On the minimum value of the Colless index and the bifurcating trees that achieve it

Tomás M. Coronado[a], Mareike Fischer[b], Lina Herbst[b], Francesc Rosselló[a], Kristina Wicke[b]

[a]*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*
[b]*Institute of Mathematics and Computer Science, University of Greifswald, Greifswald, Germany*

Abstract

Measures of tree balance play an important role in the analysis of phylogenetic trees. One of the oldest and most popular indices in this regard is the Colless index for rooted bifurcating trees, introduced by Colless [8]. While many of its statistical properties under different probabilistic models for phylogenetic trees have already been established, little is known about its minimum value and the trees that achieve it. In this manuscript, we fill this gap in the literature. To begin with, we derive both recursive and closed expressions for the minimum Colless index of a tree with n leaves. Surprisingly, these expressions show a connection between the minimum Colless index and the so-called Blancmange curve, a fractal curve. We then fully characterize the trees that achieve this minimum value and we introduce both an algorithm to generate them and a recurrence to count them. After focusing on two extremal classes of trees with minimum Colless index (the maximally balanced trees and the greedy from the bottom trees), we conclude by showing that all trees with minimum Colless index also have minimum Sackin index, another popular balance index.

Keywords: Phylogenetic tree, Tree balance, Colless index, Sackin index, Blancmange curve, Takagi curve

1. Introduction

One of the main goals of evolutionary biology is to understand which factors influence evolutionary processes and their effect on them. Since phylogenetic trees are the standard representation of joint evolutionary histories of groups of species, it is natural to look for the imprint of these factors in the shapes of phylogenetic trees [23, 30]. 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 then used to test evolutionary models [4, 10, 18, 23, 29], to compare tree shapes [3, 15] or simply to describe phylogenies [7, 20], among other applications. Since the early observation by Willis and Yule [32] that taxonomic trees tend to be asymmetric, with many small clades and only 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. In this way, the *imbalance* of a phylogenetic tree reflects the propensity of evolutionary events to occur along specific lineages [24].

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, [8, 9, 13, 18, 19, 21, 22, 26, 27], and the section "Measures of overall asymmetry" in [11] (pp. 562–563). Among them, the *Colless index*, introduced by Colless [8], is one of the oldest and most popular. It is defined, on a rooted bifurcating tree T, as the sum, over all the internal nodes v of T, of the absolute value of the difference between the numbers of descendant leaves of

Email addresses: t.martinez@uib.eu (Tomás M. Coronado), email@mareikefischer.de (Mareike Fischer), lina.herbst@uni-greifswald.de (Lina Herbst), cesc.rossello@uib.eu (Francesc Rosselló), kristina.wicke@uni-greifswald.de (Kristina Wicke)

the pair of children of v. Its statistical properties under several probabilistic models for phylogenetic trees have been thoroughly studied: see, for instance, [5, 6, 14, 16].

In this manuscript we focus on the extremal properties of the Colless index. More specifically, we solve several open problems related to the minimum Colless index for rooted bifurcating trees with a given number of leaves. Let us mention here that, as far as the maximum Colless index for a given number of leaves n goes, it is folklore knowledge that it is reached at the *caterpillar tree*, or *comb*: the unique rooted bifurcating tree with n leaves where all internal nodes have different numbers of descendant leaves (cf. Figure 2.(a)). Caterpillars are considered since the early paper by Sackin [26] to be the most imbalanced type of phylogenetic trees, and the fact that they have the maximum Colless index for any number of leaves n was already hinted at by Colless [8], but he gave a wrong value for their Colless index, which was later corrected by Heard [16], giving the correct maximum value of $(n-1)(n-2)/2$. As a matter of fact, 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 [22].

In contrast, the analysis of the minimum value of the Colless index is much more involved. On the one hand, despite its popularity and wide use, the minimum Colless index of a bifurcating tree with n leaves is unknown beyond the often stated straightforward result that for numbers of leaves that are powers of 2 it is reached at the fully symmetric trees, which clearly have Colless index 0; see for instance [16, 18, 23]. To have a closed formula for this minimum value is essential 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, for instance, by Shao and Sokal [27] or Stam [29]. On the other hand, this minimum value may be achieved by several trees, which raises the questions of characterizing these "most balanced trees" according to the Colless index and counting them.

In this manuscript, we fill these gaps in the literature. To be precise, we first prove a recursive formula and two closed expressions for the minimum Colless index for a given number n of leaves. One of the closed expressions is related to a fractal curve, namely the so-called Blancmange, or Takagi, curve, thus showing the fractal structure and symmetry of the minimum Colless index. Next, we fully characterize all rooted bifurcating trees with n leaves that have minimum Colless index and we provide an efficient algorithm to generate them and a recursive formula to count them. We also focus on two particular classes of trees with minimum Colless index: the *maximally balanced trees* [21] and a class that we call *greedy from the bottom trees*. These two classes of trees turn out to be extremal in the following sense: for every m, the difference (in absolute value) between the numbers of descendant leaves of the pair of children of an internal node with m descendant leaves in a tree T with minimum Colless index achieves its minimum value when T is maximally balanced and its maximum value when T is greedy from the bottom. We conclude by showing that all trees with minimum Colless index also have minimum Sackin index, another popular index of phylogenetic tree balance introduced by Sackin [26].

2. Basic definitions and preliminary results

Before we can present our results, we need to introduce some definitions and notations. Throughout this manuscript, by a *tree* we mean a *rooted tree*: a tree $T = (V(T), E(T))$ with node set $V(T)$ and edge set $E(T)$ where one node is designated as the root (denoted henceforth by ρ). We shall always understand a rooted tree T as a directed graph, with its edges directed away from the root. We use $V_L(T) \subseteq V(T)$ to denote the leaf set of T (i.e. $V_L(T) = \{v \in V \mid \deg_{out}(v) = 0\}$) and by $\mathring{V}(T)$ we denote the set of internal nodes, i.e. $\mathring{V}(T) = V(T) \setminus V_L(T)$. If $|V(T)| = 1$, T consists of only one node, which is at the same time the root and the only leaf of the tree, and no edge. Whenever there is no ambiguity we simply denote $E(T)$, $V(T)$, $\mathring{V}(T)$, and $V_L(T)$ as E, V, $\mathring{V}$, and V_L, respectively. To simplify the language, we shall often say that two trees are *equal* when they are actually only isomorphic as rooted trees; we shall also use the expression *to have the same shape* as a synonym of being isomorphic.

Now, a *bifurcating tree* is a rooted tree where all internal nodes have out-degree 2. We denote by $\mathcal{T}_n$, for every $n \geqslant 1$, the set of (isomorphism classes of) bifurcating trees with n leaves. Note that, for $n = 1$, $\mathcal{T}_1$ consists only of the tree with one node and no edge.

Whenever there exists a path from u to v in a tree T, we say that u is an *ancestor* of v and that v is a *descendant* of u. In addition, whenever there exists an edge from u to v, we say that v is a *child* of u and that u is the *parent* of v. Note that in a bifurcating tree with $n \geqslant 2$ leaves, each internal node has exactly two children. Two leaves x and y are said to form a *cherry* when they have the same parent. Given a node v of T, we denote by T_v the subtree of T rooted at v and by $\kappa_T(v)$ the number of leaves of T_v, i.e. the number of descendant leaves of v.

The *depth* $\delta_T(v)$ of a node v is the number of edges on the path from ρ to v and the *height* $h(T)$ of a tree T is the maximum depth of any leaf in it.

A bifurcating tree T can be decomposed into its two *maximal pending subtrees* T_a and T_b rooted at the children a and b of ρ, and we shall denote this decomposition by $T = (T_a, T_b)$; cf. Figure 1. We shall usually denote by n_a and n_b the numbers of leaves of T_a and T_b, respectively, and without any loss of generality we shall always assume, usually without any further notice, that $n_a \geqslant n_b \geqslant 1$.

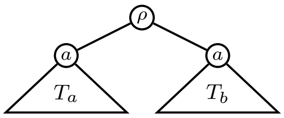

Figure 1: The decomposition $T = (T_a, T_b)$ of a bifurcating tree into its two maximal pending subtrees.

Given a bifurcating tree T and an internal node $v \in \mathring{V}$ with children v_1 and v_2, the *balance value* of v is defined as $bal_T(v) = |\kappa_T(v_1) - \kappa_T(v_2)|$. We call an internal node v *balanced* if $bal_T(v) \leqslant 1$, i.e. when its two children have $\lceil \kappa_T(v)/2 \rceil$ and $\lfloor \kappa_T(v)/2 \rfloor$ descendant leaves, respectively. Based on this we call a tree *maximally balanced* if all its internal nodes are balanced (cf. Figure 2.(b)). Recursively, a bifurcating tree is maximally balanced if its root is balanced and its two maximal pending subtrees are maximally balanced. This easily implies that any rooted subtree of a maximally balanced tree is again maximally balanced, by induction on the depth of the root of the subtree. It also implies that, for every $n \in \mathbb{N}$, there exists a unique maximally balanced tree with n leaves, which we denote by T_n^{mb}, and that, as we have just mentioned, $T_n^{mb} = (T_{\lceil n/2 \rceil}^{mb}, T_{\lfloor n/2 \rfloor}^{mb})$.

Two other particular trees appearing in this manuscript are the caterpillar trees and the fully symmetric trees (cf. Figure 2.(a) and (c)). The *caterpillar tree* with n leaves, T_n^{cat}, is the unique bifurcating tree with n leaves all whose internal nodes have different numbers of descendant leaves. As to the *fully symmetric tree of height* k, T_k^{fs}, it is the unique tree with $n = 2^k$ leaves in which all leaves have depth k. Note that $T_k^{fs} = (T_{k-1}^{fs}, T_{k-1}^{fs})$, i.e. the maximal pending subtrees of a fully symmetric tree of height k are fully symmetric trees of height $k-1$. Note also that $T_k^{fs} = T_{2^k}^{mb}$, because in the special case when $n = 2^k$, T_k^{fs} is the unique tree all whose internal nodes have balance value 0.

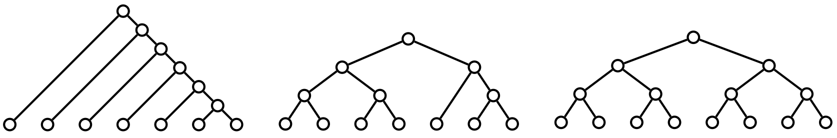
(a) T_7^{cat} (b) T_7^{mb} (c) $T_3^{fs} = T_8^{mb}$

Figure 2: The caterpillar tree T_7^{cat} with 7 leaves, the maximally balanced tree T_7^{mb} with 7 leaves, and the fully symmetric tree $T_3^{fs} = T_8^{mb}$ of depth 3, with $2^3 = 8$ leaves.

We are now in a position to define the focus of this manuscript:

Definition 1. The *Colless index* of a bifurcating tree T is the sum of the balance values of its internal nodes:

$$\mathcal{C}(T) = \sum_{v \in \overset{\circ}{V}(T)} bal_T(v) = \sum_{v \in \overset{\circ}{V}(T)} |\kappa_T(v_1) - \kappa_T(v_2)|,$$

where v_1 and v_2 denote the children of each $v \in \overset{\circ}{V}(T)$.

Note that $\mathcal{C}(T) \geqslant 0$, because it is defined as a sum of absolute values. For instance, consider the three trees depicted in Figure 2. Here, we have: $\mathcal{C}(T_7^{cat}) = 15$, $\mathcal{C}(T_7^{mb}) = 2$, and $\mathcal{C}(T_3^{fs}) = 0$.

Since the Colless index of a tree measures its global imbalance, the smaller the Colless index of a tree, the *more balanced* we consider it. In other words, for every pair of trees $T_1, T_2 \in \mathcal{T}_n$, if $\mathcal{C}(T_1) < \mathcal{C}(T_2)$, then T_1 is *more balanced* than T_2. For example, in Figure 2, T_7^{mb} is more balanced than T_7^{cat}.

It is easy to see that the Colless index satisfies the following recurrence [25].

Lemma 1. *If $T = (T_a, T_b)$ is a bifurcating tree with $T_a \in \mathcal{T}_{n_a}$ and $T_b \in \mathcal{T}_{n_b}$, where $n_a \geqslant n_b$, then*

$$\mathcal{C}(T) = \mathcal{C}(T_a) + \mathcal{C}(T_b) + n_a - n_b.$$

3. The minimum Colless index

We shall denote throughout this manuscript by c_n the *minimum Colless index of a bifurcating tree with n leaves*:

$$c_n = \min\big\{\mathcal{C}(T) \mid T \in \mathcal{T}_n\big\}.$$

The main aim of this section is to study the sequence c_n. We derive both a recurrence and two closed formulas for this sequence and we point out both its fractal structure and its symmetry. We start by showing that if a bifurcating tree $T = (T_a, T_b)$ has minimum Colless index, its two maximal pending subtrees also have minimum Colless index.

Lemma 2. *Let $T = (T_a, T_b)$ be a bifurcating tree with n leaves. If T has minimum Colless index on $\mathcal{T}_n$, then T_a and T_b have minimum Colless indices on $\mathcal{T}_{n_a}$ and $\mathcal{T}_{n_b}$, respectively.*

Proof. Assume that $\mathcal{C}(T_a)$ is not minimal; the case when $\mathcal{C}(T_b)$ is not minimal is symmetrical. Then, there exists $\widehat{T} \in \mathcal{T}_{n_a}$ such that $\mathcal{C}(\widehat{T}) < \mathcal{C}(T_a)$. Consider the tree $\widetilde{T} = (\widehat{T}, T_b) \in \mathcal{T}_n$ obtained by replacing in T the rooted subtree T_a by $\widehat{T}$. Then, by Lemma 1,

$$\mathcal{C}(\widetilde{T}) = \mathcal{C}(\widehat{T}) + \mathcal{C}(T_b) + n_a - n_b < \mathcal{C}(T_a) + \mathcal{C}(T_b) + n_a - n_b = \mathcal{C}(T),$$

which implies that $\mathcal{C}(T)$ is not minimal. Thus, if $\mathcal{C}(T)$ is minimal, $\mathcal{C}(T_a)$ must be minimal, too. □

Remark 1. Lemma 2 easily implies that every rooted subtree of a tree with minimum Colles index has also minimum Colless index, by induction on the depth of the root of the subtree.

Lemmas 1 and 2 directly imply that

$$c_n = \min\{c_{n_a} + c_{n_b} + n_a - n_b \mid n_a \geqslant n_b \geqslant 1,\ n_a + n_b = n\}. \tag{1}$$

In particular,

$$c_n \leqslant c_{n_a} + c_{n_b} + n_a - n_b \text{ for every } n_a \geqslant n_b \geqslant 1 \text{ with } n_a + n_b = n, \tag{2}$$

a fact that will be useful in subsequent proofs.

3.1. The maximally balanced trees have minimum Colless index

In this subsection we prove that the Colless index of a maximally balanced tree T_n^{mb} is c_n. The proof relies on the following lemma, which shows that the sequence $\mathcal{C}(T_n^{mb})$ also satisfies the Inequalities (2).

Lemma 3. *For every $n \in \mathbb{N}_{\geqslant 2}$ and for every $n_a \geqslant n_b \geqslant 1$ such that $n_a + n_b = n$,*

$$\mathcal{C}(T_n^{mb}) \leqslant \mathcal{C}(T_{n_a}^{mb}) + \mathcal{C}(T_{n_b}^{mb}) + n_a - n_b.$$

Proof. To simplify the notations, throughout this proof we shall denote $\mathcal{C}(T_n^{mb})$ by $C(n)$. By Lemma 1 and the equality $T_n^{mb} = (T_{\lceil n/2\rceil}^{mb}, T_{\lfloor n/2\rfloor}^{mb})$, we have that, for every $n \geqslant 2$,

$$C(n) = C(\lceil n/2\rceil) + C(\lfloor n/2\rfloor) + \lceil n/2\rceil - \lfloor n/2\rfloor,$$

or, equivalently, for every $n \geqslant 1$,

$$C(2n) = 2C(n) \quad \text{and} \quad C(2n+1) = C(n+1) + C(n) + 1. \tag{3}$$

We shall use this recurrence to prove by induction on m that, for every $m \geqslant 1$, the inequality

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

holds for every $s \in \mathbb{N}$. Taking $n_a = m+s$ and $n_b = m$, this clearly entails the statement.

Since $C(1) = 0$, the base case $m = 1$ 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 now consider 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 is even, say $s = 2s'$ with $s' \geqslant 1$, then, by Eqns. (3),

  $$C(1+s) + s = C(2s'+1) + 2s' = C(s'+1) + C(s') + 1 + 2s'$$

  and

  $$C(2+s) = C(2s'+2) = 2C(s'+1)$$

  and the desired Inequality (5) holds because, by the induction hypothesis,

  $$\begin{aligned} C(s') + 2s' + 1 &= C(1 + (s'-1)) + (s'-1) + s' + 2 \\ &\geqslant C(2 + (s'-1)) + s' + 2 > C(s'+1). \end{aligned}$$

- If s is odd, say $s = 2s'+1$ with $s' \geqslant 1$, then, by Eqns. (3),

  $$C(1+s) + s = C(2s'+2) + 2s' + 1 = 2C(s'+1) + 2s' + 1$$

  and

  $$C(2+s) = C(2s'+3) = C(s'+2) + C(s'+1) + 1$$

  and then (5) holds because, by the induction hypothesis,

  $$C(s'+1) + 2s' \geqslant C(s'+2) + s' > C(s'+2).$$

This completes the proof of the base case $m = 1$. Let us consider now the case $m \geqslant 2$ and let us assume that $C(m'+s) + C(m') + s \geqslant C(2m'+s)$ for every $1 \leqslant m' < m$ and $s \geqslant 0$. To prove that (4) is true for every $s \in \mathbb{N}$ we distinguish 4 cases:

- m and s even: say, $m = 2m'$ and $s = 2s'$. Then, by Eqns. (3),

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

  and the desired Inequality (4) is true because, by induction,

$$C(m'+s')+C(m')+s' \geqslant C(2m'+s').$$

- m even and s odd: say, $m = 2m'$ and $s = 2s'+1$. Then, by Eqns. (3),

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

  and (4) holds because, by induction,

$$C(m'+s'+1)+C(m')+s'+1 \geqslant C(2m'+s'+1)$$

  and

$$C(m'+s')+C(m')+s' \geqslant C(2m'+s').$$

- m odd and s even: say, $m = 2m'+1$ and $s = 2s'$. If $s' = 0$, the desired Inequality (4) amounts to $C(m)+C(m) \geqslant C(2m)$, which is true because it is actually an equality. So, assume that $s' \geqslant 1$. Then, by Eqns. (3),

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

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

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

- m and s odd: say, $m = 2m'+1$ and $s = 2s'+1$. Then, by Eqns. (3),

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

  and (4) is true because, by induction,

$$C(m'+s'+1)+C(m'+1)+s' \geqslant C(2m'+s'+2)$$

  and

$$C(m'+s'+1)+C(m')+s'+1 \geqslant C(2m'+s'+1).$$

This completes the proof of the inductive step. □

We are now in a position to establish our first main result.

Theorem 1. *For every $n \geqslant 1$, $\mathcal{C}(T_n^{mb}) = c_n$.*

Proof. We shall prove by induction on n that $\mathcal{C}(T) \geqslant \mathcal{C}(T_n^{mb})$ for every $T \in \mathcal{T}_n$. The case when $n = 1$ is obvious, because $\mathcal{T}_1 = \{T_1^{mb}\}$. Assume now that the assertion is true for every number of leaves smaller than n and let $T = (T_a, T_b) \in \mathcal{T}_n$, with $T_a \in \mathcal{T}_{n_a}$ and $T_b \in \mathcal{T}_{n_b}$. Then, by Lemma 1,

$$\mathcal{C}(T) = \mathcal{C}(T_a) + \mathcal{C}(T_b) + n_a - n_b \geqslant \mathcal{C}(T_a^{mb}) + \mathcal{C}(T_b^{mb}) + n_a - n_b \geqslant \mathcal{C}(T_n^{mb}),$$

where the first inequality holds by the induction hypothesis and the second inequality by the previous lemma. □

Next corollary says that the sequence c_n is the sequence A296062 in the *On-Line Encyclopedia of Integer Sequences* [28].

Corollary 1. *Let $A(T_n^{mb})$ be the number of automorphisms of T_n^{mb}. Then, $c_n = n - 1 - \log_2(A(T_n^{mb}))$.*

Proof. Since, by definition, the balance value of every internal node in T_n^{mb} is 0 or 1, $c_n = \mathcal{C}(T_n^{mb})$ is equal to the number of internal nodes of T_n^{mb} with non zero balance value. Now, for every internal node u of T_n^{mb}, its balance value is 0 if, and only if, the subtrees of T_n^{mb} rooted at its children are isomorphic, that is, with the notations of [14], if, and only if, u is a symmetric branch point. Indeed, as we mentioned in Section 2, the subtrees rooted at the children of u are again maximally balanced, and therefore they have the same numbers of leaves if, and only if, they are isomorphic.

So, the number of symmetric branch points in T_n^{mb} is $n - 1 - c_n$, which implies, by Lemma 31 in [14], that $A(T_n^{mb}) = 2^{n-1-c_n}$, as stated. □

Theorem 1, together with Lemma 1, directly imply the following recurrence for c_n, which was already used, for $\mathcal{C}(T_n^{mb})$, in the proof of Lemma 3: cf. Eqns. (3).

Corollary 2. *The sequence c_n satisfies that $c_1 = 0$ and, for every $n \geqslant 2$,*

$$c_n = c_{\lceil n/2 \rceil} + c_{\lfloor n/2 \rfloor} + \lceil n/2 \rceil - \lfloor n/2 \rfloor$$

or, equivalently, $c_{2n} = 2c_n$ and $c_{2n+1} = c_{n+1} + c_n + 1$ for every $n \geqslant 1$.

3.2. Two closed formulas for the minimum Colless index

Corollary 2 implies that we can recurrently compute c_n for any desired n. In this subsection, however, we derive from that recurrence two different closed expressions for c_n and we prove some properties of this sequence. Our first closed formula for c_n is given in terms of the binary expansion of n.

Theorem 2. *If $n = \sum_{j=1}^{\ell} 2^{m_j}$ with $m_1, \ldots, m_\ell \in \mathbb{N}$ such that $m_1 > \cdots > m_\ell$, then*

$$c_n = \sum_{j=2}^{\ell} 2^{m_j}(m_1 - m_j - 2(j-2)).$$

Proof. For every $n \geqslant 1$, let $\bar{c}_n = \sum_{j=2}^{\ell} 2^{m_j}(m_1 - m_j - 2(j-2))$, where $n = \sum_{j=1}^{\ell} 2^{m_j}$ with $m_1 > \cdots > m_\ell$. We shall prove that $c_n = \bar{c}_n$ by induction on n.

If $n = 1$, $\bar{c}_1 = \bar{c}_{2^0} = 0 = c_1$, which proves the base case of the induction. Now, we assume that the claim holds for every $n' \leqslant n - 1$ and we prove it for n by distinguishing two cases: n even and n odd.

If n is even, i.e. if $m_\ell > 0$, we have $\lfloor n/2 \rfloor = \lceil n/2 \rceil = n/2 = \sum_{j=1}^{\ell} 2^{m_j - 1}$ with $m_1 - 1 > \cdots > m_\ell - 1 \geqslant 0$ and thus

$$\begin{aligned}
c_n &= 2 \cdot c_{n/2} \quad \text{(by Corollary 2)} \\
&= 2 \cdot \overline{c}_{n/2} \quad \text{(by the induction hypothesis)} \\
&= 2 \cdot \sum_{j=2}^{\ell} 2^{m_j - 1}\big(m_1 - 1 - (m_j - 1) - 2(j-2)\big) \\
&= \sum_{j=2}^{\ell} 2^{m_j}(m_1 - m_j - 2(j-2)) = \overline{c}_n.
\end{aligned}$$

Assume now that n is odd, i.e. that $m_\ell = 0$. 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 $m_1 - 1 > \cdots > m_{k-1} - 1 > \ell - k \geqslant 0$. In this case,

$$\begin{aligned}
c_n &= c_{\lceil n/2 \rceil} + c_{\lfloor n/2 \rfloor} + \lceil n/2 \rceil - \lfloor n/2 \rfloor \quad \text{(by Corollary 2)} \\
&= \overline{c}_{\lceil n/2 \rceil} + \overline{c}_{\lfloor n/2 \rfloor} + \lceil n/2 \rceil - \lfloor n/2 \rfloor \quad \text{(by the induction hypothesis)} \\
&= \sum_{j=2}^{k-1} 2^{m_j - 1}\big((m_1 - 1) - (m_j - 1) - 2(j-2)) \\
&\quad + 2^{\ell - k}(m_1 - 1 - (\ell - k) - 2(k-2)\big) \\
&\quad + \sum_{j=2}^{\ell-1} 2^{m_j - 1}\big((m_1 - 1) - (m_j - 1) - 2(j-2)\big) + 1 \\
&= \sum_{j=2}^{k-1} 2^{m_j - 1}(m_1 - m_j - 2(k-2)) + 2^{m_k}(m_1 - m_k - 2(k-2)) - 2^{\ell - k} \\
&\quad + \sum_{j=2}^{k-1} 2^{m_j - 1}(m_1 - m_j - 2(j-2)) \\
&\quad + \sum_{j=k}^{\ell-1} 2^{\ell - j - 1}(m_1 - (\ell - j) - 2(j-2)) + 1 \\
&\text{(because } m_j = \ell - j \text{ for every } j \geqslant k) \\
&= \sum_{j=2}^{k} 2^{m_j}(m_1 - m_j - 2(j-2)) \\
&\quad + \sum_{j=k}^{\ell-1} 2^{\ell - j - 1}(m_1 - (\ell - j) - 2(j-2)) + 1 - 2^{\ell - k}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=2}^{k} 2^{m_j}(m_1 - m_j - 2(j-2)) \\
&\qquad + \sum_{i=k+1}^{\ell} 2^{\ell-i}(m_1 - (\ell - i) - 2(i-2) + 1) + 1 - 2^{\ell-k} \\
&= \sum_{j=2}^{k} 2^{m_j}(m_1 - m_j - 2(j-2)) \\
&\qquad + \sum_{i=k+1}^{\ell} 2^{m_i}(m_1 - m_i - 2(i-2)) + \sum_{i=k+1}^{\ell} 2^{\ell-i} + 1 - 2^{\ell-k} \\
&= \sum_{j=2}^{\ell} 2^{m_j}(m_1 - m_j - 2(j-2)) = \overline{c}_n.
\end{aligned}
$$

This completes the proof of the inductive step. □

Corollary 3. *For every $n \geqslant 1$, $c_n = 0$ if, and only if, n is a power of 2. Moreover, for every $n \geqslant 1$ and $T \in \mathcal{T}_n$, $\mathcal{C}(T) = 0$ if, and only if, T is fully symmetric.*

Proof. Let $n = \sum\limits_{j=1}^{\ell} 2^{m_j} \geqslant 1$ with $m_1, \dots, m_\ell \in \mathbb{N}$ such that $m_1 > \cdots > m_\ell$. Since $2^{m_j}(m_1 - m_j - 2(j-2)) > 0$ if $j > 1$, we have by Theorem 2 that $c_n = 0$ if, and only if, $\sum\limits_{j=2}^{\ell} 2^{m_j}(m_1 - m_j - 2(j-2))$ is an empty sum, which is equivalent to $\ell = 1$, i.e. to $n = 2^{m_1}$. This completes the proof of the first part of the statement. The second part now follows by an easy argument by induction using Lemma 1. □

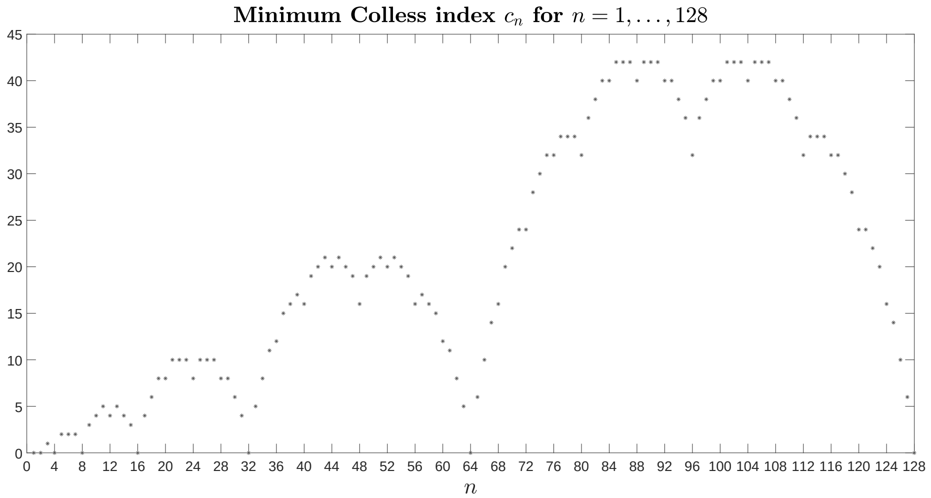

Figure 3: Plot of c_n for $n = 1, \dots, 128$.

Figure 3 depicts the value of c_n for $n = 1, \dots, 128$. Surprisingly, the minimum Colless index exhibits a fractal structure. In the next theorem we provide a second closed formula for c_n that explains this fractal structure by showing a connection between the sequence c_n and the so-called *Blancmange curve*, a fractal curve also known as the *Takagi curve* (cf. [31]). This curve plays an important role in different areas such as combinatorics, number theory and analysis [2] and it is defined as the graph of the function $T : [0, 1] \to \mathbb{R}$

with

$$T(x)=\sum_{i=0}^{\infty} 2^{-i}\cdot s(2^{i}\cdot x),$$

where $s(x)=\min_{z\in\mathbb{Z}}|x-z|$ is the distance from x to its nearest integer. Recall that this function s satisfies the following straightforward properties: $s(n)=0$ for every $n\in\mathbb{Z}$; $s(n+x)=s(x)$ for every $n\in\mathbb{Z}$ and $x\in\mathbb{R}$; $s(x)=s(-x)$ for every $x\in\mathbb{R}$; if $0\leqslant x\leqslant 1/2$, then $s(x)=x$; and if $1/2\leqslant x\leqslant 1$, then $s(x)=1-x$.

Theorem 3. *For every $n\geqslant 1$, let $k_n:=\lceil\log_2(n)\rceil$. Then,*

$$c_n=\sum_{j=1}^{k_n-1} 2^{j}\cdot s(2^{-j}\cdot n),$$

where $s(x)$ is the distance from $x\in\mathbb{R}$ to its nearest integer.

Proof. We shall prove that the expression for c_n given in the statement is equal to the expression provided in Theorem 2. In this proof, it is convenient to write the binary expansion of n as $n=\sum_{i=1}^{\ell}2^{n_i}$ with $n_1<\cdots<n_\ell$. With these notations, the formula given in Theorem 2 becomes

$$c_n=\sum_{i=1}^{\ell-1}2^{n_i}(n_\ell-n_i-2(\ell-i-1)).$$

With these notations, for every $j\in\mathbb{N}$, if $j\leqslant n_1$, then $2^{-j}\cdot n\in\mathbb{N}$ and thus $s(2^{-j}\cdot n)=0$, while if $n_t<j\leqslant n_{t+1}$ for some $t=1,\dots,\ell-1$, then

$$2^{-j}\cdot n=\sum_{i=1}^{t}2^{n_i-j}+\sum_{i=t+1}^{\ell}2^{n_i-j},$$

where $\sum_{i=t+1}^{\ell}2^{n_i-j}\in\mathbb{N}$ and, as far as $\sum_{i=1}^{t}2^{n_i-j}$ goes:

- If $j>n_t+1$
$$\sum_{i=1}^{t}2^{n_i-j}=\frac{\sum_{i=1}^{t}2^{n_i-n_1}}{2^{j-n_1}}\leqslant\frac{\sum_{s=0}^{n_t-n_1}2^{s}}{2^{n_t+2-n_1}}=\frac{2^{n_t-n_1+1}-1}{2^{n_t-n_1+2}}<\frac{1}{2}$$
- If $j=n_t+1$
$$\sum_{i=1}^{t}2^{n_i-j}=\sum_{i=1}^{t}2^{n_i-n_t-1}=\frac{1}{2}+\sum_{i=1}^{t-1}2^{n_i-n_t-1}$$
where
$$0\leqslant\sum_{i=1}^{t-1}2^{n_i-n_t-1}\leqslant\frac{\sum_{s=0}^{n_{t-1}}2^{s}}{2^{n_t+1}}=\frac{2^{n_{t-1}+1}-1}{2^{n_t+1}}<\frac{1}{2}$$
and therefore in this case $1/2\leqslant\sum_{i=1}^{t}2^{n_i-j}<1$.

This implies that, if $n_t+1<j\leqslant n_{t+1}$,

$$2^{j}\cdot s(2^{-j}\cdot n)=2^{j}\sum_{i=1}^{t}2^{n_i-j}=\sum_{i=1}^{t}2^{n_i}\qquad(6)$$

and if $j = n_t + 1$,

$$2^{n_t+1} \cdot s(2^{-n_t-1} \cdot n) = 2^{n_t+1}\Big(\frac{1}{2} - \sum_{i=1}^{t-1} 2^{n_i-n_t-1}\Big) = 2^{n_t} - \sum_{i=1}^{t-1} 2^{n_i}. \tag{7}$$

Now, on the one hand, if n is a power of 2, i.e. if $n = 2^{n_1}$, then $k_n = n_1$ and the previous discussion shows that $s(2^{-j} \cdot n) = 0$ for every $j \leqslant n_1 - 1$, which implies that

$$\sum_{j=1}^{k_n-1} 2^j \cdot s(2^{-j} \cdot n) = 0 = c_n.$$

On the other hand, if n is not a power of 2, i.e. if $\ell > 1$, then $k_n = n_\ell + 1$ and, by the previous discussion,

$$\sum_{j=1}^{k_n-1} 2^j \cdot s(2^{-j} \cdot n) = \sum_{j=n_1+1}^{n_\ell} 2^j \cdot s(2^{-j} \cdot n) = \sum_{t=1}^{\ell-1} \sum_{j=n_t+1}^{n_{t+1}} 2^j \cdot s(2^{-j} \cdot n)$$

where, for each $t = 1, \ldots, \ell - 1$,

$$\begin{aligned}\sum_{j=n_t+1}^{n_{t+1}} 2^j \cdot s(2^{-j} \cdot n) &= 2^{n_t+1} \cdot s(2^{-n_t-1} \cdot n) + \sum_{j=n_t+2}^{n_{t+1}} 2^j \cdot s(2^{-j} \cdot n) \\ &= 2^{n_t} - \sum_{i=1}^{t-1} 2^{n_i} + (n_{t+1} - n_t - 1)\sum_{i=1}^{t} 2^{n_i} \quad \text{(by Eqns. (6) and (7))} \\ &= (n_{t+1} - n_t)2^{n_t} + (n_{t+1} - n_t - 2)\sum_{i=1}^{t-1} 2^{n_i}.\end{aligned}$$

Therefore

$$\sum_{j=1}^{k_n-1} 2^j \cdot s(2^{-j} \cdot n) = \sum_{t=1}^{\ell-1} \Big((n_{t+1} - n_t)2^{n_t} + (n_{t+1} - n_t - 2)\sum_{i=1}^{t-1} 2^{n_i}\Big)$$

and the coefficient of each 2^{n_i}, for $i = 1, \ldots, \ell - 1$, in this expression is

$$n_{i+1} - n_i + \sum_{j=i+1}^{\ell-1} (n_{j+1} - n_j - 2) = n_\ell - n_i - 2(\ell - i - 1)$$

which proves that

$$\sum_{j=1}^{k_n-1} 2^j \cdot s(2^{-j} \cdot n) = \sum_{i=1}^{\ell-1} 2^{n_i}(n_\ell - n_i - 2(\ell - i - 1)) = c_n$$

as we claimed. □

We close this section with the following result, which establishes some properties of the minimum Colless index c_n that are reflected in Figure 3, in particular its symmetry.

Corollary 4. *The sequence c_n satisfies the following properties:*

(a) For every $m \geqslant 0$, $c_{2^m+1} = m$.

(b) For every $m \geqslant 0$ and for every $p = 2, \ldots, 2^m - 1$, $c_{2^m+p} < 2^m$.

(c) For every $m \geqslant 1$ and for every $p = 1, \ldots, 2^m - 1$, $c_{2^m+p} = c_{2^{m+1}-p}$.

Proof. Assertion (a) is a direct consequence of Theorem 2. Indeed, if $n = 2^m + 1$ then, with the notations of that theorem, $\ell = 2$, $m_1 = m$ and $m_2 = 0$, and therefore $c_{2^m+1} = 2^0(m - 0 - 2(2-2)) = m$.

As to (b), if $n = 2^m + p$ with $2 \leqslant p \leqslant 2^m - 1$, by Theorem 3, and recalling that, in this case, $\lceil \log_2(n) \rceil = m + 1$, and that $s(x) \leqslant 1/2$ for every $x \in \mathbb{R}$,

$$c_n = \sum_{j=1}^{m} 2^j \cdot s(2^{-j} \cdot n) \leqslant \sum_{j=1}^{m} 2^j \cdot \frac{1}{2} = \sum_{j=0}^{m-1} 2^j = 2^m - 1.$$

Finally, as far as (c) goes, let $n = 2^m + p$ for some $p = 1, \ldots, 2^m - 1$. Then:

$$\begin{aligned}
c_{2^m+p} &= \sum_{j=1}^{m} 2^j \cdot s(2^{-j}(2^m + p)) \quad \text{(by Theorem 3)}\\
&= \sum_{j=1}^{m} 2^j \cdot s(2^{m-j} + 2^{-j} \cdot p)\\
&= \sum_{j=1}^{m} 2^j \cdot s(2^{-j} \cdot p) \quad \text{(because each } 2^{m-j} \in \mathbb{N})\\
&= \sum_{j=1}^{m} 2^j \cdot s(-2^{-j} \cdot p) \quad \text{(because } s(x) = s(-x))\\
&= \sum_{j=1}^{m} 2^j \cdot s(2^{m+1-j} - 2^{-j} \cdot p) \quad \text{(because each } 2^{m+1-j} \in \mathbb{N})\\
&= \sum_{j=1}^{m} 2^j \cdot s(2^{-j}(2^{m+1} - p)) = c_{2^{m+1}-p} \quad \text{(again by Theorem 3).}
\end{aligned}$$

□

Notice that the bound given in point (b) in this corollary is sharper than the upper bound $c_n \leqslant n - 1$ that stems from Corollary 1.

4. Minimal Colless trees

We now turn our attention to the trees that achieve the minimum Colless index for their number of leaves, which we shall call henceforth *minimal Colless* trees. While we have already seen in Theorem 1 that, for every n, the maximally balanced tree T_n^{mb} has minimum Colless index and in Corollary 3 that when n is a power of 2 this is the only minimal Colless tree, for numbers n of leaves that are not powers of 2 there may exist other minimal Colless trees in $\mathcal{T}_n$. For instance, $c_6 = 2$ is reached at both trees depicted in Figure 4. Actually, for numbers of leaves n that differ more than 1 from a power of 2 there *always* exist at least two minimal Colless trees (see Corollary 6 below). So, the main goal of this section is to characterize all minimal Colless trees and to provide an efficient way of generating them for any given number n of leaves as well as a recurrence to count them.

4.1. Characterizing and generating minimal Colless trees

Recall from Eqn. (1) that

$$c_n = \min\{c_{n_a} + c_{n_b} + n_a - n_b \mid n_a \geqslant n_b \geqslant 1, n_a + n_b = n\}.$$

To simplify the language, for every $n \geqslant 2$, let

$$QB(n) := \big\{(n_a, n_b) \in \mathbb{N}^2 \mid \quad \begin{aligned} &n_a \geqslant n_b \geqslant 1,\ n_a + n_b = n,\\ &c_{n_a} + c_{n_b} + n_a - n_b = c_n\big\}. \end{aligned}$$

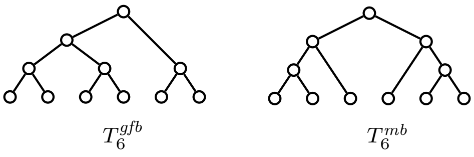

Figure 4: The GFB tree T_6^{gfb} (cf. Subsection 4.3) and the maximally balanced tree T_6^{mb} with 6 leaves. Both trees have minimum Colless index in $\mathcal{T}_6$, namely $c_6 = \mathcal{C}(T_6^{gfb}) = \mathcal{C}(T_6^{mb}) = 2$, and they are the only trees in $\mathcal{T}_6$ with Colless index 2.

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

The next proposition gives a characterization of the minimal Colless trees in terms of the sets QB that will allow us to efficiently generate them.

Proposition 1. *Let $T \in \mathcal{T}_n$. Then, $\mathcal{C}(T) = c_n$ if, and only if, $(\kappa_T(v_1), \kappa_T(v_2)) \in QB(\kappa_T(v))$ for every $v \in \mathring{V}(T)$ with children v_1, v_2 so that $\kappa_T(v_1) \geqslant \kappa_T(v_2)$.*

Proof. $\Longrightarrow$) Assume that there exists some $v \in \mathring{V}(T)$ with children v_1, v_2 so that $\kappa_T(v_2) \leqslant \kappa_T(v_1)$ and

$$c_{\kappa_T(v_1)} + c_{\kappa_T(v_2)} + \kappa_T(v_1) - \kappa_T(v_2) \neq c_{\kappa_T(v)}.$$

We shall prove that $\mathcal{C}(T) > c_n$. Indeed, by Eqn. (2), this inequality 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 the maximally balanced tree $T^{mb}_{\kappa_T(v)}$ and leaving the rest of T untouched. In this way, $T'_v = T^{mb}_{\kappa_T(v)}$ and $bal_T(x) = bal_{T'}(x)$ for every $x \in \mathring{V}(T) \setminus \mathring{V}(T_v) = \mathring{V}(T') \setminus \mathring{V}(T'_v)$; let us denote by W this last set of nodes. Then

$$\begin{aligned}
\mathcal{C}(T) &= \sum_{x \in \mathring{V}(T)} bal_T(x) = \sum_{x \in W} bal_T(x) + \sum_{x \in \mathring{V}(T_v)} bal_T(x) \\
&= \sum_{x \in W} bal_T(x) + \mathcal{C}(T_v) \\
&= \sum_{x \in W} bal_T(x) + \mathcal{C}(T_{v_1}) + \mathcal{C}(T_{v_2}) + \kappa_T(v_1) - \kappa_T(v_2) \\
&\geqslant \sum_{x \in W} bal_T(x) + c_{\kappa_T(v_1)} + c_{\kappa_T(v_2)} + \kappa_T(v_1) - \kappa_T(v_2) \\
&> \sum_{x \in W} bal_T(x) + c_{\kappa_T(v)} = \sum_{x \in W} bal_{T'}(x) + \mathcal{C}(T'_v) = \mathcal{C}(T') \geqslant c_n.
\end{aligned}$$

This proves the "only if" implication.

$\Longleftarrow$) We prove the "if" implication, i.e. that if $c_{\kappa_T(v_1)} + c_{\kappa_T(v_2)} + \kappa_T(v_1) - \kappa_T(v_2) = c_{\kappa_T(v)}$ for every $v \in \mathring{V}(T)$ with children v_1, v_2 so that $\kappa_T(v_1) \geqslant \kappa_T(v_2)$, then $\mathcal{C}(T) = c_n$, by induction on n. The case when $n = 1$ is obvious, because $\mathcal{T}_1 = \{T_1^{mb}\}$. 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 $v \in \mathring{V}(T)$,

$$c_{\kappa_T(v_1)} + c_{\kappa_T(v_2)} + \kappa_T(v_1) - \kappa_T(v_2) = c_{\kappa_T(v)},$$

where v_1, v_2 stand for the children of v so that $\kappa_T(v_1) \geqslant \kappa_T(v_2)$. Let $T = (T_a, T_b)$ be the decomposition of T into its maximal pending subtrees, with $T_a \in \mathcal{T}_{n_a}$ and $T_b \in \mathcal{T}_{n_b}$ so that $n_a \geqslant n_b$. Then, for every $v \in \mathring{V}(T_a)$, with children v_1, v_2 so that $\kappa_T(v_1) \geqslant \kappa_T(v_2)$,

$$\begin{aligned}
&c_{\kappa_{T_a}(v_1)} + c_{\kappa_{T_a}(v_2)} + \kappa_{T_a}(v_1) - \kappa_{T_a}(v_2) \\
&\quad = c_{\kappa_T(v_1)} + c_{\kappa_T(v_2)} + \kappa_T(v_1) - \kappa_T(v_2) = c_{\kappa_T(v)} = c_{\kappa_{T_a}(v)}.
\end{aligned}$$

This implies, by the induction hypothesis, that $\mathcal{C}(T_a) = c_{\kappa_T(a)}$. By symmetry, we also have that $\mathcal{C}(T_b) = c_{\kappa_T(b)}$. Finally,

$$\begin{aligned}\mathcal{C}(T) &= \mathcal{C}(T_a) + \mathcal{C}(T_b) + n_a - n_b\\ &= c_{\kappa_T(a)} + c_{\kappa_T(b)} + \kappa_T(a) - \kappa_T(b) = c_{\kappa_T(\rho)} = c_n\end{aligned}$$

as we wanted to prove. □

Next result provides a characterization of the pairs $(n_a, n_b) \in QB(n)$, for every $n \geqslant 2$. Since its proof is long and it relies on several lemmas, in order not to lose the thread of the manuscript we postpone it until Appendix A.1.

Proposition 2. *For every $n \geqslant 2$ and for every $n_a, n_b \in \mathbb{N}_{\geqslant 1}$ such that $n_a \geqslant n_b$ and $n_a + n_b = n$:*

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

(2) If $n_a > n_b$, then $(n_a, n_b) \in QB(n)$ if, and only if, one of the following three conditions is satisfied:

- *There exist $k \in \mathbb{N}$ and $p \in \mathbb{N}_{\geqslant 1}$ such that $n = 2^k(2p+1)$, $n_a = 2^k(p+1)$ and $n_b = 2^k p$.*
- *There exist $k \in \mathbb{N}$, $l \in \mathbb{N}_{\geqslant 2}$, $p \in \mathbb{N}_{\geqslant 1}$, and $t \in \mathbb{N}$, $0 \leqslant t < 2^{l-2}$, such that $n = 2^k(2^l(2p+1)+2t+1)$, $n_a = 2^{k+l}(p+1)$, and $n_b = 2^k(2^l p + 2t + 1)$.*
- *There exist $k \in \mathbb{N}$, $l \in \mathbb{N}_{\geqslant 2}$, $p \in \mathbb{N}_{\geqslant 1}$, and $t \in \mathbb{N}$, $0 \leqslant t < 2^{l-2}$, such that $n = 2^k(2^l(2p+1)-(2t+1))$, $n_a = 2^k(2^l(p+1) - (2t+1))$, and $n_b = 2^{k+l}p$.*

We now translate this proposition into an explicit and non-redundant description of $QB(n)$ from the binary expansion of n.

Proposition 3. *For every $n \geqslant 2$, let $k \in \mathbb{N}$ 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 $(n/2, n/2)$.

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

Proof. Assertion (a) is a consequence of the fact that if $QB(n)$ contains some (n_a, n_b) with $n_a > n_b$, then by (2) in the last proposition n cannot be a power of 2. So, assume henceforth that $\ell > 1$. Let now $(n_a, n_b) \in \mathbb{N}^2$ be such that $n = n_a + n_b$ and $1 \leqslant n_b < n_a$. Then, by Proposition 2, $(n_a, n_b) \in QB(n)$ if, and only if, one of the following three conditions is satisfied:

(b.1) There exist $k \in \mathbb{N}$ and $p \in \mathbb{N}_{\geqslant 1}$ such that $n_0 = 2p+1$, $n_a = 2^k(p+1)$, and $n_b = 2^k p$. In this case

$$p = \frac{n_0 - 1}{2} = \sum_{i=1}^{\ell-1} 2^{m_i - 1}$$

and this contributes to $QB(n)$ the pair (n_a, n_b) with

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

(b.2) There exist $k \in \mathbb{N}$, $l \in \mathbb{N}_{\geqslant 2}$, $p \in \mathbb{N}_{\geqslant 1}$, and $t \in \mathbb{N}$, $0 \leqslant t < 2^{l-2}$, such that $n_0 = 2^{l+1}p + 2^l + 2t + 1$ and $n_a = 2^{k+l}(p+1)$. 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, $m_j = l \geqslant 2$ and $m_{j+1} < l-1$ for some $j = 2, \ldots, \ell-1$, in which case

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

This contributes to $QB(n)$ the pairs (n_a, n_b) of the form

$$\begin{aligned} n_a &= 2^{k+m_j} \Big(\frac{\sum_{i=1}^{j-1} 2^{m_i}}{2^{m_j+1}} + 1 \Big) = 2^k \Big(\sum_{i=1}^{j-1} 2^{m_i-1} + 2^{m_j} \Big), \\ n_b &= n - 2^k \Big(\sum_{i=1}^{j-1} 2^{m_i-1} + 2^{m_j} \Big), \end{aligned} \tag{8}$$

with $j = 2, \ldots, \ell-1$ and $m_j \geqslant 2$ such that $m_{j+1} < m_j - 1$. All these pairs are different, because $\sum_{i=1}^{h-1} 2^{m_i-1} + 2^{m_h}$ is monotonously non-increasing on h (because $m_{h+1} < m_h$) and

$$\begin{aligned} \sum_{i=1}^{h-1} 2^{m_i-1} + 2^{m_h} = \sum_{i=1}^{h} 2^{m_i-1} + 2^{m_{h+1}} &\iff 2^{m_h} = 2^{m_h-1} + 2^{m_{h+1}} \\ &\iff 2^{m_h-1} = 2^{m_{h+1}} \iff m_h = m_{h+1} + 1. \end{aligned}$$

(b.3) There exist $k \in \mathbb{N}$, $l \in \mathbb{N}_{\geqslant 2}$, $p \in \mathbb{N}_{\geqslant 1}$, and $t \in \mathbb{N}$, $0 \leqslant t < 2^{l-2}$ such that $n_0 = 2^{l+1}p + 2^l - (2t+1)$ and $n_b = 2^{k+l}p$. 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 $j = 2, \ldots, \ell-1$ such that $m_{j-1} \geqslant l+1 = m_j + 2$, and then

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

This contributes to $QB(n)$ all pairs (n_a, n_b) of the form

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

with $j = 2, \ldots, \ell-1$ such that $m_j < m_{j-1} - 1$, belong to $QB(n)$, and they are pairwise different because n_b is strictly increasing on j.

This gives all pairs (n_a, n_b) in $QB(n)$ with $n_a > n_b$. 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. The pairs of the form (b.2) and different from the pair (b.1) because, since $j \leqslant \ell - 1$, their entry n_a is strictly smaller than the entry n_a in (b.1). Also, the pairs of the form (b.3) are different from the pair (b.1) because their entry n_b is strictly smaller than the entry n_b 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_a, n_b) is a pair of the form (b.2), then $n_a/2^k$ is even and $n_b/2^k$ is odd, while if (n_a, n_b) is a pair of the form (b.3), then $n_a/2^k$ is odd and $n_b/2^k$ is even. Therefore, no pair can simultaneously be of the form (b.2) and (b.3). □

Example 1. 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 5. *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

For simplicity, we call 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 *forbidden.* Using this notation we have

$$|QB(n)| = 2M_0(n) + 1 \quad \begin{array}{l}\text{minus the number of forbidden maximal}\\ \text{sequences of zeroes in } n_{(2)}.\end{array} \tag{9}$$

In the subtraction in this formula 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 or in the units position 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 1, together with Corollary 3, provide the following algorithm to produce all minimal Colless trees in $\mathcal{T}_n$, which is reminiscent of Aldous' β-model [1].

Algorithm 1: MinColless

```
Start with a single node labeled $n$;
while the current tree contains labeled leaves do
    Choose a leaf with label $m$;
    if $m$ is a power of 2 then
        replace this leaf by a fully symmetric tree $T^{fs}_{\log_2(m)}$ with its nodes unlabeled;
    end
    else
        Find a pair of integers $(m_a, m_b) \in QB(m)$;
        Split the leaf labeled $m$ into a cherry with unlabeled root and its leaves labeled $m_a$ and $m_b$,
          respectively.
    end
end
```

Example 2. Let us use this Algorithm MinColless to find all minimal Colless trees with 20 leaves; 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] See `http://evolution.genetics.washington.edu/phylip/newicktree.html`

1) We start with a single node labeled 20.

2) Since $QB(20) = \{(10,10),(12,8)\}$, this node can split 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 from these trees through all possible combinations of splitting the leaves in the trees:

$$\begin{aligned}
&((5,5),(5,5)) \Rightarrow (((3,2),(3,2)),((3,2),(3,2)))\\
&\quad \Rightarrow ((((2,1),2),((2,1),2)),(((2,1),2),((2,1),2)))\\
&\quad \Rightarrow (((((\cdot,\cdot),\cdot),(\cdot,\cdot)),(((\cdot,\cdot),\cdot),(\cdot,\cdot))),((((\cdot,\cdot),\cdot),(\cdot,\cdot)),(((\cdot,\cdot),\cdot),(\cdot,\cdot))))\\
&((5,5),(6,4)) \Rightarrow (((3,2),(3,2)),((3,3),4))\\
&\quad \Rightarrow ((((2,1),2),((2,1),2)),(((2,1),(2,1)),4))\\
&\quad \Rightarrow (((((\cdot,\cdot),\cdot),(\cdot,\cdot)),(((\cdot,\cdot),\cdot),(\cdot,\cdot))),((((\cdot,\cdot),\cdot),((\cdot,\cdot),\cdot)),((\cdot,\cdot),(\cdot,\cdot)))\\
&((5,5),(6,4)) \Rightarrow (((3,2),(3,2)),((4,2),4))\\
&\quad \Rightarrow ((((2,1),2),((2,1),2)),((4,2),4))\\
&\quad \Rightarrow (((((\cdot,\cdot),\cdot),(\cdot,\cdot)),(((\cdot,\cdot),\cdot),(\cdot,\cdot))),((((\cdot,\cdot),(\cdot,\cdot)),(\cdot,\cdot)),((\cdot,\cdot),(\cdot,\cdot))))\\
&((6,4),(6,4)) \Rightarrow (((3,3),4),((3,3),4))\\
&\quad \Rightarrow ((((2,1),(2,1)),4),(((2,1),(2,1)),4))\\
&\quad \Rightarrow (((((\cdot,\cdot),\cdot),((\cdot,\cdot),\cdot)),((\cdot,\cdot),(\cdot,\cdot))),((((\cdot,\cdot),\cdot),((\cdot,\cdot),\cdot)),((\cdot,\cdot),(\cdot,\cdot))))\\
&((6,4),(6,4)) \Rightarrow (((3,3),4),((4,2),4))\\
&\quad \Rightarrow ((((2,1),(2,1)),4),((4,2),4))\\
&\quad \Rightarrow (((((\cdot,\cdot),\cdot),((\cdot,\cdot),\cdot)),((\cdot,\cdot),(\cdot,\cdot))),((((\cdot,\cdot),(\cdot,\cdot)),(\cdot,\cdot)),((\cdot,\cdot),(\cdot,\cdot))))\\
&((6,4),(6,4)) \Rightarrow (((4,2),4),((4,2),4))\\
&\quad \Rightarrow (((((\cdot,\cdot),(\cdot,\cdot)),(\cdot,\cdot)),((\cdot,\cdot),(\cdot,\cdot))),((((\cdot,\cdot),(\cdot,\cdot)),(\cdot,\cdot)),((\cdot,\cdot),(\cdot,\cdot))))
\end{aligned}$$

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 $QB(3) = \{(2,1)\}$,

$$\begin{aligned}
&((6,6),8) \Rightarrow (((3,3),(3,3)),8) \Rightarrow ((((2,1),(2,1)),((2,1),(2,1))),8)\\
&\quad \Rightarrow (((((\cdot,\cdot),\cdot),((\cdot,\cdot),\cdot)),(((\cdot,\cdot),\cdot),((\cdot,\cdot),\cdot))),(((\cdot,\cdot),(\cdot,\cdot)),((\cdot,\cdot),(\cdot,\cdot))))\\
&((6,6),8) \Rightarrow (((3,3),(4,2)),8) \Rightarrow ((((2,1),(2,1)),(4,2)),8)\\
&\quad \Rightarrow (((((\cdot,\cdot),\cdot),((\cdot,\cdot),\cdot)),(((\cdot,\cdot),(\cdot,\cdot)),(\cdot,\cdot))),(((\cdot,\cdot),(\cdot,\cdot)),((\cdot,\cdot),(\cdot,\cdot))))\\
&((6,6),8) \Rightarrow (((4,2),(4,2)),8)\\
&\quad \Rightarrow (((((\cdot,\cdot),(\cdot,\cdot)),(\cdot,\cdot)),(((\cdot,\cdot),(\cdot,\cdot)),(\cdot,\cdot))),(((\cdot,\cdot),(\cdot,\cdot)),((\cdot,\cdot),(\cdot,\cdot))))\\
&((8,4),8)\\
&\quad \Rightarrow (((((\cdot,\cdot),(\cdot,\cdot)),((\cdot,\cdot),(\cdot,\cdot))),((\cdot,\cdot),(\cdot,\cdot))),(((\cdot,\cdot),(\cdot,\cdot)),((\cdot,\cdot),(\cdot,\cdot))))
\end{aligned}$$

So, there are 10 different minimal Colless trees in $\mathcal{T}_{20}$. We depict them in Figure 5 below.

We have implemented the Algorithm MinColless, with step 8 efficiently carried out by means of Proposition 3, in a Python script that generates, for every n, the Newick description of all minimal Colless trees in $\mathcal{T}_n$. 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 1 to 128 all such minimal Colless trees in $\mathcal{T}_n$. Figure 6 shows their number for every n. These numbers are in agreement with those provided by the recurrence established in Proposition 4 in the next subsection.

4.2. Counting minimal Colless trees

Let $\widetilde{\mathcal{MC}}_n$ denote the set of all minimal Colless trees in $\mathcal{T}_n$ and $\widetilde{c}(n) := \big|\widetilde{\mathcal{MC}}_n\big|$ its cardinality. To simplify the notations, set

$$\widetilde{QB}(n) := \{(n_a, n_b) \in QB(n) \mid n_a > n_b\}.$$

We have the following recursive formula for $\widetilde{c}(n)$:

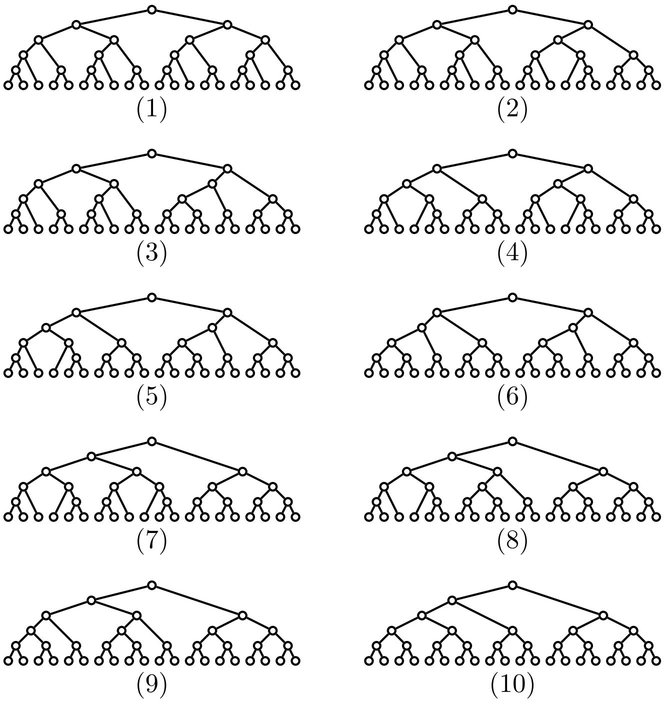
(1) (2) (3) (4) (5) (6) (7) (8) (9) (10)

Figure 5: The 10 trees in $\mathcal{T}_{20}$ with minimum Colless index, 8. They are enumerated in the same order as they have been produced in Example 2.

Proposition 4. *The sequence $\widetilde{c}(n)$ satisfies that $\widetilde{c}(1) = 1$ and, for every $n \geqslant 2$,*

$$\widetilde{c}(n) = \sum_{(n_a,n_b)\in\widetilde{QB}(n)} \widetilde{c}(n_a)\cdot\widetilde{c}(n_b) + \binom{\widetilde{c}(n/2)+1}{2}\cdot\delta_{even}(n)$$

where $\delta_{even}(n) = 1$ if n is even and 0 otherwise.

Proof. By Lemma 2 and Proposition 1, $T = (T_a, T_b) \in \widetilde{\mathcal{MC}}_n$ if, and only if, $(n_a, n_b) \in QB(n)$, $T_a \in \widetilde{\mathcal{MC}}_{n_a}$ and $T_b \in \widetilde{\mathcal{MC}}_{n_b}$. The correctness of the formula in the statement stems then from the following three facts:

- If n is odd, $\widetilde{\mathcal{MC}}_n$ is in bijection with the set

  $$X_n = \big\{(n_a, n_b, T_a, T_b) \mid (n_a, n_b) \in \widetilde{QB}(n), T_a \in \widetilde{\mathcal{MC}}_{n_a}, T_b \in \widetilde{\mathcal{MC}}_{n_b}\big\},$$

  through the relation

  $$T = (T_a, T_b) \in \widetilde{\mathcal{MC}}_n \Longleftrightarrow (n_a, n_b, T_a, T_b) \in X_n.$$

- If n is even, $\widetilde{\mathcal{MC}}_n$ is in bijection with the set

  $$X_n \sqcup \big\{\{T_a, T_b\} \mid T_a, T_b \in \widetilde{\mathcal{MC}}_{n/2}, T_a \neq T_b\big\} \sqcup \big\{T_a \mid T_a \in \widetilde{\mathcal{MC}}_{n/2}\big\}$$

  through the relation

  $$T = (T_a, T_b) \in \widetilde{\mathcal{MC}}_n \Longleftrightarrow \begin{cases} n_a > n_b \text{ and } (n_a, n_b, T_a, T_b) \in X_n \\ n_a = n_b, T_a \neq T_b, \text{ and } T_a, T_b \in \widetilde{\mathcal{MC}}_{n/2} \\ n_a = n_b, T_a = T_b \in \widetilde{\mathcal{MC}}_{n/2} \end{cases}$$

Number $\widetilde{c}(n)$ of trees with n leaves and minimum Colless index for $n = 1, \dots, 128$

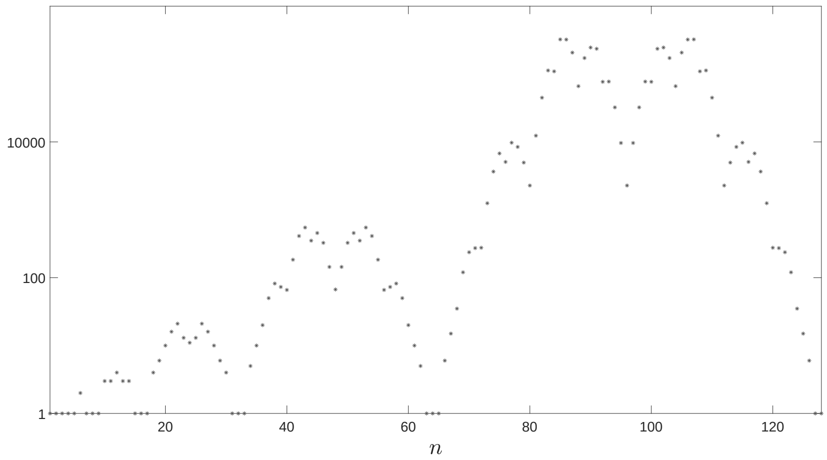

Figure 6: Plot of $\widetilde{c}(n)$ for $n = 1, \dots, 128$.

- The cardinality of X_n is $\sum\limits_{(n_a,n_b)\in\widetilde{QB}(n)} \widetilde{c}(n_a) \cdot \widetilde{c}(n_b)$ and the cardinality of
$$\big\{\{T_a, T_b\} \mid T_a, T_b \in \widetilde{\mathcal{MC}}_{n/2}, T_a \neq T_b\big\} \cup \big\{T_a \mid T_a \in \widetilde{\mathcal{MC}}_{n/2}\big\}$$
is $\binom{\widetilde{c}(n/2)+1}{2}$.

□

The sequence $\widetilde{c}(n)$ seems to be new in the literature, and it has been added to the *Online Encyclopedia of Integer Sequences* [28] as sequence A307689. It would definitely be of interest to find an explicit formula for $\widetilde{c}(n)$ and to analyze the fractal structure suggested by Figure 6, which continues for larger values of n and seems also related to the Blancmange curve (compare Figure 6 with Figure 3).

4.3. Another family of minimal Colless trees

As we have seen in Theorem 1, the maximally balanced trees T_n^{mb} have minimum Colless index. These trees are obtained through the recursive strategy suggested by Corollary 2: given a number n of leaves, we split n into $n_a = \lceil n/2 \rceil$ and $n_b = \lfloor n/2 \rfloor$ and we produce a tree $T = (T_a, T_b)$ with $T_a \in \mathcal{T}_{n_a}$ and $T_b \in \mathcal{T}_{n_b}$ constructed recursively through the same procedure. This strategy could be understood as “greedy from the top” because, starting at the root and going towards the leaves, we bipartition the leaf set of each rooted subtree into two sets so that the difference of their cardinalities is minimized.

There is another strategy for building minimal Colless trees, which we call “greedy from the bottom”, where instead of minimally splitting the sets of leaves, one minimally joins rooted subtrees by pending them from a common parent of their roots, as in the coalescent process [17]. More specifically, these trees are constructed by means of the following algorithm:

Algorithm 2: Greedy from the bottom

```
$n \leftarrow$ number of taxa;
$treeset \leftarrow n$ trees consisting of one node each;
$min \leftarrow 1$                // minimal number of leaves of all trees in treeset;
while $|treeset| > 1$ do
    $u \leftarrow$ tree from $treeset$ with $min$ leaves;
    $treeset = treeset \setminus \{u\}$;
    $min \leftarrow$ minimal number of leaves of all trees in $treeset$;
    $v \leftarrow$ tree from $treeset$ with $min$ leaves;
    $treeset = treeset \setminus \{v\}$;
    $newtree \leftarrow$ tree consisting of new root $\rho_{uv}$ and maximal pending subtrees $u$ and $v$;
    $treeset \leftarrow treeset \cup \{newtree\}$;
    $min \leftarrow$ minimal number of leaves of all trees in $treeset$;
end
$finaltree \leftarrow treeset[1]$                // i.e.  only remaining element of treeset;
return $finaltree$;
```

We shall call henceforth *greedy from the bottom*, or simply *GFB*, any bifurcating tree with n leaves that results from Algorithm 2, and we shall denote it by T_n^{gfb}. This notation leads to no ambiguity, because of the following lemma.

Lemma 4. *For every $n \geqslant 1$, there exists only one GFB tree with n leaves (up to isomorphisms).*

Proof. When $n = 1$, Algorithm 2 skips the **while** loop and it returns the only tree in $\mathcal{T}_1$. Assume now that $n \geqslant 2$. With the notations of Algorithm 2, let us denote by $treeset_k$, for $k = 1, \ldots, n-1$, the content of the auxiliary tree multiset *treeset* after the k-th iteration of the **while** loop. We shall prove by induction on k that, for every two applications of Algorithm 2 with input n (whose *treesets* will be distinguished henceforth with superscripts (1) and (2)):

(a) We have the equality of tree multisets $treeset_k^{(1)} = treeset_k^{(2)}$, which means that these two multisets of trees have the same elements with the same multiplicities; and

(b) For every $2 \leqslant m \leqslant n$, all trees with m leaves created in the first k iterations of the loop in both applications of the algorithm have the same shape.

This will imply that when, after $n-1$ iterations of the loop, both multisets $treeset_{n-1}^{(i)}$, $i = 1, 2$, consist of a single tree with n leaves, these two trees are the same.

The base case $k = 1$ is obvious, because $treeset_1$ always consists of a cherry and $n-2$ isolated nodes. Assume now that the statement is true for the $(k-1)$-th iteration, and in particular that, immediately before the k-th iteration, $treeset_{k-1}^{(1)} = treeset_{k-1}^{(2)}$ (by (a)) and this multiset contains trees of only one shape for each present number of leaves (by (b)). This implies that the minimal number of leaves of a tree in $treeset_{k-1}^{(1)}$ and $treeset_{k-1}^{(2)}$ is the same, let us call it m_1, and that all trees with m_1 leaves in both *treeset* have the same shape. Moreover, if we remove one tree with m_1 leaves from each *treeset* (which will be the same tree —up to isomorphisms— in both applications of the algorithm), the resulting multisets are equal again, and therefore the minimal number of leaves of a tree in each one of them is again the same, let us call it m_2, and all trees with m_2 leaves in both multisets are equal. Then, in the k-th iteration of the loop in each application of the algorithm, we remove from the corresponding *treeset* the same tree with m_1 leaves and the same tree with m_2 leaves and we add the same tree with $m_1 + m_2$ leaves, obtained by pending the removed trees to a common root. This proves that $treeset_k^{(1)} = treeset_k^{(2)}$, i.e. assertion (a).

To prove that (b) also holds, it remains to check that if some $treeset_j^{(1)}$ with $j \leqslant k-1$ already contained some tree T' with $m_1 + m_2$ leaves, then it has the same shape as the new one. Assume that such a tree T' with $m_1 + m_2$ leaves has been created in the j-th iteration of the loop. Let m_1' and m_2', with $m_1' \leqslant m_2'$, be the numbers of leaves of the maximal pending subtrees of T'. By construction, this means that the minimal number of leaves of any tree in the multiset $\mathit{treeset}^{(1)}_{j-1}$ was m'_1, and the second minimal number of leaves was m'_2. Now, remember that, in each iteration of the loop, two trees are removed from the *treeset* and replaced by a tree with number of leaves the sum of the numbers of leaves of the removed trees. This clearly implies that the minimal and second minimal numbers of leaves of members of the *treeset* cannot decrease in any such iteration. Therefore, $m'_1 \leqslant m_1$, because if $m_1 < m'_1$, then $\mathit{treeset}^{(1)}_{j-1}$ cannot contain any tree with m_1 leaves (as m'_1 is the minimal number of leaves of a member of $\mathit{treeset}^{(1)}_{j-1}$) and such a tree cannot be added in further iterations of the loop, but there is at least one such tree in $\mathit{treeset}^{(1)}_{k-1}$. Since $m_1+m_2 = m'_1+m'_2$, if $m'_1 < m_1$ then $m'_2 > m_2$, but a similar argument shows that this inequality is in contradiction with the fact that m'_2 is the smallest number of leaves of a tree in $\mathit{treeset}^{(1)}_{j-1}$ after removing a tree with m'_1 leaves. Therefore, $m'_1 = m_1$ and hence $m'_2 = m_2$, too. But then, by (b) in the induction hypothesis, the trees with m_1 and m_2 leaves combined in the j-th iteration of the first application of the algorithm have the same shape as the trees with m_1 and m_2 leaves combined in the k-th iteration, and therefore the tree with m_1+m_2 leaves that already existed in $\mathit{treeset}^{(1)}_{j}$ has the same shape as the one added in the k-th iteration. This completes the proof of the inductive step. □

Note that Algorithm 2 greedily clusters trees of minimal numbers of leaves starting with single nodes and proceeding until only one tree is left, which is the reason we call the resulting trees "greedy from the bottom." Our main goal in this subsection is to prove that they are also minimal Colless and, in general, different from the maximally balanced trees with the same number of leaves (cf. Figure 4).

Next result easily implies that any rooted subtree of a GFB tree is also a GFB tree, by induction on the depth of the subtree's root.

Lemma 5. *If $T = (T_a, T_b)$ is a GFB tree, then T_a and T_b are also GFB trees.*

Proof. Let $T = (T_a, T_b)$ be a GFB tree and let n_a and n_b denote the numbers of leaves of T_a and T_b, respectively. This entails that Algorithm 2 induces a bipartition of the n leaves into two disjoint sets of sizes n_a and n_b, respectively, in the sense that all iterations of the **while** loop except for the very last one combine pairs of subtrees with both sets of leaves contained either in $V_L(T_a)$ or in $V_L(T_b)$.

Now, when in an iteration of the algorithm a pair of subtrees of T_a is combined, it is because their numbers of leaves are the two smallest ones in the global *treeset*, and hence also in the submultiset of *treeset* consisting only of trees with leaves in $V_L(T_a)$. This shows that T_a is obtained through the application of Algorithm 2 to n_a leaves, i.e. $T_a = T^{gfb}_{n_a}$, and by symmetry $T_b = T^{gfb}_{n_b}$. □

The next proposition characterizes the pairs of numbers of leaves of the maximal pending subtrees of a GFB tree. Besides allowing the construction of GFB trees through an alternative top-to-bottom procedure, by splitting clusters into subclusters of suitable sizes, this characterization easily entails that the GFB trees almost never are maximally balanced, and moreover it will allow us to use Proposition 1 to prove that the GFB trees are minimal Colless (see Theorem 6 below).

Proposition 5. *Let $T^{gfb}_n = (T_a, T_b)$ be a GFB tree with $n \geqslant 2$, $T_a \in \mathcal{T}_{n_a}$, $T_b \in \mathcal{T}_{n_b}$ and $n_a \geqslant n_b$. Let $n = 2^m + p$ with $m = \lfloor \log_2(n) \rfloor$ and $0 \leqslant p < 2^m$. Then, we have:*

(i) *If $0 \leqslant p \leqslant 2^{m-1}$, then $n_a = 2^{m-1} + p$, $n_b = 2^{m-1}$ and T_b is fully symmetric.*

(ii) *If $2^{m-1} \leqslant p < 2^m$, $n_a = 2^m$, $n_b = p$ and T_a is fully symmetric.*

Since the proof of this proposition is quite long, we postpone it until Appendix A.2 at the end of the manuscript.

Remark 2. We want to point out here that as a byproduct of the proof of Proposition 5 provided in Appendix A.2 we obtain that if $n \geqslant 3$ is any odd number of leaves, then the GFB trees T^{gfb}_{n-1}, T^{gfb}_{n}, and T^{gfb}_{n+1} have a maximal pending subtree in common, which is moreover fully symmetric. Using that the maximal pending subtrees of a GFB tree are again GFB (Lemma 5), their explicit numbers of leaves provided by

Proposition 5, and the next proposition, which clearly implies that the GFB trees with numbers of leaves that are powers of 2 must be fully symmetric, this curious result on T^{gfb}_{n-1}, T^{gfb}_{n}, and T^{gfb}_{n+1} is easily extended to any number of leaves n that is not of the form 3×2^m.

Now, as we announced, we next use Proposition 5 to prove that the GFB trees always have minimum Colless index:

Proposition 6. *Let T^{gfb}_n be the GFB tree with n leaves. Then, $\mathcal{C}(T^{gfb}_n) = c_n$.*

Proof. We prove that T^{gfb}_n is Colless minimal by induction on the number of leaves n. The base case $n = 1$ is obvious, because there is only one tree in $\mathcal{T}_1$. Assume now that every GFB tree with at most $n-1$ leaves is Colless minimal and consider the tree T^{gfb}_n. By Lemma 5, if $T^{gfb}_n = (T_a, T_b)$, then T_a and T_b are GFB trees and then, by the induction hypothesis, they are Colless minimal and in particular $\mathcal{C}(T_a) = c_{n_a}$ and $\mathcal{C}(T_b) = c_{n_b}$. Let us write n as $2^m + p$ with $m = \lfloor \log_2(n) \rfloor$ and $0 \leqslant p < 2^m$, and consider its binary expansion $n = \sum\limits_{j=1}^{\ell} 2^{m_j}$ with $m_1 > \cdots > m_\ell$, so that $m_1 = m$ and $p = \sum\limits_{j=2}^{\ell} 2^{m_j}$ is the binary expansion of p if $p > 0$. Now:

(i) If $p = 0$, then, by Proposition 5, $n_a = n_b = 2^{m-1}$, and then, by Lemma 1 and the induction hypothesis,

$$\mathcal{C}(T^{gfb}_n) = \mathcal{C}(T^{gfb}_{n_a}) + \mathcal{C}(T^{gfb}_{n_b}) + n_a - n_b = c_{n_a} + c_{n_b} + n_a - n_b = 0 = c_n.$$

(ii) If $1 \leqslant p < 2^{m-1}$, then, by Proposition 5, $n_a = 2^{m-1} + p$ and $n_b = 2^{m-1}$. In this case, $m_2 < m - 1 = m_1 - 1$ and thus $n_a = 2^{m_1 - 1} + \sum\limits_{j=2}^{\ell} 2^{m_j}$ is the binary expansion of n_a. So, by Theorem 2 and the induction hypothesis, $\mathcal{C}(T^{gfb}_{n_b}) = c_{n_b} = 0$ and

$$\mathcal{C}(T^{gfb}_{n_a}) = c_{n_a} = \sum_{j=2}^{\ell} 2^{m_j}(m_1 - 1 - m_j - 2(j-2))$$

and then, by Lemma 1,

$$\begin{aligned}\mathcal{C}(T^{gfb}_n) &= \mathcal{C}(T^{gfb}_{n_a}) + \mathcal{C}(T^{gfb}_{n_b}) + n_a - n_b \\ &= \sum_{j=2}^{\ell} 2^{m_j}(m_1 - 1 - m_j - 2(j-2)) + \sum_{j=2}^{\ell} 2^{m_j} \\ &= \sum_{j=2}^{\ell} 2^{m_j}(m_1 - m_j - 2(j-2)) = c_n.\end{aligned}$$

(iii) If $p = 2^{m-1}$, so that $n = 2^m + 2^{m-1}$ is the binary expansion of n, then, by Proposition 5, $n_a = 2^m$ and $n_b = 2^{m-1}$. In this case, by the induction hypothesis, $\mathcal{C}(T_a) = c_{n_a} = 0$ and $\mathcal{C}(T_b) = c_{n_b} = 0$, and then, by Lemma 1,

$$\mathcal{C}(T^{gfb}_n) = \mathcal{C}(T^{gfb}_{n_a}) + \mathcal{C}(T^{gfb}_{n_b}) + n_a - n_b = 2^{m-1} = c_n$$

by Theorem 2.

(iv) Finally, assume that $p > 2^{m-1}$, so that its binary expansion is $p = 2^{m-1} + 2^{m_3} + \cdots + 2^{m_\ell}$, and in particular $m_2 = m - 1$. In this case, $n_a = 2^m$ and $n_b = p$, so that $n_a - n_b = 2^m - p = 2^{m-1} - (2^{m_3} +$

$\cdots + 2^{m_\ell}$), and, by the induction hypothesis, $\mathcal{C}(T^{gfb}_{n_a}) = c_{n_a} = 0$ and

$$\begin{aligned}
C(T^{gfb}_{n_b}) = c_{n_b} &= \sum_{j=3}^{\ell} 2^{m_j}(m_2 - m_j - 2(j-1-2)) \\
&= \sum_{j=3}^{\ell} 2^{m_j}(m - 1 - m_j - 2(j-2) + 2) \\
&= \sum_{j=3}^{\ell} 2^{m_j}(m - m_j - 2(j-2)) + \sum_{j=3}^{\ell} 2^{m_j}
\end{aligned}$$

Then, by Lemma 1,

$$\begin{aligned}
\mathcal{C}(T^{gfb}_{n}) &= \mathcal{C}(T^{gfb}_{n_a}) + \mathcal{C}(T^{gfb}_{n_b}) + n_a - n_b \\
&= \sum_{j=3}^{\ell} 2^{m_j}(m_1 - m_j - 2(j-2)) + \sum_{j=3}^{\ell} 2^{m_j} + 2^{m-1} - \sum_{j=3}^{\ell} 2^{m_j} \\
&= \sum_{j=3}^{\ell} 2^{m_j}(m_1 - m_j - 2(j-2)) + 2^{m_2} \\
&= \sum_{j=2}^{\ell} 2^{m_j}(m_1 - m_j - 2(j-2)) = c_n
\end{aligned}$$

(in the third and fourth equalities we use that $m = m_1$ and $m_2 = m - 1 = m_1 - 1$) as we wanted to show.

□

So, for any given number n of leaves, both the maximally balanced trees and the GFB trees have minimum Colless index. Moreover, while the balance value of the root of T^{mb}_n is by definition at most 1, Proposition 5 implies that if $n = 2^m + p$ with $m = \lfloor \log_2(n) \rfloor$, the balance value of the root of T^{gfb}_n is $\min\{p, 2^m - p\}$ and therefore $T^{mb}_n \neq T^{gfb}_n$ if $p \neq 0, 1, 2^m - 1$. On the other hand, we already know (cf. Corollary 3) that if $n = 2^m$, then there is only one minimal Colless tree with n leaves and therefore in this case $T^{mb}_n = T^{gfb}_n$, and it is straightforward to prove by induction on m, using Proposition 4 and the fact that $QB(2^m - 1) = \{(2^{m-1}, 2^{m-1} - 1)\}$ and $QB(2^m + 1) = \{(2^{m-1} + 1, 2^{m-1})\}$, that if n has the form $2^m \pm 1$, then there is only one minimal Colless tree in $\mathcal{T}_n$, too. In summary, this proves the following result.

Corollary 6. *For every $n \geqslant 1$, if $n \notin \{2^m - 1, 2^m, 2^m + 1\}$ for any $m \in \mathbb{N}_{\geqslant 1}$, then $T^{mb}_n \neq T^{gfb}_n$, while if $n \in \{2^m - 1, 2^m, 2^m + 1\}$ for some $m \in \mathbb{N}_{\geqslant 1}$, then there is only one minimal Colless tree in $\mathcal{T}_n$.*

The next result entails that the GFB trees can also be built through a top-down strategy as follows: we start with a cluster of n leaves, and build a hierarchical clustering by splitting clusters into pairs of subclusters of suitable cardinalities.

Corollary 7. *For every $T \in \mathcal{T}_n$, $T = T^{gfb}_n$ if, and only if, for every $v \in \mathring{V}(T)$, if we write $\kappa_T(v) = 2^k + s$ with $k = \lfloor \log_2(\kappa_T(v)) \rfloor$ and $0 \leqslant s < 2^k$, then the numbers of descendant leaves of the children of v are, respectively, $2^{k-1} + s$ and 2^{k-1}, if $0 \leqslant s \leqslant 2^{k-1}$, or 2^k and s, if $2^{k-1} \leqslant s < 2^k$.*

Proof. The "only if" implication is a direct consequence of Proposition 5 and the fact that, as as a consequence of Lemma 5, any rooted subtree of a GFB tree is again GFB. We prove now the "if" implication by induction on n. The base case when $n = 1$ is obvious, because there is only one tree with 1 leaf. Assume now that this implication is true for every $1 \leqslant n' < n$, and let $T \in \mathcal{T}_n$ be such that, for every $v \in \mathring{V}(T)$, if we write $\kappa_T(v) = 2^k + s$ with $k = \lfloor \log_2(\kappa_T(v)) \rfloor$ and $0 \leqslant s < 2^k$, then the numbers of descendant leaves of the children of v are, respectively, $2^{k-1}+s$ and 2^{k-1}, if $0\leqslant s\leqslant 2^{k-1}$, or 2^k and s, if $2^{k-1}\leqslant s<2^k$. Consider the decomposition $T=(T_a,T_b)$ of T into its two maximal pending subtrees, with $T_a\in\mathcal{T}_{n_a}$ and $T_b\in\mathcal{T}_{n_b}$, $n_a\geqslant n_b$. Then, on the one hand, the internal nodes of both T_a and T_b satisfy the aforementioned property on the numbers of descendant leaves of their children, which implies by the induction hypothesis that $T_a=T^{gfb}_{n_a}$ and $T_b=T^{gfb}_{n_b}$. And, on the other hand, by hypothesis n_a and n_b satisfy that if we write $n=2^m+p$, with $m=\lfloor\log_2(n)\rfloor$ and $0\leqslant p<2^m$, then $n_a=2^{m-1}+p$ and $n_b=2^{m-1}$, if $0\leqslant p\leqslant 2^{m-1}$, or $n_a=2^m$ and $n_b=p$, if $2^{m-1}\leqslant p<2^m$. But then, by Proposition 5 and Lemma 5, the decomposition of T^{gfb}_n into its maximal pending subtrees is $(T^{gfb}_{n_a},T^{gfb}_{n_b})$ with n_a and n_b precisely given by these formulas. This implies that $T=T^{gfb}_n$.

□

The maximally balanced trees and the GFB trees turn out to be extremal among the minimal Colless trees in the sense that no minimal Colless tree can have a smaller difference between the number of leaves of its maximal pending subtrees than the maximally balanced tree or a larger difference between these numbers than the GFB tree. The assertion on the maximally balanced trees being obvious, because that difference is the least possible one (0 or 1, depending on whether the number of leaves is even or odd, respectively), we must prove the assertion on the GFB trees.

Proposition 7. *Let $T^{gfb}_n=(T^{gfb}_{n^{gfb}_a},T^{gfb}_{n^{gfb}_b})$ be the decomposition of a GFB tree with n leaves into its maximal pending subtrees. If $T=(T_a,T_b)$, with $T_a\in\mathcal{T}_{n_a}$ and $T_b\in\mathcal{T}_{n_b}$, is another minimal Colless tree with n leaves, then $n_a-n_b\leqslant n^{gfb}_a-n^{gfb}_b$.*

Proof. Write n as $n=2^m+p$ with $m=\lfloor\log_2(n)\rfloor$ and $0\leqslant p<2^m-1$. We know from Proposition 5 that if $0\leqslant p\leqslant 2^{m-1}$, then $(n^{gfb}_a,n^{gfb}_b)=(2^{m-1}+p,2^{m-1})$ and hence $n^{gfb}_a-n^{gfb}_b=p$, and if $2^{m-1}\leqslant p<2^m$, then $(n^{gfb}_a,n^{gfb}_b)=(2^m,p)$ and hence $n^{gfb}_a-n^{gfb}_b=2^m-p$. Moreover, if $p\in\{0,1,2^m-1\}$, we know from Corollary 6 that there is only one minimal Colless tree in $\mathcal{T}_n$, and therefore we can assume henceforth that $2\leqslant p\leqslant 2^m-2$.

Now, if $T=(T_a,T_b)\in\mathcal{T}_n$ is Colless minimal, then, by Proposition 1, $(n_a,n_b)\in QB(n)$. Therefore, it is enough to prove that if $(n_a,n_b)\in QB(n)$, then $n_a-n_b\leqslant\min\{p,2^m-p\}$. We shall do it using the explicit description of $QB(n)$ given in Proposition 3. So, let 2^k be the largest power of 2 that divides n, which is also the largest power of 2 that divides p, and let $2^{m_1}+\cdots+2^{m_\ell}$, with $m_1=m-k>\cdots>m_\ell=1$ be the binary expansion of $n_0=n/2^k$, so that $p=2^k(2^{m_2}+\cdots+2^{m_\ell})$.

Then, using the same notations as in Proposition 3:

(a) Since n is not a power of 2, this case cannot happen.

(b.1) If (n_a,n_b) has the form

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

then

$$n_a-n_b=2^k\leqslant\min\{p,2^m-p\}$$

because 2^k divides both p and 2^m-p.

(b.2) If (n_a,n_b) has the form

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

for some $j=2,\ldots,\ell-1$ such that $m_j>m_{j+1}+1$, then

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

and this is smaller or equal than $\min\{p, 2^m - p\}$ because, on the one hand,

$$\begin{aligned}
2^k\Big(\sum_{i=1}^{j-1} 2^{m_i} + 2^{m_j+1}\Big) - n &\leqslant 2^k\Big(\sum_{i=1}^{\ell} 2^{m_i} + 2^{m_j}\Big) - n = 2^k(n_0 + 2^{m_j}) - n\\
&= 2^k \cdot 2^{m_j} \leqslant 2^k \sum_{i=2}^{\ell} 2^{m_i} = p
\end{aligned}$$

and, on the other hand,

$$\begin{aligned}
2^k\Big(\sum_{i=1}^{j-1} 2^{m_i} + 2^{m_j+1}\Big) - n &\leqslant 2^k\Big(\sum_{i=1}^{j-1} 2^{m_i} + 2^{m_{j-1}}\Big) - n\\
&\leqslant 2^k\Big(\sum_{i=m_{j-1}}^{m_1} 2^{i} + 2^{m_{j-1}}\Big) - n = 2^k \cdot 2^{m_1+1} - n = 2^{m+1} - n = 2^m - p.
\end{aligned}$$

(b.3) If (n_a, n_b) has the form

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

for some $j = 2, \dots, \ell - 1$ such that $m_j < m_{j-1} - 1$, then

$$n_a - n_b = n - 2 \cdot 2^k \sum_{i=1}^{j-1} 2^{m_i-1} = n - 2^k \sum_{i=1}^{j-1} 2^{m_i} \leqslant \min\{p, 2^m - p\}$$

because, on the one hand

$$n - 2^k \sum_{i=1}^{j-1} 2^{m_i} \leqslant n - 2^k \cdot 2^{m_1} = n - 2^m = p,$$

and, on the other hand,

$$\begin{aligned}
n - 2^k \sum_{i=1}^{j-1} 2^{m_i} &= 2^k \sum_{i=1}^{\ell} 2^{m_i} - 2^k \sum_{i=1}^{j-1} 2^{m_i} = 2^k \sum_{i=j}^{\ell} 2^{m_i}\\
&\leqslant 2^k \sum_{i=0}^{m_j} 2^i = 2^k(2^{m_j+1} - 1) < 2^k\Big(2^{m_1} - \sum_{i=2}^{\ell} 2^{m_i}\Big) = 2^m - p,
\end{aligned}$$

where the last inequality holds because $m_j < m_j + 1 < m_{j-1}$ implies that

$$2^{m_j+1} + \sum_{i=2}^{\ell} 2^{m_i} \leqslant \sum_{s=0}^{m_2} 2^s = 2^{m_2+1} - 1 < 2^{m_1}.$$

(b.4) If $(n_a, n_b) = (n/2, n/2)$, then $n_a - n_b = 0 < \min\{p, 2^m - p\}$.

□

We now immediately have:

Corollary 8. *Let T_n^{mb} be the maximally balanced tree with $n \geqslant 2$ leaves and $n_a^{mb} \geqslant n_b^{mb}$ the numbers of leaves of its maximal pending subtrees. Let T_n^{gfb} be the GFB tree with n leaves and $n_a^{gfb} \geqslant n_b^{gfb}$ the numbers of leaves of its maximal pending subtrees. Then, for every minimal Colless tree $T \in \mathcal{T}_n$, if $n_a \geqslant n_b$ are the numbers of leaves of its maximal pending subtrees,*

$$n_b^{gfb} \leqslant n_b \leqslant n_b^{mb} \leqslant n_a^{mb} \leqslant n_a \leqslant n_a^{gfb}.$$

Proof. Assume that $n_a < n_a^{mb}$. Since $n_a + n_b = n_a^{mb} + n_b^{mb}$, this would imply that $n_b > n_b^{mb}$. But since $n_a^{mb} = n_b^{mb} = n/2$ for n even and $n_a^{mb} = (n+1)/2$ and $n_b^{mb} = (n-1)/2$ for n odd, this would contradict the assumption that $n_a \geqslant n_b$. Thus, $n_a \geqslant n_a^{mb}$. A similar argument shows that $n_b \leqslant n_b^{mb}$.

Assume now that $n_a > n_a^{gfb}$. Then, since $n_a + n_b = n_a^{gfb} + n_b^{gfb}$, this would imply that $n_b < n_b^{gfb}$ and hence that $n_a - n_b > n_a^{gfb} - n_b^{gfb}$, which contradicts Theorem 7. A similar argument shows that $n_b^{gfb} \leqslant n_b$. □

Remark 3. Since any rooted subtree of a minimal Colless tree (respectively, of a maximally balanced tree or a GFB tree) is again minimal Colless (respectively, maximally balanced or GFB), the last corollary applies not only to the numbers of leaves of the maximal pending subtrees of a minimal Colless tree, but also to the numbers of descendant leaves of the children of any internal node v in minimal Colless trees, relative to the number of descendant leaves of v.

4.4. The minimal Colless trees have also minimum Sackin index

We shortly focus next on another popular index of tree balance, namely the so-called Sackin index [26, 27]. Recall that the *Sackin index* of a (not necessarily bifurcating) rooted tree is defined as the sum of the depths of its leaves:

$$\mathcal{S}(T) = \sum_{x \in V_L(T)} \delta_T(x).$$

Equivalently [4], it is equal to the sum of the numbers of descendant leaves of the internal nodes of T:

$$\mathcal{S}(T) = \sum_{v \in \mathring{V}(T)} \kappa_T(v).$$

The bifurcating trees with n leaves that achieve the maximum Sackin index are exactly the caterpillars [12, 27]. As to those achieving its minimum value, they have been recently characterized by Fischer [12] and in particular they include the fully symmetric trees (cf. Theorem 5 therein). We shall generalize this result by showing that they actually include all minimal Colless trees. We shall use from Fischer's paper the following result (cf. Corollary 4 therein):

Lemma 6. *Let $T = (T_a, T_b)$ be a bifurcating tree with $n \in \mathbb{N}_{\geqslant 2}$ leaves and let $k_n = \lceil \log_2(n) \rceil$. Then, T has minimal Sackin index if, and only if, T_a and T_b have minimal Sackin index and $n_a - n_b \leqslant \min\{n - 2^{k_n - 1}, 2^{k_n} - n\}$.*

Based on this lemma we can prove the following statement.

Proposition 8. *For every $n \geqslant 1$, if $T \in \mathcal{T}_n$ is a minimal Colless tree, then it has minimum Sackin index.*

Proof. We show the statement by induction on n. For $n = 1$, it is, as always, obvious because there is only one tree in $\mathcal{T}_1$. Assume now that the claim holds for every $1 \leqslant n' < n$ and let $T = (T_a, T_b) \in \mathcal{T}_n$ be a minimal Colless tree with n leaves, with $T_a \in \mathcal{T}_{n_a}$ and $T_b \in \mathcal{T}_{n_b}$. Write $n = 2^m + p$, with $m = \lfloor \log_2(n) \rfloor$ and $0 \leqslant p < 2^m$. If $p = 0$, there is only one minimal Colless tree, which is fully symmetric and therefore it has minimum Sackin index. So, we assume that $p > 0$, in which case $k_n = \lceil \log_2(n) \rceil = m + 1$. By Lemma 2, both T_a and T_b are minimal Colless trees and therefore, by the induction hypothesis, they have minimum Sackin index. Thus, by Lemma 6, to prove that T has minimum Colless index it is enough to prove that

$$n_a - n_b \leqslant \min\{n - 2^{k_n - 1}, 2^{k_n} - n\} = \{n - 2^m, 2^{m+1} - n\} = \{p, 2^m - p\}.$$

But this has already been proved in the proof of Proposition 7. □

The converse implication is not true. For example, the tree T_2 depicted in Figure 7 has minimum Sackin index, but it does not have minimum Colless index.

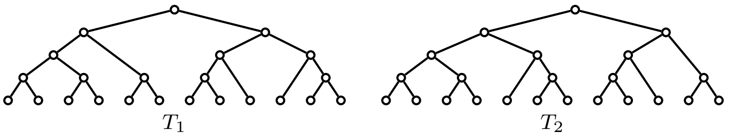

Figure 7: Trees T_1 and T_2 with 12 leaves. We have $\mathcal{C}(T_1) = 4 = c_{12}$ and $\mathcal{C}(T_2) = 6$. Thus, T_1 has minimum Colless index, while T_2 does not. Note, however, that their Sackin indices are $\mathcal{S}(T_1) = \mathcal{S}(T_2) = 44$, which can be shown to be minimal (cf. [12, Thm. 3]).

5. Discussion

The Colless index $\mathcal{C}(T)$ of a rooted bifurcating phylogenetic tree T is a measure of the total imbalance of T, and it is one of the oldest and most popular balance indices for 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 achieved were known so far. This paper fills this gap in the literature, with two main contributions.

First, we have established both a recursive and two different closed expressions for the minimum value c_n of the Colless index on the space $\mathcal{T}_n$ of bifurcating trees with n leaves. Knowing this minimum value, as well as its maximum value, which is reached at the caterpillars and is equal to $\binom{n-1}{2}$, allows one to normalize the Colless index so that its range becomes the unit interval $[0, 1]$, by means of the usual affine transformation

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

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 $\mathcal{C}$ because its value tends to grow with n.

Our expressions for c_n have been obtained by first proving that the *maximally balanced* trees are *minimal Colless*, i.e. they have minimum Colless index for their number of leaves. This result is not surprising, because, in words of Shao and Sokal [27], they are considered to be the "most balanced" bifurcating trees. But it turns out that for almost all values of n there are minimal Colless trees that are not maximally balanced. So, our second main contribution has been an alternative characterization of the minimal Colless trees, an efficient algorithm to produce all of them for any number n of leaves, and a recurrence to count them for every n. Unfortunately, we have not been able so far to find a closed expression for the number $\widetilde{c}(n)$ of different minimal Colless trees with n leaves.

Moreover, we have described a second family of minimal Colless trees, that we have called *greedy from the bottom*, *GFB*, with a member in every space $\mathcal{T}_n$. These GFB trees are different from the maximally balanced trees for numbers of leaves that differ at least 2 from any power of 2. 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 GFB trees are also Colless minimal. So, in general, the total imbalance of a phylogenetic tree does not capture the local imbalance at each internal node.

We would like to point out that one of our expressions for c_n entails a fractal structure for the graph of (n, c_n) related to the fractal Blancmange curve (cf. Figure 3). It turns out that a similar fractal structure seems to appear also in the graph of $(n, \widetilde{c}(n))$ (cf. Figure 6). It would definitely be of interest to find an explicit formula for $\widetilde{c}(n)$ and to analyze whether this seemingly fractal structure is real or not and its possible relationship with that of the sequence (n, c_n).

We have concluded by showing that every Colless minimal tree also has minimum Sackin index, while the converse is not true. This implies that the Sackin index classifies more trees as "most balanced" than the Colless index. The Colless index, on the other hand, considers more trees as "most balanced" than for example the so-called *total cophenetic index* [21], for which the minimum value is uniquely achieved by the maximally balanced tree on n leaves.

Acknowledgements. Tomás M. Coronado and Francesc Rosselló thank the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund for partial support for this research through projects DPI2015-67082-P and PGC2018-096956-B-C43 (MINECO/FEDER). Lina Herbst thanks the state Mecklenburg-Western Pomerania for a Landesgraduierten-Studentship and Kristina Wicke thanks the German Academic Scholarship Foundation for a studentship.

References

[1] Aldous D (1996) Probability distributions on cladograms. In: Aldous D, Pemantle R (eds) Random Discrete Structures. The IMA Volumes in Mathematics and its Applications, vol 76. Springer, New York, pp 1–18.

[2] Allaart PC, Kawamura K (2012) The Takagi function: a survey. Real Analysis Exchange 37:1–54.

[3] Avino M, Garway TN, et al (2018) Tree shape-based approaches for the comparative study of cophylogeny. bioRxiv DOI 10.1101/388116.

[4] Blum MG, François O (2005) On statistical tests of phylogenetic tree imbalance: The Sackin and other indices revisited. Mathematical Biosciences 195:141–153.

[5] Blum MGB, François O, Janson S (2006) The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. Annals of Applied Probability 16:2195–2214.

[6] Cardona G, Mir A, Rosselló F (2013) Exact formulas for the variance of several balance indices under the Yule model. Journal of Mathematical Biology 67:1833–1846.

[7] Chalmandrier L, Albouy C, et al (2018) Comparing spatial diversification and meta-population models in the Indo-Australian Archipelago. Royal Society Open Science 5:171366.

[8] Colless D (1982) Review of "Phylogenetics: the theory and practice of phylogenetic systematics". Systematic Zoology 31:100–104.

[9] Coronado TM, Mir A, Rosselló F, Valiente G (2019) A balance index for phylogenetic trees based on rooted quartets. Journal of Mathematical Biology (to appear), DOI 10.1007/s00285-019-01377-w (Epub ahead of print).

[10] Duchene S, Bouckaert R, Duchene DA, Stadler T, Drummond AJ (2018) Phylodynamic model adequacy using posterior predictive simulations. Systematic Biology 68:358–364.

[11] Felsenstein J (2004) Inferring Phylogenies. Oxford University Press.

[12] Fischer M (2018) Extremal values of the Sackin balance index for rooted binary trees. arXiv preprint arXiv:1801.10418v3.

[13] Fischer M, Liebscher V (2015) On the Balance of Unrooted Trees. arXiv preprint arXiv:1510.07882.

[14] Ford DJ (2005) Probabilities on cladograms: introduction to the alpha model. PhD thesis, Stanford University. arXiv preprint arXiv:math/0511246.

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

[16] Heard SB (1992) Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. Evolution 46:1818–1826.

[17] Kingman JFC (1982) The coalescent. Stochastic processes and their applications 13:235–248.

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

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

[20] Metzig C, Ratmann O, Bezemer D, Colijn C (2019) Phylogenies from dynamic networks. PLoS Computational Biology 15:e1006761.

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

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

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

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

[25] Rogers JS (1993) Response of Colless's tree imbalance to number of terminal taxa. Systematic Biology 42:102.

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

[27] Shao K, Sokal R (1990) Tree balance. Systematic Zoology 39:266–276.

[28] Sloane NJA (1964) The On-Line Encyclopedia of Integer Sequences (OEIS). `http://oeis.org`. Last accessed, July 8, 2019.

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

[30] Stich M, Manrubia SC (2009) Topological properties of phylogenetic trees in evolutionary models. The European Physical Journal B 70:583–592.

[31] Takagi T (1901) A simple example of continuous function without derivative. Tokyo Sugaku-Butsurigakkwai Hokoku 1:F176–F177.

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

Appendices

A.1 Proof of Proposition 2

This appendix is devoted to establish the following result.

Proposition 2. *For every $n \geqslant 2$ and for every $n_a, n_b \in \mathbb{N}_{\geqslant 1}$ such that $n_a \geqslant n_b$ and $n_a + n_b = n$:*

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

(2) If $n_a > n_b$, then $(n_a, n_b) \in QB(n)$ if, and only if, one of the following three conditions is satisfied:

- *There exist $k \in \mathbb{N}$ and $p \in \mathbb{N}_{\geqslant 1}$ such that $n = 2^k(2p+1)$, $n_a = 2^k(p+1)$ and $n_b = 2^k p$.*
- *There exist $k \in \mathbb{N}$, $l \in \mathbb{N}_{\geqslant 2}$, $p \in \mathbb{N}_{\geqslant 1}$, and $t \in \mathbb{N}$, $0 \leqslant t < 2^{l-2}$, such that $n = 2^k(2^l(2p+1)+2t+1)$, $n_a = 2^{k+l}(p+1)$, and $n_b = 2^k(2^l p + 2t + 1)$.*
- *There exist $k \in \mathbb{N}$, $l \in \mathbb{N}_{\geqslant 2}$, $p \in \mathbb{N}_{\geqslant 1}$, and $t \in \mathbb{N}$, $0 \leqslant t < 2^{l-2}$, such that $n = 2^k(2^l(2p+1)-(2t+1))$, $n_a = 2^k(2^l(p+1) - (2t+1))$, and $n_b = 2^{k+l}p$.*

The proof of this proposition relies on several auxiliary lemmas. In order to simplify the language in their statements and proofs, throughout this section we systematically assume, without any further notice, that the symbols j, k, m, n, p, s, t, and x, possibly with subscripts or superscripts, always represent natural numbers.

Remark 4. Note that in the proof of Lemma 3 we have established the following two facts (for $\mathcal{C}(T_n^{mb})$), which we state here for further reference:

(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 m is odd, then $c_{m+s} + c_m + s > c_{2m+s}$.

Lemma 7. *Let $s = 2^k s_0$ with $k \geqslant 1$ and $s_0 \geqslant 1$. Then, for every $m \geqslant 1$, $(m+s, m) \in QB(2m+s)$ if, and only if, $m = 2^k m_0$, for some $m_0 \geqslant 1$ such that $(m_0 + s_0, m_0) \in QB(2m_0 + s_0)$.*

Proof. We prove the equivalence in the statement by induction on the exponent $k \geqslant 1$. Recall that, by Remark 4.(b), if $s \geqslant 1$ is even and $c_{m+s} + c_m + s = c_{2m+s}$, then m must be even, too. Therefore, if $s = 2t_0$, then $m = 2m_1$ for some $m_1 \geqslant 1$, and then, since

$$c_{2m_1+2t_0} + c_{2m_1} + 2t_0 = 2(c_{m_1+t_0} + c_{m_1} + t_0)$$

and $c_{4m_1+2t_0} = 2c_{2m_1+t_0}$, the equality $c_{m+s}+c_m+s = c_{2m+s}$ is equivalent to the equality $c_{m_1+t_0}+c_{m_1}+t_0 = c_{2m_1+t_0}$. This proves the equivalence in the statement when $k = 1$.

Now, assume that this equivalence is true for the exponent $k-1$, and let $s = 2^k s_0$. Then, by the case $k = 1$, $c_{m+s} + c_m + s = c_{2m+s}$ if, and only if, $m = 2m_1$ for some $m_1 \geqslant 1$ such that

$$c_{m_1+2^{k-1}s_0} + c_{m_1} + 2^{k-1}s_0 = c_{2m_1+2^{k-1}s_0},$$

and, by the induction hypothesis, this last equality holds if, and only if, $m_1 = 2^{k-1}m_0$ for some $m_0 \geqslant 1$ such that $c_{m_0+s_0} + c_{m_0} + s_0 = c_{2m_0+s_0}$. Combining both equivalences we obtain the equivalence in the statement, thus proving the inductive step. □

Lemma 8. *Let $s = 2^{j+1} - (2t+1)$ be an odd integer, with $j = \lfloor \log_2(s) \rfloor$ and $0 \leqslant t < 2^{j-1}$. Then, for every $m \geqslant 1$, $(2m+s, 2m) \in QB(4m+s)$ if, and only if, $m = 2^j p$ for some $p \geqslant 1$.*

Proof. We prove the equivalence in the statement by induction on s. When $s = 1 = 2^1 - 1$, so that $j = t = 0$, the equivalence says that

$$c_{2m+1} + c_{2m} + 1 = c_{4m+1}$$

for every $m \geqslant 1$, which is true by Corollary 2.

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

$$\begin{aligned} &c_{2m+2^{j+1}-2t-1} + c_{2m} + 2^{j+1} - 2t - 1 \\ &\qquad = \big(c_{m+2^j-t} + c_m + 2^j - t\big) + \big(c_{m+2^j-t-1} + c_m + 2^j - t - 1\big) + 1 \\ &c_{4m+2^{j+1}-2t-1} = c_{2m+2^j-t} + c_{2m+2^j-t-1} + 1 \end{aligned}$$

and since, by Eqn. (2), $c_{m+2^j-t}+c_m+2^j-t \geqslant c_{2m+2^j-t}$ and $c_{m+2^j-t-1}+c_m+2^j-t-1 \geqslant c_{2m+2^j-t-1}$, we have that $c_{2m+s}+c_{2m}+s=c_{4m+s}$ if, and only if, the following two identities are satisfied:

$$c_{m+2^j-t}+c_m+2^j-t = c_{2m+2^j-t} \tag{10}$$

$$c_{m+2^j-t-1}+c_m+2^j-t-1 = c_{2m+2^j-t-1} \tag{11}$$

So, we must prove that Eqns. (10) and (11) hold if, and only if, $m=2^j p$ for some $p \geqslant 1$. We distinguish two subcases, depending on the parity of t:

- If $t=2x$ for some $0 \leqslant x < 2^{j-2}$, then Eqn. (10) and Lemma 7 imply that m is even, say $m=2m_0$, and then (11) says
$$c_{2m_0+2^j-2x-1}+c_{2m_0}+2^j-2x-1 = c_{4m_0+2^j-2x-1}, \tag{12}$$
which, by induction, is equivalent to $m_0=2^{j-1}p$ for some $p \geqslant 1$, i.e. to $m=2^j p$ for some $p \geqslant 1$. So, to complete the proof of the desired equivalence, it remains to prove that if $m=2^j p$, then Eqn. (10) holds. If $t=0$, this equality says
$$c_{2^j p+2^j}+c_{2^j p}+2^j = c_{2^{j+1}p+2^j}$$
and it is a direct consequence of Lemma 7 and Corollary 2. So, assume that $t>0$ and write it as $t=2^i(2x_0+1)$ with $1 \leqslant i < j-1$ and $x_0 < 2^{j-i-2}$. Then
$$\begin{aligned} & c_{m+2^j-t}+c_m+2^j-t \\ & \quad = c_{2^j p+2^j-2^i(2x_0+1)}+c_{2^j p}+2^j-2^i(2x_0+1) \\ & \quad = 2^i\big(c_{2^{j-i}p+2^{j-i}-2x_0-1}+c_{2^{j-i}p}+2^{j-i}-2x_0-1\big) \\ & \quad = 2^i c_{2^{j-i+1}p+2^{j-i}-2x_0-1} \text{ (by the induction hypothesis)} \\ & \quad = c_{2^{j+1}p+2^j-2^i(2x_0+1)} = c_{2m+2^j-t}. \end{aligned}$$

- If $t=2x+1$ for some $0 \leqslant x < 2^{j-2}$, then Eqn. (11) and Lemma 7 imply that m is even, say $m=2m_0$, and then it is Eqn. (10) which becomes Eqn. (12) above, which, in turn, by induction is equivalent to $m_0=2^{j-1}p$ for some $p \geqslant 1$, that is, to $m=2^j p$ for some $p \geqslant 1$. Thus, to complete the proof of the desired equivalence, it remains to prove that if $m=2^j p$, then (11) holds. Now:
$$\begin{aligned} & c_{m+2^j-t-1}+c_m+2^j-t-1 \\ & \quad = c_{2^j p+2^j-2x-2}+c_{2^j p}+2^j-2x-2 \\ & \quad = 2\big(c_{2^{j-1}p+2^{j-1}-x-1}+c_{2^{j-1}p}+2^{j-1}-x-1\big) \end{aligned}$$

  If x is even, say $x=2x_0$, then, since $x_0 < 2^{j-3}$, the induction hypothesis implies that
$$\begin{aligned} & 2\big(c_{2^{j-1}p+2^{j-1}-x-1}+c_{2^{j-1}p}+2^{j-1}-x-1\big) \\ & \quad = 2c_{2^j p+2^{j-1}-x-1} = c_{2^{j+1}p+2^j-2x-2} = c_{2m+2^j-t-1}. \end{aligned}$$

  And if x is odd, write it as $x=2^i(2t_0+1)-1$ for some $1 \leqslant i < j-1$ (and notice that $x < 2^{j-2}$ implies $t_0 < 2^{j-i-3}$) and then
$$\begin{aligned} & 2\big(c_{2^{j-1}p+2^{j-1}-x-1}+c_{2^{j-1}p}+2^{j-1}-x-1\big) \\ & \quad = 2\big(c_{2^{j-1}p+2^{j-1}-2^i(2t_0+1)}+c_{2^{j-1}p}+2^{j-1}-2^i(2t_0+1)\big) \\ & \quad = 2\cdot 2^i\big(c_{2^{j-i-1}p+2^{j-i-1}-(2t_0+1)}+c_{2^{j-i-1}p}+2^{j-i-1}-(2t_0+1)\big) \\ & \quad = 2^{i+1}c_{2^{j-i}p+2^{j-i-1}-(2t_0+1)} \text{ (by the induction hypothesis)} \\ & \quad = c_{2^{j+1}p+2^j-2^{i+1}(2t_0+1)} = c_{2^{j+1}p+2^j-2x-2} \\ & \quad = c_{2m+2^j-t-1} \end{aligned}$$

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

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

Lemma 9. *Let $s = 2^{j+1} - (2t+1)$ be an odd integer, with $j = \lfloor \log_2(s) \rfloor$ and $0 \leqslant t < 2^{j-1}$. Then, for every $m \geqslant 0$, $(2m+1+s, 2m+1) \in QB(4m+2+s)$ if, and only if, either $m = 2^j p + t$ for some $p \geqslant 1$ or $s = 1$ (i.e. $j = t = 0$) and $m = 0$.*

Proof. We prove also 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 Corollary 2.

Assume now that the equivalence is true for every odd natural number $1 \leqslant s' < s$ and for every $m \geqslant 0$, and let us prove it for $s = 2^{j+1} - (2t+1) \geqslant 3$ with $0 \leqslant t < 2^{j-1}$. In this case, m cannot be 0, because, by Remark 4.(a), $(s+1, s) \in QB(s+2)$ if, and only if, $s = 1$. So, we can consider only the case $m \geqslant 1$. Then, we have that

$$\begin{aligned} &c_{2m+1+2^{j+1}-2t-1} + c_{2m+1} + 2^{j+1} - 2t - 1 \\ &\qquad = \big(c_{m+2^j-t} + c_m + 2^j - t\big) + \big(c_{m+2^j-t} + c_{m+1} + 2^j - t - 1\big) + 1 \\ &c_{4m+2+2^{j+1}-2t-1} = c_{2m+2^j-t} + c_{2m+2^j-t+1} + 1 \end{aligned}$$

and since, by Eqn. (2), $c_{m+2^j-t} + c_m + 2^j - t \geqslant c_{2m+2^j-t}$ and $c_{m+2^j-t} + c_{m+1} + 2^j - t - 1 \geqslant c_{2m+2^j-t+1}$, we have that $c_{2m+1+s} + c_{2m+1} + s = c_{4m+2+s}$ if, and only if,

$$c_{m+2^j-t} + c_m + 2^j - t = c_{2m+2^j-t} \tag{13}$$

$$c_{m+2^j-t} + c_{m+1} + 2^j - t - 1 = c_{2m+2^j-t+1} \tag{14}$$

So, we must prove that Eqns. (13) and (14) hold for $m \geqslant 1$ if, and only if, $m = 2^j p + t$ for some $p \geqslant 1$. We distinguish again two subcases, depending on the parity of t:

- If $t = 2x$ for some $0 \leqslant x < 2^{j-2}$, then Eqn. (13) and Lemma 7 imply that m is even, say $m = 2m_0$ with $m_0 \geqslant 1$, and then Eqn. (14) can be written

  $$c_{2m_0+1+2^j-2x-1} + c_{2m_0+1} + 2^j - 2x - 1 = c_{4m_0+2+2^j-2x-1}$$

  which, by induction, is equivalent to $m_0 = 2^{j-1}p + x$ for some $p \geqslant 1$, that is, to $m = 2m_0 = 2^j p + t$ for some $p \geqslant 1$. So, to complete the proof of the desired equivalence, it remains to check that if $m = 2^j p + t$, then Eqn. (13) holds. Now, if $x = 0$, so that $m = 2^j p$, Corollary 2 and Lemma 7 clearly imply Eqn. (13) (cf. the case when t is even in the proof of Lemma 8). So, assume that $x > 0$ and write it as $x = 2^i(2y_0+1)$ with $0 \leqslant i < j-2$ and $y_0 < 2^{j-i-3}$. Then

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

  as we wanted to prove.

- If $t = 2x+1$ for some $0 \leqslant x < 2^{j-2}$, Eqn. (14) and Lemma 7 imply that $m+1$ is even, and then m is odd, say $m = 2m_0 + 1$ for some $m_0 \geqslant 0$, and Eqn. (13) can be written

  $$c_{2m_0+1+2^j-2x-1} + c_{2m_0+1} + 2^j - 2x - 1 = c_{4m_0+2+2^j-2x-1}. \tag{15}$$

  Now, if $m_0 = 0$, Remark 4.(a) implies that this equality holds if, and only if, $2^j - 2x - 1 = 1$ which, under the condition $0 \leqslant x < 2^{j-2}$, only happens when $j = 1$ and $x = 0$, but then $t = 1 = 2^{j-1}$ against the assumption that $t < 2^{j-1}$. Therefore m_0 must be at least 1.

Then, by induction, Identity (15) is equivalent to $m_0 = 2^{j-1}p + x$ for some $p \geqslant 1$, that is, to $m = 2m_0 + 1 = 2^j p + 2x + 1 = 2^j p + t$ for some $p \geqslant 1$. So, to complete the proof of the desired equivalence, it remains to check that if $m = 2^j p + t$, then Eqn. (14) holds. Now, in the current situation:

$$\begin{aligned}
&c_{m+2^j-t} + c_{m+1} + 2^j - t - 1\\
&\qquad = c_{2^j p+2x+1+2^j-2x-1} + c_{2^j p+2x+2} + 2^j - 2x - 2\\
&\qquad = c_{2^j p+2^j} + c_{2^j p+2x+2} + 2^j - 2x - 2\\
&\qquad = 2\big(c_{2^{j-1}p+2^{j-1}} + c_{2^{j-1}p+x+1} + 2^{j-1} - x - 1\big)\\
&\qquad = 2\big(c_{(2^{j-1}p+x+1)+(2^{j-1}-x-1)} + c_{2^{j-1}p+x+1} + 2^{j-1} - x - 1\big) = (**)
\end{aligned}$$

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

$$\begin{aligned}
(**) &= 2\big(c_{(2^{j-1}p+2x_0+1)+(2^{j-1}-2x_0-1)} + c_{2^{j-1}p+2x_0+1} + 2^{j-1} - 2x_0 - 1\big)\\
&= 2c_{2^j p+2(2x_0+1)+2^{j-1}-(2x_0+1)} \qquad \text{(by the induction hypothesis)}\\
&= c_{2^{j+1}p+2^j+4x_0+2} = c_{2m+2^j-t+1}.
\end{aligned}$$

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

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

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

□

We are now in a position to proceed with the proof of Proposition 2. Assertion (1) in it is a direct consequence of Corollary 2. So, assume $n_a > n_b$ and set $s = n_a - n_b$, so that $n_a = n_b + s$. Then:

(a) If $s = 1$, then, by Lemma 9, $c_{n_a} + c_{n_b} + n_a - n_b = c_{n_a+n_b}$ for every $n_b \geqslant 1$.

(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 8 and 9, $c_{n_a} + c_{n_b} + n_a - n_b = c_{n_a+n_b}$ if, and only if, either $n_b = 2^{j+1}p$ or $n_b = 2^{j+1}p + 2t + 1$, 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, and write the latter as $s_0 = 2^{j+1} - (2t + 1)$ with $j = \lfloor \log_2(s_0) \rfloor \geqslant 0$ and $0 \leqslant t < 2^{j-1}$. Then, by Lemma 7, $c_{n_a} + c_{n_b} + n_a - n_b = c_{n_a+n_b}$ if, and only if, $n_b = 2^k m$, for some $m \geqslant 1$ such that $c_{m+s_0} + c_m + s_0 = c_{2m+s_0}$, and then:

    - If $s_0 = 1$ (equivalently, if $j = 0$), $c_{m+s_0} + c_m + s_0 = c_{2m+s_0}$ for every $m \geqslant 1$ and therefore, in this case, $c_{n_a} + c_{n_b} + n_a - n_b = c_{n_a+n_b}$ for every $n_b = 2^k m$ with $m \geqslant 1$.
    - If $s_0 > 1$ (equivalently, if $j > 0$), Lemmas 8 and 9 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, $c_{n_a} + c_{n_b} + n_a - n_b = c_{n_a+n_b}$ if, and only if, $n_b = 2^{k+j+1}p$ or $n_b = 2^k(2^{j+1}p + 2t + 1)$, for some $p \geqslant 1$.

Combining the three cases, and taking $k = 0$ in the odd s case, we conclude that

$$c_{n_a} + c_{n_b} + n_a - n_b = c_{n_a+n_b}$$

if, and only if, writing $n_a - n_b = 2^k(2^{j+1} - (2t+1))$ (for some $k \geqslant 0$, $j \geqslant 0$, and $0 \leqslant t < 2^{j-1}$),

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

This is equivalent to the expressions for n_a and n_b in option (2) in the statement (replacing $j+1$ with $j > 0$ by $l \geqslant 2$).

This completes the proof of Proposition 2.

A. 2 Proof of Proposition 5

This appendix is devoted to establish the following result.

Proposition 5. *Let $T_n^{gfb} = (T_a, T_b)$ be a GFB tree with $n \geqslant 2$, $T_a \in \mathcal{T}_{n_a}$, $T_b \in \mathcal{T}_{n_b}$ and $n_a \geqslant n_b$. Let $n = 2^m + p$ with $m = \lfloor \log_2(n) \rfloor$ and $0 \leqslant p < 2^m$. Then, we have:*

(i) *If $0 \leqslant p \leqslant 2^{m-1}$, then $n_a = 2^{m-1} + p$, $n_b = 2^{m-1}$ and T_b is fully symmetric.*

(ii) *If $2^{m-1} \leqslant p < 2^m$, $n_a = 2^m$, $n_b = p$ and T_a is fully symmetric.*

The proof of this proposition requires of the following lemma. The leading idea of its proof is illustrated in Figure 8.

Lemma 10. *Let $n \geqslant 3$ be an odd natural number. Then, T_n^{gfb} shares a maximal pending subtree with T_{n-1}^{gfb} and a maximal pending subtree with T_{n+1}^{gfb}.*

Proof. Since $n \geqslant 3$ is odd, the first $(n-1)/2$ iterations of the loop in Algorithm 2 result in $(n-1)/2$ cherries and a single node, which in the $(n+1)/2$-th iteration is clustered with a cherry to form a tree with 3 leaves. From this moment on, as the algorithm continues clustering trees, in each i-th iteration there will be one, and only one, tree T_i^{odd} with an odd number $s(i)$ of leaves. Note now that, on the one hand, this unique tree with $s(i)$ leaves is treated by the algorithm like a tree with $s(i) - 1$ leaves, except that it is clustered as late as possible, i.e. when all other trees in *treeset* with $s(i) - 1$ leaves (if there are any) have already been clustered. On the other hand, however, this tree is also treated by the algorithm like a tree with $s(i) + 1$ leaves, except that it is clustered as early as possible, i.e. before any other elements in *treeset* with $s(i) + 1$ leaves (if there are any) get clustered. So, to summarize, after the first $i \geqslant (n+1)/2$ iterations of the loop, *treeset* contains a unique tree T_i^{odd} with an odd number $s(i)$ of leaves, which at the same time

(i) is treated like a tree with $s(i) - 1$ leaves, but is clustered as late as possible;

(ii) is treated like a tree with $s(i) + 1$ leaves, but is clustered as soon as possible.

Now, first consider Algorithm 2 for $n - 1$, which is an even number. After the first $(n-3)/2$ iterations of the loop, *treeset* contains $(n-3)/2$ trees with 2 leaves and two trees with 1 leaf, which are clustered last to form the last cherry. We keep tracking one leaf u of this cherry throughout the algorithm. The algorithm at this stage contains only cherries, which are all isomorphic, so without loss of generality, we may assume that u is contained in the one that gets clustered with another tree last, i.e. after all other cherries have been clustered. We continue like this, always assuming without loss of generality (when there is more than one tree in *treeset* of the same size as the tree that contains u) that the tree containing u is in the last one to be clustered. By (i), this means that if we replace u in T_{n-1}^{gfb} by a cherry, we derive T_n^{gfb}. This is due to

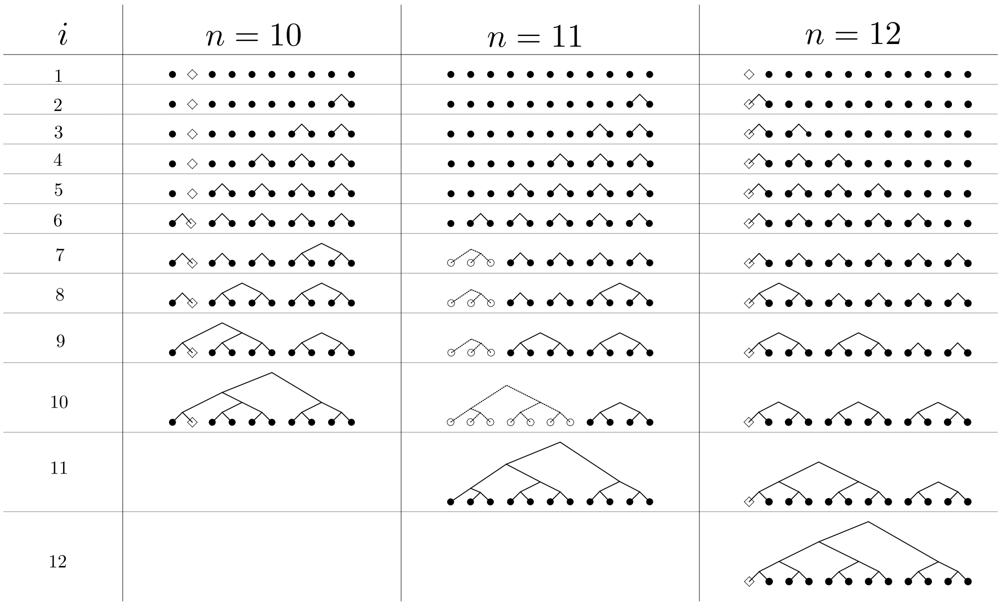

Figure 8: Content of *treeset* before the i^{th} iteration of the loop in Algorithm 2 for $n = 10, n = 11$ and $n = 12$. In case of $n = 11$, the tree depicted in dashed lines for $i = 7, \ldots, 10$, depicts the *unique* tree in *treeset* with an odd number of leaves. For $n = 10$, the leaf depicted as a diamond represents leaf u used in the proof of Lemma 10. Note that the tree containing this leaf is always clustered as late as possible. In case of $n = 12$, the leaf depicted as a diamond again represents leaf u used in the proof of Lemma 10. In this case, the tree containing this leaf is always clustered as soon as possible. The last tree depicted in each column represents the GFB tree. Note that T_n^{gfb} can be obtained from T_{n-1}^{gfb} by replacing the leaf depicted as a diamond by a cherry. Moreover, T_n^{gfb} can be obtained from T_{n+1}^{gfb} by replacing the cherry containing the diamond leaf by a single leaf.

the fact that in the analogous step where *treeset* for $n - 1$ only contains cherries, *treeset* for n will contain only cherries and a tree containing three leaves. This triplet will subsequently act like a cherry, but like the one that happens to be clustered last. So we identify the cherry of the triplet with leaf u of this last cherry to see the correspondence between T_{n-1}^{gfb} and T_n^{gfb}. Note that this directly implies that T_{n-1}^{gfb} and T_n^{gfb} share a common maximal pending subtree —namely the one that does *not* contain u.

Note that by (ii), an analogous procedure for $n + 1$ leads to T_{n+1}^{gfb} and T_n^{gfb} sharing a common maximal pending subtree. In this case, we track a cherry in T_{n+1}^{gfb}, namely the one that happens to be clustered first, and replace it by a single leaf to see the correspondence between T_{n+1}^{gfb} and T_n^{gfb}. This completes the proof. □

We can proceed with the proof of Proposition 5. Let $n = 2^m + p$ with $m = \lfloor \log_2(n) \rfloor$ and $0 \leqslant p < 2^m$. We shall prove by induction on n that if $T_n^{gfb} = (T_a, T_b)$ is a GFB tree with $n \geqslant 2$, $T_a \in \mathcal{T}_{n_a}$, $T_b \in \mathcal{T}_{n_b}$ and $n_a \geqslant n_b$ then:

(i) If $0 \leqslant p \leqslant 2^{m-1}$, $n_a = 2^{m-1} + p$ and $n_b = 2^{m-1}$ and then T_b is fully symmetric.

(ii) If $2^{m-1} \leqslant p < 2^m$, we have $n_a = 2^m$ and $n_b = p$ and then T_a is fully symmetric.

We want to point out that we understand that the conjunction of these two assertions in the case when both premises are satisfied, namely when $p = 2^{m-1}$, says that $n_a = 2^m$ and $n_b = 2^{m-1}$ and then both T_a (by (ii)) and T_b (by (i)) are fully symmetric.

The base case for (i) is when $n = 2$ and for (ii), when $n = 3$. In both cases the assertions are obvious, because there is only one bifurcating tree with $2 = 2^1 + 0$ leaves (a cherry with $n_a = n_b = 1 = 2^0$) and only one bifurcating tree with $3 = 2^1 + 1$ leaves (a caterpillar with $n_a = 2 = 2^1$ and $n_b = 1$).

Now, let $n \geqslant 4$ and assume that (i) and (ii) hold for up to $n-1$ leaves. Let $T = (T_a, T_b)$ be a GFB tree with n leaves, with $T_a \in \mathcal{T}_{n_a}$, $T_b \in \mathcal{T}_{n_b}$ and $n_a \geqslant n_b$ Recall that T_a and T_b are again GFB trees by Lemma 5. We distinguish two cases, depending on the parity of n:

- Assume that n is even, say $n = 2n_0$ with $n_0 \geqslant 2$. In this case, Algorithm 2 results in a tree T_n^{gfb} with n_0 cherries (because in each of the first n_0 iterations of the loop a pair of nodes are merged into a cherry). We now consider the tree T' with n_0 leaves that is obtained from T_n^{gfb} by replacing all cherries by single leaves. Let $T' = (T'_a, T'_b)$ be the decomposition into maximal pending subtrees, with $T'_a \in \mathcal{T}_{n'_a}$, $T'_b \in \mathcal{T}_{n'_b}$ and $n'_a \geqslant n'_b$. By construction, T_a and T_b are obtained by replacing the leaves of T'_a and T'_b by cherries, and therefore, in particular, $n_a = 2n'_a$ and $n_b = 2n'_b$. Note now that, since T_n^{gfb} is a GFB tree, so is T' (because as soon as Algorithm 2 only has cherries to choose from, they are treated like leaves). Note also that, since n is even, so is p, say $p = 2p_0$, and $n_0 = 2^{m-1} + p_0$. Then we have that:

  (i) If $0 \leqslant p \leqslant 2^{m-1}$, then $0 \leqslant p_0 \leqslant 2^{m-1-1}$ and hence, by the induction hypothesis, $n'_a = 2^{m-2} + p_0$, $n'_b = 2^{m-2}$, and T'_b is fully symmetric, which implies that $n_a = 2n'_a = 2^{m-1} + 2p_0 = 2^{m-1} + p$, $n_b = 2n'_b = 2^{m-1}$, and T_b is fully symmetric, because it is obtained from the fully symmetric tree T'_b by replacing all its leaves by cherries.

  (ii) If $2^{m-1} \leqslant p < 2^m$, then $2^{m-1-1} \leqslant p_0 \leqslant 2^{m-1}$ and hence, by the induction hypothesis, $n'_a = 2^{m-1}$, $n'_b = p_0$, and T'_a is fully symmetric, which implies that $n_a = 2n'_a = 2^m$, $n_b = 2n'_b = 2p_0 = p$, and, arguing as in (i), T_a is fully symmetric.

- Assume that n is odd, say $n = 2n_0 + 1$ with $n_0 \geqslant 2$. In this case both $n-1 = 2n_0$ and $n+1 = 2(n_0+1)$ are even. Write $n = 2^m + p$ and $p = 2p_0 + 1$, so that $n_0 = 2^{m-1} + p_0$ with $0 \leqslant p_0 < 2^{m-1}$. Let $T^1 := T_{n-1}^{gfb}$ and $T^2 := T_{n+1}^{gfb}$. The tree T^1 satisfies (i) and (ii) by the induction hypothesis, and it can be proved that T^2 also satisfies these assertions by arguing as in the previous case when n is even (i.e., replacing the pending $n_0 + 1$ cherries in T^2 by single leaves, noticing that the resulting tree is GFB, applying the induction hypothesis to it and finally returning back to T^2 by replacing leaves by cherries). Let $T^1 = (T_a^1, T_b^1)$ —with $T_a^1 \in \mathcal{T}_{n_a^1}$ and $T_b^1 \in \mathcal{T}_{n_b^1}$ and $n_a^1 \geqslant n_b^1$— and $T^2 = (T_a^2, T_b^2)$ —with $T_a^2 \in \mathcal{T}_{n_a^2}$ and $T_b^2 \in \mathcal{T}_{n_b^2}$ and $n_a^2 \geqslant n_b^2$— denote the decompositions of T^1 and T^2 into maximal pending subtrees, respectively. Note that, since n is odd, $p \neq 0, 2^{m-1}$. Now we have:

  (i) If $0 < p < 2^{m-1}$, then $n-1 = 2^m + (p-1)$ with $0 \leqslant p-1 < 2^{m-1}$ and $n+1 = 2^m + (p+1)$ with $0 < p+1 \leqslant 2^{m-1}$. Then, since T^1 and T^2 satisfy assertion (i),

  $$n_a^1 = 2^{m-1} + p - 1,\ n_b^1 = 2^{m-1},\ n_a^2 = 2^{m-1} + p + 1,\ n_b^2 = 2^{m-1}$$

  and both T_b^1 and T_b^2 are fully symmetric and hence (since they have the same numbers of leaves) $T_b^1 = T_b^2$.

  Now, we know by Lemma 10 that T shares a maximal pending subtree with T^1 and a maximal pending subtree with T^2. Looking at the numbers of leaves of the maximal pending subtrees of T^1 and T^2, one easily deduces that the only possibility for this to happen is that T shares with T^1 and T^2 the same maximal pending subtree: the fully symmetric subtree $T_b^1 = T_b^2$. (Indeed, since $T_a^1 \neq T_a^2$, because they have different numbers of leaves, if T did not share $T_b^1 = T_b^2$ with both T^1 and T^2, then it would have a maximal pending subtree in common with T^1 and the other maximal pending subtree in common with T^2, but no combination of a maximal pending subtree of T^1 and a maximal pending subtree of T^2 yields a tree with $2^m + p$ leaves.) *A fortiori*, one of the maximal pending subtrees of T is a fully symmetric tree with 2^{m-1} leaves and the other must have, thus, the remaining $2^{m-1} + p$ leaves. This shows that $n_a = 2^{m-1} + p$ and $n_b = 2^{m-1}$ and T_b is fully symmetric.

  (ii) If $2^{m-1} < p \leqslant 2^m - 3$ then $n-1 = 2^m + (p-1)$ with $2^{m-1} \leqslant p-1 < 2^m$ and $n+1 = 2^m + (p+1)$ with $2^{m-1} < p+1 < 2^m$. Then, since T^1 and T^2 satisfy assertion (ii),

  $$n_a^1 = 2^m,\ n_b^1 = p - 1,\ n_a^2 = 2^m,\ n_b^2 = p + 1$$

and both T_a^1 and T_a^2 are fully symmetric and hence (since they have the same numbers of leaves) $T_a^1 = T_a^2$. Reasoning as in the previous case, we deduce that T shares with both T^1 and T^2 the fully symmetric maximal pending subtree $T_a^1 = T_a^2$. In particular, one of its maximal pending subtrees has 2^m leaves (and it is fully symmetric) and the other must have, thus, the remaining p leaves. This shows that $n_a = 2^m$ and $n_b = p$ and T_a is fully symmetric.

(iii) Consider finally the case when $p = 2^m - 1 > 2^{m-1}$. Then, $n - 1 = 2^m + (p - 1)$ with $2^{m-1} \leqslant p - 1 < 2^m$ and $n + 1 = 2^{m+1}$. In this case, since T^1 satisfies assertion (ii) and T^2 satisfies assertion (i),

$$n_a^1 = 2^m,\ n_b^1 = 2^m - 2,\ n_a^2 = 2^m,\ n_b^2 = 2^m$$

and T_a^1, T_b^1 and T_b^2 are fully symmetric and hence (since they have the same numbers of leaves) $T_a^1 = T_b^1 = T_b^2$. Arguing as in the previous cases we conclude that T has a maximal pending subtree with 2^m leaves that is fully symmetric and the other maximal pending subtree with the remaining $2^m - 1$ leaves, and hence it satisfies assertion (ii).

This completes the proof.