

UNIVERSITY OF GREIFSWALD
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

UNIVERSITÄT GREIFSWALD
Wissen lockt. Seit 1456



Extremal values of the cherry index and the symmetry nodes index of binary rooted trees

BACHELOR THESIS

submitted in partial fulfillment of the requirements for the degree of Bachelor of
Science (B.Sc.) in Biomathematics by

Sophie Johanna Kersting

First supervisor: Prof. Dr. Mareike Fischer
Second supervisor: Prof. Dr. Volkmar Liebscher

Greifswald, 9th January 2019

Abstract

Effects like selection in evolution as well as fertility inheritance – the positive correlation of an individual's number of children and its number of siblings – in the development of populations lead to a higher degree of asymmetry in reconstructed trees than expected under a null hypothesis like neutrality or a classical Wright-Fisher population model.

To identify these influences numerous so-called balance indices were invented with various concepts on how to measure asymmetry in trees like considering the depth of the leaves or how inner nodes split the number of descending leaves.

In this bachelor thesis we take a closer look at two balance indices, the *cherry index* as well as the completely new *symmetry nodes index*. The first is using the cherries in a tree as a hint for symmetry and the latter one is taking into account the number of symmetry nodes. We analyze both indices by determining and proving exact formulas or sequences for the minimal and maximal values as well as characterizing the corresponding tree shapes and their numbers. Furthermore, we make first assumptions on the advantages and disadvantages of both indices in comparison to the popular Colless and Sackin indices.

Alongside this bachelor thesis scripts using the free programming language R are provided to calculate the cherry and symmetry nodes index value as well as the minimal symmetry nodes index value.

Table of Contents

1	Introduction	1
1.1	Importance of tree balance	1
1.1.1	Balance in phylogenetic trees	4
1.2	Preliminaries	7
1.3	Balance indices	11
2	The cherry index (CI)	15
2.1	Minimal value and number of minimal trees	15
2.1.1	Minimal trees	18
2.2	Maximal value and maximal tree	19
3	The symmetry nodes index (SNI)	21
3.1	Minimal value and number of minimal trees	21
3.1.1	Calculating first values using dynamic programming	21
3.1.2	First properties of the minimal value and minimal trees	24
3.1.3	Minimal trees and their value	26
3.1.4	Number of minimal trees	32
3.2	Maximal value and maximal tree	34
4	Discussion and Results	35
4.1	Cherry and symmetry nodes index as a function using the clade size	35
4.2	Comparison of the extremal values of CI and SNI	37
4.3	Comparison of different balance indices	37
4.4	Results	39
5	Appendix: R-Scripts for CI and SNI	41
5.1	MinSNIandNumbOfTrees.R	41
5.2	minSNI.R	42
5.3	CI.R	42
5.4	SNI.R	43
	References	44

1 Introduction

The importance of trees in evolutionary theory or also in genealogy as a tool of understanding and investigating relations between species or individuals of a population cannot be denied [3]. Reconstructing the "Tree of Life" is the goal of many scientists who want to explore how today's species are related. There are various methods about how to reconstruct a good fitting tree of e.g. genetic data of a set of species or a set of individuals. This leads to problems like how to combine trees on different taxa, how to deal with contradictions in the underlying data or how to insert a new species into a given tree.

Reconstructing a tree is one part of the problem, but then there is also the question on how to investigate it further. It was shown that effects like selection and fertility inheritance in the development of species and populations increase the asymmetry or imbalance of tree shapes [4, 17]. There are various concepts on how to measure the degree of symmetry in trees and they are called *balance indices*. Joseph Felsenstein used the term "measures of overall asymmetry" as many balance indices have higher values if the tree is more asymmetrical [6, p. 563].

From these indices a statistical test can be derived to detect whether or not a tree, that has been reconstructed from genetic data, is significantly more asymmetrical than expected under a null hypothesis. This null hypothesis can be neutrality – the natural decrease in genetic variability even under no influence of selection – in the context of evolution [15] or a classical population simulation model like the Wright-Fisher model in the context of population development [4]. An example of the use of balance indices in the latter case is presented in Section 1.1. The more problematic appliance of balance indices on phylogenetic trees is discussed as well.

In the main part of this bachelor thesis we explore two balance indices, the cherry index and the new symmetry nodes index. The first one counts how many leaves of the tree are not in a cherry and the latter one counts all interior leaves that are not symmetry nodes. Exact formulas or sequences for their minimal and maximal values are proven and the shape of the trees that have these extremal index values as well as their numbers are characterized.

Later in Chapter 4, we shortly discuss a different way to describe the indices and compare their extremal values. Moreover, we compare the cherry and the symmetry nodes index with the Sackin and the Colless index, two popular balance indices, briefly exploring if they could have new useful properties and if there are cases in which they perform not as desired.

Last but not least, we have a look on the R-scripts provided alongside this bachelor thesis that include functions to calculate the balance index value for both the cherry and the symmetry nodes index as well as the minimal symmetry nodes index value.

1.1 Importance of tree balance

Before we begin with the mathematical analysis of extremal values and trees, we look at the practical appliance of balance indices:

One example of the use of balance indices is the detection of fertility inheritance, the positive correlation between an individual's number of descendants and its number of siblings, the offspring of its parents. Fertility inheritance can be explored without balance indices using genealogical or demographic data with or without the help of genetic data [16]. In the article "Matrilineal Fertility Inheritance Detected in Hunter-Gatherer Populations Using the Imbalance

of Gene Genealogies" by Blum, Heyer, François and Austerlitz (2006) [4] however, the authors suggest a method using only genetic data by reconstructing the genealogy and measuring the imbalance of the resulting tree shape with the help of a balance index. In their research they applied it to several human populations to detect if fertility inheritance is more common in traditional hunter-gatherer populations or in food-producer populations.

This example and how the imbalance in trees is measured is shortly presented in this section to show the application of balance indices in a practical field. However, there is a difference to the balance indices that are dealt with in this bachelor thesis: The measure used here is also applicable to trees that are not fully resolved i.e. that have nodes with more than two outgoing edges. The cherry index could also be used on not fully resolved trees. The symmetry nodes index, however, depends on a binary rooted tree in which each interior node has exactly two outgoing edges.

To detect the imbalance in a tree $T = (V, E)$ the authors of the article used a method from Fusco and Cronk that was modified by Purvis et al for statistical reasons [10]: A value I' is calculated for every interior node $v \in \mathring{V}$ that is fully resolved (i.e. it has two outgoing edges) and that has more than 3 descendant leaves. Let the size of a subtree be the number of leaves in a subtree, implying that a subtree is called larger if it has more leaves than another subtree, then I' for a vertex v is defined as follows:

$$I'_v = \begin{cases} I_v & \text{if } n \text{ is even} \\ \frac{n-1}{n} \cdot I_v & \text{else} \end{cases} \quad \text{with} \quad I_v = \frac{B-m}{M-m}$$

with n being the size of the subtree rooted in v , B being the size of its larger daughter clade (pending subtree rooted in a child of v), $M = n - 1$ being the maximum value for B and $m = \lceil \frac{n}{2} \rceil$ being the minimum value for B . It measures how equally the leaves are split with 0 indicating the most equal split possible and 1 indicating the most uneven split.

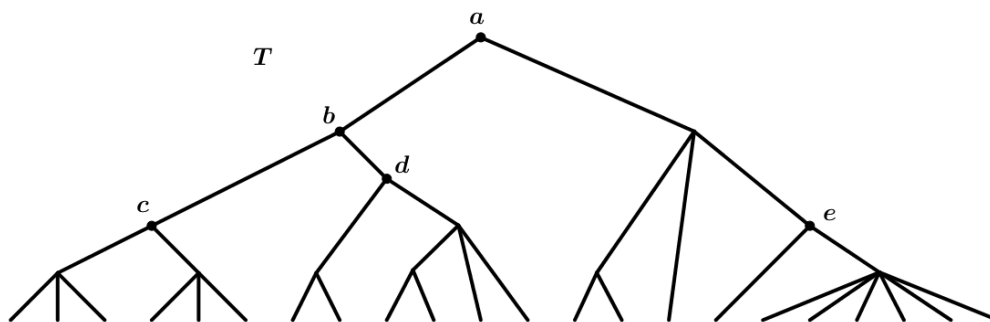


Figure 1: An example of a not fully resolved tree with 22 leaves.

An example tree with 22 leaves is depicted in Figure 1. The interior nodes in question are a ,

b , c , d and e . These are the corresponding I' values:

$$\begin{aligned} I'_a &= I_a = \frac{12 - 11}{21 - 11} = \frac{1}{10} = 0.1 \\ I'_b &= I_b = \frac{6 - 6}{11 - 6} = 0 \\ I'_c &= I_c = \frac{3 - 3}{5 - 3} = 0 \\ I'_d &= I_d = \frac{4 - 3}{5 - 3} = \frac{1}{2} = 0.5 \\ I'_e &= \frac{6}{7} \cdot I_e = \frac{6}{7} \cdot \frac{6 - 1}{6 - 1} = \frac{6}{7} \approx 0.857 \end{aligned}$$

This measure was applied to a tree reconstructed from genetic data of several individuals of a population. As the authors wanted to detect fertility inheritance in a population they used different models of population development. As a null hypothesis they chose the classical Wright-Fisher model [6, p. 410] because it does not include fertility correlation. This model describes the development of a population with constant size and for every individual in a new generation t its parent is randomly drawn from the parent generation $t - 1$ with replacement. In this model the expected value is $I' = 0.5$ for each node. The model for population development with fertility inheritance was a modified Wright-Fisher model in which each parent is not chosen randomly, but with a higher probability if he has more siblings.

The mean of all values I' was calculated and it was tested if they differed statistically significant from 0.5. A mean significantly larger than 0.5 was seen as evidence of fertility inheritance. The validity of this method was tested including power, robustness and influence of the phylogenetic method to reconstruct the tree.

In our example depicted in Figure 1 the *mean* I' is $\frac{0.1+0.5+0.857}{5} \approx 0.291$ indicating that this tree is more balanced than expected under the null hypothesis (significance not calculated) and that the corresponding hypothetical population would probably not have developed under the influence of fertility inheritance but under some opposite effect.

According to the article, this method was applied to samples of mitochondrial DNA (only maternally transmitted) of 37 human populations to study if fertility inheritance is more common in traditional hunter-gatherer populations or in food-producer populations. Every population tested contained at least 43 individuals. Their genealogies were reconstructed using a maximum-likelihood method and every population that did not contain at least 4 fully resolved interior nodes was discarded. In general, the number of resolved nodes varied noticeably for these data sets with a mean around 13, but a standard deviation of 10.

A one-sided Wilcoxon rank test revealed that the imbalance is significantly higher in hunter-gatherer populations with *mean* $I' = 0.74$ than in food-producer populations with *mean* $I' = 0.6$. Furthermore, the fraction of hunter-gatherer populations with a *mean* I' significantly higher than 0.5 is larger than in the other group.

This example shows that balance indices can produce valid results, but also that there are problems to be faced. Probably one of the biggest problems is that reconstructed trees can have vertices that are not fully resolved and thus contain less usable information. This can always happen if different individuals have similar DNA sequences, but the number of fully resolved and

therefore evaluable vertices correlates with the number of input sequences [4].

1.1.1 Balance in phylogenetic trees

The example from above leads to the question if balance indices can also be applied on phylogenetic trees that portray the relations of a set of species. The following paragraphs give a brief overview on this topic.

First, we have to ask if there is a similar effect as fertility inheritance for species. The children of inner nodes in phylogenies are not offspring but species with a common ancestor. Thus, imbalance could point to species that pass on the tendency to be more often affected by speciation events. This transfer of a speciation rate from parent to descendant species can happen in two ways:

It is possible that a species (e.g. a bacterial strain) has a genome with a higher mutation rate, maybe due to gene repairing mechanisms that work less efficient than in other species [2, p. 78]. This can lead to a more diverse gene pool, more variable phenotypes and possibly to the separation into two new species inheriting the less functioning repair mechanisms from their ancestor. These two descendant species can again be more likely to create new species.

However, there can be other possibilities that affect the rate at which speciation events happen in a line of species. For example the environment, that can either be very constant – favoring species that are the most adapted – or it can be very unstable, challenging species with constantly changing demands and therefore favoring species with a diverse gene pool as they more likely include individuals that can deal with the change. Such biotopes can be the deep sea in contrast to a group of islands. The latter one being the prime example for adaptive radiation as it happened with Darwin’s finches [12, p. 54] as well as population bottleneck [12, p. 26] through environmental events or if a small group of individuals colonizes a new island. In this version, the speciation rate is transferred to the subspecies as they are colonizing a similar habitat with similar demands.

In conclusion, the imbalance of a tree cannot point at only one reason for the evolutionary rate such as the genetic or the environmental aspect, but only at the complete set of combined effects at the most. Maybe, this set can be called selection as a whole. Since 1968 it is discussed if there is a mechanism called neutrality that results in evolutionary change even in the absence of selective pressure. Without selection, every individual has the same fitness as it is equally likely to produce offspring. This idea is portrayed by a classical Wright-Fisher model as explained in the example in Section 1.1. In such a model the genetic variability decreases with passing time and for infinite time it would result in a population that is homozygous regarding each gene. Nowadays this theory is accepted and serves as a null hypothesis. Selection as a force that increases evolutionary rate can only be assumed if neutrality is not sufficiently explaining the differences between species [15].

It has been detected that reconstructed trees of populations under the influence of selection are on average more asymmetrical than the ones with only neutrality. However, when evaluating the power of balance measures to identify trees based on selection and not on neutrality it was shown that they are less useful than other tests [17]. For example, simulations revealed that popular balance indices like the Colless index fail to detect asymmetry reliably: "Even when the speciation rates of sister clades differ by a factor of 3, there is only about a one in three chance

that a tree with 20 species will be significantly asymmetric using the most powerful statistic" [14]. Furthermore, it is questioned if balance indices can even detect differences in evolutionary rates within trees and not only bias. Two other sources for imbalance can be that the tree is incomplete or that the data from which it was reconstructed has low quality [27].

Reconstructed trees are normally incomplete due to at least two reasons: The leaves of a phylogenetic tree only represent species that still live in the present time. Hence, there is a lot of information missing about events that led to new species that died out until now. Furthermore, phylogenies are normally reconstructed from data about a set of species and should portray their relationship. Thus, it is not even guaranteed that all closely related species are considered. An example of how an incomplete set of species can affect the asymmetry in a tree and with that the interpretation of their evolution is depicted in Figure 2.

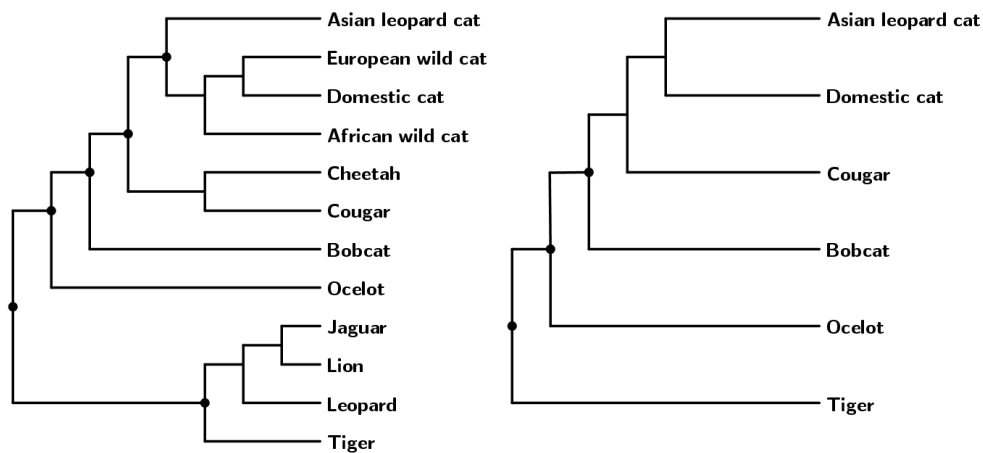


Figure 2: An already simplified version of a reconstructed phylogenetic tree of modern felidae with the 12 most well-known species (recreated from [13]) and another further modified version with an even more reduced set of species.

If we apply the same measure as before in the example in Section 1.1 on both trees, we get the following results (I' is calculated from top to bottom):

$$\begin{aligned}
 I' = I &= \frac{3-2}{3-2} = 1 \\
 I' = I &= \frac{4-3}{5-3} = \frac{1}{2} = 0.5 \\
 I' &= \frac{6}{7} \cdot I = \frac{6}{7} \cdot \frac{6-4}{6-4} = \frac{6}{7} \approx 0.857 \\
 I' = I &= \frac{7-4}{7-4} = 1 \\
 I' = I &= \frac{8-6}{11-6} = \frac{2}{5} = 0.4 \\
 I' = I &= \frac{3-2}{3-2} = 1
 \end{aligned}$$

$$\begin{aligned}
 I' = I &= \frac{5-3}{5-3} = 1 \\
 I' &= \frac{4}{5} \cdot I = \frac{4}{5} \cdot \frac{4-3}{4-3} = 0.8 \\
 I' = I &= \frac{3-2}{3-2} = 1
 \end{aligned}$$

The *mean I'* is $\frac{3+0.9+0.857}{6} \approx 0.793$ as well as $\frac{2+0.857}{3} \approx 0.933$ suggesting that both trees, but especially the modified tree, are highly unbalanced. The value of the left tree with a larger set of species still has a high, but substantially lower value. It is obvious that we could have also created a much more balanced tree with a different subset of species. This shows how important it is to test if these values are significantly higher than the expected value under the null hypothesis because the width of the rejection region depends on the number of species.

However, it can also be seen that phylogenies in contrast to reconstructed genealogies have the advantage that they are less likely to suffer under a large number of unresolved nodes as the genetic differences between species are more distinct than between individuals of the same population.

All in all, there is the influence of selection that could be detected with the help of balance indices as we also have neutrality as a null hypothesis, but it is still questionable how well balance indices can be applied to phylogenies.

1.2 Preliminaries

Before we look deeper into the cherry and the symmetry nodes index we need the basic definitions and theorems that are used throughout the next chapters. These definitions are common in literature. Here, we more closely refer to the following books:

- "Phylogeny: discrete and random processes in evolution" of M. Steel (2016) [28]
- "Phylogenetics" of C. Semple and M. Steel (2009) [22],
- "Inferring Phylogenies" of J. Felsenstein (2004) [6]

Graphs and (binary rooted) trees

The basic concept of trees and tree shapes are *graphs* $G = (V, E)$ that consist of a finite set of *vertices* or *nodes* $V \neq \emptyset$ and a set of *edges* $E \subseteq \{\{x, y\} \mid x, y \in V\}$. Furthermore, we only refer to *simple* (di-)graphs that have no parallel edges (E is not a multiset) and no loops, i.e. edges that lead from one vertex to itself. Due to the fact, that our main object that is investigated are rooted binary trees which have a natural direction induced by the root, we shortly discuss directed graphs. A directed graph (or *digraph*) $D = (V, E)$ has directed edges or arrows, this means an edge $e = (x, y)$ is an ordered pair of vertices with x as the source node and y as the target node [28, p. 3].

Two nodes u and v of a (di-)graph are *adjacent* if they are connected by an edge $e = \{u, v\}$ or $e = \{v, u\}$, in this case e is called *incident* to u and v .

The degree of a node $d(v) = |\{e \in E \mid e \text{ is incident to } v\}|$ is the number of edges that are incident to that node. For directed graphs it is split into the in-degree d^+ and out-degree d^- of a node which are the number of incoming or outgoing edges [28, p. 6].

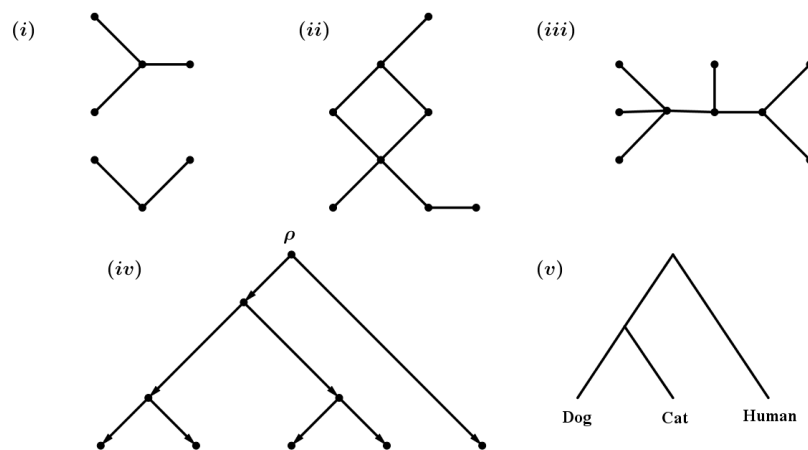


Figure 3: (i) A graph that is not connected, (ii) a graph that has a cycle, (iii) a tree, (iv) a rooted binary tree with root ρ and (v) a phylogenetic tree.

Trees $T = (V, E)$ are *connected* and *acyclic* graphs as shown in Figure 3 (iii) in contrast to (i) and (ii). The nodes of a tree can be divided into *inner* or *interior nodes* $\overset{\circ}{V}$ and leaves V^1 with degree one. A tree is called *binary* or trivalent if all inner nodes have degree 3. In a *rooted*

tree there is a vertex $\rho \in V$ distinguished as the root vertex of degree 2 [22, p. 7]. A rooted tree can be obtained by inserting the root as a new vertex into an edge of an unrooted tree. In a rooted binary tree every interior node has out-degree $d^- = 2$ and in-degree $d^+ = 1$ except the root that has only out-degree $d^- = 2$. The root gives a natural direction for the edges. All edges are directed away from the root and therefore the tree can be depicted in the common way with the root on top and leaves on the bottom. Due to that, the source node of an edge is referred to as the (direct) parent node and the target node as the (direct) child. Looking at a binary rooted tree as shown in Figure 3 (*iv*) the name "binary" gets clear as any inner node (the root ρ is also an inner node) has exactly two children or out-degree $d^- = 2$ [28, p. 6].

Phylogenetic trees and tree isomorphism

Because trees are often used to describe the process of evolution or ancestry in genealogy as in the example in Section 1.1 and because it is needed in Chapter 3, we have a short look on the concept of a phylogenetic tree or phylogeny [22, p. 19]. A *phylogenetic X-tree* $\mathcal{T} = (T, \phi)$ consists of a tree T that is called the *topology* or *tree shape* of \mathcal{T} and a bijection ϕ from the set of labels or species X to V^1 , the set of leaves of T . $RB(n)$ denotes the set of rooted binary phylogenetic X -trees with $X = \{1, \dots, n\}$. The number of such rooted binary trees is $|RB(1)| = 1$ and for $n \geq 2$ we have [22, p. 20]:

$$|RB(n)| = (2n - 3)!! = 1 \cdot 3 \cdot 5 \cdot \dots \cdot (2n - 5) \cdot (2n - 3) \quad (1.1)$$

In other words, the difference between a rooted binary tree and a phylogeny is that the leaves of the latter are labeled. Investigating tree shapes we ignore these labels [28, p. 41]. Two different phylogenetic trees that contain different information about the relation of the species can still have the same tree shape (for example the one in Figure 3 (*v*) and one that groups dogs and humans together).

Two tree shapes (or graphs in general) $T_1 = (V_1, E_1)$ and $T_2 = (V_2, E_2)$ are isomorphic if there is a bijection $f : V_1 \rightarrow V_2$ with $\{f(u), f(v)\} \in E_2 \iff \{u, v\} \in E_1$ [28, p. 4]. For rooted trees we also want $f(\rho_1) = \rho_2$.

Two phylogenetic X -trees are isomorphic if their tree shapes are isomorphic as rooted trees with a bijection f that is the identity map on the set of taxa X meaning $f(\phi(x)) = \phi(x)$ for all $x \in X$ [28, p. 9]. In the example in Figure 3 (*v*) we have a phylogenetic {Dog, Cat, Human}-tree that contains the information that dogs and cats are more closely related to each other (as they are grouped in a cherry) than they are to humans. In this example, we could swap the cat and dog thereby constructing an isomorphic phylogeny. The tree would still say they are more closely related.

Cherries and symmetry nodes

Regarding the balance indices that are explored in the next chapters, we need the following two definitions: If two leaves in a tree (or vertices of degree one in graphs) are both adjacent to a third node they are called a *cherry*. In a binary tree an interior node u with children u_1 and u_2 is called a *symmetry node* if its two pending subtrees are isomorphic. This decomposition of a binary tree into the two maximal rooted subtrees below vertex u is called the *standard*

decomposition [22, p. 21]. The direct children u_1 and u_2 will be the roots of the subtrees. Thus, the simplest symmetry node is the parent node of a cherry because the two pending subtrees are single leaves and therefore have the same tree shape. As depicted in Figure 4 the leaves u_1 and u_2 are adjacent to their parent u and form a cherry. u as well as the vertex marked with ρ_{bal} are examples for symmetry nodes.

Definition 1.1. The numbers of (disjoint) cherries and symmetry nodes in a rooted binary tree T are referred to with $c(T)$ and $s(T)$, respectively.

We know that $|V| = |E| + 1 \iff T$ is a tree and that every tree has a leaf. Let $T = (V, E)$ be a tree, then the following applies [28, p. 4]:

$$\begin{aligned} \text{If } d(v) \neq 2 \forall v \in V \text{ and } |V^1| \geq 3 \text{ then } T \text{ has at least two cherries} \\ \text{and } |V^1| \geq 4 \text{ then } T \text{ has at least two disjoint cherries} \end{aligned} \tag{1.2}$$

From Formula (1.2) about cherries in unrooted trees we can conclude:

$$\text{Every rooted binary tree with at least 2 leaves has at least one cherry.} \tag{1.3}$$

Proof. For $|V^1| = 2$ the introduction of the root into the unrooted tree will form the parent vertex of the cherry. For $|V^1| = 3$ we have as the unrooted tree a star tree with three cherries that pairwise share a leaf with each other. The insertion of the root in any edge leading from the inner node to a leaf does not affect the cherry formed by the other two leaves. Finally, for $|V^1| \geq 4$ we have two disjoint cherries in the unrooted tree and the root can only separate the leaves of one cherry, but not the other. \square

Let $T = (V, E)$ be a binary rooted tree with n leaves, then we have [28, p. 10]:

$$\begin{aligned} |V| = 2n - 1, \quad |\mathring{V}| = n - 1 \\ |E| = 2n - 2, \quad |\mathring{E}| = n - 2 \end{aligned} \tag{1.4}$$

Remark. *Unrooted binary trees with n leaves have one less (inner) node and edge than the rooted binary trees with n leaves.*

Caterpillar and fully balanced tree shapes

There are some tree shapes that deserve special attention: Let T_n^{cat} denote the so-called *caterpillar tree* with n leaves that is defined as the binary rooted tree shape that results from inserting a root in the edge of a cherry of an unrooted path graph with n leaves. The unrooted path graph or also unrooted caterpillar tree T_{ur}^{cat} has two interior nodes w and v that are adjacent to exactly one other interior node (the parent nodes of cherries) and all other interior nodes $\mathring{V} \setminus \{w, v\}$ are adjacent to exactly two interior nodes (see Figure 4 (ii)). T_n^{cat} is the unique tree shape with only one cherry (the proof can be found in Section 2.2).

The *fully balanced tree* T_k^{bal} is only defined for $n = 2^k$ with $k \in \mathbb{N}$ and every inner node is a symmetry node and splits the number of descendant leaves in half. Thus, this tree shape is the

most symmetrical and every leaf has depth $\log_2(n) = k$ with *depth* $\delta(x)$ defined as the number of edges on the unique path from the root to the leaf x [22, p. 22]. Examples for both tree shapes and their construction from unrooted trees for (i) $n = 4$ and for the caterpillar for (ii) n arbitrary can be seen in Figure 4.

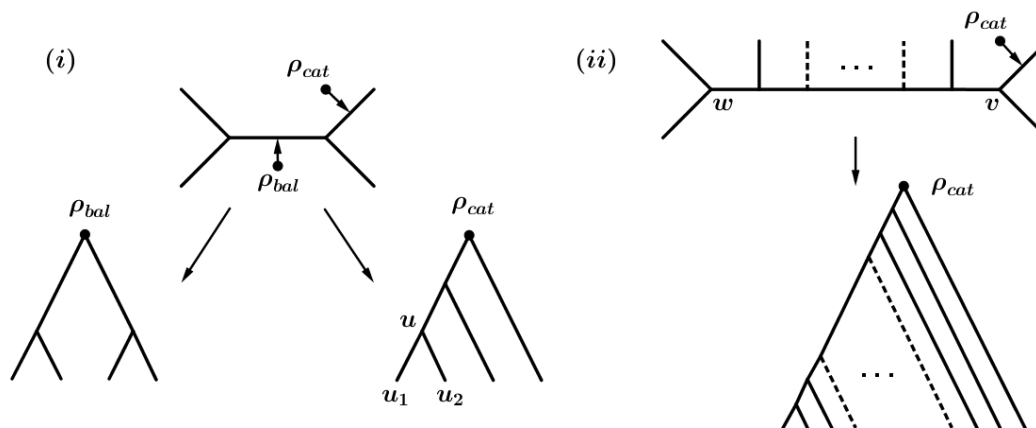


Figure 4: (i) Different roots on the same unrooted tree lead to the fully balanced tree and the caterpillar, the only possible tree shapes for binary rooted trees with four leaves. (ii) Construction of the caterpillar from a path graph with arbitrary number of leaves.

Wedderburn–Etherington numbers and the Newick tree format

Furthermore, we want to look at the *Wedderburn–Etherington numbers* [24, 29], an integer sequence named for Ivor M. H. Etherington and Joseph Wedderburn that can be used to count certain kinds of binary trees including the number of rooted binary tree shapes with n leaves. For example, we have $WE(4) = 2$ meaning that there are only two different tree shapes with 4 leaves (see Figure 4 (i)). The sequence starts with $a_1 = 1$ and is defined recursively by the equations below.

$$a_{2n-1} = \sum_{i=1}^{n-1} a_i a_{2n-i-1}$$

$$a_{2n} = \frac{a_n (a_n + 1)}{2} + \sum_{i=1}^{n-1} a_i a_{2n-i}$$

Let $WE(n) = a_n$ denote this number. The first numbers of the sequence starting with $n = 1$ are [6, p. 30]:

$$1, 1, 1, 2, 3, 6, 11, 23, 46, 98, 207, 451, 983, 2179, 4850, 10905, 24631, 56011, \dots$$

The Wedderburn–Etherington numbers $WE(n)$ can also be interpreted as the number of possible ways to insert parentheses in the term x^n when multiplication is commutative but not associative [24]. For example we have $WE(5) = 3$ with $x(x(x(xx)))$, $x((xx)(xx))$ and

$(x(xx))(xx)$. This interpretation is directly connected to trees:

The *Newick tree format* uses the "correspondence between trees and nested parentheses" [7]. For example the two trees (iv) and (v) in Figure 3 can be described by the following expressions in the Newick tree format: (iv) " $((,)(,) ,)$;" for a tree with no labels and (v) " $((Dog, Cat), Human)$;" for a phylogeny. The expressions for a tree always end with a semicolon and an inner node is represented by two matching parentheses. The leaves or their labels are always divided with a comma. The Newick format expression is commutative similar to tree isomorphism: $((Dog, Cat), Human)$; equals $(Human, (Cat, Dog))$;. Therefore, there is a one-to-one mapping up to isomorphism [7] between the Newick format, the multiplicative expressions x^n with nested parentheses (that can be interpreted as a general Newick tree format by inserting the semicolon, the commata as well as outer brackets) and tree shapes with n leaves, please refer to Figure 5.

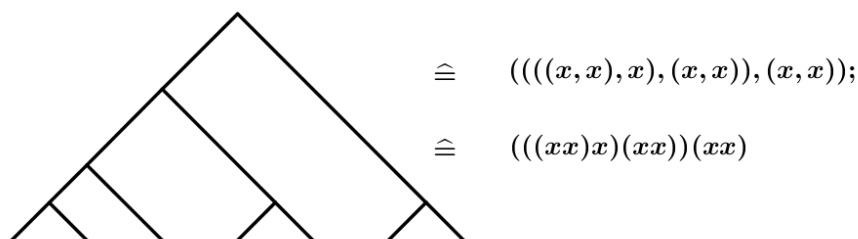


Figure 5: An example for the connection between tree, Newick tree format and the commutative multiplicative expression.

1.3 Balance indices

Throughout this bachelor thesis we will group tree shapes by their number of leaves $n = |V^1| \in \mathbb{N}$. Let M_{RB} denote the set of rooted binary trees (or tree shapes) and $M_{RB,n}$ with $n \in \mathbb{N}$ the subsets containing all trees with n leaves. A balance index for binary rooted trees can be defined as a function $\phi : M_{RB} \rightarrow [0, \infty)$ with the following properties:

- ϕ should measure some aspect of symmetry in tree shapes, with 0 or low values indicating a high degree of symmetry
- The fully balanced tree is the most symmetrical tree shape and therefore it should have the (preferably unique) minimal value (with $n = 2^k$ with $k \in \mathbb{N}$):

$$\phi(T_k^{bal}) = \min_{T \in M_{RB,n}} \phi(T)$$

- The caterpillar is commonly referred to as the most asymmetrical tree shape and should have the (preferably unique) maximal value:

$$\phi(T_n^{cat}) = \max_{T \in M_{RB,n}} \phi(T)$$

The first item in this list is admittedly very vague, but the various balance indices that already exist show that there are different ways to measure symmetry in trees and they could all have their use. It is a current research topic to investigate balance indices by, for example, determining their extremal values or their distribution under different tree simulation models which would make it possible to create statistical tests that can determine if a tree is significantly more asymmetrical than expected under a null hypothesis [6, p. 564]. It should be validated how useful a balance index is and if it has a field of application especially because methods perform differently depending on the way the imbalance was created: Trees can show "different imbalance signatures, i.e., different patterns of imbalance at different depths in the phylogeny" [1].

Additionally to the balance measure presented in 1.1 here are some examples of already existing balance indices for binary rooted trees to see the different approaches to symmetry in tree shapes:

Colless index The Colless index [28, p. 53] compares for every inner vertex v how many leaves are descendants of either v_1 or v_2 the direct children of v . This seems very intuitive as it checks how evenly distributed the leaves are. Let $\kappa(w)$ denote number of descendant leaves of an interior vertex w then the Colless index is defined as follows:

$$\mathfrak{C}(T) := \sum_{v \in \hat{V}(T)} |\kappa(v_1) - \kappa(v_2)|$$

Sackin index The idea of the Sackin index [28, p. 53] is to look at the number of descendant leaves $\kappa(y)$ of an inner vertex y or in an equivalent definition at the depth $\delta(x)$ of leaf x [8].

$$\mathfrak{S}(T) := \sum_{y \in \hat{V}(T)} \kappa(y) = \sum_{x \in V^1(T)} \delta(x)$$

Total cophenetic index Considering a tree T with n leaves this index calculates the sum of the depths of the lowest common ancestor $lca_T(v, w)$ over all pairs of different leaves v and w [18].














$$\mathfrak{T}(T) := \sum_{1 \leq v < w \leq n} \delta(lca_T(v, w))$$

Quartet index This index is also applicable on non-binary trees. It calculates the sum over the value $f(v_1, v_2, v_3, v_4)$ for all quartets (4-tupels) (v_1, v_2, v_3, v_4) of leaves of a tree T with n leaves. The value $f(v_1, v_2, v_3, v_4)$ quantifies the level of symmetry of the subtree obtained by restricting T to the leaves of the quartet. f is required to increase with the number of automorphisms on this subtree as they indicate symmetry [5].

$$\mathfrak{Q}(T) := \sum_{1 \leq v_1 < v_2 < v_3 < v_4 \leq n} f(v_1, v_2, v_3, v_4)$$

In the following chapters, we look at the extremal values and trees of two balance indices: First, the **cherry index** $CI(T)$ that uses cherries as an aspect of symmetry in trees. One version of this index is to count the cherries and it is already known [28, p. 58]. A tree like the fully balanced tree that contains many cherries would have a higher index value than for example

Table 1: Values of both the cherry and symmetry nodes index for simple tree shapes

n	$WE(n)$	tree shapes T	$CI(T)$	$SNI(T)$
1	1	•	1	0
2	1		0	0
3	1		1	1
4	2		0	0
			2	2
5	3		1	1
			1	2
			3	3
6	6		0	1
			2	2
			2	2
			2	3
			2	3
			4	4

2 The cherry index (CI)

The basic idea for the index was mentioned by M. Steel in his book "Phylogeny: discrete and random processes in evolution" [28] where he explains the advantage of looking at cherries. Even though the number of cherries "is not a particularly discriminating measure of tree shape" [28, p. 58], the positive aspect is that the number of cherries is robust to the position of the root: If we have an unrooted tree, the introduction of a root in any edge does either not change the number of cherries or separate the leaves of only one cherry if the root is placed in one of its edges. For example in Figure 4 ρ_{cat} is located in a cherry of the unrooted tree and therefore reduces the number of cherries by one. The introduction of ρ_{bal} on the other hand has no effect on the cherries.

Let $c(T)$ denote the number of (different) cherries. In order to stay close to our definition of a balance index that rewards a high level of symmetry with low values, we do not use the leaves in the cherries directly but their complement:

Definition 2.1. The *cherry index* of a binary rooted tree T with n leaves is defined as the number of leaves that are not in a cherry:

$$CI(T) := |V^1| - 2 \cdot c(T) = n - 2 \cdot c(T)$$

This obviously satisfies $CI(T) \geq 0$. Some examples can be found in Figure 6, Table 1 or in Figure 7 below.

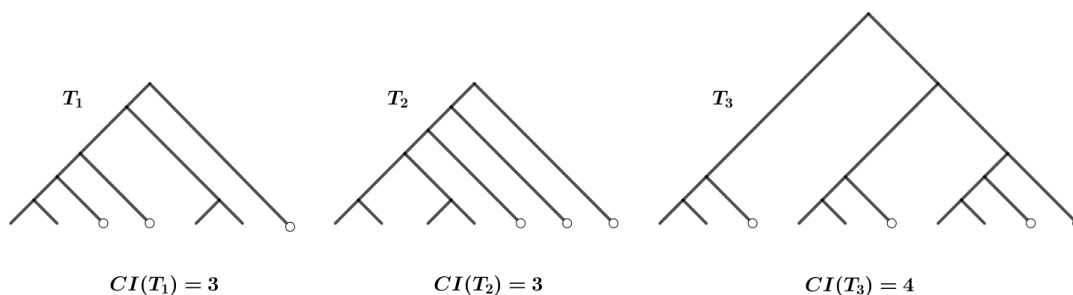


Figure 7: Examples for the cherry index. The leaves that are not in a cherry are marked.

2.1 Minimal value and number of minimal trees

First, we look at the minimum value of the cherry index. For that, we need trees that have at the most one single leaf and the rest has to be in cherries.

Theorem 2.1. For the minimal value $\min_{CI}(n)$ of the cherry index for all rooted binary trees with $n \geq 1$ leaves we have :

$$\min_{CI}(n) = \begin{cases} 0 & \text{if } n \text{ is even} \\ 1 & \text{if } n \text{ is odd} \end{cases}$$

Proof. If n is an even number, a rooted binary tree T with $CI(T) = 0$ can be constructed by any tree shape with $\frac{n}{2}$ leaves and $\frac{n}{2}$ cherries attached to its leaves. (For $n = 2$ this results in a single cherry.)

If n is odd, the leaves cannot be completely split up into cherries because there cannot be more than two leaves in a cherry in a binary tree. Therefore, the cherry index has to be ≥ 1 . For $n = 1$ there is only one possible and minimal tree shape, the single leaf. For $n \geq 3$ we can construct a minimal tree based on a minimal tree for $n - 1$ leaves by attaching one additional leaf to any edge. \square

Now we want to look at the number of trees with n leaves that obtain a minimal value. For an even number of leaves, we can use the idea of the proof above. As we use the idea of a smaller tree shape with leaves and cherries attached, we will call it the top tree.

Theorem 2.2. *Let n be even. The number of tree shapes with n leaves and minimal cherry index value can be calculated with the Wedderburn-Etherington numbers as follows:*

$$|\{T \in M_{RB,n} | CI(T) = \min_{CI}(n) = 0\}| = WE\left(\frac{n}{2}\right)$$

Proof. For $CI(T)$ to be minimal there must not be a single leaf and as we are considering a binary tree, every parent node of a cherry can only be adjacent to two leaves. Therefore, we need exactly $\frac{n}{2}$ cherries. The way they are connected does not matter for the cherry index. Thus, we can use any tree shape with $\frac{n}{2}$ leaves as a top tree. The Wedderburn-Etherington number $WE\left(\frac{n}{2}\right)$ gives us the number of all possibilities and with that the number of different minimal trees. \square

In the case of n being odd, the calculation is a bit different. In Figure 8 the approach is depicted. We follow the idea of the proof of Theorem 2.2: construct a minimal tree for $n + 1$ and delete a cherry. This means we use a tree shape with $\lceil \frac{n}{2} \rceil$ leaves with cherries attached and one cherry is removed.

In the picture, we see the two possible top trees T_4^{cat} as well as T_2^{bal} , but there are four and not two possible minimal trees because we get different trees depending on the choice which cherry is deleted in the caterpillar tree.

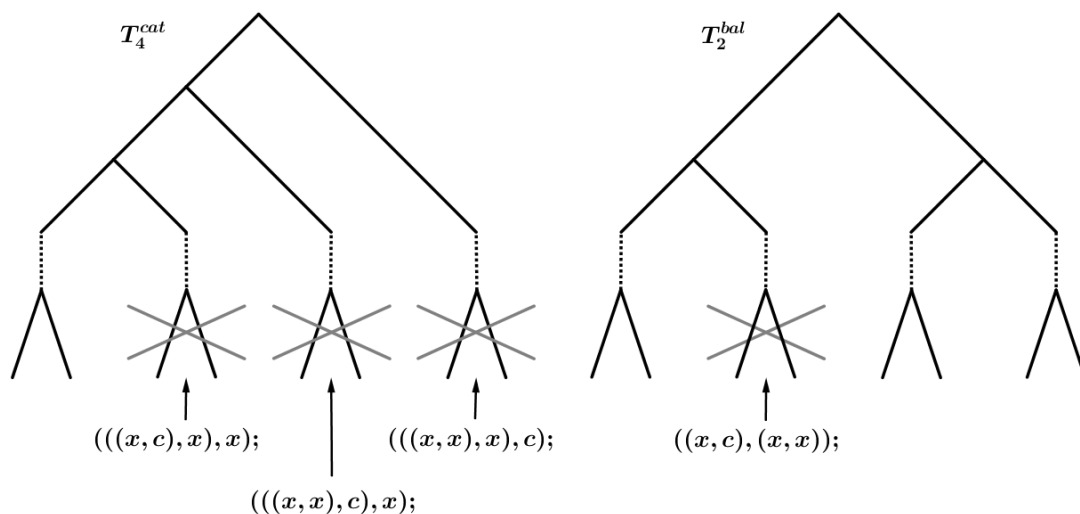


Figure 8: Visualization of the approach for $n = 7$. As the top tree, we use a tree shape with $\lceil \frac{n}{2} \rceil = 4$ leaves with cherries attached and one cherry is removed. For all possible minimal trees the corresponding expressions in the Newick format for the top tree are given with x denoting a leaf that has a cherry attached and c denoting the leaf without a cherry.

The Wedderburn–Etherington numbers $WE(n)$ correspond to the different ways to insert parentheses in x^n with multiplication being commutative, but not associative. The resulting expressions can be interpreted as trees in the Newick tree format by inserting outer parentheses, commata and the semicolon. Similar to $WE(n)$, we now want to find the number of different ways to insert parentheses in the term $x^n \cdot c$ with x denoting a leaf that has a cherry attached and c denoting the leaf without a cherry. As with $WE(n)$, there is a bijection between the commutative expressions, the corresponding Newick tree formats and the tree shapes.

The sequence exists, but unfortunately there is no explicit term for its calculation.

Definition 2.2. Let $A(n)$ denote this sequence of numbers of different ways to insert parentheses in the term $x^n \cdot c$ with multiplication being commutative, but not associative [26]. Starting with $n = 0$ the first numbers are:

$$1, 1, 2, 4, 9, 20, 46, 106, 248, 582, 1376, 3264, 7777, 18581, 44526, 106936, 257379, 620577, \dots$$

Theorem 2.3. Let n be odd. Regarding the number of tree shapes with n leaves and minimal cherry index value we have:

$$|\{T \in M_{RB,n} | CI(T) = \min_{CI}(n) = 1\}| = A\left(\frac{n-1}{2}\right)$$

Proof. The top tree has $\frac{n+1}{2}$ leaves of which $\frac{n+1}{2} - 1 = \frac{n-1}{2}$ can have cherries attached. This corresponds to the different multiplicative expressions of $x^{\frac{n-1}{2}} \cdot c$ as given by $A\left(\frac{n-1}{2}\right)$. \square

All in all, for the cherry index the sequence of numbers of minimal tree shapes with n leaves is a combination of $WE(n)$ and $A(n)$ as depicted in Table 2.

Table 2: Number of tree shapes with n leaves and minimal cherry index value

n		1	2	3	4	5	6	7	8	9	10	11	12	13	14
$min_{CI}(n)$	$A\left(\frac{n-1}{2}\right)$	1		1		2		4		9		20		46	
	$WE\left(\frac{n}{2}\right)$		1		1		1		2		3		6		11

2.1.1 Minimal trees

The construction of minimal trees as depicted in Figure 8 also shows how minimal tree shapes look like. The cherry index is only defined by the cherries, the tree shape on top can be arbitrary. Therefore the minimal trees are not unique, not even in the case $n = 2^k$ with $k \in \mathbb{N}$ in which for example the Sackin index has the unique minimal tree shape T_k^{bal} [8]. This is already observable for $n = 8 = 2^3$ as depicted in Figure 9. Generally the minimal tree shape is only unique for $n = 1, 2, 3, 4, 6$ because in all other cases are $WE\left(\frac{n}{2}\right)$ or $A\left(\frac{n-1}{2}\right)$ greater than 1 (see Table 2).

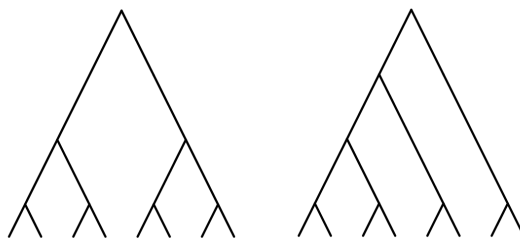


Figure 9: For $n = 8$ there are two tree shapes with minimal cherry index value. Both trees are built from four cherries and a tree shape with four leaves (T_2^{bal} and T_4^{cat}).

2.2 Maximal value and maximal tree

A tree shape has the maximal cherry index value if there are as few as possible leaves in cherries. The case of a single leaf is trivial as there is only one possible tree shape with $CI(T) = 1 = n$. For all other numbers of leaves we can state the following:

Theorem 2.4. *The unique maximal value of the cherry index is*

$$\max_{CI}(n) = (n - 2)$$

for a binary rooted tree with $n \geq 2$ leaves. The unique maximal tree shape is the caterpillar tree.

Proof. From Formula (1.3) we know that every rooted binary tree shape with at least two leaves has a cherry. Thus, the maximal value for binary rooted trees with $n \geq 2$ leaves is less or equal $|V^1| - 2 = n - 2$ for the two leaves in the cherry. This upper limit is met by the caterpillar tree shape T_n^{cat} that has exactly one cherry.

Uniqueness of the maximal tree shape: Here, we show that the binary rooted tree shape T_n^{cat} is the only tree shape with exactly one cherry. For $n = 2, 3, 4$ this is trivial, in the case of $n \geq 5$ we can argue with the unrooted version of the caterpillar tree – a binary path graph (for reference see Figure 10):

The unrooted caterpillar tree T_{ur}^{cat} has two interior nodes w and v that are adjacent to exactly one other interior node (the parent nodes of the cherries) and all other interior nodes $\mathring{V} \setminus \{w, v\}$ are adjacent to exactly two interior nodes. It is the unique unrooted binary tree shape with only two cherries because if we assume there is a binary tree $T \neq T_{ur}^{cat}$ with more than 4 leaves and exactly two disjoint cherries then we can again let w and v denote the parent nodes of the cherries. Then, as $|\mathring{V}| = n - 2 \geq 5 - 2 = 3$ (see Formula (1.4)) and $T \neq T_{ur}^{cat}$, there is (at least) one inner vertex z with $w \neq z \neq v$ that is not adjacent to exactly two inner nodes. If z is adjacent to two leaves it would be the parent node of a cherry which would be a contradiction to our assumption. As trees are connected graphs the only other case is that z is adjacent to three inner vertices. Now, consider the subtree T_z that includes z as a leaf and is obtained by cutting the two incident edges of z that are in the direction of w or v . Then T_z contains at least one inner node y adjacent to z and (including z) at least 3 leaves because $n_z = |\mathring{V}_z| + 2 \geq 1 + 2 = 3$. We can use Formula (1.2): y is a parent node of a cherry if $n_z = 3$ and if $n_z \geq 4$ there are at least two disjoint cherries of which one at least does not include z . All cases lead to a contradiction and it follows that T_{ur}^{cat} is unique with two cherries.

All rooted trees can be obtained by inserting a root on an edge of an unrooted tree. It is obvious that the insertion of a root can only break up one cherry. Because T_{ur}^{cat} is the unique tree with only two cherries and the rooted version of the caterpillar is obtained by introducing a root on an edge of any of the two cherries of the unrooted caterpillar, it follows that T^{cat} is unique as well. \square

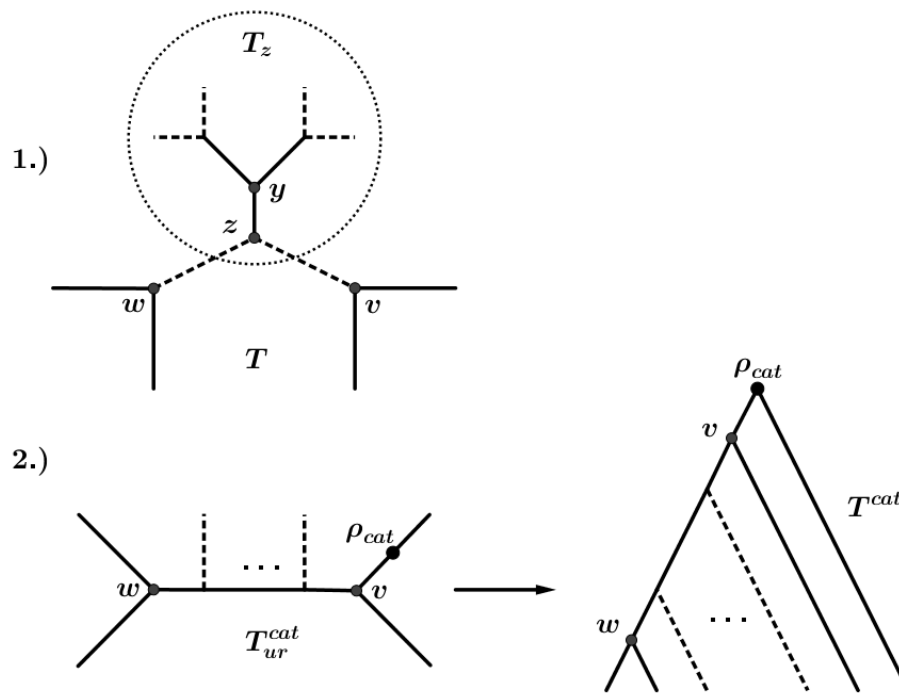


Figure 10: 1.) Depiction of T and T_z ; 2.) Construction of a rooted caterpillar tree by introducing a root in one of the two cherries of an unrooted caterpillar tree.

3 The symmetry nodes index (SNI)

Definition 3.1. The *symmetry nodes index* of a binary rooted tree T with n leaves is defined as

$$SNI(T) := (n - 1) - s(T)$$

with $s(T)$ being the number of symmetry nodes in T . In other words the index is the number of inner nodes that are not symmetry nodes.

As symmetry nodes are inner nodes and the number of inner nodes in a binary rooted tree with n leaves is $(n - 1)$, the symmetry nodes index is always greater than or equal to zero. Furthermore, it is clear that the tree shape T^{bal} that has only symmetry nodes as interior nodes has symmetry nodes index value of 0.

Some examples can be found in Figure 6, Table 1 or in Figure 11 below. In the latter one the same tree shapes are depicted as in Figure 7 with the examples for the cherry index to be comparable.

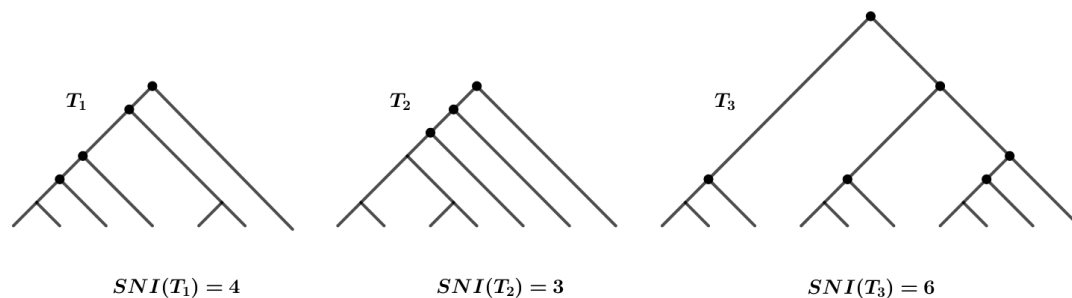


Figure 11: Examples for the cherry index. The interior vertices that are not symmetry nodes are marked.

3.1 Minimal value and number of minimal trees

Remark. *Minimal trees in this chapter are always considered being minimal regarding the symmetry nodes index.*

3.1.1 Calculating first values using dynamic programming

To calculate the minimal value of the symmetry nodes index $min_{SNI}(n)$ depending on the number of leaves n we can use dynamic programming that uses information on subproblems to generate the solution for the main problem. Here, we use the fact that all trees investigated here are rooted and the root always divides the whole tree into two pending subtrees. This is called the standard decomposition [22, p. 21]. In the simplest case, the root only separates one single leaf from the rest. The idea is to search for the minimal symmetry nodes index value by looking at all possible pairs of subtrees and their minimal values. This is depicted in Figure 12.

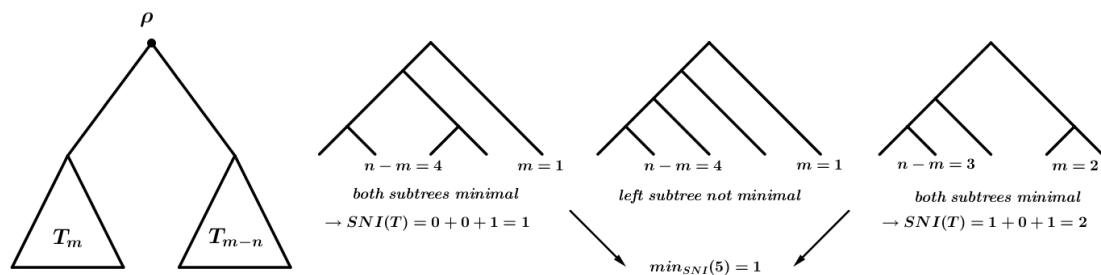


Figure 12: The idea of the algorithm as well as the three possible tree shapes with 5 leaves and their subtrees showing the calculation of $\min_{SNI}(5)$ following the second equation in (3.1).

As starting values we have the single leaf with $\min_{SNI}(1) = 0$ and the single cherry with $\min_{SNI}(2) = 0$. We look at $m = 1, 2, \dots, \lfloor \frac{n}{2} \rfloor$ with T_m being the "smaller" subtree of the standard decomposition i.e. with the smaller or equal number of leaves. To calculate $\min_{SNI} = \min \{c_m | m \in \{1, 2, \dots, \lfloor \frac{n}{2} \rfloor\}\}$ we can use the following equations:

$$\begin{aligned}
 c_m &= 2 \cdot \min_{SNI}(m) && \text{if } n \text{ is even and } 2m=n \\
 c_m &= \min_{SNI}(m) + \min_{SNI}(n-m) + 1 && \text{else}
 \end{aligned} \tag{3.1}$$

The first equation deals with the special case that both subtrees have the same size and therefore equal minimal index value. It follows that the whole tree can only be minimal iff both subtrees have the same minimal tree shape. Otherwise, the root would not be a symmetry node and would increase the value by 1. Therefore, we can just count $\min_{SNI}(m)$ two times. The second equation covers all other cases: The subtrees have different numbers of leaves and therefore different tree shapes. Hence, the root cannot be a symmetry node and increases the index value by 1.

The number of possible minimal tree shapes with n leaves increases for every possible minimal decompositions T_m and T_{n-m} with $m \in [1, 2, \dots, \lfloor \frac{n}{2} \rfloor]$. Let $\text{numbtrees}(n)$ denote the number of possible minimal tree shapes with n leaves. Again in the special case $2m = n$, the tree shape has only the minimal index value iff both subtrees have the same tree shape. Therefore, we have to count the number of possible minimal tree shapes of T_m only once. In all other cases, the number is given by the product of the possibilities of both subtrees as the root can never be a symmetry node. In short, the number of minimal trees can be calculated by the following commands written in pseudo-code derived from R for every $m \in [1, 2, \dots, \lfloor \frac{n}{2} \rfloor]$ that fulfills $c_m = \min_{SNI}$ with $\text{numbtrees}(n)$ being initialized with 0:

$$\begin{aligned}
 \text{numbtrees}(n) &\leftarrow \text{numbtrees}(n) + \text{numbtrees}(m) && \text{if } n \text{ is even and } 2m=n \\
 \text{numbtrees}(n) &\leftarrow \text{numbtrees}(n) + \text{numbtrees}(m) \cdot \text{numbtrees}(n-m) && \text{else}
 \end{aligned}$$

The following results in Table 3 were obtained with the script `MinSNIandNumOfTrees.R` using the programming language R (see Section 5.1). It shows the minimal value dependent on the number of leaves as well as the number of minimal tree shapes.

Table 3: Minimal value of the symmetry nodes index and number of minimal trees

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$min_{SNI}(n)$	0	0	1	0	1	1	2	0	1	1	2	1	2	2	3
$\#min.trees$	1	1	1	1	1	1	3	1	1	1	3	1	3	3	15
n	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
$min_{SNI}(n)$	0	1	1	2	1	2	2	3	1	2	2	3	2	3	3
$\#min.trees$	1	1	1	3	1	3	3	15	1	3	3	15	3	15	15

Additionally to the values and numbers of minimal trees of the symmetry nodes index all possible numbers of leaves ("size") m and $n - m$ of the two subtrees can also be explored. A few examples are depicted in Table 4. It is apparent that $m = \max\{k \in \mathbb{N} | 2^k \leq n\} = \lfloor \log_2(n) \rfloor$ is always a possible number of leaves of a subtree.

Table 4: Possible subtree sizes for a minimal tree with n leaves

n	possible m 's and $(n - m)$'s	n	possible m 's and $(n - m)$'s
1	0, 1	11	1, 2, 3, 8, 9, 10
2	1, 1	12	4, 8
3	1, 2	13	1, 4, 5, 8, 9, 12
4	2, 2	14	2, 4, 6, 8, 10, 12
5	1, 4	15	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14
6	2, 4	16	8, 8
7	1, 2, 3, 4, 5, 6	17	1, 16
8	4, 4	18	2, 16
9	1, 8	19	1, 2, 3, 16, 17, 18
10	2, 8	20	4, 16

3.1.2 First properties of the minimal value and minimal trees

These results lead to the assertion $\min_{SNI}(n) = wt(n) - 1$ because the first calculated values coincide with this already existing sequence [23, 25]: $wt(n) - 1$. $wt(n)$ denotes the so called binary weight of n and counts the 1's in the binary extension of n :

$$wt(n) = \sum_{i=0}^N a_i \text{ with } n = \sum_{i=0}^N a_i 2^i = (a_N, a_{N-1}, \dots, a_2, a_1, a_0)_2 \tag{3.2}$$

and $N \in \mathbb{N}_0, a_N = 1, a_0, \dots, a_{N-1} \in \{0, 1\}$

To illustrate the idea for this sequence we can look at Figure 13. In the example we use $n = 13 = 8 + 4 + 1 = (1101)_2$. We can construct the subtrees T_3^{bal}, T_2^{bal} and T_0^{bal} . Now, we can connect two of them with a temporary root and then the resulting tree again with the last subtree. The complete tree T_1 has $SNI(T_1) = 2 = 3 - 1 = wt(13)$ because the root and the temporary root are the only vertices that are not symmetry nodes.

Let $n = (a_N, a_{N-1}, \dots, a_2, a_1, a_0)_2$ be an arbitrary positive integer with N and a_i defined as above in (3.2). $I = \{i \in \{0, \dots, N\} \mid a_i = 1\}$ is the index set of the 1's and $|I| = wt(n)$. We can connect the subtrees $(T_i^{bal})_{i \in I}$ successively with (temporary) roots. Again, only these $wt(n) - 2$ temporary roots as well as the final root are no symmetry nodes and result in $SNI(T_2) = wt(n) - 2 + 1 = wt(n) - 1$. In other words we always need one less asymmetrical connection than we have balanced subtrees that directly refer to the 1's in the binary extension.

Remark. *Of course the order of the $(T_i^{bal})_{i \in I}$ does not matter for the index value. Nevertheless, they create different tree shapes. Their number is discussed in Chapter 3.1.4.*

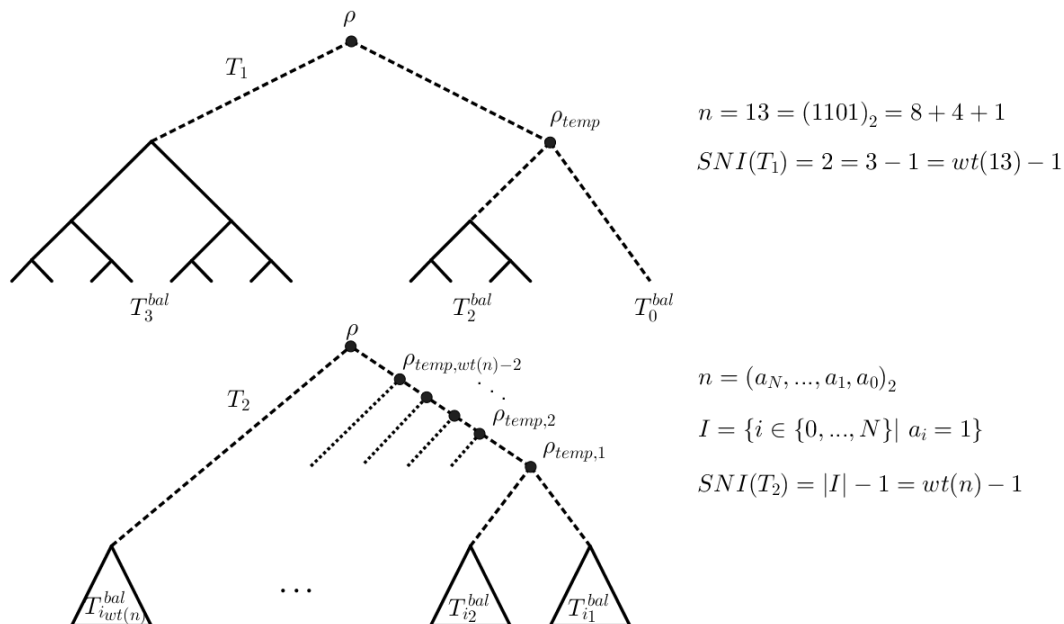


Figure 13: Construction of a tree shape T with $SNI(T) = wt(n) - 1$.

To finally prove $\min_{SNI}(n) = wt(n) - 1$ we first have to prove the following lemma as well as the structure of minimal trees:

Lemma 3.1. *Let T be a binary rooted tree that is minimal regarding the symmetry nodes index with n leaves and root ρ . Then we have the following equivalence:*

$$\rho \text{ is a symmetry node} \iff n = 2^k \text{ for a } k \in \mathbb{N}_0 \text{ and } T = T_k^{bal}$$

Proof. " \Leftarrow ": Trivial because every interior vertex of T_k^{bal} is a symmetry node.

" \Rightarrow ": This is proven by induction. Let $T_1 = (V_1, E_1)$ and $T_2 = (V_2, E_2)$ be the standard decomposition of T with roots ρ_1 and ρ_2 respectively. If ρ is a symmetry node we can conclude that n is even and T_1 and T_2 are isomorphic.

Base case ($n = 1, 2, 3$): The minimal trees for $n = 1, 2, 3$ are either fully balanced trees or the root is no symmetry node (see Table 1).

Induction hypothesis: Let the assertion hold for all trees with \tilde{n} leaves with $1 \leq \tilde{n} < n$ with $(n - 1) \in \mathbb{N}$ and $(n - 1) \geq 3$.

Induction step ($(n - 1) \rightarrow n$): Assume that a minimal tree shape T with $n \geq 4$ leaves is not T^{bal} , but its root ρ is a symmetry node. Due to the fact that $2 \leq n_j = |V_j^1| = \frac{n}{2} < n$ for $j = 1, 2$ the subtrees T_1 and T_2 contain at least one interior node (see Formula (1.4)) i.e. ρ_1 and ρ_2 are inner nodes. As $T \neq T^{bal}$ and ρ is a symmetry node there exists an inner node in each of the two isomorphic subtrees T_1 and T_2 that is not a symmetry node and therefore we know neither of the subtrees can be fully balanced as well. As $n_j \leq (n - 1)$ for $i = 1, 2$ we can use the induction hypothesis and conclude that ρ_1 and ρ_2 are no symmetry nodes.

Because $2 \leq n_j$ for $j = 1, 2$ we can decompose the tree further. Let T_{11} and T_{12} as well as T_{21} and T_{22} be the standard decomposition of T_1 and T_2 respectively. As T_1 and T_2 are isomorphic, there are two pairs of subtrees with one from each tree that have to be isomorphic. Without loss of generality let T_{1k} and T_{2k} be isomorphic for $k = 1, 2$. As ρ , but not ρ_1 and ρ_2 is a symmetry node, we have:

$$SNI(T) = 1 + 1 + SNI(T_{11}) + SNI(T_{12}) + SNI(T_{21}) + SNI(T_{22})$$

Now, consider the tree \hat{T} depicted in Figure 14 in which the subtrees are rearranged. The vertices x and y are symmetry nodes, but not the root z . Thus, the symmetry nodes index value is:

$$SNI(\hat{T}) = 1 + SNI(T_{11}) + SNI(T_{12}) + SNI(T_{21}) + SNI(T_{22}) < SNI(T)$$

This is a contradiction because \hat{T} has the same number of leaves as T that should be a minimal tree shape, but also a lower symmetry nodes index value.

□

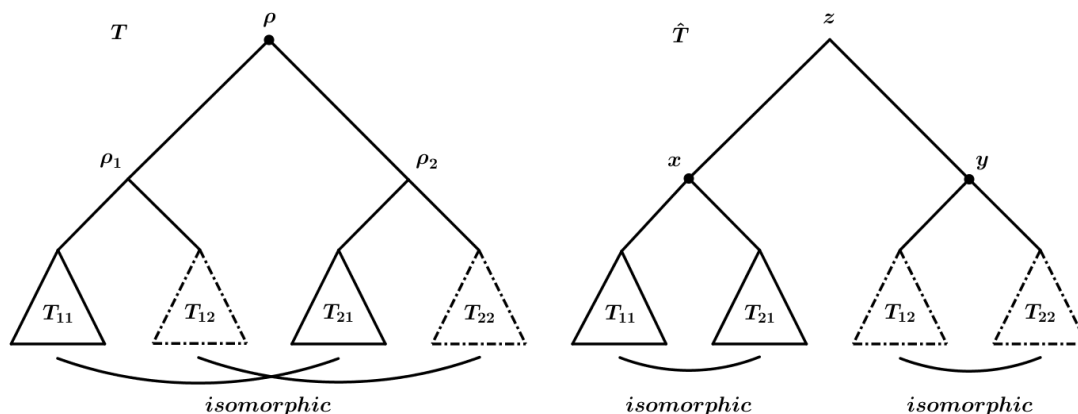


Figure 14: Depiction of T and \hat{T} from the proof of Lemma 3.1. The symmetry nodes are marked as a dot.

3.1.3 Minimal trees and their value

At the beginning of Section 3.1.2 we showed that we can construct a tree with $SNI(T) = wt(n) - 1$, but did not yet know if it is minimal. Now, we want to generalize the method as depicted in Figure 15:

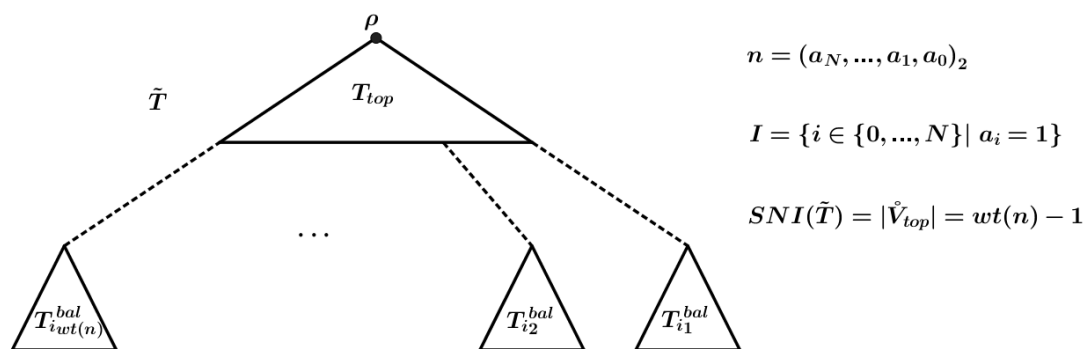


Figure 15: Construction of a minimal tree shape for the symmetry nodes index.

Definition 3.2. Define the general tree shape \tilde{T}_n as follows:

Let $n = (a_N, a_{N-1}, \dots, a_2, a_1, a_0)_2$ be an arbitrary positive integer with N and a_i defined as in Formula (3.2). $I = \{i \in \{0, \dots, N\} \mid a_i = 1\}$ is again the index set of the 1's and $|I| = wt(n)$. Then, \tilde{T}_n consists of the subtrees $(T_i^{bal})_{i \in I}$ that are connected by any "top tree" \tilde{T}_{top} with $wt(n)$ leaves.

We have to show that all interior nodes of \tilde{T}_n that are part of \tilde{T}_{top} are not symmetry nodes. Then we will have $SNI(\tilde{T}_n) = wt(n) - 1$ because \tilde{T}_{top} has $|\hat{V}_{top}| = wt(n) - 1$ inner vertices and the subtrees $(T_i^{bal})_{i \in I}$ only consist of symmetry nodes according to the definition.

Lemma 3.2. *The $wt(n) - 1$ interior nodes of a tree \tilde{T}_n constructed as explained above that are not part of the subtrees $(T_i^{bal})_{i \in I}$ are not symmetry nodes. This results in:*

$$SNI(\tilde{T}_n) = wt(n) - 1$$

Proof. In the case of $n = 2^k$ with $k \in \mathbb{N}$ the top tree does not contain any vertices and $\tilde{T}_n = T_k^{bal}$. Thus, we have $SNI(\tilde{T}_n) = wt(n) - 1 = 0$.

In all other cases let us assume there is an interior vertex $v \in V_{top}$ with children v_1 and v_2 in the top tree that is a symmetry node. This implies that the number of descendant leaves $\kappa(v_1)$ and $\kappa(v_2)$ of v_1 and v_2 should be equal.

Let $J_1 \subseteq I$ and $J_2 \subseteq I$ be index sets with $J_k = \{j \in I \mid T_j^{bal} \text{ descending of } v_k\}$. Then we have $J_1 \cap J_2 = \emptyset$ and $\kappa(v_i) = \sum_{j \in J_i} 2^j$ which leads to the contradiction $\kappa(v_1) \neq \kappa(v_2)$ because the binary expression is unique for every number. \square

Now we want to prove that all minimal trees are of the above-mentioned type \tilde{T}_n .

Theorem 3.3. *Any minimal tree for the symmetry nodes index value with n leaves is isomorphic to a tree \tilde{T}_n from Definition 3.2.*

Proof. We can prove this by induction:

Base case ($n = 1, 2$): The corresponding tree shapes are T_0^{bal} as well as T_1^{bal} and therefore fulfill the assertion.

Induction hypothesis: Let the assertion hold for all minimal trees with \tilde{n} leaves ($1 \leq \tilde{n} < n$ with $n \in \mathbb{N}$ and $n \geq 2$).

Induction step ($n - 1 \rightarrow n$): Let T be a binary rooted tree with at least 3 leaves, minimal symmetry nodes index value and root ρ . Again we can use the standard decomposition $T_1 = (V_1, E_1)$ and $T_2 = (V_2, E_2)$ of T .

Case " ρ is a symmetry node": Using Lemma 3.1 it follows that $n = 2^k$ with $k \in \mathbb{N}$ and $T = T_k^{bal}$ which is exactly \tilde{T}_n .

Case " ρ is not a symmetry node": Then T_1 and T_2 have to be minimal because T is minimal and they are not isomorphic. We have $1 \leq n_j = |V_j^1| < n$ for $j = 1, 2$ with $n = n_1 + n_2$. Using the induction hypothesis we can say that T_1 and T_2 are isomorphic to a tree \tilde{T}_{n_1} and \tilde{T}_{n_2} respectively. Now it only has to be ensured that T_1 and T_2 do not contain the same subtree T_i^{bal} with $i \in I$:

Assume that both T_1 and T_2 contain a certain subtree T_i^{bal} with $i \in I$. We have to differentiate between the shape of their top trees as they can be empty trees if T_1 or T_2 are isomorphic to T_i^{bal} ($\iff wt(n_j) = 1$ for $j = 1$ or 2) which implies that there is only a connecting edge from ρ to the root of T_i^{bal} . Furthermore, the case $wt(n_j) = 2$ must also be analyzed separately.

Case " $wt(n_j) = 1$ for $j = 1$ or $j = 2$ ": If both $wt(n_1) = wt(n_2) = 1$ then T_1 and T_2 are both isomorphic to T_i^{bal} because both of them contain exactly one subtree

which has to be T_i^{bal} according to our assertion. However, this would contradict T_1 and T_2 being non-isomorphic.

Therefore, we only have to deal with the case that only one subtree is isomorphic to T_i^{bal} . Without loss of generality let $wt(n_1) = 1$ and $wt(n_2) \geq 2$. The two cases $wt(n_2) = 2$ (i) and $wt(n_2) > 2$ (ii) are depicted in Figure 16.

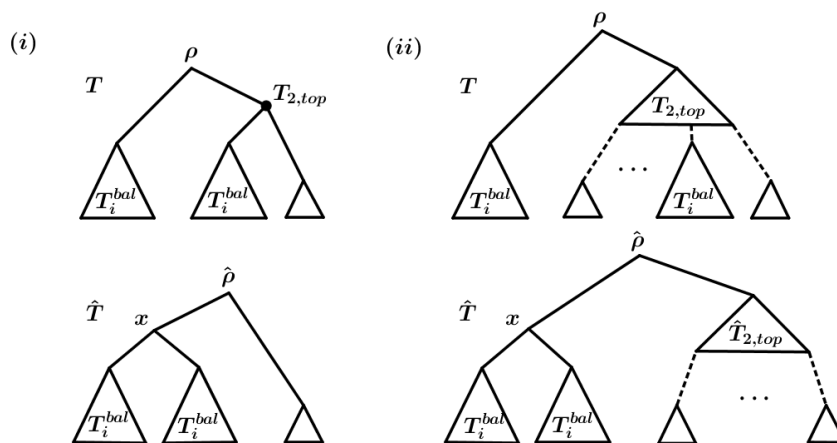


Figure 16: Depiction of T and \hat{T} for the cases that (i) : $wt(n_1) = 1$ and $wt(n_2) = 2$ as well as (ii) : $wt(n_1) = 1$ and $wt(n_2) > 2$.

(i): T has two vertices that are not symmetry nodes, but if we construct a tree \hat{T} as shown in the figure, we have increased the number of symmetry nodes by one (see vertex x). We have $SNI(T) = 2 > 1 = SNI(\hat{T})$ which contradicts T being minimal.

(ii): T has the root ρ as well as the $wt(n_2) - 1 \geq 2$ interior nodes of $T_{2,top}$ that are not symmetry nodes (see Lemma 3.2), but if we construct a tree \hat{T} as shown in the figure, we have increased the number of symmetry nodes by one because x is a symmetry node and $\hat{T}_{2,top}$ contains $wt(n_2) - 2 \geq 1$ vertices that are not symmetry nodes. This contradicts T being minimal because we have:

$$SNI(T) = 1 + wt(n_2) - 1 = wt(n_2) > wt(n_2) - 1 = 1 + wt(n_2) - 2 = SNI(\hat{T}).$$

Case " $wt(n_j) = 2$ for $j = 1$ or $j = 2$ ": Without loss of generality let $wt(n_1) = 2$ and $wt(n_2) \geq 2$ (otherwise former case). The two cases $wt(n_2) = 2$ (i) and $wt(n_2) > 2$ (ii) are depicted in Figure 17.

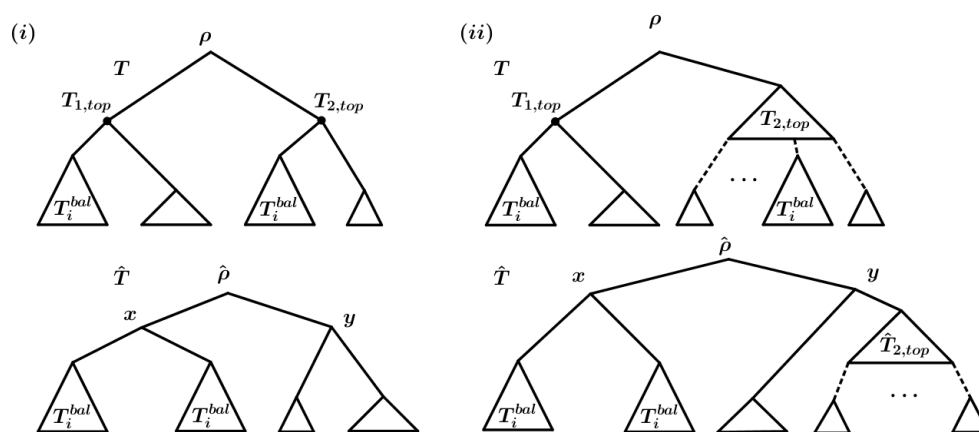


Figure 17: Depiction of T and \hat{T} for the cases that (i) : $wt(n_1) = 2$ and $wt(n_2) = 2$ as well as (ii) : $wt(n_1) = 2$ and $wt(n_2) > 2$.

(i): T has three vertices that are not symmetry nodes, but if we construct a tree \hat{T} as shown in the figure, we have increased the number of symmetry nodes by one (see vertex x). y will still be no symmetry node as the subtrees are non-isomorphic. As a result we have $SNI(T) = 3 > 2 = SNI(\hat{T})$ which contradicts T being minimal.

(ii): T has the root ρ , the one vertex of $T_{1,top}$ as well as the $wt(n_2) - 1 \geq 2$ interior nodes of $T_{2,top}$ that are not symmetry nodes (see Lemma 3.2), but if we construct a tree \hat{T} as shown in the figure, we have increased the number of symmetry nodes by one because x is a symmetry node and $\hat{T}_{2,top}$ contains $wt(n_2) - 2 \geq 1$ vertices that are not symmetry nodes. y will again be not a symmetry node as the subtrees are non-isomorphic. This contradicts T being minimal because we have:

$$SNI(T) = 1 + wt(n_2) - 1 = wt(n_2) > wt(n_2) - 1 = 1 + wt(n_2) - 2 = SNI(\hat{T}).$$

Case " $wt(n_j) > 2$ for $j = 1, 2$ ": The structure of T and its subtrees in this case are depicted in Figure 18. ρ is not a symmetry node as well as the inner nodes of the top trees whose numbers are $wt(n_1) - 1 \geq 2$ and $wt(n_2) - 1 \geq 2$ (see Lemma 3.2). This results in:

$$SNI(T) = 1 + wt(n_1) - 1 + wt(n_2) - 1 = wt(n_2) + wt(n_2) - 1.$$

Consider the tree shape \hat{T} also depicted in Figure 18. It has the same number of leaves as T , but the two subtrees T_i^{bal} and its connecting edge to their top trees were removed and rearranged so that the new node y is their parent node. The former root ρ is renamed as x and both x and y are the children of a new root $\hat{\rho}$. The new top trees $\hat{T}_{1,top}$ and $\hat{T}_{2,top}$ contain $wt(n_1) - 2 \geq 1$ and $wt(n_2) - 2 \geq 1$ vertices that are not symmetry nodes. x and $\hat{\rho}$ are no symmetry nodes because T_1 and T_2 are not isomorphic, but y definitely is a symmetry node. Because of

this we have:

$$SNI(\hat{T}) = wt(n_1) - 2 + wt(n_2) - 2 + 2 = wt(n_2) + wt(n_2) - 2 < SNI(T).$$

Again this is a contradiction because T should have been minimal. From this, we know that T_1 and T_2 contain only different fully balanced subtrees which concludes the proof.

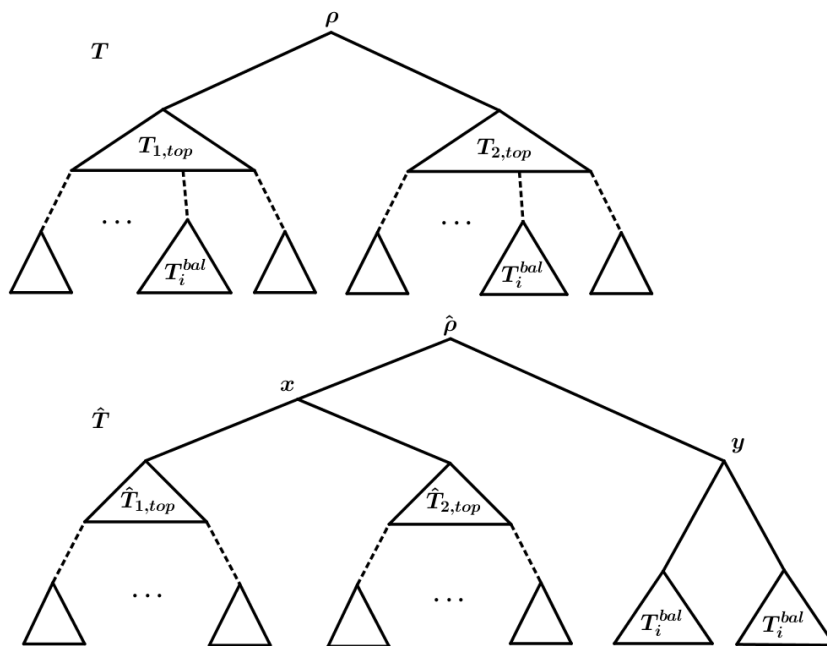


Figure 18: Depiction of T and \hat{T} for the case that $wt(n_j) > 2$ for $j = 1, 2$.

□

Remark. The thought in this proof of Theorem 3.3 that T_1 and T_2 cannot contain the same fully balanced subtree T_i^{bal} is equivalent to the binary extensions of both n_1 and n_2 not having a "1" at the same position. This results in $wt(n) = wt(n_1 + n_2) = wt(n_1) + wt(n_2)$ which can be used to show $SNI(T) = SNI(T_1) + SNI(T_2) + 1 = (wt(n_1) - 1) + (wt(n_2) - 1) + 1 = wt(n_1) + wt(n_2) - 1 = wt(n) - 1$. A proof following this approach could be shorter, but the current prove shows in an illustrative way how the symmetry nodes index works.

An example for the last case " $wt(n_j) > 2$ for $j = 1, 2$ " from the proof of Theorem 3.3 is shown in Figure 19 to comprehend the construction of $\hat{T}_{1,top}$ and $\hat{T}_{2,top}$. To keep everything clear both subtrees T_1 and T_2 contain the same (smallest) fully balanced subtrees, but they are non isomorphic so that the root ρ is not a symmetry node. $T_i^{bal} = T_1^{bal}$ was randomly chosen. Both top trees $T_{1,top}$ and $T_{2,top}$ have 2 interior nodes that are not symmetry nodes. We have $SNI(T) = 1 + wt(n_1) - 1 + wt(n_2) - 1 = 1 + 2 + 2 = 5$ but \hat{T} has a smaller index value because y and x are symmetry nodes: $SNI(\hat{T}) = wt(n_1) - 2 + wt(n_2) - 2 + 1 = 1 + 1 + 1 = 3 < SNI(T)$.

Of course, \hat{T} is not minimal as well because the left subtree of its standard decomposition is not minimal.

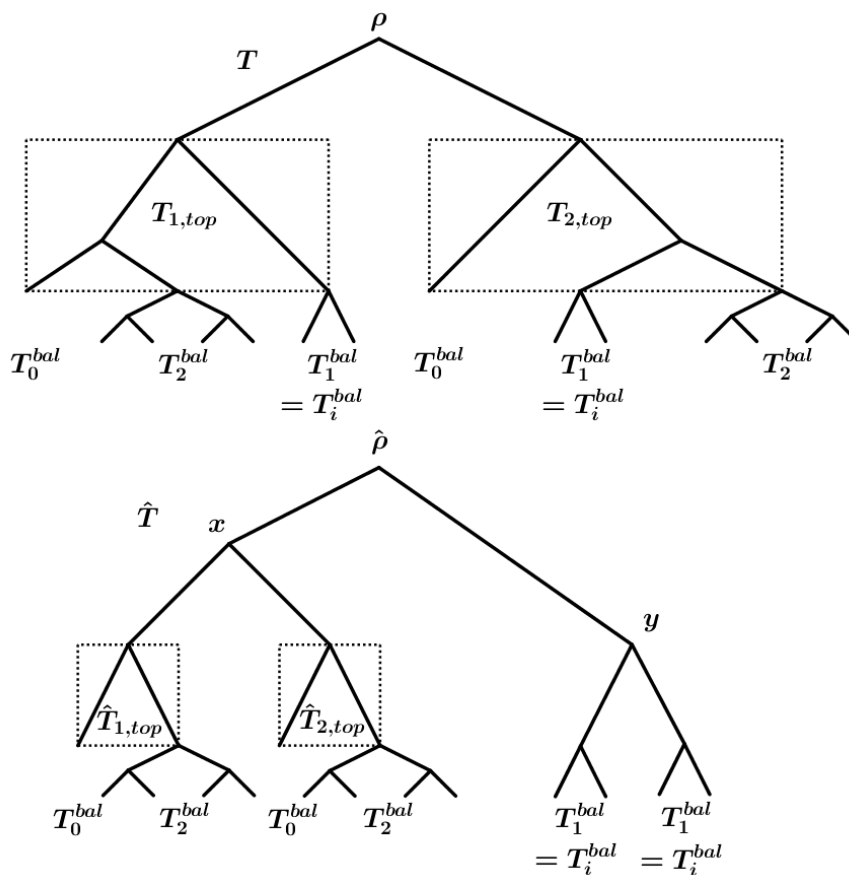


Figure 19: Example for T and \hat{T} with $n = 14$ leaves and $wt(n_j) = 3$ for $j = 1, 2$.

Due to the fact that the top tree of a tree isomorphic to \tilde{T}_n contains $wt(n) - 1$ nodes that are not symmetry nodes (see Lemma 3.2), we can now conclude:

Corollary 3.3.1. *The minimal symmetry nodes index value for a tree shape with n leaves is given by the following sequence:*

$$\min_{SNI}(n) = wt(n) - 1$$

Furthermore, we now know:

Theorem 3.4. *For the minimal symmetry nodes index value for a tree shape with n leaves we have:*

$$\min_{SNI}(n) = 0 \iff n = 2^k \text{ with } k \in \mathbb{N}_0$$

The corresponding tree shape T_k^{bal} is unique.

Proof. From the properties of the binary extension of n we can conclude:

$$\begin{aligned}
 n &= 2^k \text{ with } k \in \mathbb{N}_0 \\
 \iff n &= 1 \cdot 2^k + 0 \cdot 2^{k-1} + \dots + 0 \cdot 2^0 = a_k 2^k + a_{k-1} 2^{k-1} + \dots + a_0 2^0 \\
 \iff n &= (\dots, 0, 1, 0, \dots)_2 = (\dots, a_{k+1}, a_k, a_{k-1}, \dots, a_0)_2 \\
 \iff wt(n) &= \sum_{i=0}^k a_i = a_k = 1
 \end{aligned}$$

Now, the claim follows directly from Theorem 3.3.1:

$$min_{SNI}(n) = wt(n) - 1 = 0 \iff wt(n) = 1 \iff n = 2^k \text{ with } k \in \mathbb{N}_0$$

Every interior vertex of the corresponding tree shape has to be a symmetry node. This directly characterizes the fully balanced tree shape. \square

3.1.4 Number of minimal trees

If we depict the calculated numbers of minimal tree shapes in correspondence to the binary weight, we receive the following results (see Table 5). Here n ranges from 50 to 69 to see all binary weights from 1 to 6. The numbers of minimal tree shapes are:

$$3 = 1 \cdot 3, 15 = 1 \cdot 3 \cdot 5, 105 = 1 \cdot 3 \cdot 5 \cdot 7 \text{ and } 945 = 1 \cdot 3 \cdot 5 \cdot 7 \cdot 9.$$

Table 5: Number of minimal trees with n leaves and binary weight of n .

n	50	51	52	53	54	55	56	57	58	59
$wt(n)$	3	4	3	4	4	5	3	4	4	5
$\#min.trees$	3	15	3	15	15	105	3	15	15	105
n	60	61	62	63	64	65	66	67	68	69
$wt(n)$	4	5	5	6	1	2	2	3	2	3
$\#min.trees$	15	105	105	945	1	1	1	3	1	3

As there seems to be a dependence and with the knowledge of minimal trees from Section 3.1.3 we can put forward the following theorem using Equation (1.1):

Theorem 3.5. *The number of minimal trees with n leaves for the symmetry nodes index is given by*

$$\begin{aligned}
 |\{T \in M_{RB,n} | SNI(T) = min_{SNI}(n) = wt(n) - 1\}| &= |RB(wt(n))| \\
 &= \begin{cases} 1 & \text{if } wt(n)=1 \text{ or } 2 \\ (2 \cdot wt(n) - 3)!! & \text{else} \end{cases}
 \end{aligned}$$

with the double factorial $(2 \cdot wt(n) - 3)!! = 1 \cdot 3 \cdot 5 \cdot \dots \cdot (2 \cdot wt(n) - 5) \cdot (2 \cdot wt(n) - 3)$.

Proof. From Theorem 3.3 we know that a minimal tree T with n leaves consists of the subtrees $(T_i^{bal})_{i \in I}$ with a top tree T_{top} with $wt(n)$ leaves (see Definition 3.2). Thus, the number of minimal trees only depends on the number of possible top trees.

We can consider T_{top} as a (phylogenetic) X -tree with $X = I$ as all subtrees $(T_i^{bal})_{i \in I}$ are pairwise different. The labels $i \in I$ each mark the leaf of the top tree that has T_i^{bal} attached. Due to the fact that $|I| = wt(n)$ we can calculate the number of such labeled rooted binary trees with $|RB(wt(n))|$. \square

3.2 Maximal value and maximal tree

A tree shape with the least amount of symmetry nodes has the maximal index value.

Theorem 3.6. *The unique maximal value of the symmetry nodes index for a binary rooted tree with $n \geq 2$ leaves is*

$$\max_{SNI}(n) = |\mathring{V}| - 1 = n - 2$$

The unique maximal tree shape is the caterpillar tree.

Proof. From Formula (1.3) we know that every rooted binary tree shape with at least two leaves has a cherry and therefore a symmetry node. Hence, the upper bound for the maximum value is $|\mathring{V}| - 1$ with $|\mathring{V}| = n - 1$ including the root referring to Formula (1.4). The caterpillar tree has only one symmetry node and thus it follows $\max_{SNI}(n) = |\mathring{V}| - 1 = (n - 1) - 1 = n - 2$.

The caterpillar is also the unique tree shape with only one symmetry vertex because from proof of Theorem 2.4 we already know that T^{cat} is the unique rooted binary tree with only one cherry. Thus, any other rooted binary tree T has at least two cherries whose two parent nodes are symmetry nodes and therefore $SNI(T) \leq |\mathring{V}| - 2 = n - 3 < n - 2$. \square

4 Discussion and Results

In this chapter we first look at a different option to describe the balance indices and compare the extremal values of both indices with each other. Then, we explore the advantages and properties of the cherry and the symmetry nodes index, but also their disadvantages by comparing them with the established Sackin and Colles index in certain examples. Last but not least, we have a look at the summarized results from all chapters.

4.1 Cherry and symmetry nodes index as a function using the clade size

For unrooted trees it is possible to generate the split size sequence. Every edge induces a bipartition of the set of leaves into non-empty subsets A and B , a so called *split* $\sigma = A|B$ [22, p. 43]. An edge incident to a leaf induces a trivial split. The *size* $|\sigma|$ of a split $\sigma = A|B$ is the minimum of $|A|$ and $|B|$. The *split size sequence* is defined as the increasing sequence of split sizes except trivial splits that would have size 1 (see example in Figure 20).

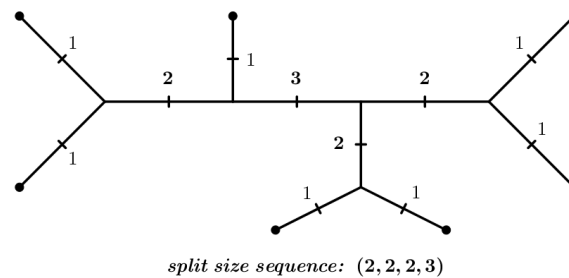


Figure 20: Example for the split size sequence. Every edge is marked with the split size of the induced split.

The connection of the split size sequence and the balance of unrooted trees has already been discussed [9]. Here, we want to explore if we can find a function for the cherry as well as the symmetry nodes index using the rooted equivalent of the split size sequence:

Every edge in a rooted tree induces a bipartition of the graph into the pending subtree and the subtree that contains the root. The former will here be called the *clade* θ [11, p. 150]. Its *size* $||\theta||$ is the number of descending leaves and the *clade size sequence* is the increasing sequence of all non-trivial clade sizes. That means all edges incident to a leaf will again be excluded as their clade has size 1 (see example in Figure 21).

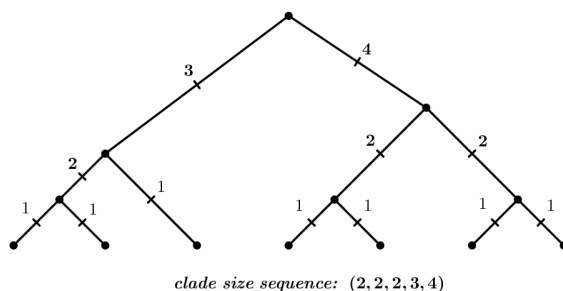


Figure 21: Example for the clade size sequence. Every edge is marked with its induced clade size.

For the cherry index it is very intuitive to count all 2's in the clade size sequence of a tree T and subtract twice their number from the number of leaves n . We have:

$$CI(T) = n - 2 \cdot \sum_{\theta \in \Theta} \mathbb{1}_{\{2\}}(|\theta|)$$

$\mathbb{1}_M(x)$ is the indicator function that is 1 if $x \in M$ and 0 in all other cases and Θ is the set of all clades induced by inner edges.

The symmetry nodes index cannot be described using the clade size sequence because there are trees with the same clade size sequence, but different symmetry nodes index values. One example is depicted in Figure 22. A rough lower and upper bound can be easily calculated though as the root of every clade with odd clade size cannot be a symmetry node, but the root of every clade with size 2 is a symmetry node. Let $Odd := \{2k + 1 \mid k \in \mathbb{N}_0\}$ denote the set of all odd integers, then we have:

$$\sum_{\theta \in \Theta} \mathbb{1}_{Odd}(|\theta|) \leq SNI(T) \leq (n - 1) - \sum_{\theta \in \Theta} \mathbb{1}_{\{2\}}(|\theta|)$$

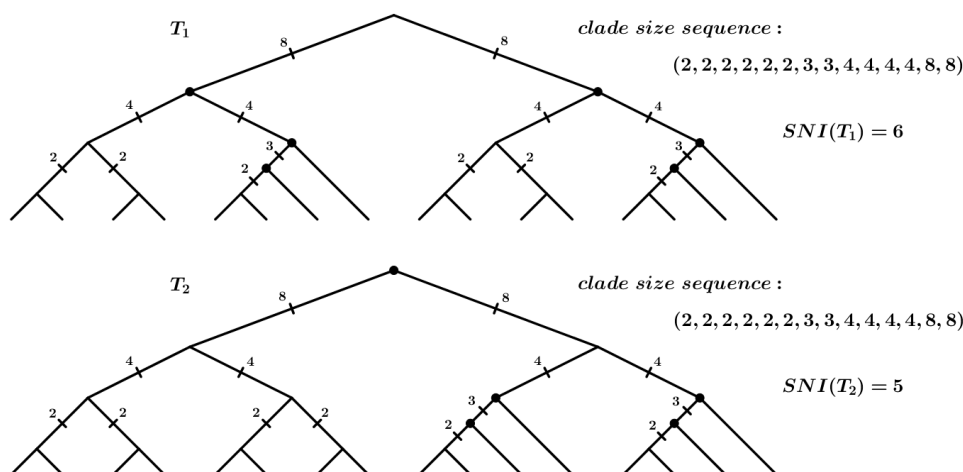


Figure 22: Example of two trees that have the same clade size, but different SNI value. The interior vertices that are not symmetry nodes are marked.

4.2 Comparison of the extremal values of CI and SNI

We have already proven that both indices have the same maximal value and maximal tree shape T_n^{cat} . Although their minimal values differ we can state the following:

Theorem 4.1. *Let T be a binary rooted tree with n leaves that is minimal regarding the symmetry nodes index. Then T is also minimal regarding the cherry index.*

Proof. Because T is minimal regarding the symmetry nodes index it follows that T is isomorphic to \tilde{T} (see Definition 3.2 and Theorem 3.3). Thus, all leaves of T are part of its subtrees $(T_i^{bal})_{i \in I}$.

If n is even, the binary extension contains $a_0 = 0$ and therefore T_i^{bal} has more than two leaves for every $i \in I$. As the fully balanced tree shape contains only symmetry nodes as interior vertices, the parent nodes of the leaves are symmetry nodes implying that every leaf is in a cherry. Thus, we have $CI(T) = 0$.

If n is odd, we have $a_0 = 1$ which means T_0^{bal} – a single leaf – is either the only leaf (for $n = 1$) or has a parent node that is not a symmetry node (see Lemma 3.2). In both cases it cannot be in a cherry. All leaves in the rest of the subtrees $(T_i^{bal})_{i \in I \setminus \{0\}}$ are again in cherries as in the even case. Therefore, T has the minimal cherry index value $CI(T) = 1$. \square

4.3 Comparison of different balance indices

Here, we want to have a look at the differences and similarities of the cherry, the symmetry nodes index and other more popular indices like the Colless and the Sackin index which are defined in Section 1.3). Here are the formulas of the Colless and Sackin index we use in the calculations:

$$\mathfrak{C}(T) := \sum_{v \in \check{V}(T)} |\kappa(v_1) - \kappa(v_2)| \qquad \mathfrak{S}(T) := \sum_{y \in \check{V}(T)} \kappa(y)$$

Definition 4.1. Let T_1 and T_2 be two binary rooted trees with n leaves and ϕ a balance index for binary rooted trees (see Section 1.3). Then we have:

$$T_1 \prec_{\phi} T_2 : \iff \phi(T_1) < \phi(T_2)$$

meaning that T_1 is more balanced than T_2 regarding the balance index ϕ . " $=_{\phi}$ " and " \succ_{ϕ} " can be defined analogously.

In Figure 23 there is an example in which the Colless and the Sackin index cannot decide which tree is more balanced than the other. However, the cherry and the symmetry nodes index can make this decision regarding T_2 that has a fully balanced subtree as the more balanced tree.

Nevertheless, even though this can be a desired property there are other examples in which both the cherry and the symmetry nodes index do not assess the degree of balance as it would be expected. That the cherry index is error-prone is easily imaginable. We only have to construct a tree from a sufficiently large caterpillar as a top tree with cherries attached. An example in which also the symmetry nodes index decides differently is depicted in Figure 24 with two trees with 15 leaves each. T_1 in this figure is constructed to be a minimal tree shape for these two indices with four fully balanced subtrees and a caterpillar as a top tree. However, intuitively and derived from the values of the Colless and Sackin index it seems to be a more imbalanced tree shape in

contrast to T_2 which nearly has the minimal Sackin index value $n(\lceil \log_2(n) \rceil + 1) - 2^{\lceil \log_2(n) \rceil} = 59$ [8]. T_2 , in turn, is regarded as less balanced from both the cherry and the symmetry nodes index.

The two index values have still some distance to their maximum $n - 2 = 13$ though and the Sackin index does not regard T_1 as completely asymmetrical because the maximum value would be $\frac{n \cdot (n+1)}{2} - 1 = 119$ [8].

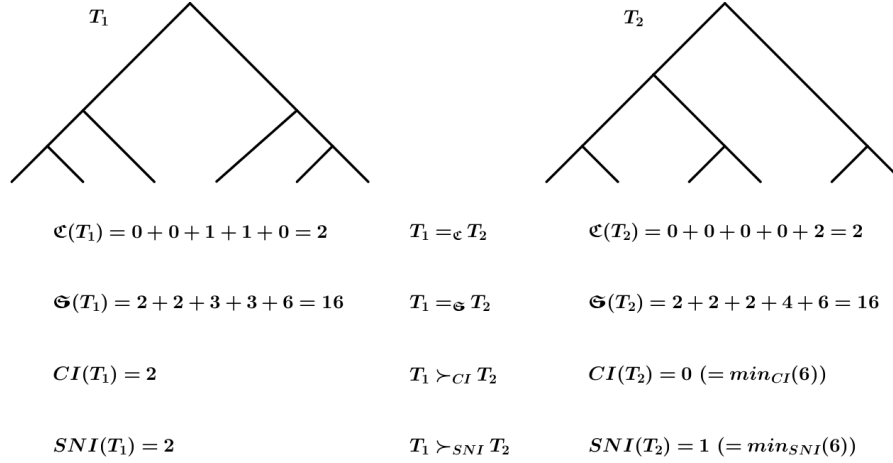


Figure 23: Two different tree shapes with 6 leaves and their balance index values.

In this second example, two trees are compared that have minimal values for one index and medium values for another. How substantial this difference is cannot be easily determined. We cannot just assume that the index values are equally distributed (this assumption is already invalid as we know that there are several minimal, but only one maximal tree for the cherry and symmetry nodes index). We would have to know their distributions for a more profound comparison.

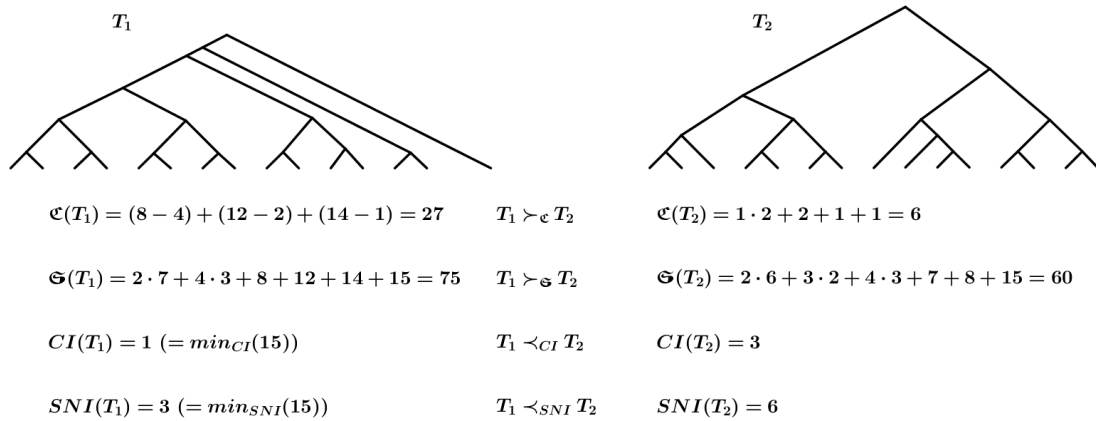


Figure 24: Two different tree shapes with 15 leaves and their balance index values.

Last but not least, it has to be noted that the advantage of the symmetry nodes index is that

it does not penalize asymmetry near the root as much as for example the Sackin index does. More important is the existence of subtrees that are highly balanced. This property could be useful if there is a task in which exactly these tree shapes have to be found.

4.4 Results

We want to shortly summarize the results of this bachelor thesis. Table 6 shows the results of Chapter 2 and 3 on the two balance indices for a tree with n leaves. The construction of the corresponding minimal trees can be found in Section 3.1.3 and 2.1.1.

Table 6: Extremal values and trees of both balance indices

	cherry index	symmetry nodes index
minimal value	$\begin{cases} 0 & \text{if } n \text{ is even} \\ 1 & \text{else} \end{cases}$	$wt(n) - 1$
minimal trees	<i>all tree shapes with $\lfloor \frac{n}{2} \rfloor$ cherries, unique only for $n = 1, 2, 3, 4, 6$</i>	$\begin{cases} T_k^{bal} \text{ unique} & \text{if } n = 2^k, k \in \mathbb{N} \\ \text{unique} & \text{for } wt(n) = 2 \\ \text{not unique} & \text{else} \end{cases}$
# min. trees	$\begin{cases} WE(\frac{n}{2}) & \text{if } n \text{ is even} \\ A(\frac{n-1}{2}) & \text{else} \end{cases}$	$ RB(wt(n)) $
maximal value	$(n - 2)$	$(n - 2)$
maximal trees	T_n^{cat}	T_n^{cat}
# max. trees	1	1

Both indices have similarities even though one looks solely at interior vertices and the other on leaves. They can be compared easily because both have a similar range of values and furthermore we proved that every minimal tree regarding the symmetry nodes index is also a minimal tree for the cherry index.

In comparison to other indices, both the cherry and the symmetry nodes index show more affinity to trees that at least contain subtrees that are highly balanced and are able to discriminate trees that other indices regarded as equally balanced.

The cherry index is easily calculable, but also prone to errors as it ignores a large part of the tree shape by only focusing on leaves. The symmetry nodes index, on the other hand, could be

a useful balance index as it seems to detect a different sort of balance pattern in trees. Thus, there are examples which it does not evaluate as expected based on values of other indices, but there possibly is an application that can make use of that. Further and more profound research on this topic would be interesting but would require knowledge of the index distribution.

5 Appendix: R-Scripts for CI and SNI

In this bachelor thesis, we used dynamic programming to calculate the minimal symmetry nodes index value as well as the number of minimal trees. Furthermore, R-scripts to calculate $CI(T)$ and $SNI(T)$ are provided. All use R version 3.4.3 [20] and additionally the latter ones also use the package `phytools` 0.6-60 [21] that includes amongst others the package `ape` 5.2 [19]. In this chapter, the functions and their ideas are presented including instructions on the usage. If all functions are loaded successfully, the working environment should look similar to Figure 25 (R-Studio).

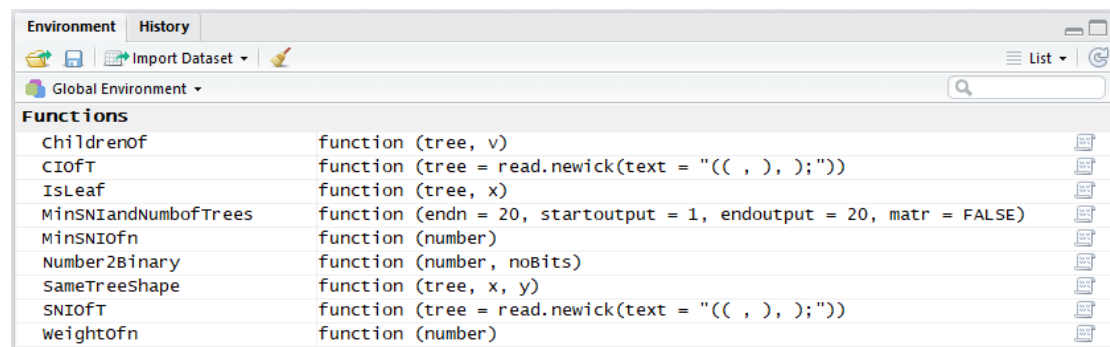


Figure 25: Complete environment if all scripts are loaded.

5.1 `MinSNIandNumbofTrees.R`

This script contains the function `MinSNIandNumbofTrees` that was used in Chapter 3 to get the first values of the symmetry nodes index (the ideas and the approach are explained in Section 3.1.1 and the code is documented as well). Due to the fact, that we could not prove short formulas for the minimal value, the number of possible tree shapes and the sizes of the subtrees, this function only exists to show how the algorithm with dynamic programming was implemented and to comprehend the results in Table 3 and 4. Of course, it can be used to calculate the minimal values, the number of possible tree shapes and their subtree sizes, but it will obviously perform less efficient than the function `MinSNIofn` in the Section 5.2 that uses the results from Chapter 3.

There are no packages required. Run the script to load the function (see Figure 25).

`MinSNIandNumbofTrees` uses four parameters: The first is `"endn"`, the maximum number of leaves for which all values are calculated. As we use dynamic programming the values of trees with lower numbers of leaves are saved during the course of the algorithm. Thus we can use the second and third parameter `"startoutput"` and `"endoutput"` to indicate which range of values should be printed. At last, there is the boolean parameter `"matr"` that can be set to `TRUE` if $min_{SNI}(n)$ and the number of minimal trees are required in the form of a matrix. For example we can call the function with `MinSNIofn(20,16,20,FALSE)` which is equivalent to `MinSNIofn(20,16,20)` as the default value for `"matr"` is `FALSE`. The results are depicted in Figure 26 below.

```
> MinSNIandNumbofTrees(20,16,20)
For n = 16 we have minSNI(n) = 0 and 1 possible tree shape(s) with subtree sizes 8, 8.
For n = 17 we have minSNI(n) = 1 and 1 possible tree shape(s) with subtree sizes 1, 16.
For n = 18 we have minSNI(n) = 1 and 1 possible tree shape(s) with subtree sizes 2, 16.
For n = 19 we have minSNI(n) = 2 and 3 possible tree shape(s) with subtree sizes 1, 2, 3, 16, 17, 18.
For n = 20 we have minSNI(n) = 1 and 1 possible tree shape(s) with subtree sizes 4, 16.
```

Figure 26: Example output of the function `MinSNIandNumbofTrees`.

5.2 minSNI.R

This script contains the function `MinSNIofn` which calculates $\min_{SNI}(n)$ using the result of Theorem 3.3 that $\min_{SNI}(n) = wt(n) - 1$. It uses two auxiliary functions: The function `WeightOfn` that calculates the binary weight of a number and the function `Number2Binary` that converts a number into a vector containing the binary extension.

There are no packages required to use all functions of the script `minSNI.R`. Run the script to load the functions (see Figure 25). Example calls of the functions are depicted in Figure 27.

`MinSNIofn` uses as input a non-negative integer "number" and returns the minimal symmetry nodes index value for trees with "number" leaves $wt(n) - 1$ (\rightarrow `WeightOfn`).

`WeightOfn` takes as input the variable "number" that should be a non-negative integer. The return value is the binary weight of this number. It sums up all entries of the binary extension (\rightarrow `Number2Binary` and the parameter "noBits" is set automatically).

`Number2Binary` is a function that uses two parameters: "number" a non-negative integer and "noBits" the number of bits or the length of the output vector that contains the binary extension of "number".

```
> MinSNIofn(55)
[1] 4
> weightofn(55)
[1] 5
> Number2Binary(55,ceiling(log2(55))+1)
[1] 0 1 1 0 1 1 1
```

Figure 27: Example output of all functions of `minSNI.R`.

5.3 CI.R

To use the script `CI.R` the package `phytools` is required. The script contains the function `CIofT` that calculates the cherry index value $CI(T)$ of a given tree. Two calls of the function with trees in the Newick tree format as input are depicted in Figure 28.

`CIofT` As input the function uses a tree with parameter name "tree" of class `phylo`, a common class for (phylogenetic) trees from the `ape` package. The return value is the number of leaves that are not in a cherry: $n - 2 \cdot c(T)$.

```
> CIofT(read.newick(text = "(( , ),((( , ), ), ),( , )))")
[1] 2
> CIofT(read.newick(text = "(( , ),((( , ),( , ( , ))) , ), )))")
[1] 3
```

Figure 28: Example output of the function `CIofT`.

5.4 SNI.R

To use the script `SNI.R` the package `phytools` is required. The script contains four functions. The main function is `SNIofT` that calculates the symmetry nodes index value $SNI(T)$ of a given tree. The other three functions `SameTreeShape`, `IsLeaf` and `ChildrenOf` are auxiliary functions, but can also be used on their own. Example calls of all functions `SNIofT` with trees in the Newick tree format as input are depicted in Figure 29. The function `SNIofT` is roughly tested and still delivers the result in a fraction of a second for $n \approx 800$ leaves.

SNIofT As input the function uses a tree with parameter name "tree" of class `phylo`. The return value is the number of interior nodes that are not symmetry nodes. The algorithm checks for every interior vertex if it is a symmetry node i.e. if the subtrees rooted in its direct children (\rightarrow `ChildrenOf`) have the same tree shape (\rightarrow `SameTreeShape`).

SameTreeShape This function uses three parameters: "tree", again a tree of class `phylo`, as well as the numbers of two nodes "x"= x and "y"= y . The recursive function determines if the subtrees rooted in x and y have the same tree shape and returns the corresponding boolean value. If both x and y are leaves they have the same tree shape, but not if only one of them is a leaf (\rightarrow `IsLeaf`). In the case of them being two interior nodes, we use the definition of tree shape isomorphism: $T_1 = (V_1, E_1)$ and $T_2 = (V_2, E_2)$ are isomorphic if there is a bijection $f : V_1 \rightarrow V_2$ with $\{f(u), f(v)\} \in E_2 \iff \{u, v\} \in E_1$ and $f(\rho_1) = \rho_2$. If there was such a bijection f between the subtrees rooted in x and y we would have $f(x) = y$ and one of the following cases: $f(x_1) = y_1 \wedge f(x_2) = y_2$ or $f(x_1) = y_2 \wedge f(x_2) = y_1$ with x_1, x_2, y_1, y_2 being the direct children of x and y respectively (\rightarrow `ChildrenOf`). The function recursively determines for these two cases if f can be extended on the rest of the subtrees. In other words it checks if either the subtrees rooted in x_1 and y_1 as well as x_2 and y_2 or in x_1 and y_2 as well as x_2 and y_1 are isomorphic.

IsLeaf This function uses two parameters: "tree", again a tree of class `phylo`, as well as the number of a vertex "x". It checks if "x" belongs to the leaves and returns the corresponding boolean value.

ChildrenOf As input the function also uses a tree "tree" of class `phylo` as well as the number of a vertex "v". The return value of the function is a vector containing all children of "v". This is not limited to rooted binary trees and can also give all adjacent (target) nodes of a directed graph with arbitrary out-degrees.

```
> SNIofT(read.newick(text = "(( , ),((( , ), ), ),( , )))")
[1] 4
> SameTreeShape(read.newick(text = "( , );"),1,3)
[1] FALSE
> IsLeaf(read.newick(text = "( , );"),1)
[1] TRUE
> ChildrenOf(read.newick(text = "( , );"),3)
[1] 1 2
```

Figure 29: Example output of the functions of `SNI.R`.

References

- [1] P. Agapow and A. Purvis. Power of eight tree shape statistics to detect nonrandom diversification: A comparison by simulation of two models of cladogenesis. *Systematic Biology*, 2002.
- [2] D. Baron, J. Braun, A. Erdmann, et al. *Genetik*. Schroedel, 2012.
- [3] D. Baum and S. Smith. *Tree Thinking*. Ben Roberts (Roberts and Company Publishers, Inc.), 2013.
- [4] M. G. B. Blum, E. Heyer, O. François, and F. Austerlitz. Matrilineal fertility inheritance detected in hunter-gatherer populations using the imbalance of gene genealogies. <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0020122>, 2006. Accessed: 08/20/18.
- [5] T. Coronadoa, A. Mira, F. Rossellóa, and G. Valiente. A balance index for phylogenetic trees based on quartets. *arXiv:1803.01651v1*, 2018.
- [6] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Inc., 2004.
- [7] J. Felsenstein, J. Archie, W. Day, W. Maddison, C. Meacham, F. Rohlf, and D. Swoford. The newick tree format. <http://evolution.genetics.washington.edu/phylip/newicktree.html>, 2000. Accessed: 11/18/18.
- [8] M. Fischer. Extremal values of the sackin balance index for rooted binary trees. *arXiv:1801.10418v2*, 2018.
- [9] M. Fischer and V. Liebscher. On the balance of unrooted trees. *arXiv:1510.07882v1*, 2015.
- [10] G. Fusco and G. Cronk. A new method for evaluating the shape of large phylogenies. *Theor Biol* 175: 235–243, 1995.
- [11] O. Gascuel and M. Steel. *Reconstructing Evolution*. Oxford University Press Inc., New York, 2007.
- [12] P. Hoff, W. Miram, and A. Paul. *Evolution*. Schroedel, 2013.
- [13] W. Johnson, E. Eizirik, J. Pecon-Slatteyrl, and W. Murphy. The Late Miocene Radiation of Modern Felidae: A Genetic Assessment. *Science, Volume 311, Issue 5757*, 2006.
- [14] M. Kirxpatrick and M. Slatkin. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution, Volume 47, Issue 4*, 1993.
- [15] J. S. Lansing and M. P. Cox. Neutrality and selection. <http://www.islandsoforder.com/neutrality-and-selection.html>. Accessed: 08/20/18.
- [16] J. S. Lansing, J. C. Watkins, B. Hallmark, M. P. Cox, T. M. Karafet, H. Sudoyo, and M. F. Hammer. Male dominance rarely skews the frequency distribution of y chromosome haplotypes in human populations. <http://www.pnas.org/content/105/33/11645>, 2008. Accessed: 08/20/18.

- [17] L. Maia, A. Colato, and J. Fontanari. Effect of selection on the topology of genealogical trees. *Journal of Theoretical Biology, Volume 226, Issue 3*, 2004.
- [18] A. Mir, F. Rosselló, and L. Rotger. A new balance index for phylogenetic trees. *Elsevier*, 2012.
- [19] E. Paradis and K. Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 2018.
- [20] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.
- [21] Liam J. Revell. phytools: An r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3:217–223, 2012.
- [22] C. Semple and M. Steel. *Phylogenetics*. Oxford University Press Inc., New York, 2009.
- [23] N. J. A. Sloane. The On-Line Encyclopedia of Integer Sequences, A0001201. <https://oeis.org/A0001201>, 2018. Accessed: 11/03/18.
- [24] N. J. A. Sloane. The On-Line Encyclopedia of Integer Sequences, A001190, Wedderburn-Etherington numbers. <https://oeis.org/A001190>, 2018. Accessed: 10/11/18.
- [25] N. J. A. Sloane. The On-Line Encyclopedia of Integer Sequences, A048881. <https://oeis.org/A048881>, 2018. Accessed: 11/03/18.
- [26] N. J. A. Sloane. The On-Line Encyclopedia of Integer Sequences, A085748. <https://oeis.org/A085748>, 2018. Accessed: 11/24/18.
- [27] E. Stam. Does imbalance in phylogenies reflect only bias? *Evolution*, 56(6), 2002.
- [28] M. Steel. *Phylogeny: discrete and random processes in evolution*. D. Marshall, 2016.
- [29] Wikipedia. Wedderburn–etherington number. https://en.wikipedia.org/wiki/Wedderburn%E2%80%93etherington_number. Accessed: 10/11/18.