A new universal system of tree shape indices

Robert Noble[1,*], Kimberley Verity[1]

[1] *Department of Mathematics, City, University of London, London, UK*

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

Abstract

The comparison and categorization of tree diagrams is fundamental to large parts of biology, linguistics, computer science, and other fields, yet the indices currently applied to describing tree shape have important flaws that complicate their interpretation and limit their scope. Here we introduce a new system of indices with no such shortcomings. Our indices account for node sizes and branch lengths and are robust to small changes in either attribute. Unlike currently popular phylogenetic diversity, phylogenetic entropy, and tree balance indices, our definitions assign interpretable values to all rooted trees and enable meaningful comparison of any pair of trees. Our self-consistent definitions further unite measures of diversity, richness, balance, symmetry, effective height, effective outdegree, and effective branch count in a coherent system, and we derive numerous simple relationships between these indices. The main practical advantages of our indices are in 1) quantifying diversity in non-ultrametric trees; 2) assessing the balance of trees that have non-uniform branch lengths or node sizes; 3) comparing the balance of trees with different leaf counts or outdegrees; 4) obtaining a coherent, generic, multidimensional quantification of tree shape that is robust to sampling error and inferential error. We illustrate these features by comparing the shapes of trees representing the evolution of HIV and of Uralic languages. Given the ubiquity of tree structures, we identify a wide range of applications across diverse domains.

Introduction

Tree shape indices that quantify key properties of rooted trees – such as the effective number of leaves, average out-degree, and balance – have myriad applications. Conservation biologists use phylogenetic diversity values to determine which actions will preserve the most biodiversity (Tucker et al., 2017; Veron et al., 2019). Tree balance indices are used to compare models and to infer parameter values in systematic biology

(Mooers and Heard, 1997; Purvis and Agapow, 2002), virology (Chindelevitch et al., 2021; Barzilai and Schrago, 2023), epdiemiology (Leventhal et al., 2012; Colijn and Gardy, 2014), and oncology (Scott et al., 2020; Noble et al., 2022). Computer scientists seek to balance binary trees to make them more efficient as data structures (Albers and Westbrook, 2005). Numerous indices designed for such tasks have previously been proposed (Pavoine and Bonsall, 2011; Tucker et al., 2017; Fischer et al., 2021).

Rather than simply adding to a profusion of indices, our aim here is to solve important open problems: How can we modify existing phylogenetic diversity and entropy indices so that they are meaningful when applied to non-ultrametric trees? How can we define a tree balance index that accounts for both branch lengths and node sizes? How can we likewise generalize the concepts of outdegree, branch count, and node count? How can we unite all these types of tree shape index in a coherent system, so that their interrelationships can be easily understood? Only by solving these problems can we arrive at a general purpose method for fairly evaluating the shape of any rooted tree.

Among current diversity indices for generic rooted trees, arguably the most sophisticated are those introduced by Chao et al. (2010), which generalize and unify previous definitions of Hill (1973), Faith (1992), Jost (2006) and Allen et al. (2009). In quantifying the effective number of types in a data set, these ${}^q\bar{D}$ indices account for both node sizes (type frequencies) and branch lengths (degree of dissimilarity between types). Nevertheless, a critical shortcoming of these indices, which limits their applications, is that they assign meaningful values only to leafy ultrametric trees (that is, trees in which the only non-zero-sized nodes are leaves, all equally distant from the root) (Chao et al., 2010; Leinster and Cobbold, 2012). We will further show that the ${}^q\bar{D}$ indices of Chao et al. (2010) are not fully self-consistent and have peculiar properties for $q > 1$. Moreover, the relationships between these diversity indices and other types of index, such as tree balance indices, are generally opaque, which thwarts multi-dimensional analysis.

Conventional tree balance and imbalance indices – including those attributed to Sackin (1972) and Colless (1982), the total cophenetic index of Mir et al. (2013), and others reviewed by Fischer et al. (2021) – are also flawed. These indices, which are meant to quantify the extent to which each internal node splits its descendants into equally sized subtrees, are not defined for all rooted trees, do not permit meaningful comparison of trees with differing leaf counts, and are highly sensitive to the addition or removal of rare types (Noble et al., 2022; Lemant et al., 2022). We recently introduced a family of tree balance indices that solve these problems and that have additional desirable properties (Lemant et al., 2022). Our indices are defined for any degree distribution, account for node sizes, and enable meaningful comparison of trees with different numbers of leaves. But our previously definitions do not account for branch lengths, which restricts their applications because branch lengths often convey important information (for example, genetic distance in virus evolution, or elapsed time in the evolution of species).

Here we define a new system of indices that resolve all the aforementioned problems by accounting for node sizes and branch lengths, being robust to small changes to the tree, assigning meaningful values to all rooted trees, and belonging to a coherent framework, so that mathematical relationships between the indices are well characterized. Our system captures fundamental properties such as diversity (effective number of leaves), tree balance (the extent to which each internal node splits its descendants into equally sized subtrees), and bushiness (average effective outdegree). Given that our indices share the desirable properties but not the flaws of prior indices, we discuss their potential to supersede current methods in a wide range of applications.

Materials and Methods

Hill numbers as a basis for defining robust, universal, interpretable tree indices

A rooted tree is a tree in which one node is designated the root and all branches are directed away from the root. Our aim is to define indices that are useful for categorizing and comparing the shapes of unlabelled rooted trees that have three attributes: tree topology, non-negative node sizes, and non-negative branch lengths. These indices should be generic and model-agnostic, meaning that they make no assumptions about what the tree represents or the process by which it was generated. In evolutionary trees, for example, the size of a node can correspond to the population size of the respective biological type, or simply to whether a type is extant (node size 1) or extinct (0), while branch lengths can represent genetic distance, morphological difference, or elapsed time. Linguists use similar structures with unequal branch lengths to study the evolution of languages (Honkola et al., 2013; Atkinson and Gray, 2005). In computing, the size of a search tree node corresponds to the probability of it being visited.

In this general context, a useful index should be robust, universal, and interpretable (RUI). A loose definition of robustness is that small changes to the tree have only small effects on the index value, except where sensitivity is desirable; universal means that the index is defined for all rooted trees; and interpretable implies a simple, consistent interpretation, enabling meaningful comparison of any pair of rooted trees. Lemant et al. (2022) provides more rigorous, axiomatic definitions. In practical terms, robustness implies that an index is relatively insensitive to the effects of issues such as sampling error, inferential error, omission of rare types, imperfect genetic sequencing, and incomplete resolution of ancestral relationships. All our indices are dimensionless but the diversity indices can be re-scaled in terms of the branch length unit where desired.

We begin by recalling the family of diversity indices attributed to Hill (1973). These Hill numbers are functions of a set of proportions $P = \{p_1, \ldots, p_n\}$ with $0 \leqslant p \leqslant 1$ for all

$p \in P$ and $\sum_{i=1}^{n} p_i = 1$. Every Hill number of order $q \geqslant 0$ can be written as

$$^{q}D(P) := \left(\sum_{i=1}^{n} p_i^q \right)^{\frac{1}{1-q}} \text{ with } {}^{1}D(P) := \lim_{q \to 1} {}^{q}D(P) = \exp\left(- \sum_{i=1}^{n} p_i \log p_i \right).$$

Hence ${}^{q}D$ is the exponential of the Rényi entropy of order q, which we will denote ${}^{q}H$, and ${}^{1}H$ is Shannon's entropy. Another important special case is

$$^{0}D(P) := |\{p \in P : p > 0\}|,$$

which is simply the number of types, or richness. Following Pielou (1966) and Jost (2010), we further define the evenness indices

$$^{q}J(P) := \begin{cases} \dfrac{\log {}^{q}D(P)}{\log {}^{0}D(P)} \in [0,1] & \text{if } {}^{0}D(P) > 1 \\ 1 & \text{otherwise.} \end{cases}$$

For completeness, we set ${}^{q}D(\emptyset) = 0$ and ${}^{q}J(\emptyset) = 1$.

We can apply these indices to a rooted tree T simply by equating $P(T) = \{p_1, \ldots, p_n\}$ to the proportional sizes of the n nodes of T. The richness index ${}^{0}D(T) = {}^{0}D(P(T))$ then quantifies the number of non-zero-sized nodes in the tree, which we will refer to as the counted nodes. In an evolutionary tree, counted nodes correspond to extant types. For each $q > 0$, the diversity index ${}^{q}D(T) = {}^{q}D(P(T))$ can be interpreted as an effective number of counted nodes, while ${}^{q}J(T) = {}^{q}J(P(T))$ gauges the evenness of the counted node sizes.

Clearly ${}^{q}D$ and ${}^{q}J$ are insensitive to small changes to proportional node sizes. For $q > 0$, ${}^{q}D$ is also generally robust to the addition or removal of relatively small nodes (and the degree of robustness increases with q), whereas ${}^{0}D$ and ${}^{q}J$ are not, as is appropriate for indices that are meant to quantify richness and evenness. ${}^{q}D$ and ${}^{q}J$ are universal because they can be applied to any set of node sizes, and they are interpretable as described above. Yet although these indices are RUI, they are clearly inadequate for assessing tree shape because they depend only on node sizes, ignoring both tree topology and branch lengths.

Many indices that capture aspects of tree shape have previously been defined (surveys include Pavoine and Bonsall (2011); Tucker et al. (2017); Fischer et al. (2021)) but, to the best of our knowledge, none is RUI (Table 1). We address this deficiency by developing new RUI tree indices that extend the basic indices ${}^{q}D$ and ${}^{q}J$ to account for tree topology and branch lengths. We do this using three types of weighted mean, which we refer to as the longitudinal mean, the node-wise mean, and the star mean (Table 2). Our consistent definitions ensure that our indices can be precisely related to each other and to ${}^{q}D$ and ${}^{q}J$ in numerous meaningful ways, so that all the indices belong to a single coherent system.

	Robust		Universal	Interpretable	
	Robustly accounts for node sizes?	Robustly accounts for branch lengths?	Defined for all rooted trees?	Has a simple, consistent interpretation?	Can meaningfully compare any pair of rooted trees?
Faith's PD	No	Yes			
Allen et al's H_P	Yes			Only for leafy ultrametric trees	
Chao et al's ${}^q\bar{D}$					
Sackin's index	No			Yes	No
Colless's index					
Total cophenetic index					
Lemant et al's J^q	Yes	No	Yes	Only if uniform branch lengths	

Table 1. Properties of some previously defined non-RUI tree indices (see main text for definitions and citations.)

Branches or nodes	Type of average	Richness	Diversity (with $q > 0$)	Evenness (with $q > 0$)
Nodes	None	${}^0D$ = number of counted nodes	${}^qD$ = effective number of counted nodes	${}^qJ$ = evenness of counted node sizes
Branches	Longitudinal mean	${}^0D_L$ = average branch count across the tree	${}^qD_L$ = effective number of maximally distant leaves	${}^qJ_L$ = evenness of branch sizes across the tree (tree symmetry if leafy and ultrametric)
Branches	Node-wise mean	${}^0D_N$ = average effective outdegree, ignoring branch sizes	${}^qD_N$ = average effective outdegree, accounting for branch sizes	${}^qJ_N$ = tree balance
Branches	Star mean	${}^0D_S$ = effective number of non-root nodes	${}^qD_S$ = effective number of branches, accounting for branch sizes	${}^qJ_S$ = evenness of all branch sizes

Table 2. Nature, notation and interpretation of RUI tree indices, including prior indices (top row) and new indices (second, third and fourth rows). Counted nodes are those with non-zero size.

Further preliminary definitions

In a rooted tree, the depth of a node is the sum of the branch lengths along the unidirectional path from the root to the node. The height of the tree is the maximum depth of its non-zero-sized nodes. Nodes with no descendants are called leaves and non-leaves are called internal nodes. We define the size of a branch as the sum of the proportional node sizes that descend (directly or indirectly) from the branch. For example, in the three-leaf tree depicted in Figure 1a, the branches descending from the root have sizes $\frac{1}{3}$ and $\frac{2}{3}$, and the other two branches each have size $\frac{1}{3}$. The size of any segment of a

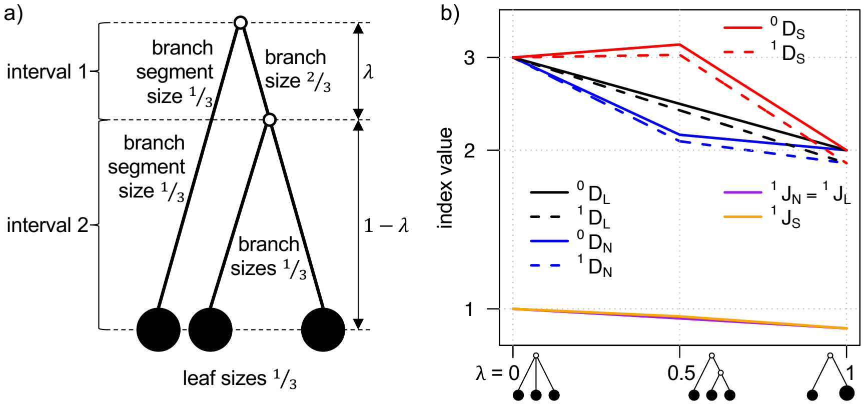


Fig. 1. a) A leafy bifurcating ultrametric tree with three equally sized leaves. In this and every subsequent tree diagram, open circles indicate zero-sized nodes. b) Index values versus branch length λ for the three-leaf tree. The y-axis is log-transformed so that the curves for all diversity indices appear piecewise linear. ${}^{1}J_S$ is slightly greater than ${}^{1}J_N$ whenever $0 < \lambda < 1$.

branch is the same as the size of the branch.

A leafy tree is such that all internal nodes have zero size (equivalently, all counted nodes are leaves). A tree is ultrametric if all its leaves have the same depth after the removal of all subtrees that contain only zero-sized branches (corresponding to extinct lineages in an evolutionary tree). A caterpillar tree is a bifurcating tree in which every internal node except one has exactly one child leaf. A star tree is a tree in which all non-zero-sized branches are attached to the root. We define a piecewise star tree as a tree that can be divided into transverse intervals such that, within each interval, all the non-zero-sized branches are attached to a common node. For example, the leafy ultrametric tree in Figure 1a is a star tree if $\lambda = 0$ or $\lambda = 1$ and is otherwise a caterpillar tree. To simplify our notation, we will usually omit the tree as a function argument (for example, writing ${}^{0}D$ instead of ${}^{0}D(T)$).

It will be helpful to recall that, for a sequence of positive real numbers $X = x_1, \ldots, x_n$, real number $r \neq 0$, and set of positive weights $W = w_1, \ldots, w_n$, the weighted power mean of exponent r is

$$M_r(X; W) := \left(\frac{\sum_{i=1}^{n} w_i x_i^r}{\sum_{i=1}^{n} w_i} \right)^{\frac{1}{r}}.$$

M_0 is defined from the limit as

$$M_0(X; W) := \exp\left(\frac{\sum_{i=1}^{n} w_i \log x_i}{\sum_{i=1}^{n} w_i} \right).$$

M_{-1}, M_0 and M_1 are respectively the weighted harmonic, geometric, and arithmetic means. $M_{-\infty}$ and M_{∞} respectively return the minimum and the maximum. Power means are closely related to Hill numbers as, for all $q \geqslant 0$ and any sequence of proportions P,

$$^{q}D(P) = [M_{q-1}(P; P)]^{-1}. \tag{0.1}$$

Prior tree balance and imbalance indices

The most popular conventional tree imbalance indices can be expressed in the form

$$I_X = \sum_{i \in V} n_i F_X(i),$$

where V is the set of all internal nodes and n_i is the number of leaves that descend from node i. For I_S (Sackin's index), I_C (Colless' index) and I_Φ (the total cophenetic index) we have

$$F_S(i) = 1, \quad F_C(i) = |p_{i_1} - p_{i_2}|, \quad F_\Phi(i) = \frac{n_i - 1}{2},$$

where p_{i_1} is the proportion of the n_i leaves that descend from the left child branch of i, and p_{i_2} is the proportion that descend from the right child branch. I_C is defined only for bifurcating trees (in which all internal nodes have outdegree two). I_S and I_Φ are defined only for trees in which all internal nodes have outdegree greater than one. By convention, each index is normalized over the set of trees on $n > 2$ leaves by subtracting its minimum value over such trees and then dividing by the difference between its maximum and its minimum. The minima of I_S, I_C and I_Φ are n, 0 and 0, and the maxima are $(n+2)(n-1)/2$, $\binom{n-1}{2}$ and $\binom{n}{3}$, respectively (Shao and Sokal, 1990; Rogers, 1993; Mir et al., 2013).

Lemant et al. (2022) proposed instead defining tree balance or imbalance indices in the form of the weighted arithmetic mean

$$\frac{1}{\sum_{i \in V} w_i} \sum_{i \in V} w_i F(i),$$

where w_i is the weight assigned to node i, and $F(i)$ quantifies the degree to which node i splits its descendants into equally sized subtrees. For example, we can obtain an alternative normalization of Colless' index by setting $w_i = n_i$ and $F(i) = F_C(i)$. The normalizing factor $\sum_{i \in V} w_i$ is then Sackin's index. An advantage of this approach is that it allows us to compare the balance of any pair of trees for which F is defined, rather than only trees with equal leaf counts.

Definition of the normalizing factor $\bar{h}$

Consistent with Lemant et al. (2022), our new index definitions are based on weighted means. Our preferred weights require us to define the normalizing factor

$$\bar{h} := \sum_{b \in B} s_b l_b \leqslant h,$$

where B is the set of all branches in the tree, $s_b \in [0,1]$ is the size of branch b, l_b is the length of branch b, and h is the tree height. We can interpret $\bar{h}$ (denoted $\bar{T}$ in Chao et al. (2010)) as the effective tree height or as the average counted node depth. In computer science, $\bar{h}$ is called the weighted path length (Albers and Westbrook, 2005). For leafy trees with uniform leaf sizes and uniform branch lengths, $\bar{h} = lI_S/{}^0D$, where l is the branch length and I_S is Sackin's index. Hence $\bar{h}$ can also be considered a generalization of Sackin's index. Indeed, we have previously argued that Sackin's index is best interpreted not as a general imbalance index but rather as a normalizing factor, which works as an imbalance index only in the special case of trees with uniform node sizes, uniform branch lengths, and uniform outdegree (Lemant et al., 2022). $\bar{h} = h$ if and only if the tree is leafy and ultrametric.

Definition of the longitudinal mean

The basic idea of the longitudinal mean is that we split the tree into transverse intervals, calculate an index value based on the proportional sizes of the branch segments within each interval, and then take a weighted average of these within-interval index values. Let I denote the set of transverse intervals created by locating an interval boundary at every node depth (dashed lines in Figure 1a), excluding intervals that contain only zero-sized branches. Each interval i then contains a set B_i of branch segments, all of the same length, which we will refer to as the interval height h_i. Let

$$S_i := \sum_{b \in B_i} s_b \in (0,1],$$

where s_b is the size of branch segment b. Then $S_i = 1$ for all intervals i if and only if the tree is leafy and ultrametric. It follows that

$$\sum_{i \in I} S_i h_i = \bar{h}.$$

Now for each $b \in B_i$, define the within-interval proportional branch size $p_b := s_b/S_i$ and let $P_i := \{p_b : b \in B_i, p_b > 0\}$. Then $\sum_{p \in P_i} p = \sum_{b \in B_i} p_b = 1$ for all intervals $i \in I$.

Finally, for index F and tree T, we define the longitudinal mean of order r of F as the functional $F \mapsto M_{long,r}(F)$ such that

$$M_{long,r}(F)(T; w) := \begin{cases} \left(\dfrac{\sum_{i \in I(T)} w_i [F(P_i)]^r}{\sum_{i \in I(T)} w_i} \right)^{\frac{1}{r}}, & \text{if } h > 0 \\ F(\emptyset) & \text{otherwise,} \end{cases} \tag{0.2}$$

where the weight $w > 0$ is a function of i that remains to be specified. Hence $M_{long,r}(F)$ is a weighted power mean of the F values assigned to the intervals. For succinctness, we will omit the argument T and specify w only where necessary.

Example 0.1 For the function $F(x_1,\ldots,x_n)=\sum_{k=1}^{n} x_k$ we have

$$M_{long,r}(F)=\left(\frac{\sum_{i\in I} w_i \left(\sum_{p\in P_i} p\right)^r}{\sum_{i\in I} w_i}\right)^{\frac{1}{r}}=1.$$

New longitudinal mean indices

We define new tree indices as longitudinal means of ${}^qD$ and ${}^qJ$ with $w_i = S_i h_i$, so that the index value assigned to each interval i is weighted by the product of the length h_i and the summed sizes S_i of the branch segments that i contains. First, we define

$$^qD_L := M_{long,0}(^qD). \tag{0.3}$$

This is equivalent to ${}^qH_L = M_{long,1}({}^qH)$ with $D_L = \exp H_L$. In particular,

$$^0D_L = \begin{cases} \exp\left(\frac{1}{\bar{h}}\sum_{i\in I} S_i h_i \log|P_i|\right) & \text{if } h>0 \\ 0 & \text{otherwise,}\end{cases}$$

$$^1D_L = \begin{cases} \exp\left(-\frac{1}{\bar{h}}\sum_{i\in I} h_i \sum_{b\in B_i} s_b \log\frac{s_b}{S_i}\right) & \text{if } h>0 \\ 0 & \text{otherwise.}\end{cases}$$

We can interpret ${}^0D_L$ as the average tree width or, more precisely, as the geometric mean number of branches counted across the tree. In an evolutionary tree where branch lengths correspond to elapsed time, ${}^0D_L$ equates to average richness across time, excluding extinct lineages. For $q>0$, ${}^qD_L$ can be interpreted as the effective number of counted nodes maximally distant from the root or – because all maximally distant counted nodes must be leaves – as the effective number of maximally distant leaves. In biological terms, this corresponds to the effective number of extant types maximally distinct from the root type.

Second, we define

$$^qJ_L := M_{long,1}(^qJ) = \begin{cases} \frac{1}{\bar{h}}\sum_{i\in I} S_i h_i {}^qJ(P_i) & \text{if } h>0 \\ 1 & \text{otherwise.}\end{cases} \tag{0.4}$$

Just as ${}^qJ$ measures the evenness of node sizes, so ${}^qJ_L$ measures the average evenness of branch sizes across the tree. If the tree is leafy and ultrametric then ${}^qJ_L = 1$ for $q>0$ if and only if the tree is fully symmetric. Hence, when applied to leafy ultrametric trees, ${}^qJ_L$ can be interpreted as a symmetry index (also known as a sound balance index (Mir et al., 2018)).

Figure 1b illustrates how ${}^0D_L,{}^1D_L$ and ${}^1J_L$ (and other index values yet to be defined) vary with branch length λ for the three-leaf tree of Figure 1a.

Definition of the node-wise mean: first special case

In the special case in which all branches have the same length l, we can obtain a node-wise mean by calculating an index value for each node, based on the node's child branch sizes, and then taking a weighted average of these node index values. We previously used this approach to define new tree balance indices (Lemant et al., 2022).

Let V denote the set of all internal nodes, excluding nodes with only zero-sized descendants. Let C_i denote the subtree containing only i and its children. For $i \in V$ and $b \in C_i$, let s_b denote the size of b and define

$$S_i = \sum_{b \in C_i} s_b \in (0, 1].$$

Then $S_i = 1$ for all nodes i if and only if the tree is a leafy piecewise star tree. It follows that

$$\sum_{i \in V} S_i l = \bar{h}.$$

Now for each $b \in C_i$, define the proportional branch size $p_b := s_b / S_i$ and let $P_i := \{p_b : b \in C_i, p_b > 0\}$. We then define the node-wise mean of order r of index F as the weighted power mean of the F values assigned to the nodes:

$$M_{node,r}(F)(T; w) := \begin{cases} \left(\dfrac{\sum_{i \in V(T)} w_i [F(P_i)]^r}{\sum_{i \in V(T)} w_i} \right)^{\frac{1}{r}}, & \text{if } h > 0 \\ F(\emptyset) & \text{otherwise,} \end{cases} \tag{0.5}$$

where the weight $w > 0$ is a function of i that remains to be specified.

Definition of the node-wise mean: second special case

In the case of a piecewise star tree with $h > 0$, we can set the index value of each internal node k as the longitudinal mean index value of the subtree C_k. We then have

$$\begin{aligned} M_{node,r,t}(F)(T; u, w) &= \left(\frac{\sum_{k \in V(T)} u_k [M_{long,r}(F)(C_k; w)]^t}{\sum_{k \in V(T)} u_k} \right)^{\frac{1}{t}} \\ &= \left(\frac{1}{\sum_{k \in V(T)} u_k} \sum_{k \in V(T)} u_k \left(\frac{\sum_{i \in I(T)} w_{ik} [F(P_{ik})]^r}{\sum_{i \in I(T)} w_{ik}} \right)^{\frac{t}{r}} \right)^{\frac{1}{t}}, \end{aligned} \tag{0.6}$$

where t is the exponent of the across-nodes power mean, $u_k > 0$ is the weight assigned to node k, P_{ik} contains the proportional sizes of all branch segments that belong to both subtree C_k and interval i, and $w_{ik} > 0$ is the weight assigned to k associated with interval i.

To keep our system internally consistent we would like, in the case of piecewise star trees, the node-wise mean of any index to be equal to the longitudinal mean of the same index. Comparing Equation 0.6 with the definition of the longitudinal mean (Equation 0.2), we see that the right-hand sides are equivalent if and only if three conditions hold:

$$r = t, \quad \sum_{i \in I(T)} w_{ik} = u_k, \quad \sum_{k \in V(T)} u_k = \sum_{i \in I(T)} w_i.$$

Under these conditions, summing index values across subtree intervals and then across nodes gives the same result as summing across tree intervals. We then have for any piecewise star tree T with $h > 0$,

$$M_{node,r}(F)(T; w) = \left(\frac{\sum_{k \in V(T)} \sum_{i \in I(T)} w_{ik} [F(P_{ik})]^r}{\sum_{i \in I(T)} w_i} \right)^{\frac{1}{r}}.$$

In the particular case $F = {}^qD$, the index value assigned to each node k (that is, the longitudinal mean index value of the subtree C_k) measures the diversity of the child branches of k. When C_k has m branches of equal length and size, the node diversity of k is m. In the case $m > 1$, as one branch length is reduced while all else is kept constant, the node diversity of k decreases continuously to $m - 1$. Decreasing instead the size of one branch has the same effect provided $q > 0$. Hence the diversity value assigned to each node can be interpreted as an effective outdegree, and the node-wise mean diversity can be interpreted as an average effective outdegree. When $q = 0$ the effective outdegree ignores branch sizes. As q increases, the effective outdegree gives less weight to branches of smaller size. We would like to retain this interpretation as we generalize the definition of the node-wise mean.

Definition of the node-wise mean: general case

In extending the definition to all rooted trees, we want to ensure that, as with the longitudinal mean, the node-wise mean changes continuously as we vary branch lengths. We illustrate this general issue with an example.

Example 0.2 Consider of a leafy ultrametric tree with six leaves such that the root has two descendant branches each of length λ, and both non-root internal nodes have three descendant branches, all of length $1 - \lambda$. When $\lambda = \frac{1}{2}$ (Figure 2a), it follows from our special-case definition (Equation 0.5) that the root has richness 2, the internal nodes each have richness 3, and the node-wise mean richness is intermediate between 2 and 3. As λ increases from $\frac{1}{2}$ to 1, the node richness values should remain unchanged but the root node richness should be given greater weight, so that the node-wise mean richness (which we will denote ${}^0D_N$) approaches 2 as $\lambda \to 1$ (Figure 2b).

At the other extreme, as $\lambda \to 0$, we must have ${}^0D_N \to 6$ (Figure 2c). And as λ decreases from $\frac{1}{2}$ to 0, we would like ${}^0D_N$ to increase continuously to 6. Given that the weight assigned to the root node richness should decrease as λ decreases, the only way to

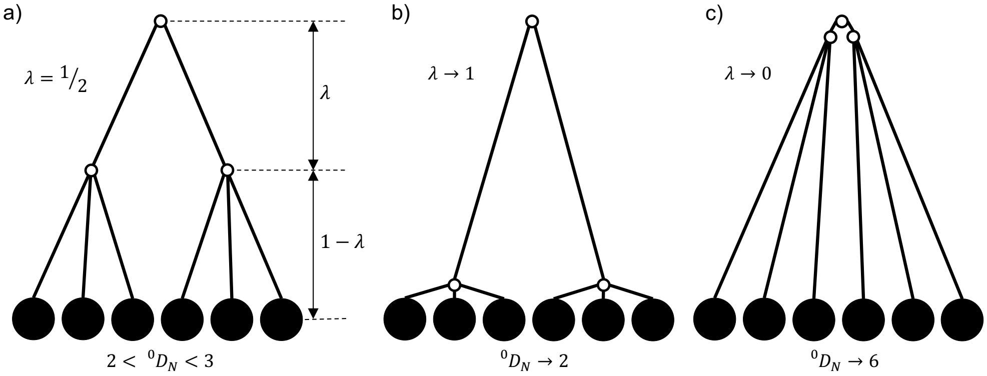


Fig. 2. a) The six-leaf tree considered in Example 0.2 with branch length $\lambda = \frac{1}{2}$. b) As $\lambda \to 1$, the tree approaches a two-leaf star tree. c) As $\lambda \to 0$, the tree approaches a six-leaf star tree.

achieve the required increase in ${}^0D_N$ is to increase the richness value assigned to each non-root internal node k. We can do this by making the richness value assigned to k depend not only on the child branches of k but also, to an increasing degree as λ decreases, on the other branches that run alongside the branches of k.

Generalizing from the example we conclude that, when the distance between node k and any ancestor j of k (in the example, the root) is less than the height of C_k (in the example, when $\lambda < \frac{1}{2}$), the index value assigned to k should depend not only on the branches of C_k (the child branches of k) but also on branch segments that descend from j and that coexist in transverse intervals with the branches of C_k. The weight assigned to k depends only on C_k but the index value assigned to k is a weighted average of index values across k and all ancestors of k.

To formalize this concept, we first define, for interval $i \in I$ and node $j \in V$,

$$S_{iT_j} = \sum_{b \in B_i \cap T_j} s_b \in [0,1], \quad S_{iC_j} = \sum_{b \in B_i \cap C_j} s_b \in [0,1],$$

where T_j is the subtree containing j and all its descendants. This implies

$$\sum_{i \in I} S_{iT_r} h_i = \sum_{i \in I} \sum_{j \in V} S_{iC_j} h_i = \bar{h},$$

where r is the root (and hence T_r is the entire tree). S_{iC_j} is a generalization of the S_i used in our previous definitions, whereas S_{iT_j} is a new concept. For each $b \in B_i \cap T_j$, let

$$p_b = \begin{cases} s_b/S_{iT_j} & \text{if } S_{iT_j} > 0 \\ 0 & \text{otherwise,} \end{cases}$$

and define $P_{ij} = \{p_b : b \in B_i \cap T_j, p_b > 0\}$. We then define the node-wise average as the triple power mean

$$M_{node,r,s,t}(F)(T;u,v,w) = \left(\frac{1}{\sum_{k\in V(T)} u_k} \sum_{k\in V(T)} u_k \left[\frac{1}{\sum_{j\in A_k} v_{jk}} \sum_{j\in A_k} v_{jk} \left(\frac{\sum_{i\in I(T)} w_{ik}[F(P_{ij})]^r}{\sum_{i\in I(T)} w_{ik}} \right)^{\frac{s}{r}} \right]^{\frac{t}{s}} \right)^{\frac{1}{t}},$$

where A_k is the set containing k and all ancestors of k, s is the exponent of the across-ancestors power mean, and v_{jk} are the ancestor weights. This expression is consistent with Equation 0.6 if and only if

$$t = s = r, \quad \sum_{j\in A_k} v_{jk} = u_k = \sum_{i\in I(T)} w_{ik}, \quad \sum_{k\in V(T)} u_k = \sum_{i\in I(T)} w_i. \tag{0.7}$$

We then arrive at a simpler general definition

$$M_{node,r}(F)(T;v,w) := \begin{cases} \left(\frac{1}{\sum_{k\in V(T)} u_k} \sum_{k\in V(T)} \sum_{j\in A_k} \frac{v_{jk}}{u_k} \sum_{i\in I(T)} w_{ik}[F(P_{ij})]^r \right)^{\frac{1}{r}} & \text{if } h > 0 \\ F(\emptyset) & \text{otherwise.} \end{cases}$$

Integral forms of the node-wise and longitudinal means

Since our preferred ancestor weights are best expressed as integrals, we will find it useful to define the longitudinal and node-wise means even more generally by integrating over depths instead of summing over intervals. Suppose we assign a non-negative density $f_b(x)$ to every branch b at every depth x, with $f_b(x) = 0$ for every x at which b is absent. Define the tree height $h := \max\{x : f_b(x) > 0, b \in B\}$, where B is the set of all branches. We can then define branch size s_b as the non-increasing function of depth x:

$$\bar{h} := \sum_{b\in B} \int_0^h f_b(x)\,dx, \quad s_b(x) := \begin{cases} \frac{1}{\bar{h}} \sum_{b\in G_b} \int_x^h f_b(t)\,dt & \text{if } h > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where G_b is the set containing b and all branches that descend from b. Let

$$S_{T_j}(x) := \sum_{b\in B_j} s_b(x) \in [0,1].$$

For each $b \in B_j$, define the proportional branch size

$$p_{bj}(x) := \begin{cases} s_b(x)/S_{T_j}(x) & \text{if } S_{T_j}(x) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Let $P_j(x) := \{p_{bj}(x) : b \in B_j, p_{bj}(x) > 0\}$. We then define the node-wise mean of an index F as

$$M_{node,r}(F)(T; v, w) := \begin{cases} \left(\dfrac{1}{\sum\limits_{k \in V(T)} u_k} \sum\limits_{k \in V(T)} \sum\limits_{j \in A_k} \dfrac{v_{jk}}{u_k} \displaystyle\int_0^h w_k(x) [F(P_j(x))]^r \, dx \right)^{\frac{1}{r}} & \text{if } h > 0 \\ F(\emptyset) & \text{otherwise,} \end{cases}$$

where $w_k(x)$ is the weight assigned to node k at depth x, and

$$u_k = \int_0^h w_k(x) \, dx.$$

The longitudinal mean can similarly be defined in terms of integrals as

$$M_{long,r}(F)(T; w) := \begin{cases} \left(\dfrac{\int_0^h w(x)[F(P(x))]^r \, dx}{\int_0^h w(x) \, dx} \right)^{\frac{1}{r}} & \text{if } h > 0 \\ F(\emptyset) & \text{otherwise,} \end{cases}$$

where $P(x) := \{p_b(x) : b \in B, p_b(x) > 0\}$,

$$p_b(x) := \begin{cases} s_b(x)/S(x) & \text{if } S(x) > 0 \\ 0 & \text{otherwise,} \end{cases} \qquad S(x) := \sum_{b \in B} s_b(x) = \sum_{j \in V} \sum_{b \in B_j} s_b(x).$$

Our previous definitions are included as special cases in which the branch density is zero except at each counted node, where it is equal to the node size. In an evolutionary tree, branch density corresponds to population size, and branch size corresponds to number of extant descendants. Although it is beyond the scope of the current manuscript, we note that the integral forms would permit us to apply our indices to a more general class of tree, such that the size of any branch is allowed to vary along its length.

New node-wise mean indices

To define new tree indices as node-wise means of ${}^qD$ and ${}^qJ$, we first set $w_k = S_{C_k}$, where

$$S_{C_k}(x) := \sum_{b \in C_k} s_b(x),$$

and we define the normalization factor

$$\bar{h}_{C_j} := \int_0^h S_{C_j}(x) \, dx, \implies \bar{h} = \sum_{j \in V} \bar{h}_{C_j} = \int_0^h S(x) \, dx.$$

Let d_k denote the depth of node k and let $d_{jk} = d_k - d_j$ denote the distance from j to k. Let j' denote the parent of node j. The ancestor weight function v should have three properties. First, as an assumption of our general definition (Equation 0.7),

$$\sum_{j \in A_k} v_{jk} = u_k = \int_0^h w_k(x)\, dx.$$

Second, v_{jk} should decrease as $d_{j'j}$ decreases. Third, v_{jk} should increase as the overlap between C_j and C_k increases. A simple way to satisfy all three conditions is to set

$$v_{jk} = \int_{\alpha_{jk}}^{\beta_{jk}} S_{C_k}(x)\, dx,$$

where $\alpha_{jk} := d_k + d_{jk}$ and

$$\beta_{jk} := \begin{cases} \alpha_{jk} + d_{j'j} & \text{if } j \text{ is not the root} \\ \infty & \text{otherwise.} \end{cases}$$

Given the above choices of w and v, we define the node-wise mean diversity of order q as

$$^{q}D_N := M_{node,0}(^{q}D). \tag{0.8}$$

This is equivalent to $^{q}H_N = M_{node,1}(^{q}H)$ with $^{q}D_N = \exp {}^{q}H_N$. In particular,

$$^{0}D_N = \begin{cases} \exp\left(\frac{1}{\bar{h}} \sum_{k \in V} \frac{1}{\bar{h}_{C_k}} \sum_{j \in A_k} v_{jk} \int_0^h S_{C_k}(x) \log|P_j(x)|\, dx\right) & \text{if } h > 0 \\ 0 & \text{otherwise,} \end{cases}$$

$$^{1}D_N = \begin{cases} \exp\left(\frac{1}{\bar{h}} \sum_{k \in V} \frac{1}{\bar{h}_{C_k}} \sum_{j \in A_k} v_{jk} \int_0^h S_{C_k}(x)^1 H(P_j(x))\, dx\right) & \text{if } h > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where

$$^{1}H(P_j(x)) = -\sum_{b \in B_j} \frac{s_b(x)}{S_{T_j}(x)} \log \frac{s_b(x)}{S_{T_j}(x)}.$$

As previously explained, we can interpret $^{q}D_N$ as an average effective outdegree (branching factor in computer science) that accounts for branch lengths only ($q = 0$) or for both branch lengths and branch sizes ($q > 0$). Less formally, $^{q}D_N$ quantifies the bushiness of the tree.

With the same w and v, we define the universal tree balance $^{q}J_N$ as

$$^{q}J_N := M_{node,0}(^{q}J) = \begin{cases} \frac{1}{\bar{h}} \sum_{k \in V} \frac{1}{\bar{h}_{C_k}} \sum_{j \in A_k} v_{jk} \int_0^h S_{C_k}(x)^q J(P_j(x))\, dx & \text{if } h > 0 \\ 1 & \text{otherwise.} \end{cases} \tag{0.9}$$

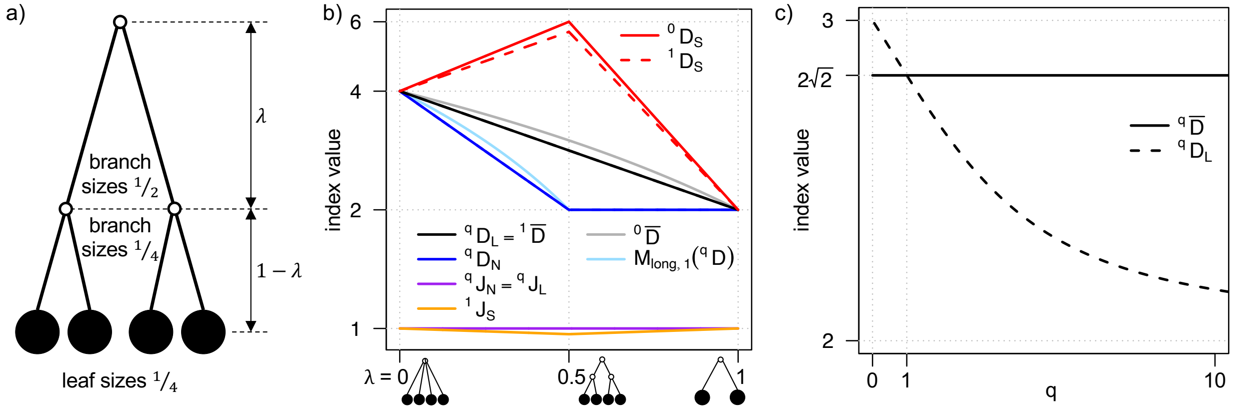


Fig. 3. a) The four-leaf tree considered in Examples 0.3 and 0.6. b) Index values versus branch length λ for the tree of Example 0.6. Curves for indices with parameter q are independent of the value of $q \geqslant 0$. The y-axis is log-transformed so that the curves for all diversity indices except ${}^{0}\bar{D}$ and $M_{long,1}({}^{q}D)$ appear piecewise linear. c) ${}^{q}\bar{D}$ and ${}^{q}D_L$ values for the four-leaf tree considered in Example 0.6, for varied q with $\lambda = \frac{1}{2}$.

In the case of uniform branch lengths, this definition simplifies to

$$ {}^{q}J_N = \frac{1}{\bar{h}} \sum_{i \in V} S_i^q J(P_i), $$

where S_i and P_i are defined as in Equation 0.5. This means that for trees with uniform branch lengths, ${}^{q}J_N$ is identical to our previous definition of the tree balance index J^q (Lemant et al., 2022), excepting one important difference. Whereas our prior index assigns a balance score of zero to any node that has outdegree 1, the above definition instead assigns a balance score of one. Therefore linear trees are considered maximally unbalanced according to J^q but maximally balanced according to ${}^{q}J_N$. This difference ensures that all our new evenness indices have consistent definitions and interpretations.

Example 0.3 Consider the perfectly balanced, bifurcating, leafy tree with four leaves and branch lengths λ (upper two branches) and $1 - \lambda$ (lower four branches), as shown in Figure 3a. For all $q \geqslant 0$, if $\lambda \geqslant \frac{1}{2}$ then ${}^{q}D_N = 2$, and otherwise ${}^{q}D_N = 4^{1-\lambda}$, as shown in Figure 3b (dark blue curve). A step-by-step derivation is in the Appendix.

The above example illustrates that, for leafy ultrametric trees, the node-wise mean diversity, like the longitudinal mean diversity, is a piecewise exponential function of branch lengths. Equivalently, the entropy indices are piecewise linear. This property depends on our defining the ancestor weight function v_{jk} as an integral of S_{C_k}. Because S_{C_k} is a step function, the integrals in all our node-wise mean index definitions are simply sums of areas of rectangles, and the widths of these rectangles are linear functions of branch lengths. Our definitions are designed so that, although Equations 0.8 and 0.9 might appear complicated, in practice they produce relatively simple expressions.

The star mean and new star mean indices

Like the longitudinal and node-wise means, the star mean is based on branch sizes. Unlike those other two means, but in common with the node-size indices ${}^qD$ and ${}^qJ$, the star mean ignores tree topology. The idea is that, in effect, we rearrange the tree by reattaching all branches to the root to form a star tree, while retaining branch sizes and lengths, and then calculate the longitudinal (equivalently node-wise) mean index value of the star tree. For index F and tree T, we define the star mean of order r of F such that

$$M_{star,r}(F)(T;w^*) := \begin{cases} \left(\dfrac{\int_0^h w^*(x)[F(P^*(x))]^r \, dx}{\int_0^h w^*(x)\, dx} \right)^{\frac{1}{r}} & \text{if } h > 0 \\ F(\emptyset) & \text{otherwise,} \end{cases} \tag{0.10}$$

where $P^*(x) := \{p_b^*(x) : b \in B, p_b^*(x) > 0\}$,

$$p_b^*(x) := \begin{cases} s_b(x + d_b)/S^*(x) & \text{if } S^*(x) > 0 \\ 0 & \text{otherwise,} \end{cases} \qquad S^*(x) := \sum_{b \in B} s_b(x + d_b),$$

and d_b is the depth of the parent node of branch b. Note that

$$\int_0^h S^*(x) = \int_0^h S(x) = \bar{h}.$$

With $w^* = S^*$, we define the star mean diversity of order q as

$${}^qD_S := M_{star,0}({}^qD), \tag{0.11}$$

which is equivalent to ${}^qH_S = M_{star,1}({}^qH)$ with ${}^qD_S = \exp {}^qH_S$. In particular,

$${}^0D_S = \begin{cases} \exp\left(\dfrac{1}{\bar{h}} \displaystyle\int_0^h S^*(x) \log|P^*(x)| \, dx \right) & \text{if } h > 0 \\ F(\emptyset) & \text{otherwise,} \end{cases}$$

$${}^1D_S = \begin{cases} \exp\left(\dfrac{1}{\bar{h}} \displaystyle\int_0^h S^*(x) {}^1H(P^*(x)) \, dx \right) & \text{if } h > 0 \\ F(\emptyset) & \text{otherwise.} \end{cases}$$

${}^qD_S$ quantifies the effective number of branches in the tree, either accounting for branch lengths only ($q = 0$) or for both branch lengths and branch sizes ($q > 0$). Because every non-root node has exactly one parent branch, and because ${}^0D_S$ accounts for branch lengths but not sizes, ${}^0D_S$ can also be interpreted as an effective number of non-root nodes. We also define an index that quantifies the evenness of all branch sizes:

$${}^qJ_S := M_{star,0}({}^qJ) = \begin{cases} \dfrac{1}{\bar{h}} \displaystyle\int_0^h S^*(x) {}^qJ(P^*(x)) \, dx & \text{if } h > 0 \\ 1 & \text{otherwise.} \end{cases} \tag{0.12}$$

Figures 1b and 3b illustrate how ${}^0D_S, {}^1 D_S$ and ${}^1J_S$ values vary with branch lengths for three- and four-leaf trees.

Non-normalized indices

Although our focus is on indices that describe shape, rather than size, we note that every longitudinal, node-wise, or star mean diversity index can be converted into a non-normalized diversity index simply by omitting the normalization factor. Such indices are useful in applications where the unit of branch length should be retained, such as when assessing loss of richness or diversity due to the removal of a node. In particular, we will find it useful to define the non-normalized entropy index

$$H'_P := -\sum_{b\in B} l_b p_b \log p_b = \sum_{i\in I} h_i \log {}^1D(P_i). \tag{0.13}$$

Results

${}^qD_L$ improves on prior indices for non-ultrametric trees

Our indices ${}^0D_L$ and ${}^1D_L$ are similar to well-known pre-existing indices but with important improvements (Table 3). The phylogenetic diversity of Faith (1992) – which is popular among conservation biologists – is defined as

$$PD := \sum_{b\in B} l_b.$$

Phylogenetic entropy (Allen et al., 2009) – a previous generalization of Shannon's entropy – is defined in our notation as

$$H_P := -\sum_{b\in B} l_b s_b \log s_b.$$

Chao et al. (2010) defined normalized versions of these indices that can be written as

$${}^0\bar{D} = \frac{PD}{\bar{h}} = \frac{1}{\bar{h}}\sum_{i\in I} h_i|B_i| = \frac{\sum_{i\in I} h_i {}^0D(Q_i)}{\sum_{i\in I} S_i h_i},$$

$${}^1\bar{D} = \exp\left(\frac{H_P}{\bar{h}}\right) = \exp\left(-\frac{1}{\bar{h}}\sum_{i\in I} h_i \sum_{b\in B_i} s_b \log s_b\right) = \exp\left(\frac{\sum_{i\in I} h_i \log {}^1D(Q_i)}{\sum_{i\in I} S_i h_i}\right),$$

where $Q_i = \{s_b : b \in B_i\}$.

A first problem with these definitions is that, for non-ultrametric trees, phylogenetic entropy lacks a clear interpretation. This issue is due to H_P being defined in terms of sets of branch sizes Q_i instead of sets of within-interval proportional branch sizes $P_i = \{p_b = s_b/S_i : b \in B_i\}$, as illustrated by the following example.

Example 0.4 Consider the three-node tree with leaf sizes p and $1 - p$, and leaf depths

Prior index	Proposed replacement	Equation	Advantages of replacement
Allen et al's H_P	H'_P	0.13	Interpretable for non-ultrametric trees
Chao et al's ${}^q\bar{D}$	${}^qD_L$	0.3	Bounded and interpretable for non-ultrametric trees; more self-consistent; more intuitive for $q > 1$
All prior tree balance and imbalance indices	${}^qJ_N$	0.9	Defined for all rooted trees; can meaningfully compare any pair of trees; accounts for node sizes and branch lengths

Table 3. Advantages of using our indices instead of previously-defined indices.

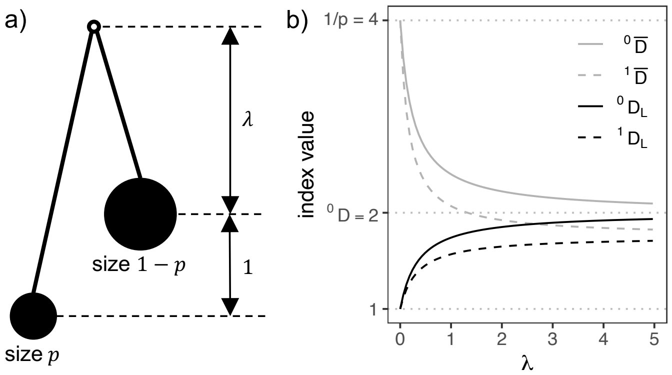


Fig. 4. a) The two-leaf tree considered in Examples 0.4 and 0.5. b) Index values for the tree of Example 0.5 with $p = \frac{1}{4}$. As branch length λ decreases, the previously defined indices ${}^0\bar{D}$ and ${}^1\bar{D}$ (grey curves) increase monotonically until both ${}^0\bar{D} > {}^0D$ and ${}^1\bar{D} > {}^0D$. In contrast, our new indices ${}^0D_L$ and ${}^1D_L$ (black curves) decrease monotonically as l decreases, with ${}^0D_L < {}^0D$ and ${}^1D_L < {}^0D$ for all values of λ.

$1 + \lambda$ and λ, respectively (Figure 4a). For this tree, as $\lambda \to 0$,

$$PD = 1 + 2\lambda \to 1,$$
$$\exp H_P = \exp[-(1+\lambda)p \log p - \lambda(1-p)\log(1-p)] \to p^{-p}.$$

Therefore PD behaves as expected but, except when $p = 0$ or $p = 1$, $\exp H_P$ approaches a limit greater than 1. Hence $\exp H_P$ (which is supposed to be a measure of diversity) is greater than PD (a measure of richness). Moreover, whereas we expect diversity to be maximal when node sizes are equal, $\exp H_P$ is maximal when the node sizes are unequal (specifically, $\exp H_P \approx 1.44$ when $p = e^{-1} \approx 0.37$). If we instead use our index H'_P (Equation 0.13) then we obtain

$$\exp H'_P = \exp[\lambda(-p\log p - (1-p)\log(1-p))] \to 1,$$

as we would expect.

A second problem is that if the tree is not ultrametric then ${}^0\bar{D}$ and ${}^1\bar{D}$ do not correspond to weighted means. If and only if the tree is leafy and ultrametric, $S_i = 1$ and

$Q_i = P_i$ for all i and so

$$^0\bar{D} = M_{long,1}(^0D), \quad ^1\bar{D} = M_{long,0}(^1D) = {}^1D_L,$$

with $w_i = S_i h_i = h_i$ in both cases. Otherwise, the numerator weights h_i are unequal to the denominator weights $S_i h_i$. As previously noted (Chao et al., 2010; Leinster and Cobbold, 2012), this implies that $^0\bar{D}$ and $^1\bar{D}$ can take values exceeding the number of counted nodes when applied to non-ultrametric (or non-leafy) trees. Therefore these normalized indices lack a universal interpretation in terms of effective numbers of counted nodes (or extant types) (Leinster and Cobbold, 2012).

We avoid both problems by defining our richness and diversity indices as weighted means of the within-interval proportional branch sizes in all cases. As illustrated by the following example, the differences between ^0D_L and $^0\bar{D}$ and between ^1D_L and $^1\bar{D}$ are generally unbounded and can be relatively large even when branch sizes and node sizes are not very unequal.

Example 0.5 Consider the three-node tree of Figure 4a with $p < \frac{1}{2}$. We have $^0D = 2, \bar{h} = p + \lambda$, and

$$^0\bar{D} = \frac{(1+\lambda)+\lambda}{p+\lambda} > \frac{1+2\lambda}{\frac{1}{2}+\lambda} = 2,$$
$$^1\bar{D} = \exp\left(\frac{-(1+\lambda)p\log p - \lambda(1-p)\log(1-p)}{p+\lambda}\right) \to \frac{1}{p} > 2 \text{ as } \lambda \to 0.$$

It follows that $^0\bar{D} > {}^0D$ for all λ, and we can choose λ sufficiently small such that also $^1\bar{D} > {}^0D$ (Figure 4b, grey curves). For the same three-node tree, our new indices are instead

$$^0D_L = \exp\left(\frac{\lambda\log 2}{p+\lambda}\right) < \exp\left(\frac{\lambda\log 2}{\lambda}\right) = 2,$$
$$^1D_L = \exp\left(\frac{\lambda(-p\log p - (1-p)\log(1-p))}{p+\lambda}\right) < \exp\left(\frac{\lambda\log 2}{\lambda}\right) = 2.$$

Therefore $^0D_L < {}^0D$ and $^1D_L < {}^0D$ for all $\lambda \geqslant 0$, as we would expect (Figure 4b, black curves). As $\lambda \to 0$, both ^0D_L and ^1D_L approach 1, consistent with the fact that the tree has exactly one non-root node when $\lambda = 0$. As $\lambda \to \infty$, the tree becomes increasingly close to being an ultrametric star tree, and hence $^0D_L \to {}^0\bar{D}$ and $^1D_L \to {}^1\bar{D}$ (convergence between dashed curves and between solid curves in Figure 4b).

^qD_L is more self-consistent and intuitive than the $^q\bar{D}$ of Chao et al. (2010)

Additional problems with the $^q\bar{D}$ indices of Chao et al. (2010) are that they are not self-consistent, and that they have counter-intuitive properties when $q > 1$. The general definition can be expressed as

$${}^q\bar{D} = \left(\frac{1}{\bar{\bar{h}}}\sum_{i\in I} h_i \sum_{b\in B_i} s_b^q\right)^{\frac{1}{1-q}},$$

which can be restructured as

$${}^q\bar{D} = \left(\frac{1}{\bar{\bar{h}}}\sum_{i\in I} h_i \left[\left(\sum_{b\in B_i} s_b^q\right)^{\frac{1}{1-q}}\right]^{1-q}\right)^{\frac{1}{1-q}} = \left(\frac{1}{\bar{\bar{h}}}\sum_{i\in I} h_i [{}^qD(Q_i)]^{1-q}\right)^{\frac{1}{1-q}}.$$

Hence for leafy ultrametric trees we have

$${}^q\bar{D} = M_{long,1-q}({}^qD),$$

with $w_i = h_i = S_i h_i$. We have thus shown that, in the case of leafy ultrametric trees, every ${}^q\bar{D}$ can be expressed as a weighted mean of within-interval diversities. But ${}^0\bar{D}$ is the weighted arithmetic mean, ${}^1\bar{D}$ is the weighted geometric mean, and in general ${}^q\bar{D}$ is the weighted power mean of exponent $1-q$. One consequence is that, for ultrametric trees in which every transverse interval contains branches of equal size, the set of within-interval values will be the same for every q value but the ${}^q\bar{D}$ values will be different. Moreover, as q becomes larger, ${}^q\bar{D}$ increasingly gives larger weight to *smaller* within-interval diversities. As $q \to \infty$, the ${}^qD$ value assigned to each interval approaches the reciprocal of the maximum branch size within the interval. Counter-intuitively, ${}^q\bar{D}$ approaches the *minimum* of these within-interval ${}^qD$ values.

These peculiar properties of ${}^q\bar{D}$ are unnecessary and have no obvious advantages. The Hill numbers ${}^qD$, which are used to assign a diversity value to each interval, necessarily relate to different types of weighted mean (Equation 0.1). But the method of averaging *between* intervals need not depend on the method of calculating diversity *within* intervals. Every Hill number ${}^qD$ can be extended to account for tree shape using the weighted arithmetic mean, the weighted geometric mean, or any other weighted power mean of the within-interval diversities by varying exponent r of the longitudinal mean diversity index

$$M_{long,r}({}^qD) = \left(\frac{1}{\bar{\bar{h}}}\sum_{i\in I} S_i h_i [{}^qD(Q_i)]^r\right)^{\frac{1}{r}}.$$

The same choice exists when defining node-wise means and star means. To avoid incompatibilities within our system, we define all our diversity indices as weighted geometric means ($r \to 0$). The following example illustrates the problem and our solution.

Example 0.6 Consider again the four-leaf tree of Example 0.3 (Figure 3a). The

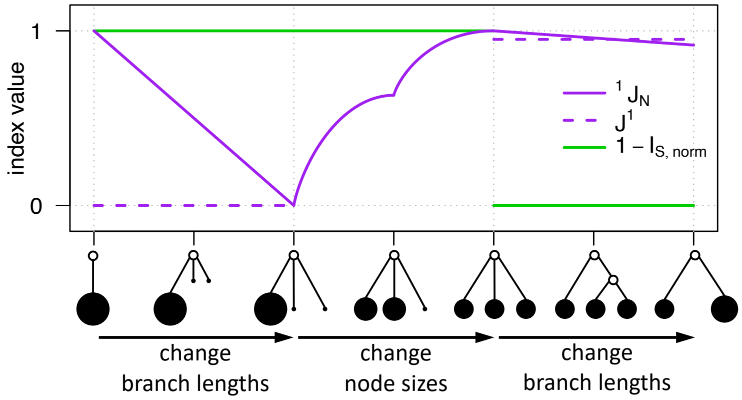


Fig. 5. Values of three tree balance indices for a tree undergoing continuous changes. J^1 is the index introduced by Lemant et al. (2022), which is equal to ^1J_N in the central third of the plot. $I_{S,norm}$ is the normalized Sackin index, which is undefined for the leftmost, linear tree. We plot $1 - I_{S,norm}$ for fair comparison because $I_{S,norm}$ is an imbalance index whereas J^1 and ^1J_N are balance indices. The normalized Colless index is equal to $I_{S,norm}$ in the rightmost third of the plot and is otherwise undefined. The normalized total cophenetic index is equal to $I_{S,norm}$ throughout the plot.

longitudinal mean diversity values assigned to this tree are

$$\begin{aligned} {}^qD_L = {}^1\bar{D} = \exp(\lambda \log 2 + (1-\lambda)\log 4) = 2^{2-\lambda} \text{ for all } q \geqslant 0, \\ \text{and } {}^0\bar{D} = 2h + 4(1-\lambda) = 2(2-\lambda), \end{aligned}$$

which are unequal except when $\lambda = 0$ or $\lambda = 1$ (Figure 3b, black and grey curves). In particular, in the case of uniform branch lengths ($\lambda = \frac{1}{2}$), we find ${}^1\bar{D} = 2\sqrt{2} \approx 2.83$ and ${}^0\bar{D} = 3$ (Figure 3c, dashed curve). As derived in Example 0.3, the node-wise mean diversity for this tree is

$${}^qD_N = \begin{cases} 4^{1-\lambda} & \text{if } \lambda < \frac{1}{2} \\ 2 & \text{otherwise,} \end{cases}$$

for all $q \geqslant 0$. Choosing the arithmetic mean instead of the geometric mean would instead give

$$M_{long,1}({}^qD) = \begin{cases} 4(1-\lambda) & \text{if } \lambda < \frac{1}{2} \\ 2 & \text{otherwise.} \end{cases}$$

${}^qD_N := M_{long,0}({}^qD) \neq M_{long,1}({}^qD)$ for all $q \geqslant 0$ and all λ with $0 < \lambda < \frac{1}{2}$ (Figure 3b, dark blue and pale blue curves). As $q \to \infty$, ${}^q\bar{D} \to 2$ (Figure 3c, dashed curve), while ${}^qD_L$ remains constant (Figure 3c, solid line).

${}^qJ_N$ improves on all prior tree balance and imbalance indices

As previously explained (Lemant et al., 2022) and as summarized in Tables 1 and 3, conventional tree balance and imbalance indices including Sackin's index, Colless' index, the total cophenetic index, and others (reviewed by Fischer et al. (2021)) have important shortcomings. In the first place, these indices account for neither node sizes nor branch lengths. This means, for example, that these indices consider all star trees maximally balanced and all caterpillar trees maximally imbalanced, even as the relative sizes of some nodes or the relative lengths of some branches approach zero (Figure 5, green lines). The tree balance index J^q defined by Lemant et al. (2022) varies continuously with changing node sizes but is independent of branch lengths (Figure 5, dashed purple curves). ${}^qJ_N$ improves on J^q by also varying continuously with branch lengths (Figure 5, solid purple curves).

Lemant et al. (2022) further showed that, even when restricted to the tree types on which conventional tree balance indices are defined, and even when all node sizes are equal, J^q enables a more meaningful comparison of trees with different degree distributions or different numbers of leaves. For example, when applied to leafy caterpillar trees with uniform branch lengths and uniform node sizes, J^q considers long trees (those with many leaves) to be less balanced than short ones, whereas conventional indices consider them equally imbalanced. ${}^qJ_N$, as an extension of J^q, shares this useful property.

Inequalities between indices

Choosing self-consistent definitions ensures that our diversity indices are related by simple sets of inequalities, which formalize and generalize the results of previous sections (Figure 6a). Hill (1973) showed that ${}^qD \geqslant {}^rD$ for all $r \geqslant q \geqslant 0$. Because ${}^qD_L$ and ${}^qD_N$ are geometric weighted means of ${}^qD$ values with weights independent of q, it follows that they obey corresponding inequalities:

Property 0.1 For all rooted trees, ${}^qD_L \geqslant {}^rD_L$ and ${}^qD_N \geqslant {}^rD_N$ for all $r \geqslant q \geqslant 0$.

Additional inequalities exist between different types of diversity index.

Proposition 1 For all rooted trees, ${}^0D \geqslant {}^qD_L$ for all $q \geqslant 0$. For all leafy ultrametric trees, but not for all rooted trees, ${}^qD \geqslant {}^qD_L$ for all $q \geqslant 0$.

Proof. See Appendix. □

Informally, the reason why the second inequality in Proposition 1 applies only to leafy ultrametric trees is that ${}^qD_L$, unlike ${}^qD$, is independent of the size of the root node (and any node arbitrarily close to the root).

Proposition 2 For all rooted trees, ${}^qD_L \geqslant {}^qD_N$ for all $q \geqslant 0$.

Proof. For every node k, every $j \in A_k$, and at every depth x, we have $P_j(x) \subseteq P(x)$ and so

${}^qH(P_j(x)) \leqslant {}^qH(P(x))$ for all $q \geqslant 0$. Hence

$$\begin{aligned} {}^qD_N &:= \exp\left(\frac{1}{\overline{h}}\sum_{k\in V}\sum_{j\in A_k}\frac{v_{jk}}{u_k}\int_0^h S_{C_k}(x)^qH(P_j(x))\,dx\right) \\ &\leqslant \exp\left(\frac{1}{\overline{h}}\sum_{k\in V}\max_{j\in A_k}\int_0^h S_{C_k}(x)^qH(P_j(x))\,dx\right) \leqslant \exp\left(\frac{1}{\overline{h}}\sum_{k\in V}\int_0^h S_{C_k}(x)^qH(P(x))\,dx\right) \\ &= \exp\left(\frac{1}{\overline{h}}\int_0^h {}^qH(P(x))\sum_{k\in V}S_{C_k}(x)\,dx\right) = \exp\left(\frac{1}{\overline{h}}\int_0^h {}^qH(P(x))S(x)\,dx\right) = {}^qD_L. \end{aligned}$$

□

No such result applies to the evenness indices.

PROPOSITION 3 For $q > 0$, no single ordering of ${}^qJ, {}^qJ_L$ and ${}^qJ_N$ applies to all leafy ultrametric trees.

Proof. The top left panel of Figure 6b shows a leafy ultrametric tree for which ${}^qJ = 1, {}^qJ_L < 1$ and ${}^qJ_N < 1$. The third panel in the top row of Figure 6b shows a leafy ultrametric tree for which ${}^qJ < 1, {}^qJ_L < 1$ and ${}^qJ_N = 1$. Now consider the four-leaf, bifurcating, leafy ultrametric tree with uniform branch lengths, such that the sizes of each pair of sibling leaves are ϵ and $\frac{1}{2} - \epsilon$ (Figure 6c). As $\epsilon \to 0$, we have ${}^qJ \to \frac{1}{2}, {}^qJ_L \to \frac{3}{4}$ and ${}^qJ_N \to \frac{1}{2}$. The different orderings of ${}^qJ, {}^qJ_L$ and ${}^qJ_N$ for three trees are inconsistent with any universal ordering. □

Special cases

Our consistent definitions further yield numerous simple equations that unite our indices in special cases. To simplify the statement of these results, we will assume that all branch sizes are greater than zero. This assumption implies no loss of generality because our index definitions are invariant to the addition or removal of subtrees containing only zero-sized branches (which in an evolutionary tree correspond to extinct lineages). The properties in this section hold for all $q, r \geqslant 0$.

We begin with cases in which diversities based on the same type of average but with different q values are equal. These first four properties, which are illustrated by simple examples in the top row of Figure 6b, follow immediately from the definitions.

Property 0.2 ${}^qD = {}^rD \Leftrightarrow {}^qJ = 1$ if and only if all counted nodes have equal size.

Property 0.3 ${}^qD_L = {}^rD_L \Leftrightarrow {}^qJ_L = 1$ if and only if the branch sizes at every depth are equal. This also implies ${}^qD_N = {}^rD_N$ and ${}^qJ_N = 1$.

Property 0.4 ${}^qD_N = {}^rD_N \Leftrightarrow {}^qJ_N = 1$ if and only if every internal node's child branches

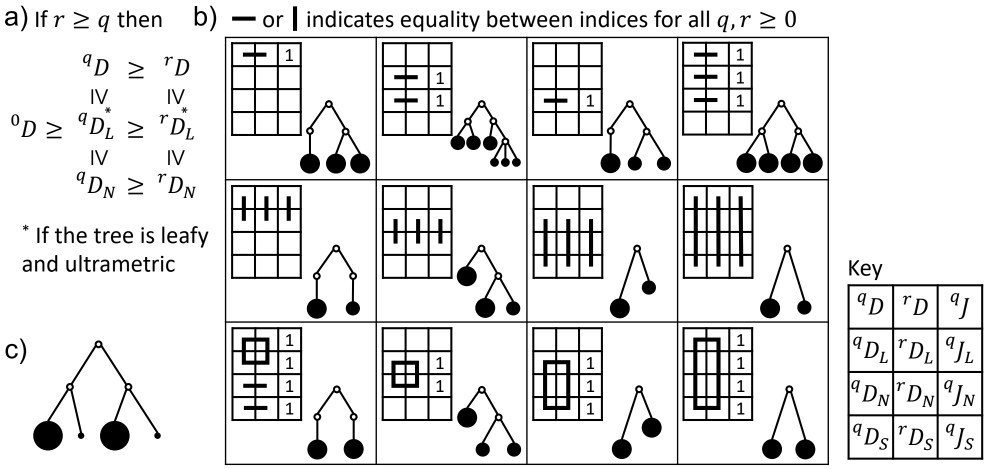

a) If $r \geq q$ then
${}^qD \geq {}^rD$
IV IV
${}^0D \geq {}^qD_L^* \geq {}^rD_L^*$
IV IV
${}^qD_N \geq {}^rD_N$
* If the tree is leafy and ultrametric
b) — or | indicates equality between indices for all $q, r \geq 0$
1
Key
${}^qD$ ${}^rD$ ${}^qJ$
${}^qD_L$ ${}^rD_L$ ${}^qJ_L$
${}^qD_N$ ${}^rD_N$ ${}^qJ_N$
${}^qD_S$ ${}^rD_S$ ${}^qJ_S$
c)

Fig. 6. a) Inequalities between diversity indices for all $q \geqslant 0$ and all $r \geqslant q$. b) Examples of leafy trees with uniform branch lengths for which various index values are equal for all $q, r \geqslant 0$. The top left corner of each panel contains a grid, whose twelve squares correspond to the twelve indices shown in the key. A line connecting two grid squares indicates that the corresponding indices are equal for the tree shown in the panel. Instances where evenness indices are equal to 1 are indicated in the third grid column. c) A tree for which ${}^qJ_L > {}^qJ$ and ${}^qJ_L > {}^qJ_N$.

have equal size.

Property 0.5 $({}^qD = {}^rD$ and ${}^qD_L = {}^rD_L) \Leftrightarrow ({}^qJ = 1$ and ${}^qJ_N = 1)$ if and only if the branch sizes at every depth are equal and all node sizes are equal. This implies that the tree is ultrametric and perfectly symmetric, and that ${}^qD_N = {}^rD_N$ and ${}^qJ_N = 1$.

In other special cases, we find equality among diversities of different types but with equal q values. Again, these properties are directly implied by the definitions. Simple examples are shown in the middle row of Figure 6b.

Property 0.6 ${}^qD_L = {}^qD$ if and only if the tree is a leafy ultrametric tree in which no non-root node has outdegree greater than 1. This also implies ${}^qJ_L = {}^qJ$.

Property 0.7 ${}^qD_N = {}^qD_L$ if and only if the tree is a piecewise star tree. This also implies ${}^qJ_N = {}^qJ_L$.

Property 0.8 ${}^qD_S = {}^qD_N = {}^qD_L$ if and only if the tree is a star tree. This also implies ${}^qJ_S = {}^qJ_N = {}^qJ_L$.

Property 0.9 ${}^qD_S = {}^qD_N = {}^qD_L = {}^qD$ if and only if the tree is a leafy ultrametric star tree. This also implies ${}^qJ_S = {}^qJ_N = {}^qJ_L = {}^qJ$.

It follows that equality both within and between types applies under more restrictive conditions, as illustrated in the bottom row of Figure 6b:

Property 0.10 ${}^qD_L = {}^rD$ if and only if the tree is a leafy ultrametric tree with equally sized leaves in which only the root has outdegree greater than 1. This also implies ${}^qD_N = {}^rD_N$, ${}^qD_S = {}^rD_S$ and ${}^qJ_S = {}^qJ_N = {}^qJ_L = {}^qJ = 1$.

Property 0.11 ${}^qD_N = {}^rD_L$ if and only if the tree is a piecewise star tree with equal branch sizes at every depth. This also implies ${}^qJ_N = {}^qJ_L = 1$.

Property 0.12 ${}^qD_S = {}^qD_N = {}^qD_L$ if and only if the tree is a star tree with equally sized leaves. This also implies ${}^qJ_S = {}^qJ_N = {}^qJ_L = 1$.

Property 0.13 ${}^qD_S = {}^qD_N = {}^qD_L = {}^rD$ if and only if the tree is a leafy ultrametric star tree with equally sized leaves. This also implies ${}^qJ_S = {}^qJ_N = {}^qJ_L = {}^qJ = 1$.

In yet another set of special cases, the evenness formulas simplify to ratios. The following two results are immediate consequences of ${}^0D_L$ or ${}^0D_N$ being constant under the specified conditions.

Property 0.14 If the branch count across the tree is constant and greater than one then

$$ {}^qJ_L = \frac{\log {}^qD_L}{\log {}^0D_L}. $$

Property 0.15 If the tree has uniform outdegree greater than one and the branches present at every depth in the tree have equal lengths then

$$ {}^qJ_N = \frac{\log {}^qD_N}{\log {}^0D_N}. $$

All properties described in this section would also hold if we were to define all our richness and diversity indices as weighted arithmetic, rather than geometric, means of interval or node values (or indeed any other weighted power mean). Our preference for geometric means will be justified in the next section.

The leafy tree identity

For an important class of trees, our index definitions lead to a surprisingly simple, fundamental connection between tree balance, Shannon's diversity index, Sackin's index, and outdegree. This result is less obvious than the properties of the previous section and requires a more substantial proof. We term this unifying relationship the leafy tree identity.

Lemma 0.7 If the tree is leafy and all branches have equal length $l > 0$ then

$$\log {}^1D_N = \frac{{}^1Hl}{\bar{h}}.$$

If additionally all n leaves have equal size then

$$\log {}^1D_N = \frac{n \log n}{I_S},$$

where I_S is Sackin's index.

Proof. The proof is identical to the proof of Proposition 6 in Lemant et al. (2022), except for the base of the logarithms and the additional factor l. □

Proposition 4 (The leafy tree identity; generalization of Proposition 6 in Lemant et al. (2022)) If the tree is leafy and has uniform branch lengths and all internal nodes have outdegree $m > 1$ then

$$ {}^1J_N = \frac{{}^1Hl}{\bar{h}\log m}. \tag{0.14}$$

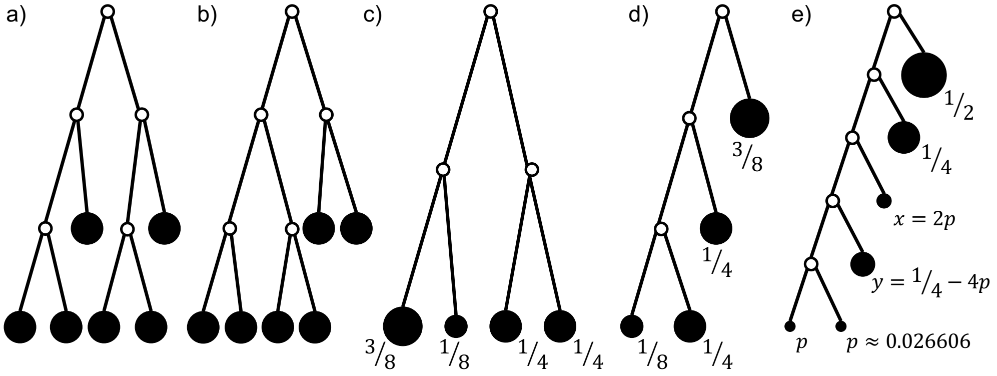
If additionally all n leaves have equal size then

$$ {}^1J_N = \frac{n \log_m n}{I_S}. \tag{0.15}$$

Proof. The result follows immediately from Lemma 0.7 and Property 0.15. □

The leafy tree identity implies that, among leafy trees with uniform branch lengths and uniform outdegrees, tree balance depends only on node sizes and node depths. If two such trees have equal effective heights relative to branch length ($\bar{h}/l$), equal outdegrees (m), and equal node size Shannon entropy values (${}^1H$) then they must have equal balance (${}^1J_N$), irrespective of topology and number of leaves. For example, Figure 7a and 7b show a pair of bifurcating leafy ultrametric trees with uniform leaf sizes and uniform branch lengths. Because these trees have equal outdegrees, leaf counts, and Sackin's index values, the special form of the leafy tree identity (Equation 0.15) implies they must be equally balanced (other equal index values are recorded in Figure 7f). The following example applies the more general form of the leafy tree identity (Equation 0.14) to trees that are less obviously similar.

Example 0.8 Consider the bifurcating leafy ultrametric tree with four leaves, uniform branch lengths, and leaf sizes $\frac{3}{8}$, $\frac{1}{8}$, $\frac{1}{4}$ and $\frac{1}{4}$ (Figure 7c). Now suppose we retain the leaf sizes but rearrange the nodes and branches to form a caterpillar tree with the node of size $\frac{3}{8}$ at depth l and one of the nodes of size $\frac{1}{4}$ at depth $2l$ (Figure 7d). Finally, consider a six-leaf caterpillar tree with uniform branch lengths and proportional leaf sizes (in order of



f)

	^0D	^1D	^qD	^1J	^qJ	^0D_L	^1D_L	^qD_L	^1J_L	^qJ_L	^0D_N	^1D_N	^qD_N	^1J_N	^qJ_N	^0D_S	^1D_S	^qD_S	^1J_S	^qJ_S
a, b	=	=	=	=	=	=	≠	≠	≠	≠	=	=	≠	=	≠	=	≠	≠	≠	≠
c, d	=	=	=	=	=	≠	≠	≠	≠	≠	=	=	≠	=	≠	=	≠	≠	≠	≠
d, e	≠	=	≠	≠	≠	=	=	≠	=	≠	=	=	≠	=	≠	≠	≠	≠	≠	≠

Fig. 7. a-b) Two leafy bifurcating trees with uniform node sizes and uniform branch lengths, which differ in topology but are equally balanced. c-e) Three leafy bifurcating trees with uniform branch lengths, which differ in topology and number of leaves but are equally balanced. Nodes are labelled with their sizes. f) Table recording where pairs of trees have equal or unequal index values. Parameter q can take any non-negative value.

increasing depth) $\frac{1}{2}, \frac{1}{4}, x, y, p$ and p, with $p \approx 0.026606$ (Figure 7e). All three trees have identical values of $m, \bar{h}$ and ^1H (see Appendix for derivation). Hence the leafy tree identify implies that they have equal ^1D_N and ^1J_N values. All three trees also have equal values of $^1D = \exp {}^1H \approx 3.75$ and $^0D_N = m = 2$. Other index values shared by pairs of trees are indicated in Figure 7f.

Equation 0.15 is especially useful because the numerator $n \log_m n$ is the minimum value that I_S can attain on leafy n-leaf trees with uniform branch lengths, uniform node sizes, and uniform outdegree $m > 1$. Hence $(n \log_m n)/I_S$ lies between 0 and 1 and is equal to 1 if and only if the tree is fully balanced. We previously showed (Proposition 7 in Lemant et al. (2022)) that, among all node-wise arithmetic mean indices with $w_i = n_i$, ^1J_N is the only index that satisfies Equation 0.15. Our previous proof can be straightforwardly generalized to show that Equation 0.15 cannot hold for any index of the form $M_{node,r}(^qJ)$ with $r \neq 1$ or $q \neq 1$. Therefore ^1J_N is the only tree balance index for which this useful, unifying identity holds.

An example cross-disciplinary application

We illustrate the universality of our methods by using them to compare the shapes of two trees from different fields of research, representing dissimilar processes and constructed using different methods. The first of these trees depicts the evolution of the Human Immunodeficiency Virus (HIV) within a host, as inferred from molecular data and

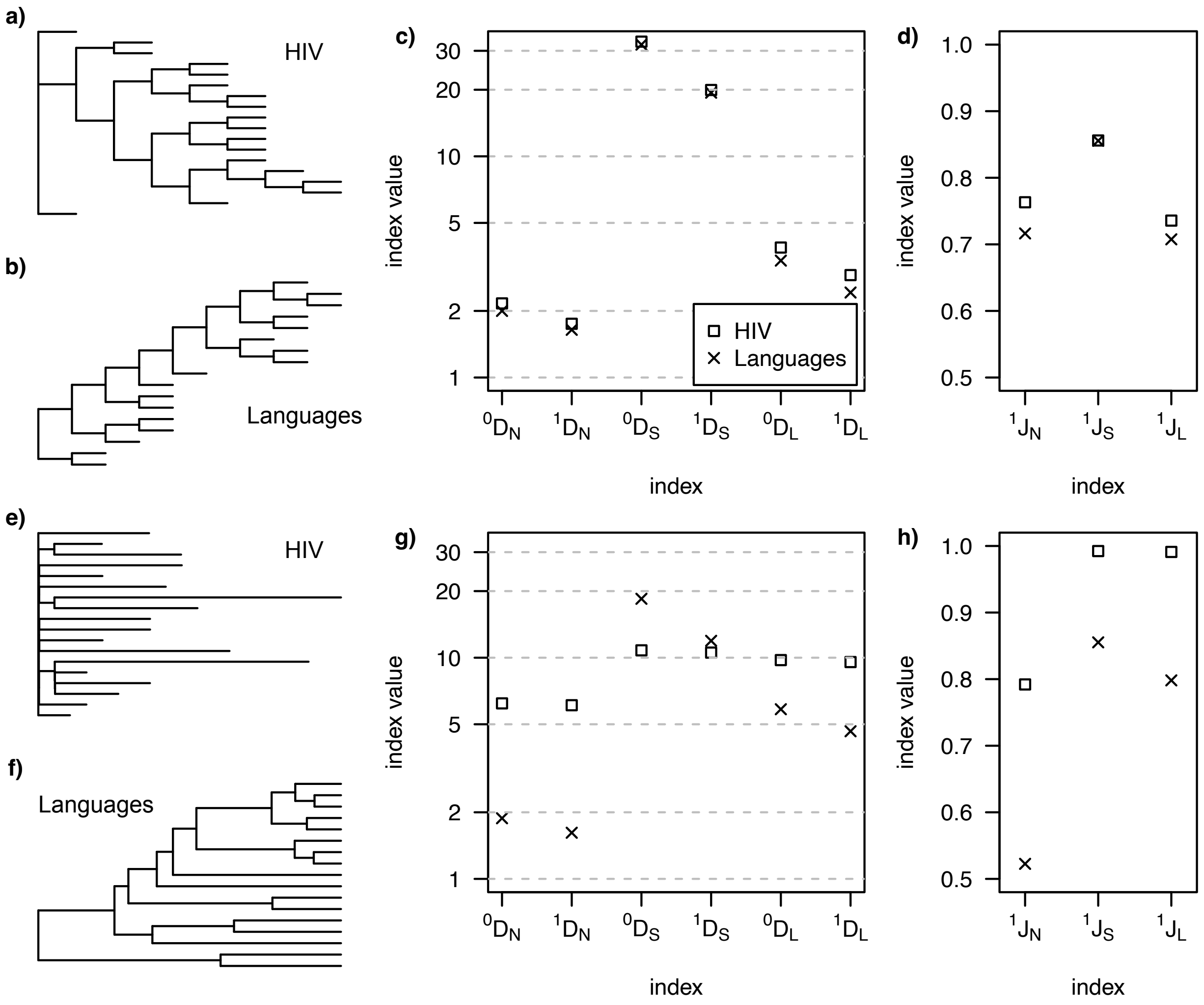


Fig. 8. a-b) Trees with equalized branch lengths representing the within-host evolution of HIV (a) and the evolutionary history of the Uralic languages (b). c) Diversity index values for the two trees with equalized branch lengths. d) Evenness index values for the two trees with equalized branch lengths. e-f) The same trees but with the originally inferred branch lengths. g) Diversity index values, accounting for branch lengths. h) Evenness index values, accounting for branch lengths. In all cases, leaves are assigned equal size and internal nodes are assigned size zero. The HIV tree was sourced from the GitHub repository associated with Barzilai and Schrago (2023) (file PIC38051.tre) and the languages tree from the D-PLACE database (Kirby et al., 2016) (folder honkola_et_al2013).

as used in another recent study of tree shape indices (Barzilai and Schrago, 2023). The second tree represents the diversification of the Uralic language family (Honkola et al., 2013). To simplify the exposition we assign size zero to all internal nodes and an equal size to all leaves.

If we disregard the inferred branch lengths then it is difficult by eye to assess which tree is the more diverse or more balanced (Figure 8a, b). These apparent similarities are borne out in the shape index values (Figure 8c, d). Excepting one node, both trees are bifurcating and therefore both have $^0D_N \approx 2$. The two trees have similar branch counts in total (${}^0D_S = 33$ and 32) and at each depth ($3 < {}^0D_L < 4$). The ${}^1D_N, {}^1D_S$ and ${}^1D_L$ values are somewhat lower than the corresponding richness values due to imbalances, as captured by our evenness indices, which are likewise similar for the two trees (${}^1J_N$ and ${}^1J_L$ between 0.7 and 0.8; ${}^1J_S \approx 0.86$). Lemma 0.7 further implies similar I_S values (93 and 97).

When we restore the inferred branch lengths, the two trees no longer look alike (Figure 8e, f). The HIV phylogeny approximates a non-ultrametric star tree, with long branches originating close to the root. The average effective out-degree of the HIV tree, accounting for unequal branch lengths, is substantially higher than two (${}^0D_N \approx 6$); the effective number of branches is three times lower than when branch lengths are ignored (${}^0D_S \approx 11$); and there are more than twice as many parallel branches (${}^0D_L \approx 10$). Because the HIV tree is approximately a star tree with equal node sizes, all its diversity indices are approximately equal and all its evenness indices are close to one (Property 0.12). In the case of the languages tree, accounting for the inferred branch lengths – which are approximately exponentially distributed and not nearly so depth-dependent – has only a small effect on most index values. The diversity indices for the languages tree remain far from equal. Altogether our indices thus show that the HIV tree is much bushier, has a larger number of effective types, and is in every sense more balanced than the languages tree (Figure 8g, h).

In summary, the clear differences between these two trees, implying different modes of evolution, are captured only by indices that account for their different branch length distributions. An analysis based on prior tree balance indices, which ignore branch lengths, would incorrectly conclude that the trees have very similar shapes and plausibly resulted from similar processes.

Discussion

The seminal paper of Hill (1973) cautions that “almost unlimited scope for mathematical generality in relation to measures of diversity and taxonomic difference” and therefore “Simple and well-understood indices should be used”. In accordance with this advice, here we have constructed new tree shape indices as weighted means of the most standard, basic diversity and evenness indices. This systematic approach ensures that all our indices are not only robust and universally applicable but also have simple, consistent interpretations and clear interrelationships.

Some of the indices we have defined here are refinements of prior approaches to assessing tree shape. Our ${}^qD_L$ and H'_P are similar to the ${}^q\bar{D}$ of Chao et al. (2010) and the phylogenetic entropy H_P of Allen et al. (2009), respectively, but are more self-consistent and can be meaningfully applied to non-ultrametric trees. ${}^qJ_N$ builds on the ideas of Lemant et al. (2022) but, by accounting for branch lengths – a key advantage of prior phylogenetic diversity and phylogenetic entropy indices, not shared by any prior tree balance indices – generalizes the concept of tree balance to a wider class of trees. These new indices share all the desirable properties but not the shortcomings of their predecessors and can therefore universally supersede them (Table 3). For the remainder of our indices describing average effective out-degree, effective numbers of nodes and branches, and evenness of branch sizes, we know of no precedents. In combination, our indices provide a more sophisticated, general, multidimensional description of tree shape than has previously been possible.

Whereas we have focussed on a system built around ${}^qD$ and ${}^qJ$, it is easy to use our general definitions of the longitudinal, node-wise, and star means to quantify other aspects of tree shape. A parallel, self-consistent system of indices can be defined by setting $w_i = h_i$ instead of $w_i = S_i h_i$ in Equations 0.3 and 0.4, and setting $w_k = 1$ instead of $w_k = S_{C_k}$ in Equations 0.8 and 0.9. These indices, which are robust to small changes in branch lengths but not node sizes, are normalized by dividing by h instead of $\bar{h}$. Alternatively, ${}^qD$ can be replaced by another basic diversity index, or ${}^qJ$ by another evenness index, such as the ratio ${}^qD/{}^0D$ preferred by Hill (1973) (see also Smith and Wilson (1996); Jost (2010); Tuomisto (2012)). Based on the means, we can also straightforwardly derive expressions for higher moments to obtain indices that, for example, quantify how much effective out-degree varies across all nodes or varies with node depth.

There are nevertheless several reasons for preferring our specific definitions. First, the foundational ${}^qD$ and ${}^qJ$ are the most popular diversity and evenness indices among biologists (Tucker et al., 2017; Tuomisto, 2012). Second, defining entropy and evenness indices as weighted arithmetic means, and diversity indices as weighted geometric means, results in relatively simple expressions, especially in the case of leafy ultrametric trees. Third, ${}^1J_N$ is the only universal tree balance index for which the unifying leafy tree identity holds. In summary, we have taken the best of the existing indices, improved them, unified them, and filled in the gaps to create a coherent system (Table 2).

Given the ubiquity of tree structures, we expect our multidimensional method of describing tree shape to empower research and inform decision making in diverse domains. Our initial development of universal, robust indices was motivated by the need to compare and categorize non-leafy, non-ultrametric trees representing the clonal evolution of human tumours, where node sizes (corresponding to cell subpopulation sizes) and branch lengths (genetic distances) convey valuable information (Noble et al., 2022). Tree structures with node sizes and branch lengths are likewise centrally important in community ecology, conservation biology, systematic biology, and the study of microbial evolution. For instance, our indices can be used instead of conventional tree balance indices to evaluate alternative models of speciation, or to investigate how the mode of evolution of a pathogenic virus varies with geographical location, time period, or strain. In place of phylogenetic diversity and phylogenetic entropy, our non-normalized diversity indices could be used to inform policy making by quantifying how different actions would affect biodiversity. Beyond biology, obvious subjects for analysis include phylogenetic trees of language evolution, hierarchical organizational structures, and the tree data structures that abound in computing. As we have illustrated, our generic indices can be used not only within but also across domains to uncover similarities and differences in, say, the evolution of organisms, languages, and technologies.

One key topic for further theoretical research is to derive the expected values and covariances of our indices under standard tree generation models, such as the uniform model and the Yule process, for comparison with empirical data. Relationships between our indices and distance-based metrics such as the mean pairwise distance (which lacks a universal normalization (Tsirogiannis et al., 2012)) also remain to be examined. In the same vein as Figure 7, we are investigating sets of distinct trees to which our indices assign equal values, to determine whether additional indices might ever be needed to distinguish between trees in typical applications. Towards establishing a universal standard for describing tree shape, we are developing software packages for calculating index values that can be integrated with popular tree inference methods. Just as the first step in analysing a set of measurements is to calculate the mean and variance, so we propose that, whenever one encounters a rooted tree, a useful first step will be to describe its shape by evaluating our indices.

Acknowledgements

We are grateful to Kerry Manson for helpful comments on an earlier draft of this manuscript, and to Lucia Barzilai and Chiara Barbieri for helping us obtain suitable empirical tree data.

Code availability

Open source code to calculate our new tree shape indices for trees in Newick or NEXUS format, or phylo objects, is at https://github.com/robjohnnoble/TreeIndices.

References

Susanne Albers and Jeffery Westbrook. Self-organizing data structures. *Online Algorithms: The state of the art*, pages 13–51, 2005.

Benjamin Allen, Mark Kon, and Yaneer Bar-Yam. A New Phylogenetic Diversity Measure Generalizing the Shannon Index and Its Application to Phyllostomid Bats. *The American Naturalist*, 174(2):236–243, 2009.

Quentin D. Atkinson and Russell D. Gray. Curious Parallels and Curious

Connections—Phylogenetic Thinking in Biology and Historical Linguistics. *Systematic Biology*, 54(4):513–526, 2005.

Lucia P Barzilai and Carlos G Schrago. Signatures of natural selection in tree topology shape of serially sampled viral phylogenies. *Molecular Phylogenetics and Evolution*, 183: 107776, 2023.

Anne Chao, Chun Huo Chiu, and Lou Jost. Phylogenetic diversity measures based on Hill numbers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365 (1558):3599–3609, 2010.

Leonid Chindelevitch, Maryam Hayati, Art FY Poon, and Caroline Colijn. Network science inspires novel tree shape statistics. *Plos one*, 16(12):e0259877, 2021.

Caroline Colijn and Jennifer Gardy. Phylogenetic tree shapes resolve disease transmission patterns. *Evolution, medicine, and public health*, 2014(1):96–108, 2014.

Donald H Colless. Review of Phylogenetics, The Theory and Practice of Phylogenetic Systematics. *Systematic Zoology*, 31(1):100–104, 1982.

Daniel P. Faith. Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61(1):1–10, 1992.

Mareike Fischer, Lina Herbst, Sophie Kersting, Luise Kühn, and Kristina Wicke. Tree balance indices: a comprehensive survey, 2021. arXiv: 2109.12281.

Mark Hill. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology*, 54:427–432, 1973.

Terhi Honkola, Outi Vesakoski, Kalle Korhonen, Jyri Lehtinen, Kaj Syrjänen, and Niklas Wahlberg. Cultural and climatic changes shape the evolutionary history of the uralic languages. *Journal of Evolutionary Biology*, 26(6):1244–1253, 2013.

Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.

Lou Jost. The Relation between Evenness and Diversity. *Diversity*, 2(2):207–232, 2010.

Kathryn R Kirby, Russell D Gray, Simon J Greenhill, Fiona M Jordan, Stephanie Gomes-Ng, Hans-Jörg Bibiko, Damián E Blasi, Carlos A Botero, Claire Bowern, Carol R Ember, et al. D-place: A global database of cultural, linguistic and environmental diversity. *PloS one*, 11(7):e0158391, 2016.

Tom Leinster and Christina A. Cobbold. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489, 2012.

Jeanne Lemant, Cécile Le Sueur, Veselin Manojlović, and Robert Noble. Robust, Universal Tree Balance Indices. *Systematic Biology*, 71(5):1210–1224, 2022.

Gabriel E Leventhal, Roger Kouyos, Tanja Stadler, Viktor Von Wyl, Sabine Yerly, Jürg Böni, Cristina Cellerai, Thomas Klimkait, Huldrych F Günthard, and Sebastian Bonhoeffer. Inferring epidemic contact structure from phylogenetic trees. *PLoS computational biology*, 8(3):e1002413, 2012.

Arnau Mir, Francesc Rosselló, and Lucía Rotger. A new balance index for phylogenetic trees. *Mathematical Biosciences*, 241(1):125–136, 2013. arXiv: 1202.1223.

Arnau Mir, Lucía Rotger, and Francesc Rosselló. Sound Colless-like balance indices for multifurcating trees. *PLoS ONE*, 13(9):559–560, 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, 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 & Evolution*, 6(2):207–217, 2022.

Sandrine Pavoine and Michael B Bonsall. Measuring biodiversity to explain community assembly: a unified approach. *Biological Reviews*, 86(4):792–812, 2011.

E. C. Pielou. The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology*, 13:131–144, 1966.

Andy Purvis and Paul-Michael Agapow. Phylogeny imbalance: taxonomic level matters. *Systematic Biology*, 51(6):844–854, 2002.

James S. Rogers. Response of Colless's Tree Imbalance to Number of Terminal Taxa. *Systematic Biology*, 42(1):102–105, 1993.

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.

Benjamin Smith and J. Bastow Wilson. A Consumer's Guide to Evenness Indices. *Oikos*, 76(1):70–82, 1996.

Constantinos Tsirogiannis, Brody Sandel, and Dimitris Cheliotis. Efficient computation of popular phylogenetic tree measures. In *International Workshop on Algorithms in Bioinformatics*, pages 30–43. Springer, 2012.

Caroline M. Tucker, Marc W. Cadotte, Silvia B. Carvalho, T. Jonathan Davies, Simon Ferrier, Susanne A. Fritz, Rich Grenyer, Matthew R. Helmus, Lanna S. Jin, Arne O. Mooers, Sandrine Pavoine, Oliver Purschke, David W. Redding, Dan F. Rosauer, Marten Winter, and Florent Mazel. A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological Reviews*, 92(2):698–715, 2017.

Hanna Tuomisto. An updated consumer's guide to evenness and related indices. *Oikos*, 121(8):1203–1218, 2012.

Simon Veron, Victor Saito, Nélida Padilla-García, Félix Forest, and Yves Bertheau. The use of phylogenetic diversity in conservation biology and community ecology: A common base but different approaches. *Quarterly Review of Biology*, 94(2):123, 2019.

APPENDIX

Derivation of ${}^qD_N$ in Example 0.3

For the root r, we have $\bar{h}_{C_r} = \lambda$ and $A_r = \{r\}$. For either of the other internal nodes k, we have $\bar{h}_{C_k} = (1-\lambda)/2$ and $A_k = \{k, r\}$. The subtree weights are

$$S_{C_r}(x) = \begin{cases} 1 \text{ if } 0 \leqslant x < \lambda, \\ 0 \text{ otherwise,} \end{cases} \qquad S_{C_k}(x) = \begin{cases} \frac{1}{2} \text{ if } \lambda \leqslant x < 1, \\ 0 \text{ otherwise.} \end{cases}$$

The ancestor weights are

$$v_{rr} = \int_0^\infty S_{C_r}(x)\,dx = \int_0^\lambda 1\,dx = \lambda, \quad v_{kk} = \int_\lambda^{2\lambda} S_{C_k}(x)\,dx = \begin{cases} \int_\lambda^{2\lambda} \frac{1}{2}\,dx = \frac{\lambda}{2} \text{ if } \lambda < \frac{1}{2}, \\ \int_\lambda^1 \frac{1}{2}\,dx = \frac{1-\lambda}{2} \text{ otherwise,} \end{cases}$$

$$v_{rk} = \int_{2\lambda}^\infty S_{C_k}(x)\,dx = \begin{cases} \int_{2\lambda}^1 \frac{1}{2}\,dx = \frac{1-2\lambda}{2} \text{ if } \lambda < \frac{1}{2}, \\ 0 \text{ otherwise.} \end{cases}$$

The node diversity values are, for all $q \geqslant 0$,

$${}^qD(P_r(x)) = \begin{cases} 2 \text{ if } 0 \leqslant x < \lambda, \\ 4 \text{ if } \lambda \leqslant x < 1, \\ 0 \text{ otherwise.} \end{cases} \qquad {}^qD(P_k(x)) = \begin{cases} 2 \text{ if } \lambda \leqslant x < 1, \\ 0 \text{ otherwise.} \end{cases}$$

Hence

$$\int_0^h S_{C_r}(x)^q H_N(P_r(x))\,dx = \int_0^\lambda 1 \log 2\,dx = \lambda \log 2,$$

$$\int_0^h S_{C_k}(x)^q H_N(P_k(x))\,dx = \int_\lambda^1 \frac{1}{2} \log 2\,dx = \frac{(1-\lambda)\log 2}{2},$$

$$\int_0^h S_{C_k}(x)^q H_N(P_r(x))\,dx = \int_\lambda^1 \frac{1}{2} \log 4\,dx = (1-\lambda)\log 2.$$

For all $q \geqslant 0$, it follows from Equation 0.8 that if $\lambda \geqslant \frac{1}{2}$ then

$${}^{q}D_N = \exp\left(\frac{1}{\lambda}\lambda(\lambda\log 2) + 2 \times \frac{1}{(1-\lambda)/2}\left(\frac{1-\lambda}{2}\frac{(1-\lambda)\log 2}{2}\right)\right) = 2,$$

and otherwise

$${}^{q}D_N = \exp\left(\frac{1}{\lambda}\lambda(\lambda\log 2) + 2 \times \frac{1}{(1-\lambda)/2}\left(\frac{\lambda}{2}\frac{(1-\lambda)\log 2}{2} + \frac{2-2\lambda}{2}(1-\lambda)\log 2\right)\right) = 4^{1-\lambda}.$$

Proof of Proposition 1

Proof. For $q \geqslant 0$, let $k(q) \in I(T)$ such that ${}^{q}D(P_{k(q)}) \geqslant {}^{q}D(P_i)$ for all $i \in I(T)$, and let $b_1, \ldots, b_{|P_{k(q)}|}$ denote the non-zero-sized branches in the interval $k(q)$. Then, by a basic property of generalized means,

$${}^{q}D_L(T) := M_{long,0}({}^{q}D)(T) \leqslant \lim_{r\to\infty} M_{long,r}({}^{q}D)(T) = \max_{i\in I(T)} {}^{q}D(P_i) = {}^{q}D(P_{k(q)}).$$

For the first part of the proposition, we note that for any interval $i \in I(T)$, the number of non-zero-sized branches in i is $|P_i|$, which cannot exceed the number of counted nodes ${}^{0}D(T)$. Hence, for any rooted tree T and all $q \geqslant 0$,

$${}^{q}D_L(T) \leqslant {}^{q}D(P_{k(q)}) := \left(\sum_{b\in B_{k(q)}} p_b^q\right)^{\frac{1}{1-q}} = \left(\sum_{j=1}^{|P_{k(q)}|} p_{b_j}^q\right)^{\frac{1}{1-q}} \leqslant \left(\sum_{j=1}^{{}^{0}D} p_{b_j}^q\right)^{\frac{1}{1-q}} \leqslant {}^{0}D(T).$$

We now turn to the second part. By definition, for all $i \in I$, if $b \in B_i$ then branch size $s_b = \sum_{x\in V_b} f_x$, where V_b is the set of all nodes that descend from b, and f_x is the proportional size of node x. For all rooted trees we have $V_{b_1} \cap V_{b_2} = \emptyset$ for all $b_1, b_2 \in B_i$ with $b_1 \neq b_2$. For ultrametric trees, $\bigcup_{b\in B_i} V_b = L$, where L is the set of all leaves in the tree. For leafy ultrametric trees, $S_i = 1$ for all i and hence $p_b = s_b$ for all $b \in B_i$. Then for any leafy ultrametric tree T and all $q \geqslant 0$,

$$\begin{aligned}{}^{q}D_L(T) \leqslant {}^{q}D(P_{k(q)}) := \left(\sum_{b\in B_{k(q)}} p_b^q\right)^{\frac{1}{1-q}} &= \left(\sum_{b\in B_{k(q)}} s_b^q\right)^{\frac{1}{1-q}} = \left(\sum_{b\in B_{k(q)}}\left(\sum_{x\in V_b} f_x\right)^q\right)^{\frac{1}{1-q}} \\ &\leqslant \left(\sum_{b\in B_{k(q)}}\sum_{x\in V_b} f_x^q\right)^{\frac{1}{1-q}} = \left(\sum_{x\in L(T)} f_x^q\right)^{\frac{1}{1-q}} = {}^{q}D(T).\end{aligned}$$

Finally we will prove that this inequality does not hold for all rooted trees. We will do so in a more general context to show that the result is independent of our choice of weight function w and exponent r. Let ${}^{q}D_{L,r,w} = M_{long,r}({}^{q}D; w)$ for real r, where w is a continuous, monotonically increasing function of S_i and h_i such that $w_i = w(S_i, h_i) > 0$

when $S_i > 0$ or $h_i > 0$, and $w_i \to 0$ as $S_i \to 0$ or $h_i \to 0$. First consider the leafy but non-ultrametric three-leaf star tree T_1 in which one leaf has size $1-p$ and depth λ, and the other two leaves have size $\frac{p}{2}$ and depth $1+\lambda$ (as in Figure 4 but with one more leaf). Now

$$^{q}D_{L,r,w}(T_1) = \left(\frac{\sum_{i\in I(T_!)} w_i [^{q}D(P_i)]^r}{\sum_{i\in I(T_1)} w_i} \right)^{\frac{1}{r}} = \left(\frac{w_1 \left((1-p)^q + 2\left(\frac{p}{2}\right)^q \right)^{\frac{r}{1-q}} + w_2 \left(2\left(\frac{1}{2}\right)^q \right)^{\frac{r}{1-q}}}{w_1 + w_2} \right)^{\frac{1}{r}}.$$

Since w_1 depends only on λ and w_2 depends only on p, we can make λ a function of p such that $\frac{w_1}{w_2} \to 0$ as $\lambda \to 0$ and $p \to 0$, in which case $^{q}D_{L,r,w}(T_1) \to 2$ as $\lambda \to 0$ and $p \to 0$. Also, for all $q > 0$, $^{q}D(T_1) \to 1$ as $p \to 0$. Hence $^{q}D_{L,r,w}(T_1) > {}^{q}D(T_1)$ as $\lambda \to 0$ and $p \to 0$. Instead setting $\lambda = 0$ makes T_1 ultrametric but non-leafy, with $^{q}D_{L,r,w}(T_1) \to 2$ and $^{q}D(T_1) \to 1$ as $p \to 0$ as before. □

Derivation of index values in Example 0.2

Since the tree of Figure 7c is ultrametric, $\bar{h}/l = 2$, where l is the branch length. The node size entropy is

$$^{1}H = -\frac{3}{8}\log\frac{3}{8} - 2\left(\frac{1}{4}\log\frac{1}{4}\right) - \frac{1}{8}\log\frac{1}{8} = \frac{5}{2}\log 2 - \frac{3}{8}\log 3 \approx 1.32.$$

The tree of Figure 7d has the same node sizes and therefore the same ^{1}H value as the previous tree. It also has the same relative effective height, as

$$\frac{\bar{h}}{l} = \frac{3}{8} + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{8} + \frac{1}{4}\right) = 2.$$

For the tree of Figure 7e, since proportional node sizes must sum to unity we have

$$1 = \frac{1}{2} + \frac{1}{4} + x + y + 2p \implies x + y + 2p = \frac{1}{4}.$$

For this tree to have the same $\bar{h}$ and ^{1}H values as the four-leaf trees we additionally require

$$2 = \frac{\bar{h}}{l} = \frac{1}{2} + 2\left(\frac{1}{4}\right) + 3x + 4y + 5(2p) \implies 3x + 4y + 10p = 1,$$
$$\frac{5}{2}\log 2 - \frac{3}{8}\log 3 =^{1} H = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - x\log x - y\log y - 2p\log p.$$

The first two equations together imply $x = 2p$ and $y = \frac{1}{4} - 4p$. After substituting these results into the third equation we obtain the numerical solution $p \approx 0.026606$. Since all three trees have identical values of $m, \bar{h}$ and ^{1}H, the leafy tree identify implies that they have equal $^{1}D_N$ values and equal balance:

$$^{1}D_N = \exp\left(\frac{^{1}Hl}{\bar{h}}\right) = \exp\left(\frac{^{1}H}{2}\right) \approx 1.94, \quad ^{1}J_N = \frac{^{1}Hl}{\bar{h}\log m} = \frac{^{1}H}{2\log 2} \approx 0.95.$$