The minimum value of the Colless index

Tomás M. Coronado, Francesc Rosselló

Dept. of Mathematics and Computer Science, University of the Balearic Islands, E-07122 Palma, Spain, and Balearic Islands Health Research Institute (IdISBa), E-07010 Palma, Spain

Abstract

The Colless index is one of the oldest and most widely used balance indices for rooted bifurcating trees. Despite its popularity, its minimum value on the space $\mathcal{T}_n$ of rooted bifurcating trees with n leaves is only known when n is a power of 2. In this paper we fill this gap in the literature, by providing a formula that computes, for each n, the minimum Colless index on $\mathcal{T}_n$, and characterizing those trees where this minimum value is reached.

Keywords: Phylogenetic tree, Colless index, Balance index

1. Introduction

One of the main goals of evolutionary biology is to understand what factors influence evolutionary processes and their effect. Since phylogenetic trees are the standard representation of joint evolutive histories of groups of species, it is natural to look for the imprint of these factors in the shapes of phylogenetic trees [14, 21]. This has motivated the introduction of various indices that quantify topological features of tree shapes supposedly related to properties of the evolutionary processes represented by the trees. These indices have been used then to test evolutionary hypothesis [3, 11, 14, 20] and to compare tree shapes [2, 9], among other applications. Since the early observation by Willis and Yule [23] that taxonomic trees tend to be asymmetric, with many small clades and a few large ones at every taxonomic level, the most popular topological feature used to describe the shape of a phylogenetic tree has been its *balance*, the tendency of the children of any given node to have the same number of descendant leaves. The *imbalance* of a phylogenetic tree reflects the propensity of evolutive events to occur along specific lineages [15].

Several *balance indices* have been proposed so far to quantify the balance (or actually, in most cases, the imbalance) of a phylogenetic tree; see, for instance, [4, 5, 11, 12, 17, 18, 22] and the section "Measures of overall asymmetry" in Felsenstein's book [6] (pp. 562–563). Among them, the Colless index, introduced by D. H. Colless [4], is one of the oldest and most popular. It is defined, on a bifurcating tree T, as the sum, over all the internal nodes v of T, of the absolute value of the difference of the numbers of descendant leaves of the pair of children of v. Although it was defined primarily for bifurcating phylogenetic trees, Shao and Sokal [18] proposed to extend it to multifurcating trees by taking into account in this sum only the bifurcating nodes; a more meaningful extension to arbitrary trees has been proposed recently by Mir et al [13]. Anyway, in this paper we deal with the classical Colless index for bifurcating trees, and since it is a *shape index*, in the sense that it does not depend on the actual labels of the tree's leaves, we shall not take into account these labels and we shall consider this index defined on unlabeled rooted bifurcating trees, or simply *bifurcating trees* for short.

By its very definition, the Colless index measures directly the total amount of imbalance of a bifurcating tree, with the notion of balance recalled above. And, indeed, for every number n of leaves, it is folklore knowledge that the bifurcating tree with n leaves and largest Colless index is the *comb*: the bifurcating tree

Email addresses: t.martinez@uib.es (Tomás M. Coronado), cesc.rossello@uib.es (Francesc Rosselló)

such that each internal node has a leaf child, considered since the early paper by Sackin [17] to be the most imbalanced type of phylogenetic tree. This fact was already hinted by Colless [4], but he gave a wrong value for the Colless index of a comb which was later corrected by Heard [10], giving the correct maximum value of $(n-1)(n-2)/2$. Actually, to our knowledge, no explicit direct proof of the maximality of this Colless value has been provided in the literature, but it can be easily deduced as a particular case of Thm. 18 in [13].

But, in spite of its popularity and wide use, the minimum Colless index of a bifurcating tree with n leaves remains unknown beyond the often stated straighforward result that for numbers of leaves that are powers of 2 it is reached at the fully symmetric bifurcating trees, which clearly have Colless index 0; see for example [10, 11, 14]. To have a closed formula for this minimum value is necessary, for instance, in order to normalize the Colless index to the range $[0, 1]$ for every number of leaves, making its value independent of its size as it is recommended by Shao and Sokal [18] or Stam [20].

The goal of this paper is then to fill this gap in the literature. In Sections 2 and 3 we prove that, for every n, the minimum Colles index on the space $\mathcal{T}_n$ of bifurcating trees with n leaves is reached at the *maximally balanced trees*, those trees such that, for every internal node v, the numbers of descendant leaves of its two children differ at most of 1, and we provide a formula that computes this minimum value for each n (see Proposition 2). This result is not surprising, because these maximally balanced trees are considered, as their name hints, the "most balanced binary trees" [18], and many other balance indices reach their smallest values at them: for instance, Fischer [7] has recently proved it for the Sackin index introduced in [17, 18], and it is also the case for the total cophenetic index [12] or the rooted quartet index [5] (in fact, the maximally balanced trees have the largest rooted quartet index, because this index, unlike the others, grows from most imbalanced to most balanced). Unfortunately, there are trees that are not maximally balanced but that also yield the minimum Colless index for their number of leaves (see, for instance, Fig. 2 below). Then, in Section 4 we characterize the trees in $\mathcal{T}_n$ having minimum Colless index and we provide an efficient algorithm that produces them for any n.

Notations. In this paper, by a *tree* we mean a rooted tree, understood as a directed graph with its arcs pointing away from the root. Such a tree is *bifurcating* when all its nodes have out-degree either 0 (its *leaves*) or 2 (its *internal nodes*). We shall denote by $\mathcal{T}_n$ the set of (isomorphism classes of) bifurcating trees with n leaves. A *cherry* is a tree consisting only of a root and two leaves.

Given a bifurcating tree $T \in \mathcal{T}_n$, we shall denote by $V(T)$ its set of nodes and by $V_{int}(T)$ its set of internal nodes. For every $u, v \in V(T)$, v is a *child* of u when T contains the arc (u, v), and v is a *descendant* of u when T contains a (directed) path from u to v. For every $v \in V(T)$, the *subtree T_v of T rooted at v* is the subgraph of T induced by the descendants of v and $\kappa_T(v)$ is the number of leaves of T_v. Given two bifurcating trees $T \in \mathcal{T}_{n_1}$ and $T' \in \mathcal{T}_{n_2}$, their *root join* is the tree $T \star T' \in \mathcal{T}_{n_1+n_2}$ whose subtrees rooted at the children of the root are T and T'.

For every $v \in V_{int}(T)$, say with children v_1, v_2, the *balance value* of v is $bal_T(v) = |\kappa_T(v_1) - \kappa_T(v_2)|$. We shall say that an internal node of T is *balanced* when its balance value is 0, and *imbalanced* otherwise. The *Colless index* [4] of $T \in \mathcal{T}_n$ is the sum of the balance values of its internal nodes:

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

It is easy to see that the Colless index satisfies the following recurrence [16]: if $T = T_1 \star T_2$, with $T_1 \in \mathcal{T}_{n_1}$ and $T_2 \in \mathcal{T}_{n_2}$, then

$$C(T) = C(T_1) + C(T_2) + |n_1 - n_2|. \tag{1}$$

2. Colless indices of maximally balanced trees

An internal node v in a bifurcating tree T is *balanced* when $bal_T(v) \leqslant 1$: i. e., when its two children have $\lceil \kappa_T(v)/2 \rceil$ and $\lfloor \kappa_T(v)/2 \rfloor$ descendant leaves, respectively. A tree $T \in \mathcal{T}_n$ is *maximally balanced* when all its internal nodes are balanced. Therefore, a tree is maximally balanced when its root is balanced and both subtrees rooted at the children of the root are maximally balanced. This easily implies, on the one hand,

that, for any number n of leaves, there is only one maximally balanced tree in $\mathcal{T}_n$, which we shall denote by B_n, and, on the other hand, that the maximally balanced trees satisfy the following recurrence:

$$B_n = B_{\lceil n/2 \rceil} \star B_{\lfloor n/2 \rfloor}. \tag{2}$$

When n is a power of 2, B_n is the *fully symmetric bifurcating tree*, where, for each internal node, the pair of subtrees rooted at its children are isomorphic. Fig. 1 depicts the trees B_6, B_7 and B_8.

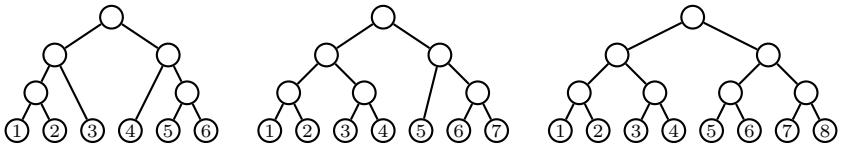

Figure 1: Three maximally balanced trees. The tree with 8 leaves is fully symmetric.

We shall denote by $C(n)$ the Colless index of B_n. The sequence $C(n)$ satisfies that $C(1) = 0$ and, by (1) and (2), for every $n \geqslant 2$

$$C(n) = C(\lceil n/2 \rceil) + C(\lfloor n/2 \rfloor) + \lceil n/2 \rceil - \lfloor n/2 \rfloor. \tag{3}$$

The following well-known lemma is easily proved by induction on n, using recurrence (1).

Lemma 1. *For every $n \geqslant 1$ and $T \in \mathcal{T}_n$, $C(T) = 0$ if, and only if, n is a power of 2 and $T = B_n$.* □

Nex result gives a closed formula for $C(n)$ in terms of the binary expansion of n.

Proposition 2. *If $n = \sum_{j=1}^{\ell} 2^{m_j}$ with $m_1 > \cdots > m_\ell$, then $C(n) = \sum_{j=2}^{\ell} 2^{m_j}(m_1 - m_j - 2(j-2))$.*

Proof. For every $n \geqslant 1$, let $\overline{C}(n) = \sum_{j=2}^{\ell} 2^{m_j}(m_1 - m_j - 2(j-2))$, where $n = \sum_{j=1}^{\ell} 2^{m_j}$ with $m_1 > \cdots > m_\ell$. Since $\overline{C}(1) = \overline{C}(2^0) = 0 = C(1)$, to prove that $\overline{C}(n) = C(n)$ for every $n \geqslant 1$ it is enough to prove that the sequence $\overline{C}(n)$ satisfies the recurrence (3). Now, to prove the latter, we must distinguish two cases, depending on the parity of n.

Assume first that n is even, i. e., $m_\ell > 0$. In this case, $\lfloor n/2 \rfloor = \lceil n/2 \rceil = \sum_{j=1}^{\ell} 2^{m_j - 1}$ with $m_1 - 1 > \cdots > m_\ell - 1$ and then

$$\begin{aligned}\overline{C}(\lceil n/2 \rceil) + \overline{C}(\lfloor n/2 \rfloor) + \lceil n/2 \rceil - \lfloor n/2 \rfloor &= 2\overline{C}\Big(\sum_{j=1}^{\ell} 2^{m_j - 1}\Big) \\ = 2\Big(\sum_{j=2}^{\ell} 2^{m_j - 1}\big(m_1 - 1 - (m_j - 1) - 2(j-2)\big)\Big) &= \sum_{j=2}^{\ell} 2^{m_j}(m_1 - m_j - 2(j-2)) = \overline{C}(n).\end{aligned}$$

Assume now that n is odd, i. e., $m_\ell = 0$. Let $k = \min\{j \mid m_j = \ell - j\}$ (which exists because $m_\ell = \ell - \ell$). Then, $\lfloor n/2 \rfloor = \sum_{j=1}^{\ell-1} 2^{m_j - 1}$, with $m_1 - 1 > \cdots > m_{\ell-1} - 1$, and

$$\lceil n/2 \rceil = \sum_{j=1}^{\ell-1} 2^{m_j - 1} + 1 = \sum_{j=1}^{k-1} 2^{m_j - 1} + \sum_{j=k}^{\ell-1} 2^{\ell - j - 1} + 1 = \sum_{j=1}^{k-1} 2^{m_j - 1} + 2^{\ell - k}$$

with $m_1 - 1 > \cdots > m_{k-1} - 1 > \ell - k$. In this case,

$$\begin{aligned}
&\overline{C}(\lceil n/2\rceil) + \overline{C}(\lfloor n/2\rfloor) + \lceil n/2\rceil - \lfloor n/2\rfloor \\
&\quad = \sum_{j=2}^{k-1} 2^{m_j-1}\big((m_1-1) - (m_j-1) - 2(j-2)\big) + 2^{\ell-k}(m_1 - 1 - (\ell-k) - 2(k-2)\big) \\
&\qquad + \sum_{j=2}^{\ell-1} 2^{m_j-1}\big((m_1-1) - (m_j-1) - 2(j-2)\big) + 1 \\
&\quad = \sum_{j=2}^{k-1} 2^{m_j-1}(m_1 - m_j - 2(k-2)) + 2^{m_k}(m_1 - m_k - 2(k-2)) - 2^{\ell-k} \\
&\qquad + \sum_{j=2}^{k-1} 2^{m_j-1}(m_1 - m_j - 2(j-2)) + \sum_{j=k}^{\ell-1} 2^{\ell-j-1}(m_1 - (\ell-j) - 2(j-2)) + 1 \\
&\quad \text{(because } m_j = \ell - j \text{ for every } j \geqslant k) \\
&\quad = \sum_{j=2}^{k} 2^{m_j}(m_1 - m_j - 2(j-2)) + \sum_{j=k}^{\ell-1} 2^{\ell-j-1}(m_1 - (\ell-j) - 2(j-2)) + 1 - 2^{\ell-k} \\
&\quad = \sum_{j=2}^{k} 2^{m_j}(m_1 - m_j - 2(j-2)) + \sum_{i=k+1}^{\ell} 2^{\ell-i}(m_1 - (\ell-i) - 2(i-2) + 1) + 1 - 2^{\ell-k} \\
&\quad = \sum_{j=2}^{k} 2^{m_j}(m_1 - m_j - 2(j-2)) + \sum_{i=k+1}^{\ell} 2^{m_i}(m_1 - m_i - 2(i-2)) + \sum_{i=k+1}^{\ell} 2^{\ell-i} + 1 - 2^{\ell-k} \\
&\quad = \sum_{j=2}^{\ell} 2^{m_j}(m_1 - m_j - 2(j-2)) = \overline{C}(n)
\end{aligned}$$

This completes the proof that $\overline{C}(n)$ satisfies (3), and hence that $C(n) = \overline{C}(n)$ for every $n \geqslant 1$. □

Remark 3. Since the balance of every internal node in a maximally balanced tree B_n is at most 1, $C(n)$ is equal to the number of imbalanced nodes in B_n. Now, an internal node in B_n is balanced if, and only if, the subtrees rooted at their children are isomorphic: that is, with the notations in [8], if, and only if, it is a *symmetric branch point*. Therefore, the number of symmetric branch points in B_n is $n - 1 - C(n)$, which implies, by Lemma 31 in [8], that the number of automorphisms of B_n is $2^{n-1-C(n)}$. This says that the sequence $C(n)$ is equal to the sequence A296062 in Sloane's *On-Line Encyclopedia of Integer Sequences* [19]. So, Proposition 2 gives an explicit formula for that sequence.

3. The minimum Colless index

In this section we prove that $C(n)$ is the minimum value of the Colless index on $\mathcal{T}_n$.

Lemma 4. *For every $(n, s) \in \mathbb{N}^2$ with $n \geqslant 1$, $C(n+s) + C(n) + s \geqslant C(2n+s)$.*

Proof. We shall prove by induction on n that, for every $n \geqslant 1$, the inequality

$$C(n+s) + C(n) + s \geqslant C(2n+s) \tag{4}$$

holds for every $s \geqslant 0$. Since $C(1) = 0$, the base case $n = 1$ says that, for every $s \geqslant 0$,

$$C(1+s) + s \geqslant C(2+s). \tag{5}$$

We prove it by induction on s. The cases $s = 0$ and $s = 1$ are obviously true, because $C(1) + 0 = 0 = C(2)$ and $C(2)+1 = 1 = C(3)$. Let us consider now the case $s \geqslant 2$ and let us assume that $C(1+s')+s' \geqslant C(2+s')$ for every $s' < s$. To prove the induction step, we distinguish two cases.

- If $s \in 2\mathbb{N}$, say $s = 2t$ with $t \geqslant 1$, then $C(1+s)+s = C(2t+1)+2t = C(t+1)+C(t)+1+2t$ and $C(2+s) = C(2t+2) = 2C(t+1)$ and the desired inequality (5) holds because, by the induction hypothesis, $C(t)+2t+1 = C(1+(t-1))+(t-1)+t+2 \geqslant C(2+(t-1))+t+2 > C(t+1)$.
- If $s \notin 2\mathbb{N}$, say $s = 2t+1$ with $t \geqslant 1$, then $C(1+s)+s = C(2t+2)+2t+1 = 2C(t+1)+2t+1$ and $C(2+s) = C(2t+3) = C(t+2)+C(t+1)+1$ and the desired inequality (5) holds because, by the induction hypothesis, $C(t+1)+2t \geqslant C(t+2)+t > C(t+2)$.

This completes the proof of the base case $n = 1$. Let us consider now the case $n \geqslant 2$ and let us assume that $C(n'+s)+C(n')+s \geqslant C(2n'+s)$ for every $1 \leqslant n' < n$ and $s \in \mathbb{N}$. To prove that (4) is true for every $s \in \mathbb{N}$ we distinguish 4 cases.

- If $n \in 2\mathbb{N}$ and $s \in 2\mathbb{N}$, say, $n = 2m$ and $s = 2t$, then

$$\begin{aligned} C(n+s)+C(n)+s &= C(2m+2t)+C(2m)+2t = 2(C(m+t)+C(m)+t) \\ C(2n+s) &= C(4m+2t) = 2C(2m+t) \end{aligned}$$

  and the desired inequality (4) is true because, by induction, $C(m+t)+C(m)+t \geqslant C(2m+t)$.

- If $n \in 2\mathbb{N}$ and $s \notin 2\mathbb{N}$, say, $n = 2m$ and $s = 2t+1$, then

$$\begin{aligned} C(n+s)+C(n)+s &= C(2m+2t+1)+C(2m)+2t+1 \\ &= C(m+t+1)+C(m+t)+1+2C(m)+2t+1 \\ C(2n+s) &= C(4m+2t+1) = C(2m+t+1)+C(2m+t)+1 \end{aligned}$$

  and (4) holds because, by induction, $C(m+t+1)+C(m)+t+1 \geqslant C(2m+t+1)$ and $C(m+t)+C(m)+t \geqslant C(2m+t)$

- If $n \notin 2\mathbb{N}$ and $s \notin 2\mathbb{N}$, say, $n = 2m+1$ and $s = 2t+1$, then

$$\begin{aligned} C(n+s)+C(n)+s &= C(2m+2t+2)+C(2m+1)+2t+1 \\ &= 2C(m+t+1)+C(m+1)+C(m)+1+2t+1 \\ C(2n+s) &= C(4m+2t+3) = C(2m+t+2)+C(2m+t+1)+1 \end{aligned}$$

  and (4) is true because, by induction, $C(m+t+1)+C(m+1)+t \geqslant C(2m+2+t)$ and $C(m+t+1)+C(m)+t+1 \geqslant C(2m+t+1)$.

- Assume finally that $n \notin 2\mathbb{N}$ and $s \in 2\mathbb{N}$: say, $n = 2m+1$ and $s = 2t$. If $t = 0$, then the desired inequality (4) amounts to $C(n)+C(n) \geqslant C(2n)$, which is true because it is actually an equality. So, assume that $t \geqslant 1$. In this case,

$$\begin{aligned} C(n+s)+C(n)+s &= C(2m+2t+1)+C(2m+1)+2t \\ &= C(m+t+1)+C(m+t)+1+C(m+1)+C(m)+1+2t \\ C(2n+s) &= C(4m+2+2t) = 2C(2m+t+1) \end{aligned}$$

  and the (4) holds because, by induction, $C(m+t+1)+C(m)+t+1 \geqslant C(2m+t+1)$ and

$$\begin{aligned} C(m+t)+C(m+1)+t+1 &= C((m+1)+(t-1))+C(m+1)+t-1+2 \\ &\geqslant C(2(m+1)+t-1)+2 > C(2m+t+1) \end{aligned}$$

This completes the proof of the inductive step. □

Remark 5. Notice that in the proof of Lemma 4 we have proved that the following two facts:

(a) $C(1+s)+C(1)+s = C(2+s)$ if, and only if, $s \leqslant 1$.

(b) If $s \geqslant 2$ is even and n is odd, then $C(n+s)+C(n)+s > C(2n+s)$.

Theorem 6. *For every $n \geqslant 1$, the minimum Colless index on $\mathcal{T}_n$ is $C(n)$.*

Proof. We shall prove by induction on n that $C(T) \geqslant C(n)$ for every $T \in \mathcal{T}_n$. The case when $n = 1$ is obvious, because $\mathcal{T}_1 = \{B_1\}$. Assume now that the assertion is true for every number of leaves smaller than n and let $T \in \mathcal{T}_n$. Let $T_1 \in \mathcal{T}_{n_1}$ and $T_2 \in \mathcal{T}_{n_2}$ be its subtrees rooted at the children of its root and assume that $n_1 \geqslant n_2$. Then

$$C(T) = C(T_1) + C(T_2) + n_1 - n_2 \geqslant C(n_1) + C(n_2) + n_1 - n_2 \geqslant C(n_1 + n_2) = C(n)$$

where the first inequality holds by the induction hypothesis and the second inequality by the previous lemma (taking in it $s = n_1 - n_2$). □

This theorem says that the minimum Colless index on $\mathcal{T}_n$ is reached at the maximally balanced bifurcating trees. When n is a power of 2, Lemma 1 guarantees that this minimum is reached exactly at these trees, which in this case are the fully symmetric bifurcating trees. But for arbitrary values of n there may be other trees whose Colles index is $C(n)$. For instance, the minimum Colless index on $\mathcal{T}_6$, $C(6) = 2$, is reached at the trees B_6 and $B_2 \star B_4$ depicted in Figure 2.

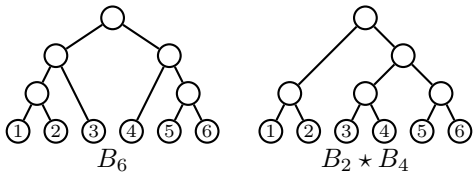

Figure 2: The trees in $\mathcal{T}_6$ with minimum Colless index.

4. Which trees have Colless index $C(n)$?

In this section we provide a way of generating all bifurcating trees with minimum Colless index among all bifurcating trees with their number of leaves.

Lemma 7. *Let $T \in \mathcal{T}_n$. Then, $C(T) = C(n)$ if, and only if, for every $v \in V_{int}(T)$, say with children v_1, v_2, $C(\kappa_T(v_1)) + C(\kappa_T(v_2)) + |\kappa_T(v_1) - \kappa_T(v_2)| = C(\kappa_T(v))$.*

Proof. $\Longrightarrow$) Assume that there exists some $v \in V_{int}(T)$, with children v_1, v_2, such that $C(\kappa_T(v_1)) + C(\kappa_T(v_2)) + |\kappa_T(v_1) - \kappa_T(v_2)| \neq C(\kappa_T(v))$. By Lemma 4, this implies that $C(\kappa_T(v_1)) + C(\kappa_T(v_2)) + |\kappa_T(v_1) - \kappa_T(v_2)| > C(\kappa_T(v))$. Let $T' \in \mathcal{T}_n$ be the tree obtained by replacing in T the rooted subtree T_v by a maximally balanced tree $B_{\kappa_T(v)}$ and leaving the rest of T untouched. In this way, $T'_v = B_{\kappa_T(v)}$ and $bal_T(x) = bal_{T'}(x)$ for every $x \in V_{int}(T) \setminus V_{int}(T_v) = V_{int}(T') \setminus V_{int}(T'_v)$; let us denote by W this last set of nodes. Then

$$\begin{aligned} C(T) &= \sum_{x \in W} bal_T(x) + C(T_v) = \sum_{x \in W} bal_T(x) + C(T_{v_1}) + C(T_{v_2}) + |\kappa_T(v_1) - \kappa_T(v_2)| \\ &\geqslant \sum_{x \in W} bal_T(x) + C(\kappa_T(v_1)) + C(\kappa_T(v_2)) + |\kappa_T(v_1) - \kappa_T(v_2)| \\ &> \sum_{x \in W} bal_T(x) + C(\kappa_T(v)) = \sum_{x \in W} bal_{T'}(x) + C(T'_v) = C(T') \geqslant C(n) \end{aligned}$$

This proves the "only if" implication.
$\Longleftarrow$) We prove the "if" implication by induction on n. The case when $n = 1$ is obvious, because $\mathcal{T}_1 = \{B_1\}$. Assume now that this implication is true for every tree in $\mathcal{T}_{n'}$ with $n' < n$, and let $T \in \mathcal{T}_n$ be such that,

for every $v \in V_{int}(T)$, $C(\kappa_T(v_1)) + C(\kappa_T(v_2)) + |\kappa_T(v_1) - \kappa_T(v_2)| = C(\kappa_T(v))$, where v_1, v_2 stand for the children of v. Let r be the root of T and x_1, x_2 its two children. Then, for every $v \in V_{int}(T_{x_1})$, with children v_1, v_2,

$$\begin{aligned} &C(\kappa_{T_{x_1}}(v_1)) + C(\kappa_{T_{x_1}}(v_2)) + |\kappa_{T_{x_1}}(v_1) - \kappa_{T_{x_1}}(v_2)| \\ &\quad = C(\kappa_T(v_1)) + C(\kappa_T(v_2)) + |\kappa_T(v_1) - \kappa_T(v_2)| = C(\kappa_T(v)) = C(\kappa_{T_{x_1}}(v)). \end{aligned}$$

This implies, by the induction hypothesis, that $C(T_{x_1}) = C(\kappa_T(x_1))$. By symmetry, we also have that $C(T_{x_2}) = C(\kappa_T(x_2))$. Finally,

$$\begin{aligned} C(T) &= C(T_{x_1}) + C(T_{x_2}) + |\kappa_T(x_1) - \kappa_T(x_2)| \\ &= C(\kappa_T(x_1)) + C(\kappa_T(x_2)) + |\kappa_T(x_1) - \kappa_T(x_2)| = C(\kappa_T(r)) = C(n) \end{aligned}$$

as we wanted to prove. □

This lemma, together with Lemma 1, provide the following algorithm to produce all trees $T \in \mathcal{T}_n$ such that $C(T) = C(n)$. In it, and henceforth, for every $n \geqslant 2$, let

$$QB(n) = \big\{(n_1, n_2) \in \mathbb{N}^2 \mid 1 \leqslant n_2 \leqslant n_1,\ n_1 + n_2 = n, C(n_1) + C(n_2) + n_1 - n_2 = C(n)\big\}.$$

Notice that $QB(n) \neq \emptyset$, because $(\lceil n/2 \rceil, \lfloor n/2 \rfloor) \in QB(n)$.

Algorithm 1.

1) Start with a single node labeled n.

2) While the current tree contains labeled leaves, choose a leaf with label m.

   2.a) If m is a power of 2, replace this leaf by a fully bifurcating tree B_m with its leaves unlabeled.

   2.b) If m is a not a power of 2:

      2.b.i) Find a pair of integers $(m_1, m_2) \in QB(m)$.

      2.b.ii) Split the leaf labeled m into a cherry with leaves labeled m_1 and m_2, respectively.

Next result provides a characterization of the pairs $(n_1, n_2) \in QB(n)$ which will allow to perform efficiently step (2.b.i) in this algorithm.

Proposition 8. *For every $n \geqslant 2$ and for every $1 \leqslant n_2 \leqslant n_1$ such that $n_1 + n_2 = n$:*

(1) If $n_1 = n_2 = n/2$, then $(n_1, n_2) \in QB(n)$ always.

(2) If $n_1 > n_2$, then $(n_1, n_2) \in QB(n)$ if, and only if, there exist $k \geqslant 0$, $l \geqslant 1$, $p \geqslant 1$, and $0 \leqslant t < 2^{l-2}$ such that one of the following three conditions holds:

- *There exist $k \geqslant 0$ and $p \geqslant 1$ such that $n = 2^k(2p+1)$, $n_1 = 2^k(p+1)$ and $n_2 = 2^k p$.*
- *There exist $k \geqslant 0$, $l \geqslant 2$, $p \geqslant 1$, and $0 \leqslant t < 2^{l-2}$ such that $n = 2^k(2^l(2p+1) - (2t+1))$, $n_1 = 2^k(2^l(p+1) - (2t+1))$, and $n_2 = 2^{k+l}p$.*
- *There exist $k \geqslant 0$, $l \geqslant 2$, $p \geqslant 1$, and $0 \leqslant t < 2^{l-2}$ such that $n = 2^k(2^l(2p+1) + 2t + 1)$, $n_1 = 2^{k+l}(p+1)$, and $n_2 = 2^k(2^l p + 2t + 1)$.*

We split the proof of this result into several auxiliary lemmas.

Lemma 9. *Let $s = 2^k t$ with $k \geqslant 1$ and $t \geqslant 1$. Then, for every $n \geqslant 1$, $(n+s, n) \in QB(2n+s)$ if, and only if, $n = 2^k m$, for some $m \geqslant 1$ such that $(m+t, m) \in QB(2m+t)$.*

Proof. We prove the equivalence in the statement by induction on the exponent $k \geqslant 1$. Recall that, by Remark 5.(b), if $s \geqslant 1$ is even and $C(n+s)+C(n)+s = C(2n+s)$, then n must be even, too. Therefore, if $s = 2t_0$, then $n = 2m_0$ for some $m_0 \geqslant 1$, and then, since $C(2m_0+2t_0)+C(2m_0)+2t_0 = 2(C(m_0+t_0)+C(m_0)+t_0)$ and $C(4m_0+2t_0) = 2C(2m_0+t_0)$, the equality $C(n+s)+C(n)+s = C(2n+s)$ is equivalent to the equality $C(m_0+t_0)+C(m_0)+t_0 = C(2m_0+t_0)$. This proves the equivalence in the statement when $k = 1$.

Now, assume that this equivalence is true for the exponent $k-1$, and let $s = 2^k t$. Then, by the case $k = 1$, $C(n+s)+C(n)+s = C(2n+s)$ if, and only if, $n = 2m_0$ for some $m_0 \geqslant 1$ such that $C(m_0+2^{k-1}t)+C(m_0)+2^{k-1}t = C(2m_0+2^{k-1}t)$, and, by the induction hypothesis, this last equality holds if, and only if, $m_0 = 2^{k-1}m$ for some $m \geqslant 1$ such that $C(m+t)+C(m)+t = C(2m+t)$. Combining both equivalences we obtain the equivalence in the statement, thus proving the inductive step. □

Lemma 10. *Let $s = 2^{k+1}-(2t+1)$, with $k = \lfloor \log_2(s) \rfloor$ and $0 \leqslant t < 2^{k-1}$, be an odd integer. Then, for every $m \geqslant 1$, $(2m+s, 2m) \in QB(4m+s)$ if, and only if, $m = 2^k p$ for some $p \geqslant 1$.*

Proof. We prove the equivalence in the statement by induction on s. When $s = 1 = 2^1 - 1$, the equivalence says that $C(2m+1)+C(2m)+1 = C(4m+1)$ for every $m \geqslant 1$, which is true by (3).

Assume now that the equivalence is true for every odd natural number $s' < s$ and for every m, and let us prove it for $s = 2^{k+1}-(2t+1)$ with $0 \leqslant t < 2^{k-1}$. We have that

$$\begin{aligned} &C(2m+2^{k+1}-2t-1)+C(2m)+2^{k+1}-2t-1 \\ &\quad = \big(C(m+2^k-t)+C(m)+2^k-t\big)+\big(C(m+2^k-t-1)+C(m)+2^k-t-1\big)+1 \\ &C(4m+2^{k+1}-2t-1) = C(2m+2^k-t)+C(2m+2^k-t-1)+1 \end{aligned}$$

and since, by Lemma 4, $C(m+2^k-t)+C(m)+2^k-t \geqslant C(2m+2^k-t)$ and $C(m+2^k-t-1)+C(m)+2^k-t-1 \geqslant C(2m+2^k-t-1)$, we have that $C(2m+s)+C(2m)+s = C(4m+s)$ if, and only if, the following two identities are satisfied:

$$C(m+2^k-t)+C(m)+2^k-t = C(2m+2^k-t) \tag{6}$$

$$C(m+2^k-t-1)+C(m)+2^k-t-1 = C(2m+2^k-t-1) \tag{7}$$

So, we must prove that (6) and (7) hold if, and only if, $m = 2^k p$ for some $p \geqslant 1$. We distinguish two subcases, depending on the parity of t:

- If $t = 2x$ for some $0 \leqslant x < 2^{k-2}$, then (6) and Lemma 9 imply that m is even, say $m = 2m_0$, and then (7) says
$$C(2m_0+2^k-2x-1)+C(2m_0)+2^k-2x-1 = C(4m_0+2^k-2x-1), \tag{8}$$
which, by induction, is equivalent to $m_0 = 2^{k-1}p$ for some $p \geqslant 1$, i. e., to $m = 2^k p$ for some $p \geqslant 1$. So, to complete the proof of the desired equivalence, it remains to prove that if $m = 2^k p$, then (6) holds. If $t = 0$, this equality is a direct consequence of Lemma 9 and (3), so assume that $t > 0$ and write it as $t = 2^j(2x_0+1)$ with $1 \leqslant j < k-1$ and $x_0 < 2^{k-j-2}$. Then

$$\begin{aligned} C(m+2^k-t)+C(m)+2^k-t &= C(2^k p+2^k-2^j(2x_0+1))+C(2^k p)+2^k-2^j(2x_0+1) \\ &= 2^j\big(C(2^{k-j}p+2^{k-j}-2x_0-1)+C(2^{k-j}p)+2^{k-j}-2x_0-1\big) \\ &= 2^j C(2^{k-j+1}p+2^{k-j}-2x_0-1) \text{ (by the induction hypothesis)} \\ &= C(2^{k+1}p+2^k-2^j(2x_0+1)) = C(2m+2^k-t) \end{aligned}$$

- If $t = 2x+1$ for some $0 \leqslant x < 2^{k-2}$, then (7) and Lemma 9 imply that m is even, say $m = 2m_0$, and then it is (6) which is equivalent to equation (8) above, which, on its turn, by induction is equivalent to $m_0 = 2^{k-1}p$ for some $p \geqslant 1$, that is, to $m = 2^k p$ for some $p \geqslant 1$ Thus, to complete the proof of the desired equivalence, it remains to prove that if $m = 2^k p$, then (7) holds. Now:

$$\begin{aligned} C(m+2^k-t-1)+C(m)+2^k-t-1 &= C(2^k p+2^k-2x-2)+C(2^k p)+2^k-2x-2 \\ &= 2\big(C(2^{k-1}p+2^{k-1}-x-1)+C(2^{k-1}p)+2^{k-1}-x-1\big) \end{aligned}$$

Now, if x is even, say $x = 2x_0$, then, since $x_0 < 2^{k-3}$, the induction hypothesis implies that

$$\begin{aligned}2\big(C(2^{k-1}p + 2^{k-1} - x - 1) + C(2^{k-1}p) + 2^{k-1} - x - 1\big) &= 2C(2^k p + 2^{k-1} - x - 1)\\ &= C(2^{k+1}p + 2^k - 2x - 2) = C(2m + 2^k - t - 1)\end{aligned}$$

And if x is odd, write it as $x = 2^j(2t_0 + 1) - 1$ for some $1 \leqslant j < k - 1$ (and notice that $x < 2^{k-2}$ implies $t_0 < 2^{k-j-3}$) and then

$$\begin{aligned}&2\big(C(2^{k-1}p + 2^{k-1} - x - 1) + C(2^{k-1}p) + 2^{k-1} - x - 1\big)\\ &\quad= 2\big(C(2^{k-1}p + 2^{k-1} - 2^j(2t_0 + 1)) + C(2^{k-1}p) + 2^{k-1} - 2^j(2t_0 + 1)\big)\\ &\quad= 2 \cdot 2^j\big(C(2^{k-j-1}p + 2^{k-j-1} - (2t_0 + 1)) + C(2^{k-j-1}p) + 2^{k-j-1} - (2t_0 + 1)\big)\\ &\quad= 2^{j+1}C(2^{k-j}p + 2^{k-j-1} - (2t_0 + 1)) \text{ (by the induction hypothesis)}\\ &\quad= C(2^{k+1}p + 2^k - 2^{j+1}(2t_0 + 1)) = C(2^{k+1}p + 2^k - 2x - 2) = C(2m + 2^k - t - 1)\end{aligned}$$

This completes the proof of the desired equivalence when t is odd.

So, the inductive step is true in all cases. □

Lemma 11. *Let $s = 2^{k+1} - (2t + 1)$, with $k = \lfloor \log_2(s) \rfloor$ and $0 \leqslant t < 2^{k-1}$, be an odd integer. Then, for every $m \geqslant 0$, $(2m + 1 + s, 2m + 1) \in QB(4m + 2 + s)$ if, and only if, either $m = 2^k p + t$ for some $p \geqslant 1$ or $s = 1$ (i.e., $k = t = 0$) and $m = 0$.*

Proof. We prove the equivalence in the statement by induction on s. When $s = 1 = 2^1 - 1$, the equivalence says that $C(2m + 2) + C(2m + 1) + 1 = C(4m + 3)$ for every $m \geqslant 0$, which is true by (3).

Assume now that the equivalence is true for every odd natural number $s' < s$ and for every $m \geqslant 0$, and let us prove it for $s = 2^{k+1} - (2t + 1)$ with $0 \leqslant t < 2^{k-1}$. In this case, if $m = 0$, then by Remark 5.(a) we know that $(s + 1, s) \in QB(s + 2)$ if, and only if, $s = 1$. So, assume that $m \geqslant 1$. Then, we have that

$$\begin{aligned}&C(2m + 1 + 2^{k+1} - 2t - 1) + C(2m + 1) + 2^{k+1} - 2t - 1\\ &\quad= \big(C(m + 2^k - t) + C(m) + 2^k - t\big) + \big(C(m + 2^k - t) + C(m + 1) + 2^k - t - 1\big) + 1\\ &C(4m + 2 + 2^{k+1} - 2t - 1) = C(2m + 2^k - t) + C(2m + 2^k - t + 1) + 1\end{aligned}$$

and since, by Lemma 4, $C(m+2^k-t)+C(m)+2^k-t \geqslant C(2m+2^k-t)$ and $C(m+2^k-t)+C(m+1)+2^k-t-1 \geqslant C(2m + 2^k - t + 1)$, we have that $C(2m + 1 + s) + C(2m + 1) + s = C(4m + 2 + s)$ if, and only if,

$$C(m + 2^k - t) + C(m) + 2^k - t = C(2m + 2^k - t) \tag{9}$$

$$C(m + 2^k - t) + C(m + 1) + 2^k - t - 1 = C(2m + 2^k - t + 1) \tag{10}$$

So, we must prove that (9) and (10) hold for $m \geqslant 1$ if, and only if, $m = 2^k p + t$ for some $p \geqslant 1$. We distinguish again two subcases, depending on the parity of t:

- If $t = 2x$ for some $0 \leqslant x < 2^{k-2}$, then (9) and Lemma 9 imply that m is even, say $m = 2m_0$ with $m_0 \geqslant 1$, and then (10) can be written

  $$C(2m_0 + 1 + 2^k - 2x - 1) + C(2m_0 + 1) + 2^k - 2x - 1 = C(4m_0 + 2 + 2^k - 2x - 1)$$

  which, by induction, is equivalent to $m_0 = 2^{k-1}p + x$ for some $p \geqslant 1$, that is, to $m = 2^k p + t$ for some $p \geqslant 1$. So, to complete the proof of the desired equivalence, it remains to check that if $m = 2^k p + t$, then (9) holds. Now, if $x = 0$, so that $m = 2^k p$, (3) and Lemma 9 clearly imply (9). So, assume that $x > 0$ and write it as $x = 2^j(2y_0 + 1)$ with $0 \leqslant j < k - 2$ and $y_0 < 2^{k-j-3}$. Then

  $$\begin{aligned}&C(m + 2^k - t) + C(m) + 2^k - t = C(2^k p + 2x + 2^k - 2x) + C(2^k p + 2x) + 2^k - 2x\\ &\quad= C(2^k p + 2^{j+1}(2y_0 + 1) + 2^k - 2^{j+1}(2y_0 + 1)) + C(2^k p + 2^{j+1}(2y_0 + 1)) + 2^k - 2^{j+1}(2y_0 + 1)\\ &\quad= 2^{j+1}\big(C(2^{k-j-1}p + 2y_0 + 1 + 2^{k-j-1} - (2y_0 + 1)) + C(2^{k-j-1}p + 2y_0 + 1) + 2^{k-j-1} - (2y_0 + 1)\big)\\ &\quad= 2^{j+1}C(2^{k-j}p + 4y_0 + 2 + 2^{k-j-1} - (2y_0 + 1)) \text{ (by the induction hypothesis)}\\ &\quad= C(2^{k+1}p + 2^k + 2^{j+1}(2y_0 + 1)) = C(2^{k+1}p + 2^k + 2x) = C(2m + 2^k - t)\end{aligned}$$

  as we wanted to prove.

- If $t = 2x + 1$ for some $0 \leqslant x < 2^{k-2}$, (10) and Lemma 9 imply that $m + 1$ is even, and then m is odd, say $m = 2m_0 + 1$ for some $m_0 \geqslant 0$, and (9) can be written

$$C((2m_0 + 1) + 2^k - 2x - 1) + C(2m_0 + 1) + 2^k - 2x - 1 = C(4m_0 + 2 + 2^k - 2x - 1). \qquad (11)$$

Now, if $m_0 = 0$, Remark 5.(a) implies that this equality holds if, and only if, $2^k - 2x - 1 = 1$ which, under the condition $0 \leqslant x < 2^{k-2}$, only happens when $k = 1$ and $x = 0$, but then $t = 1 = 2^{k-1}$ against the assumption that $t < 2^{k-1}$. Therefore m_0 must be at least 1.

Then, by induction, identity (11) is equivalent to $m_0 = 2^{k-1}p+x$ for some $p \geqslant 1$, that is, to $m = 2^k p+t$ for some $p \geqslant 1$. So, to complete the proof of the desired equivalence, it remains to check that if $m = 2^k p + t$, then (10) holds. Now, in the current situation:

$$\begin{aligned} &C(m + 2^k - t) + C(m + 1) + 2^k - t - 1 \\ &\quad = C(2^k p + 2x + 1 + 2^k - 2x - 1) + C(2^k p + 2x + 2) + 2^k - 2x - 2 \\ &\quad = 2\big(C((2^{k-1}p + x + 1) + (2^{k-1} - x - 1)) + C(2^{k-1}p + x + 1) + 2^{k-1} - x - 1\big) = (**) \end{aligned}$$

If x is even, say $x = 2x_0$ with $0 \leqslant x_0 < 2^{k-3}$, then

$$\begin{aligned} (**) &= 2\big(C((2^{k-1}p + 2x_0 + 1) + (2^{k-1} - 2x_0 - 1)) + C(2^{k-1}p + 2x_0 + 1) + 2^{k-1} - 2x_0 - 1\big) \\ &= 2C(2^k p + 2(2x_0 + 1) + 2^{k-1} - (2x_0 + 1)) \text{ (by induction)} \\ &= C(2^{k+1}p + 2^k + 4x_0 + 2) = C(2m + 2^k - t + 1) \end{aligned}$$

And if x is odd, write it as $x = 2^j(2t_0 + 1) - 1$ with $1 \leqslant j < k - 1$ and $t_0 < 2^{k-j-3}$, and then

$$\begin{aligned} (**) &= 2\big(C(2^{k-1}p + 2^j(2t_0 + 1) + 2^{k-1} - 2^j(2t_0 + 1)) + C(2^{k-1}p + 2^j(2t_0 + 1)) + 2^{k-1} - 2^j(2t_0 + 1)\big) \\ &= 2^{j+1}\big(C(2^{k-j-1}p + 2t_0 + 1 + 2^{k-j-1} - (2t_0 + 1)) + C(2^{k-j-1}p + 2t_0 + 1) + 2^{k-j-1} - (2t_0 + 1)\big) \\ &= 2^{j+1}C(2^{k-j}p + 4t_0 + 2 + 2^{k-j-1} - (2t_0 + 1)) \text{ (by the induction hypothesis)} \\ &= C(2^{k+1}p + 2^{j+1}(2t_0 + 1) + 2^k) = C(2^{k+1}p + 2x + 2 + 2^k) = C(2m + 2^k - t + 1) \end{aligned}$$

This completes the proof of the desired equivalence when t is odd.

□

We can return now to Proposition 8.

Proof of Proposition 8. Assertion (1) is a direct consequence of identity (3). So, assume $n_1 > n_2$ and set $s = n_1 - n_2$, so that $n_1 = n_2 + s$. Then:

(a) If $s = 1$, then, by Lemma 11, $C(n_1) + C(n_2) + n_1 - n_2 = C(n_1 + n_2)$ for every $n_2 \geqslant 1$.

(b) If $s > 1$ is odd, write it as $s = 2^{j+1} - (2t + 1)$, with $j = \lfloor \log_2(s) \rfloor \geqslant 1$ and $0 \leqslant t < 2^{j-1}$. Then, by Lemmas 10 and 11, $C(n_1) + C(n_2) + n_1 - n_2 = C(n_1 + n_2)$ if, and only if, either $n_2 = 2^{j+1}p$ or $n_2 = 2^{j+1}p + 2t + 1$, for some $p \geqslant 1$.

(c) If $s \geqslant 2$ is even, write it as $s = 2^k s_0$, with $k \geqslant 1$ the largest exponent of a power of 2 that divides s and s_0 an odd integer, which we write $s_0 = 2^{j+1} - (2t + 1)$ with $j = \lfloor \log_2(s_0) \rfloor$ and $0 \leqslant t < 2^{j-1}$. Then, by Lemma 9, $C(n_1) + C(n_2) + n_1 - n_2 = C(n_1 + n_2)$ if, and only if, $n_2 = 2^k m$, for some $m \geqslant 1$ such that $C(m + s_0) + C(m) + s_0 = C(2m + s_0)$, and then:

- If $s_0 = 1$, $C(m + s_0) + C(m) + s_0 = C(2m + s_0)$ for every $m \geqslant 1$ and therefore, in this case, $C(n_1) + C(n_2) + n_1 - n_2 = C(n_1 + n_2)$ for every $n_2 = 2^k m$ with $m \geqslant 1$.
- If $s_0 > 1$, Lemmas 10 and 11 imply that $C(m + s_0) + C(m) + s_0 = C(2m + s_0)$ if, and only if, $m = 2^{j+1}p$ or $m = 2^{j+1}p+2t+1$, for some $p \geqslant 1$. Therefore, in this case, $C(n_1)+C(n_2)+n_1-n_2 = C(n_1 + n_2)$ if, and only if, $n_2 = 2^{k+j+1}p$ or $n_2 = 2^k(2^{j+1}p + 2t + 1)$, for some $p \geqslant 1$.

Combining the three cases, and taking $k = 0$ in the odd s case, we conclude that $C(n_1)+C(n_2)+n_1-n_2 = C(n_1+n_2)$ if, and only if, $n_1 - n_2 = 2^k(2^{j+1} - (2t+1))$ (for some $k \geqslant 0$, $j \geqslant 0$, and $0 \leqslant t < 2^{j-1}$) and

- If $j = t = 0$, then $n_2 = 2^k p$ for some $p \geqslant 1$, in which case $n_1 = 2^k p + 1$ and $n = 2^{k+1}p + 1$
- If $j > 0$ or $t > 0$, then one of the following two conditions holds for some $p \geqslant 1$:
  - $n_2 = 2^{k+j+1}p$, in which case $n_1 = 2^k(2^{j+1}(p+1) - (2t+1))$ and $n = 2^k(2^{j+1}(2p+1) - (2t+1))$; or
  - $n_2 = 2^k(2^{j+1}p + 2t + 1)$, $n_1 = 2^{k+j+1}(p+1)$ and $n = 2^k(2^{j+1}(2p+1) + 2t + 1)$

This is equivalent to the expressions for n_1 and n_2 in option (2) in the statement (replacing $j+1$ by $l \geqslant 1$). □

Proposition 12. *For every $n \geqslant 2$, let $k \geqslant 0$ be the exponent of the largest power of 2 that divides n, let $n_0 = n/2^k$, and let $n_0 = \sum_{i=1}^{\ell} 2^{m_i}$, with $m_1 > \cdots > m_{\ell-1} > m_\ell = 0$, be the binary expansion of n_0. Then*

(a) If $\ell = 1$, i.e., if $n = 2^k$, then $QB(n) = \{(n/2, n/2)\}$.

(b) If $\ell > 1$:

(b.1) $QB(n)$ always contains the pair

$$\Big(2^k\Big(\sum_{i=1}^{\ell-1} 2^{m_i-1} + 1\Big), 2^k \sum_{i=1}^{\ell-1} 2^{m_i-1}\Big).$$

(b.2) For every $j = 2, \ldots, \ell-1$ such that $m_j > m_{j+1} + 1$, $QB(n)$ contains the pair

$$\Big(2^k\Big(\sum_{i=1}^{j-1} 2^{m_i-1} + 2^{m_j}\Big), n - 2^k\Big(\sum_{i=1}^{j-1} 2^{m_i-1} + 2^{m_j}\Big)\Big).$$

(b.3) For every $j = 2, \ldots, \ell-1$ such that $m_j < m_{j-1} - 1$, $QB(n)$ contains the pair

$$\Big(n - 2^k \sum_{i=1}^{j-1} 2^{m_i-1}, 2^k \sum_{i=1}^{j-1} 2^{m_i-1}\Big).$$

(b.4) If $k \geqslant 1$, then $QB(n)$ contains the pair $(n/2, n/2)$.

The pairs described in (b.1) to (b.4) are pairwise different, and $QB(n)$ contains no other member.

Proof. Assertion (a) is obvious by Lemma 1. So, assume henceforth that $\ell > 1$. Let now (n_1, n_2) such that $n = n_1 + n_2$ and $1 \leqslant n_2 < n_1$. Then, by Proposition 8, $(n_1, n_2) \in QB(n)$ if, and only if, one of the following three conditions is satisfied:

(b.1) There exist $k \geqslant 0$ and $p \geqslant 1$ such that $n_0 = 2p+1$, $n_1 = 2^k(p+1)$, and $n_2 = 2^k p$. In this case $p = (n_0 - 1)/2 = \sum_{i=1}^{\ell-1} 2^{m_i-1}$ and this contributes to $QB(n)$ the pair (n_1, n_2) with

$$n_1 = 2^k\Big(\sum_{i=1}^{\ell-1} 2^{m_i-1} + 1\Big), \quad n_2 = 2^k \sum_{i=1}^{\ell-1} 2^{m_i-1}.$$

(b.2) There exist $k \geqslant 0$, $l \geqslant 2$, $p \geqslant 1$, and $0 \leqslant t < 2^{l-2}$ such that $n_0 = 2^{l+1}p + 2^l + 2t + 1$ and $n_1 = 2^{k+l}(p+1)$. Now, if $t < 2^{l-2}$ and $p \geqslant 1$, then $2t + 1 < 2^{l-1}$ and $2^{l+1}p \geqslant 2^{l+1}$. Therefore, the equality

$$2^{l+1}p + 2^l + 2t + 1 = \sum_{i=1}^{\ell} 2^{m_i}$$

holds for some $p \geqslant 1$ and $t < 2^{l-2}$ if, and only if, $m_j = l \geqslant 2$ for some $j = 2, \dots, \ell - 1$, in which case $p = \big(\sum_{i=1}^{j-1} 2^{m_i}\big)/2^{m_j+1}$. This contributes to $QB(n)$ all pairs (n_1, n_2) of the form

$$n_1 = 2^{k+m_j}\Big(\frac{\sum_{i=1}^{j-1} 2^{m_i}}{2^{m_j+1}} + 1\Big) = 2^k\Big(\sum_{i=1}^{j-1} 2^{m_i-1} + 2^{m_j}\Big),\ n_2 = n - 2^k\Big(\sum_{i=1}^{j-1} 2^{m_i-1} + 2^{m_j}\Big), \tag{12}$$

with $j = 2, \dots, \ell - 1$ and $m_j \geqslant 2$. But not all these pairs are different, because $\sum_{i=1}^{j-1} 2^{m_i-1} + 2^{m_j} = \sum_{i=1}^{j} 2^{m_i-1} + 2^{m_{j+1}}$ if, and only if, $m_j = m_{j+1} + 1$. Indeed,

$$\sum_{i=1}^{j-1} 2^{m_i-1} + 2^{m_j} = \sum_{i=1}^{j} 2^{m_i-1} + 2^{m_{j+1}} \iff 2^{m_j} = 2^{m_j-1} + 2^{m_{j+1}}$$
$$\iff 2^{m_j-1} = 2^{m_{j+1}} \iff m_j = m_{j+1} + 1.$$

This entails that each sequence of values of consecutive exponents m_j that fiffer only in 1 yield the same n_1, and hence the same pair (n_1, n_2). On the other hand, $\sum_{i=1}^{j-1} 2^{m_i-1} + 2^{m_j}$ is clearly monotonously non-increasing on j (because $m_{j+1} < m_j$), and therefore, by the previous equivalence, its value jumps at each j such that $m_j > m_{j+1} + 1$. Finally, if it happens that $m_{\ell-2} = 2$, then $m_{\ell-1} = 1 = m_{\ell-2} - 1$ and $m_\ell = 0 = m_{\ell-1} - 1$ and hence, as we have just seen,

$$\sum_{i=1}^{\ell-3} 2^{m_i-1} + 2^{m_{\ell-2}} = \sum_{i=1}^{\ell-2} 2^{m_i-1} + 2^{m_{\ell-1}} = \sum_{i=1}^{\ell-1} 2^{m_i-1} + 2^{m_\ell} = \sum_{i=1}^{\ell-1} 2^{m_i-1} + 1$$

and the value of n_1 that we obtain in this case, $2^k\big(\sum_{i=1}^{\ell-1} 2^{m_i-1} + 1\big)$, is the one already given in (b.1), so we can omit it. In summary, the new pairwise different pairs (n_1, n_2) of this form are obtained by taking in (12) as l the exponents m_j with $j = 2, \dots, \ell - 1$ such that $m_j > m_{j+1} + 1$.

(b.3) There exist $k \geqslant 0$, $l \geqslant 2$, $p \geqslant 1$, and $0 \leqslant t < 2^{l-2}$ such that $n_0 = 2^{l+1}p + 2^l - (2t+1)$ and $n_2 = 2^{k+l}p$. Since $t < 2^{l-2}$, we have that $n_0 = 2^{l+1}p + 2^{l-1} + 2t_0 + 1$ with $2t_0 + 1 < 2^{l-1}$. Then, the equality

$$2^{l+1}p + 2^{l-1} + 2t_0 + 1 = \sum_{i=1}^{\ell} 2^{m_i}$$

holds for some $p \geqslant 1$ and $t_0 < 2^{l-2}$ if, and only if, $l - 1 = m_j$ for some $j = 2, \dots, \ell - 1$ such that $m_{j-1} > m_j + 1$, and then

$$p = \frac{\sum_{i=1}^{j-1} 2^{m_i}}{2^{m_j+2}}.$$

This contributes to $QB(n)$ all pairs (n_1, n_2) of the form

$$n_2 = 2^{k+m_j+1}\Big(\frac{\sum_{i=1}^{j-1} 2^{m_i}}{2^{m_j+2}}\Big) = 2^k \sum_{i=1}^{j-1} 2^{m_i-1}, \quad n_1 = n - 2^k \sum_{i=1}^{j-1} 2^{m_i-1},$$

with $j = 2, \dots, \ell - 1$ such that $m_j < m_{j-1} - 1$, belong to $QB(n)$, and they are pairwise different.

This gives all pairs (n_1, n_2) in $QB(n)$ with $n_1 > n_2$. If n is even, we must add moreover to $QB(n)$ the pair $(n/2, n/2)$ and this completes the set of pairs belonging to $QB(n)$. To finish the proof of the statement, we must check that these pairs are pairwise different.

Now, along our construction we have already checked that the pairs of the form (b.2), as well as those of the form (b.3), are pairwise different. We have also checked that the pairs of the form (b.2) and different from the pair (b.1). The pairs of the form (b.3) are also different from the pair (b.1), because, since $j \leqslant \ell - 1$, their entry n_2 is strictly smaller than the corresponding entry in (b.1). On the other hand, if the pair $(n/2, n/2)$

is added to $QB(n)$, it is not of the form (b.1) to (b.3), because all these pairs have both entries divisible by 2^k, while the maximum power of 2 that divides $n/2$ is 2^{k-1}. Finally, if (n_1, n_2) is a pair of the form (b.2), then $n_1/2^k$ is even and $n_2/2^k$ is odd, while if (n_1, n_2) is a pair of the form (b.3), then $n_1/2^k$ is odd and $n_2/2^k$ is even. Therefore, no pair can be simultaneously of the form (b.2) and (b.3). □

Example 13. Let us find $QB(214)$. Since $214 = 2(2^6 + 2^5 + 2^3 + 2 + 1)$, with the notations of the last corollary we have that $k = 1$, $\ell = 5$, $m_1 = 6$, $m_2 = 5$, $m_3 = 3$, $m_4 = 1$, and $m_5 = 0$. Then:

(b.1) The pair of this type in $QB(214)$ is $\big(2^k(\sum_{i=1}^4 2^{m_i-1} + 1), 2^k \sum_{i=1}^4 2^{m_i-1}\big) = (108, 106)$.

(b.2) The indices $j \in \{2, 3, 4\}$ such that $m_j > m_{j+1} + 1$ are 2 and 3. Therefore, the pairs of this type in $QB(214)$ are:

- For $j = 2$, $\big(2^k(2^{m_1-1} + 2^{m_2}), n - 2^k(2^{m_1-1} + 2^{m_2})\big) = (128, 86)$.
- For $j = 3$, $\big(2^k(2^{m_1-1} + 2^{m_2-1} + 2^{m_3}), n - 2^k(2^{m_1-1} + 2^{m_2-1} + 2^{m_3})\big) = (112, 102)$.

(b.3) The indices $j \in \{2, 3, 4\}$ such $m_j < m_{j-1} - 1$ are 3 and 4. Therefore, the pairs of this type in $QB(214)$ are:

- For $j = 3$, $\big(n - 2^k(2^{m_1-1} + 2^{m_2-1}), 2^k(2^{m_1-1} + 2^{m_2-1})\big) = (118, 96)$.
- For $j = 4$, $\big(n - 2^k(2^{m_1-1} + 2^{m_2-1} + 2^{m_3-1}), 2^k(2^{m_1-1} + 2^{m_2-1} + 2^{m_3-1})\big) = (110, 104)$.

(b.4) Since $214 = 2 \cdot 107$ is even, $QB(214)$ contains the pair $(107, 107)$.

Therefore

$$QB(214) = \big\{(107, 107), (108, 106), (110, 104), (112, 102), (118, 96), (128, 86)\big\}.$$

Corollary 14. *For every $n \geqslant 2$, the cardinality of $QB(n)$ is at most $\lfloor \log_2(n) \rfloor$.*

Proof. Let $n_{(2)}$ denote the binary representation of n. If n is a power of 2, then $|QB(n)| = 1 \leqslant \lfloor \log_2(n) \rfloor$. Assume henceforth that n is not a power of 2. In this case, by construction, the number of pairs of type (b.2) in $QB(n)$ is the number of maximal sequences of zeroes in $n_{(2)}$ that do not end immediately before the last 1 or in the units position; the number of pairs of type (b.3) in $QB(n)$ is the number of maximal sequences of zeroes in $n_{(2)}$ that do not start immediately after the leading 1 or that do not end in the units position; there is one pair of type (b.4) in $QB(n)$ if $n_{(2)}$ contains a sequence of zeroes ending in the units position; and $QB(n)$ always contains a pair of the form (b.1). So, if we denote by $M_0(n)$ the number of maximal sequences of zeroes in $n_{(2)}$, to compute the cardinality $|QB(n)|$:

- We count twice the number of maximal sequences of zeroes in $n_{(2)}$ plus 1, $2M_0(n) + 1$
- We subtract 1 if $n_{(2)}$ contains a maximal sequence of zeroes starting immediately after the leading 1
- We subtract 1 if $n_{(2)}$ contains a maximal sequence of zeroes ending immediately before the last 1
- We subtract 2 and we add 1 (i.e., we subtract 1) if $n_{(2)}$ contains a maximal sequence of zeroes ending in the units position

So, if, to simplify the language, we call *forbidden* any maximal sequence of zeroes in $n_{(2)}$ that starts immediately after the leading 1 or ends immediately before the last 1 or in the units position, we have the formula

$$|QB(n)| = 2M_0(n) + 1 \quad \begin{array}{l}\text{minus the number of forbidden maximal sequences} \\ \text{of zeroes in } n_{(2)},\end{array}$$

where in the subtraction we count each forbidden maximal sequence of zeroes as many times as it satisfies a "forbidden" property. So, a maximal sequence of zeroes starting immediately after the leading 1 and ending immediately before the last 1 subtracts 2.

Now, on the one hand, if $\lfloor \log_2(n) \rfloor$ is an even number, by the pigeonhole principle we have that $M_0(n) \leqslant \lfloor \log_2(n) \rfloor/2$. But if $n_{(2)}$ does not contain any forbidden maximal sequence of zeroes, then $n_{(2)}$ starts with 11 and ends with 11 and the number of maximal sequences of zeroes in such an $n_{(2)}$ is at most $\lfloor \log_2(n) \rfloor/2-1$. So, if $M_0(n) = \lfloor \log_2(n) \rfloor/2$, then $n_{(2)}$ contains some forbidden maximal sequence of zeroes and then $|QB(n)| \leqslant 2M_0(n) = \lfloor \log_2(n) \rfloor$, while if $M_0(n) \leqslant \lfloor \log_2(n) \rfloor/2 - 1$, then $|QB(n)| \leqslant 2M_0(n) + 1 \leqslant \lfloor \log_2(n) \rfloor - 1$.

On the other hand, if $\lfloor \log_2(n) \rfloor$ is an odd number, again by the pigeonhole principle we have that $M_0(n) \leqslant (\lfloor \log_2(n) \rfloor + 1)/2$. Now, if $M_0(n) = (\lfloor \log_2(n) \rfloor + 1)/2$, then $n_{(2)}$ contains at least 2 forbidden maximal sequences of zeroes. Indeed, let $\lfloor \log_2(n) \rfloor = 2s + 1$. If $n_{(2)}$ starts with 11, avoiding a forbidden maximal sequence of zeroes at the beginning, then $M_0(n) \leqslant s = (\lfloor \log_2(n) \rfloor - 1)/2$. On the other hand, if it ends in 11, avoiding a forbidden maximal sequence of zeroes at the end, then again $M_0(n) \leqslant s = (\lfloor \log_2(n) \rfloor - 1)/2$. So, to reach the maximum value of $M_0(n)$, $n_{(2)}$ must start with 10 and end with 10, 01 or 00, thus having at least 2 forbidden maximal sequences of zeroes. Thus, if $M_0(n) = (\lfloor \log_2(n) \rfloor + 1)/2$, then $|QB(n)| \leqslant 2M_0(n) - 1 = \lfloor \log_2(n) \rfloor$, while if $M_0(n) \leqslant (\lfloor \log_2(n) \rfloor + 1)/2 - 1$, then $|QB(n)| \leqslant 2M_0(n) + 1 \leqslant \lfloor \log_2(n) \rfloor$. □

Proposition 7, together with Lemma 1, provide the following algorithm to produce all trees $T \in \mathcal{T}_n$ such that $C(T) = C(n)$, which is reminiscent of Aldous' β-model [1].

Algorithm MinColless.

1) Start with a single node labeled n.

2) While the current tree contains labeled leaves, choose a leaf with label m.

   2.a) If m is a power of 2, replace this leaf by a fully bifurcating tree B_m with its nodes unlabeled.

   2.b) If m is a not a power of 2:

      2.b.i) Find a pair of integers $(m_1, m_2) \in QB(m)$.

      2.b.ii) Split the leaf labeled m into a cherry with unlabeled root and its leaves labeled m_1 and m_2, respectively.

Example 15. Let us use Algorithm MinColless to find all bifurcating trees with 20 leaves and minimum Colless index; we describe the trees by means of the usual Newick format,[1] with the unlabeled leaves represented by a symbol $*$ and omitting the semicolon ending mark in order not to confuse it with a punctuation mark.

1) We start with a single node labeled 20.

2) Since $QB(20) = \{(10, 10), (12, 8)\}$, this node splits into the cherries $(10, 10)$ and $(12, 8)$.

3.1) Since $QB(10) = \{(5, 5), (6, 4)\}$, the different ways of splitting the leaves of the tree $(10, 10)$ produce the trees $((5, 5), (5, 5))$, $((5, 5), (6, 4))$, and $((6, 4), (6, 4))$. Now, since $QB(5) = \{(3, 2)\}$, $QB(6) = \{(3, 3), (4, 2)\}$, and $QB(3) = \{(2, 1)\}$, and 1, 2, and 4 are powers of 2, we have the following derivations

[1] See `http://evolution.genetics.washington.edu/phylip/newicktree.html`

from these trees through all possible combinations of splitting the leaves in the trees:

$$((5,5),(5,5)) \Rightarrow (((3,2),(3,2)),((3,2),(3,2))) \Rightarrow ((((2,1),2),((2,1),2)),(((2,1),2),((2,1),2)))$$
$$\Rightarrow (((((*,*),*),(*,*)),(((*,*),*),(*,*))),((((*,*),*),(*,*)),(((*,*),*),(*,*))))$$
$$((5,5),(6,4)) \Rightarrow (((3,2),(3,2)),((3,3),4)) \Rightarrow ((((2,1),2),((2,1),2)),(((2,1),(2,1)),4))$$
$$\Rightarrow (((((*,*),*),(*,*)),(((*,*),*),(*,*))),((((*,*),*),((*,*),*)),((*,*),(*,*)))$$
$$((5,5),(6,4)) \Rightarrow (((3,2),(3,2)),((4,2),4)) \Rightarrow ((((2,1),2),((2,1),2)),((4,2),4))$$
$$\Rightarrow (((((*,*),*),(*,*)),(((*,*),*),(*,*))),((((*,*),(*,*)),(*,*)),((*,*),(*,*))))$$
$$((6,4),(6,4)) \Rightarrow (((3,3),4),((3,3),4)) \Rightarrow ((((2,1),(2,1)),4),(((2,1),(2,1)),4))$$
$$\Rightarrow (((((*,*),*),((*,*),*)),((*,*),(*,*))),((((*,*),*),((*,*),*)),((*,*),(*,*))))$$
$$((6,4),(6,4)) \Rightarrow (((3,3),4),((4,2),4)) \Rightarrow ((((2,1),(2,1)),4),((4,2),4))$$
$$\Rightarrow (((((*,*),*),((*,*),*)),((*,*),(*,*))),((((*,*),(*,*)),(*,*)),((*,*),(*,*))))$$
$$((6,4),(6,4)) \Rightarrow (((4,2),4),((4,2),4))$$
$$\Rightarrow (((((*,*),(*,*)),(*,*)),((*,*),(*,*))),((((*,*),(*,*)),(*,*)),((*,*),(*,*))))$$

3.2) Since $QB(12) = \{(6,6),(8,4)\}$ and 8 is a power of 2, the tree $(12,8)$ gives rise to the trees $((6,6),8)$ and $((8,4),8)$, and then, using $QB(6) = \{(3,3),(4,2)\}$ and $QB(3) = \{(2,1)\}$,

$$((6,6),8) \Rightarrow (((3,3),(3,3)),8) \Rightarrow ((((2,1),(2,1)),((2,1),(2,1))),8)$$
$$\Rightarrow (((((*,*),*),((*,*),*)),(((*,*),*),((*,*),*))),(((*,*),(*,*)),((*,*),(*,*))))$$
$$((6,6),8) \Rightarrow (((3,3),(4,2)),8) \Rightarrow ((((2,1),(2,1)),(4,2)),8)$$
$$\Rightarrow (((((*,*),*),((*,*),*)),(((*,*),(*,*)),(*,*))),(((*,*),(*,*)),((*,*),(*,*))))$$
$$((6,6),8) \Rightarrow (((4,2),(4,2)),8)$$
$$\Rightarrow (((((*,*),(*,*)),(*,*)),(((*,*),(*,*)),(*,*))),(((*,*),(*,*)),((*,*),(*,*))))$$
$$((8,4),8) \Rightarrow (((((*,*),(*,*)),((*,*),(*,*))),((*,*),(*,*))),(((*,*),(*,*)),((*,*),(*,*))))$$

So, there are 10 different trees in $\mathcal{T}_{20}$ with minimum Colles index.

We have implemented the Algorithm MinColless, with step (2.b.i) efficiently carried out by means of Proposition 12, in a Python script that generates, for every n, the Newick description of all bifurcating trees in $\mathcal{T}_n$ with minimum Colles index. It is available at the GitHub repository `https://github.com/biocom-uib/Colless`. As a proof of concept, we have computed for every n from 4 to $2^6 = 128$ all such bifurcating trees in $\mathcal{T}_n$. Figure 4 shows their number for every n.

5. Conclusions

The Colless index $C(T)$ of a bifurcating phylogenetic tree T measures the total amount of imbalance of T, and it is one of the oldest and most popular balance indices for rooted bifurcating phylogenetic trees. But, despite its popularity, neither its minimum value for any given number of leaves nor the trees where this minimum value is reached were known so far. This paper fills this gap in the literature, with two main contributions.

First, we have proved that, for every number n of leaves, the minimum Colless index of a bifurcating phylogenetic tree with n leaves is reached at the maximally balanced bifurcating trees and we have provided an explicit formula for its vale $C(n)$. Knowing this minimum value, as well as its maximum value, which is reached at the combs and it is $(n-1)(n-2)/2$, allows one to normalize the Colless index so that its range becomes the unit interval $[0,1]$, by means of the usual affine transformation

$$\widetilde{C}(T) = \frac{C(T) - C(n)}{\frac{(n-1)(n-2)}{2} - C(n)}.$$

This normalized index allows the comparison of the balance of trees with different numbers of leaves, which cannot be done directly with the unnormalized Colles index C because its value tends to grow with n.

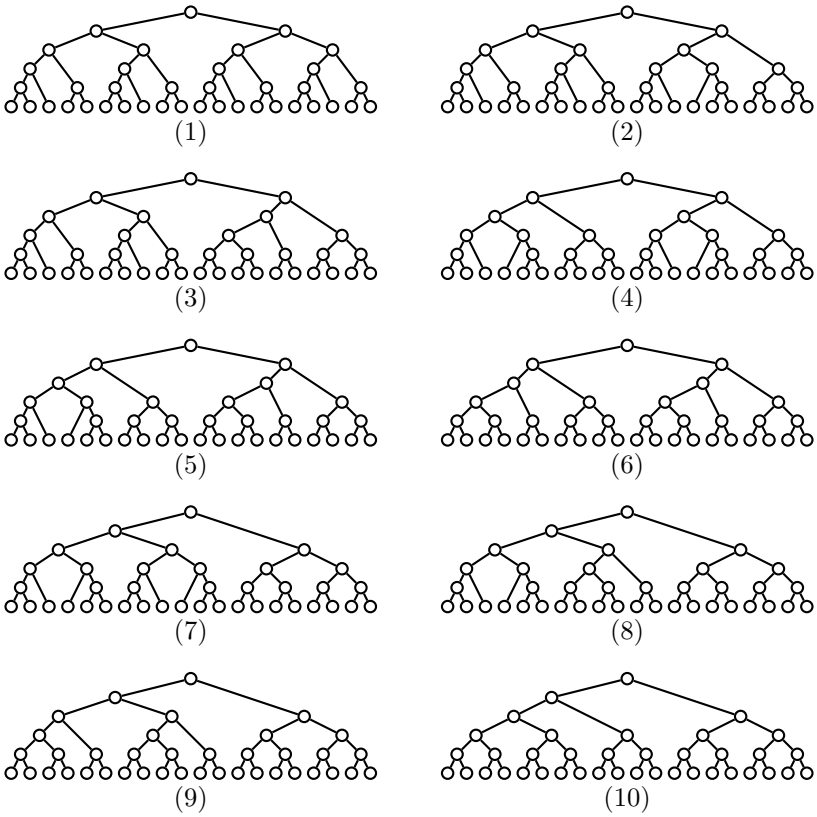

Figure 3: The 10 trees in $\mathcal{T}_{20}$ with minimum Colles index, 8. They are enumerated in the same order as they have been produced in Example 15.

The fact that the maximally balanced bifurcating trees have the minimum Colless index for their number of leaves is not surprising, because, in words of Shao and Sokal [18], they are considered the "most balanced" bifurcating trees. But it turns out that for almost all values of n there are bifurcating trees with n leaves that are not maximally balanced but whose Colless index is also minimal. So, our second main contribution is an alternative characterization of these trees and an efficient algorithm to produce all of them for any number n of leaves. Notice that, in spite of not being considered the "most balanced" ones because they have internal nodes whose imbalance is not minimal, the fact is that these trees also have the minimum total amount of imbalance. So, the total imbalance of a phylogenetic tree does not capture the local imbalance at each internal node.

The Colless index shares the drawback of classifying as "most balanced" many bifurcating trees that are not maximally balanced with the other most popular balance index, the Sackin index [7, 17, 18]. This problem can be avoided by also using other balance indices, like the total cophenetic index [12] or the rooted quartet index [5]. Let us recall in this connection that Shao and Sokal [18] already advised to use more than one balance index to quantify tree balance.

We find remarkable the regularities displayed by the sequence of numbers of trees in $\mathcal{T}_n$ with Colless index $C(n)$ hinted in Fig. 4 and that continue for larger values of n. We plan to study in a future paper the properties of this sequence.

Acknowledgements. This research was partially supported by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund through projects DPI2015-67082-P and PGC2018-096956-B-C43 (MINECO/FEDER).

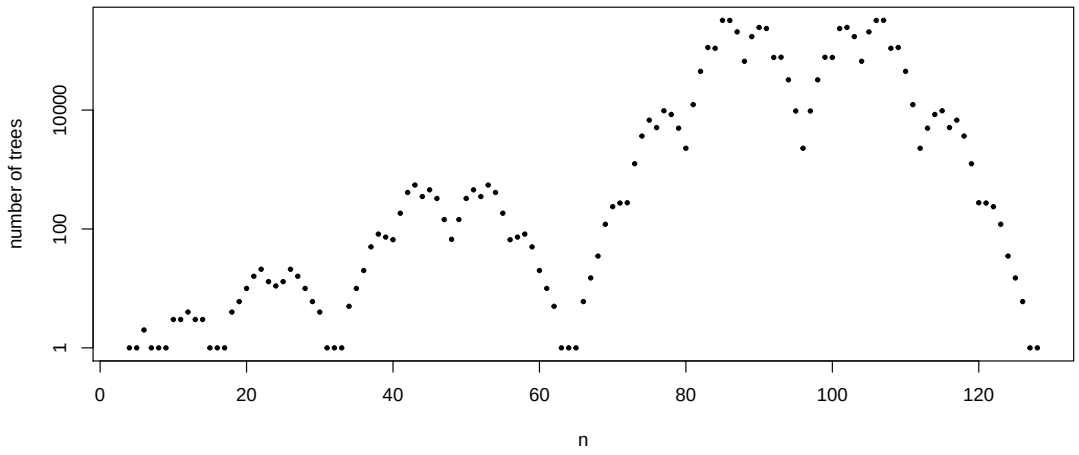

Figure 4: Scatterplot of of the number of trees in $\mathcal{T}_n$ with minimum Colless index, for $n = 4, \ldots, 128$.

References

[1] Aldous D (1996) Probability distributions on cladograms. In: Random discrete structures, The IMA Voumes in Mathematics and its Applications, vol 76, Springer, pp 1–18

[2] Avino M, Ng GT, He Y et al (2018) Tree shape-based approaches for the comparative study of cophylogeny. BioRxiv, 388116.

[3] Blum MGB, François O (2005) On statistical tests of phylogenetic tree imbalance: the Sackin and other indices revisited. Mathematical Biosciences 195:141–53

[4] Colless DH (1982) Review of "Phylogenetics: The theory and practice of phylogenetic systematics". Systematic Zoology 31:100–104

[5] Coronado TM, Mir A, Rosselló F, Valiente G. (2018) A balance index for phylogenetic trees based on rooted quartets. arXiv preprint arXiv:1803.01651, to appear in Journal of Mathematical Biology.

[6] Felsenstein J (2004) *Inferring Phylogenies*. Sinauer Associates Inc

[7] Fischer M (2018) Extremal values of the Sackin balance index for rooted binary trees. arXiv preprint arXiv:1801.10418.

[8] Ford D (2005) Probabilities on cladograms: Introduction to the alpha model. PhD Thesis (Stanford University). arXiv preprint arXiv:math/0511246 [math.PR].

[9] Goloboff PA, Arias JS, Szumik CA (2017) Comparing tree shapes: beyond symmetry. Zoologica Scripta 46: 637–648

[10] Heard SB (1992) Patterns in Tree Balance among Cladistic, Phenetic, and Randomly Generated Phylogenetic Trees. Evolution 46, 1818–1826

[11] Kirkpatrick M, Slatkin M (1993) Searching for evolutionary patterns in the shape of a phylogenetic tree. Evolution 47:1171–1181

[12] Mir A, Rosselló F, Rotger L (2013) A new balance index for phylogenetic trees. Mathematical Biosciences 241:125–136

[13] Mir A, Rotger L, Rosselló F (2018) Sound Colless-like balance indices for multifurcating trees. PLoS ONE 13(9):e0203401

[14] Mooers A, Heard SB (1997) Inferring evolutionary process from phylogenetic tree shape. The Quarterly Review of Biology 72:31–54

[15] Nelson MI, Holmes EC (2007) The evolution of epidemic influenza. Nature Revieq Genetics 8:196–205.

[16] Rogers JS (1993) Response of tree imbalance to number of terminal taxa. Sys. Biol. 42, 102–105.

[17] Sackin MJ (1972) Good and "bad" phenograms. Systematic Zoology 21:225–226

[18] Shao KT, Sokal R (1990) Tree balance. Systematic Zoology 39:226–276

[19] Sloane NJA (2010) *The On-Line Encyclopedia of Integer Sequences*. `http://oeis.org/`

[20] Stam, E (2002) Does imbalance in phylogenies reflect only bias? Evolution 56:1292–1295

[21] Stich M, Manrubia S (2009) Topological properties of phylogenetic trees in evolutionary models. European Physics Journal B 70:583–592

[22] Steel M, McKenzie A (2000) Distributions of cherries for two models of trees. Mathematical Bioscences 164:81–92

[23] Willis JC, Yule GU (1922) Some statistics of evolution and geographical distribution in plants and animals, and their significance. Nature 109:177–179