}{\sum_{k\in\widetilde{V}} S_k} \sum_{i\in L} f(i)(\log_m S_{a(i,0)} - \log_m S_{a(i,\nu_i)}).$$

Now since i is a leaf, $\log_m S_{a(i,0)} = \log_m S_i = \log_m f(i)$. Also $\log_m S_{a(i,\nu_i)} = \log_m S_r = \log_m S(T)$. Hence

$$\begin{aligned} J^1(T) &= \frac{-1}{\sum_{k\in\widetilde{V}} S_k} \sum_{i\in L} f(i)(\log_m f(i) - \log_m S(T)) \\ &= \frac{-1}{\sum_{k\in\widetilde{V}} S_k} \sum_{i\in L} f(i) \log_m \frac{f(i)}{S(T)} \\ &= \frac{{}^1H_m(T) S(T)}{\sum_{k\in\widetilde{V}} S_k} = \frac{{}^1H_m(T) S(T)}{I_{S,gen}(T)}. \end{aligned}$$

If additionally all leaves i of T have the same size $f(i) = f_0$ then $S(T) = nf_0$, ${}^1H_m(T) = \log_m n$, and $I_{S,gen}(T) = f_0 I_S(T)$, which implies $J^1(T) = n \log_m n / I_S(T)$. □

Proof of Proposition 0.7

Proof. Since $\sum_{k\in\widetilde{V}} n_k = I_S(T)$, the conditions are equivalent to

$$I_S(T)J(T) = \sum_{i\in\widetilde{V}} n_i W_i = n \log_m n, \quad \text{with } W_i = W\left(\frac{n_{i_1}}{n_i}, \ldots, \frac{n_{i_{p(i)}}}{n_i}\right),$$

where $n_{i_1}, \dots n_{p(i)}$ are the children of i. Let T be a tree in $\mathcal{T}^*_{n,m}$ and i be an internal node of T. Then $T_i \in \mathcal{T}^*_{n_i,m}$ and $T_j \in \mathcal{T}^*_{n_j,m}$ for every child j of i. Therefore

$$I_S(T_i)J(T_i) = n_i W_i + \sum_{j \in C(i)} J(T_j) = n_i W_i + \sum_{j \in C(i)} n_j \log_m n_j.$$

Also, $I_S(T_i)J(T_i) = n_i \log_m n_i$, so we have

$$\begin{aligned} & n_i W_i + \sum_{j \in C(i)} n_j \log_m n_j = n_i \log_m n_i \\ \implies & W_i = \log_m n_i - \sum_{j \in C(i)} \frac{n_j}{n_i} \log_m n_j. \end{aligned}$$

Since $\sum_{j \in C(i)} n_j = n_i$, this implies

$$W_i = \sum_{k \in C(i)} \frac{n_k}{n_i} \log_m n_i - \sum_{j \in C(i)} \frac{n_j}{n_i} \log_m n_j = - \sum_{j \in C(i)} \frac{n_j}{n_i} \log_m \frac{n_j}{n_i} = W^1_i.$$

□

Proof that J_S is a sound tree balance index

Proof. By the definition of Mir et al. (2018), a *sound* tree balance index J is such that $J(T)$ is maximal if and only if T is fully symmetric. The fully symmetric full m-ary tree on n leaves is the unique tree that minimises I_S among full m-ary trees on n leaves. This minimum value is $\min_{n,m} I_S = n \log_m n$ (since every leaf l has the same depth $\nu_l = \log_m n$). Because $J_S \coloneqq n \log_m n / I_S$ is defined only on full m-ary trees, if follows that $J_S(T)$ is maximal if and only if T is fully symmetric. □