

Extremal values of the Sackin balance index for rooted binary trees

Mareike Fischer*

Institute of Mathematics and Computer Science, Greifswald University, Greifswald, Germany

Abstract

Tree balance plays an important role in different research areas like theoretical computer science and mathematical phylogenetics. For example, it has long been known that under the Yule model, a pure birth process, imbalanced trees are more likely than balanced ones. Therefore, different methods to measure the balance of trees were introduced. The Sackin index is one of the most frequently used measures for this purpose. In many contexts, statements about the minimal and maximal values of this index have been discussed, but formal proofs have never been provided. Moreover, while the number of trees with maximal Sackin index as well as the number of trees with minimal Sackin index when the number of leaves is a power of 2 are relatively easy to understand, the number of trees with minimal Sackin index for all other numbers of leaves was completely unknown. In this manuscript, we fully characterize trees with minimal and maximal Sackin index and also provide formulas to explicitly calculate the number of such trees.

Keywords: tree balance, Sackin index, rooted binary tree

1. Introduction

Rooted trees, and binary ones in particular, play a fundamental role in many sciences as they can be used as a basis for search algorithms as well as, amongst others, as a model for evolution. In many cases where these trees occur, probability distributions are not always uniform concerning the degree of tree balance – for instance, the Yule model in phylogenetics, which is a pure birth process, has long been known to lead to more imbalanced trees. A simple example is depicted in Figure 1, where it can be seen that if all leaves of a tree with three leaves are equally likely to give rise to a new leaf, then two of them lead to the same (‘imbalanced’) tree, whereas the other possible tree (the ‘balanced’ one) occurs only once. So in order to understand such processes and

*Corresponding author

Email address: email@mareikefischer.de (Mareike Fischer)

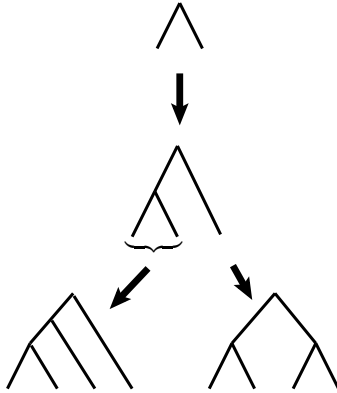


Figure 1: The Yule process splits one leaf at a time uniformly at random to form a so-called cherry. It can be easily seen that this leads to a tree shape bias already when there are $n = 4$ leaves. This is due to the fact that the only rooted binary tree on three leaves has two leaves that give rise to the tree on the left, which is considered ‘imbalanced’ (it is the so-called caterpillar tree T_4^{cat} on 4 leaves), whereas only one leaf leads to the so-called fully balanced tree T_2^{bal} of height 2.

their possible bias towards imbalanced (or, in other cases, balanced) trees, one has to be able to classify the degree of balance in more detail than just in a binary way (‘balanced’ versus ‘imbalanced’). Therefore, various balance indices were introduced and have been used over the years, e.g. [2, 3, 6, 7, 10]. One of the most frequently used and discussed such indices is the Sackin index [7].

This index has been observed to have some very nice properties – for instance, it has been stated that its maximum is achieved by the caterpillar tree (the tree with only one ‘cherry’, i.e. with only one internal node whose two descendants are both leaves) and that, whenever the number n of leaves equals a power of 2, i.e. when $n = 2^k$ for some k , the minimum is achieved by the so-called fully balanced tree of height k , i.e. the tree in which all leaves have distance k to the root [8]. However, while these statements can be found in the literature, rigorous proofs of them are nowhere to be found. It is one aim of this manuscript to present proofs and thus to show that the Sackin index indeed has these properties. Note that these properties are desirable for a good tree balance index, as the caterpillar tree is normally perceived as very ‘imbalanced’, whereas the fully balanced tree is normally referred to as very ‘balanced’ (which explains its name).

The idea of the Sackin index is to assign a small number to trees that are perceived as balanced and a high number to more imbalanced trees – i.e. the higher the Sackin index, the more imbalanced the tree. So while some statements on the maximum of the Sackin index and its minimum in the special case $n = 2^k$ can already be found in the literature, even if without proofs, little is known about trees with the minimum Sackin index for $n \neq 2^k$. Moreover, while it has long been known that the tree achieving this minimum in such cases need not be unique, the number of most balanced or most imbalanced trees has never

been formally investigated.

The main aim of this manuscript is to state and rigorously prove the following statements:

1. For all leaf numbers n , the caterpillar is the unique tree with maximal Sackin index.
2. For $n \in \{2^m - 1, 2^m, 2^m + 1\}$ ($m \in \mathbb{N}$), the minimum of the Sackin index is achieved by a unique tree. If $n = 2^m$, this tree is the fully balanced tree.
3. If there is **no** $m \in \mathbb{N}$ such that $n \in \{2^m - 1, 2^m, 2^m + 1\}$, then the minimum of the Sackin index is achieved by more than one tree, i.e. it is not unique.

Moreover, we will present an algorithm which constructs *all* trees with minimal Sackin index. Last but not least, we use this algorithm to derive a recursive formula for the number of such trees, which leads to a sequence that has only now been submitted to the Online Encyclopedia of Integer Sequences, i.e. it has not occurred elsewhere in the literature before.

2. Preliminaries

Before we can start to discuss the Sackin tree balance index, we first need to introduce all concepts used in this manuscript. We start with trees: *Trees* are connected, acyclic graphs with node set V and edge set E . We use V^1 in order to denote the set of *leaves* of a tree, i.e. the set of nodes of degree at most 1. All nodes v that are not leaves, i.e. $v \in V \setminus V^1$, are called *internal nodes*. The set of internal nodes of a tree T will be denoted by $\hat{V}(T)$, or, whenever there is no ambiguity, simply by \hat{V} .

All trees T in this manuscript are assumed to be *rooted* and *binary*, i.e. if they have an internal node at all, they have one root node ρ of degree 2 and all other internal nodes have degree 3. The only rooted binary tree which does not have an internal node is the tree that consists of only one node and no edge – in this special case, the only node is for technical reasons at the same time defined to be the root and the only leaf of the tree, so it is the only case where the root is not an internal node. Furthermore, for technical reasons all tree edges in this manuscript are implicitly assumed to be *directed* from the root to the leaves. Thus, for an edge $e = (u, v)$ of T , it makes sense to refer to u as the *direct ancestor* or *parent* of v (and v as the *direct descendant* or *child* of u). More generally, when there is a directed path from ρ to v employing u , u is called *ancestor* of v (and v descendant of u). Two leaves v and w are said to form a *cherry*, denoted by $[v, w]$, if v and w have the same parent, i.e. if there exists an internal node u in V such that (u, v) and (u, w) are edges in E . Note that every rooted binary tree with at least 2 leaves has at least one cherry.

Let T be a rooted binary tree with root ρ , and let $x \in V^1$ be a leaf of T . Then we denote by δ_x the *depth* of x in T , which is the number of edges on the unique shortest path from ρ to x . Then, the *height* of T is defined as $h(T) = \max_{x \in V^1} \delta_x$, i.e. as the maximum of these distances. Note that whenever a leaf v has maximal depth, i.e. whenever $h(T) = \delta_v$, v is element of a cherry.

This is due to the fact that if the other direct descendant, say w , of the parent of v , say u , was not a leaf but an internal node, it would have descending nodes of a greater depth than $\delta_v = \delta_w$, which would contradict the maximality of δ_v . This is why in this manuscript, instead of considering both v and w separately as leaves of maximal depth, we sometimes refer to a cherry $[v, w]$ as *cherry of maximal depth*.

Moreover, recall that a rooted binary tree T can be decomposed into its two maximal pending subtrees T_a and T_b rooted at the direct descendants a and b of ρ , and we denote this by $T = (T_a, T_b)$.

Last but not least, we want to introduce two particular trees which play a crucial role in this manuscript, namely the so-called *caterpillar tree* T_n^{cat} and the so-called *fully balanced tree* T_k^{bal} , respectively. T_n^{cat} denotes the unique rooted binary tree with n leaves that has only one cherry, while T_k^{bal} denotes the unique tree with $n = 2^k$ leaves in which all leaves have depth precisely k . Whenever there is no ambiguity concerning n or k , we also write T^{cat} and T^{bal} instead of T_n^{cat} and T_k^{bal} . T_4^{cat} and T_2^{bal} are depicted in the bottom row of Figure 1. Note that without loss of generality, the unique rooted binary tree with only one leaf, which consists of only one node and no edges, is defined to be T_1^{cat} , and it is thus the only caterpillar tree which does not contain a cherry. This tree is at the same time equal to T_0^{bal} , i.e. the fully balanced tree of height 0. This technicality enables inductive proofs concerning T^{cat} and T^{bal} to start at $n = 1$.

We are now in a position to define the central concept of this manuscript, namely the Sackin index. As there are four different versions of this index to be found in the literature, we will define all of them first and subsequently investigate their respective relationships. Note, however, that we focus on the first two definitions in the present manuscript, which can be shown to be equivalent. Therefore, we will not refer to the other two definitions as Sackin index, but give them modified names instead.

Definition 1. [6, 10] The *Sackin index* of a rooted binary tree T is defined as $\mathcal{S}(T) = \sum_{u \in \tilde{V}(T)} n_u$, where n_u denotes the number of leaves in the subtree of T rooted at u .

Note that in the following, whenever we have for two trees T_1 and T_2 that $\mathcal{S}(T_1) < \mathcal{S}(T_2)$, then T_1 is called *more balanced* than T_2 .

Definition 2. [6, 10] The *Sackin index* of a rooted binary tree T is defined as $\tilde{\mathcal{S}}(T) = \sum_{x \in V^1(T)} \delta_x$.

Definition 3. [10, (3.10)] The *Sackin index* of a rooted binary tree T with root ρ is defined as $\hat{\mathcal{S}}(T) = \sum_{u \in V(T) \setminus \{\rho\}} n_u$, where n_u denotes the number of leaves in the subtree of T rooted at u .

Note that in the original paper by Sackin [7], in fact no index is defined at all. Instead, a sequence b of leaf depths is defined, which implies that Definition

2 is probably most closely related to what Sackin originally intended. However, we will show in Lemma 1 that the first three definitions are in fact equivalent, which does not hold for the following definition.

Definition 4. [11] The *normalized Sackin index* of a rooted binary tree T with n leaves is defined as $\widehat{\mathcal{S}}(T) = \frac{1}{n} \sum_{x \in V^1(T)} \delta_x$, where δ_x denotes the depth of leaf x .

We now state a first lemma, which has already been partially stated (albeit without proof) in the literature [6].

Lemma 1. *Definitions 1, 2 and 3 are equivalent, i.e. for any rooted binary tree T , we have $\mathcal{S}(T) = \bar{\mathcal{S}}(T) = \widetilde{\mathcal{S}}(T)$.*

Proof. We first prove $\mathcal{S}(T) = \bar{\mathcal{S}}(T)$. Therefore, consider $\bar{\mathcal{S}}(T) = \sum_{x \in V^1(T)} \delta_x$. Note that while δ_x by definition denotes the number of edges separating leaf x from the root ρ of T , this is equivalent to the number of internal nodes on the path from ρ to x (including ρ). This leads to:

$$\begin{aligned} \bar{\mathcal{S}}(T) &= \sum_{x \in V^1(T)} \delta_x = \sum_{x \in V^1(T)} |\{u \in \mathring{V}(T) : u \text{ is an ancestor of } x\}| \\ &= |\{(x, u) : x \in V^1(T), u \in \mathring{V}(T) \text{ and } u \text{ is an ancestor of } x\}| \\ &= |\{(x, u) : x \in V^1(T), u \in \mathring{V}(T) \text{ and } x \text{ is a descendant of } u\}| \\ &= \sum_{u \in \mathring{V}(T)} |\{x \in V^1(T) : x \text{ is a descendant of } u\}| = \sum_{u \in \mathring{V}(T)} n_u = \mathcal{S}(T). \end{aligned}$$

So now we have $\mathcal{S}(T) = \bar{\mathcal{S}}(T)$, and next we show that $\mathcal{S}(T) = \widetilde{\mathcal{S}}(T)$. We have

$$\mathcal{S}(T) = \sum_{u \in \mathring{V}(T)} n_u = \sum_{u \in V(T)} n_u - \sum_{x \in V^1(T)} n_x = \left(\sum_{u \in V(T)} n_u \right) - n.$$

The latter equality is due to the fact that the only leaf belonging to a subtree rooted at a leaf is the leaf itself, so as we have n leaves, this gives n summands that contribute 1 to the sum. Now recall that $n_\rho = n$, so this leads to

$$\mathcal{S}(T) = \left(\sum_{u \in V(T)} n_u \right) - n_\rho = \sum_{u \in V(T) \setminus \{\rho\}} n_u = \widetilde{\mathcal{S}}(T).$$

This completes the proof. □

So because Definitions 1, 2 and 3 are equivalent, we do not have to distinguish between them and use them interchangeably. In fact, we will focus on in this manuscript mainly on the first two definitions.

The normalized Sackin index, however, is a modification of the Sackin index whenever trees with different numbers of leaves are considered, because the ranking induced by the normalized Sackin index can even reverse the ranking induced by the Sackin index. For instance, consider the two trees $T_1 = T_{37}^{cat}$, i.e. the caterpillar tree with 37 leaves, and $T_2 = T_9^{bal}$ the fully balanced tree with $2^9 = 512$ leaves. Then, it can easily be verified that we have $\mathcal{S}(T_1) = 702 < 4608 = \mathcal{S}(T_2)$, but $\widehat{\mathcal{S}}(T_1) \approx 18.97 > 9 = \widehat{\mathcal{S}}(T_2)$.¹ In fact, this ranking modification is no artifact but the very purpose of the normalization: The effect that many leaves automatically may lead to more ‘imbalance’ shall be eliminated. So in fact, \mathcal{S} and $\widehat{\mathcal{S}}$ can be very different – but only when different leaf numbers are considered! As long as n is fixed, the induced rankings of the two indices are of course equivalent, and in this case, $\widehat{\mathcal{S}}$ is just \mathcal{S} divided by the constant factor n . So when we discuss for instance the question how many trees with n leaves exist that have maximal or minimal Sackin index, the answers for \mathcal{S} and $\widehat{\mathcal{S}}$ will be the same.

Therefore, as this is sufficient for the number of minima and maxima, we focus in this manuscript on Definitions 1 and 2.

3. Results

It is the main aim of this manuscript to fully characterize trees with minimal and maximal Sackin index, respectively, and to count such trees. We will start with the easier case, which is the maximum, before we consider the more involved and therefore more interesting case of the minimum.

3.1. Maximally imbalanced trees / Minimally balanced trees

In this section, we present an upper bound for $\mathcal{S}(T)$ and prove that this bound is tight as it is achieved by T^{cat} . Moreover, we will show that for all values of n , T_n^{cat} is even the unique tree maximizing \mathcal{S} , i.e. the unique most imbalanced tree. This result has been stated in the literature before, e.g. in [6], but so far, a formal proof has not been stated anywhere.

However, we start by establishing the upper bound on $\mathcal{S}(T)$ before we can proceed as explained.

Theorem 1. *Let T be a rooted binary tree with n leaves. Then, we have:*

$$\mathcal{S}(T) \leq \frac{n \cdot (n + 1)}{2} - 1.$$

Before we can prove this theorem, we need one more lemma.

¹Note that these numbers can also be verified later on by using Propositions 1 and 2.

Lemma 2. *Let T be a rooted binary tree with $n \geq 2$ leaves. Let $[u, v]$ be a cherry of maximal depth in T with parent w , and let \tilde{T} be the tree derived by T by deleting u and v as well as the edges (w, u) and (w, v) . Then, for the Sackin indices of T and \tilde{T} we have:*

$$\mathcal{S}(T) = \mathcal{S}(\tilde{T}) + \delta_u + 1.$$

Proof. By Definition 2 and Lemma 1, we have $\mathcal{S}(T) = \sum_{x \in V^1} \delta_x$. By construction of \tilde{T} , as u and v have been deleted and w is a leaf in \tilde{T} but was an internal node in T , we have $\mathcal{S}(\tilde{T}) = \mathcal{S}(T) - \delta_u - \delta_v + \delta_w$. Note that as u and v form a cherry, we have $\delta_u = \delta_v$, and as w was the parent of u and v in T , we have $\delta_w = \delta_u - 1$. Thus, altogether we get $\mathcal{S}(\tilde{T}) = \mathcal{S}(T) - 2\delta_u + (\delta_u - 1) = \mathcal{S}(T) - \delta_u - 1$. Rearranging the latter term yields the desired result. \square

Now we can use Lemma 2 to prove Theorem 1.

Proof of Theorem 1. We prove the statement by induction on n . For $n = 1$ there is only one tree T , which consists of only one leaf and has no internal nodes. Thus, by Definition 1, $\mathcal{S}(T) = 0$ (as the sum in the definition of $\mathcal{S}(T)$ is empty). Moreover, $0 = \frac{1 \cdot 2}{2} - 1 = \frac{n \cdot (n+1)}{2} - 1$, which completes the base case of the induction.

Now we assume that for all trees with n leaves the claim is proven and consider a tree T with $n + 1$ leaves. It remains to show $\mathcal{S}(T) \leq \frac{(n+1)(n+2)}{2} - 1$. We construct a tree \tilde{T} by taking a cherry $[u, v]$ of T of maximal depth $\delta_u = \delta_v$ and deleting u and v as well as the edges leading from u and v to their direct ancestor, say w . Thus, \tilde{T} has $(n + 1) - 2 + 1 = n$ leaves, because leaves u and v have been deleted, but w , which is an internal node in T , is a leaf in \tilde{T} . By Lemma 2, we have $\mathcal{S}(T) = \mathcal{S}(\tilde{T}) + \delta_u + 1$. We use the inductive hypothesis for \tilde{T} and derive:

$$\mathcal{S}(T) = \mathcal{S}(\tilde{T}) + \delta_u + 1 \leq \frac{n(n+1)}{2} - 1 + \delta_u + 1 = \frac{n(n+1)}{2} + \delta_u.$$

Note that for all leaves, including those of maximal depth, their depths are bounded by one less than the number of leaves (cf. Lemma 6 in the appendix). We use this to conclude that $\delta_u \leq (n + 1) - 1 = n$. Therefore, we have:

$$\mathcal{S}(T) \leq \frac{n(n+1)}{2} + n = \frac{n^2 + 3n}{2} = \frac{n^2 + 3n}{2} + \frac{2}{2} - 1 = \frac{(n+1)(n+2)}{2} - 1.$$

This completes the proof. \square

We now consider the caterpillar tree T^{cat} and show that it achieves the bound induced by Theorem 1, which implies that this bound is indeed tight.

Proposition 1. *Let $T^{cat} = T_n^{cat}$ be the caterpillar tree with n leaves. Then, we have: $\mathcal{S}(T^{cat}) = \frac{n \cdot (n+1)}{2} - 1$, i.e. T^{cat} takes on the upper bound provided by Theorem 1.*

Proof. If $n = 1$, T^{cat} consists only of a leaf and has no internal nodes, so the sum over all internal nodes in Definition 1 is empty, which implies $\mathcal{S}(T^{cat}) = 0$. On the other hand, in this case we have $\frac{n \cdot (n+1)}{2} - 1 = \frac{1 \cdot (1+1)}{2} - 1 = 0 = \mathcal{S}(T^{cat})$. This completes the case $n = 1$.

Now consider the case $n \geq 2$. By Definition 2, we have $\mathcal{S}(T^{cat}) = \sum_{x \in V^1(T^{cat})} \delta_x$.

But note that in T^{cat} there is precisely one leaf of depth 1, one leaf of depth 2 and so forth. In the unique cherry of T^{cat} , both leaves have depth $n - 1$. So in total, we have:

$$\mathcal{S}(T^{cat}) = \sum_{x \in V^1(T^{cat})} \delta_x = 1 + 2 + \dots + (n - 1) + (n - 1) = \left(\sum_{i=1}^{n-1} i \right) + (n - 1).$$

Using the Gaussian sum, this leads to:

$$\mathcal{S}(T^{cat}) = \frac{(n-1)n}{2} + (n-1) = \frac{n^2 - n}{2} + \frac{2n - 2}{2} = \frac{n^2 + n - 2}{2} = \frac{n(n+1)}{2} - 1.$$

This completes the proof. \square

So by Proposition 1 we know that the caterpillar tree assumes the maximal Sackin index for all possible leaf numbers n . However, Proposition 1 does not make a statement about whether there exist other trees with maximal Sackin index. We will now show that this is not the case, i.e. the caterpillar tree is the unique most imbalanced tree.

Theorem 2. *Let $n \in \mathbb{N}_{\geq 1}$. Then, there is only one rooted binary tree T with maximal Sackin index, i.e. with $\mathcal{S}(T) = \frac{n(n+1)}{2} - 1$, and we have $T = T^{cat}$. Thus, the caterpillar tree is the unique tree maximizing the Sackin index.*

Before we can prove this theorem, we need one more lemma.

Lemma 3. *Let T be a rooted binary tree with n leaves and with maximal (or minimal) Sackin index for n , i.e. for all other trees \tilde{T} with n leaves we have $\mathcal{S}(T) \geq \mathcal{S}(\tilde{T})$ (or $\mathcal{S}(T) \leq \mathcal{S}(\tilde{T})$, respectively). Let $T = (T_a, T_b)$ be the standard decomposition of T into its two maximal pending subtrees T_a and T_b , with n_a and n_b leaves, respectively. Then, $\mathcal{S}(T_a)$ and $\mathcal{S}(T_b)$ are maximal (minimal) for n_a and n_b , respectively.*

Proof. We consider the case of maximality. The case of minimality can be shown analogously. Now, assume that $\mathcal{S}(T)$ is maximal. Using Definition 1, it is easy to see that $\mathcal{S}(T) = \mathcal{S}(T_a) + \mathcal{S}(T_b) + n$. Now assume that $\mathcal{S}(T_a)$ is not maximal, i.e. there is a tree \hat{T} with n_a leaves such that $\mathcal{S}(\hat{T}) > \mathcal{S}(T_a)$. Then we can construct a tree \tilde{T} on n leaves such that $\tilde{T} = (\hat{T}, T_b)$, i.e. we replace T_a in T by \hat{T} to derive \tilde{T} . Now for \tilde{T} we have by Definition 1: $\mathcal{S}(\tilde{T}) = \mathcal{S}(\hat{T}) + \mathcal{S}(T_b) + n > \mathcal{S}(T_a) + \mathcal{S}(T_b) + n = \mathcal{S}(T)$. This contradicts the maximality of $\mathcal{S}(T)$, which implies that the assumption was wrong. So $\mathcal{S}(T_a)$ has to be maximal, and analogously, $\mathcal{S}(T_b)$ has to be maximal, too. This completes the proof. \square

We are now in a position to prove Theorem 2.

Proof of Theorem 2. By Proposition 1, we have $\mathcal{S}(T^{cat}) = \frac{n(n+1)}{2} - 1$, which is maximal according to Theorem 1. Now assume there is another tree T with $\mathcal{S}(T) = \frac{n(n+1)}{2} - 1$. We prove by induction on n that then T must equal T^{cat} . For $n = 1$, there is only one rooted binary tree, which is by definition a caterpillar, so there is nothing to show. This completes the base case of the induction.

Next we assume that the statement is already proven for all leaf numbers up to $n-1$ and consider a rooted binary tree T with n leaves and $\mathcal{S}(T) = \frac{n(n+1)}{2} - 1$. Note that without loss of generality, $n \geq 2$ (else we consider the base case of the induction again). As $n \geq 2$, we can consider the standard decomposition of T into its two maximal pending subtrees T_a and T_b with n_a and n_b leaves, respectively. Note that $n = n_a + n_b$ and that we may assume without loss of generality that $n_a \geq n_b$. Moreover, recall that we have $\mathcal{S}(T) = \mathcal{S}(T_a) + \mathcal{S}(T_b) + n$ by Definition 1. By Lemma 3, $\mathcal{S}(T_a)$ and $\mathcal{S}(T_b)$ are also maximal (as $\mathcal{S}(T)$ is maximal), and thus, by induction, T_a and T_b are both caterpillars.

So we can conclude by Proposition 1 that $\mathcal{S}(T_a) = \frac{n_a(n_a+1)}{2} - 1$ and $\mathcal{S}(T_b) = \frac{n_b(n_b+1)}{2} - 1$, which gives

$$\mathcal{S}(T) = \mathcal{S}(T_a) + \mathcal{S}(T_b) + n = \frac{n_a(n_a+1)}{2} - 1 + \frac{n_b(n_b+1)}{2} - 1 + n.$$

Using $n_b = n - n_a$ and expanding all terms leads to

$$\mathcal{S}(T) = \frac{1}{2}n^2 - n \cdot n_a + n_a^2 + \frac{3}{2}n - 2. \quad (1)$$

We now consider this term as a function of n_a and analyze the values of this function more in-depth, i.e. we consider $f(n_a) = \frac{1}{2}n^2 - n \cdot n_a + n_a^2 + \frac{3}{2}n - 2$. Note that as we have $n_a + n_b = n$ and as we assume without loss of generality that $n_a \geq n_b$, we have $n_a \geq \lceil \frac{n}{2} \rceil$ (and $n_b \leq \lfloor \frac{n}{2} \rfloor$). Moreover, note that $n_a \leq n - 1$, because $n_b \geq 1$.

We now show that for $n_a \in \{\lceil \frac{n}{2} \rceil, \dots, n - 1\}$, $f(n_a)$ is strictly monotonically increasing. Therefore, consider the first derivative of f : $f'(n_a) = -n + 2n_a$, which equals 0 precisely if $n_a = \frac{n}{2}$ and is strictly larger than 0 for all $n_a > \frac{n}{2}$. In total, this implies that we have a unique minimum at $\frac{n}{2}$ and that f strictly increases after n_a passes this minimum.

So indeed, f is strictly monotonically increasing on $\{\lceil \frac{n}{2} \rceil, \dots, n - 1\}$, which implies that its unique maximum is assumed at $n - 1$. This implies that $\mathcal{S}(T)$ is maximal if $n_a = n - 1$, i.e. if T_a is the caterpillar on $n - 1$ leaves. Using $n = n_a + n_b$, this implies that $n_b = 1$. So in total, T_a is a caterpillar on $n - 1$ leaves and T_b consists of only one leaf. This implies that T is a caterpillar on n leaves, which completes the proof. \square

In total, we conclude that the bound provided by Theorem 1 is tight, as Proposition 1 implies that for each n , the caterpillar reaches this bound, and by Theorem 2 we know that the caterpillar is unique with this property, i.e. the

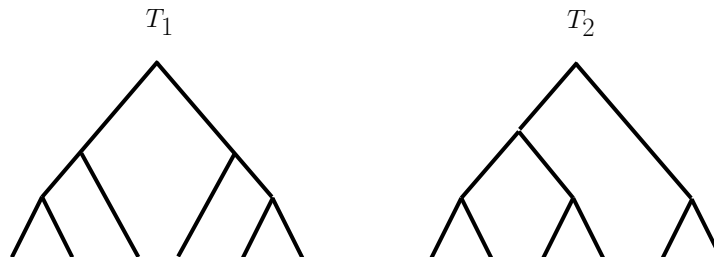


Figure 2: Two trees with $n = 6$ leaves which both have Sackin index 16, which can be shown to be minimal for $n = 6$.

caterpillar is the unique tree with maximal Sackin index. So for all n , there is precisely one tree with maximal Sackin index. We will show in the following section that this is different for the minimal Sackin index, as it can be taken on by various trees (depending on n).

We conclude this section by noting that the sequence $(a_n)_{n \in \mathbb{N}_{\geq 1}}$ with $a_n = \mathcal{S}(T_n^{cat}) = \frac{n(n+1)}{2} - 1$ for $i \in \mathbb{N} \geq 1$, which starts with 2, 5, 9, 14, 20, 27, 35, 44, 54, 65, 77, 90, 104, 119, 135, 152, 170, 189, 209, \dots , corresponds to sequence A000096 in the Online Encyclopedia of Integer Sequences OEIS [9, Sequence A000096], when the index is shifted by 1 (i.e. the i^{th} entry of the OEIS sequence corresponds to the $(i + 1)^{\text{st}}$ entry of our sequence). So this sequence has already occurred in other contexts, which might link the maximal Sackin index to other areas of research like the study of prime polyominoes or the traveling salesman polytope [9, Sequence A000096].

3.2. Maximally balanced trees

In this section, it is our aim to achieve the same results for trees with minimal Sackin index that the previous section stated for trees with maximal Sackin index. In particular, we want to find a tight bound for the minimal Sackin index and we want to characterize the trees that achieve it. However, it turns out that – as opposed to the previous section – the case of minimality is far more involved. In fact, depending on the number n of leaves, the tree with the minimal Sackin index need not be unique, so counting these trees is more complicated. Note that while examples for the fact that the Sackin index can be minimized by more than one tree have been presented before, e.g. in [6], see Figure 2, it has so far not been investigated for which values of n this happens and how many minima there are. It is the main aim of this section to present an algorithm that is able to systematically generate all trees with minimal Sackin index and that therefore also leads to a recursive formula for counting such trees.

However, we start as in the previous section and state the first result, which provides a lower bound on the Sackin index of trees with n leaves.

Theorem 3. *Let T be a rooted binary tree with n leaves. Let $k = \lceil \log_2(n) \rceil$. Then, $\mathcal{S}(T) \geq -2^k + n(k + 1)$.*

Proof. Note that as $k = \lceil \log_2(n) \rceil$, we have $n \leq 2^k$, and k is the smallest integer with this property. In other words, we have $k = \min\{\widehat{k} : n \leq 2^{\widehat{k}}\}$. We now prove the theorem by induction on n . For $n = 1$, we have $k = 0$, as $1 \leq 2^0$. So the term on the right hand side equals $-2^k + n(k+1) = -2^0 + 1 \cdot (0+1) = -1 + 1 = 0$. As $\mathcal{S}(T) \geq 0$ for all T by the definition of the Sackin index, this condition is clearly fulfilled. This completes the base case of the induction.

Let us now consider a tree T with n leaves. We assume that for all trees with $n - 1$ leaves, the statement already holds, and we have to show that now $\mathcal{S}(T) \geq -2^k + n(k+1)$. We consider a cherry $[u, v]$ of T of maximal depth $\delta_u = \delta_v$ and its parent node w . We consider tree \widetilde{T} which we derive from T as follows: We remove leaves u and v together with the edges (w, u) and (w, v) . By Lemma 2, this leads to $\mathcal{S}(T) = \mathcal{S}(\widetilde{T}) + \delta_u + 1$. Note that as \widetilde{T} was derived from T by deleting two leaves (u and v) but creating one new one (w), for the number of leaves \widetilde{n} of \widetilde{T} we have $\widetilde{n} = n - 1$. Let $\widetilde{k} = \lceil \log_2(\widetilde{n}) \rceil = \min\{\widetilde{k} : \widetilde{n} \leq 2^{\widetilde{k}}\}$. Then there are two cases: either $\widetilde{k} = k$ or $\widetilde{k} = k - 1$, as the deletion of one leaf might result in a smaller power of 2 being necessary to cover all leaves, but it cannot decrease by more than 1.

1. We first consider the case $\widetilde{k} = k - 1$. This case implies that the deletion of one leaf leads to a smaller power of 2 necessary to cover $n - 1$ than n . This is precisely the case if $n = 2^{k-1} + 1 > 2^{k-1} = 2^{\widetilde{k}}$, because then $n - 1 = 2^{k-1} = 2^{\widetilde{k}}$, and in this case, the smallest \widehat{k} such that $2^{\widehat{k}}$ is at least n would be $(k - 1) + 1 = k$.

So now we have $\mathcal{S}(T) = \mathcal{S}(\widetilde{T}) + \delta_u + 1$ and, by the inductive hypothesis, $\mathcal{S}(\widetilde{T}) \geq -2^{\widetilde{k}} + (n-1)(\widetilde{k}+1) = -2^{k-1} + 2^{k-1}((k-1)+1) = -2^{k-1} + 2^{k-1}k = 2^{k-1}(k-1)$. Combining these two observations, we derive

$$\mathcal{S}(T) = \mathcal{S}(\widetilde{T}) + \delta_u + 1 \geq 2^{k-1}(k-1) + \delta_u + 1. \quad (2)$$

It remains to show that the latter term in Equation 2 is at least $-2^k + n(k+1)$. Note that δ_u is the maximal depth of T , because u is an element of a cherry of maximal depth. It can easily be seen that as $n = 2^{k-1} + 1$, we have $\delta_u \geq k$ (see Lemma 5 in the appendix). So this leads to:

$$\delta_u \geq k = -2^k + 2^{k-1}(k+1 - (k-1)) + k \quad (3)$$

Now we combine Equations (2) and (3) to derive the desired result:

$$\mathcal{S}(T) \geq 2^{k-1}(k-1) - 2^k + 2^{k-1}(k+1 - (k-1)) + k + 1 = -2^k + n(k+1).$$

This completes the proof of this case.

2. Now we consider the case $\widetilde{k} = k$. In this case, we have $n \leq 2^k$ and $n - 1 > 2^{k-1}$. This implies that $n > 2^{k-1} + 1$ and thus, again by Lemma 5, we have $\delta_u \geq k$. Moreover, we still have $\mathcal{S}(T) = \mathcal{S}(\widetilde{T}) + \delta_u + 1$, and by

the inductive hypothesis we have that $\mathcal{S}(\tilde{T}) \geq -2^k + (n-1) \cdot k + (n-1)$. Combining all these observations leads to:

$$\mathcal{S}(T) = \mathcal{S}(\tilde{T}) + \delta_u + 1 \geq -2^k + (n-1) \cdot k + (n-1) + k + 1 = -2^k + n(k+1). \quad (4)$$

This completes the proof. □

We now consider the so-called fully balanced tree T_k^{bal} and show that its name is indeed justified in the sense that it achieves the bound induced by Theorem 3 whenever it is defined, i.e. whenever the number n of leaves equals 2^k .

Proposition 2. *Let T_k^{bal} be the fully balanced tree of height k with $n = 2^k$ leaves. Then, we have: $\mathcal{S}(T_k^{bal}) = k \cdot 2^k$.*

Proof. By Definition 2 and Lemma 1, we have $\mathcal{S}(T_k^{bal}) = \sum_{x \in V^1(T)} \delta_x$. Note that $|V^1| = n = 2^k$ and that each of the 2^k leaves in T_k^{bal} has depth k . This leads to $\mathcal{S}(T_k^{bal}) = k \cdot 2^k$, which completes the proof. □

Corollary 1. *Let T_k^{bal} be the fully balanced tree of level k with $n = 2^k$ leaves. Then, T_k^{bal} is maximally balanced, i.e. for all rooted binary trees T with $n = 2^k$ leaves we have: $\mathcal{S}(T) \geq \mathcal{S}(T_k^{bal})$.*

Proof. By Proposition 2 we have $\mathcal{S}(T_k^{bal}) = k \cdot 2^k$. Let T be a rooted binary tree with $n = 2^k$ leaves. By Theorem 3, we then have:

$$\mathcal{S}(T) \geq -2^k + n(k+1) = -2^k + \underbrace{2^k}_{=n} (k+1) = -2^k + 2^k \cdot k + 2^k = 2^k \cdot k = \mathcal{S}(T_k^{bal}).$$

This completes the proof. □

Corollary 1 shows that when $n = 2^k$, the bound provided by Theorem 3 is tight, i.e. there is in fact a tree that obtains the lower bound of the Sackin index, but it does not consider any other values of n . We next want to show that the bound is in fact tight for all n , even if $n \neq 2^k$, and we will present an explicit construction to find such maximally balanced trees, i.e. trees with a minimal Sackin index. Therefore, we need the following Algorithm 1. The high-level idea of this algorithm is to start with a maximally balanced tree (and we know already by Corollary 1 that T_k^{bal} has that property), possibly of too many leaves, and then deleting leaves, one at a time, in a way that keeps the balance maximal. In particular, we will show that this can be achieved by deleting the leaves of a cherry of maximum depth.

Algorithm 1: Construction of maximally balanced rooted binary trees with n leaves

Input: number n of leaves

Output: tree T with n leaves and $\mathcal{S}(T)$ minimal.

Initialization: $k := \lceil \log_2(n) \rceil$; $T := T_k^{bal}$; $\hat{n} := 2^k$;

while $\hat{n} > n$ **do**

Find a cherry $[u, v]$ of maximal depth in T and call its parent node w ;

Delete leaves u and v and edges (w, u) and (w, v) from T ;

$\hat{n} := \hat{n} - 1$;

return T

Theorem 4. Let $n \in \mathbb{N}$ and let $k = \lceil \log_2(n) \rceil$. Then, Algorithm 1 returns a tree T with $\mathcal{S}(T) = -2^k + n(k + 1)$. So Algorithm 1 finds a tree with minimal $\mathcal{S}(T)$; i.e. a maximally balanced tree.

Proof. Note that as before, $k = \lceil \log_2(n) \rceil$ implies $k = \min\{\tilde{k} : n \leq 2^{\tilde{k}}\}$. We now distinguish two cases. If $n = 2^k$, the algorithm does not enter the while-loop, as $\hat{n} = n$. Therefore, the algorithm, returns the tree from the initialization step, which is the fully balanced tree T_k^{bal} . We know by Proposition 2 that $\mathcal{S}(T_k^{bal}) = k \cdot 2^k$, which is minimal by Corollary 1. Moreover, $\mathcal{S}(T_k^{bal}) = k \cdot 2^k = -2^k + (k \cdot 2^k) + 2^k = -2^k + 2^k(k + 1) = -2^k + n(k + 1)$. Here, the last equality is due to $n = 2^k$. This completes the proof of this case.

Now let us consider the case where $n < 2^k$. Note that we additionally have $n > 2^{k-1}$, because otherwise k would not be minimal with the property $n \leq 2^k$.

We start with T_k^{bal} employing $\hat{n} = 2^k$ leaves, and the while-loop reduces \hat{n} in each step by 1 and repeats that as long as $\hat{n} > n$. So in total, this step is repeated $2^k - n$ times.

In each step, the leaves of a cherry $[u, v]$ of maximal depth are deleted along with their respective pending edges. Let us call the tree before performing this step T and the resulting tree after the cherry deletion \tilde{T} . Note that the direct ancestor of u and v in T , say w , is a leaf in \tilde{T} . So in total, \tilde{T} has one leaf less than T . So after repeating this step $2^k - n$ times, we end up with a tree with $2^k - (2^k - n) = n$ leaves.

Moreover, as $[u, v]$ is a cherry of maximal depth, we know from Lemma 2 that $\mathcal{S}(\tilde{T}) = \mathcal{S}(T) - \delta_u - 1$. So each time we run through the while-loop, the Sackin index gets reduced by $(\delta_u + 1)$.

It is now crucial to note that the value of δ_u does not change throughout the procedure! This is because we start with $\hat{n} = 2^k$ and stop when $\hat{n} = n > 2^{k-1}$, and we always take a cherry of maximal depth. In the beginning, we have $\frac{2^k}{2} = 2^{k-1}$ such cherries in T_k^{bal} to choose from, and in each step, one of those cherries gets replaced by a leaf. So as long as one of the original cherries of T_k^{bal} remains unchanged, the value of δ_u does not change. But if all of the original cherries were replaced by a leaf, we would end up with T_{k-1}^{bal} and thus with $\hat{n} = 2^{k-1} < n$, which by definition of Algorithm 1 cannot happen.

So δ_u is the same each time we run the while-loop, and thus, in total, on the way from T_k^{bal} to the final tree T , the Sackin index gets reduced by $(2^k - n)(\delta_u + 1)$.

1). So we have $\mathcal{S}(T) = \mathcal{S}(T_k^{bal}) - (2^k - n)(\delta_u + 1)$. Now note that the depth δ_u of a leaf of maximal depth in T_k^{bal} is precisely k . In total, using Proposition 2, this leads to:

$$\mathcal{S}(T) = \mathcal{S}(T_k^{bal}) - (2^k - n)(\delta_u + 1) = k \cdot 2^k - (2^k - n)(k + 1) = -2^k + n(k + 1).$$

This completes the proof. \square

Remark. Note that Theorem 4 proves that the bound provided by Theorem 3 is in fact tight for all n , and that Algorithm 1 even provides a way of finding a maximally balanced tree for all n . The tight bound also implies that the sequence of minimal values of the Sackin index, starting at $n = 1$, is 1, 2, 5, 8, 12, 16, 20, 24, \dots . This corresponds to Sequence A003314 in the Online Encyclopedia of Integer Sequences OEIS [9, Sequence A003314], which is also often referred to as binary entropy function.

Note that by Theorem 4, it is also clear why trees with minimal Sackin index are not necessarily unique with this property: In the while-loop, any cherry of maximum depth in a maximally balanced tree with \hat{n} leaves will provide a maximally balanced tree with $\hat{n} - 1$ leaves – so the choice of the particular cherry of maximum depth might result in different trees.

Example 1. Consider the two trees T_1 and T_2 with $n = 6$ leaves as depicted in Figure 2. These two trees both have Sackin index 16: $\mathcal{S}(T_1) = 2 + 2 + 3 + 3 + 6 = 16 = 2 + 2 + 2 + 4 + 6 = \mathcal{S}(T_2)$. Let $k = \lceil \log_2(6) \rceil = 3$.

Now it is easy to see that $\mathcal{S}(T_1) = \mathcal{S}(T_2) = 16$ is indeed minimal, as for $n = 6$ we have by Theorem 3 that $\mathcal{S}(T) \geq -2^k + n(k + 1) = -2^3 + 6(3 + 1) = 16$. And the reason why we have two maximally balanced trees with 6 leaves is due to the (unique) maximally balanced tree T with 7 leaves (see Figure 3). In this tree we have three cherries of maximal depth. However, due to symmetry it does not make a difference if we choose either one of the two underbraced ones in the while-loop of Algorithm 1, because the resulting tree is in any case T_1 . But if we choose the third cherry of maximal depth, the resulting tree is T_2 .

Remark. It can be easily seen that instead of using the top-down principle as presented in Algorithm 1, it is equivalent to use a bottom-up principle. In particular, instead of constructing a maximally balanced tree with n leaves according to the Sackin index by starting at T_k^{bal} for $k = \lceil \log_2(n) \rceil$ and deleting cherries of maximal depth, one could instead start at T_{k-1}^{bal} and replace leaves of minimal depth with a cherry. Figure 3 illustrates this as all arrows could simply be reversed. However, as both algorithms can easily be shown to lead to equivalent results, we omit the proof for the second algorithm.

We will now consider the set of trees of minimal Sackin index, i.e. the maximally balanced ones according to this index. As we will show, if the number of leaves is a power of 2, i.e. $n = 2^k$, the maximally balanced tree is unique,

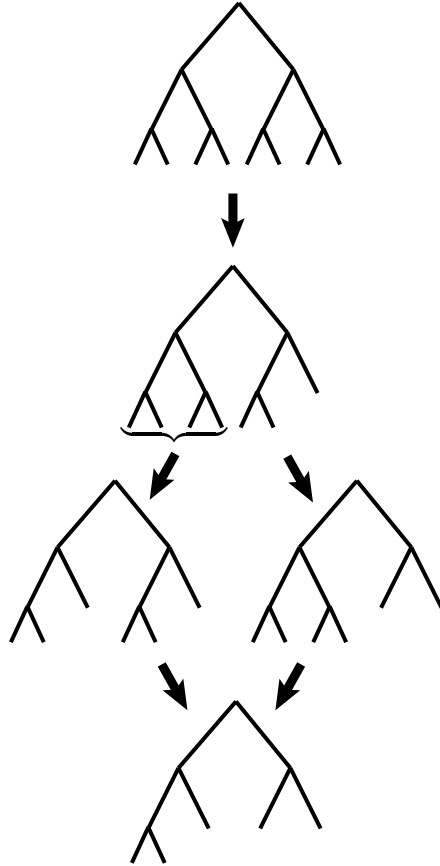


Figure 3: Illustration of Algorithm 1: For $n = 5$, we consider $k = \lceil \log_2(5) \rceil = 3$ and thus start at $2^3 = 8$ leaves. Then, one by one, we delete a cherry of maximal depth until we reach $n = 5$ leaves. In this case, when we go from 7 to 6, the particular choice of such a cherry can lead to different trees.

Note that alternatively, instead of going from 2^k leaves down to n , one could start at 2^{k-1} leaves and go up to n , i.e. the arrows in the figure could be reversed.

namely T_k^{bal} . This has frequently been observed in the literature (cf. [5, 8]), but to the best of our knowledge, no formal proof has been presented so far. Moreover, no statement on the number of maximally balanced trees for leaf numbers that are not a power of 2 has been made so far. We will provide a recursive formula to calculate this number in the following. But before we can do that, we have to consider the easier case with $n = 2^k$.

Theorem 5. *Let $n = 2^k$ for some $k \in \mathbb{N}$. Then, T_k^{bal} is the **unique** tree with minimal Sackin index, i.e. for any other tree T with $n = 2^k$ leaves we have $\mathcal{S}(T) > \mathcal{S}(T_k^{bal})$.*

Proof. We know by Corollary 1 that T_k^{bal} has minimum Sackin index, i.e. $\mathcal{S}(T_k^{bal}) = k \cdot 2^k$. So all that remains to be shown is that T_k^{bal} is unique with this property. We prove this statement by induction on k . For $k \leq 1$ there is nothing to show as then there is only one tree to consider, which gives therefore of course the unique minimum. Now we assume the statement holds for $k - 1$ and we consider a tree T with $n = 2^k$ leaves and $\mathcal{S}(T)$ minimal, i.e. by Proposition 2 and Corollary 1 we have $\mathcal{S}(T) = k \cdot 2^k$. We have to show that $T = T_k^{bal}$.

Let $T = (T_a, T_b)$ be the standard decomposition of T into its two maximal pending subtrees T_a and T_b with leaf numbers n_a and n_b , respectively, such that $n = n_a + n_b$. Without loss of generality, we may assume $n_a \geq n_b$, i.e. $n_a \geq \frac{n}{2}$. Note that as T has minimum Sackin index and as by definition we have $\mathcal{S}(T) = \mathcal{S}(T_a) + \mathcal{S}(T_b) + n$, by Lemma 3, $\mathcal{S}(T_a)$ and $\mathcal{S}(T_b)$ must be minimal, too. We now distinguish two cases.

1. $n_a = n_b = \frac{n}{2} = 2^{k-1}$. In this case, we know by the inductive hypothesis that T_a and T_b must equal T_{k-1}^{bal} , as this is the unique tree with minimum Sackin index and 2^{k-1} leaves. So $T = (T_a, T_b)$, and T_a and T_b are both equal to T_{k-1}^{bal} , which implies $T = T_k^{bal}$. This completes the proof of this case.
2. $n_a > n_b$, i.e. $n_a > \frac{n}{2} = 2^{k-1}$. Then, there exists an $s \in \mathbb{N}$, $s \geq 1$, such that $n_a = \frac{n}{2} + s = 2^{k-1} + s$. Due to $2^k = n = n_a + n_b = 2^{k-1} + s + n_b$ we conclude $2^k - 2^{k-1} = s + n_b$ and thus $n_b = 2^{k-1} - s$.

As in the first case we have $\mathcal{S}(T) = \mathcal{S}(T_a) + \mathcal{S}(T_b) + n$ with $\mathcal{S}(T_a)$ and $\mathcal{S}(T_b)$ minimal. By Theorems 3 and 4 we conclude that $\mathcal{S}(T_a) = -2^{k_a} + n_a \cdot (k_a + 1)$ and $\mathcal{S}(T_b) = -2^{k_b} + n_b \cdot (k_b + 1)$, where $k_a = \lceil \log_2(n_a) \rceil$ and $k_b = \lceil \log_2(n_b) \rceil$. Now, note that due to $n_a = 2^{k-1} + s$ and $n_b = 2^{k-1} - s$ for $s \geq 1$ we have:

- $k_a = k$ (as $n_a > 2^{k-1}$ and $n_a < n = 2^k$) and
- $0 \leq k_b \leq k - 1$ (as $n_b < 2^{k-1}$) and
- $s \leq 2^{k-1} - 1$, because as $n_b \geq 1$ (otherwise, if $n < 2$, we would be in the base case of the induction, so now we have $n \geq 2$ and therefore $n_b \geq 1$), we have $n_a \leq n - 1$. This implies $\frac{n}{2} + s \leq n - 1$, which yields $s \leq \frac{n}{2} - 1$. Using $n = 2^k$, we derive $s \leq 2^{k-1} - 1$.

So, in total we get:

$$\mathcal{S}(T) = \mathcal{S}(T_a) + \mathcal{S}(T_b) + n = -2^{k_a} + n_a(k_a + 1) - 2^{k_b} + n_b(k_b + 1) + 2^k.$$

Using $k_a = k$, $n_a = 2^{k-1} + s$ and $n_b = 2^{k-1} - s$, we get:

$$\begin{aligned}\mathcal{S}(T) &= -2^k + (2^{k-1} + s) \cdot (k+1) - 2^{k_b} + (2^{k-1} - s) \cdot (k_b + 1) + 2^k \\ &= 2^{k-1}(k + k_b) + s(k - k_b) + 2^k - 2^{k_b}.\end{aligned}$$

Using $k_b = \lceil \log_2(n_b) \rceil = \lceil \log_2(2^{k-1} - s) \rceil$, this leads to:

$$\mathcal{S}(T) = 2^{k-1} \cdot (k + \lceil \log_2(2^{k-1} - s) \rceil) + s(k - \lceil \log_2(2^{k-1} - s) \rceil) + 2^k - 2^{\lceil \log_2(2^{k-1} - s) \rceil}. \quad (5)$$

We now consider the term $k_b = \lceil \log_2(2^{k-1} - s) \rceil$ in more detail. For a fixed value of k , this term depends only on s , so we will refer to it as $k_b(s)$. $k_b(s)$ can assume values ranging from $0 = k - k$ (if $n_b = 1$, i.e. if T_b consists only of one leaf, which is the case when $s = 2^{k-1} - 1$) to $k - 1$ if $s \in \{1, \dots, 2^{k-2} - 1\}$ (note that as we are in the case where $n_a \neq n_b$, we already know that $s > 0$).

Moreover, as $k_b(s) = \lceil \log_2(2^{k-1} - s) \rceil$, we can conclude $s \geq 2^{k-1} - 2^{k_b(s)}$.

This implies:

$$k_b(s) = k - a \text{ if } s \in \{2^{k-1} - 2^{k-a}, \dots, 2^{k-1} - 2^{k-a-1} - 1\},$$

where a ranges from 1 to k . The only exceptions are the cases $a = 1$, where the possible choices for s start at 1, not at $2^{k-1}(2^0 - 1) = 0$ (because we already know that $s > 0$), and $a = k$, because $2^{k-(k+1)}(2^k - 1) - 1 = 2^{k-1} - \frac{3}{2} \notin \mathbb{N}$, so s cannot assume this value (which is why in this case, the set of possible choices of s contains only one element).

Now we use these insights concerning $k_b(s)$ to find minimal values of $\mathcal{S}(T)$ using Equation (5).

First, we consider two trees T_1 and T_2 such that $s_1, s_2 \in \{2^{k-a}(2^{a-1} - 1), \dots, 2^{k-(a+1)}(2^a - 1) - 1\}$ for some value of $a \in \{1, \dots, k\}$. As both values s_1 and s_2 correspond to the same value of a , we have $k_b(s_1) = k_b(s_2) = k - a$ and therefore we get by Equation (5):

$$\begin{aligned}\mathcal{S}(T_1) &= 2^{k-1} \cdot (k + (k - a)) + s_1(k - (k - a)) + 2^k - 2^{k-a} \\ &= 2^{k-1} \cdot (2k - a) + s_1 \cdot a + 2^k - 2^{k-a}.\end{aligned}$$

Analogously, we get:

$$\begin{aligned}\mathcal{S}(T_2) &= 2^{k-1} \cdot (k + (k - a)) + s_2(k - (k - a)) + 2^k - 2^{k-a} \\ &= 2^{k-1} \cdot (2k - a) + s_2 \cdot a + 2^k - 2^{k-a}.\end{aligned}$$

This implies that $\mathcal{S}(T_1) < \mathcal{S}(T_2)$ if and only if $s_1 < s_2$. So in order to minimize the Sackin index, in each possible line of Equation (??), i.e. for each possible value of a , the lower bound of the possible values for s is the only candidate for a minimum.

Now let us summarize what we have: If $a > 1$, we know that $s = 2^{k-a}(2^{a-1} - 1)$ minimizes \mathcal{S} , and if $a = 1$, $s = 1$ minimizes \mathcal{S} . Next,

we compare two trees T_1 and T_2 with different values a_1 and a_2 , respectively, i.e. let a_1 and a_2 be in $\{2, \dots, k\}$ such that $a_1 < a_2$ (the case where one of the values is 1 will be considered later). Then, in order for \mathcal{S} to be as small as possible, we have already seen that we must use the smallest possible values of s_1 and s_2 , respectively, i.e. $s_1 = 2^{k-a_1}(2^{a_1-1} - 1)$ and $s_2 = 2^{k-a_2}(2^{a_2-1} - 1)$. Using this and considering again Equation (5), we derive:

$$\begin{aligned}\mathcal{S}(T_1) &= 2^{k-1} \cdot (k + (k - a_1)) + s_1(k - (k - a_1)) + 2^k - 2^{k-a_1} \\ &= 2^k(k + 1) - 2^{k-a_1} \cdot (a_1 + 1),\end{aligned}$$

$$\mathcal{S}(T_2) = 2^k(k + 1) \underbrace{- 2^{k-a_2}}_{> -2^{k-a_1}} \cdot \underbrace{(a_2 + 1)}_{> (a_1+1)} > \mathcal{S}(T_1).$$

So if $a_1 < a_2$ (for $a_1, a_2 \in \{2, \dots, k\}$), we have $\mathcal{S}(T_1) < \mathcal{S}(T_2)$. This implies that for a minimal value of \mathcal{S} , a has to be minimal. So the only candidate for a from the set $\{2, \dots, k\}$ is $a = 2$.

However, we have not investigated the case $a = 1$ yet, which is different because in this case we are not allowed to choose $s = 2^{k-1}(2^0 - 1) = 0$ as we are in the case where $s \geq 1$. So what remains to be done is to compare the case $a = 1$ and $s = 1$ with the case $a = 2$ and $s = 2^{k-2}(2^1 - 1) = 2^{k-2}$. Therefore, now let T_1 be such that $s_1 = 1$ and thus $a_1 = 1$, and let T_2 be such that $s_2 = 2^{k-2}$ and $a_2 = 2$. Then, we get by Equation (5):

$$\mathcal{S}(T_1) = 2^{k-1} \cdot (k + (k - a_1)) + s_1(k - (k - a_1)) + 2^k - 2^{k-a_1} = 2^k k + 1,$$

and analogously

$$\mathcal{S}(T_2) = 2^{k-1} \cdot (k + (k - a_2)) + s_2(k - (k - a_2)) + 2^k - 2^{k-a_2} = 2^k k + 2^{k-2}.$$

So as we are in the case where $k \geq 2$ (else consider the base case of the induction again), this leads to $\mathcal{S}(T_2) \geq \mathcal{S}(T_1) = 2^k k + 1$. So the minimal value of \mathcal{S} is achieved by the choice of $s = 1$ and thus $a = 1$, i.e. $k_b = k - 1$, but even for a tree fulfilling these conditions we have $\mathcal{S}(T) = 2^k k + 1 > 2^k k = \mathcal{S}(T_k^{bal})$. The latter equality holds due to Corollary 1. So in total, we obtain $\mathcal{S}(T) > \mathcal{S}(T_k^{bal})$, which contradicts the minimality of $\mathcal{S}(T)$. Therefore, the case $n_a \neq n_b$ leads to a contradiction, which is why it cannot occur. Thus we only have to consider the first case, for which we have already shown $T = T_k^{bal}$. This completes the proof. \square

So now we know that in case of $n = 2^k$, the minimum of the Sackin index is unique. But it has long been known that this need not be the case for other values of n . In fact, already for $n = 6$, there are two minima, which are depicted

in Figure 2. As stated before, this example is not new; it can for instance already be found in [6]. However, so far the explicit number of Sackin minima for $n \neq 2^k$ has not been investigated. In order to derive such a formula, we first need the following theorem, which characterizes all Sackin minima.

Theorem 6. *Let T be a rooted binary tree with $n \geq 2$ leaves and $\mathcal{S}(T)$ minimal, i.e. there is no tree on n leaves with a smaller Sackin index. Then, we have: T has height $h_T = k$ with $k = \lceil \log_2(n) \rceil$, and T has precisely $n - 2^{k-1}$ cherries of depth k .*

Proof. We prove the statement by induction on n . Throughout this proof, for a rooted binary tree T , we denote by c_T the number of cherries of maximal depth.

Now for $n = 2$, the only binary rooted tree T consists only of a cherry, and this tree has $h_T = 1 = \lceil \log_2(2) \rceil = k$, so the height statement holds. Moreover, T has only one cherry, so $c_T = 1 = 2 - 2^{1-1} = n - 2^{k-1}$. This completes the base case.

Next we assume that the statement is true for up to n leaves and now consider a tree T with $n + 1$ leaves and with minimal Sackin index. First, in case $n + 1 = 2^k$ for some k , by Theorem 5 we know that $T = T_k^{bal}$. This immediately implies $h_T = \log_2(n + 1) = k$ and $c_T = 2^{k-1} = (n + 1) - 2^{k-1}$.

This completes the proof of the case where $n + 1$ is a power of 2.

So now we consider the case $2^{k-1} < n + 1 < 2^k$ with $k = \lceil \log_2(n + 1) \rceil$. Let $[u, v]$ be a cherry of maximal depth in T with direct ancestor w . We remove u, v as well as the edges (w, u) and (w, v) to get a tree \tilde{T} with n leaves (u and v have been deleted, but w , which is in T in internal node, is a leaf in \tilde{T}). Now, again there are two cases:

1. $n = 2^{k-1}$, which is precisely the case if $n + 1 = 2^{k-1} + 1$, or
2. $n > 2^{k-1}$.

1. We now consider the first (and easier) case: If $n = 2^{k-1}$, there are again only two possibilities:

- (a) Either \tilde{T} has minimal Sackin index, in which case (by Theorem 5) we know that $\tilde{T} = T_{k-1}^{bal}$, or
- (b) $\mathcal{S}(\tilde{T}) > \mathcal{S}(T_{k-1}^{bal})$.

- (a) If $\tilde{T} = T_{k-1}^{bal}$, we have $h_{\tilde{T}} = k - 1$, and as \tilde{T} was derived by T by deleting one cherry of maximal depth, T must look as depicted in Figure 4. In particular, $h_T = (k - 1) + 1 = k$, and T has only one cherry of depth k (namely $[u, v]$), as otherwise $h_{\tilde{T}}$ would equal h_T . So $c_T = 1 = (2^{k-1} + 1) - 2^{k-1} = (n + 1) - 2^{k-1}$. This completes the proof for the case $\tilde{T} = T_{k-1}^{bal}$.

- (b) Now assume that we have $\mathcal{S}(\tilde{T}) > \mathcal{S}(T_{k-1}^{bal})$, i.e. in summary we now have that T with $n + 1 = 2^{k-1} + 1$ leaves has $\mathcal{S}(T)$ minimal, but \tilde{T} with $n = 2^{k-1}$ leaves has a Sackin index strictly larger than that of T_{k-1}^{bal} , i.e. $\mathcal{S}(\tilde{T}) > \mathcal{S}(T_{k-1}^{bal}) = 2^{k-1} \cdot (k - 1)$. The last equality is due

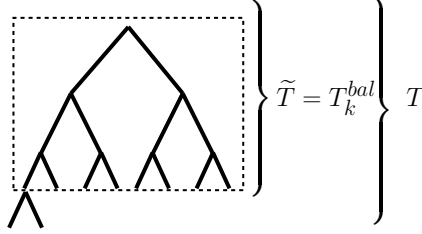


Figure 4: Illustration of the scenario described in Case (1a).

to Proposition 2. Note that by construction of \tilde{T} and Lemma 2, we have $\mathcal{S}(\tilde{T}) = \mathcal{S}(T) - \delta_u - 1$. Combining both statements on $\mathcal{S}(\tilde{T})$, we conclude

$$\mathcal{S}(\tilde{T}) = \mathcal{S}(T) - \delta_u - 1 > 2^{k-1} \cdot (k-1) = \mathcal{S}(T_{k-1}^{bal}). \quad (6)$$

On the other hand, as $\mathcal{S}(T)$ is minimal by assumption, we know from Theorems 3 and 4 that

$$\mathcal{S}(T) = -2^k + (n+1)(k+1) = -2^k + nk + n + k + 1. \quad (7)$$

Combining (6) with (7), we obtain:

$$-2^k + nk + n + k + 1 - \delta_u - 1 > k \cdot 2^{k-1} - 2^{k-1}.$$

Using $n = 2^{k-1}$, we note that this holds precisely if

$$-2^k + k \cdot 2^{k-1} + 2^{k-1} + k + 1 - \delta_u - 1 > k \cdot 2^{k-1} - 2^{k-1}.$$

Rearranging the inequality shows that this is fulfilled if and only if

$$-2^k + 2 \cdot 2^{k-1} + k > \delta_u,$$

i.e. precisely if $k > \delta_u$. But recall that u was part of a cherry of maximal depth, i.e. u has maximal depth in T , and T has $2^{k-1} + 1 > 2^{k-1}$ leaves. Thus, it can easily be seen (cf. Lemma 5) that $h_T \geq k$, which implies $\delta_u \geq k$. This contradicts $k > \delta_u$, which shows that this case cannot happen. So if \tilde{T} has $n = 2^{k-1}$ leaves, \tilde{T} must equal T_{k-1}^{bal} , for which we have already proven the statement in (1a). This completes the proof of this case.

2. Now what remains to be considered is the case $2^{k-1} < n < n+1 < 2^k$, which implies $k = \lceil \log_2(n) \rceil = \lceil \log_2(n+1) \rceil$. As above, we know that $\mathcal{S}(T)$ is minimal, i.e. by Theorems 3 and 4 we have $\mathcal{S}(T) = -2^k + (n+1)(k+1)$, and again, by construction of \tilde{T} , we have $\mathcal{S}(\tilde{T}) = \mathcal{S}(T) - \delta_u - 1$. Again, we distinguish two cases:
 - (a) $\mathcal{S}(\tilde{T})$ is minimal for n , or
 - (b) there is another tree \hat{T} on n leaves with $\mathcal{S}(\hat{T}) < \mathcal{S}(\tilde{T})$.

- (a) If $\mathcal{S}(\tilde{T})$ is minimal for n , we know by Theorems 3 and 4 that $\mathcal{S}(\tilde{T}) = -2^k + n(k+1)$. Combining this with $\mathcal{S}(\tilde{T}) = \mathcal{S}(T) - \delta_u - 1$ yields

$$\mathcal{S}(T) = -2^k + nk + n + \delta_u + 1. \quad (8)$$

On the other hand, we also know that $\mathcal{S}(T)$ is minimal, so we have

$$\mathcal{S}(T) = -2^k + (n+1)(k+1). \quad (9)$$

Combining Equations (8) and (9) leads to $-2^k + nk + n + \delta_u + 1 = -2^k + nk + n + k + 1$, which immediately gives $\delta_u = k$. As u was a leaf of maximal depth, we have $\delta_u = h_T$ and thus $h_T = k$. What remains to be shown is that $c_T = (n+1) - 2^{k-1}$. But note that as $\mathcal{S}(\tilde{T})$ is minimal and \tilde{T} has n leaves, we know by induction that $c_{\tilde{T}} = n - 2^{k-1}$ and $h_{\tilde{T}} = k$. We have already shown that $h_T = k$, too. So T and \tilde{T} have the same height k , but \tilde{T} was constructed by deleting one cherry of maximal depth from T . So we have $c_T = c_{\tilde{T}} + 1 = n - 2^{k-1} + 1 = (n+1) - 2^{k-1}$. This completes the proof of this case.

- (b) On the other hand, if $\mathcal{S}(\tilde{T})$ is not minimal, we know by Theorems 3 and 4 that $\mathcal{S}(\tilde{T}) > -2^k + n(k+1)$. As above, we also know that $\mathcal{S}(T) = \mathcal{S}(\tilde{T}) + \delta_u + 1$, and thus we conclude

$$\mathcal{S}(T) > -2^k + n(k+1) + \delta_u + 1. \quad (10)$$

On the other hand, though, $\mathcal{S}(T)$ is minimal, so again by Theorems 3 and 4 we have

$$\mathcal{S}(T) = -2^k + (n+1)(k+1). \quad (11)$$

Combining Inequality (10) with Equation (11), we conclude

$$-2^k + (n+1)(k+1) > -2^k + n(k+1) + \delta_u + 1,$$

which holds precisely if $k > \delta_u$. However, as u was a leaf of maximum depth in T , we have $h_T = \delta_u$ and thus $h_T < k$. But this is a contradiction, because as T has more than 2^{k-1} leaves, it has at least height k , i.e. $h_T \geq k$ (cf. Lemma 5). This implies that this case cannot happen, so in fact, $\mathcal{S}(\tilde{T})$ has to be minimal and we must be in case (2a), for which we have already proven the statement. This completes the proof. □

Note that Theorem 6 characterizes trees with minimal Sackin index in terms of their height, which must be minimal, and in terms of the number of cherries of maximal depth, which must be maximal (both with respect to the number of leaves). So trees with minimal Sackin index can be easily identified with Theorem 6.

Example 2. Consider again the two trees on $n = 6$ taxa depicted in Figure 2, which – as already stated above and as can be seen by exhaustive search through the space of all rooted binary trees with 6 leaves – have minimal Sackin index for $n = 6$. We have $k = \lceil \log_2(6) \rceil = 3$, and indeed, both trees have height 3 and 2 cherries of maximum depth 3, respectively, and thus $h_T = k = 3$ and $c_T = n - 2^{k-1} = 6 - 2^{3-1} = 6 - 4 = 2$.

Note that so far we have shown that all trees with minimal Sackin index have the height and cherry properties described in Theorem 6, but as the previous example shows, we still have not shown that all trees with these properties are also automatically minimal (we had to hint at an exhaustive search in order to verify that the two trees in the example indeed were maximally balanced). But in fact, it turns out that the opposite is also true, i.e. if the height is minimal and if the number of cherries of maximal depth is maximal, then the tree under consideration has minimal Sackin index. Indeed, this must be true as we will show now that such a tree will be discovered by Algorithm 1, which by Theorems 3 and 4 returns only trees with minimal Sackin index.

Theorem 7. *Let T be a rooted binary tree with n leaves and height $h_T = k = \lceil \log_2(n) \rceil$ and with $c_T = n - 2^{k-1}$ cherries of maximal depth k . Then, $\mathcal{S}(T)$ is minimal, i.e. there is no tree \hat{T} with n leaves and $S(\hat{T}) < \mathcal{S}(T)$.*

Proof. Let T be as described in the theorem. We will first prove that T can be discovered by Algorithm 1. Note that $2^{k-1} < n \leq 2^k$ for $k = \lceil \log_2(n) \rceil$. Recall that Algorithm 1 starts with T_k^{bal} and deletes one cherry of maximal depth at a time until n leaves are reached. Moreover, as long as $n > 2^{k-1}$ (which must be the case because otherwise Algorithm 1 would start at T_{k-1}^{bal} instead), we only delete at most $2^k - (2^{k-1} + 1) = 2^k - 2^{k-1} - 1 = 2^{k-1}(2 - 1) - 1 = 2^{k-1} - 1$ cherries² of the originally 2^{k-1} cherries of maximum depth in T_k^{bal} . So any tree on n leaves with $2^{k-1} < n \leq 2^k$ that the algorithm recovers has height k (as at least one cherry of maximum depth k is kept throughout the algorithm!). Moreover, the number of cherries of maximum depth in a tree of n leaves that are left by Algorithm 1 is $c_T = 2^{k-1} - (2^k - n)$, as we start with the 2^{k-1} cherries of T_k^{bal} and make $2^k - n$ steps, in each of which we delete exactly one cherry of maximum depth. So we end up with $c_T = 2^{k-1} - (2^k - n) = 2^{k-1} - 2^k + n = 2^{k-1}(1 - 2) + n = n - 2^{k-1}$. So, in summary, all trees recovered by Algorithm 1 have height $h_T = k$ and $c_T = n - 2^{k-1}$. And, more importantly, as the particular choice of cherry of maximum depth that the algorithm deletes is arbitrary, *all* such trees can be recovered. So T will be found by Algorithm 1, as all trees of height k and with $c_T = n - 2^{k-1}$ leaves of maximum depth can be reached by starting with T_k^{bal} and deleting $2^{k-1} - (2^k - n)$ cherries of maximum depth.

So T can be recovered by Algorithm 1. But, on the other hand, we know by Theorem 4 that any tree recovered by Algorithm 1 has minimal Sackin index. This completes the proof. \square

²Note that this extreme case would correspond to $n = 2^{k-1} + 1$.

This immediately leads to the following corollary, which is a direct conclusion of Theorems 6 and 7.

Corollary 2. *Let T be a rooted binary tree. Then, $\mathcal{S}(T)$ is minimal if and only if T can be recovered by Algorithm 1.*

Example 3. Consider again the two trees with $n = 6$ taxa, which are depicted in Figure 2. As both trees can be recovered by Algorithm 1 (c.f. Figure 3), both trees must by Theorem 4 indeed have minimal Sackin index, and as they are the only two trees with six leaves that are recovered by Algorithm 1, they are by Corollary 2 indeed the only two trees with this property. So an exhaustive search through all trees on six taxa as proposed above is not necessary anymore – instead, all trees with 6 leaves that can be recovered by Algorithm 1 are maximally balanced, and all other trees cannot be. In this sense, Corollary 2 provides a complete characterization of trees with minimal Sackin index.

As the previous example shows, for some values of n there is more than one tree with minimal Sackin index. However, without explicitly listing and enumerating all possible outputs of Algorithm 1, it is not easy to see why for six leaves there are two such trees and for, say, twelve leaves there are five such trees. Of course, by Corollary 2 it all has to do with the number of cherries of maximum depth we can choose from – but that is not all. Some choices might lead to the same tree. For instance, consider T_3^{bal} as depicted in Figure 3 at the top. It can easily be checked that no matter which of the cherries we choose to delete in order to recover a tree with $n = 7$ leaves and minimal Sackin index, we always end up with the same tree, namely with the tree depicted in the second line of the same figure. This is due to the symmetries of T_3^{bal} . However, when we go from $n = 7$ down to $n = 6$ and choose one of the underbraced cherries, we will end up with the same tree, namely with the one depicted in the third line of Figure 3 on the left-hand side. But if we choose the third cherry of maximum depth in this tree, this will lead to another tree – namely precisely the one depicted on the right-hand side of the same line in Figure 3. This is due to the fact that the two underbraced cherries are in some sense ‘symmetric’ (in the sense that there is a graph isomorphism which exchanges them), while the other cherry clearly is not symmetric to the two first ones. However, it is not trivial to count such symmetries or isomorphisms. But luckily, it is quite easy to characterize the cases where the minimum is unique, as can be seen in the following corollary.

Corollary 3. *Let $n \in \mathbb{N}$. Then, there is only one tree T with minimum Sackin index if and only if there exists an $m \in \mathbb{N}$ such that $n \in \{2^m - 1, 2^m, 2^m + 1\}$.*

Proof. If $n = 2^m$, then $m = \lceil \log_2(n) \rceil$, so $m = k$ in Theorem 5, which implies that the minimum is indeed unique. If $n = 2^m - 1$, then again for $k = \lceil \log_2(n) \rceil$, we have $m = k$. So in this case, Algorithm 1 would start with tree T_k^{bal} and delete precisely one cherry. It can be easily seen that due to symmetry, all cherries lead to the same tree. So for $n = 2^m - 1$, we again have uniqueness of the most balanced tree. However, if $n = 2^m + 1$, for $k = \lceil \log_2(n) \rceil$ we have

$k = m + 1$. In this case, the algorithm would start at T_{k+1}^{bal} and delete all but one cherries of maximal depth. Again, due to symmetry, it does not matter which of these cherries remains in the tree (the scenario is equivalent to the one depicted in Figure 4). So whenever $n \in \{2^m - 1, 2^m, 2^m + 1\}$, Algorithm 1 returns precisely one tree, which by Corollary 2 is the unique tree that minimizes the Sackin index in these cases.

Now, assume that $n \notin \{2^m - 1, 2^m, 2^m + 1\}$ for any m . In particular, for $k = \lceil \log_2(n) \rceil$, this implies that $n \in \{2^{k-1} + 2, \dots, 2^k - 2\}$. We partition this set into two cases: either $n \in \{2^{k-1} + 2, \dots, 2^k - 2^{k-2} - 1\}$ or $n \in \{2^k - 2^{k-2}, \dots, 2^k - 2\}$. Recall that Algorithm 1 starts with T_k^{bal} with 2^k leaves that form 2^{k-1} cherries. Now in the first case, if $n \in \{2^{k-1} + 2, \dots, 2^k - 2^{k-2} - 1\}$, at least $2^k - (2^k - 2^{k-2} - 1) = 2^{k-2} + 1$ cherries of T_k^{bal} get replaced by single leaves in the course of the algorithm. This implies that at most $2^{k-1} - (2^{k-2} + 1) = 2^{k-2} - 1$ cherries of maximal depth remain. So the remaining cherries can either all be together in one of the maximal pending subtrees of the tree, which we call T_a , or, for instance, one of them is in the other maximal pending subtree, which we call T_b . This already gives two options on how to choose a tree that can be recovered by Algorithm 1, which by Corollary 2 implies at least two optima.

On the other hand, if $n \in \{2^k - 2^{k-2}, \dots, 2^k - 2\}$, this implies that at most $2^k - (2^k - 2^{k-2}) = 2^{k-2}$ cherries get replaced by single leaves. These replacements can either all happen in the same maximal pending subtree of the tree, which we then call T_b , or, for instance, all can happen in T_b except for one which happens in the other maximal pending subtree, which we then call T_a . This again gives at least two options on how to choose a tree that can be recovered by Algorithm 1, which by Corollary 2 implies at least two optima.

So in summary, the minimum is unique if and only if $n \in \{2^m - 1, 2^m, 2^m + 1\}$ for some $m \in \mathbb{N}$. This completes the proof. \square

Note that Algorithm 1 guarantees that the difference between the two maximal pending subtrees T_a and T_b of the standard decomposition of a minimum Sackin tree does not get too large (in terms of the number of leaves). We will quantify this in the following corollary, which is based on Corollary 2.

Corollary 4. *Let T be a rooted binary tree with $n \in \mathbb{N}_{\geq 2}$ leaves. Moreover, let $T = (T_a, T_b)$ be the standard decomposition of T into its two maximal pending subtrees, let n_i denote the number of leaves in T_i for $i \in \{a, b\}$, respectively, such that $n_a \geq n_b$. Let $k = \lceil \log_2 n \rceil$. Then, the following equivalence holds: T has minimum Sackin index if and only if T_a and T_b have minimum Sackin index and $n_a - n_b \leq \min\{n - 2^{k-1}, 2^k - n\}$.*

Proof. We first consider the case where T has minimum Sackin index. It is clear that T_a and T_b then also have minimum Sackin index. Moreover, by considering Algorithm 1, one can easily see that the difference between n_a and n_b is maximal when you perform as many cherry deletions in T_b as possible. If all cherry deletions happen in T_b , which implies $n \geq 2^k - 2^{k-2} = 3 \cdot 2^{k-2}$, T_a equals T_{k-1}^{bal} and thus $n_a = 2^{k-1}$ and thus $n_b = n - 2^{k-1}$. So if $n \geq 2^k - 2^{k-2}$, we have $n_a - n_b \leq 2^{k-1} - n + 2^{k-1} = 2^k - n$. Moreover, as $n \geq 3 \cdot 2^{k-2}$ also

implies that $2^k - n \leq n - 2^{k-1}$, this completes the proof for the case where all cherry deletions happen in T_b .

However, if more than 2^{k-2} cherry deletions happen, T_b equals T_{k-2}^{bal} and thus we have $n_b = 2^{k-2}$. Moreover, in this case we have $n_a = n - 2^{k-2}$, and $n < 3 \cdot 2^{k-2}$. Thus, on the one hand we have $n_a - n_b = n - 2^{k-1}$ and $n - 2^{k-1} < 3 \cdot 2^{k-2} - 2^{k-1} = 2^{k-2}$. On the other hand, we have $2^k - n > 2^k - 3 \cdot 2^{k-2}$. This implies $n_a - n_b = n - 2^{k-1} < 2^k - n$, which completes the first direction of the proof.

Now let us consider $T = (T_a, T_b)$ such that T_a and T_b are Sackin minima and $n_a - n_b \leq \min\{n - 2^{k-1}, 2^k - n\}$. We need to show that T is also a Sackin minimum. First note that we have $n_a - n_b \leq 2^k - n$, which implies $n_a \leq 2^{k-1}$, and that we also have $n_a - n_b \leq n - 2^{k-1}$, which implies $n_b \geq 2^{k-2}$. As $n = n_a + n_b$ and as $n_a \geq n_b$, we have $n_a \geq \frac{n}{2}$ and $n_b \leq \frac{n}{2}$. So altogether we get $\frac{n}{2} \leq n_a \leq 2^{k-1}$ and $2^{k-2} \leq n_b \leq \frac{n}{2}$. It is easy to see that these leaf numbers precisely correspond to the trees recovered by Algorithm 1. This completes the proof. \square

As stated above, counting the symmetries that remain when cherries of maximal depth are deleted is not trivial, even if Corollary 3 already provides some insight into whether there is more than one tree that minimizes the Sackin index for a given value of n . So in order to conclude this section, we will show in the following that the number of trees with minimal Sackin index can be explicitly counted by a recursive formula. The proof of this formula exploits Corollary 2.

Theorem 8. *Let $s(n)$ denote the number of binary rooted trees with n leaves and with minimal Sackin index and let $k = \lceil \log_2(n) \rceil$. For any partition of n into two integers n_a, n_b , i.e. $n = n_a + n_b$, we use k_a and k_b to denote $\lceil \log_2(n_a) \rceil$ and $\lceil \log_2(n_b) \rceil = \lceil \log_2(n - n_a) \rceil$, respectively. Moreover, let*

$$f(n) = \begin{cases} 0 & \text{if } n \text{ is odd} \\ \binom{s(\frac{n}{2})+1}{2} & \text{else.} \end{cases}$$

Then, the following recursion holds:

- $s(1) = 1$
- $s(n) = \sum_{\substack{(n_a, n_b): \\ n_a + n_b = n, \\ n_a \geq \frac{n}{2}, \\ k_a = k-1, \\ k_b = k-1, \\ n_a \neq n_b}} s(n_a) \cdot s(n_b) + f(n) + s(n - 2^{k-2})$

Proof. First consider $n = 1$. In this case, it is clear that there is only one rooted binary tree, namely the one consisting of only one node, which therefore has minimal Sackin index, which implies $s(1) = 1$.

Before we continue with the recursion, let us analyze Algorithm 1 a little bit more in-depth. We know by Corollary 2 that if we want to count all trees with n leaves with minimal Sackin index, we only need to count the ones that

can be recovered by Algorithm 1. However, we have also already seen that all trees recovered by this algorithm have height $k = \lceil \log_2(n) \rceil$. Moreover, let T be such a tree recovered by Algorithm 1 and consider its standard decomposition $T = (T_a, T_b)$ with n_a and n_b leaves, respectively, such that $n = n_a + n_b$. As before, we can assume without loss of generality that $n_a \geq n_b$, i.e. that $n_a \geq \frac{n}{2}$. Then, Algorithm 1 will assure that the larger of the two subtrees has height $k - 1$. This is due to the fact that we start with T_k^{bal} , i.e. in the beginning both subtrees have height $k - 1$, and then we delete at most all but one cherry of maximum depth (because otherwise the algorithm would have started at T_{k-1}^{bal}). So the larger of the two subtrees, T_a , always keeps its initial height, which is $k - 1$, so we always have $k_a = k - 1$. Therefore, by Corollary 2, no trees with $k_a \neq k - 1$ can have minimal Sackin index.

Now consider T_b . For T_k^{bal} , both maximal pending subtrees have height $k - 1$, but then some cherries of maximal depth are deleted. This way, it may happen that the height k_b of T_b is at some stage less than $k - 1$ (given that the difference between n and 2^k is large enough). However, note that we always have $k_b \geq k - 2$ (and thus $n_b \geq 2^{k-2}$), because as soon as the height of T_b is only $k_b = k - 2$, there is no longer a cherry of maximum depth k of T to be found in T_b . So when the algorithm continues, it will never choose a cherry from T_b again, because all cherries of maximum depth would now be in T_a . So in total, we have $k_b \in \{k - 2, k - 1\}$.

So in summary, all trees T recovered by Algorithm 1, and thus all trees with a minimal Sackin index, have the property that $k_a = k - 1$ and $k_b \in \{k - 2, k - 1\}$ for $n = n_a + n_b$ and $n_a \geq \frac{n}{2}$. Trees with n leaves which do not have these properties can by Corollary 2 thus not have minimal Sackin index.

Moreover, we know that (as $\mathcal{S}(T) = \mathcal{S}(T_a) + \mathcal{S}(T_b) + n$ by definition) if $\mathcal{S}(T)$ is minimal, $\mathcal{S}(T_a)$ and $\mathcal{S}(T_b)$ must be minimal, too.

Now we want to calculate $s(n)$. Therefore, we consider all integer partitions of n into precisely 2 summands, i.e. $n = n_a + n_b$ for some $n_a, n_b \in \mathbb{N}$ such that $n_a \geq n_b$, i.e. $n_a \geq \frac{n}{2}$. Now we set $k := \lceil \log_2(n) \rceil$, $k_a := \lceil \log_2(n_a) \rceil$ and $k_b := \lceil \log_2(n_b) \rceil$. Note that as $n_a \geq \frac{n}{2}$, $k_a \geq \lceil \log_2(\frac{n}{2}) \rceil = \lceil \log_2(n) - \log_2(2) \rceil = \lceil \log_2(n) - 1 \rceil = \lceil \log_2(n) \rceil - 1 = k - 1$, so $k_a \in \{k - 1, k\}$.

We distinguish the only three possible cases:

1. $k_a = k - 1$ and $k_b = k - 1$ and $n_a \neq n_b$,
2. $k_a = k - 1$ and $k_b = k - 1$ and $n_a = n_b$,
3. $k_a = k - 1$ and $k_b = k - 2$,

1. Let n_a be such that $k_a = k - 1$ and $k_b = k - 1$ and $n_a \neq n_b$. Any tree with n leaves and minimum Sackin index and standard decomposition $T = (T_a, T_b)$ with $n = n_a + n_b$, $n_a \geq \frac{n}{2}$, $k_a = k - 1$, $k_b = k - 1$ and $n_a \neq n_b$ must have the property that T_a and T_b have minimum Sackin index, too, as $\mathcal{S}(T) = \mathcal{S}(T_a) + \mathcal{S}(T_b) + n$. This implies by Theorems 3 and 4 that $\mathcal{S}(T_a) = -2^{k_a} + n_a(k_a + 1)$ and $\mathcal{S}(T_b) = -2^{k_b} + n_b(k_b + 1)$. Using this together with $k_a = k - 1$ and $k_b = k - 1$ and $n_b = n - n_a$, we derive:

$$\mathcal{S}(T) = \mathcal{S}(T_a) + \mathcal{S}(T_b) + n = -2^k + n(k+1).$$

So $\mathcal{S}(T) = -2^k + n(k+1)$, which we know must hold for any tree T with n leaves and minimum Sackin index (again due to Theorems 3 and 4). So *every* combination of T_a and T_b with the properties that T_a and T_b have minimum Sackin index and $n = n_a + n_b$, $k_a = k - 1$, $k_b = k - 1$ and $n_a \neq n_b$ leads to a valid tree $T = (T_a, T_b)$ on n leaves, i.e. a tree which indeed has minimum Sackin index. So every such combination has to be considered. Thus, we sum over all those pairs (n_a, n_b) and consider for each of them all possibilities to combine choices from the $s(n_a)$ trees with n_a leaves with minimum Sackin index with the $s(n_b)$ trees with n_b leaves and minimum Sackin index. There are $s(n_a) \cdot s(n_b)$ such combinations. This explains the first sum of the recursion.

2. Now if n is even, there is another summand: If $k_a = k - 1$ and $k_b = k - 1$ and $n_a = n_b$, we can proceed as in Case 1 but here, we have to consider all combinations of 2 trees from the $s(n_a) = s(n_b)$ trees with $n_a = n_b$ leaves and with minimum Sackin index, and there are $\binom{s(n_a)-1+2}{2}$ such combinations (unordered sampling of two trees from $s(n_a)$ trees with replacement, as the two trees may be equal and as their order does not matter), see for instance [1, Equation (1.4.4)] for more details. This explains the second summand in the recursion.
3. Next, consider the case $k_a = k - 1$ and $k_b = k - 2$. Note that if $k_b = \lceil \log_2(n_b) \rceil = k - 2$, we have $n_b \leq 2^{k-2}$. But as explained above, for a tree recovered by Algorithm 1, we always have $n_b \geq 2^{k-2}$, so now we must have $n_b = 2^{k-2}$. By Theorem 5, this implies that the only choice for T_b with minimum Sackin index is T_{k-2}^{bal} . So for T_b , there is no alternative, but we can combine the only choice of T_b with all possible choices of T_a , of which there are $s(n_a) = s(n - n_b) = s(n - 2^{k-2})$ many. This explains the last summand in the above recursion.

As we have considered all cases and added all contributions of each possible integer partition of n , there is nothing more to show. This completes the proof. \square

Remark. Starting at $n = 1$ and continuing up to $n = 32$, the sequence $s(n)$ of numbers of trees with n leaves and with minimal Sackin index is 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 3, 5, 3, 3, 1, 1, 1, 4, 6, 14, 17, 27, 28, 35, 28, 27, 17, 14, 6, 4, 1, 1. We have calculated the values of $s(n)$ for up to $n = 1024$. These data can be found online at [4]. Note that this sequence is new to the Online Encyclopedia of Integer Sequences OEIS [9, Sequence A299037]; it has been submitted in the scope of this manuscript. It had previously not been contained in the OEIS, i.e. this sequence has so far apparently not occurred in any other context.

4. Discussion

In this manuscript, we first proved some results which are maybe not surprising as they have been stated in the literature before, but as they still had to

be proved, we delivered the mathematical arguments to back those results up. We also explicitly derived the sequences for the values $\mathcal{S}(T)$ (where T is a tree with n leaves) of the minimum and maximum Sackin indices for growing values of n , and showed that they are known sequences as they are already contained in OEIS. This might lead to future research as it connects the optimal Sackin index to other problems that can be found in the literature.

Subsequently and more importantly, we provided two algorithms which can be used to explicitly find all trees with minimal Sackin index even in the complicated case where $n \neq 2^k$, about which previously little had been known. We then used these algorithms to characterize trees with a minimal Sackin index and to derive a recursion for the sequence of numbers of such trees, which is new to the OEIS. A question for future research could be to find a closed formula or a generating function describing this sequence.

Another area of interest for future research might be the investigation of other well-known and frequently used balance indices like, for instance, the Colless index [3]. Moreover, the implications of our findings, i.e. of the number of extremal trees concerning the Sackin index, on evolutionary models and their induced probability distributions on the tree space are also of high interest.

Acknowledgements

The author wishes to thank Lina Herbst and Kristina Wicke for very helpful discussions on the general topic and for comments concerning an earlier version of the manuscript. The author also wishes to thank Mike Steel for very helpful discussions and some suggestions, in particular concerning a shorter version of the first part of the proof of Lemma 1 and a shorter version of the proof of Proposition 1.

Appendix

Here we state and prove some basic combinatorial and graph theoretic results that are helpful to understand the proofs in the main part of the manuscript.

Lemma 4. *Let T be a rooted binary tree with n leaves. Let $h = h(T)$ be the height of T , i.e. the number of edges on the longest path from the root to any of the leaves of T . Let $k \in \mathbb{N}$ be such that $n > 2^k$. Then we have: $h > k$.*

Proof. Recall that in a rooted binary tree T with node set V and n leaves, we have $|V| = 2n - 1$. So by assumption, we here have $|V| = 2n - 1 > 2 \cdot 2^k - 1 = 2^{k+1} - 1$, as we assume $n > 2^k$. Now let a_i denote the number of nodes in V with level i , i.e. the number of nodes in V whose shortest path to the root employs i edges. Note that then, a_0 is 1, because the root ρ itself is the only node with distance 0 to the root, and the node with maximal level has level h – it will be precisely one of the leaves whose depths defines the height of T . We now use the a_i values to derive a second bound for $|V|$, namely by summing up

over all possible levels: $|V| = \sum_{i=0}^h a_i$. Note that it can be easily seen that as T is binary, we have $a_i \leq 2^i$ for each $i = 0, \dots, h$, which leads to:

$$|V| = \sum_{i=0}^h a_i \leq \sum_{i=0}^h 2^i = 2^{h+1} - 1.$$

The last equality uses the geometric series property.

So, in total we conclude $2^{k+1} - 1 < |V| \leq 2^{h+1} - 1$. This implies $k < h$ and thus completes the proof. \square

Lemma 5. *Let T be a rooted binary tree with $n \geq 2^{k-1} + 1$ leaves for some $k \in \mathbb{N}_0$, and let u be a leaf of maximal depth δ_u in T , i.e. $\delta_u = h(T)$. Then, $\delta_u = h(T) \geq k$. In particular, if $n = 2^k$, we have $h(T) \geq k$.*

Proof. By Lemma 4, as $n \geq 2^{k-1} + 1 > 2^{k-1}$, we conclude that $h(T) > k - 1$ and thus $\delta_u > k - 1$. As $\delta_u \in \mathbb{N}$, we conclude $\delta_u \geq k$. Moreover, if $n = 2^k$, this implies $n \geq 2^{k-1} + 1$, so this completes the proof. \square

Lemma 6. *Let T be a rooted binary tree with $n \geq 1$ leaves. Let u be a leaf of maximal depth δ_u in T . Then, $\delta_u \leq n - 1$.*

Proof. Considering that a rooted binary tree with n leaves has $2n - 1$ nodes in total and therefore $n - 1$ internal nodes, the stated bound is obvious: The extreme case would be if a leaf had *all* internal nodes on its path to the root, which implies $\delta_u \leq n - 1$ for all leaves u of T . This completes the proof. \square

References

- [1] Ash, R. 2008. *Basic Probability Theory*. Dover Books.
- [2] Blum, M. and Francois, O. 2005. On statistical tests of phylogenetic tree imbalance: The Sackin and other indices revisited. *Mathematical Biosciences*, 195(2): 141 – 153.
- [3] Colless, D. 1982. Review of "Phylogenetics: the theory and practice of phylogenetic systematics". *Systematic Zoology*, 31: 100.
- [4] Fischer, M. 2018. Number of rooted binary trees with $n \leq 1024$ leaves and minimal Sackin index. <http://mareikefischer.de/SupplementaryMaterial/Sackin.txt>.
- [5] Heard, S. 1992. Patterns in tree balance among cladistic, phonetic, and randomly generated phylogenetic trees. *Evolution*, 46(6): 1818–1826.
- [6] Mir, A., Rossello, F., and Rotger, L. 2013. A new balance index for phylogenetic trees. *Mathematical Biosciences*, 241(1): 125 – 136.

- [7] Sackin, M. 1972. "good" and "bad" phenograms. *Systematic Zoology*, 21: 225.
- [8] Shao, K.-T. and Sokal, R. 1990. Tree balance. *Systematic Zoology*, 39(3): 266–276.
- [9] Sloane, N. 2018. The On-Line Encyclopedia of Integer Sequences OEIS. <https://oeis.org>.
- [10] Steel, M. 2016. *Phylogeny: Discrete and random processes in evolution*. CBMS-NSF Regional conference series in Applied Mathematics. SIAM.
- [11] Than, C. and Rosenberg, N. 2014. Mean deep coalescence cost under exchangeable probability distributions. *Discrete Applied Mathematics*, 174: 11–26.