On the Balance of Unrooted Trees

Mareike Fischer, Volkmar Liebscher

Ernst-Moritz-Arndt University Greifswald, Institute for Mathematics and Computer Science, Walther-Rathenau-Straße 47 D-17487 Greifswald/Germany

Abstract

We solve a class of optimization problems for (phylogenetic) X-trees or their shapes. These problems have recently appeared in different contexts, e.g. in the context of the impact of tree shapes on the size of TBR neighborhoods, but so far these problems have not been characterized and solved in a systematic way. In this work we generalize the concept and also present several applications. Moreover, our results give rise to a nice notion of balance for trees. Unsurprisingly, so-called caterpillars are the most unbalanced tree shapes, but it turns out that balanced tree shapes cannot be described so easily as they need not even be unique.

Keywords: phylogenetic trees, splits, caterpillars, semi-regular trees, NNI-moves

1. Introduction

When phylogenetic trees are considered, i.e. trees describing the evolutionary history of n present-day species which label the leaves of the tree, one is often confronted with the need to find the extreme values of the expression

$$\Phi_f(\tau) = \sum_{\sigma \in \Sigma^*(\tau)} f(\|\sigma\|).$$

Here, τ varies over all phylogenetic trees with n leaves, and $\Sigma^*(\tau)$ is the set of so-called non-trivial splits of $X = \{1,\ldots,n\}$ induced by τ. Recall that a split σ of a set X is just a bipartition of this set into two non-empty subsets A and B (we then write $\sigma = A|B$), and a split is called non-trivial whenever both parts of the bipartition have cardinality at least 2. Splits of the species set X play an important role in mathematical phylogenetics, because every edge of a phylogenetic tree induces such a split, and splits are non-trivial if and only if they are induced by an inner edge of the tree (i.e. not by an edge connected to a leaf). Moreover, in the above definition of $\Phi_f(\tau)$, $\|A|B\| = \min(|A|,|B|)$ denotes the cardinality of the smaller part of the split.

Such expressions recently appeared in different contexts, each time with a different choice of the monotone function $f : \{2,\ldots,\lfloor n/2 \rfloor\} \to \mathbb{R}_{\geq 0}$:

Email addresses: email@mareikefischer.de (Mareike Fischer), volkmar.liebscher@uni-greifswald.de (Volkmar Liebscher)

1. The size of a TBR-neighborhood of the tree τ with $f(k) = k(n-k)$, see [2]. This example is also related to the so-called Wiener index of the tree [5]. For this index, you need to assign edge lengths to all edges of τ. Then, the Wiener index is defined as $W(\tau) = \sum_{u,v \in V(T)} d_\tau(u,v)$, where $V(\tau)$ refers to the vertex set of τ and d_τ refers to the pairwise distances between any two vertices induced by the edge lengths.
2. An estimate of the diameter of the (unweighted) tree space under Gromov-type distance measures introduced in [4] with $f(k) = k$ (ℓ^1-Gromov) and $f(k) = \sqrt{k(n-k)}$ (ℓ^2-Gromov).
3. The number of so-called cherries of a tree, i.e. splits $\sigma = A|B$ with $\|A|B\| = \min(|A|,|B|) = 2$, can be described with $f(k) = \begin{cases} 1 & k = 2 \\ 0 & k > 2 \end{cases}$.

In many situations, it turns out that the extremal shapes (giving the minimal or maximal values of the functional) are so-called caterpillars (i.e. rooted trees with just one cherry or unrooted trees with just two cherries) and so-called semi-regular trees, i.e. trees which have at most one non-leaf vertex that is not of maximum degree and that such a vertex, if it exists, cannot be adjacent to more than one non-leaf vertex [5]. While the first kind of trees, the caterpillars, are often considered the most unbalanced tree shapes, the latter kind, i.e. the semi-regular trees, are considered to be most balanced. This reoccurence is challenging, as there seems to be no study of the whole family of functionals with f broadly varying. We regard the present note as the beginning of this study.

But how is $\Phi_f(\tau)$ measuring the balance of the shape of τ? Intuitively, splits $\sigma \in \Sigma^*(\tau)$ with a high value of $\|\sigma\|$ are very balanced, and it seems that the maximum of $\Phi_f(\tau)$ should be attained by the most balanced shapes. But, more importantly, the balanced tree shapes display *more* splits σ with a small $\|\sigma\|$, for example cherries. Thus they realise a smaller value of Φ_f than caterpillars.

In fact, the Arxiv version of [1], namely [2], provided already the main idea for deriving the maximal value for increasing functions f. As this Arxiv version provides a few more details than the published manuscript, we will subsequently refer to that version. Anyway, a lot of structure is required for deriving the minimal value, whereas the maximum is significantly easier to prove, cf. [2, 5].

We will show in the following that for general functions f, the main principle becomes even more transparent when considering particular functionals. Last but not least, we also give further applications to topological indices.

2. Preliminaries

Let $\mathscr{T}_n$ be the set of all (unrooted) phylogenetic trees (i.e. connected acyclic graphs with leaves labelled by a so-called taxon set X) with $|X| = n \geq 6$ leaves, and $\mathscr{T}_n^2$ the subset of $\mathscr{T}_n$ which contains all fully resolved (i.e. binary, bifurcating) trees in $\mathscr{T}_n$; i.e. the trees in $\mathscr{T}_n^2$ have the property that all vertices have either degree 3 (inner nodes) or 1 (leaves). When there is no ambiguity, we often just say tree when referring to a phylogenetic tree or, when the leaf labeling is not important, its so-called tree shape, respectively.

Caterpillars with n leaves are binary phylogenetic trees with two leaves, say $1, n$, such that every vertex of the tree is on the path from 1 to n or adjacent to a vertex on this path. In the so-called Newick format [8], which gives a nested list of all leaves such that leaves which are

separated by fewer edges in the tree are also separated by fewer brackets, caterpillars may be denoted by the expression

$$\tau_c = (((1,2),3),\ldots,n).$$

If $\tau \in \mathscr{T}_n$, let $\Sigma(\tau)$ denote the set of all splits, i.e. all partitions of the leaf set X into two non-empty subsets A and B, and let $\Sigma^*(\tau)$ be the set of all non-trivial splits $\sigma = A|B$ induced by inner edges of τ, i.e. $\Sigma^*(\tau)$ contains all splits for which both $|A| \geq 2$ and $|B| \geq 2$. For a split $\sigma = A|B$ let $\|\sigma\| = \|A|B\| = \min(|A|,|B|)$ denote its size.

Now we introduce for any function $f : \{2,\ldots,\lfloor n/2 \rfloor\} \to \mathbb{R}_{\geq 0}$ the functional $\Phi_f : \mathscr{T}_n \to \mathbb{R}_{\geq 0}$ via

$$\Phi_f(\tau) = \sum_{\sigma \in \Sigma^*(\tau)} f(\|\sigma\|).$$

Clearly, a tree and its contraction, obtained by suppressing all inner vertices of degree 2, share the same value of Φ_f. A rotation of the tree, which is obtained by permuting the leaf labels, does not alter the value of Φ_f either. So Φ_f is just a function of the phylogenetic tree shape of τ. Particularly, all caterpillars get the same value under Φ_f.

The last concept we need to introduce before we can present our results are so-called NNI-moves on binary trees. NNI stands for Nearest Neighbor Interchange, and in order to perform an NNI-move on a binary tree τ, you fix an inner edge of τ. This edge is connected to four subtrees, two on either side. You then swap two of these subtrees from opposite sides of the edge. This procedure is called NNI-move, and the resulting tree τ' is called an NNI-neighbor of τ. You can also define a metric based on NNI in order to measure the distance between two trees τ, τ' such that $d(\tau,\tau')$ equals the minimum number of NNI moves needed to get from τ to τ'.

3. Results for increasing functions f

In this section, we consider a function $f : \{1,\ldots,\lfloor n/2 \rfloor\} \to \mathbb{R}_{\geq 0}$ with the additional assumption that f is monotonously increasing, i.e. if $x > y$, then $f(x) \geq f(y)$. The following theorem, which is based on ideas presented in [2], shows that caterpillars maximize Φ_f in this case.

Theorem 1. *Let $\tau_c \in \mathscr{T}_n$ be a caterpillar and $f : \{2,\ldots,\lfloor n/2 \rfloor\} \to \mathbb{R}_{\geq 0}$ monotonously increasing. Then for all $\tau \in \mathscr{T}_n$*

$$\Phi_f(\tau_c) \geq \Phi_f(\tau).$$

If f is strictly increasing and $\tau \in \mathscr{T}_n$ is a point of maximum of Φ_f, then τ is a caterpillar.

Before presenting the proof, we consider the following elementary lemma, which we will need subsequently.

Lemma 1. *Let $f : \{2,\ldots,\lfloor n/2 \rfloor\} \to \mathbb{R}_{\geq 0}$ be strictly monotonously increasing. Then for all $m \in \{2,\ldots,n-2\}$ and $p \in \{1,\ldots,n-3-m\}$*

$$\sum_{l=2}^{m} f(\min(l,n-l)) < \sum_{l=2+p}^{m+p} f(\min(l,n-l)).$$

Proof. For symmetry reasons we may assume that $p \leq \lfloor (n-2-m)/2 \rfloor$ and consider the function $g : \{0,\ldots,\lfloor (n-2-m)/2 \rfloor\} \to \mathbb{R}_{\geq 0}$,

$$g(p) = \sum_{l=2+p}^{m+p} f(\min(l,n-l)).$$

We find

$$g(p+1)-g(p)=f(\min(m+p+1,n-m-p-1))-f(\min(2+p,n-2-p))>0$$

since $\min(m+p+1,n-m-p-1)>\min(2+p,n-2-p)$. □

Proof of Theorem 1. Consider first the case that f is strictly increasing and τ is such that $\Phi_f(\tau)$ is maximal. We now follow directly the arguments in the proof of [2, Lemma 4.1].

Fix two cherries x_1,x_2 and x_3,x_4 of τ and partition $\{1,\dots,n\}$ by the edges of the path from x_1 to x_4, see Fig. 1.

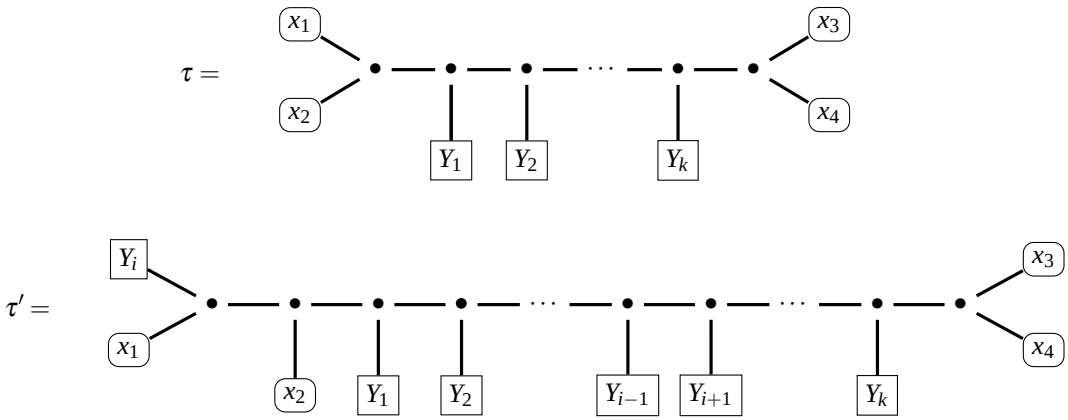


Figure 1: The structure of the trees τ and τ' in the proof of Theorem 1. The labels with round edges refer to leaves $x_1,\dots,x_4$, whereas the squares refer to subtrees $Y_1,\dots,Y_k$.

We just want to prove that the number of leaves in Y_i, which we denote by $|Y_i|=y_i$, equals 1 for all $i=1,\dots,k$, because this is equivalent to τ being a caterpillar. So let us assume there is an i with $y_i\geq 2$. We choose i minimal, i.e. $y_1=y_2=\cdots=y_{i-1}=1$. We now construct a new tree τ' by moving Y_i such that it is now a neighbor of x_1, see again Figure 1. We now consider several kinds of non-trivial splits in τ':

1. Splits σ which are not on the path from x_1 to x_4. For those splits $\|\sigma\|$ is the same for τ' as for τ.
2. Splits σ on the path from x_1 to x_4 before the edge splitting Y_{i-1} from Y_{i+1} (i.e. in Figure 1 to the left of this split). Enumerating these splits by their distance from x_1 in τ and τ', the size of the l^{th} split changes from $\min(l+1,n-l-1)$ to $\min(l+y_i,n-l-y_i)$.
3. The edge on the path from x_1 to x_4 which separates Y_{i-1} from Y_{i+1}. Here, the size of σ in τ' equals the size of the split splitting Y_i from Y_{i+1} in τ.
4. Splits σ on the path from x_1 to x_4 after the edge splitting Y_{i-1} from Y_{i+1}, i.e. in Figure 1 at the right-hand side. Here, the size of σ remains unchanged from τ to τ'.

Thus we find

$$\Phi_f(\tau')-\Phi_f(\tau)=\sum_{l=1}^{i} f(\min(l+y_i,n-l-y_i))-f(\min(l+1,n-l-1))>0$$

by Lemma 1. However, this implies that τ is not optimal. This is a contradiction, because τ was chosen to be optimal. So we conclude that the assumption is wrong and $y_i = 1$ for all $i = 1,\dots,k$. Therefore, τ is a caterpillar.

If f is increasing but not strictly increasing take an arbitrary tree τ and a caterpillar τ_c. For an arbitrary $\varepsilon > 0$ consider the function $f_\varepsilon : \{1,\dots,\lfloor n/2 \rfloor\} \to \mathbb{R}_{\geq 0}$,

$$f_\varepsilon(k) = f(k) + \varepsilon k \text{ for all } k = 1,\dots,\lfloor \frac{n}{2} \rfloor.$$

f_ε is strictly increasing and we obtain from the previous arguments

$$\Phi_{f_\varepsilon}(\tau_c) \geq \Phi_{f_\varepsilon}(\tau).$$

For $\varepsilon \downarrow 0$ we have $\Phi_{f_\varepsilon}(\tau_c) \downarrow \Phi_f(\tau_c)$ and $\Phi_{f_\varepsilon}(\tau) \downarrow \Phi_f(\tau)$, which completes the proof. □

Looking for an unconstrained minimum over $\mathscr{T}_n$ does not make much sense since it is attained at the most unresolved tree τ_0 with no inner splits: $\Sigma^*(\tau) = \emptyset$. Note that such a tree is also often referred to as a star-tree. If we restrict our consideration to $\mathscr{T}_n^2$, instead of considering the minimum of an increasing function, we could equivalently consider decreasing functions f and look again at the maximum. Anyway, the analysis is much more difficult, which is why we first need to present an important concept needed in this context, namely the so-called split size sequences.

4. Results for split size sequences and minima

From now on, we focus on binary trees, and we now introduce so-called split size sequences. In order to obtain these, we first associate to every binary tree $\tau \in \mathscr{T}_n^2$ the (multi-)set of split sizes $\{\|\sigma\| : \sigma \in \Sigma^*(\tau)\}$. However, for ease of notation, it is better to work with ordered $n-3$ tuples instead of (unordered) sets, which is why we continue with the following definition.

Definition 1. *Let $\tau \in \mathscr{T}_n^2$. Then, τ induces $n-3$ non-trivial splits, i.e. $|\Sigma^*(\tau)| = n-3$. We assume an arbitrary ordering $\sigma_1,\dots,\sigma_{n-3}$ of these splits and define the $(n-3)$-tuple $\tilde{s}(\tau)$ as follows:*

$$\tilde{s}(\tau)_i = \|\sigma_i\| \text{ for all } i = 1,\dots,n-3.$$

We then define the split size sequence $s(\tau)$ as follows: Order the $n-3$ entries of $\tilde{s}(\tau)$ increasingly and call the resulting ordered sequence $s(\tau)$. This is the split size sequence. Moreover, we denote by $\mathscr{S}_n = \left\{s(\tau) : \tau \in \mathscr{T}_n^2\right\}$ the set of all split size sequences on n taxa.

As an example for the split size sequence, we now consider the caterpillar tree. Consider again Figure 1. If τ in this figure is a caterpillar, denoted τ_c, it has $|Y_1| = \dots = |Y_k| = 1$. In order to get an (unordered) sequence of all split sizes, we start at one cherry, say x_1, x_2 and subsequently consider the splits $\{x_1,x_2,Y_1\}|X \setminus \{x_1,x_2,Y_1\},\dots,\{x_1,x_2,Y_1,\dots,Y_k\}|X \setminus \{x_1,x_2,Y_1,\dots,Y_k\}$. It is clear that the cherry contributes the value 2 to $\tilde{s}(\tau_c)$, then we get a 3, 4, 5, and so forth. However, as soon as there are fewer leaves on the right-hand side, say in set B, than on the left-hand side, say in set A, where $\sigma = A|B$, the sequence will continue with $|B| = n - |A|$ instead of $|A|$. In the end, we order the elements of $\tilde{s}(\tau_c)$ increasingly in order to derive $s(\tau_c)$. For example, if the caterpillar has $n = 6$ leaves, we get $s(\tau_c) = (2,2,3)$, if $n = 7$, we get $s(\tau_c) = (2,2,3,3)$, if $n = 8$, we get $s(\tau_c) = (2,2,3,3,4)$, and so forth.

Note that there is an alternative equivalent definition of $s(\tau)$. At first, it might seem a little less intuitive, but it proves to be useful in the following. Consider the increasing sequence $m(\tau) \in \mathbb{N}^{\{2,\dots,\lfloor n/2 \rfloor\}}$ whose entries are defined as follows:

$$m(\tau)_j = |\{\sigma \in \Sigma^*(\tau) : \|\sigma\| \leq j\}| \quad \text{for all } j = 2,\dots,\lfloor n/2 \rfloor.$$

Then, $s(\tau) \in \mathbb{N}^{n-3}$ is just the left inverse of $m(\tau)$:

$$s(\tau)_i = \min\left\{ j : m(\tau)_j \geq i \right\}.$$

This means

$$s(\tau)_i = j \qquad \text{if } m(\tau)_j \geq i \text{ and } m(\tau)_{j-1} < i.$$

The function m will be used in the subsequent proofs.

Note that it is not so clear how to characterize $\mathscr{S}_n$, i.e. all possible split size sequences. However, in the following we investigate them a bit further. Therefore, first of all we order $\mathscr{S}_n$ in the pointwise sense. So we say that $\tau \in \mathscr{T}_n^2$ is more balanced than $\tau' \in \mathscr{T}_n^2$, denoted $\tau \prec \tau'$, if

$$s(\tau)_i \leq s(\tau')_i \qquad \text{for all } i = 1,\dots,n-3,$$

or equivalently,

$$m(\tau)_j \geq m(\tau')_j \qquad \text{for all } j = 2,\dots,\lfloor n/2 \rfloor.$$

This leads to the following theorem.

Theorem 2. *Let $\tau, \tau' \in \mathscr{T}_n^2$. Then, we have $\tau \prec \tau'$ if and only if for all monotonously increasing $f : \{2,\dots,\lfloor n/2 \rfloor\} \to \mathbb{R}_{\geq 0}$*

$$\Phi_f(\tau) \leq \Phi_f(\tau'). \tag{1}$$

Proof. By definition, we have

$$\Phi_f(\tau) = \sum_{i=1}^{n-3} f(s(\tau)_i) \leq \sum_{i=1}^{n-3} f(s(\tau')_i) = \Phi_f(\tau'),$$

where the inequality is due to $\tau \prec \tau'$, which implies $s(\tau)_i \leq s(\tau')_i$ for all $i = 1,\dots,n-3$, and f being monotonously increasing. This completes the first part of the proof.

Now assume we have (1) for all monotonously increasing $f : \{2,\dots,\lfloor n/2 \rfloor\} \to \mathbb{R}_{\geq 0}$. Then, in particular (1) holds for the following family of monotonously increasing functions: For each $j \in \{2,\dots,\lfloor n/2 \rfloor\}$, define $f_j(k) = \begin{cases} 1 & k \geq j \\ 0 & k < j \end{cases}$. We obtain

$$\sum_{i=1}^{n-3} f_j(s(\tau)_i) = \Phi_{f_j}(\tau) \leq \Phi_{f_j}(\tau') = \sum_{i=1}^{n-3} f_j(s(\tau')_i) \quad \text{for all } j = 2,\dots,\lfloor n/2 \rfloor.$$

By the definition of f_j, this implies

$$|\{\sigma \in \Sigma^*(\tau) : \|\sigma\| \geq j\}| \geq \left|\left\{\sigma \in \Sigma^*(\tau') : \|\sigma\| \geq j\right\}\right|.$$

This immediately leads to $m(\tau)_j \geq m(\tau')_j$ for all $j = 2,\dots,\lfloor n/2 \rfloor$ and thus $\tau \prec \tau'$. □

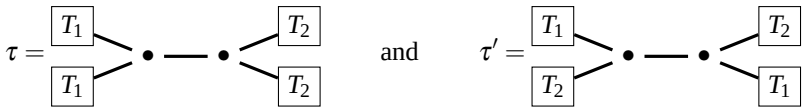
Remark 1. *From an abstract point of view the result is almost obvious: The functions* $f_j(k) = \begin{cases} 0 & k < j \\ 1 & k \geq j \end{cases}$ *, which we used in the proof of the above theorem, generate the extremal rays of the convex cone* $\{f : \{2,\dots,\lfloor n/2 \rfloor\} \to \mathbb{R}_{\geq 0} : f \nearrow\}$ *and thus determine the order* $\prec$*. On the other hand,* $\Phi_{f_j}(\tau) = n-3-m_{j-1}(\tau)$*.*

Theorem 3. *The only maximal point of* $\mathscr{S}_n$ *is derived from the sequence* $m(\tau)_j = \min(2j-2, n-3)$ *or* $s(\tau)_i = \lfloor (i+3)/2 \rfloor$ *corresponding to caterpillar trees* τ*.*

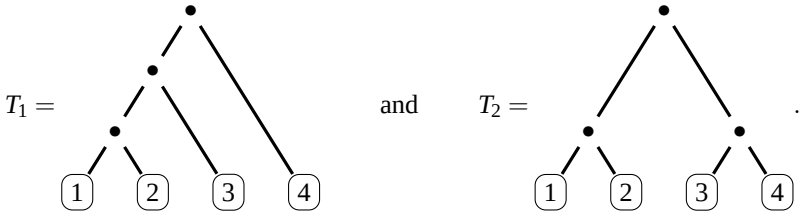
Proof. That caterpillars give the only maximal point on $\mathscr{S}_n$ is derived easily from the previous two theorems, as for a caterpillar τ_c we have $\Phi_f(\tau_c) \geq \Phi_f(\tau)$ for all $\tau \in \mathscr{T}_n^2$ by Theorem 1, and thus, by Theorem 2 we conclude $\tau \prec \tau_c$ for all τ. The split size sequence of caterpillars has already been described above. □

Thus, the Pareto maximum of $\mathscr{S}_n$ is unique. It even corresponds to a unique tree shape. Unfortunately, $s(\tau)$ does not determine the shape of τ in general. This means $\prec$ does not induce a partial order on tree shapes. We illustrate this with the following example.

Example 1. The basis for this example are the following two trees:

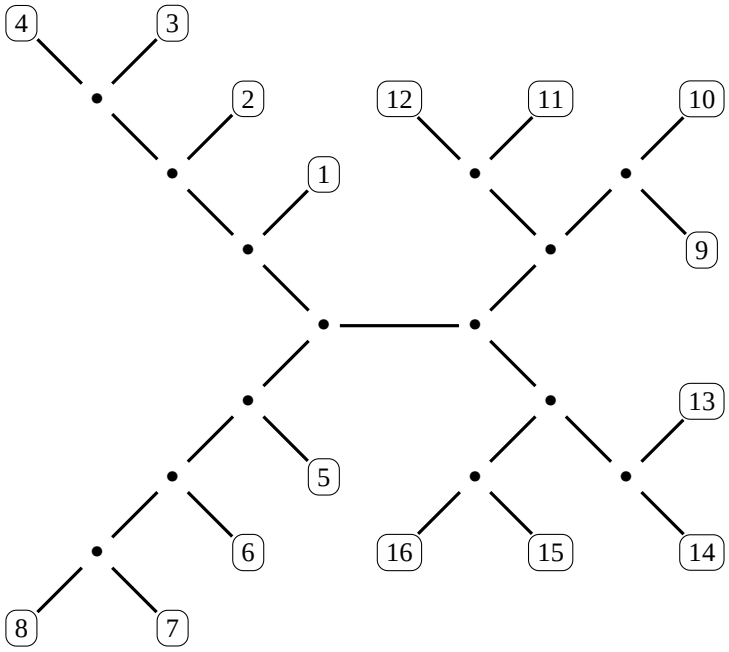


for two (rooted) subtrees T_1, T_2 on the same number of leaves with different shape. τ and τ' are just one NNI-move apart, which means that only two subtrees have to swap their position across one inner edge. Note that τ and τ' show the same split size sequences: First of all, both trees contain the splits induced by edges inside T_1 and T_2, as well as the four splits separating either of the two copies of T_1 or either of the two copies of T_2 from the rest of the tree. However, τ and τ' are different because they differ in the split induced by the central edge: this split separates the two copies of T_1 from the two copies of T_2 in τ and thus differs from the edge separating one copy of both T_1 and T_2 from another copy of T_1 and T_2. But even for these two splits the split size is the same, namely $n/2$, because T_1 and T_2 have the same number of leaves. Clearly, τ and τ' differ in shape if T_1, T_2 do so. The simplest situation is for $|T_1| = |T_2| = 4$,



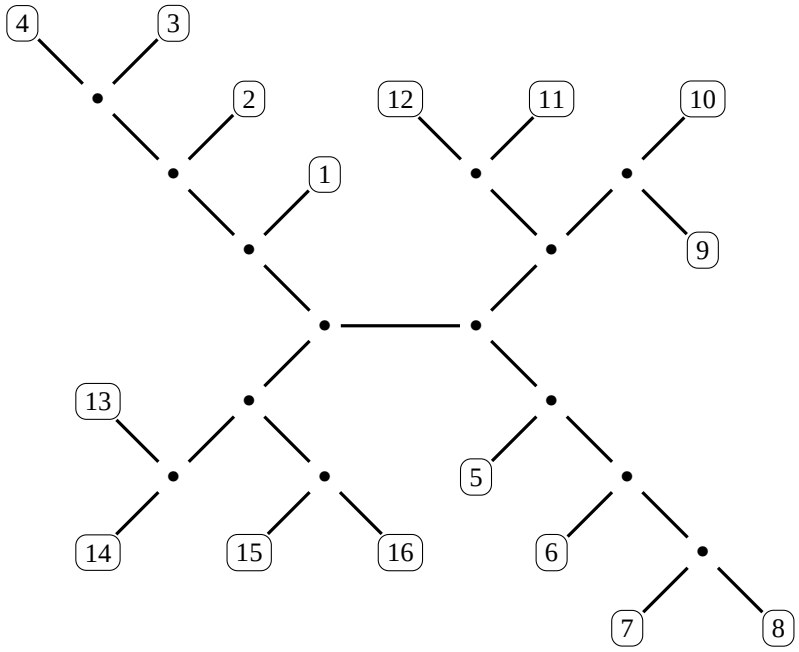
Thus we obtain

$\tau =$



and

$\tau' =$



which display both the split size sequence

$$s(\tau) = s(\tau') = (2,2,2,2,2,2,3,3,4,4,4,4,8).$$

So τ and τ' have the same split size sequence but differ in shape. By exhaustive search we found different (more complex) tree shapes with the same split size sequence already for $n = 11$. For $n \leq 10$, all binary trees with the same split size sequence have the same shape.

So far we have analyzed the maximum of $\mathscr{S}_n$ and have seen that it is achieved uniquely by the caterpillar tree. However, we also want to find the most balanced tree shape(s), i.e. Pareto minima of $\mathscr{S}_n$. Surprisingly, they are not unique, as we will demonstrate now.

Example 2. Let $n = 8$ and consider the trees $\tau, \tau' \in \mathscr{T}_8^2$ with split size sequences

$$\begin{aligned} s(\tau) &= (2,2,2,2,4) \\ s(\tau') &= (2,2,2,3,3) \end{aligned}$$

which are depicted in Figure 2. One could think of τ being mostly balanced with respect to the edge 1234|5678 and τ' being mostly balanced with respect to the vertex 123|45|678. For the monotonously increasing function f with

$$f(k) = \begin{cases} 1 & k \geq 4 \\ 0 & \text{otherwise} \end{cases},$$

we derive $\Phi_f(\tau) = 1 > 0 = \Phi_f(\tau')$. On the other hand, for the monotonously increasing function g with

$$g(k) = \begin{cases} 1 & k \geq 3 \\ 0 & \text{otherwise} \end{cases},$$

we derive $\Phi_g(\tau) = 1 < 2 = \Phi_g(\tau')$.

So depending on the underlying monotonously increasing function, either tree can be better than the other one. However, that the split size sequences induced by these trees are both Pareto minimal, i.e. that there is no tree which is more balanced than these two trees, can be seen when considering the hypothetical split size sequence $(2,2,2,2,3)$. This sequence would clearly dominate the above sequences, but $(2,2,2,2,3) \notin \mathscr{S}_8$: There are only eight leaves, so if there are four splits of size 2, this implies that all leaves should be in cherries. However, this contradicts the existence of a split of size 3.

As we have seen above, Pareto minima are not necessarily unique. The above example employed eight leaves. The following example shows that for $n = 12$, there are even more Pareto minima, namely three.

Example 3. There are more than 2 Pareto minima for large n. E.g., for $n = 12$,

$$\begin{array}{l} (2,2,2,2,2,2,4,4,4) \\ (2,2,2,2,2,3,3,4,5) \\ (2,2,2,2,3,3,3,3,6) \end{array} \tag{2}$$

are all minimal with respect to $\prec$.

We explicitly calculated the number of Pareto minima of $\mathscr{S}_n$ up to $n = 22$. Surprisingly, the number of Pareto minima is not monotonous, see Figure 3. The sequence of the numbers of Pareto minima is not contained in the online encyclopedia of integer sequences [13], which means that it seems to be unrelated to problems inducing other known integer sequences.

Next we want to focus on NNI-moves and the neighborhoods they induce.

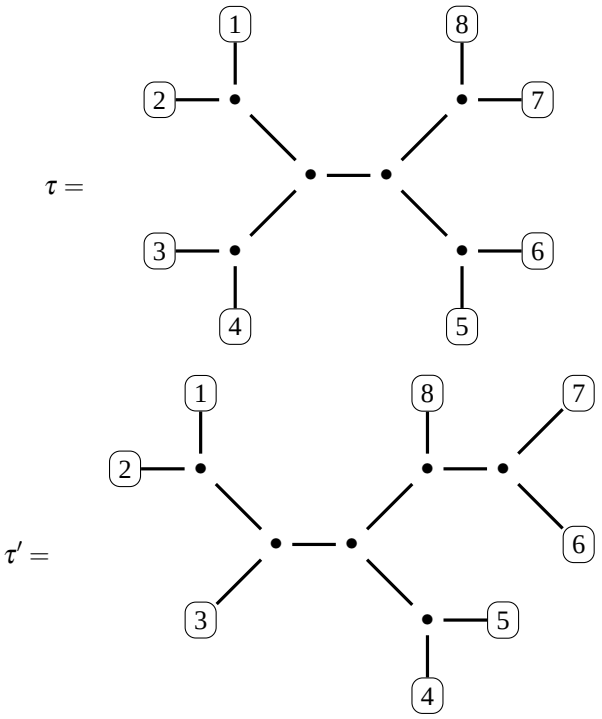


Figure 2: The structure of the trees τ and τ' with different Pareto minima of $\mathscr{S}_8$.

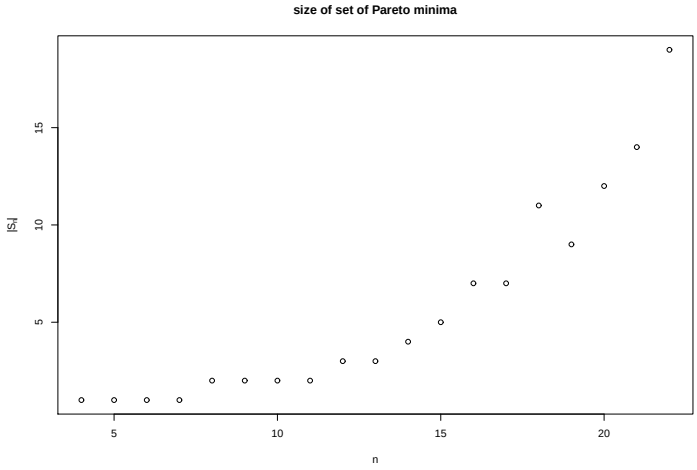
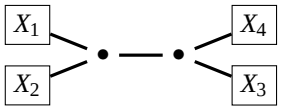


Figure 3: Number of Pareto minima of $\mathscr{S}_n$ for $n = 4, 5, \ldots, 22$.

Definition 2. *Let f be increasing. Then $\tau \in \mathscr{T}_n$ is called an NNI-local minimum of Φ_f if for τ' in the 1-NNI-neighborhood of τ, i.e. in the set of trees which can be reached from τ by performing one NNI-move, we have $\Phi_f(\tau') \geq \Phi_f(\tau)$.*

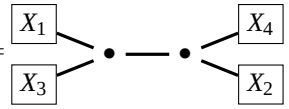
Lemma 2. *Let $\tau \in \mathscr{T}_n$ be a tree with its leaves labelled by X, $|X| = n$, and f be strictly increasing. Then $\tau \in \mathscr{T}_n$ is an NNI-local minimum of Φ_f if and only if for all splits σ such that $\sigma = X_1 \cup X_2 | X_3 \cup X_4$ with $X_i \cap X_j = \emptyset$ for all $i, j \in \{1,2,3,4\}$, $i \neq j$, and $\bigcup_{i=1}^{4} X_i = X$,*

$\tau =$ 

we have

$$\min(x_1 + x_2, x_3 + x_4) \leq \min \left\{ \begin{array}{c} \min(x_1 + x_3, x_2 + x_4), \\ \min(x_1 + x_4, x_2 + x_3) \end{array} \right\} \tag{3}$$

Proof. τ' in the 1-NNI-neighborhood of τ contains without loss of generality the edge $X_1 \cup X_3 | X_2 \cup X_4$ as depicted below.

$\tau' =$ 

Note that τ and τ' differ only in this edge, i.e. all other splits are the same in $\Sigma(\tau)$ and $\Sigma(\tau')$. Thus, in order to conclude that τ is an NNI-local minimum, we need

$$f(\min(x_1 + x_2, x_3 + x_4)) \leq f(\min(x_1 + x_3, x_2 + x_4)).$$

Since f is strictly increasing, this follows directly from $\min(x_1 + x_2, x_3 + x_4) \leq \min(x_1 + x_3, x_2 + x_4)$. Using an analogous argument for the other possible tree $\tau'' = X_1 \cup X_4 | X_2 \cup X_3$ leads to the desired result. □

Remark 2. *Note that (3) is just NNI-local minimality with respect to the function $f(k) = k$.*

We now consider the role of NNI concerning the Pareto minima discussed earlier.

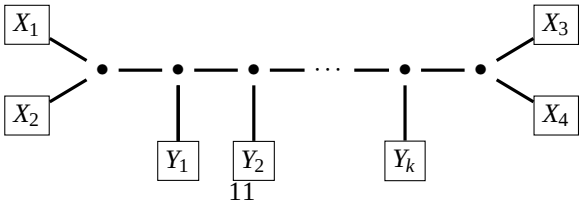
Lemma 3. *Every Pareto minimum is an NNI-local minimum.*

Proof. Any NNI-move changes at most one split size. Thus, Inequality (3) follows immediately from Pareto minimality. □

The following lemma is a direct conclusion of the above findings.

Lemma 4. *Let f be strictly monotonously increasing. Then for any minimal point $\tau \in \mathscr{T}_n$ of Φ_f the split size sequence $s(\tau)$ is a Pareto minimum of $\mathscr{S}_n$.*

According to [5], we call a tree τ semi-regular, if for all representations

$\tau =$ 

(4)

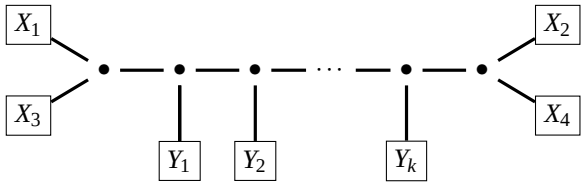
with (without loss of generality) $x_1 \leq x_2$ and $x_3 \leq x_4$, we additionally have $x_2 \leq x_3$ or $x_4 \leq x_1$. There is a unique semi-regular tree shape for every $n \geq 4$ which is completely characterized.

Remark 3. *Inequality (3) is just semi-regularity of τ as defined above and in [5], but restricted to adjacent vertices.*

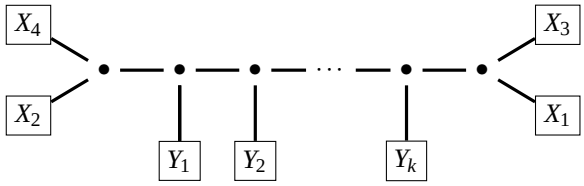
We are now in the position to state and prove the following theorem.

Theorem 4. *Let $f : \{1,\ldots,\lfloor n/2 \rfloor\} \to \mathbb{R}_{\geq 0}$ be strictly increasing and strictly concave, i.e. $2f(k) \geq f(k-1)+f(k+1)$ for all $k = 2,\ldots,\lfloor n/2 \rfloor - 1$. If τ is a minimal point of Φ_f then τ is semi-regular.*

Proof. First observe that we can extend f to a concave function $f : \{1,\ldots,n-1\} \to \mathbb{R}_{\geq 0}$ with $f(n-k) = f(k)$. Now consider the situation in (4) and assume $x_2 > x_3$ and $x_4 > x_1$. Fix the trees

$\tau_1 =$ 

and

$\tau_2 =$ 

We set now $\lambda = \frac{x_2 - x_3}{x_2 - x_3 + x_4 - x_1}$, and $z_j = \sum_{m=1}^{j} y_j$ where $z_0 = 0$. This way we obtain from strict concavity and

$$x_1 + x_2 + z_j = \lambda(x_1 + x_3 + z_j) + (1-\lambda)(x_4 + x_2 + z_j)$$

that

$$\begin{aligned} &\Phi_f(\tau) - \lambda\Phi_f(\tau_1) - (1-\lambda)\Phi_f(\tau_2) \\ &= \sum_{j=0}^{k} f(x_1+x_2+z_j) - \lambda f(x_1+x_3+z_j) - (1-\lambda) f(x_4+x_2+z_j) \\ &> 0. \end{aligned}$$

This contradicts minimality of τ, so the assumption was wrong. This leads to the desired result. □

Remark 4. *Interestingly, the trees τ_1, τ_2 are exactly at NNI distance 2 from τ.*

It looks like strict concavity is not needed, but we are still lacking a proof for this case. Furthermore, we see that the NNI-neighborhood for two NNI-moves gives a sufficient criterion for minimality (and maximality) and that the NNI-neighborhood for one NNI-move is not sufficient for $n \geq 8$. This is consistent with the empirical observation that tree search algorithms using (too small) NNI-neighborhoods often get stuck in local minima [3].

Example 4. Also for non-concave f the semi-regular pattern can be the minimal point. For $n = 12$ we saw that there are 3 different Pareto minima. For the convex function $f(k) = k^2$ we obtain the values $72, 79, 88$ for the split sequences in (2). This means the semi-regular pattern $(2,2,2,2,2,2,4,4,4)$ is still the minimizer.

We conclude this section with the following lemma which is used in a later example.

Lemma 5. *Let $\tau \in \mathscr{T}_n^2$ with $n \geq 4$. Then, we have $\lceil n/3 \rceil \leq s(\tau)_{n-3} \leq \lfloor n/2 \rfloor$.*

Proof. Let σ be the split the size of which is $s(\tau)_{n-3}$. Consider the edge e of τ corresponding to σ. Note that e splits τ into two parts, which we call the left-hand part and the right-hand part. Moreover, e must be an inner edge (as $n \geq 4$ and thus there is at least one inner edge giving rise to one non-trivial split, so the split of size $s(\tau)_{n-3}$, which is maximal, has to refer to a non-trivial split). Without loss of generality assume that the left-hand part of τ corresponds to $s(\tau)_{n-3}$. Then, the left-hand side end vertex of edge e must have degree 3 (as e is an inner edge). The splits corresponding to the two other edges coincident with this vertex must have a size not larger than $s(\tau)_{n-3}$. Thus $3 \cdot s(\tau)_{n-3} \geq n$ and the lower bound is derived. The upper bound is trivial. □

5. Applications

As we already mentioned in the introduction, functionals of the kind analyzed in the present manuscript have recently occurred in various contexts, some of which we want to mention here.

Example 5. In [2] the authors considered the functional

$$\Gamma(\tau) = \sum_{A|B \in \Sigma^*(\tau)} |A| \cdot |B|,$$

which is equivalent to Φ_f for $f(k) = k \cdot (n-k)$. They considered the functional only on $\mathscr{T}_n^2$, but for the maximum there is no difference anyway. The application of Theorem 1 is just the same as [2, Lemma 4.1], where we adapted our proof of Theorem 1 from. Further, the minimum is attained for the semi-regular shape, since f is even strictly increasing.

There is also a close relation between $\Gamma(\tau)$ and the Wiener index of τ as described in the next section, see also [5].

Example 6. Recently, the maximum parsimony distance was defined independently in [9] and [7]. We will now briefly explain this concept before we show how it is related to the topic of this manuscript.

Recall that a character χ is a function $\chi : X_\tau \to \mathscr{C}$ from the leaf set X_τ of τ to an alphabet $\mathscr{C}$. In biology, $\mathscr{C}$ often refers to the four nucleotides in the DNA alphabet, but we do not restrict the definition to this case. Then, an extension $\overline{\chi} : V(\tau) \to \mathscr{C}$ is a function from the vertex set $V(\tau)$ of τ to $\mathscr{C}$ which agrees with χ on the leaf set X. For a character χ with extension $\overline{\chi}$, the changing number of $\overline{\chi}$, denoted $ch(\overline{\chi})$, is defined as the number of edges $e = \{u,v\}$ such that $\overline{\chi}(u) \neq \overline{\chi}(u)$, and the parsimony score $PS(\chi,\tau)$ is then defined as the minimum number of changes over all extensions $\overline{\chi}$ of χ on τ, i.e. $PS(\chi,\tau) = \min_{\overline{\chi}} ch(\overline{\chi})$. Note that the parsimony score of a character on a tree can be easily calculated with the famous Fitch algorithm [11].

Now, the parsimony distance between two trees τ_1 and τ_2 is defined as follows: $d_{MP}(\tau_1,\tau_2) = \max_\chi |PS(\chi,\tau_1) - PS(\chi,\tau_2)|$. It has been shown [9, 10] that it is NP-hard to calculate this distance for two given trees (even if both trees are binary). We now consider neighborhoods induced by this metric.

In [7], the size of the 1-neighborhood of a tree $\tau \in \mathscr{T}_n$ with respect to d_{MP} was derived to be

$$n_p(\tau) = 4 \sum_{A|B \in \Sigma^*(\tau)} |A| \cdot |B| - 4(n-2)(n-3) + 2|V_2(\tau)| + 6|V_3(\tau)|$$

where $V_q(\tau)$, $q = 1,2,3$ is the set of vertices of degree q after removal of all pendant edges. Clearly, $V_1(\tau)$ are just all cherries. Unfortunately, $|V_2(\tau)|$ is not of the form Φ_f, but we can provide bounds on the minimal and maximal values, since we have

$$4\Gamma(\tau) + 2(n-2-|V_1(\tau)|) \leq n_p(\tau) + 4(n-2)(n-3) \leq 4\Gamma(\tau) + 6(n-2-|V_1(\tau)|).$$

So we get

$$4\Gamma(\tau) - c|V_1(\tau)| = \Phi_f(\tau)$$

for

$$f(k) = \begin{cases} 4k \cdot (n-k) & k > 2, \\ 8(n-2) - c & k = 2. \end{cases}$$

Clearly, for $c < 8(n-2)$, particularly for $c < 32$, f is strictly increasing and we find maximal values at caterpillars.

Note that if you restrict the parsimony distance to binary characters, denoted d^2_{MP}, and define $f(\tau) = d^2_{MP}(\tau, \tau_0)$, where τ_0 denotes the star-tree, i.e. the tree with no inner edges, then by Lemma 5, we have $f(\tau) \leq \lfloor n/2 \rfloor - 1$ and $f(\tau) \geq \lfloor n/3 \rfloor - 1$. It can easily be seen that f is maximized by all trees containing an edge inducing a split $\sigma = A|B$ with $|A| = \lfloor n/2 \rfloor$ and $|B| = \lceil n/2 \rceil$. Caterpillar trees have this property, but they are not the only ones. On the other hand, f is minimized by all trees containing a node which is adjacent to three subtrees of size at least $\lfloor n/3 \rfloor$. Therefore, the minimum is not unique, either.

Example 7. Consider the functional $\tau \mapsto |V_1(\tau)|$, the number of cherries. Clearly, we have to choose $f(k) = \begin{cases} 1 & k = 2, \\ 0 & \text{otherwise.} \end{cases}$ Since f is decreasing and convex (on $\{2, \ldots, \lfloor n/2 \rfloor\}$), the minimum number 2 of cherries is attained by caterpillars only. The maximal number of cherries is yielded by the semi-regular shapes.

Example 8. Another application are estimates of the diameter of $\mathscr{T}_n$ with respect to the Gromov-type ℓ^p-distances on $\mathscr{T}_n$ introduced in [4] for $p = 1,2$. We do not want to repeat the lengthy definition of these distances here. It is enough to note that Example 3 from that paper computed $D_p(\tau, \tau')$ if τ and τ' differ by one split, say $A|B$. Then

$$D_p(\tau, \tau') = \begin{cases} \min(|A|, |B|) & \text{for } p = 1, \\ \sqrt{|A| \cdot |B|/n} & \text{for } p = 2. \end{cases}$$

So we fix the functions $\tilde{f}_p$, $p = 1,2$, $\tilde{f}_1(k) = k$, $\tilde{f}_2(k) = \sqrt{k \cdot (1 - k/n)}$ which are both strictly increasing and concave on $\{2, \ldots, \lfloor n/2 \rfloor\}$. Let $\tau_0 \in \mathscr{T}_n$ denote the star tree, i.e. $\Sigma^*(\tau_0) = \emptyset$. We see for all trees $\tau \in \mathscr{T}_n$:

$$D_p(\tau_0, \tau) \leq \Phi_{\tilde{f}_p}(\tau).$$

Theorem 1 gives us for even n the estimates

$$\operatorname{diam}_{D_1}(\mathscr{T}_n) \leq 2 \sum_{k=2}^{n/2} 2k = n^2 - 2n - 4$$

and

$$\mathrm{diam}_{D_2}(\mathscr{T}_n) \leq 2\sum_{k=2}^{n/2} 2\sqrt{k\cdot(1-k/n)} \sim 4n^{3/2}\int_0^{1/2}\sqrt{x(1-x)}\mathrm{d}x = \pi n^{3/2}$$

on the diameter of $\mathscr{T}_n$.

Unfortunately, these bounds are less tight that the ones derived in [4, Lemma 9]. Nevertheless the same technique would apply to any other tree distance if we had estimates for Robinson-Foulds moves [12] in terms of $\|\sigma\|$.

At least, Theorem 1 supports now the conjecture that the maximal distance is attained between two caterpillars. But note that we are just dealing with upper bounds on the diameter here.

Example 9. Another simple functional is defined through

$$\Phi(\tau) = \max\{\|\sigma\| : \sigma \in \Sigma^*(\tau)\} = s(\tau)_{n-3}.$$

This functional is not of the form Φ_f. Still, it is monotonously increasing with respect to $\prec$. Its extremal values $\lfloor n/2\rfloor, \lceil n/3\rceil$ were derived in Lemma 5. By the previous section, they are achieved by caterpillars and some Pareto minimum, possibly among others.

For $n = 8$ we computed the Pareto minima, see Example 2. $(2,2,2,3,3)$ realizes the minimal value 3 for Φ. But, the other (semi-regular) split size sequence $(2,2,2,2,4)$ realizes the *maximal* value for Φ.

6. Discussion

We derived necessary and sufficient criteria to compute minimal and maximal points of the functionals Φ_f. There were a lot of specific functionals of this kind considered in the past, see the previous section. By our theory, it is now less surprising that quite often caterpillars yielded maximal values and the semi-regular trees yielded minimal ones. But, we also saw that sometimes the minimum of Φ_f could be achieved by a different Pareto minimum. So the set of split size sequences $\mathscr{S}_n$ and their Pareto minima seems quite interesting to study. For instance, it would be nice to know whether all Pareto minima, not only the semi-regular one, correspond to a unique tree shape.

The $\Gamma-$index from Example 5 is closely related to the Wiener index of graphs, it is its leaf-restricted form. Essentially it holds

$$\Gamma(\tau) = \sum_{u,v\in L(\tau)} d_\tau(u,v),$$

where d_τ is the metric induced by τ and $L(\tau)$ are the leaves of τ, see [5]. This kind of index is natural, unique and of course a shape invariant of τ.

It is easy to see that we can derive a whole family of similar topological indices if we introduce another metric on the tree, still just depending on the shape of τ. More precisely, introduce for any $\sigma \in \Sigma(\tau)$ a weight $w(\sigma,\tau) = g(\|\sigma\|) \geq 0$ and the corresponding (semi-)metric d_τ^w on $L(\tau)$. Then, setting

$$\Gamma^w(\tau) = \sum_{u,v\in L(\tau)} d_T^w(u,v)$$

we obtain $\Gamma^w(\tau) = \Phi_f(\tau)$ for $f(k) = g(k)k(n-k)$. If g is increasing, f is increasing as well. Unfortunately, concavity is not so easy to derive. Nevertheless, our theory will apply to many of these Γ^w-indices.

Surely, these indices are linear in d_τ. It is easy to derive similar functionals which are quadratic in d_τ or depending on triple partitions $A|B|C$ compatible with the tree τ. For instance, the functional $\Phi(\tau) = |V_2(\tau)|$ from Example 6 is of this kind. It would be quite valuable to extend our theory to this type of functionals.

References

[1] P.J. Humphreys and T. Wu, On the Neighborhoods of Trees, IEEE Trans. Comp. Biol. Bioinf. **10**(3):721–728, 2013.

[2] P.J. Humphreys and T. Wu, On the neighbourhoods of trees. preprint 2012 `arXiv:1202.2203` (`Arxiv` version of [1])

[3] S. Kläre and J. Leigh, personal communication

[4] V. Liebscher, Gromov meets Phylogenetics — new Animals for the Zoo of Metrics on Tree Space. submitted 2015 `arXiv:1504.05795`

[5] L.A. Székely, H. Wang, and T. Wu, The sum of the distances between the leaves of a tree and the 'semi-regular' property. Discr. Math. **311**(13):1197–1203, 2011.

[6] L.A. Székely and H. Wang, On subtrees of trees, Adv. Appl. Math.**34**(1): 138–155, 2005

[7] V. Moulton and T. Wu, A parsimony-based metric for phylogenetic trees, Adv. Appl. Math. **66**: 22–45, 2015.

[8] J. Felsenstein, J. Archie, W. Day, W. Maddison, C. Meacham, F. Rohlf and D. Swofford, The newick tree format. http://evolution.genetics.washington.edu/phylip/newicktree.html, 2000.

[9] M. Fischer and S. Kelk, On the Maximum Parsimony distance between phylogenetic trees, in press at Annals of Combinatorics.

[10] S. Kelk and M. Fischer, On the complexity of computing MP distance between binary phylogenetic trees, submitted to Annals of Combinatorics.

[11] W.M. Fitch, Toward defining the course of evolution: minimum change for a specific tree topology, Syst. Zool.**20**(4): 406–416, 1971.

[12] D.R. Robinson, L.R. Foulds, Comparison of phylogenetic trees, Math. Biosciences **53**: 131–147, 1981.

[13] The On-Line Encyclopedia of Integer Sequences, published electronically at `https://oeis.org/`, 2015