2018

Tropical Geometry of Phylogenetic Tree Space A Statistical Perspective

Monod, Anthea; Lin, Bo; Yoshida, Ruriko; Kang, Qiwen

ArXiv

Monod, Anthea, et al. "Tropical Geometry of Phylogenetic Tree Space: A Statistical Perspective." arXiv (2018): arXiv-1805.
http://hdl.handle.net/10945/64568

Tropical Geometry of Phylogenetic Tree Space: A Statistical Perspective

Anthea Monod[1,†], Bo Lin[2], Ruriko Yoshida[3], and Qiwen Kang[4]

1 Department of Applied Mathematics, Tel Aviv University, Tel Aviv, Israel
2 Department of Mathematics, University of Texas, Austin, TX, USA
3 Department of Operations Research, Naval Postgraduate School, Monterey, CA, USA
4 Medpace, Inc. and Department of Statistics, University of Kentucky, Lexington, KY, USA

† **Corresponding e-mail: antheam@tauex.tau.ac.il**

Abstract

Phylogenetic trees are the fundamental mathematical representation of evolutionary processes in biology. As data objects, they are characterized by the challenges associated with "big data," as well as the complication that their discrete geometric structure results in a non-Euclidean phylogenetic tree space, which poses computational and statistical limitations. We propose and study a novel framework based on tropical geometry and discuss its implications in the statistical analysis of evolutionary biological processes represented by phylogenetic trees. Our setting exhibits analytic, geometric, and topological properties that are desirable for theoretical studies in probability and statistics, as well as increased computational efficiency over the current state-of-the-art. We demonstrate our approach on seasonal influenza data.

Keywords: BHV tree space; phylogenetic trees; tropical geometry; tropical line segment; tropical metric.

1 Introduction

Evolutionary relationships describing how organisms are related by a common ancestor are represented in a branching diagram known as a *phylogenetic tree*. Phylogenetic trees model many important and diverse biological phenomena, such as speciation, the spread of pathogens, and cancer evolution. Methodology to analyze phylogenetic data has been under active research for several decades for two important reasons. First, explicit computations directly on collections of phylogenetic trees are challenging due to extremely high dimensionality. Second, standard statistical methodologies are not directly applicable due to the non-Euclidean nature of the setting. Significant previous work addresses various classical statistical interests, however a fundamental breakthrough for quantitative studies on sets of trees emerged through studying the *geometry* of the set of all phylogenetic trees (Billera et al., 2001).

Referred to in the literature as *BHV tree space*, the set of all phylogenetic trees is constructed as a moduli space, where each phylogenetic tree is represented as an individual point. The geometry is characterized by a unique geodesic between any two points; its length defines a metric on the space. Since its introduction in 2001, it has been actively studied in various wide-reaching domains, including algebraic geometry, category theory, computational biology, and statistical genetics (e.g. Baez and Otter (2017); Devadoss and Morava (2015); Nye et al. (2017); Weyenberg and Yoshida (2016)). Despite its indisputable significance, the BHV geometry nevertheless poses important data-analytic complications for both descriptive and inferential statistics.

Considering an alternative approach based on tropical geometry alleviates such complications. The first formal relation between tropical geometry and mathematical phylogenetics arises in a homeomorphism between the space of phylogenetic trees and a tropical variety (Speyer and Sturmfels, 2004). This coincidence has been further studied in theoretical research (e.g. Ardila and Klivans (2006); Manon (2011)), however, its implication and potential in applied work remain largely untapped.

In this paper, we harness the linearizing properties of the tropical geometric interpretation of phylogenetic tree space for quantitative and statistical analysis within the framework of metric geometry. Our approach is characterized by an explicit tropical metric; we refer to the resulting metric space as *palm tree space* (tropical

tree space). We show that mathematical properties of palm tree space provide the existence of fundamental statistical quantities—such as probability measures, means, and variances—ensuring that statistical and probabilistic studies are well defined, and more generally, are desirable for analytics and computation. Our framework lends a novel computational perspective to the set of all phylogenetic trees, which effectively results in a more natural setting for statistics: in the same way that linear algebra underlies classical statistics, the tropical metric allows for a tropical interpretation of linear algebra on tree space for *tropical statistics*.

The remainder of this paper is organized as follows. In Section 2, we provide detail on the mathematical setting for our contributions. Specifically, we give the relationship between phylogenetic trees and metrics in the context of tropical geometry. This section also provides a literature overview, including uses of tropical geometry in other computational settings and applications. In Section 3, we introduce and provide formal properties of the tropical metric on the space of phylogenetic trees and formally define palm tree space. We study its geometry, topology, and analytic properties. Section 4 gives a detailed treatment of the special case of equidistant trees and studies the behavior of their tree topologies in the setting of tropical line segments. In Section 5, we give some examples of probability measures that are often used and important in statistics and especially probability theory. Section 6 illustrates a statistical analysis on real data in the setting of palm tree space. We close in Section 7 with a discussion, and some ideas for future research.

2 Tropical Geometry and Tree Metrics

In this section, we introduce the essentials of tropical geometry for our study, and we provide a review of occurrences of tropical geometry in data analytic and statistical settings. We also discuss the representation of phylogenetic trees as metrics and give some examples of important and commonly-occurring tree metrics used in applications.

2.1 Basics of Tropical Geometry

Tropical algebra is a branch of abstract algebra based on specific semirings, which we now define. The extension of tropical algebra to tropical geometry is a variant of algebraic geometry: in algebraic geometry, geometric properties of zero sets of systems of multivariate polynomials are studied; in tropical geometry, these polynomials are defined in the tropical algebra.

Definition 1. The *tropical (max-plus) semiring* is $(\mathbb{R}\cup\{-\infty\}, \boxplus, \odot)$, with addition and multiplication given respectively by

$$a \boxplus b := \max(a, b) \quad \text{and} \quad a \odot b := a + b.$$

Remark 2. In the literature on tropical geometry (e.g. Speyer and Sturmfels (2009)), the term "tropical" typically refers to min-plus addition (where addition is given by the minimum between two elements, rather than their maximum), denoted by $\oplus$, where the semiring is $(\mathbb{R} \cup \{\infty\}, \oplus, \odot)$, isomorphic to the max-plus semiring. As will be further discussed in the context of phylogenetic tree metrics, we adopt the max-plus convention for consistency.

Note that tropical subtraction is not defined in this semiring. *Tropicalization* means passing from standard arithmetic to tropical arithmetic, replacing the standard arithmetic operations with their tropical counterparts. Using these arithmetic operations, usual mathematical relations and objects of interest may be studied, such as lines, functions, and sets. Tropicalized mathematical objects are piecewise linear constructions due to the linearizing binary operations on the tropical semiring. For example, the tropically quadratic polynomial $x^2 \boxplus 1$ is given by $x \odot x \boxplus 1 = \max(2x, 1)$ in standard arithmetic. These linearizing properties are desirable for increased computational efficiency. See Maclagan and Sturmfels (2015) for a complete, formal treatment of tropical geometry in full detail with examples.

2.2 Phylogenetic Trees and Metrics

A *phylogenetic tree* is an acyclic connected graph, $T = (V, E)$, where V is the set of labeled terminal nodes called *leaves* with degree 1, and E is the set of *edges* or *branches*, each with positive length and representing

evolutionary time. Non-leaf vertices in a phylogenetic tree have degree at least 2. Edges connecting to leaves are called *external edges*; otherwise, they are *internal edges*. Phylogenetic trees model the dynamics of many biological processes. When the biological process is believed to originate from a common ancestor, the tree is *rooted*, where the root represents the common ancestor and is a unique node of degree 2. The evolution then progresses by series of bifurcations from the root (edges), ending in terminal nodes (leaves).

Biologically, phylogenetic trees come from an input of molecular sequence data from a finite number of species. The task is then to reconstruct their evolutionary phylogeny and graphically represent it as a tree. Tree reconstruction is an extremely challenging problem that has been under active research for several decades. Techniques for tree reconstruction largely fall into two categories: statistical methods and distance-based methods.

Statistical methods are primarily based on the optimization of some specified optimality criteria, which are based on the idea that nucleotides in DNA evolution substitute one another according to a continuous-time Markov chain, which has been developed to a general independent time-reversible model for finite sites (Tavaré, 1986). Classical statistical optimization methodology of such criteria are then implemented, such as maximum parsimony or likelihood (Felsenstein, 1981; Fitch, 1971). Bayesian methods for tree reconstruction also exist to estimate the posterior distributions of trees (Edwards, 1970; Rannala and Yang, 1996). The underlying need for statistical methods is based on the fact that there is an extremely high number of possibilities for tree combinatorial types or *topologies* (i.e. the branching configuration together with a leaf labeling scheme): for rooted binary trees with N leaves, there are

$$(2N-3)!! = (2N-3) \times (2N-5) \times \cdots \times 3$$

tree topologies (Schröder, 1870). The sample complexity for likelihood-based phylogenetic inference methods is known to match, up to constants, the information-theoretic lower bound of classical models of evolution (Roch and Sly, 2017). For parsimony-based methods under the assumption of uniform distributivity, the optimization problem has been shown to be an NP-complete Steiner tree problem (Foulds and Graham, 1982). Restricted cases of the Steiner tree problem, such as the minimum spanning tree problem, as well as other variants are solvable in polynomial time and by other more efficient and scaleable solutions (e.g. Juhl et al. (2014)). In biological applications when other distributional assumptions are reasonable, such as identifiability of mixture distributions, statistical methods are easier and more computationally efficient to implement (Allman and Rhodes, 2008; Long and Sullivant, 2015; Rhodes and Sullivant, 2012).

Distance-based methods deal with reconstructing trees from distance matrices (e.g. Fitch and Margoliash (1967); Saitou and Nei (1987); Sokal and Michener (1958)). These methods rely on the specification of genetic distance, such as Hamming or Jukes–Cantor distances, which measure distances between all pairs of sequences. Phylogenetic trees are then reconstructed from this matrix by positioning sequences that are closely related under the same node, with branch lengths faithfully corresponding to the observed distances between sequences. These pairwise distances between leaves in a tree can be stored in vectors, effectively representing trees as vectors. Such vectors computed from various trees can then be used for comparative statistical and computational studies on sets of trees. Other approaches that also represent trees as metrics have been studied where the tree itself is considered as a probability space arising from some stochastic process (on the sequence level), and Wasserstein distances as metrics on spaces of probability measures are then studied (Evans and Matsen, 2012; Sommerfeld and Munk, 2018). In these studies, the space is restricted to the tree and the measures studied are those on sequences that make up the tree, while in our setting, we study sets of trees, rather than a single tree. Thus, the space we work in is therefore the collection of all possible sets of trees.

We now provide the formal setting of phylogenetic trees in the context of distances; further details to supplement what follows may be found in Pachter and Sturmfels (2005).

Notation. Throughout this paper, $N \in \mathbb{N}$, and we denote $[N] := \{1, 2, \ldots, N\}$ and $n := \binom{N}{2}$. We also write interchangeably $w(i,j) = w_{ij} = w_{\{i,j\}}$.

Definition 3. A *dissimilarity map* w is a function $w : [N] \times [N] \to \mathbb{R}_{\geq 0}$ such that

$$w(i,i) = 0 \quad \text{and} \quad w(i,j) = w(j,i) \geq 0$$

for every $i, j \in [N]$. If a dissimilarity map w additionally satisfies the triangle inequality, $w(i,j) \leq w(i,k) + w(k,j)$ for all $i, j, k \in [N]$, then w is called a *metric*.

The relationship between dissimilarity maps and metrics is intrinsically tropical, which can be seen in the following result.

Lemma 4 (Butkovič (2010))**.** *Let $w : [N] \times [N] \to \mathbb{R}_{\geq 0}$ be a dissimilarity map and let W be its corresponding matrix. Then w is a metric if and only if $-W \odot -W = -W$.*

2.2.1 Tree Metrics

For a tree with N leaves, the pairwise distances between leaves are sufficient to specify a phylogenetic tree.

Definition 5. Let T be a phylogenetic tree with N leaves, with labels $[N]$, and assign a branch length $b_e \in \mathbb{R}_{\geq 0}$ to each edge e in T. Let

$$\begin{aligned} w : [N] \times [N] &\to \mathbb{R}_{\geq 0} \\ w(i,j) &\mapsto \bigodot_{e \in P_{\{i,j\}}} b_e, \end{aligned}$$

where $P_{\{i,j\}}$ is the unique path from leaf i to leaf j. The map w is then called a *tree distance*. If $w(i,j) \geq 0$ for all i, $j \in [N]$, then w is a *tree metric*.

Note that tree metrics may be represented as *cophenetic vectors* in the following manner (Cardona et al., 2013):

$$(w(1,2),\ w(1,3),\ \ldots\ ,\ w(N-1,N)) \in \mathbb{R}^n_{\geq 0}. \tag{1}$$

Cophenetic vectors are also called *tree metrics* (in vectorial representation) in the literature on mathematical phylogenetics. The following definition gives a technical condition for tree metrics.

Definition 6. The space $\mathcal{T}_N$ of tree metrics for phylogenetic trees with N leaves consists of all n-tuples $\big\{w(i,j)\big\}_{1 \leq i < j \leq N}$ where the maximum among the following tropically quadratic *Plücker relations*:

$$\begin{aligned} &w(i,j) \odot w(k,l), \\ &w(i,k) \odot w(j,l), \\ &w(i,l) \odot w(j,k) \end{aligned} \tag{2}$$

is attained at least twice for $1 \leq i < j < k < l \leq N$. This condition for a tree metric is called the *four-point condition* (Buneman, 1974). This condition may equivalently be expressed by requiring that the nonnegative, symmetric $N \times N$ matrix $W = (w(i,j))$ have zero entries on the diagonal, and that

$$w(i,j) \odot w(k,l) \leq w(i,k) \odot w(j,l) \boxplus w(i,l) \odot w(j,k) \tag{3}$$

for all distinct $i, j, k, l \in [N]$.

Remark 7. The space $\mathcal{T}_N$ consists of all trees: by definition, in order to be a tree and an element of $\mathcal{T}_N$, it must satisfy the four-point condition. The four-point condition makes no distinction between rooted or unrooted trees; in the case of rooted trees, the root (i.e. leaf label 0) is simply ignored and the condition applies to the N leaves.

Notice that the conditions in Definition 6 are expressed in terms of maxima, which motivates our use of the max-plus convention in this work, as in Definition 1.

The four-point condition is a stronger constraint than the triangle inequality, and is the defining difference between a general metric and a tree metric. In other words, for a distance matrix to be a tree metric, it must satisfy not only the triangle inequality, but also the four-point condition. The implication is that general biological distance matrices do not necessarily give rise to phylogenetic trees: there may be biological processes with evolutionary behavior measured by differences, which may be captured by distance matrices (such as Hamming or Jukes–Cantor distances), however such processes may not necessarily be realized as phylogenetic trees since not all distance matrices satisfy the four-point condition. We now give two examples to illustrate this distinction.

Example 8. Let $N = 4$ and consider the matrix W with 0 on diagonal entries, and $w(1,2) = w(1,3) = w(1,4) = 1$, and $w(2,3) = w(2,4) = 1$, but $w(3,4) = 2$. Then

$$\begin{aligned} w(1,2) \odot w(3,4) &\not\leq w(1,3) \odot w(2,4) \boxplus w(1,4) \odot w(2,3) \\ w(1,2) + w(3,4) &\not\leq \max\big(w(1,3) + w(2,4),\ w(1,4) + w(2,3)\big) \\ 3 &\not\leq \max(2,2) = 2. \end{aligned}$$

Thus, (3) is not satisfied, and it is therefore impossible to construct a phylogenetic tree with 4 leaves such that for $1 \leq i < j \leq 4$, the length of the unique path connecting leaves i and j is equal to $w(i,j)$ as specified above.

Example 9. The tree metric $w \in \mathbb{R}^6$ for the tree in Figure 1 is $(w_{PQ}, w_{PR}, w_{PS}, w_{QR}, w_{QS}, w_{RS})$, expressed as a vector. As a matrix W, it is

$$\begin{pmatrix} 0 & w_{PQ} & w_{PR} & w_{PS} \\ & 0 & w_{QR} & w_{QS} \\ & & 0 & w_{RS} \\ & & & 0 \end{pmatrix} = \begin{pmatrix} 0 & a+b & a+c+d & a+c+e \\ & 0 & b+c+d & b+c+e \\ & & 0 & d+e \\ & & & 0 \end{pmatrix}.$$

The Plücker relations (2) associated with W are

$$\begin{aligned} A &:= w_{PQ} \odot w_{RS} = a+b+d+e, \\ B &:= w_{QR} \odot w_{PS} = a+b+2c+d+e, \\ C &:= w_{PR} \odot w_{QS} = a+b+2c+d+e. \end{aligned}$$

The maximum $B = C$ is achieved exactly twice, and $B - A = 2c > 0$. Also, (3) holds: $A \leq B \boxplus C = \max\{B, C\} = B$.

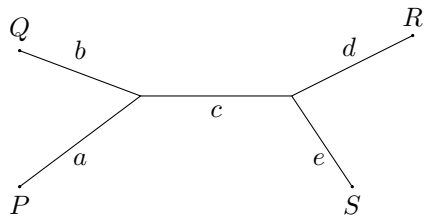

Figure 1: Example of an unrooted phylogenetic tree to illustrate the four-point condition.

2.2.2 Tree Ultrametrics

A strengthening of the triangle inequality for general metrics gives rise to an important class of tree metrics, which we will now discuss.

Definition 10. Let $w : [N] \times [N] \to \mathbb{R}_{\geq 0}$ be a metric. If $w(i,k) \leq w(i,j) \boxplus w(j,k)$ for each choice of distinct $i, j, k \in [N]$, then w is an *ultrametric*.

If w is a tree metric, and the maximum among

$$w(i,j), \quad w(i,k), \quad w(j,k),$$

is achieved at least twice for $1 \leq i < j < k \leq N$, then w satisfies the *three-point condition*, and hence is a *tree ultrametric* (Jardine et al., 1967). The space $\mathcal{U}_N$ of tree ultrametrics for phylogenetic trees with N leaves consists of all n-tuples $\{w(i,j)\}_{1 \leq i < j \leq N}$ satisfying the three-point condition for tree ultrametrics.

As in the case above for metrics to define tree metrics, the strengthening of the triangle inequality for a metric to an ultrametric applies to general metric functions, but analogously, in order for an ultrametric to be a tree ultrametric, it must satisfy the three-point condition. Trees satisfying the three-point condition imply also that the four-point condition is satisfied.

Definition 11. A rooted phylogenetic tree T with N leaves is called *equidistant* if the distance from every leaf to its root is a constant.

Proposition 12. *A dissimilarity map computed from all pairwise distances between all pairs of leaves in a phylogenetic tree T is a tree ultrametric if and only if T is equidistant.*

Proof. For any two points X, Y on a tree, we denote by $w(X, Y)$ the length of the unique path connecting X and Y. Suppose a tree T is equidistant with root R. Then for any three leaves A, B, C, we have that $w(R, A) = w(R, B) = w(R, C)$. Since R, A, B, C satisfies the four-point condition, the maximum among (2):

$$w(R, A) + w(B, C), \quad w(R, B) + w(A, C), \quad w(R, C) + w(A, B),$$

is attained at least twice. Thus, the maximum among $w(B, C), w(A, C), w(A, B)$ is also attained at least twice, satisfying the three-point condition, and T is therefore a tree ultrametric.

Conversely, suppose T is a tree ultrametric. Then there are finitely many leaves in T, so we can choose a pair of leaves A, B such that $w(A, B)$ is maximal among all such pairs. Along the unique path from A to B, there is a unique point R such that $w(R, A) = w(R, B)$. For any other leaf C, consider the distance $w(R, C)$. Since the paths from R to A and B only intersect at R, the path from R to C intersects at least one of them only at R. Suppose without loss of generality that the path from R to A is such a path. Then $w(A, C) = w(R, A) + w(R, C)$. Since $w(A, C) \leq w(A, B)$, we have $w(R, C) \leq w(R, B)$. If $w(R, C) < w(R, B)$, then $w(A, C) < w(R, A) + w(R, B) = 2w(R, B)$, and $w(B, C) \leq w(R, B) + w(R, C) < 2w(R, B)$, so the maximum among $w(A, B)$, $w(A, C)$, $w(B, C)$ is $w(A, C) = 2w(R, B)$ and it is only attained once—a contradiction, since T was assumed a tree ultrametric. Hence, $w(R, C) = w(R, B)$, and R has equal distance to all leaves of T. Therefore T is equidistant with root R. □

2.2.3 The Tropical Projective Torus

Given that the three-point condition is stronger than the four-point condition, we have that the space of tree ultrametrics is contained in the space of tree metrics, $\mathcal{U}_N \subset \mathcal{T}_N$, which itself is a subset of the *tropical projective torus*. Denoted by $\mathbb{R}^n/\mathbb{R}\mathbf{1}$, it is the quotient space generated by the following equivalence relation $\sim$ on $\mathbb{R}^n$:

$$x \sim y \Leftrightarrow x_1 - y_1 = x_2 - y_2 = \cdots = x_n - y_n.$$

In other words, for two points $x, y \in \mathbb{R}^n$, $x \sim y$ if and only if all coordinates of their difference $x - y$ are equal. In the context of trees, the quotient normalizes evolutionary time between trees; tree distances differ from tree metrics by scalar multiples of $\mathbf{1}$, the vector of all ones.

The tropical projective torus $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ embeds into $\mathbb{R}^{n-1}$ by considering representatives of the equivalence classes whose first coordinate is zero, $(x_2 - x_1, x_3 - x_1, \ldots, x_n - x_1) \in \mathbb{R}^{n-1}$.

The tropical projective torus $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ may also be generated by a group action. Let $G := \{(c, c, \ldots, c) \in \mathbb{R}^n \mid c \in \mathbb{R}\}$ with coordinate-wise addition, then G is an additive group. G acts on $\mathbb{R}^n$ as follows: for $g \in G$ and $x \in \mathbb{R}$,

$$g \circ x = (x_1 + g_1,\, x_2 + g_2,\, \ldots,\, x_n + g_n).$$

Each point in $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ is then exactly one orbit under the group action of G on $\mathbb{R}^n$.

2.3 Some Common Tree Metrics

In this framework of tree metrics and ultrametrics, we now provide some examples of tree metrics used in the literature on phylogenetic trees. The selection presented here is far from exhaustive; for a more comprehensive survey, see Weyenberg and Yoshida (2016) and references therein. An important metric that is particularly relevant to our work is the *BHV geodesic metric*, which will be discussed in greater detail, since much of the contributions presented in this paper are comparative to the BHV setting.

2.3.1 Inner Product Distances

The *path difference* d_P, *quartet distance* d_Q, and *Robinson–Foulds distance (or splits distance)* d_S are commonly-occurring tree distances, which can be formulated as a form of inner products of vectors. We

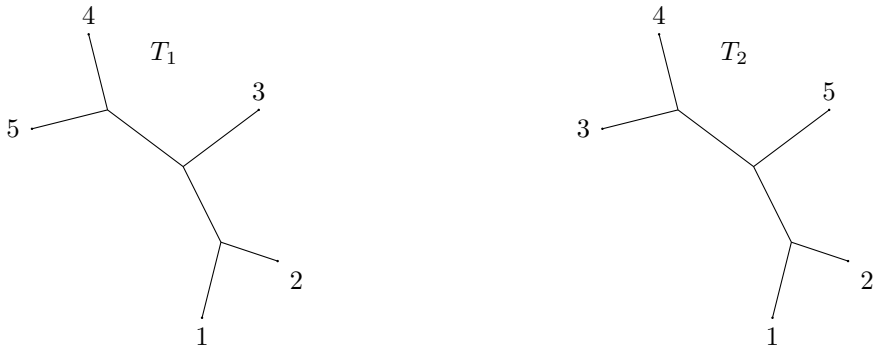

Figure 2: Example trees T_1 and T_2 for inner-product tree distances.

now briefly describe these distances and illustrate them with the running example of two trees T_1, T_2 depicted in Figure 2.

The *path difference* between two trees T and T' is the Euclidean distance

$$d_P(T, T') = \big\| v_P(T) - v_P(T') \big\|_2$$

where $v_P(T) \in \mathbb{R}^n$ is the vector whose (i, j)th entry counts the number of edges between leaves i and j in T Steel and Penny (1993). For example, for the trees T_1 and T_2 in Figure 2, the coordinates of $v_P(T_1)$ and $v_P(T_2)$ are given by

$$\big(v_{\{1,2\}},\, v_{\{1,3\}},\, v_{\{1,4\}},\, v_{\{1,5\}},\, v_{\{2,3\}},\, v_{\{2,4\}},\, \ldots,\, v_{\{4,5\}}\big),$$

where $v_{i,j}$ is the number of edges between leaf i and j. Thus,

$$\begin{aligned} v_P(T_1) &= (2, 3, 4, 4, 3, 4, 4, 3, 3, 2), \\ v_P(T_2) &= (2, 4, 4, 3, 4, 4, 3, 2, 3, 3), \end{aligned}$$

and

$$d_P(T_1, T_2) = \big\| v_P(T_1) - v_P(T_2) \big\|_2 = \sqrt{6}.$$

A *quartet* in a tree T is a subtree on four leaves induced by removing all other leaves from T. For each choice of 4 leaves, there are four possibilities for the tree topology of the induced quartet. Let $Q(T)$ denote the set of quartets induced by a tree T. The quartet distance $d_Q(T, T')$ is half the size of the symmetric difference of the trees' quartets,

$$d_Q(T, T') = \frac{1}{2} \# \big\{ \big(Q(T) - Q(T')\big) \cup \big(Q(T') - Q(T)\big) \big\}.$$

Referring to Figure 2, $d_Q(T_1, T_2) = 2$ since all quartets in T_1, T_2 are the same except for the quartets whose leaves are $\{1, 2, 3, 4\}$ and $\{1, 2, 4, 5\}$. Since d_Q is the square of a Euclidean distance, the distance $\sqrt{d_Q}$ is used as a matter of convention.

A *split* in a tree T is a bipartition of the leaves induced by removing one edge from T (splitting T into two trees defining the bipartition). Let $S(T)$ denote the set of all splits for a tree T. The Robinson–Foulds (or splits) distance (Robinson and Foulds, 1981) $d_S(T, T')$ is half the size of the symmetric difference of the trees' splits,

$$d_S(T, T') = \frac{1}{2} \# \big\{ \big(S(T) - S(T')\big) \cup \big(S(T') - S(T)\big) \big\}.$$

For example, for the trees T_1 and T_2 in Figure 2, we have $d_S(T_1, T_2) = 1$ since all splits in T_1, T_2 are the same, except for the splits obtained by removing the middle edge between 3 and 5. Since d_S is the square of a Euclidean distance, the distance $\sqrt{d_S}$ is also used by convention.

2.3.2 Cophenetic Metrics and Interleaving Distances

The copehentic vectorization of tree metrics (1) represents trees as points in the Euclidean space $\mathbb{R}^n$, which may be compared using the *cophenetic metric*, which is simply the ℓ^∞ distance (Cardona et al., 2013). In topological data analysis, the *interleaving distance* is constructed in the context of category theory and used to compare persistence modules. Recent work by Munch and Stefanou (2019) shows that the ℓ^∞-cophenetic metric is in fact an interleaving distance, connecting the fields of mathematical phylogenetics and topological data analysis.

2.3.3 The BHV Geodesic Metric

A pioneering contribution to the study of sets of phylogenetic trees is due to Billera, Holmes, and Vogtmann, who introduced a moduli space that considers equidistant rooted phylogenetic trees on a fixed set of leaves as points with geometry characterized by geodesics which is now commonly referred to in the literature as BHV space (Billera et al., 2001). Due to its intuitive and elegant construction, which then results in desirable properties of the geodesic, BHV space has been suggested and proposed as the decisive setting for statistics on phylogenetic tree space (Benner et al., 2014; Gavryushkin and Drummond, 2016).

In BHV space, trees are expressed only by the lengths of their internal edges. External or pendant edges are not considered, since taking them into account does not affect the geometry of the space: including external edges amounts to taking the product of tree space with an N-dimensional Euclidean space. Any two points (i.e. two rooted trees with a fixed set of leaves) are connected by a geodesic, with the distance between the points being defined by the length of the distance connecting them. In this paper, we denote the general BHV space of trees with N leaves by $\mathcal{T}_N^{\text{BHV}}$, and the case where rooted phylogenetic trees with N leaves with zero-length external edges and finite branch lengths from the root to the leaves are normalized to length 1 by $\mathcal{U}_N^{\text{BHV}} \subset \mathcal{T}_N^{\text{BHV}}$. The latter case of $\mathcal{U}_N^{\text{BHV}}$ is also widely studied (e.g. Gavryushkin and Drummond (2016), Section 5 of Holmes (2003), and Lin et al. (2017)), including in Section 4.4 of the original paper by Billera et al. (2001). The $\mathcal{U}_N^{\text{BHV}}$ standardization to unit edge lengths presents only a mild simplification of the geometry of $\mathcal{T}_N^{\text{BHV}}$ (Holmes, 2003). Geometrically, the local behavior of BHV space is intuitive and behaves like Euclidean space.

The construction of BHV space proceeds as follows: Consider a rooted tree with N leaves. Such a tree has at most $2N-2$ edges: N terminal edges, which are connected to leaves, and at most $N-2$ internal edges. When a rooted tree is *binary* (that is, it is a bifurcating tree that has exactly two descendants stemming from each interior node), then the number of edges is exactly $N-2$; the number of edges is lower than $N-2$ if it is not binary. To each distinct tree combinatorial type or topology, a Euclidean orthant of dimension $N-2$ (i.e. the number of internal edges) is associated. In this setting, an orthant may be regarded as the polyhedral cone of $\mathbb{R}^{N-2}$ where all coordinates are nonnegative. Thus, for each tree topology, the orthant coordinates correspond to the internal edge lengths in the tree. Since each of the coordinates in an orthant corresponds to an internal edge length, the orthant *boundaries* (where at least one coordinate is zero) represent trees with collapsed internal edges. These points can be thought of as trees with slightly different, though closely related, tree topologies. The BHV space is constructed by noting that the boundary trees from two different orthants may describe the same polytomic topology, i.e. a *split*, and thus grafting orthant boundaries together when the trees they represent coincide. The grafting of orthants always occurs at right-angles. The BHV space is thus the union of $(2N-3)!!$ polyhedra, each with dimension $N-2$. Each polyhedron may also be thought of as the cone $\mathbb{R}_{\geq 0}^{N-2}$. Note that an equivalent construction may also be outlined for $N-3$ internal edges for the case of unrooted trees.

Geodesics are unique in BHV space—a property based on the CAT(0) property of flag complexes, which is a result on the curvature of the space and derived from the right-angle grafting of cones. To compute a geodesic in $\mathcal{T}_N^{\text{BHV}}$, first, the geodesic distance between two trees on the BHV tree space is computed, and then the terminal branch lengths are considered to compute the overall geodesic distance between two trees, by taking the differences between terminal branch lengths. Since each orthant is locally viewed as a Euclidean space, the shortest path between two points within a single orthant is a straight Euclidean line. The difficulty appears in establishing which sequence of orthants joining the two topologies contains the geodesic. In the case of four leaves, this can be readily determined using a systematic grid search, but such a search is intractable with larger trees. Owen and Provan (2011) present a quartic-time algorithm (in the number of leaves N) for finding the geodesic path between any two points in BHV space, which is the

currently the fastest available method (although other methods also exist, such as those by Amenta et al. (2007) (which is an approximative method) and Kupczok et al. (2008)). The algorithm proceeds as follows: For $T_1, T_2 \in \mathcal{T}_N^{\text{BHV}}$ belonging to the same orthant, the geodesic between them is simply the Euclidean line segment connecting them in the same orthant. If not, then there exist k paths with finite line segments, $z_0 \to z_1 \to z_2 \to \cdots \to z_{k-1} \to z_k$ joining T_1 and T_2. For each $1 \leq i \leq k-1$, the node $z_i \in \mathcal{T}_N^{\text{BHV}}$ lies on the intersection of two orthants in $\mathcal{T}_N^{\text{BHV}}$ in such a manner that any two consecutive nodes along the geodesic belong to a common orthant in $\mathcal{T}_N^{\text{BHV}}$. Once the geodesic is known, its length, and thus the distance between the trees, is readily computable: the length of this path is simply the sum of the lengths of these line segments. In this paper, we denote the BHV metric by d_{BHV}. Complete details on the computation of geodesic distances are given in Owen and Provan (2011). Intuitively, for two points that are not in the same orthant, the shortest path between them depends on whether the points are in neighboring orthants or not. In the former case, then the shortest path from one point to the other is through the unique boundary that separates them; in the latter case, the shortest path passes through the origin (i.e. the tree where all internal edges are of zero length, or the *star tree*) referred to as the *cone path*.

2.4 Relating Tropical Geometry to Phylogenetic Trees

The connection of tropical geometry to phylogenetic tree space arises in two occurrences. Classically, as discussed in Maclagan and Sturmfels (2015), the space of phylogenetic trees with N leaves is the moduli space of tropical curves of genus 0 and N marked points. A more recent result upon which our work is based identifies a homeomorphism between the space of phylogenetic trees with N leaves and a tropical construction of the Grassmannian of 2-planes in N dimensions (Speyer and Sturmfels, 2004). The Grassmannian is the space of all k-dimensional subspaces of an n-dimensional vector space, with $k < n$, which can be mapped to a projective variety—a particular zero set of polynomials—via a construction known as the *Plücker embedding*. Tropicalizing the Plücker embedding for the Grassmannian of 2-planes in N dimensions recovers the four-point condition (see Definition 6) completely and explicitly describing phylogenetic trees with N leaves as tree metrics. Essentially, this result endows the space of phylogenetic trees with a tropical structure: the homeomorphism provides a purely tropical parameterization of the space of phylogenetic trees via the Plücker relations given in Definition 6. In this work, we study characteristics of this tropical structure for computational and especially data-analytic applications involving sets of phylogenetic trees.

Other Applications of Tropical Geometry

Tropical geometry also arises in other applied settings, specifically in computer science, statistics, and economics: Maclagan and Sturmfels (2015) and Pachter and Sturmfels (2005) discuss the use of tropical mathematics to reinterpret the dynamic programming approach to the problem of sequence alignment for molecular data in computational biology. In the context of statistics—specifically algebraic statistics—tropical geometry arises in the reinterpretation of various stochastic models. As a field of research, one of the core principles of algebraic statistics is the fact that algebraic varieties and semi-algebraic sets are statistical models (Améndola et al., 2018; Sullivant, 2018). For statistical models, such as graphical models, tropicalized statistical models (i.e. tropicalized algebraic varieties) are fundamental in parametric inference, which was specifically demonstrated on hidden Markov models and general Markov models on binary trees (Pachter and Sturmfels, 2004). In computational biology, this algebraic statistical approach was adapted to study invariants of joint probability distributions on leaf labels of phylogenetic trees (Sturmfels and Sullivant, 2005). In economics and finance, tropical geometry arises in game theoretic settings and max-linear models for financial data (Einmahl et al., 2018; Lin and Tran, 2019).

3 Palm Tree Space

Our main contribution in this section is a metric measure space construction for the set of all phylogenetic trees. Here, we prove geometric, topological, and analytic properties of the space that are not only desirable for computation, but also turn out to result in probabilistic and statistical advantages. We also give a comparison of these properties relative to BHV space. We begin by motivating our study.

3.1 Motivation: Statistics in BHV Space

While results enabling statistical analysis on BHV space exist (e.g. Benner et al. (2014); Holmes (2005)), they are not always straightforward to implement nor interpret. There are two important reasons underlying these difficulties. First, polytopes in BHV space with edges defined by geodesics, where the geodesics are calculated by the Owen–Provan algorithm (Owen and Provan, 2011), are unbounded in dimension (Lin et al., 2017). In the most general case, these can be polytopes whose edges are all cone paths, which have maximal *depth*. The depth of a point x in the relative interior of some k-dimensional polyhedron is $N-2-k$; the depth of a geodesic is the largest depth among all point son the geodesic, or, in other words, the largest codimension of all cones traversed.

Corollary 13 (Lin et al. (2017))**.** *There exist three ultrametric trees with* $2d+2$ *leaves whose geodesic triangle in ultrametric BHV space* $\mathcal{U}^{\text{BHV}}_{2d+2}$ *has dimension at least* d*,* $d \geq 1$*.*

The data-analytic implication is that there exist no useful, lower-dimensional representations (projections) of data on BHV planes. This is restrictive for standard dimensionality-reduction techniques in classical statistics and machine learning, such as principal component analysis (PCA), which is also an important procedure in descriptive statistics used to visualize data and reduce dimensionality. Sophisticated methods have since been developed to bypass these limitations (Feragen et al., 2013; Nye, 2011; Nye et al., 2017).

Large depths of geodesics also give rise to the phenomenon of *stickiness* (Miller et al., 2015), where the Fréchet mean lies stably at a singularity. The statistical implication is that perturbing any of the data points results in no change in the mean, and thus, there exists no exact asymptotic distribution for the mean. This is the second difficulty with statistical analyses and parametric inference on BHV space, which has been well-studied in various settings, including BHV space in particular (Barden and Le, 2018; Hotz et al., 2013), but also in more general settings of hyperbolic planar singularities (Huckemann et al., 2015). This issue has been bypassed in approximations using the delta method to obtain asymptotic behavior and yield a central limit theorem (Barden et al., 2013), but parametric statistical inference in classical statistics, such as hypothesis testing, fundamentally depends on an exact asymptotic distribution. Approximate and other methods that are nonparametric in nature have been proposed which leverage the results by Barden et al. (2013) (e.g. Willis (2019); Willis and Bell (2018)). It should be noted, however, that stickiness only occurs when the Fréchet mean lies on a lower-dimensional orthant (see Example 15 that follows).

Definition 14 (Stephan Huckemann, personal communication)**.** Let $\mathcal{M}$ be a space of probability measures on a topological space Q, and suppose that $\mathcal{M}$ is endowed with a topology. Further, consider a map $E : \mathcal{M} \to \mathbf{Q}$ to closed subsets of Q. Then a measure $\mu \in \mathcal{M}$ is said to *stick* to a set $M \subseteq Q$ if $E(\mu) = M$, and for every neighborhood U' of μ in $\mathcal{M}$, there is an open subset $U \subset U'$ such that $E(\nu) \subseteq M$ for all $\nu \in U$.

Example 15. In Figure 3, we position three unit masses on the 3-spider, which is the stratified space of three $\mathbb{R}_{\geq 0}$ rays joined at the origin. This is precisely the BHV space of phylogenetic trees with three leaves and fixed pendant edge lengths, since, as mentioned previously, coordinates in BHV space are given only by internal edge lengths. The position x of the barycenter (Fréchet mean) is calculated by minimizing $2(1+x)^2 + (a-x)^2$. The solution is $x = 0$ for $a < 2$, and $x = (a-2)/3$ for $a \geq 2$. The Fréchet mean tends to stick to lower-dimensional strata.

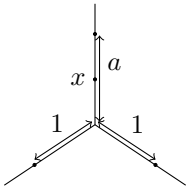

Figure 3: 3-spider to illustrate stickiness.

These limitations have been the motivation for the recent work adopting the tropical interpretation of tree space (Lin et al., 2017; Yoshida et al., 2019). Our contribution in the present paper is a formal study of mathematical properties of the metric geometry of this interpretation, and specifically, those that give rise to desirable settings for studies in probability and statistics.

3.2 The Tropical Metric

We define an intrinsically tropical generalized Hilbert projective metric function (Akian et al., 2011; Cohen et al., 2004), d_{tr}, on the tropical projective torus $\mathbb{R}^n/\mathbb{R}\mathbf{1}$. This metric has been used in recent literature on tropical geometric tree spaces (Lin et al., 2017; Yoshida et al., 2019).

Definition 16. For any point $x \in \mathbb{R}^n$, denote its coordinates $x_1, x_2, \ldots, x_n$ and its representation in $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ by $\bar{x}$. For $\bar{x}, \bar{y} \in \mathbb{R}^n/\mathbb{R}\mathbf{1}$, set the distance between $\bar{x}$ and $\bar{y}$ to be

$$\begin{aligned} d_{\mathrm{tr}}(\bar{x}, \bar{y}) &:= \max_{1 \le i < j \le n} \left|(x_i - y_i) - (x_j - y_j)\right| \\ &= \max_{1 \le i \le n} (x_i - y_i) - \min_{1 \le i \le n} (x_i - y_i). \end{aligned}$$

We refer to the function d_{tr} as the *tropical metric*.

Proposition 17. *The function d_{tr} is a well-defined metric function on $\mathbb{R}^n/\mathbb{R}\mathbf{1}$.*

Proof. We verify that the defining properties of metrics are satisfied. By definition, for $u, v \in \mathbb{R}^n$, $d_{\mathrm{tr}}(\bar{u}, \bar{v}) = d_{\mathrm{tr}}(\bar{v}, \bar{u})$, satisfying symmetry. The tropical metric is nonnegative, since $\left|(u_i - v_i) - (u_j - v_j)\right| \ge 0$, so is $d_{\mathrm{tr}}(\bar{u}, \bar{v}) \ge 0$. If $d_{\mathrm{tr}}(\bar{u}, \bar{v}) = 0$, then $u_i - v_i$ are equal for all $1 \le i \le n$, thus $\bar{u} = \bar{v}$, so indiscernibles are identifiable.

For $u, v, w \in \mathbb{R}^n$, we now show that triangle inequality is satisfied: $d_{\mathrm{tr}}(\bar{u}, \bar{w}) \le d_{\mathrm{tr}}(\bar{u}, \bar{v}) + d_{\mathrm{tr}}(\bar{v}, \bar{w})$. Suppose $1 \le i' < j' \le n$ such that $\left|(u_{i'} - w_{i'}) - (u_{j'} - w_{j'})\right| = \max_{1 \le i < j \le n} |u_i - w_i - u_j + w_j|$, then $d_{\mathrm{tr}}(\bar{u}, \bar{w}) = |u_{i'} - w_{i'} - u_{j'} + w_{j'}|$. Note that

$$u_{i'} - w_{i'} - u_{j'} + w_{j'} = (u_{i'} - v_{i'} - u_{j'} + v_{j'}) + (v_{i'} - w_{i'} - v_{j'} + w_{j'}).$$

Hence

$$\begin{aligned} d_{\mathrm{tr}}(\bar{u}, \bar{w}) = |u_{i'} - w_{i'} - u_{j'} + w_{j'}| &\le |u_{i'} - v_{i'} - u_{j'} + v_{j'}| + |v_{i'} - w_{i'} - v_{j'} + w_{j'}| \\ &\le d_{\mathrm{tr}}(\bar{u}, \bar{v}) + d_{\mathrm{tr}}(\bar{v}, \bar{w}). \end{aligned}$$

Thus, d_{tr} is a metric function on $\mathbb{R}^n/\mathbb{R}\mathbf{1}$. □

With this metric, we formally define palm tree space as follows.

Definition 18. For a positive integer N, let $\mathcal{T}_N$ be the space of phylogenetic trees as in Definition 6. The metric space $\mathcal{P}_N := (\mathcal{T}_N, d_{\mathrm{tr}})$ is called the *palm tree space*.

Remark 19. Palm tree space is defined in the general setting of non-equidistant trees (see Definition 11), in which case the rootedness of the trees is not a significant matter. When there is no condition of equidistance, there is no constraint on pairwise distances between leaves in the tree and thus any point in the tree may be treated as a root. In other words, palm tree space may be defined in terms of either rooted or unrooted non-equidistant trees.

The spaces $\mathcal{T}_N^{\mathrm{BHV}}$ and $\mathcal{P}_N$ are not isometric, however, the tropical metric d_{tr} is nevertheless stable. This means that perturbations of points in BHV space, measured by the BHV metric d_{BHV}, correspond to bounded perturbations of their images in palm tree space, measured by the tropical metric. This stability property is desirable, since it allows for interpretable comparisons between the two spaces, and allows for "translations" in the widely-used BHV framework over to palm tree space.

The following lemma ensures coordinate-wise stability of the tropical metric in $\mathcal{P}_N$.

Lemma 20. *Let $u \in \mathbb{R}^n$. For $1 \le i \le n$, if we perturb the ith coordinate of u by ε to obtain another point $u' \in \mathbb{R}^n$, then in $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ we have*

$$d_{\mathrm{tr}}(\bar{u}, \bar{u}') = |\varepsilon|.$$

Proof. For $1 \le j \le n$, the difference $u'_j - u_j = 0$ if $j \ne i$, and $u'_i - u_i = \pm\varepsilon$. The set of these differences is then either $\{0, \varepsilon\}$ or $\{0, -\varepsilon\}$. By Definition 16, $d_{\mathrm{tr}}(\bar{u}, \bar{u}') = |0 - \pm\varepsilon| = |\varepsilon|$. □

Theorem 21 (Stability)**.** *Let N be the number of leaves in palm tree space and BHV space. Let u and u' be two unrooted phylogenetic trees with N leaves. Then the following inequality holds:*

$$d_{\mathrm{tr}}(u,u') \leq \sqrt{N+1}\cdot d_{\mathrm{BHV}}(u,u').$$

Moreover, the smallest possible constant is $\sqrt{N+1}$.

Proof. We first prove that for any two unrooted trees u, u' with N leaves, $d_{\mathrm{tr}}(u,u') \leq \sqrt{N+1}\cdot d_{\mathrm{BHV}}(u,u')$. First, assume that u, u' belong to the same orthant in BHV space. Then no matter what the tree topology is, if we denote the differences of the lengths of the $N-3$ internal edges in u and u' by $d_1, d_2, \ldots, d_{N-3}$, and the differences of the length of the N external edges by $p_1, p_2, \ldots, p_N$, we always have

$$d_{\mathrm{BHV}}(u,u') = \sqrt{\sum_{i=1}^{N-3} d_i^2 + \sum_{i=1}^{N} p_i^2}$$

(e.g. Lin et al. (2017); Owen and Provan (2011)).

For every pair of leaves i, j in both trees, the distance between them is a sum of the length of some internal edges and two external edges. In other words, all differences $w^u_{\{i,j\}} - w^{u'}_{\{i,j\}}$ are of the form of the sum between some d_k, and $p_i + p_j$. Thus, the maximum of these differences is at most the sum of all positive d_i values, plus the two greatest p_i values (take these to be p_{i_1} and p_{i_2}), while the minimum of these differences is at least the sum of all negative d_i values, plus two smallest p_i values (take these to be p_{i_3} and p_{i_4}). By definition, $d_{\mathrm{tr}}(u,u')$ is the maximum minus the minimum of these differences, so we have

$$d_{\mathrm{tr}}(u,u') \leq \sum_{i=1}^{N-3} |d_i| + |p_{i_1}| + |p_{i_2}| + |p_{i_3}| + |p_{i_4}|.$$

By the Cauchy–Schwarz inequality (Cauchy, 1821; Schwarz, 1890),

$$(N+1)\cdot\left(\sum_{i=1}^{N-3} |d_i|^2 + |p_{i_1}|^2 + |p_{i_2}|^2 + |p_{i_3}|^2 + |p_{i_4}|^2\right) \geq \left(\sum_{i=1}^{N-3} |d_i| + \sum_{i=1}^{N} |p_i|\right)^2.$$

Hence

$$\begin{aligned}
d_{\mathrm{tr}}(u,u') &\leq \sum_{i=1}^{N-3} |d_i| + |p_{i_1}| + |p_{i_2}| + |p_{i_3}| + |p_{i_4}| \\
&\leq \sqrt{N+1}\cdot\sqrt{\sum_{i=1}^{N-3} |d_i|^2 + |p_{i_1}|^2 + |p_{i_2}|^2 + |p_{i_3}|^2 + |p_{i_4}|^2} \\
&\leq \sqrt{N+1}\cdot\left(\sum_{i=1}^{N-3} |d_i|^2 + \sum_{i=1}^{N} p_i^2\right) \\
&= \sqrt{N+1}\cdot d_{\mathrm{BHV}}(u,u').
\end{aligned}$$

Now, for u, u' with distinct tree topologies, we consider the unique geodesic connecting them: there exist finitely many points $u^1, \ldots, u^{k-1}$ in BHV space such that u^i and u^{i+1} belong to the same polyhedron corresponding to a tree topology for $0 \leq i \leq k-1$, where $u^0 = u$ and $u^k = u'$, and $d_{\mathrm{BHV}}(u,u') = \sum_{i=0}^{k-1} d_{\mathrm{BHV}}(u^i, u^{i+1})$. For $1 \leq i \leq k-1$, by the proof above, we have that

$$d_{\mathrm{tr}}(u^i, u^{i+1}) \leq \sqrt{N+1}\cdot d_{\mathrm{BHV}}(u^i, u^{i+1}) \quad \forall \quad 1 \leq i \leq k-1.$$

Thus,

$$d_{\mathrm{tr}}(u,u') \leq \sum_{i=0}^{k-1} d_{\mathrm{tr}}(u^i, u^{i+1}) \leq \sum_{i=0}^{k-1} \sqrt{N+1}\cdot d_{\mathrm{BHV}}(u^i, u^{i+1}) = \sqrt{N+1}\cdot d_{\mathrm{BHV}}(u,u').$$

Next, we consider the case where the equality holds: consider two trees T_1 and T_2 with N leaves and the same tree topology, given by the following nested sets

$$\big\{\{1,2\},\{1,2,3\},\ldots,\{1,2,\ldots,N-2\}\big\}.$$

Suppose in T_1, the internal edges have lengths

$$b^1(e_i)=\begin{cases}2, & \text{if } 1\le i\le N-4;\\ 1, & \text{if } i=N-3.\end{cases}$$

Similarly, in T_2, the internal edges have lengths

$$b^2(e_i)=\begin{cases}1, & \text{if } 1\le i\le N-4;\\ 2, & \text{if } i=N-3.\end{cases}$$

The external edge lengths of T_1 and T_2 are

$$p_j^i=\begin{cases}1, & \text{if } (i,j)=(1,2),(1,N-2),(2,N-1),(2,N);\\ 0, & \text{otherwise.}\end{cases}$$

Then

$$d_{\mathrm{BHV}}(T_1,T_2)=\sqrt{(N-4)\cdot(2-1)^2+(1-2)^2+2\cdot(1-0)^2+2\cdot(0-1)^2}=\sqrt{N+1}.$$

For $1\le i<j\le N$, in either tree the distance $w_{\{i,j\}}$ is the sum of the edge lengths of

$$p_i, e_{\max(i-1,1)}, e_{\max(i-1,1)+1}, \ldots, e_{\max(j-2,N-3)}, p_j.$$

Since $b^1(e_i)>b^2(e_i)$ for $i<N-3$ and $b^1(e_i)<b^2(e_i)$ for $i=N-3$, the maximum of all differences $w^1_{\{i,j\}}-w^2_{\{i,j\}}$ is

$$w^1_{\{2,N-2\}}-w^2_{\{2,N-2\}}=\big((N-4)\cdot 2+2\cdot 1\big)-(N-4)\cdot 1=N-2;$$

and the minimum of all differences $w^1_{\{i,j\}}-w^2_{\{i,j\}}$ is

$$w^1_{\{N-2,N-1\}}-w^2_{\{N-2,N-1\}}=1-(2+1+1)=-3.$$

By definition, $d_{\mathrm{tr}}(T_1,T_2)=(N-2)-(-3)=N+1=\sqrt{N+1}\cdot d_{\mathrm{BHV}}(T_1,T_2)$ in this case. Thus, $\sqrt{N+1}$ is the smallest possible stability constant. □

There are several important remarks concerning this stability result to discuss.

Remark 22. The stability constant $\sqrt{N+1}$ is the best possible constant, however, it is not universal, since it depends on the number of leaves in a tree. For a fixed number N of leaves, however, the value is indeed constant.

Remark 23. It is important to note here that explicit calculations involving geodesics between trees in the original paper by Billera et al. (2001) do not include pendant (external) edges, since these do not modify the geometry of the space. Indeed, their inclusion only amounts to an additional Euclidean factor, since the tree space then becomes the cross product of BHV space of trees with internal edges only, and $\mathbb{R}^N_{\ge 0}$. Geodesic distances, which depend directly on geodesic paths (the former is the length of the latter), considered in Billera et al. (2001) also do not include pendant edges.

The statement of Theorem 21 treats the most general case of unrooted trees with pendant edges included in both tree spaces. Our reasoning for considering this case, which differs from what appears in Billera et al. (2001) as described above, is twofold. First, the procedure for calculating geodesic distances between trees in BHV space described in Section 2.3.3 follows Owen and Provan (2011), who explicitly consider pendant edges in computing geodesics, and whose algorithm is considered to be the current standard. Since the tropical metric includes pendant edges in its calculation, a study relative to the Owen–Provan algorithm, rather than the procedure given in Billera et al. (2001) where pendant edges are excluded, provides a more valid comparison. Second, the present paper builds upon previous work on the tropical metric in Lin et al. (2017), which also deals with comparative studies relative to the Owen–Provan algorithm. We follow in this same vein for consistency.

In terms of statistical interpretation, Theorem 21 provides an important comparative measure and guarantees that quantitative results from BHV space are bounded in palm tree space. For example, in single-linkage clustering, where clusters are fully determined by distance thresholds, the stability result means that a given clustering pattern in BHV space will be preserved in palm tree space, thus maintaining interpretability of clustering behavior.

3.3 Geometry of Palm Tree Space

The uniqueness property of geodesics in BHV space, used in the proof of Theorem 21, leads naturally to the study of similar geometric properties that characterize palm tree space as well as important differences between the two spaces. These characteristics will now be developed in this section.

3.3.1 Geodesics in Palm Tree Space

Definition 24. A *geodesic* between any two points x and y in a metric space is a path between x and y with the shortest length (measured by the metric) that is fully contained within the metric space.

In palm tree space, geodesics are in general not unique, which is a common occurrence in various metric spaces.

Example 25. Consider the following union of three 2-dimensional orthants

$$O := \big\{(x, y, z) \in \mathbb{R}^3 \mid x, y, z \geq 0, xyz = 0\big\},$$

where the distance function is the usual Euclidean metric within each orthant. For two points $v, w \in O$ where $v = (2, 0, 0)$ and $w = (0, 2, 2)$, there exist two shortest paths: one that connects v to w by passing through $(0, 1, 0)$, and the other passing through $(0, 0, 1)$. The length of both paths is

$$\sqrt{(2-0)^2 + (0-1)^2} + \sqrt{(2-1)^2 + (2-0)^2} = 2\sqrt{5}.$$

There exists, however, a unique path joining two points in palm tree space, which is also a geodesic—the tropical line segment.

Definition 26. Given $x, y \in \mathbb{R}^n/\mathbb{R}\mathbf{1}$, the *tropical line segment* with endpoints x and y is the set

$$\{a \odot x \boxplus b \odot y \in \mathbb{R}^n/\mathbb{R}\mathbf{1} \mid a, b \in \mathbb{R}\}.$$

Here, max-plus addition $\boxplus$ for two vectors is performed coordinate-wise.

Proposition 27. *For points $x, y \in \mathcal{P}_N$, the tropical line segment connecting x and y is a geodesic.*

Proof. It suffices to show that for any $a, b \in \mathbb{R}$, we have that

$$d_{\mathrm{tr}}(z, x) + d_{\mathrm{tr}}(z, y) = d_{\mathrm{tr}}(x, y),$$

where $z = a \odot x \boxplus b \odot y$ is the tropical line segment. We may assume that $x_i - y_i \leq x_{i+1} - y_{i+1}$ for $1 \leq i \leq n-1$. Under this assumption, $d_{\mathrm{tr}}(x, y) = (x_n - y_n) - (x_1 - y_1)$. Now if $0 \leq j \leq n$ is the largest index such that $x_j - y_j \leq b - a$, then for some $i \geq j+1$, $z_i = b + y_i$ and, analogously, $z_i = a + x_i$. If $j = 0$ or $j = n$, then z is equal to either x or y and the claim is apparent. We may thus assume $1 \leq j \leq n-1$.

The set of all differences $x_i - z_i$ contains $-a$ and the greater values $x_i - y_i - b > -a$ for $i \geq j+1$. So,

$$d_{\mathrm{tr}}(z, x) = (x_n - y_n - b) - (-a) = (x_n - y_n) + (a - b).$$

Similarly, the set of all differences $z_i - y_i$ contains b and the smaller values $(x_i - y_i) + a < b$ for $i \leq j$. So,

$$d_{\mathrm{tr}}(z, y) = b - (x_1 - y_1 + a) = (b - a) - (x_1 - y_1).$$

Therefore, $d_{\mathrm{tr}}(z, x) + d_{\mathrm{tr}}(z, y) = d_{\mathrm{tr}}(x, y)$, and the tropical line segment connecting x and y is a geodesic. □

In addition, it turns out that tropical line segments are easy and fast to compute. In particular, the time complexity to compute them is lower than that of the Owen–Provan algorithm, which is currently the fastest available algorithm to compute the unique geodesics in BHV space with a time complexity of $O(N^4)$.

Proposition 28. *(Maclagan and Sturmfels, 2015, Proposition 5.2.5) The time complexity to compute the tropical line segment connecting two points in $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ is $O(n \log n) = O(N^2 \log N)$.*

3.3.2 Structure of Palm Tree Space

In the same way that $\mathcal{T}_N^{\text{BHV}}$ is constructed as the union of polyhedra, the geometry of $\mathcal{P}_N$ is also given by such a union.

Proposition 29. *(Maclagan and Sturmfels, 2015, Proposition 4.3.10) The space $\mathcal{T}_N$ is the union of $(2N-5)!!$ polyhedra in $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ with dimension $N-3$.*

Here, the cone in consideration is the usual convex cone as a subset of a vector space on an ordered field, closed under linear combinations with positive coefficients (or equivalently, the set spanned by conical combinations of vectors).

Example 30. When $N = 5$, $\mathcal{T}_5$ has $5!! = 15$ cones, which correspond to the edges of the *Petersen graph*, depicted in Figure 4. Here, the root is considered to be the leaf labeled 0, and the other leaf labels are numbered 1 through 4. The nodes of the graph represent the commonality between the coinciding edges.

To interpret the correspondence to trees in this Petersen graph, consider the leftmost upper graph vertex on the outer hexagon and notice that there are three edges (cones) that meet at this vertex. The figures of the trees associated with these cones (illustrated in the circles) share the property that the edge to leaf 1 is the longest edge, and the remaining three leaves are permuted. Now, consider the graph vertex at the very center of the graph, inside the hexagon: the property common to the trees associated with the three cones meeting at this vertex is that the pair of leaves labeled 1 and 3 are coupled, symmetric, and are joined at the same internal node of the tree that is not the root. The remaining graph vertices may be interpreted in a similar manner, in the sense that they all share one of these two commonalities, under symmetry of and up to leaf labeling scheme. Thus, the 10 vertices of the Petersen graph here are given by the number of ways to choose a label for the longest branch length in a tree that connects directly to the root, among 4 choices of leaf labels (i.e. $\binom{4}{2}$); and the number of ways to choose pairs of leaf labels that are coupled, symmetric, and correspond to the same internal node of the tree that is not directly linked to the root (i.e. $\binom{4}{1}$), so $\binom{4}{2} + \binom{4}{1} = 10$. The intuition of the graph edges as cones lies in the property of closedness under scaling of branch lengths, for each tree type associated with each graph edge in the figure (illustrated in the circles).

The Petersen graph also coincides in the context of BHV space when considering $N = 4$ leaves, as the so-called *link of the origin*; see Billera et al. (2001) for further details.

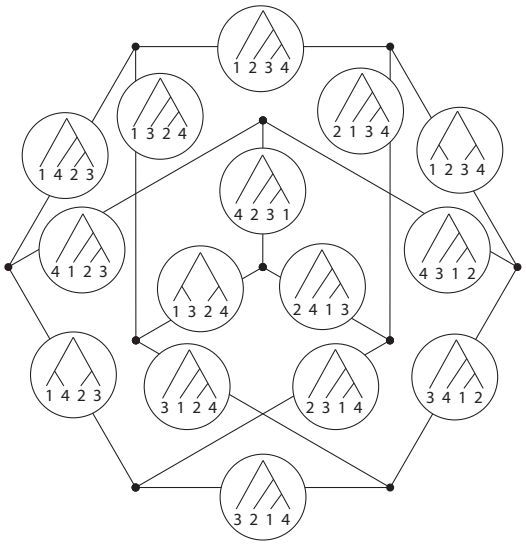

Figure 4: The Petersen graph depicting the 15 cones in $\mathcal{T}_5$.

3.4 Topology of Palm Tree Space

The measure of a space is relevant in probabilistic studies; the measure of a topological space, in particular, results in a desirable compatibility where the topology of a space may be interpreted in terms of measures (for

example, Radon measures may also be interpreted as linear functionals on the space of continuous functions with compact support, which is locally convex, by e.g. Bourbaki (2004), Chapter 3). This motivates our study of the topology of palm tree space.

The following two lemmas allow us to characterize the topology of palm tree space. Recall that for $x \in \mathbb{R}^n$, the set $B(x,r) = \{y \in \mathbb{R}^n \mid |y-x| < r\}$ is the open ball centered at x with radius r. By identifying $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ with $\mathbb{R}^{n-1}$ as discussed in Section 2.2.3, an equivalent set may be correspondingly defined in palm tree space as follows.

Definition 31. Under the tropical metric d_{tr}, we define $B_{\mathrm{tr}}(x,r) = \{y \in \mathbb{R}^n \mid d_{\mathrm{tr}}((0,\bar{y}),(0,\bar{x})) < r\}$ to be the open *tropical ball* centered at x with radius r.

Lemma 32. *For $x \in \mathbb{R}^{n-1}$ and $r > 0$, the open tropical ball $B_{\mathrm{tr}}(x,r)$ is the open convex polytope defined by the following strict inequalities for $1 \le i < j \le n-1$:*

$$\begin{aligned} y_i &> x_i - r, \\ y_i &< x_i + r, \\ y_i - y_j &> x_i - x_j - r, \\ y_i - y_j &< x_i - x_j + r. \end{aligned} \tag{4}$$

Proof. For $y \in \mathbb{R}^{n-1}$, $y \in B_{\mathrm{tr}}(x,r)$ if and only if $d_{\mathrm{tr}}((0,\bar{x}),(0,\bar{y})) < r$. Definition 16 admits the strict inequalities in (4). □

Lemma 33. *For $r > 0$ and $x \in \mathbb{R}^{n-1}$, $B(x,r) \subseteq B_{\mathrm{tr}}(x,2r)$ and $B_{\mathrm{tr}}(x,r) \subseteq B(x,\sqrt{n-1}r)$.*

Proof. By Lemma 32, if a point y lies in $B_{\mathrm{tr}}(x,r)$, then for $1 \le i \le n-1$, $|y_i - x_i| < r$, thus $y \in B(x,\sqrt{n-1}r)$. Conversely, if a point y lies in $B(x,r)$, then for $1 \le i \le n-1$, we have that $|y_i - x_i| < r$. Therefore,

$$\big|(y_i - y_j) - (x_i - x_j)\big| = \big|(y_i - x_i) - (y_j - x_j)\big| < 2r.$$

Hence $y \in B_{\mathrm{tr}}(x,2r)$. □

Theorem 34. *On $\mathbb{R}^{n-1}$, the family of open balls $B(x,r)$ and the family of open tropical balls $B_{\mathrm{tr}}(x,r)$ define the same topology.*

Proof. Suppose for all $r > 0$ and $x \in \mathbb{R}^{n-1}$ that the open balls $B(x,r)$ form a topological basis. For any $y \in \mathbb{R}^{n-1}$ and $s > 0$, we consider the ball $B_{\mathrm{tr}}(y,s)$: For any point $z \in B_{\mathrm{tr}}(y,s)$, we have that $d_{\mathrm{tr}}(z,y) < s$. Let $\varepsilon = \dfrac{s - d_{\mathrm{tr}}(z,y)}{2} > 0$. Then $B_{\mathrm{tr}}(z,2\varepsilon) \subseteq B_{\mathrm{tr}}(y,s)$. By Lemma 33, we have $B(z,\varepsilon) \subseteq B_{\mathrm{tr}}(z,2\varepsilon) \subseteq B_{\mathrm{tr}}(y,s)$. Therefore, $B_{\mathrm{tr}}(y,s)$ is also an open set. The other direction is proved in the same manner. □

Example 35. Figure 5 illustrates the unit balls in Euclidean, BHV, and palm tree space. Here, the number of leaves is fixed to be 3. There are three 1-dimensional cones in BHV space, and they share the origin. The palm tree space $\mathcal{P}_3 = \{w = (w_{\{1,2\}}, w_{\{1,3\}}, w_{\{2,3\}}) \in \mathbb{R}^3/\mathbb{R}\mathbf{1} \mid \max(w) \text{ is attained at least twice}\}$ may be embedded in $\mathbb{R}^2$.

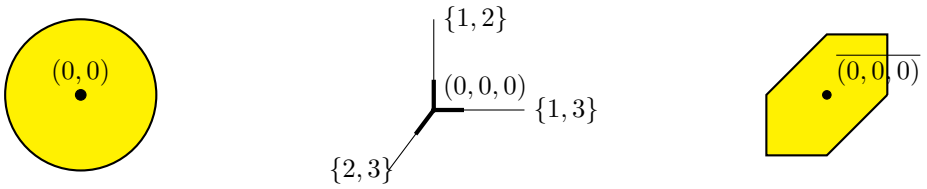

Figure 5: Comparison of unit balls in Euclidean, BHV, and palm tree space for $N = 3$ leaves. The leftmost figure is the Unit ball $B((0,0),1)$ in $\mathbb{R}^2$; the center figure is the unit ball centered at the origin with radius 1 in a BHV space with 3 leaves; the rightmost figure is the unit ball $B_{\mathrm{tr}}(\overline{(0,0,0)},1)$ in $\mathcal{P}_3$.

3.5 Palm Tree Space is a Polish Space

We have thus far established palm tree space as a metric space when equipped with the tropical metric, and moreover, established its topology. In this section, we prove additional analytic properties of palm tree space that are desirable for probabilistic and statistical analysis. Specifically, we prove that palm tree space is a separable, completely metrizable topological space, and thus a Polish space, by definition.

Polish spaces are important settings for studies in probability due to the fact that classical results maybe formulated and generalized in a well-behaved manner; some examples are the construction of conditional expectations, Kolmogorov's extension theorem (which guarantees the definition of a stochastic process from a series of finite-dimensional distributions), and Prokhorov's theorem (which guarantees weak convergence by relating tightness of measures to compactness in a probability space) (Parthasarathy, 1967).

Proposition 36. *$\mathcal{P}_N$ is complete.*

Proof. For convenience, when considering points in $\mathcal{P}_N$, we always choose their unique preimage in $\mathbb{R}^n$ whose first coordinate is 0. Then, we may denote each point in $\mathcal{P}_N$ by an $(n-1)$-tuple in $\mathbb{R}^{n-1}$. Let $T_1, T_2, \ldots \in \mathbb{R}^{n-1}$ be a Cauchy sequence of points in $\mathcal{P}_N$. For $1 \leq i \leq n-1$, we claim that $(T_k^i)_{k\geq 1}$ also form a Cauchy sequence in $\mathbb{R}$: For any $\varepsilon > 0$, there exists M such that for $k_1, k_2 > M$, we have $d_{\text{tr}}(T_{k_1}, T_{k_2}) < \varepsilon$. By Definition 16, $d_{\text{tr}}(T_{k_1}, T_{k_2}) \geq |0 - 0 - T_{k_2}^i + T_{k_1}^i| = |T_{k_1}^i - T_{k_2}^i|$. Thus, for $k_1, k_2 > M$, we have

$$\left|T_{k_1}^i - T_{k_2}^i\right| < \varepsilon.$$

Suppose now that the Cauchy sequence $(T_k^i)_{k\geq 1}$ converges to $T_0^i \in \mathbb{R}$. It suffices to show that

(i) $T_0 = (T_0^1, T_0^2, \ldots, T_0^{n-1})$ represents a point in $\mathcal{P}_N$;

(ii) $\lim_{k\to\infty} d_{\text{tr}}(T_k, T_0) = 0$.

To show (ii), we argue that since $(T_k^i)_{k\geq 1}$ converges to T_0^i for all $1 \leq i \leq n-1$, then for any $\varepsilon > 0$ there exists M such that for $k > M$, we have $|T_k^i - T_0^i| < \frac{\varepsilon}{2}$ for all $1 \leq i \leq n-1$. Then by Definition 16,

$$\begin{aligned} d_{\text{tr}}(T_k, T_0) &= \max_{1\leq i\leq n-1} \left(0, T_k^i - T_0^i\right) - \min_{1\leq i\leq n-1} \left(0, T_k^i - T_0^i\right) \\ &< \frac{\varepsilon}{2} - \left(-\frac{\varepsilon}{2}\right) = \varepsilon. \end{aligned}$$

So $\lim_{k\to\infty} d_{\text{tr}}(T_k - T_0) = 0$.

To show (i), note that each coordinate of T_0, including the first, is the limit of the corresponding coordinates of $(T_k)_{k\geq 1}$. Suppose $T_0 \notin \mathcal{P}_N$, then there exists $1 \leq i < j < k < l \leq N$ such that one term of T_0 in (2) is strictly greater than the remaining two. Then there exists M_2 such that for all $k > M_2$, the one term of T_k in (2) is also strictly greater than the remaining two, thus $T_k \notin \mathcal{P}_N$—a contradiction. Hence (i) holds, and $\mathcal{P}_N$ is complete. □

Proposition 37. *$\mathcal{P}_N$ is separable.*

Proof. We claim that the set of all trees with all rational coordinates is dense in $\mathcal{P}_N$: Fix any tree $T = (w_{ij}) \in \mathcal{P}_N$. By Proposition 29, T belongs to a polyhedron and there exists a tree topology with $(N-3)$ internal edges. Then the distance between any two leaves is the sum of the lengths of the edges along the unique path connecting them. The number of edges along each path is at most $(N-1)$. For any $\varepsilon > 0$ and length b_k of each edge of the tree T, since $\mathbb{Q}$ is dense in $\mathbb{R}$, we can find a rational number q_k such that $|q_k - b_k| < \frac{1}{2(N-1)}\varepsilon$. Now, construct another tree $T' = (w'_{ij})$ with the same topology as T, and with corresponding edge lengths q_k. Then for any $1 \leq i < j \leq N$ we have that $|w'_{ij} - w_{ij}| < \frac{\varepsilon}{2}$. Thus

$$d_{\text{tr}}(T', T) = \max_{1\leq i<j\leq n} \left(|w'_{ij} - w_{ij}|\right) - \min_{1\leq i<j\leq n} \left(|w'_{ij} - w_{ij}|\right) < \varepsilon,$$

and all coordinates of q_k are rational. Thus, $\mathcal{P}_N$ is separable. □

In addition, we have the following characterization of compact subsets in $\mathbb{R}^n$.

Theorem 38 (Heine–Borel Theorem (Conway, 2014, Theorem 1.4.8)). *In the Euclidean space $\mathbb{R}^n$, a subset is compact if and only if it is closed and bounded.*

Thus, for palm tree space, the following corollary holds, and there exist compact subsets in palm tree space.

Corollary 39. *In $\mathcal{P}_N$, a subset is compact if and only if it is closed and bounded.*

4 Tropical Line Segments and Tree Ultrametrics

In this section, we treat in detail the case of equidistant trees (recall Definition 11), which, by Proposition 12, are tree ultrametrics. The relevance for a study dedicated to ultrametrics specifically in this paper is twofold. First, it allows for a parallel, geometric comparison between BHV space and the tropical interpretation of phylogenetic tree space: As mentioned above in Section 2.3.3 and in Remark 23, the geometric significance of BHV space lies in its construction when only internal edges are considered. Its structure is based on the union of orthants, where each orthant corresponds to a distinct, rooted tree topology (i.e. combinatorial type, together with leaf labeling scheme). In other words, the definition of BHV space inherently relies on tree topologies. The tropical construction of tree space, while also polyhedral (see Proposition 29), has a more complex algebraic structure. This section provides details on rooted equidistant tree topologies and their occurrence within the tropical construction of tree space. We remark that in this section, the tropical metric that defines palm tree space is not specifically used in the study (indeed, a metric is not explicitly required to deduce the results in this section), however, our discourse remains fundamentally tropical. In particular, we consider tropical line segments as a framework for our study of the behavior of ultrametric tree topologies.

Second, in the context of applications, ultrametrics correspond to *coalescent processes*, which model important biological phenomena, such as cancer evolution (Kingman, 2000). In phylogenomics, the *coalescent model* is often used to model gene trees given a species tree (see e.g. Knowles (2009); Rosenberg (2003); Tian and Kubatko (2014) for further details). The coalescent model takes two parameters, the *population size* and *species depth*, i.e. the number of generators from the most recent common ancester (MRCA) of all individuals at present. The species depth coincides with the height of each gene tree from the root (that is, the MRCA) to each leaf representing an individual in the present time. The output of the coalescent model is a set of equidistant gene trees, since the number of generations from their MRCA to each individual in the present time are the same by model construction. Understanding the structure of ultrametrics is an important step towards the modeling and analysis of coalescent biological processes.

Notation. For a positive integer N, let p_N be the set of all pairs in $[N]$. In this section, for convenience, we denote a tree ultrametric with N leaves by $(w_p)_{p\in p_N}$, where for $p = \{i, j\}$, $w_p = w(\min(i, j), \max(i, j))$.

4.1 Geometry of the Space of Tree Ultrametrics

We begin by providing definitions of tropical geometric line and set objects and outlining important properties of these objects in the space of tree ultrametrics. In what follows, we denote the space of tree ultrametrics for equidistant trees with N leaves by $\mathcal{U}_N$.

Definition 40. Consider the subspace of $L_N \subseteq \mathbb{R}^n$ defined by the linear equations

$$w_{ij} - w_{ik} + w_{jk} = 0 \tag{5}$$

for $1 \leq i < j < k \leq N$ in tree metrics w. For the linear equations (5) spanning L_N, their (max-plus) tropicalization is $w_{ij} \boxplus w_{ik} \boxplus w_{jk}$: recall that under the trivial valuation, all coefficients are disregarded when tropicalizing. This tropicalization of L_N is denoted by $\mathrm{Trop}(L_N) \subseteq \mathbb{R}^n/\mathbb{R}\mathbf{1}$ and is referred to as the *tropical linear space* with points $(w_p)_{p\in p_N}$ where $\max\left(w_{ij}, w_{ik}, w_{jk}\right)$ is obtained at least twice for all triples $i, j, k \in [N]$.

This is equivalent to the three-point condition for ultrametrics given in Definition 10. An observation from tropical geometry gives a correspondence between the tropical linear space $\text{Trop}(L_N)$ and the graphic matroid of a complete graph with N vertices.

We have the following geometric characterization of the space $\mathcal{U}_N$ and corresponding characterizations of tropical line segments between ultrametrics.

Theorem 41 (Ardila and Klivans (2006))**.** *The image of $\mathcal{U}_N$ in the tropical projective torus $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ coincides with $\text{Trop}(L_N)$. That is, $\text{Trop}(L_N) = \mathcal{U}_N$.*

Definition 42. Let $S \subset \mathbb{R}^n$. If tropical line segments (recall from Definition 26) $a \odot x \boxplus b \odot y \in S$ for all $x, y \in S$ and all $a, b \in \mathbb{R}$, then S is said to be *tropically convex*.

The *tropical convex hull* or *tropical polytope* of a given subset $V \subset \mathbb{R}^n$ is the smallest tropically-convex subset containing $V \subset \mathbb{R}^n$; it is denoted by $\text{tconv}(V)$. The tropical convex hull of V may be also written as the set of all tropical linear combinations:

$$\text{tconv}(V) = \{a_1 \odot v_1 \boxplus a_2 \odot v_2 \boxplus \cdots \boxplus a_r \odot v_r \mid v_1, \ldots, v_r \in V \text{ and } a_1, \ldots, a_r \in \mathbb{R}\}.$$

Proposition 43. *For two tree ultrametrics $T_1, T_2 \in \mathcal{U}_N$, the tropical line segment generated by T_1 and T_2,*

$$a \odot T_1 \boxplus b \odot T_2 \ \forall\ a, b \in \mathbb{R},$$

is contained in $\mathcal{U}_N$. In other words, $\mathcal{U}_N$ is tropically convex.

Proof. Since T_1 is a tree ultrametric, $a \odot T_1$ remains a tree ultrametric; this is also true for $b \odot T_2$. Thus, we may assume $a = b = 0$.

Suppose $T_1 = (w_{\{i,j\}})_{1 \le i < j \le N}$ and $T_2 = (w'_{\{i,j\}})_{1 \le i < j \le N}$, then $T_1 \boxplus T_2 = (w_{\{i,j\}} \boxplus w'_{\{i,j\}})_{1 \le i < j \le N}$. Let $z_{\{i,j\}} = w_{\{i,j\}} \boxplus w'_{\{i,j\}}$. It suffices to show that for any $1 \le i < j < k \le n$, we have that the maximum among $z_{\{i,j\}}$, $z_{\{i,k\}}$, $z_{\{j,k\}}$ is attained at least twice in order for $T_1 \boxplus T_2$ to be a tree ultrametric. Let M be this maximum, and set $M := z_{\{i,j\}}$. Then either $w_{\{i,j\}} = M$ or $w'_{\{i,j\}} = M$. If $w_{\{i,j\}} = M$, at least one of $w_{\{i,k\}}$, $w_{\{j,k\}}$ is equal to M since T_1 is a tree ultrametric, therefore at least one of $z_{\{i,k\}}$, $z_{\{j,k\}}$ is also equal to M. Similarly, if $w'_{\{i,j\}} = M$, at least one of $w'_{\{i,k\}}$, $w'_{\{j,k\}}$ is equal to M, since T_2 is a tree ultrametric, therefore at least one of $z_{\{i,k\}}$, $z_{\{j,k\}}$ is also equal to M. Thus, in either case the maximum among $z_{\{i,j\}}$, $z_{\{i,k\}}$, $z_{\{j,k\}}$ is attained at least twice, hence $T_1 \boxplus T_2$ is a tree ultrametric. □

This result generalizes outside the context of trees; in general, tropical linear spaces in the tropical projective torus are tropically convex (see Proposition 5.2.8 of Maclagan and Sturmfels (2015)).

An algorithm for computing tropical line segments between tree ultrametrics is given in Appendix A1, Algorithm 1.

4.2 Tree Topologies Along Tropical Line Segments

Thus far, we have shown that the tropical linear space coincides with the space of tree ultrametrics, and hence, that tropically-convex sets (and therefore tropical line segments) are also fully contained in the space of tree ultrametrics. In other words, this endows the space of tree ultrametrics with a tropical structure, and now allows us to study the behavior of points (trees) along tropical line segments. In particular, this allows us to characterize ultrametric tree topologies geometrically, thereby providing a description of the tropically-constructed tree space that is comparable to the geometry of BHV space for rooted trees and zero-length external edges.

The strategy that we implement in this section is largely combinatorial. We first formalize the definition of a tree topology first introduced in Section 2.2 as a collection of subsets of leaves, and use these subsets to define notions of size, and in particular, largest and smallest subsets. Given these upper and lower bounds, we then define an equivalence relation and a partial order that allow us to iteratively and combinatorially partition and compare leaf subsets and tree topologies. This gives us a framework to study types of trees that may or may not exist within the setting of tropical line segments: Specifically, Theorem 57 is a combinatorial search procedure to give the possible tree topologies that exist along a tropical line segment.

Definition 44. A *tree topology* F on $[N]$ is a collection of *clades* $S \subseteq [N]$, where $2 \leq |S| \leq N-1$ and for any two distinct clades $S_1, S_2 \in F$, exactly one of the following *nested set conditions* holds:

$$\begin{aligned} S_1 &\subsetneq S_2, \\ S_2 &\subsetneq S_1, \\ S_1 \cap S_2 &= \emptyset. \end{aligned} \tag{6}$$

F is said to be *full dimensional* if $|F| = N-2$.

Clades always belong to $[N]$ rather than $[N] \cup \{0\}$, since we may always choose the clade excluding the root, i.e. the leaf with label 0. In this manner, they allow for an alternative representation of trees over tree metric vectors w or matrices W.

Example 45. For the tree in Figure 6, there are two ways to express this tree:

1. As a tree ultrametric in $\mathcal{U}_5$: $(16, 40, 40, 40, 40, 40, 40, 20, 20, 10)$
2. As a vector in an ambient space, in terms of lengths of internal edges:
   $(0, \ldots, 0, 12, 10, 5, 0, \ldots, 0)$, where the values of the nonzero coordinates in this vector are the internal edge lengths leading to clades $\{A, B\}$, $\{C, D, E\}$ and $\{D, E\}$.

In general, using the internal edges to represent trees allows for an iterative construction of a family of clades satisfying one of the nested set conditions (6). In the case of Figure 6, this family is $\{A, B\}$, $\{C, D, E\}$ and $\{D, E\}$.

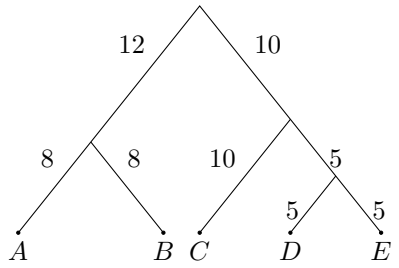

Figure 6: Example of an equidistant tree. This tree is also an ultrametric.

The following lemma provides intuition on the definition of full-dimensionality given above in Definition 44.

Lemma 46. *Let $N \geq 2$ be an integer. For any tree topology F on a ground set of N elements, we have $|F| \leq N-2$.*

Proof. We proceed by induction on N. When $N = 2$, since $2 > N-1$, F is necessarily empty because clades S cannot satisfy $2 \leq |S| \leq 1$. When $N = 3$, all clades of F must have cardinality 2, because the cardinality is at least 2 and at most 2, and thus are among $\{1, 2\}, \{1, 3\}, \{2, 3\}$. But any two of these clades do not satisfy one of the nested set conditions (6). Hence $|F| \leq 1$, and the base case for $N = 2, 3$ holds.

Next, suppose Lemma 46 holds for $3 \leq N \leq m$ where $m \geq 3$. Consider the case when $N = m+1$: We take all clades S belonging to F that are maximal in terms of inclusion. There are two cases:

(i) There is a unique maximal clade $S_{\max}$ in F: $|S_{\max}| \leq m$ and all clades of F are subsets of $S_{\max}$. If $|S_{\max}| \leq 2$, then F has a unique clade, which is $S_{\max}$ and therefore $|F| = 1 \leq m-1 = N-2$; otherwise $|S_{\max}| \geq 3$. Consider the family $F \backslash \{S_{\max}\}$ on the ground set $S_{\max}$. For any $S \in F \backslash \{S_{\max}\}$, since S is a proper subset of $S_{\max}$, we have $2 \leq |S| \leq |S_{\max}| - 1$. Since $F \backslash \{S_{\max}\}$ is still a nested set, it is a tree topology on $S_{\max}$. By the induction hypothesis, $|F \backslash \{S_{\max}\}| \leq |S_{\max}| - 2 \leq m-2$. Thus, $|F| = |F \backslash \{S_{\max}\}| + 1 \leq m-1 = N-2$ and Lemma 46 holds for $N = m+1$.

(ii) There are at least two maximal clades in F: Let these clades be $S_1, S_2, \ldots, S_k$, with $k \geq 2$, then $2 \leq |S_i| \leq m$ for $1 \leq i \leq k$. Since F is a nested set, these S_i are pairwise disjoint. So

$$\sum_{i=1}^{m} |S_i| \leq m + 1.$$

Let c_i be the number of proper subsets of S_i that belong to F. Then $|F| = k + \sum_{i=1}^{k} c_i$. Notice that c_i is also the cardinality of a tree topology on S_i, so by the induction hypothesis, $c_i \leq |S_i| - 2$. Hence

$$|F| = k + \sum_{i=1}^{k} c_i \leq \sum_{i=1}^{m} |S_i| - k \leq (m+1) - 2 = N - 2. \tag{7}$$

Lemma 46 thus holds for $N = m + 1$.

This concludes the transition step, and the proof. □

Conversely, now, we consider the minimal clades of a tree topology F.

Lemma 47. *Let F be a tree topology on $[N]$. For any pair $p \in p_N$, let $F(p) = \{S \in F \mid S \supseteq p\}$. For $p \in p_N$, if $F(p) \neq \emptyset$, then the intersection of all clades $S \in F(p)$ is also an element of $F(p)$.*

Proof. Suppose $p \in p_N$ and $F(p) \neq \emptyset$. For any two clades $S_1, S_2 \in F(p)$, since $p \subseteq S_1, S_2$, then S_1 and S_2 cannot be disjoint. Thus, by the nested set condition in Definition 44, either S_1 contains S_2 or vice versa. This means that all clades in $F(p)$ form a completely ordered set with respect to set inclusion. Since $F(p)$ is finite, it has a minimal element that must be contained in all other elements. This minimal element is the intersection of all clades in $F(p)$. □

Definition 48. The minimal element that is the intersection of all clades $S \in F(p)$ given in Lemma 47 is called the *closure* of p in F. We denote this closure by $\mathrm{cl}_F(p)$. If $F(p) = \emptyset$, we set $\mathrm{cl}_F(p)$ to be $[N]$.

Lemma 47 intuitively gives us the result that $\mathrm{cl}_F(p)$ is the minimal subset in $F \cup [N]$ that contains p.

Given the definition of a tree topology F, and notions of maximal and minimal clades of F, we now proceed to study the behavior of varying tree topologies. We shall construct the setting for such a study via the definitions of an equivalence relation and a partial order on a tree topology F in terms of pairs $p \in p_N$ as follows.

Definition 49. Let F be a tree topology on $[N]$. We define an equivalence relation $=_F$ on p_N by

$$p_1 =_F p_2 \text{ if } \mathrm{cl}_F(p_1) = \mathrm{cl}_F(p_2),$$

and a partial order $<_F$ on p_N by

$$p_1 <_F p_2 \text{ if } \mathrm{cl}_F(p_1) \subsetneq \mathrm{cl}_F(p_2),$$

for distinct pairs $p_1, p_2 \in p_N$.

By the definition of the equivalence relation $=_F$, every equivalence class with respect to $=_F$ corresponds to an element of F or $[N]$. Non-distinct pairs p_N are always comparable under $=_F$ or $<_F$.

Lemma 50. *Let F be a tree topology on $[N]$. For any distinct elements $i, j, k \in [N]$, exactly one of the following holds:*

$$\{i,j\} =_F \{i,k\},$$
$$\{i,j\} <_F \{i,k\},$$
$$\{i,k\} <_F \{i,j\}.$$

Proof. Suppose for contradiction that Lemma 50 does not hold for some distinct i, j, k. Then $F(\{i,j\})$ and $F(\{i,k\})$ do not contain each other and there exist $S_1, S_2 \in F$ such that $S_1 \in F(\{i,j\}) \backslash F(\{i,k\})$ and $S_2 \in F(\{i,k\}) \backslash F(\{i,j\})$. Thus $j \in S_1 \backslash S_2$ and $k \in S_2 \backslash S_1$, contradicting that F is a nested set. Hence Lemma 50 holds. □

Given this framework to compare pairs of elements, we now give two notions of bipartitioning of trees that are closely related. It turns out, as we will see in Theorem 53, in the context of tree topologies, they coincide.

Definition 51. A rooted phylogenetic tree T is said to be *binary* if every vertex of T is either a leaf or trivalent.

Definition 52. Let F be a tree topology on $[N]$ and $\bar{F} = \{S \in F \mid |S| \geq 3\} \cup [N]$. F is said to be *bifurcated* if for every $S \in \bar{F}$, exactly one of the following holds:

(a) there exists a proper subset $S' \subset S$ such that $S' \in F$ and $|S'| = |S| - 1$; or

(b) there exist two proper subsets $S', S'' \subset S$ such that $S', S'' \in F$ and $S' \cup S'' = S$.

Note that in (b), we must have that $S' \cap S'' = \emptyset$.

Theorem 53. *Let F be a tree topology on $[N]$. The following are equivalent:*

(1) F is full dimensional;

(2) for $1 \leq i < j < k \leq N$, two of the pairs $\{i,j\}, \{i,k\}, \{j,k\}$ are $=_F$, and the third pair is $<_F$ than the two F-equivalent pairs;

(3) F is bifurcated;

(4) every phylogenetic tree with tree topology F is binary.

We defer the proof of Theorem 53 to Appendix A2.

These concepts of bipartitioning in terms of tree topologies are important in understanding the combinatorial aspects of ultrametrics representing equidistant trees. Recall from Section 4.1, however, that in addition to the fact that tree ultrametrics are equidistant trees, but also that every point along any tropical line segment between two equidistant trees is also itself an equidistant tree. The equivalence relation $=_F$ and partial order $<_F$ completely define the set of all ultrametrics for a given tree topology F, which we present and formalize in the following results.

Proposition 54. *Given a tree topology F on $[N]$, let $ut(F)$ be the set of all ultrametrics in $\mathcal{U}_N$ corresponding to a tree with tree topology F. Then*

$$ut(F) = \{(w_p)_{p \in p_N} \in \mathcal{U}_N \mid w_{p_1} = w_{p_2} \text{ if } p_1 =_F p_2 \text{ and } w_{p_1} < w_{p_2} \text{ if } p_1 <_F p_2\}. \tag{8}$$

Proof. Fix a tree topology F on $[N]$ and a corresponding equidistant tree T. For each internal edge of T indexed by a clade S, let $\ell(S) > 0$ be its length. For the external edge of T, connecting the ith leaf to the root of T, let ℓ_i be its length. Let h be the height of T. Then, by definition, the distance of each leaf to the root of T is h. There exists a unique path from the ith leaf to the root, consisting of the external edge connecting the ith leaf, and some internal edges: an internal edge indexed by S appears on this path if and only if the ith leaf and the root are separated by the internal edge itself. This necessarily means that $i \in S$. Then

$$h = \ell_i + \sum_{i \in S} \ell(S). \tag{9}$$

Next we consider the path connecting the ith and jth leaves. This path consists of the two external edges and some internal edges. An internal edge indexed by S appears on this path if and only if the ith leaf and the jth leaf are separated by the edge. Equivalently, this means that $|\{i,j\} \cap S| = 1$. Hence

$$\begin{aligned} w_{\{i,j\}} &= \ell_i + \ell_j \sum_{|\{i,j\} \cap S| = 1} \ell(S) \\ &= 2h - \sum_{i \in S} \ell(S) - \sum_{j \in S} \ell(S) + \sum_{|\{i,j\} \cap S| = 1} \ell(S) \\ &= 2h - 2 \sum_{i,j \in S} \ell(S) \\ &= 2h - 2 \sum_{S \in F(\{i,j\})} \ell(S). \end{aligned}$$

Therefore w satisfies the condition defining the set (8).

Conversely, suppose a vector $w \in \mathbb{R}^n/\mathbb{R}\mathbf{1}$ satisfies the conditions defining the set (8). Then the system of linear equations

$$w_{\{i,j\}} = 2h - 2 \sum_{S \in F(\{i,j\})} x_S \quad \forall\, \{i,j\} \in p_N$$

has a solution such that $h \in \mathbb{R}$ and $x_S > 0$ for all $S \in F$. By a translation in $\mathbb{R}$, we may choose a sufficiently large h such that all ℓ_i in (9) are positive. Then w is the ultrametric of an equidistant tree with external edge lengths ℓ_i and internal edge lengths x_S, whose tree topology is F. □

Corollary 55. *Let F be a tree topology on $[N]$ and $w \in ut(F)$. Then for any pairs $p, q \in p_N$, if $p \cap q \neq \emptyset$, then $w_p = w_q$ implies $p =_F q$ and $w_p < w_q$ implies $p <_F q$.*

Proof. If $p = q$, then $w_p = w_q$ holds and $p =_F q$ also holds. Otherwise we may assume $p = \{i,j\}$ and $q = \{i,k\}$. By Lemma 50, one of the following relationships holds: $p <_F q$, or $p =_F q$, or $q <_F p$. By Proposition 54, these three relationships imply $w_p < w_q$, or $w_p = w_q$, or $w_p > w_q$ respectively. Hence the converse implications also hold. □

Thus far, we have introduced means to study tree structures by studying subsets of leaves, and iteratively dividing and comparing these subsets. We have also determined the set of tree ultrametrics defined by the comparison framework set up in Definition 49. Given this framework, we now determine when we have compatibility of sets in order to present the main combinatorial result of this section, which gives all possible tree topologies along a tropical line segment.

It should be noted that the geometric and combinatorial procedure we present and the framework we construct is a natural approach that has been previously implemented in other tree settings and, more generally, in finite metric spaces, e.g. Bandelt and Dress (1992); Dress (1984). Our approach differs in that we proceed in the setting of tropical geometry, within the framework of tropical line segments.

Definition 56. Let F_1 and F_2 be tree topologies on $[N]$. Define the set of *compatible* tree topologies of F_1, F_2 to consist of tree topologies F where there exist tree ultrametrics $w_1 \in ut(F_1)$ and $w_2 \in ut(F_2)$, such that the tree ultrametric $w_1 \boxplus w_2 \in ut(F)$. We denote this set by $C(F_1, F_2)$.

Theorem 57. *Let F_1, F_2, F be full-dimensional tree topologies on $[N]$. If $F \in C(F_1, F_2)$, then any equivalence class $C \subseteq p_N$ with respect to $=_F$ is contained in an equivalence class with respect to either $=_{F_1}$ or $=_{F_2}$.*

Put differently, for each $S \in \bar{F}$, there exists either $S_1 \in F_1$ such that for $p \in p_N$, if $\mathrm{cl}_F(p) = S$, then $\mathrm{cl}_{F_1}(p) = S_1$; or $S_2 \in F_2$ such that for $p \in p_N$, if $\mathrm{cl}_F(p) = S$, then $\mathrm{cl}_{F_2}(p) = S_2$.

Proof. Suppose $F_3 \in C(F_1, F_2)$. There exist ultrametrics w^1, w^2, w^3 such that for $p \in p_N$, $w^3_p = \max(w^1_p, w^2_p)$ and $w^i \in ut(F_i)$ for $i = 1, 2, 3$. Choose $S \in \bar{F} = \big\{S \in F \mid |S| \geq 3\big\} \cup [N]$. By Theorem 53, F is bifurcated, so condition (a) or (b) of Definition 52 holds for S. If (a) holds, we set $X = S'$ and $Y = S \backslash S'$; if (b) holds, we set $X = S_1$ and $Y = S_2$. Then for $p = \{i,j\} \in p_N$, $\mathrm{cl}_F(p) = S$ if and only if $(i,j) \in X \times Y$ or $(j,i) \in X \times Y$. Let $M = w^3_{\{i,j\}}$ for all $i \in X$ and $j \in Y$, then

$$\max\left(w^1_{\{i,j\}}, w^2_{\{i,j\}}\right) = M \tag{10}$$

for all $i \in X$ and $j \in Y$.

Consider a complete, bipartite graph $G := K_{|X|,|Y|}$ with vertices $X \cup Y$. Recall that the vertices of a bipartite graph can be partitioned into two disjoint, independent sets where every graph edge connects a vertex in one set to one in the other. Thus, for $(i,j) \in X \times Y$, if $w^1_{\{i,j\}} = M$, then we call the edge (i,j) of G pink; if $w^2_{\{i,j\}} = M$, then we call the edge (i,j) of G purple. Then, each edge in G is pink, purple, or both pink and purple. We claim that in fact, either all edges of G are pink, or all edges of G are purple.

Suppose there exists a non-pink edge $(i,j) \in X \times Y$ in G. Then this edge is purple: $M = w^2_{\{i,j\}} \neq w^1_{\{i,j\}}$. By (10), $w^1_{\{i,j\}} < M$. For any other element $j' \in Y, j' \neq j$, if (i,j') is not purple, then $w^2_{\{i,j'\}} < M = w^1_{\{i,j'\}}$. Note that $w^1_{\{i,j\}} < w^1_{\{i,j'\}}$, thus by Corollary 55, $\{i,j\} <_{F_1} \{i,j'\}$. Similarly, since $w^2_{\{i,j'\}} < w^2_{\{i,j\}}$, we have $\{i,j'\} <_{F_2} \{i,j\}$. By Theorem 53, $\{j,j'\} =_{F_1} \{i,j'\}$ and $\{j,j'\} =_{F_2} \{i,j\}$. Then $w^1_{\{j,j'\}} = w^2_{\{j,j'\}} = M$ and

$w_{\{j,j'\}} = M$. By Corollary 55, this means that $\{i,j\} =_F \{i,j'\} =_F \{j,j'\}$, which contradicts that F is full dimensional. Therefore (i,j') must be purple. Symmetrically, for any $i' \in X$ with $i' \neq i$, (i',j) must be purple. Then for such i' and j', we have

$$w^2_{\{i,j\}} = w^2_{\{i,j'\}} = w^2_{\{i',j\}} = M.$$

By Corollary 55, we have $\{i,j\} =_{F_2} \{i,j'\} =_{F_2} \{i',j\}$. Since F_2 is full dimensional, Theorem 53 gives us that $\{j,j'\} <_{F_2} \{i,j\} =_{F_2} \{i,j'\}$. Then $\{j,j'\} <_{F_2} \{i',j\}$, and thus $\{i',j'\} =_{F_2} \{i',j\}$. Hence $w^2_{\{i',j'\}} = M$ and the edge (i',j') in G is purple too. So all edges in G are purple, thus proving our claim.

Finally, if all edges in G are pink, then all $w^1_{\{i,j\}}$ are equal for $(i,j) \in X \times Y$. By Corollary 55, all these $\{i,j\}$ belong to the same equivalence class with respect to $=_{F_1}$; symmetrically, if all edges in G are purple, then all $w^2_{\{i,j\}}$ are equal for $(i,j) \in X \times Y$ and all these $\{i,j\}$ belong to the same equivalence class with respect to $=_{F_2}$. □

There are two important remarks concerning this result to highlight:

Remark 58. Theorem 57 may still hold when one of the tree topologies is not full dimensional, say F_1. For example, let $N = 5$ and

$$F_1 = \{3,4\}, \quad F_2 = \big\{\{1,4\},\{2,3\},\{1,2,3,4\}\big\}.$$

For $F = \big\{\{1,4\},\{1,3,4\},\{2,5\}\big\}$, $F \notin C(F_1,F_2)$: Suppose $w^1 \in ut(F_1)$ and $w^2 \in ut(F_2)$. Since $\mathrm{cl}_{F_i}(\{2,5\}) = [5]$ for $i = 1,2$, by definition $(w^i)_{\{2,5\}} = \max_{p \in p_5} (w^i)_p$ for $i = 1,2$. Let $w = w^1 \boxplus w^2$. Then $w_{\{2,5\}} = \max_{p \in p_5} w_p$ also. However, since $\{2,5\} \in F$, $\mathrm{cl}_F\big(\{2,5\}\big) \neq [5]$ and thus $w \notin ut(F)$. Nevertheless, Theorem 57 holds for F, since for $p \in p_5$, we have:

$$\begin{array}{lcclcl}
\mathrm{cl}_F(p) & = & [5] & \text{implies} & \mathrm{cl}_{F_1}(p) & = & [5]; \\
\mathrm{cl}_F(p) & = & \{1,3,4\} & \text{implies} & \mathrm{cl}_{F_2}(p) & = & \{1,2,3,4\}; \\
\mathrm{cl}_F(p) & = & \{1,4\} & \text{implies} & \mathrm{cl}_{F_2}(p) & = & \{1,4\}; \\
\mathrm{cl}_F(p) & = & \{2,5\} & \text{implies} & \mathrm{cl}_{F_1}(p) & = & [5].
\end{array}$$

Remark 59. Theorem 57 provides a combinatorial way of searching for all possible tree topologies a tropical line segment of tree ultrametrics can traverse; Example 60 illustrates this procedure. However, the converse is not true: given full-dimensional tree topologies satisfying the conditions of Theorem 57, it is not always possible to construct tropical line segments that traverse these topologies; see Example 61 for details.

Example 60. Let $N = 5$ and choose the following two full-dimensional tree topologies on $[5]$:

$$F_1 = \big\{\{1,2,3\},\{1,2\},\{4,5\}\big\} \quad \text{and} \quad F_2 = \big\{\{1,3,4,5\},\{1,3,5\},\{1,5\}\big\}.$$

Then the full-dimensional tree topologies in $C(F_1,F_2)$ are F_1, F_2 themselves, and three others:

$$\big\{\{1,3,4,5\},\{1,3,5\},\{1,3\}\big\}, \quad \big\{\{1,3,4,5\},\{4,5\},\{1,3\}\big\}, \quad \big\{\{1,2,3\},\{1,3\},\{4,5\}\big\}.$$

Note that $F_3 = \big\{\{1,2,3\},\{2,3\},\{4,5\}\big\}$ does not belong to $C(F_1,F_2)$, because $\{1,2\}$ and $\{1,3\}$ form an equivalence class with respect to $=_F$, but $\{1,3\} <_{F_i} \{1,2\}$ for $i = 1,2$. This gives an example of the non-existence of certain types of trees.

Example 61. Following Remark 59, let $N = 12$. Consider the following three tree topologies F_1, F_2, F, depicted in Figure 9:

$$\begin{aligned}
F_1 = \big\{&\{1,2,7,8,9,12\},\ \{1,7,9\},\ \{2,8,12\},\ \{1,7\},\ \{2,8\},\\
&\{3,4,5,6,10,11\},\ \{3,5,11\},\ \{4,6,10\},\ \{3,5\},\ \{4,6\}\big\}\\
F_2 = \big\{&\{1,2,3,4,9,10\},\ \{2,3,4,9,10\},\ \{2,3,4,10\},\ \{3,4,10\},\ \{3,10\},\\
&\{5,6,7,8,11,12\},\ \{5,7,12\},\ \{6,8,11\},\ \{5,7\},\ \{6,8\}\big\}\\
F = \big\{&\{1,2,3,4,9,10\},\ \{1,2,9\},\ \{3,4,10\},\ \{1,9\},\ \{3,10\},\\
&\{5,6,7,8,11,12\},\ \{5,6,11\},\ \{7,8,12\},\ \{5,11\},\ \{7,12\}\big\}.
\end{aligned}$$

Note that $|F_1| = |F_2| = |F| = 10$, so F_1, F_2, F are all full dimensional. Moreover, Theorem 57 holds for F, since for $p \in p_{12}$, we have

$$\begin{array}{lcclcl}
\mathrm{cl}_F(p) & = & [12] & \text{implies} & \mathrm{cl}_{F_2}(p) & = & [12]; \\
\mathrm{cl}_F(p) & = & \{1,2,3,4,9,10\} & \text{implies} & \mathrm{cl}_{F_1}(p) & = & \{1,2,3,4,9,10\}; \\
\mathrm{cl}_F(p) & = & \{5,6,7,8,11,12\} & \text{implies} & \mathrm{cl}_{F_1}(p) & = & \{5,6,7,8,11,12\}; \\
\mathrm{cl}_F(p) & = & \{1,2,9\} & \text{implies} & \mathrm{cl}_{F_1}(p) & = & \{1,2,7,8,9,12\}; \\
\mathrm{cl}_F(p) & = & \{3,4,10\} & \text{implies} & \mathrm{cl}_{F_2}(p) & = & \{3,4,10\}; \\
\mathrm{cl}_F(p) & = & \{5,6,11\} & \text{implies} & \mathrm{cl}_{F_1}(p) & = & \{3,4,5,6,10,11\}; \\
\mathrm{cl}_F(p) & = & \{7,8,12\} & \text{implies} & \mathrm{cl}_{F_2}(p) & = & \{5,6,7,8,11,12\}; \\
\mathrm{cl}_F(p) & = & \{1,9\} & \text{implies} & \mathrm{cl}_{F_1}(p) & = & \{1,7,9\}; \\
\mathrm{cl}_F(p) & = & \{3,10\} & \text{implies} & \mathrm{cl}_{F_2}(p) & = & \{3,10\}; \\
\mathrm{cl}_F(p) & = & \{5,11\} & \text{implies} & \mathrm{cl}_{F_1}(p) & = & \{3,5,11\}; \\
\mathrm{cl}_F(p) & = & \{7,12\} & \text{implies} & \mathrm{cl}_{F_2}(p) & = & \{5,7,12\}.
\end{array}$$

However, $F \notin C(F_1, F_2)$: suppose there exist $w \in ut(F), w^1 \in ut(F_1), w^2 \in ut(F_2)$ such that $w = w^1 \boxplus w^2$. Note that $\{1,2\} =_F \{2,9\}$ and $\{1,2\} =_{F_1} \{2,9\}$ but $\{2,9\} <_{F_2} \{1,2\}$. Then

$$w^2_{\{2,9\}} < w^2_{\{1,2\}} \leq \max\left(w^1_{\{1,2\}}, w^2_{\{1,2\}}\right) = w_{\{1,2\}} = w_{\{2,9\}} = \max\left(w^1_{\{2,9\}}, w^2_{\{2,9\}}\right).$$

Hence $w^1_{\{2,9\}} > w^2_{\{2,9\}}$, and we have

$$w^1_{\{1,2\}} = w^1_{\{2,9\}} = w_{\{2,9\}} = w_{\{1,2\}} = \max\left(w^1_{\{1,2\}}, w^2_{\{1,2\}}\right).$$

Thus,

$$w^1_{\{1,2\}} \geq w^2_{\{1,2\}}. \tag{11}$$

Similarly, $\{4,10\} <_{F_1} \{3,4\}$ implies that

$$w^2_{\{3,4\}} \geq w^1_{\{3,4\}}. \tag{12}$$

We also have $\{6,11\} <_{F_2} \{5,6\}$, which implies that

$$w^1_{\{5,6\}} \geq w^2_{\{5,6\}}, \tag{13}$$

and $\{8,12\} <_{F_1} \{7,8\}$ implies that

$$w^2_{\{7,8\}} \geq w^1_{\{7,8\}}. \tag{14}$$

Then Proposition 54 together with (11), (12), (13), and (14), we have the following chain of inequalities:

$$w^1_{\{1,2\}} \geq w^2_{\{1,2\}} > w^2_{\{3,4\}} \geq w^1_{\{3,4\}} = w^1_{\{5,6\}} \geq w^2_{\{5,6\}} = w^2_{\{7,8\}} \geq w^1_{\{7,8\}} = w^1_{\{1,2\}}$$

—a contradiction. Hence such tree ultrametrics w, w^1, w^2 do not exist.

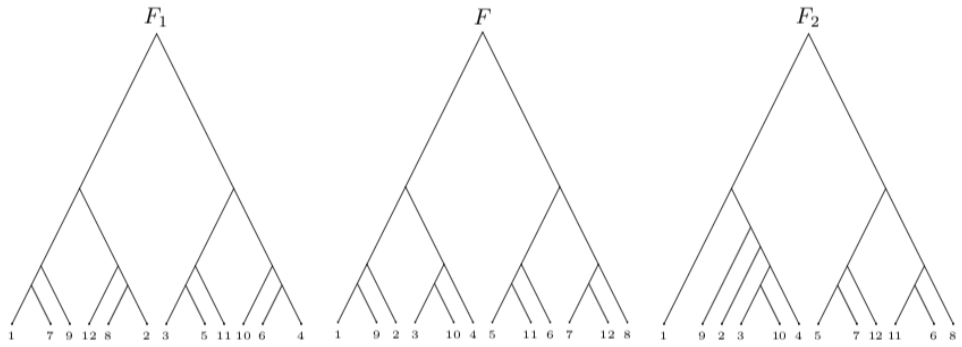

Figure 7: Tree topologies F_1, F_2, F with 12 leaves in Example 61.

4.3 Symmetry on Tropical Line Segments

In contrast to the previous section, which dealt with how tree topologies vary along tropical line segments, we now turn our focus to understanding when and how *invariance* arises in the space of ultrametrics. To do so, we define the notion of *symmetry* on ultrametrics in terms of leaf relabeling. The natural setting for such a study is the action of the symmetric group $\mathrm{Sym}(N)$ on N labels on $[N]$, given by permuting the coordinates (positions) of the labels of the leaves. This will be the focus of the results derived in this section.

In our study, we consider the map $\Sigma : \mathcal{U}_N \to \mathcal{U}_N$ defined by

$$\Sigma(w, \sigma) = \big(w_{\{\sigma_1,\sigma_2\}}, w_{\{\sigma_1,\sigma_3\}}, \ldots, w_{\{\sigma_{N-1},\sigma_N\}}\big),$$

where w is an ultrametric and $\sigma \in \mathrm{Sym}(N)$ is an Nth-order permutation of the symmetric group.

Definition 62. Let T be an equidistant tree with N leaves and let $w_T \in \mathcal{U}_N$ be a tree metric associated with T. Define the equivalence relation $=_\sigma$ between equidistant trees T and T' with N leaves by $T =_\sigma T'$ if and only if T and T' have the same tree topology and branch lengths, but the labels of leaves in T' are permuted by $\sigma \in \mathrm{Sym}(N)$.

Example 63. Let T_1 and T_2 be the equidistant trees shown in Figure 8. Here, $T_1 =_\sigma T_2$, where $\sigma = (2, 3, 1, 4)$.

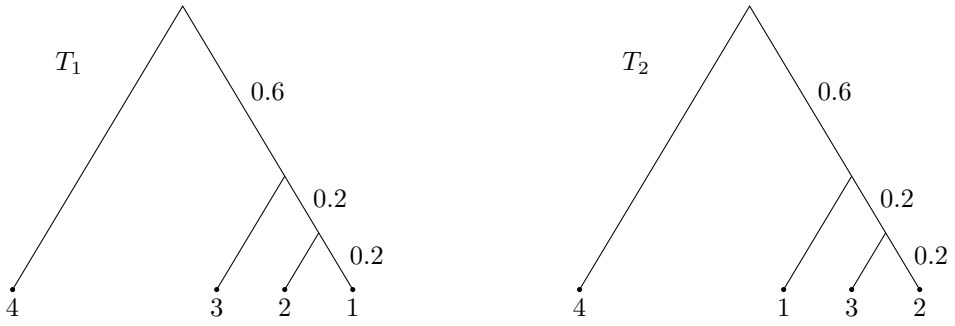

Figure 8: T_1 and T_2 in Example 63.

Lemma 64. *Let T_1, T_2 be equidistant trees with N leaves. Let w^1 be a tree ultrametric for T_1 and w^2 be a tree ultrametric for T_2. Then $T_1 =_\sigma T_2$ if and only if $\Sigma(w^1, \sigma^{-1}) = w^2$ or equivalently, $\Sigma(w^2, \sigma) = w^1$.*

Proof. Suppose $T_1 =_\sigma T_2$. Let $w^1 = (w^1_{\{1,2\}}, \ldots, w^1_{\{N-1,N\}})$ and, similarly, $w^2 = (w^2_{\{1,2\}}, \ldots, w^2_{\{N-1,N\}})$. We have that $w^1_{\{i,j\}}$ is the distance from a leaf i to a leaf j on T_1, and $w^2_{\{i,j\}}$ is the distance from a leaf i to a leaf j on T_2, by the definition of a tree metric. Since $T_1 =_\sigma T_2$, we have that $w^1_{\{i,j\}} = w^2_{\{\sigma_i,\sigma_j\}}$. Thus, $\Sigma(w^1, \sigma^{-1}) = w^2$.

Now suppose $\Sigma(w^1, \sigma^{-1}) = w^2$. Then we have $w^1_{\{i,j\}} = w^2_{\{\sigma_i,\sigma_j\}}$ for any pair of leaves $i, j \in [N]$. Since w^1, $w^2 \in \mathcal{U}_N$, there exist equidistant trees T_1, T_2 with N leaves, respectively, by Proposition 12. Then, $w^1_{\{i,j\}}$ is the distance from a leaf i to a leaf j on T_1, and $w^2_{\{i,j\}}$ is the distance from a leaf i to a leaf j on T_2, by the definition of a tree metric. Since $\Sigma(w^1, \sigma^{-1}) = w^2$, we have that $w^1_{\{i,j\}} = w^2_{\{\sigma_i,\sigma_j\}}$ for all pair of leaves i and j. Therefore, $T_1 =_\sigma T_2$. □

Example 65. To illustrate Lemma 64, we revisit the equidistant trees T_1 and T_2 shown in Figure 8. Their ultrametrics are

$$w^1 = (0.4,\ 0.8,\ 2,\ 0.8,\ 2,\ 2),$$
$$w^2 = (0.8,\ 0.8,\ 2,\ 0.4,\ 2,\ 2).$$

Then we have $\sigma = (2, 3, 1, 4)$ and $\sigma^{-1} = (3, 1, 2, 4)$, thus

$$\Sigma(w^2, \sigma) = w^1,$$
$$\Sigma(w^1, \sigma^{-1}) = w^2.$$

Note that for T_1 :

$$\begin{array}{ccccccccc}
w^1(1,2) & = & w^2(\sigma_1,\sigma_2) & = & w^2(2,3) & = & & & 0.4 \\
w^1(1,3) & = & w^2(\sigma_1,\sigma_3) & = & w^2(2,1) & = & w^2(1,2) & = & 0.8 \\
w^1(1,4) & = & w^2(\sigma_1,\sigma_4) & = & w^2(2,4) & = & & & 2 \\
w^1(2,3) & = & w^2(\sigma_2,\sigma_3) & = & w^2(3,1) & = & w^2(1,3) & = & 0.8 \\
w^1(2,4) & = & w^2(\sigma_2,\sigma_4) & = & w^2(3,4) & = & & & 2 \\
w^1(3,4) & = & w^2(\sigma_3,\sigma_4) & = & w^2(1,4) & = & & & 2
\end{array}$$

and similarly, for T_2:

$$\begin{array}{ccccccccc}
w^2(1,2) & = & w^1(\sigma_1^{-1},\sigma_2^{-1}) & = & w^1(3,1) & = & w^1(1,3) & = & 0.8 \\
w^2(1,3) & = & w^1(\sigma_1^{-1},\sigma_3^{-1}) & = & w^1(3,2) & = & w^1(2,3) & = & 0.8 \\
w^2(1,4) & = & w^1(\sigma_1^{-1},\sigma_4^{-1}) & = & w^1(3,4) & = & & & 2 \\
w^2(2,3) & = & w^1(\sigma_2^{-1},\sigma_3^{-1}) & = & w^1(1,2) & = & & & 0.4 \\
w^2(2,4) & = & w^1(\sigma_2^{-1},\sigma_4^{-1}) & = & w^1(1,4) & = & & & 2 \\
w^2(3,4) & = & w^1(\sigma_3^{-1},\sigma_4^{-1}) & = & w^1(2,4) & = & & & 2
\end{array}$$

With this definition of symmetry given by permutation of leaf relabeling, we now study how symmetry behaves on tropical line segments. Let $\Gamma(w^T)$ be a tropical line segment from the origin $(0, 0, \ldots, 0)$ to the ultrametric $w^T \in \mathcal{U}_N$ associated with an equidistant tree T. Also, let $\Gamma(w^T, w^{T_0})$ be a tropical line segment from the ultrametric $w^{T_0} \in \mathcal{U}_N$ associated with an equidistant tree T_0, to the ultrametric $w^T \in \mathcal{U}_N$ associated with an equidistant tree T. These definitions give the following result.

Proposition 66. *Suppose T_1 is an equidistant tree with N leaves and let T_2 be an equidistant tree such that $T_1 =_\sigma T_2$. Then*

$$\Sigma\big(\Gamma(w^1), \sigma\big) = \Gamma(w^2)$$

and

$$\Sigma\big(\Gamma(w^2), \sigma^{-1}\big) = \Gamma(w^1).$$

Proof. Let $w^1 = (w^1_{\{1,2\}}, w^1_{\{1,3\}}, \ldots, w^1_{\{N-1,N\}})$ be an ultrametric computed from an equidistant tree T_1, and similarly, $w^2 = (w^2_{\{1,2\}}, w^2_{\{1,3\}}, \ldots, w^2_{\{N-1,N\}})$ be an ultrametric computed from an equidistant tree T_2. Order the coordinates of both w^1 and w^2 from the smallest and largest, and let $(w^1_{(1)}, \ldots, w^1_{(n)})$ and $(w^2_{(1)}, \ldots, w^2_{(n)})$ be the ultrametrics after ordering the coordinates of w^1 and w^2, respectively, from the smallest to the largest coordinates. By Lemma 64, $\Sigma(w^1, \sigma^{-1}) = w^2$ since $T_1 =_\sigma T_2$. Thus, $w^1_{(i)} = w^2_{(i)}$ for $i = 1, \ldots, n$. By applying the algorithm in the proof of Proposition 5.2.5 in Maclagan and Sturmfels (2015) on the line segment from the origin $(0, 0, \ldots, 0)$ to $(w^1_{(1)}, \ldots, w^1_{(n)})$, we obtain a tropical line segment from the origin to w^1. Since $w^1_{(i)} = w^2_{(i)}$ for $i = 1, \ldots, n$, we have that $\Sigma(\Gamma(w^1), \sigma) = \Gamma(w^2)$ or equivalently $\Sigma(\Gamma(w^2), \sigma^{-1}) = \Gamma(w^1)$. □

The following results provide a formalization of invariance under the action of permutation of leaf labels in terms of tree topologies. If two tree ultrametrics have the same topology and branch lengths, but differ by a permutation of leaf labels, and, correspondingly, two other tree ultrametrics have the same properties, and the orbit of the symmetric group action permuting the labels is the same for both sets of leaf-permuted trees, then the tropical line segments connecting these pairs coincide.

Theorem 67. *Suppose T_0 and T_0' are equidistant trees with N leaves such that $T_0 =_\sigma T_0'$. Also, suppose T and T' are equidistant trees with N leaves such that $T =_\sigma T'$. Then*

$$\Sigma(\Gamma(w^T, w^{T_0}), \sigma) = \Gamma(w^{T'}, w^{T_0'})$$

and

$$\Sigma(\Gamma(w^{T'}, w^{T_0'}), \sigma^{-1}) = \Gamma(w^T, w^{T_0}).$$

Proof. Since $T_0 =_\sigma T_0'$, we have that the differences $w^T - w^{T_0}$ and $w^{T'} - w^{T_0'}$ are equal after ordering the coordinates of $w^T - w^{T_0}$ and $w^{T'} - w^{T_0'}$ from the smallest and largest. The remainder of the proof follows the proof of Proposition 66. □

Corollary 68. *Fix T_0, and let T and T' be equidistant trees with N leaves such that $T =_\sigma T'$. Then*

$$\Sigma(\Gamma(w^T, w^{T_0}), \sigma) = \Gamma(w^{T'}, w^{T_0})$$

and

$$\Sigma(\Gamma(w^{T'}, w^{T_0}), \sigma^{-1}) = \Gamma(w^T, w^{T_0}).$$

Example 69. Let T_1, T_2 be equidistant trees in Figure 9. Here, $\sigma = \sigma^{-1} = (4, 3, 2, 1)$. Notice that $\Sigma(\Gamma(w^1, w^0), \sigma) = \Gamma(w^2, w^0)$ and $\Sigma(\Gamma(w^2, w^0), \sigma^{-1}) = \Gamma(w^1, w^0)$.

In this figure, the black points represent the trees T_1 and T_2. These trees have the same tree topology and the same branch lengths, but leaves are labeled differently. In BHV space, T_1 and T_2 are distinct trees, and thus belong to different orthants in this example, since there are only two internal edges in these trees: recall that in BHV space, within each orthant, trees are stored by their internal edge lengths, which represent coordinates. The BHV distance between T_1 and T_2 is then the unique geodesic that is the cone path, traversing the origin, illustrated by the red line. (See Figure 4 for the configuration of BHV_5 and to see that T_1 and T_2 are not in neighboring orthants.)

We have $T_1 =_\sigma T_2$, and the tropical line segment is illustrated by the blue line. See the Appendix A1 for an explicit computation of the tropical line segment for this example. The orange points represent the three remaining trees in the figure: notice that in these trees, there is only one internal edge length, and thus the orange points lie on the 1-dimensional strata partitioning the quadrants. The line segments connecting the points traverse the quadrants, and at every point along the blue lines, there are two internal edge lengths until the subsequent orange point is reached, where one internal edge contracts completely into one internal node.

5 Probability Measures and Means in Palm Tree Space

We showed in Section 3.5 that palm tree space is a Polish space, and thus exhibits desirable (though not necessary) properties for rigorous probability and statistics. Such properties ensure well-behaved measure-theoretic properties, and in particular, allow for classical probabilistic and statistical studies, such as convergence in various modes, as well as ensuring that stochastic processes are well defined. In this section, we discuss the existence of important probabilistic and statistical quantities for parametric data analysis, such as probability measures and Fréchet means and variances.

5.1 Tropical Measures of Central Tendency

For distributions in general metric spaces, there are various measures of central tendency. These may be framed in palm tree space as follows (and may be generalized by replacing the tropical metric d_{tr} with any well-defined metric).

Definition 70. Given a probability space $(\mathcal{T}_N, \mathcal{B}(\mathcal{T}_N), \mathbb{P}_{\mathcal{T}_N})$, the quantity

$$\text{Var}_{\mathbb{P}_{\mathcal{T}_N}}(y) = \int_{\mathcal{T}_N} d_{\text{tr}}(y, x)^2 d\nu(x) < \infty \tag{15}$$

is known as the *tropical Fréchet variance*. The minimizer of the quantity (15) is the *tropical Fréchet population mean* or *barycenter* μ_F of a distribution ν:

$$\mu_{\text{tr}}^F = \arg\min_y \int_{\mathcal{T}_N} d_{\text{tr}}(y, x)^2 d\nu(x) < \infty. \tag{16}$$

Definition 71. The *tropical Fermat–Weber point* of a distribution is a similarly-defined measure of central tendency, and can be thought of as a generalized median of a distribution in a general metric space:

$$\mu_{\text{tr}}^{FW} = \arg\min_y \int_{\mathcal{T}_N} d_{\text{tr}}(y, x) d\nu(x). \tag{17}$$

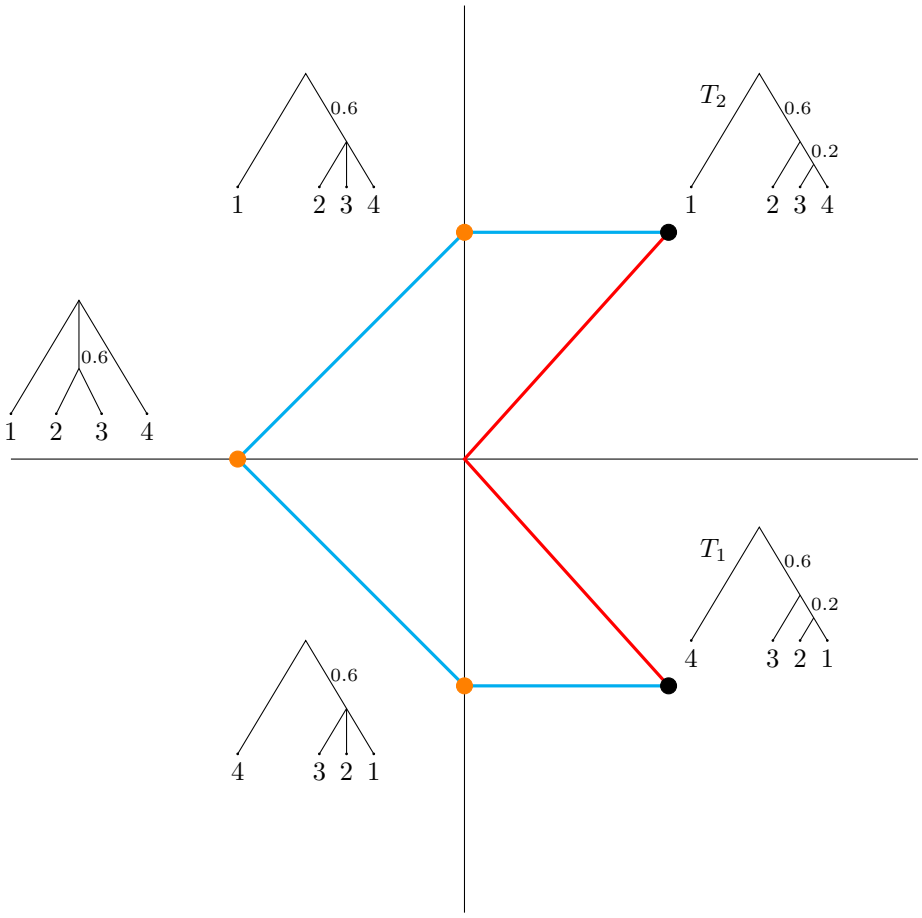

Figure 9: Equidistant trees T_1 and T_2, with tropical line segment (blue line) and the BHV geodesic (red line) for Examples 69 and 76.

For general metric spaces, neither existence nor uniqueness of (16) nor (17) are guaranteed. Ohta (2012) proves a condition under which barycenters are guaranteed to exist.

Lemma 72. *(Ohta, 2012, Lemma 3.2) If (M, d) is a proper metric space, then any distribution ν where $\int_M d(y,x)^2 d\nu(x) < \infty$ has a barycenter.*

As a consequence of the Heine–Borel theorem (see Corollary 39), palm tree space is a proper metric space. Thus, (16) evaluated according to the tropical metric is guaranteed to exist. However, since geodesics are not unique in palm tree space, Fermat–Weber points and Fréchet means will also, in general, not be unique. It is known that the set of tropical Fermat–Weber points is a classical convex polytope; Lin and Yoshida (2018) present a formal and complete treatment of the Fermat–Weber point under the tropical metric.

Understanding properties of a Fréchet mean for sets of phylogenetic trees is an important question that has been a topic of recent active research. Algorithms to compute the Fréchet mean have been proposed in the BHV setting (Skwerer et al., 2018); properties of the Fréchet mean in the extrinsic case where the mean lies outside of tree space have been recently studied (Anaya et al., 2019), and in the case where it is computed by Sturm's algorithm (Sturm, 2003), the convergence behavior remains unknown.

5.2 Tropical Probability Measures

Probability measures on combinatorial and phylogenetic trees have been previously discussed, for example by Aldous (1996) and Billera et al. (2001). This section is dedicated to an analogous discussion on palm tree

space. In $\mathcal{P}_N$, the Borel σ-algebra $\mathcal{B}(\mathcal{T}_N)$ is the σ-algebra generated by the open tropical balls B_{tr} of $\mathcal{T}_N$, given in Definition 31. We begin by providing the existence of probability measures on $\mathcal{P}_N$.

Definition 73. A *finite tropical Borel measure* on $\mathcal{T}_N$ is a map $\mu : \mathcal{B}(\mathcal{T}_N) \to [0, \infty)$ such that $\mu(\emptyset) = 0$, and for mutually disjoint Borel sets $A_1, A_2, \ldots \in \mathcal{B}(\mathcal{T}_N)$ implies that $\mu(\bigcup_{i=1}^{\infty} B_{\text{tr}}^{(i)}) = \sum_{i=1}^{\infty} \mu(B_{\text{tr}}^{(i)})$. If in addition $\mu(\mathcal{T}_N) = 1$, then μ is a *tropical Borel probability measure* on $\mathcal{T}_N$.

Since $\mathcal{T}_N$ is a finite union of polyhedra in $\mathbb{R}^n/\mathbb{R}1$ (see Proposition 29), tropical Borel probability measures exist if finite tropical Borel measures μ exist on $\mathcal{T}_N$ by an appropriate scaling of the value of μ on each polyhedron.

Alternatively, the existence of probability measures on $\mathcal{T}_N$ can be seen by considering an arbitrary probability space $(\Omega, \mathcal{F}, \mathbb{P})$ together with a measurable map $X : \Omega \to \mathcal{T}_N$. Such maps exist, since we have shown in Section 3 that $\mathcal{T}_N$ is a Polish space and thus $(\mathcal{T}_N, \mathcal{B}(\mathcal{T}_N))$ is a standard measurable space (e.g. Taylor (2012)). The probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a measure space by assumption, thus also measurable. In this case, the map X is a random variable taking values in $\mathcal{T}_N$. Then, X induces a probability measure $\mathbb{P}_{\mathcal{T}_N}$ on $(\mathcal{T}_N, \mathcal{B}(\mathcal{T}_N))$ by the pushforward measure $X_*\mathbb{P}$ of $\mathbb{P}$ under X, known as the *distribution*, for all Borel sets $A \in \mathcal{B}(\mathcal{T}_N)$:

$$X_*\mathbb{P}(A) := \mathbb{P}(X^{-1}(A)) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in A\}).$$

Example 74. If $\mathbb{P}_\vartheta$ is a probability measure parameterized by ϑ, then we may obtain parametric probability distributions on $(\mathcal{T}_N, \mathcal{B}(\mathcal{T}_N))$ induced by $\mathbb{P}_\vartheta$. In the Bayesian setting, if the measure λ gives the prior distribution of ϑ, then the joint probability measure $\mathbb{P}(\mathcal{T}_N, \vartheta)$ is given by the product measure $\mathbb{P}_{\mathcal{T}_N} \times \lambda$. Similarly, the conditional measure $\mathbb{P}(\vartheta \mid \mathcal{T}_N)$ is also proportional to this same product measure, which can be seen by Bayes' rule, where for events $B_1, B_2, \ldots$ that partition the sample space, then for any event A,

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A)}.$$

Additionally, for continuous random variables Y and Z, the conditional density $f_{Z|Y}(z \mid y)$ satisfies

$$f_{Z|Y}(z \mid y) = \frac{f_{Y|Z}(y \mid z)f_Z(z)}{f_Y(y).}$$

Example 75. Haar measures are commonly-occurring quantities in probability theory as the natural generalization of Lebesgue measure on spaces with a specified group structure. Haar measures exist on $\mathcal{P}_N$: By Proposition 43, tropical line segments between ultrametrics are themselves ultrametrics. In addition, Theorem 67 gives us a setting for the coincidence of any two tropical line segments connecting two trees of the same topology, with equal branch lengths among pairs. For measures on $\mathcal{U}_N \subset \mathcal{P}_N$, there exists a group structure on the space $\mathcal{U}_N$ given by the symmetric group $\text{Sym}(N)$. The group action is permutation of the N leaf labels, and the invariance is on tropical line segments between ultrametrics.

Specific probability measures analogous to classical probability distributions exist on $\mathcal{P}_N$, which we now exemplify. We note that these measures may be generalized to arbitrary compact subspaces of $\mathcal{T}_N$ by Corollary 39.

By Proposition 29, since $\mathcal{T}_N$ is a finite union of polyhedra in $\mathbb{R}^n/\mathbb{R}\mathbf{1}$, the *base measure* on palm tree space can be defined by assigning a probability of $1/(2N-5)!!$ to each polyhedron, and a uniform distribution within each polyhedron. Since each polyhedron is unbounded, the uniform distribution here would be an improper prior. A proper uniform distribution may be obtained by a rescaling of each polyhedron to be unitary.

An exponential family-type analog for palm tree space may be defined using the tropical metric as follows:

$$f(w) = C \cdot \exp\{d_{\text{tr}}(w, \mu_{\text{tr}}^0)\}, \tag{18}$$

where C is a normalizing constant, and μ_0 is taken to be a measure of central tendency, as discussed above in Section 5.1. Measures of the form (18) give rise to families of distributions concentrated on a tropical central tree, μ_{tr}^0.

6 Application to Data: Seasonal Influenza

We now provide an example of a statistical study in palm tree space following Yoshida et al. (2019). We demonstrate and compare statistical studies of tropical versus BHV principal component analysis (PCA) of the seasonal influenza virus by studying itsdiversity over twenty years of collected longitudinal data.

Influenza is an RNA virus affecting up to 10% of adults and 30% of children worldwide and on an annual basis, resulting in more than half a million deaths (WHO, 2016). Because of the rapid evolution of the virus genome, the development of an effective vaccine critically relies on being able to effectively visualize, analyze, and predict the viral evolution patterns in a statistically-rigorous setting.

6.1 Influenza Data

We focus on the influenza type A virus, which is an RNA virus that is classified by subtype according to the two proteins occurring on the surface of the virus: hemagglutinin (HA) and neuraminidase (NA). Here, we focus on HA, which tends to be the most variable protein in genomic evolution, in terms of changing the antigenic make-up of surface proteins. Such antigenic variability (referred to as *antigenic drift*) is an important driving factor behind vaccine failure. We restrict our study to the subtype H3N2, which is becoming increasingly abundant, and a dominant factor studied in developing flu vaccines due to its recently increasing resistance to standard antiviral drugs (Altman, 2006). It was also the cause of a recent epidemic due to vaccine failure in 2002-2003 (CDC, 2004).

Genomic data for 1089 full length sequences of hemagglutinin (HA) for influenza A H3N2 from 1993 to 2017 in the state of New York were obtained from the GI-SAID EpiFlu^{TM} database (`www.gisaid.org`) and aligned with MUSCLE (Edgar, 2004) using default settings. HA sequences from each season were related to those of the preceding season. We then applied *tree dimensionality reduction* (Zairis et al., 2016) using temporal windows of 5 consecutive seasons to create 21 datasets. The date of each dataset corresponds to the first season; for example, the dataset dated 2013 consists of 5-leaved trees where the leaves come from seasons 2013 through 2017. Each unrooted tree in these datasets was constructed using the neighbor-joining method (Saitou and Nei, 1987) with Hamming distance. Outliers were then removed from each season using KDETREES (Schardl et al., 2014). On average, there were approximately 20,000 remaining trees in each dataset. Finally, PCA was performed under the tropical metric (Yoshida et al., 2019) and under the BHV metric (Nye et al., 2017).

6.1.1 Tree Dimensionality Reduction

The influenza virus is assumed to emerge and evolve from a common ancestor: although the virus mutates each season and within each patient, each of these seasonal and patient-specific mutations can be traced back to a single virus (e.g. Liu et al. (2009)). This evolutionary pattern is depicted in a large phylogenetic tree. Tree dimensionality reduction is a sampling method that generates a data set of smaller trees via a structural and systematic sampling of the larger tree (Zairis et al., 2016). In other words, the method produces a collection of smaller trees that faithfully represents the evolutionary behavior of the single large tree, and thus allows the evolutionary information to be treated as a data set with multiple points, rather than viewing the large tree as a single datum. Specifically, this is done by randomly sampling one individual from each season over the length of the desired temporal window; each individual then gives one leaf of the reduced tree.

In the applications presented in Zairis et al. (2016), smaller trees were constructed with three, four, and five leaves; we follow within this size scale for consistency of the original method, but choose the largest number of leaves, since the two tree PCA methods were implemented in Nye et al. (2017) and Yoshida et al. (2019) for trees with a larger number of leaves.

6.2 PCA in Tree Spaces

PCA is a fundamental technique in descriptive and exploratory statistics that visualizes relationships within the data by reducing their dimensionality. As such, PCA has many important implications—for example, it projects to the subspace of the solution of k-means clustering (Ding and He, 2004)—and may be interpreted in several different ways. One interpretation may be seen as searching for a lower-dimensional plane that

minimizes the sum of squared distances from the data points to the plane, and then finding an orthogonal projection from the data points onto the plane to visualize them. More precisely, given a set X of data points $\{x_1, x_2, \ldots, x_n\}$ where each $x_i \in \mathbb{R}^m$, $i = 1, \ldots, n$, we may consider a subset of $k+1$ of these points $V = \{v_0, \ldots, v_k\} \subset \mathbb{R}^m$ and define

$$\Pi(V) := \left\{ \sum_{i=0}^{k} p_i v_i \mid p_0, \ldots, p_k \in \mathbb{R} \text{ such that } p_0 + \cdots + p_k = 1 \right\},$$

where $\Pi(V)$ is the affine subspace of $\mathbb{R}^m$ containing V. The orthogonal L^2 distance between any point $y \in \mathbb{R}^m$ and $\Pi(V)$ is denoted by $d(y, \Pi(V))$; the squared of projected distances for the data X onto $\Pi(V)$ is then

$$D_X^2(\Pi(V)) = \sum_{i=1}^{n} d(x_i, \Pi(V))^2.$$

The kth principal component Π_k is the choice of V minimizing D_X^2. In Euclidean space, Π_0 corresponds to the sample mean, while Π_1 is the regression line passing through the mean, and so on, for higher dimensions. These principal components are nested: $\Pi_0 \subset \Pi_1 \subset \Pi_2 \subset \cdots$. This interpretation relies heavily on the setting of a vector space, since $\Pi(V)$ is a linear combination of vectors, and visualizing the data on $\Pi(V)$ relies on orthogonal projection, both of which are inherently linear-algebraic notions, and hence difficult to directly implement in tree spaces.

It is important to emphasize that PCA is a descriptive and exploratory technique for data, and a form of unsupervised learning. As such, there is no *a priori* assumption on the distribution of the data, despite the coincidence of the above-mentioned description of the technique with classical linear regression, which is an inferential tool (and thus, classically, assumes some distributional properties). PCA can always be used to reduce the dimensionality of and visualize data, no matter what their distribution (and no matter what interpretation of PCA is adapted to implement the procedure).

Respective adaptations of the above-mentioned interpretation of PCA to palm tree space and BHV space are given by Nye et al. (2017) and Yoshida et al. (2019): convex, triangular regions—the tropical triangle and the *locus* of weighted Fréchet means computed with respect to the BHV metric, respectively—define notions of second principal components in the respective spaces. Following Nye (2014), in the setting of phylogenetic tree space, we seek a convex hull of $k+1$ points to represent the kth principal component, $\Pi(V)$ where now each v_i in $V = \{v_0, \ldots, v_k\}$ is a tree. Since any geodesic segment is the convex hull of its endpoints, a natural extension to the plane, or second principal component, would be to use the convex hull of three points. This also gives us a representation with projections that are readily visualizable.

In other words, we seek a 2-plane defined by three ultrametric trees and visualize the data by restricting and projecting onto the convex hull of these three points. In palm tree space, this is straightforward since tropical triangles are always 2-dimensional (Lin et al., 2017). In BHV space, however, this entails finding a locus, which is constructed from the set of discrete Fréchet means weighted by values on the probability simplex. As opposed to BHV triangles, the dimensionality of loci of weighted Fréchet means computed under the BHV metric is well behaved (Nye et al., 2017).

6.2.1 Tropical PCA

The (tropically-) convex hull we seek in palm tree space by implementing the method of Yoshida et al. (2019) represents the second principal component is the tropical triangle. Edges are given by tropical line segments between their three vertices. Projections of data points onto the tropical triangle are exact and unique.

To build the tropical triangles, we seek three points within the dataset; in this sense, the PCA method is approximate and not exact. For smaller datasets, such a search may be combinatorially tractable. For larger datasets, we implement a Markov chain Monte Carlo (MCMC) algorithm to find these three points (Kang and Yoshida, 2018). Briefly, the idea is to first choose three points from the dataset at random and calculate the sums of the tropical distances of the remaining points to the tropical triangle initially defined by the three points. One point among these vertices is then randomly chosen to be iteratively replaced by other points within the dataset to find a smaller sum of tropical distances to the resulting triangles until a minimal sum of tropical distances is found; the corresponding three vertices of the triangle are then retained

to define the second tropical principal components. Unlike in classical Euclidean PCA where the components are mutually orthogonal, the vertices of the tropical triangle onto which the data points are projected are in general not. The projections of the data points onto the tropical triangle are tropically-convex combinations of the vertices. This procedure was run 5 times on each of the 21 datasets; each iteration was run in parallel on 18 CPU cores—Intel(R) Xeon(R) W-2155 CPU @ 3.30 GHz—and took approximately two hours to terminate, for a total of around 210 CPU hours to complete the tropical PCA procedure.

6.2.2 BHV PCA

Since geodesic triangles (and more generally, polytopes) in BHV space are not well-behaved convex hulls (see Corollary 13), they are thus problematic candidates for BHV second principal components. An alternative method proposed by Nye et al. (2017) searches for the *locus* of weighted, discrete, Fréchet means computed with respect to the BHV metric as a BHV kth principal component.

Briefly, for the weighted, discrete, BHV Fréchet mean,

$$\mu(V, w) = \arg \min_{y \in \mathcal{T}_N^{\mathrm{BHV}}} \sum_{i=0}^{k} w_i \cdot d_{\mathrm{BHV}}(y, v_i)^2,$$

if the weights are given by elements in the k-dimensional simplex of probability vectors,

$$\mathcal{S}^k := \left\{ (p_0, \ldots, p_k) \mid p_i \geq 0,\, i = 0, \ldots, k,\, \sum_{i=0}^{k} p_i = 1 \right\},$$

then $\Pi(V)$ defined by

$$\Pi(V) := \{\mu(V, p) \mid p \in \mathcal{S}^k\}$$

is the locus of the Fréchet mean of V. Geometric properties of the locus, such as convexity and dimensionality, are well-behaved, as opposed to BHV convex hulls. In our application, we implement the algorithm by Bačák to compute the weighted, discrete, BHV Fréchet means and thus construct the locus (Bačák, 2014). Projections of data points onto the locus are approximate and need not be unique.

Similar to the tropical PCA case described above, the procedure is approximate and uses trees from the dataset to construct the locus and its boundaries. We follow the algorithm proposed in the original reference, which is also a stochastic optimization algorithm. As in the tropical PCA case, the procedure was run 5 times on each of the 21 datasets; each iteration was run in parallel on 18 CPU cores—Intel(R) Xeon(R) W-2155 CPU @ 3.30 GHz—and also took approximately two hours to terminate, for a total of also around 210 CPU hours to complete the BHV PCA procedure.

Software and Data Availability

Software to compute both tropical and BHV PCA is publicly available in R and Java code. Their implementation to the influenza data described in this paper is located on the FluPCA GitHub repository at `https://github.com/antheamonod/FluPCA`.

The data used in this paper were obtained from publicly available sources and preprocessed, as detailed in Section 6.1. The final version used in the analyses in this paper are also publicly available on the FluPCA GitHub repository.

The resulting figures from both BHV and tropical PCA projections for all 21 data sets are also available on the FluPCA GitHub repository.

6.3 Interpretation of Tree PCA

In the tropical case, the second principal component represented by the tropical triangle—whose vertices are given by three ultrametric trees, and whose edges are given by the tropical line segments between them—divides into cells, which are determined by the tree topologies that an edge (tropical line segment) traverses, as given by Theorem 57. Trees in the dataset are then projected into the cells corresponding to

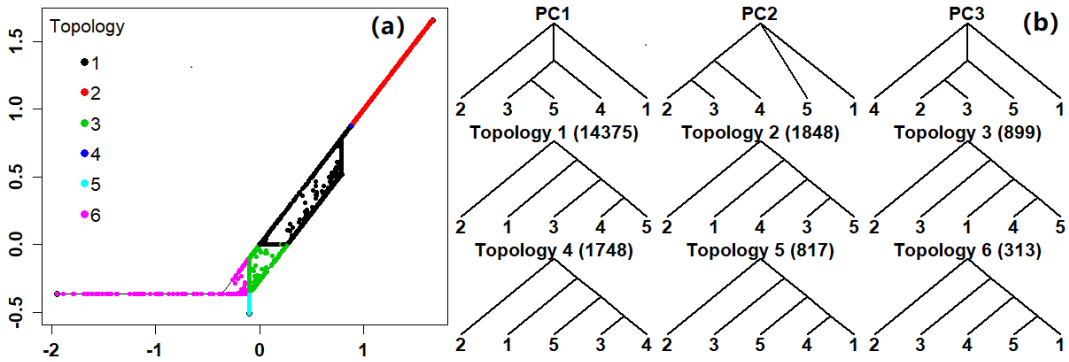

Figure 10: 2008: The tropical triangle as the second tropical principal component. (a) Tropical triangle and projected data points; (b) Vertices of the tropical triangle and projected tree topologies.

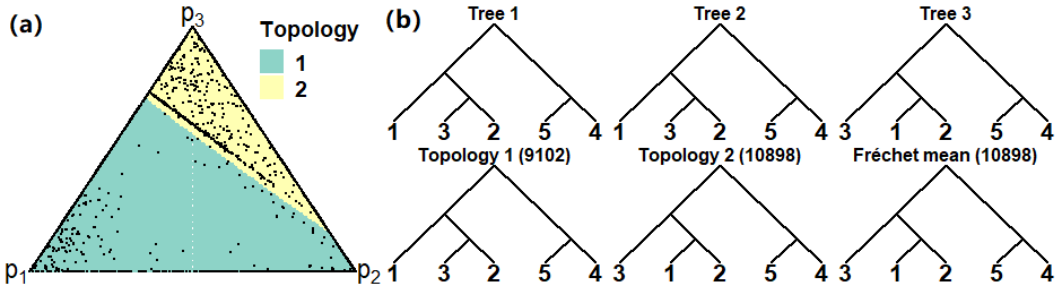

Figure 11: 2008: The locus of BHV Fréchet means as the second principal component. (a) The simplex shaded by topology of corresponding points on the affine subspace; (b) Trees 1, 2, and 3 correspond to three weighted Fréchet means.

their topologies in the PCA visualization. The simplest case in both tropical PCA is where all three vertices of the triangle are of the same tree topology, then there will only be one cell and all projections will be of the same topology; if two points are trees of the same tree topology, then every point on the tropical line segment connecting them will also be of the same tree topology. An example can be seen in the 1993 data set, available at `https://github.com/antheamonod/FluPCA/Figures`.

A more interesting example can be seen in the 2008 dataset in Figure 10. The tropical triangle divides into six cells; a complete discussion on tropical polytopes, their decomposition into cells and their self-duality can be found in Section 5.2 of Maclagan and Sturmfels (2015), Theorem 5.2.21 in particular gives the isomorphism from a tropical polytope to its self-dual. As conjectured in Yoshida et al. (2019), one possible biological interpretation of the cells is that neighboring cells represent similar yet distinct tree structures differing by tree rearrangements of one move. Tree rearrangements are used in algorithms where the goal is to search for the optimal tree structure in statistical tree reconstruction methods (see e.g. Bordewich and Semple (2005); Felsenstein (2004)). The tropical line segment forming the topmost edge of the triangle traverses five different tree topologies (indicated by the pink, green, black, blue, and red points along the topmost line segment), which may be calculated via Theorem 57, indicating that the three vertices of the triangle are quite different from one another in terms of tree topology.

The BHV locus, which represents the second principal component in BHV space, is also generated by three vertices (ultrametric trees), and depicted in the BHV PCA plots by a triangle. Here, varying tree topologies are depicted by the multicolored patches within the triangular region, indicating that the locus straddles several orthants in BHV space. In the 2008 dataset in Figure 11, the locus straddles two BHV orthants and we see two tree topologies occurring among the projected points.

In the 2008 dataset, and as can be seen in Example 76 implementing Algorithm 1 for computing tropical line segments, versus the BHV geodesic for the same trees discussed in Example 69, palm tree space and the tropical geometric approach to computations in phylogenetic tree space appears to allow for occurrences of richer and more subtle structures and methods: in these examples, we see tropical PCA projections of six

Table 1: Proportion of Explained Variance for Tropical and BHV PCA

	1993	1994	1995	1996	1997	1998	1999
Tropical	0.7269	0.8505	**0.9577**	0.7482	0.8437	0.8790	0.8564
BHV	0.3019	0.4347	0.3151	0.5025	0.0505	0.6408	**0.9524**
	2000	2001	2002	2003	2004	2005	2006
Tropical	0.7942	0.8302	0.9525	0.8622	0.7931	0.8304	0.7300
BHV	**0.0014**	0.9488	0.8962	0.4927	0.3651	0.3634	0.2383
	2007	2008	2009	2010	2011	2012	2013
Tropical	0.6995	**0.4637**	0.6289	0.6665	0.5920	0.5568	0.5624
BHV	0.2727	0.0460	0.1563	0.1935	0.2771	0.1998	0.1279

different topologies, versus two in the BHV case; and in the case of computing tropical line segments, we see the sister leaves labeled 2 and 3 extending a nonzero-length branch of the root in Figure 12, as well as a line segment that bypasses the origin at a lower computational complexity.

6.4 Results: Proportion of Explained Variance R^2

In terms of explained variance given in Table 1, we see that in general, tropical PCA is able to explain more of the variance in the data than does BHV PCA. BHV PCA results also have a higher variability over a wider range than tropical PCA: BHV PCA explains between ∼0.0% and 95% of the variance, while tropical PCA explains between 46% and 96%.

7 Discussion

In this paper, we defined palm tree space as the space of phylogenetic trees with N leaves, endowed with the tropical metric. We gave results on its analytic, topological, geometric, and combinatorial properties and showed that they are conducive to rigorous treatments and studies in probability, and demonstrated that descriptive statistical analysis on real data is also feasible. We showed that in certain respects, certain properties of palm tree space are more natural and amenable to statistical analyses than BHV space. An important difference, however, is that geodesics in palm tree space are not unique.

In applied and structural settings, we see that palm tree space suggests a setting for richer and more interesting behavior of tree topologies. This can be concretely seen in the algorithm computing tropical line segments and in our real data analyses of tropical PCA. Our work thus invites the reinterpretation of existing statistical methods in terms of the tropical metric to make a wider array of exact analyses readily available and interpretable to phylogenetic research with potential impact for biological discoveries.

Biological and Statistical Implications

Despite the statistical challenges of arbitrariness of dimension of BHV polytopes and stickiness described in Section 3.1, the BHV parameterization has been successfully implemented to reveal important biological findings (e.g. Zairis et al. (2014)). In terms of interpretation, the unresolved singularities of BHV Fréchet means translate to "indecisiveness" of which branching patterns or tree topologies are "preferred," which is consistent with what is often seen in some biological settings where the trees arise from sequence alignment. However, mathematically, trees are used to model other biological phenomena, such as pulmonary paths as airway trees (e.g. Feragen et al. (2013, 2015)); brain growth and structure (e.g. Yan and Yan (2013)); and neuronal morphologies (e.g. Kanari et al. (2018)). Such a probabilistic assumption may not be reasonable in these other settings. Given recent research interest in developing methods to bypass these difficulties support the goal of our work, which is that alternative representations is an important direction (e.g. Anaya et al. (2019); Skwerer et al. (2018)). An important and interesting direction for future research is the identification of non-uniform probability distributions in the tropical setting, which is challenging yet promising: Tran (2018) outlines various ways in which tropical Gaussian distributions may be constructed.

In the context of shape statistics (Kendall, 1989) and computational anatomy (Grenander and Miller, 1998), the data objects of interest are often modeled as elements of algebraic spaces. In particular, these algebraic spaces are quotient spaces generated by group actions. Recent work has studied the behavior of estimators on such spaces, uncovering undesirable properties, such as biasedness and inconsistency when the group actions are random (that is, when the quotient spaces are generated by elements of the group are chosen at random to act on the topological space) or continuous (as in the case of Lie groups acting on Riemannian manifolds) (Devilliers et al., 2017; Miolane and Pennec, 2015). Nonparametric methods have been developed to bypass the problem of inconsistency by Bhattacharya and Patrangenaru (2014). As previously mentioned in Section 2.2.3, the tropical projective torus $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ is a quotient space that may be generated by a group action, however the biasedness and inconsistency in previous work arise due to the poor behavior of the transformed metric after it is mapped into the quotient space, which results in a pseudometric. In our case, the tropical metric is well-behaved and defined directly on the quotient space, therefore differing in setting to previous work.

It should also be noted (as discussed in Section 3.3.1) that the non-uniqueness property of geodesics in palm tree space poses computational difficulties, but does not prohibit statistical analysis and can still yield useful descriptive as well as inferential information, for example, on clustering behavior. Another important setting where geodesics are not unique is that of positively-curved spaces. Bhattacharya and Bhattacharya (2008) study asymptotic behavior and distributions on Riemannian manifolds, including positively curved manifolds. Recent work such as that by Kobayashi and Wynn (2019) develops techniques for data analysis on curved spaces by tuning the geodesic metrics accordingly. In particular, a general Fréchet function is defined, and its parameters are chosen accordingly, depending on the goal: for example, one geodesic metric may be transformed into another to control the curvature of the space for data analysis. There are large bodies of existing work in related areas on curved spaces, for example, in the case of manifold learning; shape statistics; Wasserstein spaces for probability measures; and information geometry. Though these domains each have their own specific goals and studies, data analysis and computation play a central role in these settings. Moreover, there are known settings in these related areas where positive curvature, and thus non-uniqueness of geodesics, arises (for example, the 2-Wasserstein space for Gaussian measures is positively curved (Takatsu, 2011)). Adapting existing techniques in these settings to statistical analysis in palm tree space is an important direction of research that merits exploration.

Acknowledgments

The authors are grateful to Robert J. Adler, Omer Bobrowski, Ioan Filip, Maia Fraser, Stephan Huckemann, Elliot Paquette, Juan Ángel Patiño-Galindo, Yitzchak (Elchanan) Solomon, and Bernd Sturmfels for helpful discussions. We are especially indebted to Carlos Améndola and Hossein Khiabanian for their extensive commentary, suggestions, and support throughout the course of this project.

A.M. and R.Y. wish to acknowledge the Mathematisches Forschunginstitut Oberwolfach (MFO) for hosting their visit in January 2018, which inspired this work. B.L. wishes to acknowledge support by the Simons Foundation and the hospitality of the Institut Mittag-Leffler during the spring of 2018, where a large part of this research was carried out.

The authors would also like to acknowledge the GI-SAID EpiFlu™ initiative for making the data used in this study publicly available.

R.Y. is supported in part by NSF DMS #1622369 and #1916037. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the funders.

Author Contributions Statement

A.M. conceived the study and designed the research. A.M., B.L., and R.Y. performed research. A.M. and R.Y. developed the algorithms; Q.K. implemented the software and performed the analyses. A.M. wrote the manuscript.

Competing Financial Interests Statement

The authors declare that no competing financial interests exist.

Appendix

A1 Algorithm for Computing Tropical Line Segments Between Ultrametrics

We now give an algorithm giving tropical line segments between equidistant trees with N leaves. It takes as input two ultrametrics: $w^1 = \left(w^1_{\{1,2\}}, \ldots, w^1_{\{N-1,N\}}\right)$ associated with an equidistant tree T_1 with N leaves; and $w^2 = \left(w^2_{\{1,2\}}, \ldots, w^2_{\{N-1,N\}}\right)$ associated with an equidistant tree T_2 with N leaves. It returns the tropical line segment between T_1 and T_2 in $\mathcal{U}_N$.

Algorithm 1 Algorithm for Computing Tropical Line Segments Between Ultrametrics

procedure TropLine(w^1, w^2) ▷ Tropical line segment between ultrametrics w^1 and w^2
  $w^1 \leftarrow \left(w^1_{\{1,2\}}, \ldots, w^1_{\{N-1,N\}}\right)$
  $w^2 \leftarrow \left(w^2_{\{1,2\}}, \ldots, w^2_{\{N-1,N\}}\right)$
  $\lambda' \leftarrow \left(w^2_{\{1,2\}} - w^1_{\{1,2\}}, \ldots, w^2_{\{N-1,N\}} - w^1_{\{N-1,N\}}\right)$ ▷ Compute the difference between T_2 and T_1
  $\lambda \leftarrow \left(\lambda'_{(1)}, \ldots, \lambda'_{(n)}\right)$ ▷ Reorder the elements of λ' from smallest to largest
  $L = \emptyset$
  for $i = 1, \ldots, n$ **do**
    $w_{\{i,i+1\}} \leftarrow \lambda_i \odot w^1_{\{i,i+1\}}$
    $y_{\{i,i+1\}} \leftarrow w_{\{i,i+1\}} \boxplus w^2_{\{i,i+1\}}$
  $y \leftarrow \left(y_{\{1,2\}}, \ldots, y_{\{N-1,N\}}\right)$
  if $\exists\ y_i > 2$ **then**
    increment $\leftarrow \max(y_i) - 2$
    for $i = 1, \ldots, n$ **do**
      $y_i \leftarrow y_i -$ increment ▷ Rescale branch lengths to obtain a unitary equidistant tree
      $y \leftarrow \left(y_1, \ldots, y_n\right)$
  else
    $y \leftarrow \left(y_{\{1,2\}}, \ldots, y_{\{N-1,N\}}\right)$
  $L \leftarrow L \cup \{y\}$
  return L ▷ Tropical line segments connecting the ultrametrics in L

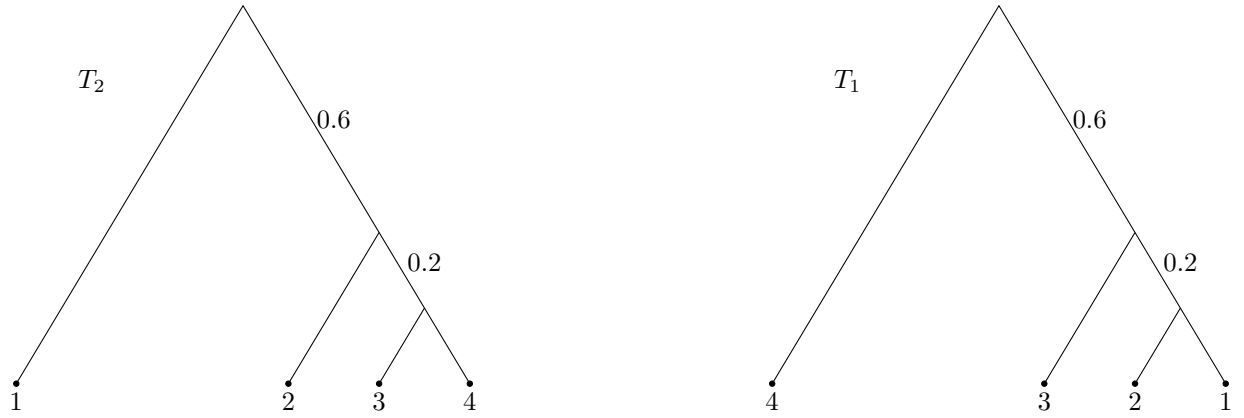

Figure 12: Trees T_2 and T_1 from Figure 9.

Example 76. We revisit the equidistant trees T_2 and T_1 from Figure 9, and compute the tropical line segment (blue line) in the figure using Algorithm 1, from T_2 to T_1. Their ultrametrics are

$$w^2 = (2,\ 2,\ 2,\ 0.8,\ 0.8,\ 0.4),$$
$$w^1 = (0.4,\ 0.8,\ 2,\ 0.8,\ 2,\ 2).$$

First, we calculate the difference between w^2 and w^1, $\lambda' = (1.6,\ 1.2,\ 0,\ 0,\ -1.2,\ -1.6)$, and reorder the elements of λ' from smallest to largest: $\lambda = (-1.6,\ -1.2,\ 0,\ 1.2,\ 1.6)$.

1. For $\lambda_1 = -1.6$, we have

$$\begin{aligned} w_1 := \lambda_1 \odot w^1 &= (-1.6+0.4,\ -1.6+0.8,\ -1.6+2,\ -1.6+0.8,\ -1.6+2,\ -1.6+2) \\ &= (-1.2,\ -0.8,\ 0.4,\ -0.8,\ 0.4,\ 0.4) \\ y_1 := w_1 \boxplus w^2 &= \big(\max(-1.2,2),\ \max(-0.8,2),\ \max(0.4,2),\ \max(-0.8,0.8),\ \max(0.4,0.8),\ \max(0.4,0.4)\big) \\ &= w^2 = (2,\ 2,\ 2,\ 0.8,\ 0.8,\ 0.4) \end{aligned}$$

2. For $\lambda_2 = -1.2$, we have

$$\begin{aligned} w_2 := \lambda_2 \odot w^1 &= (-1.2+0.4,\ -1.2+0.8,\ -1.2+2,\ -1.2+0.8,\ -1.2+2,\ -1.2+2) \\ &= (-0.8,\ -0.4,\ 0.8,\ -0.4,\ 0.8,\ 0.8) \\ y_2 := w_2 \boxplus w^2 &= \big(\max(-0.8,2),\ \max(-0.4,2),\ \max(0.8,2),\ \max(-0.4,0.8),\ \max(0.8,0.8),\ \max(0.8,0.4)\big) \\ &= (2,\ 2,\ 2,\ 0.8,\ 0.8,\ 0.8) \end{aligned}$$

3. For $\lambda_3 = 0$, we have

$$\begin{aligned} w_3 := \lambda_3 \odot w^1 = w^1 &= (0.4,\ 0.8,\ 2,\ 0.8,\ 2,\ 2) \\ y_3 := w_3 \boxplus w^2 &= \big(\max(0.4,2),\ \max(0.8,2),\ \max(2,2),\ \max(0.8,0.8),\ \max(2,0.8),\ \max(2,0.4)\big) \\ &= (2,\ 2,\ 2,\ 0.8,\ 2,\ 2) \end{aligned}$$

4. For $\lambda_4 = 1.2$, we have

$$\begin{aligned} w_4 := \lambda_4 \odot w^1 &= (1.2+0.4,\ 1.2+0.8,\ 1.2+2,\ 1.2+0.8,\ 1.2+2,\ 1.2+2) \\ &= (1.6,\ 2,\ 3.2,\ 2,\ 3.2,\ 3.2) \\ y_4 := w_4 \boxplus w^2 &= \big(\max(1.6,2),\ \max(2,2),\ \max(3.2,2),\ \max(2,0.8),\ \max(3.2,0.8),\ \max(3.2,0.4)\big) \\ &= (2,\ 2,\ 3.2,\ 2,\ 3.2,\ 3.2) \\ &= (0.8,\ 0.8,\ 2,\ 0.8,\ 2,\ 2) \end{aligned}$$

5. For $\lambda_5 = 1.6$, we have

$$\begin{aligned} w_5 := \lambda_5 \odot w^1 &= (1.6+0.4,\ 1.6+0.8,\ 1.6+2,\ 1.6+0.8,\ 1.6+2,\ 1.6+2) \\ &= (2,\ 2.4,\ 3.6,\ 2.4,\ 3.6,\ 3.6) \\ y_5 := w_5 \boxplus w^2 &= \big(\max(2,2),\ \max(2.4,2),\ \max(3.6,2),\ \max(2.4,0.8),\ \max(3.6,0.8),\ \max(3.6,0.4)\big) \\ &= (2,\ 2.4,\ 3.6,\ 2.4,\ 3.6,\ 3.6) \\ &= (0.4,\ 0.8,\ 2,\ 0.8,\ 2,\ 2) = w^1 \end{aligned}$$

The trees y_i that we obtain for each value of λ_i are shown in sequence in Figure 13. Notice that $w^2 = y_1 =_\sigma y_5 = w^1$, and $y_2 =_\sigma y_4$.

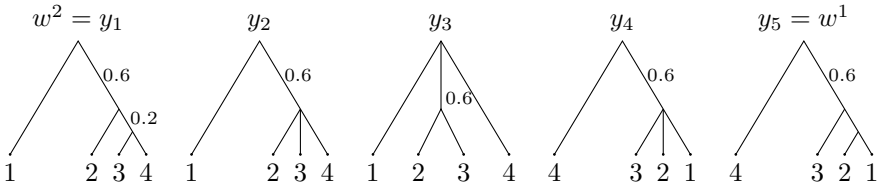

Figure 13: Trees y_i for $i = 1, \ldots, 5$ given by Algorithm 1 from T_2 to T_1 given in Figure 12.

A2 Proof of Theorem 53

We recall Definition 52 here, which appears frequently in the proof of Theorem 53, which we restate and now prove.

Definition. Let F be a tree topology on $[N]$ and $\bar{F} = \{S \in F \mid |S| \geq 3\} \cup \{[N]\}$. F is said to be *bifurcated* if for every $S \in \bar{F}$, exactly one of the following holds:

(a) there exists a proper subset $S' \subset S$ such that $S' \in F$ and $|S'| = |S| - 1$; or

(b) there exist two proper subsets $S', S'' \subset S$ such that $S', S'' \in F$ and $S' \cup S'' = S$.

Note that in (b), we must have that $S' \cap S'' = \emptyset$.

Theorem. *Let F be a tree topology on $[N]$. The following are equivalent:*

(1) F is full dimensional;

(2) for $1 \leq i < j < k \leq N$, two of the pairs $\{i,j\}, \{i,k\}, \{j,k\}$ are $=_F$, and the third pair is $<_F$ than the two F-equivalent pairs;

(3) F is bifurcated;

(4) every phylogenetic tree with tree topology F is binary.

Proof. (3) $\Rightarrow$ (2): Suppose F is bifurcated and consider distinct elements $i, j, k \in [N]$. By Lemma 50, any two of $\{i,j\}, \{i,k\}, \{j,k\}$ are comparable with respect to $=_F$ or $<_F$. If the three pairs are all $=_F$, then by Definition 49, their closures in F are equal. Let this closure be $S \in F \cup \{[N]\}$, then $i, j, k \in S$ and $|S| \geq 3$, thus $S \in \bar{F}$. Since F is bifurcated, condition (a) or (b) in Definition 52 holds for S. If (a) holds, then there exists $S' \in F$ such that S' is a proper subset of S with $|S'| = |S| - 1$. In this case, at least two of i, j, k belong to S', and the closure of the pair formed by these elements is contained in S'—a contradiction. So we may assume $\{i,k\} <_F \{i,j\}$.

We now need to show F-equivalence between $\{i,j\}$ and $\{j,k\}$: by Definition 49, there exists $S_1 \in F$ such that $\{i,k\} \subseteq S_1$ but $\{i,j\} \not\subseteq S_1$. Then $i, k \in S_1$ and $j \notin S_1$. Now for any $S_2 \in F$, if $\{j,k\} \subseteq S_2$, then S_2 and S_1 are not disjoint. Since $j \in S_2 \backslash S_1$, we must have $S_1 \subseteq S_2$. Then $i \in S_2$, and $\{i,k\} \subseteq S_2$. Since $\{j,k\} \not\subseteq S_1$, by definition, $\{i,k\} <_F \{j,k\}$. In addition, if an element of F is a superset of $\{i,j\}$, then it also contains k and thus is also a superset of $\{j,k\}$. Conversely, being a superset of $\{j,k\}$ implies that it also contains i, and thus is also a superset of $\{i,j\}$. Therefore $\{i,j\} =_F \{j,k\}$, and (2) holds.

(2) $\Rightarrow$ (3): Suppose (2) holds for F. We will show that F is bifurcated: for any subset $S \in \bar{F}$, consider all maximal proper subsets $M_1, \ldots, M_m$ of S that are clades in F. For any two such maximal subsets, since neither can be a subset of the other by definition, they must be disjoint. If $m \geq 3$, we can choose $i, j, k \in [N]$ from M_1, M_2, M_3 respectively. Then $\mathrm{cl}_F(\{i,j\}) = \mathrm{cl}_F(\{i,k\}) = \mathrm{cl}_F(\{j,k\}) = S$ and thus $\{i,j\} =_F \{i,k\} =_F \{j,k\}$—a contradiction. Therefore, m must be either 1 or 2.

If $m = 1$, then $2 \leq |M_1| \leq |S| - 1$. Suppose $|M_1| \leq |S| - 2$, we can choose two elements $i, j \in S \backslash M_1$ and another element $k \in M_1$. Then we also have $\mathrm{cl}_F(\{i,j\}) = \mathrm{cl}_F(\{i,k\}) = \mathrm{cl}_F(\{j,k\}) = S$ and $\{i,j\} =_F \{i,k\} =_F \{j,k\}$—a contradiction. Hence $|M_1| = |S| - 1$ and condition (a) holds. If $m = 2$, we already have disjoint $M_1, M_2 \in F$ that are proper subsets of S. If $M_1 \cup M_2 \neq S$, we may choose $i \in S \backslash (M_1 \cup M_2)$

and $j \in M_1, k \in M_2$. Then $\mathrm{cl}_F(\{i,j\}) = \mathrm{cl}_F(\{i,k\}) = \mathrm{cl}_F(\{j,k\}) = S$ and $\{i,j\} =_F \{i,k\} =_F \{j,k\}$—a contradiction. Hence $M_1 \cup M_2 = S$ and condition (b) holds. Hence F is bifurcated.

(1) $\Rightarrow$ (3): We proceed by induction on N. When $N = 3$ and F is full dimensional, then $|F| = 1$ and F consists of one 2-element subset. Thus, condition (a) holds for $\{1,2,3\}$ and F is bifurcated.

Suppose (1) $\Rightarrow$ (3) holds for eligible integers less than N. Let F be a full-dimensional nested set on $[N]$, then $|F| = N-2$. Consider the maximal elements $S_1, \ldots, S_k \in F$, $k \geq 1$, with respect to set inclusion. Then by the case (ii) in the proof of Lemma 46, $|F| \leq N - k$. Hence $k \leq 2$.

If $k = 1$, there exists a unique maximal element $S \in F$ with $|S| \leq N-1$. Then $F\backslash\{S\}$ is a nested set on the ground set S, and $N - 3 = |F\backslash\{S\}| \leq |S| - 2 \leq N - 1 - 2 = N - 3$. So $|S| = N-1$, and condition (a) holds for $[N]$. In addition, $|F\backslash\{S\}| = |S| - 2$, so $F\backslash\{S\}$ is full dimensional. By the induction hypothesis, all elements in $\bar{F}\backslash\{S\}$ satisfy either condition (a) or (b). Note that $\bar{F} = \bar{F}\backslash\{S\} \cup \{[N]\}$, so (3) also holds for F.

If $k = 2$, all equalities hold in (7), so there are two maximal elements $S_1, S_2 \in F$ with $S_1 \cap S_2 = \emptyset$ and $|S_1| + |S_2| = N$. So condition (b) holds for $[N]$. Let F_i be the set of the proper subsets of S_i that belong to F for $i = 1, 2$. Then both F_i are full-dimensional tree topologies on their respective ground sets S_i. By the induction hypothesis, both F_i are bifurcated and all elements in $\bar{F}_i$ satisfy either condition (a) or (b). Note that $\bar{F} = F \cup \{[N]\} = \bar{F}_1 \cup \bar{F}_2 \cup \{[N]\}$, hence F is also bifurcated, which completes the transition step.

(3) $\Rightarrow$ (1): We proceed by induction on N. When $N = 3$ and F is bifurcated, then condition (a) holds for $\{1,2,3\}$ and F only contains one 2-element subset, so F is full dimensional.

Suppose (3) $\Rightarrow$ (1) holds for all eligible integers less than N. For any bifurcated nested set F on $[N]$, either condition (a) or (b) holds for the set $[N]$. If (a) holds, then there exists $S \in F$ such that $|S| = N-1$. Then all elements in $F\backslash\{S\}$ are proper subsets of S and they thus form a nested set on the ground set S. This nested set is also bifurcated, by the induction hypothesis, so it is full dimensional. So $|F\backslash\{S\}| = |S| - 1 = N - 2$ and $|F| = |F\backslash\{S\}| + 1 = N - 1$, F is full dimensional.

If condition (b) holds, then there exist disjoint $S_1, S_2 \in F$ such that $S_1 \cup S_2 = [N]$. Let F_i be the elements of F that are proper subsets of S_i for $i = 1, 2$. Then $F = F_1 \cup F_2 \cup \{S_1, S_2\}$. Each F_i is a nested set on the ground set S_i (and may be empty when $|S_i| = 2$); it is still bifurcated. By the induction hypothesis, F_i is full dimensional and $|F_i| = |S_i| - 2$. Then $|F| = (|S_1| - 2) + (|S_2| - 2) + 2 = N - 2$, so F is still full dimensional. This completes the transition step.

(3) $\Rightarrow$ (4): Suppose F is bifurcated and a rooted phylogenetic tree T has tree topology F. Let v be a non-leaf node. It suffices to show that v has degree 3. We consider two cases:

(i) Suppose v is not the root of T. Then there is a unique path from the root of T to v. Along the path, there is an edge connecting v, and this edge corresponds to a clade S in F. Since F is bifurcated, S satisfies either conditions in Definition 51. If there exists a proper subset S' such that S' is also a clade of F and $|S'| = |S| - 1$, then all other edges connecting v include one edge connecting to the leaf in $S\backslash S'$ and one edge corresponding to S'. Otherwise there exist clades S', S'' of F such that $S' \cup S'' = S$, then all other edges connecting v include the two edges corresponding to S' and S''. In either case, v is trivalent.

(ii) Suppose v is the root of T. Then v is connected to the virtual leaf 0. Since $[N] \in \bar{F}$, $[N]$ satisfies either conditions in Definition 51 and v is connecting to two other edges in either case, so v is also trivalent.

(4) $\Rightarrow$ (3): Suppose a binary rooted tree T has tree topology F. For any clade $S \in \bar{F}$, S corresponds to an edge e of T. Let v be the vertex of e with greater distance to the root of T. Since T is binary, v is trivalent and connects to other two edges e' and e''. Each leaf in S has a unique path to v, which must contain e' or e''. This admits a partition of S into two nonempty subsets. If both subsets have cardinality of at least 2, then they are both clades of F, and we have $S' \cap S'' = S$. Otherwise one of them is a singleton, say S'', and thus S' is a clade of F with $|S'| = |S| - 1$. Thus S satisfies the condition in Definition 51. For $[N] \in \bar{F}$, since T is binary, the root of T is also connected by two edges other than the edge to the virtual leaf. The reasoning above applies to $[N]$, and F is bifurcated. □

References

Akian, M., S. Gaubert, V. Niţică, and I. Singer (2011). Best Approximation in Max-plus Semimodules. *Linear Algebra and its Applications 435*(12), 3261–3296.

Aldous, D. (1996). Probability Distributions on Cladograms. In D. Aldous and R. Pemantle (Eds.), *Random Discrete Structures*, New York, NY, pp. 1–18. Springer New York.

Allman, E. S. and J. A. Rhodes (2008). Phylogenetic ideals and varieties for the general Markov model. *Advances in Applied Mathematics 40*(2), 127–148.

Altman, L. K. (2006). This Season's Flu Virus Is Resistant to 2 Standard Drugs. https://www.nytimes.com/2006/01/15/health/this-seasons-flu-virus-is-resistant-to-2-standard-drugs.html.

Améndola, C., M. Casanellas, and L. D. G. Puente (2018). Tapas of Algebraic Statistics. *Notices of the AMS 65*(8).

Amenta, N., M. Godwin, N. Postarnakevich, and K. S. John (2007). Approximating Geodesic Tree Distance. *Information Processing Letters 103*(2), 61–65.

Anaya, M., O. Anipchenko-Ulaj, A. Ashfaq, J. Chiu, M. Kaiser, M. S. Ohsawa, M. Owen, E. Pavlechko, K. St. John, S. Suleria, K. Thompson, and C. Yap (2019). Properties for the Fréchet Mean in Billera–Holmes–Vogtmann Treespace. *arXiv:1907.05937*.

Ardila, F. and C. J. Klivans (2006). The Bergman Complex of a Matroid and Phylogenetic Trees. *Journal of Combinatorial Theory, Series B 96*(1), 38–49.

Baez, J. C. and N. Otter (2017). Operads and phylogenetic trees. *Theory and Applications of Categories 32*(40), 1397–1453.

Bandelt, H.-J. and A. W. Dress (1992). A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics 92*(1), 47–105.

Barden, D. and H. Le (2018). The logarithm map, its limits and Fréchet means in orthant spaces. *Proceedings of the London Mathematical Society 117*(4), 751–789.

Barden, D., H. Le, M. Owen, et al. (2013). Central limit theorems for fréchet means in the space of phylogenetic trees. *Electronic journal of probability 18*.

Bačák, M. (2014). Computing Medians and Means in Hadamard Spaces. *SIAM Journal on Optimization 24*(3), 1542–1566.

Benner, P., M. Bačák, and P.-Y. Bourguignon (2014). Point Estimates in Phylogenetic Reconstructions. *Bioinformatics 30*(17), i534–i540.

Bhattacharya, A. and R. Bhattacharya (2008). Statistics on Riemannian Manifolds: Asymptotic Distribution and Curvature. *Proceedings of the American Mathematical Society 136*(8), 2959–2967.

Bhattacharya, R. and V. Patrangenaru (2014). Statistics on manifolds and landmarks based image analysis: A nonparametric theory with applications. *Journal of Statistical Planning and Inference 145*, 1–22.

Billera, L. J., S. P. Holmes, and K. Vogtmann (2001). Geometry of the Space of Phylogenetic Trees. *Advances in Applied Mathematics 27*(4), 733–767.

Bordewich, M. and C. Semple (2005, Jan). On the Computational Complexity of the Rooted Subtree Prune and Regraft Distance. *Annals of Combinatorics 8*(4), 409–423.

Bourbaki, N. (2004). Elements of Mathematics: Integration I.

Buneman, P. (1974). A Note on the Metric Properties of Trees. *Journal of Combinatorial Theory, Series B 17*(1), 48–50.

Butkovič, P. (2010). *Max-Linear Systems: Theory and Algorithms.* Springer Monographs in Mathematics. London: Springer–Verlag.

Cardona, G., A. Mir, F. Rosselló, L. Rotger, and D. Sánchez (2013). Cophenetic Metrics for Phylogenetic Trees, After Sokal and Rohlf. *BMC Bioinformatics 14*(1), 3.

Cauchy, A.-L. (1821). Sur les formules qui résultent de l'emploie du signe et sur $>$ ou $<$, et sur les moyennes entre plusieurs quantités. *Cours d'Analyse, 1er Partie: Analyse algébrique*, 373–377.

CDC (2004). Preliminary assessment of the effectiveness of the 2003-04 inactivated influenza vaccine–Colorado, December 2003. *MMWR. Morbidity and mortality weekly report 53*(1), 8.

Cohen, G., S. Gaubert, and J.-P. Quadrat (2004). Duality and Separation Theorems in Idempotent Semimodules. *Linear Algebra and its Applications 379*, 395–422. Special Issue on the Tenth ILAS Conference (Auburn, 2002).

Conway, J. B. (2014). *A Course in Point Set Topology.* Undergraduate Texts in Mathematics. Springer, Cham.

Devadoss, S. L. and J. Morava (2015). Navigation in Tree Spaces. *Advances in Applied Mathematics 67*, 75–95.

Devilliers, L., S. Allassonnière, A. Trouvé, and X. Pennec (2017). Template Estimation in Computational Anatomy: Fréchet Means Top and Quotient Spaces Are Not Consistent. *SIAM Journal on Imaging Sciences 10*(3), 1139–1169.

Ding, C. and X. He (2004). K-means clustering via principal component analysis. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pp. 29. ACM.

Dress, A. W. (1984). Trees, tight extensions of metric spaces, and the cohomological dimension of certain groups: A note on combinatorial properties of metric spaces. *Advances in Mathematics 53*(3), 321–402.

Edgar, R. C. (2004, 03). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research 32*(5), 1792–1797.

Edwards, A. W. F. (1970). Estimation of the Branch Points of a Branching Diffusion Process. *Journal of the Royal Statistical Society. Series B (Methodological) 32*(2), 155–174.

Einmahl, J. H. J., A. Kiriliouk, and J. Segers (2018, Jun). A continuous updating weighted least squares estimator of tail dependence in high dimensions. *Extremes 21*(2), 205–233.

Evans, S. N. and F. A. Matsen (2012). The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(3), 569–592.

Felsenstein, J. (1981). Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. *Journal of Molecular Evolution 17*(6), 368–376.

Felsenstein, J. (2004). *Inferring phylogenies*, Volume 2. Sinauer associates Sunderland, MA.

Feragen, A., P. Lo, M. de Bruijne, M. Nielsen, and F. Lauze (2013, Aug). Toward a Theory of Statistical Tree-Shape Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*(8), 2008–2021.

Feragen, A., M. Owen, J. Petersen, M. M. W. Wille, L. H. Thomsen, A. Dirksen, and M. de Bruijne (2013). Tree-Space Statistics and Approximations for Large-Scale Analysis of Anatomical Trees. In J. C. Gee, S. Joshi, K. M. Pohl, W. M. Wells, and L. Zöllei (Eds.), *Information Processing in Medical Imaging*, Berlin, Heidelberg, pp. 74–85. Springer Berlin Heidelberg.

Feragen, A., J. Petersen, M. Owen, P. Lo, L. H. Thomsen, M. M. W. Wille, A. Dirksen, and M. de Bruijne (2015, June). Geodesic Atlas-Based Labeling of Anatomical Trees: Application and Evaluation on Airways Extracted From CT. *IEEE Transactions on Medical Imaging 34*(6), 1212–1226.

Fitch, W. M. (1971). Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Biology 20*(4), 406–416.

Fitch, W. M. and E. Margoliash (1967). Construction of Phylogenetic Trees. *Science 155*(3760), 279–284.

Foulds, L. R. and R. L. Graham (1982). The Steiner Problem in Phylogeny is NP-Complete. *Advances in Applied Mathematics 3*(1), 43–49.

Gavryushkin, A. and A. J. Drummond (2016). The Space of Ultrametric Phylogenetic Trees. *Journal of Theoretical Biology 403*, 197–208.

Grenander, U. and M. I. Miller (1998). Computational anatomy: An emerging discipline. *Quarterly of applied mathematics 56*(4), 617–694.

Holmes, S. (2003). Statistics for Phylogenetic Trees. *Theoretical Population Biology 63*(1), 17–32.

Holmes, S. (2005). Statistical Approach to Tests Involving Phylogenies. *Mathematics of Evolution and Phylogeny*, 91–120.

Hotz, T., S. Huckemann, H. Le, J. S. Marron, J. C. Mattingly, E. Miller, J. Nolen, M. Owen, V. Patrangenaru, and S. Skwerer (2013). Sticky Central Limit Theorems on Open Books. *The Annals of Applied Probability 23*(6), 2238–2258.

Huckemann, S., J. Mattingly, E. Miller, and J. Nolen (2015). Sticky central limit theorems at isolated hyperbolic planar singularities. *Electron. J. Probab. 20*, 34 pp.

Jardine, C., N. Jardine, and R. Sibson (1967). The Structure and Construction of Taxonomic Hierarchies. *Mathematical Biosciences 1*(2), 173–179.

Juhl, D. D., D. M. Warme, P. Winter, and M. Zachariasen (2014). The GeoSteiner software package for computing Steiner trees in the plane: An updated computational study.

Kanari, L., P. Dłotko, M. Scolamiero, R. Levi, J. Shillcock, K. Hess, and H. Markram (2018, Jan). A Topological Representation of Branching Neuronal Morphologies. *Neuroinformatics 16*(1), 3–13.

Kang, Q. and R. Yoshida (2018). Estimating Tropical Principal Components Using Metropolis Hastings Algorithm. In *International Congress on Mathematical Software*, pp. 272–279. Springer.

Kendall, D. G. (1989). A Survey of the Statistical Theory of Shape. *Statistical Science 4*(2), 87–99.

Kingman, J. F. C. (2000). Origins of the Coalescent: 1974-1982. *Genetics 156*(4), 1461–1463.

Knowles, L. (2009). Statistical Phylogeography. *Annual Review of Ecology, Evolution, and Systematics 40*, 593–612.

Kobayashi, K. and H. P. Wynn (2019, Feb). Empirical geodesic graphs and CAT(k) metrics for data analysis. *Statistics and Computing*.

Kupczok, A., A. von Haeseler, and S. Klaere (2008). An Exact Algorithm for the Geodesic Distance Between Phylogenetic Trees. *J. Comput. Biol. 15*(6), 577–591.

Lin, B., B. Sturmfels, X. Tang, and R. Yoshida (2017). Convexity in Tree Spaces. *SIAM Journal on Discrete Mathematics 31*(3), 2015–2038.

Lin, B. and N. M. Tran (2019). Two-player incentive compatible outcome functions are affine maximizers. *Linear Algebra and its Applications 578*, 133–152.

Lin, B. and R. Yoshida (2018). Tropical Fermat–Weber Points. *SIAM Journal on Discrete Mathematics*. To appear. Available at arXiv:1604.04674.

Liu, S., K. Ji, J. Chen, D. Tai, W. Jiang, G. Hou, J. Chen, J. Li, and B. Huang (2009, 03). Panorama Phylogenetic Diversity and Distribution of Type A Influenza Virus. *PLOS ONE 4*(3), 1–20.

Long, C. and S. Sullivant (2015). Identifiability of 3-class Jukes–Cantor mixtures. *Advances in Applied Mathematics 64*, 89–110.

Maclagan, D. and B. Sturmfels (2015). *Introduction to Tropical Geometry (Graduate Studies in Mathematics)*. American Mathematical Society.

Manon, C. (2011). Dissimilarity Maps on Trees and the Representation Theory of $\mathrm{SL}_m(\mathbb{C})$. *Journal of Algebraic Combinatorics 33*(2), 199–213.

Miller, E., M. Owen, and J. S. Provan (2015). Polyhedral computational geometry for averaging metric phylogenetic trees. *Advances in Applied Mathematics 68*, 51–91.

Miolane, N. and X. Pennec (2015). Biased Estimators on Quotient Spaces. In F. Nielsen and F. Barbaresco (Eds.), *Geometric Science of Information*, Cham, pp. 130–139. Springer International Publishing.

Munch, E. and A. Stefanou (2019). The ℓ^∞-Cophenetic Metric for Phylogenetic Trees as an Interleaving Distance. In *Research in Data Science*, pp. 109–127. Springer.

Nye, T. M. W. (2011). Principal Components Analysis in the Space of Phylogenetic Trees. *The Annals of Statistics 39*(5), 2716–2739.

Nye, T. M. W. (2014, March). An Algorithm for Constructing Principal Geodesics in Phylogenetic Treespace. *IEEE/ACM Trans. Comput. Biol. Bioinformatics 11*(2), 304–315.

Nye, T. M. W., X. Tang, G. Weyenberg, and R. Yoshida (2017). Principal Component Analysis and the Locus of the Fréchet Mean in the Space of Phylogenetic Trees. *Biometrika 104*(4), 901–922.

Ohta, S.-I. (2012). Barycenters in Alexandrov Spaces of Curvature Bounded Below. *Advances in Geometry 14*(4).

Owen, M. and J. S. Provan (2011). A Fast Algorithm for Computing Geodesic Distances in Tree Space. *IEEE/ACM Trans. Comput. Biol. Bioinformatics 8*(1), 2–13.

Pachter, L. and B. Sturmfels (2004). Tropical geometry of statistical models. *Proceedings of the National Academy of Sciences 101*(46), 16132–16137.

Pachter, L. and B. Sturmfels (2005). *Algebraic Statistics for Computational Biology*, Volume 13. Cambridge University Press.

Parthasarathy, K. (1967). Probability and Mathematical Statistics: A Series of Monographs and Textbooks. In *Probability Measures on Metric Spaces*, Probability and Mathematical Statistics: A Series of Monographs and Textbooks, pp. ii. Academic Press.

Rannala, B. and Z. Yang (1996). Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference. *Journal of Molecular Evolution 43*(3), 304–311.

Rhodes, J. A. and S. Sullivant (2012). Identifiability of large phylogenetic mixture models. *Bulletin of mathematical biology 74*(1), 212–231.

Robinson, D. F. and L. R. Foulds (1981). Comparison of Phylogenetic Trees. *Mathematical Biosciences 53*(1), 131–147.

Roch, S. and A. Sly (2017). Phase Transition in the Sample Complexity of Likelihood-Based Phylogeny Inference. *Probability Theory and Related Fields 169*(1), 3–62.

Rosenberg, N. A. (2003). The Shapes of Neutral Gene Genealogies in Two Species: Probabilities of Monophyly, Paraphyly, and Polyphyly in a Coalescent Model. *Evolution 57*, 1465–1477.

Saitou, N. and M. Nei (1987). The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution 4*(4), 406–425.

Schardl, C. L., D. K. Howe, G. Weyenberg, P. M. Huggins, and R. Yoshida (2014, 04). KDETREES: Non-parametric estimation of phylogenetic tree distributions. *Bioinformatics 30*(16), 2280–2287.

Schröder, E. (1870). Vier kombinatorische Probleme. *Z. Math. Phys 15*, 361–376.

Schwarz, H. A. (1890). *Ueber ein die Flächen kleinsten Flächeninhalts betreffendes Problem der Variationsrechnung*, pp. 223–269. Berlin, Heidelberg: Springer Berlin Heidelberg.

Skwerer, S., S. Provan, and J. Marron (2018). Relative Optimality Conditions and Algorithms for Treespace Fréchet Means. *SIAM Journal on Optimization 28*(2), 959–988.

Sokal, R. R. and C. D. Michener (1958). A Statistical Method for Evaluating Systematic Relationships. *University of Kansas Science Bulletin 38*, 1409–1438.

Sommerfeld, M. and A. Munk (2018). Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(1), 219–238.

Speyer, D. and B. Sturmfels (2004). The Tropical Grassmannian. *Advances in Geometry 4*(3).

Speyer, D. and B. Sturmfels (2009). Tropical mathematics. *Mathematics Magazine 82*(3), 163–173.

Steel, M. and D. Penny (1993). Distributions of Tree Comparison Metrics: Some New Results. *Syst. Biol. 42*(2), 126–141.

Sturm, K.-T. (2003). Probability measures on metric spaces of nonpositive curvature. *Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces: Lecture Notes from a Quarter Program on Heat Kernels, Random Walks, and Analysis on Manifolds and Graphs: April 16-July 13, 2002, Emile Borel Centre of the Henri Poincaré Institute, Paris, France 338*, 357.

Sturmfels, B. and S. Sullivant (2005). Toric ideals of phylogenetic invariants. *Journal of Computational Biology 12*(2), 204–228.

Sullivant, S. (2018). *Algebraic Statistics*, Volume 194. American Mathematical Soc.

Takatsu, A. (2011, 12). Wasserstein geometry of Gaussian measures. *Osaka J. Math. 48*(4), 1005–1026.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences 17*(2), 57–86.

Taylor, J. C. (2012). *An Introduction to Measure and Probability*. Springer Science & Business Media.

Tian, Y. and L. Kubatko (2014). Gene Tree Rooting Methods Give Distributions that Mimic the Coalescent Process. *Molecular Phylogenetics and Evolution 70*, 63–69.

Tran, N. M. (2018). Tropical Gaussians: A Brief Survey. *arXiv preprint arXiv:1808.10843*.

Weyenberg, G. and R. Yoshida (2016). Phylogenetic Tree Distances. In *Encyclopedia of Evolutionary Biology*, pp. 285–290. Elsevier.

WHO (2016). World Health Organization — Influenza. https://www.who.int/biologicals/vaccines/influenza/en/.

Willis, A. (2019). Confidence sets for phylogenetic trees. *Journal of the American Statistical Association 114*(525), 235–244.

Willis, A. and R. Bell (2018). Uncertainty in phylogenetic tree estimates. *Journal of Computational and Graphical Statistics 27*(3), 542–552.

Yan, B. C. and J. F. Yan (2013). A tree-like model for brain growth and structure. *J Biophys 2013*, 241612.

Yoshida, R., L. Zhang, and X. Zhang (2019, Feb). Tropical Principal Component Analysis and Its Application to Phylogenetics. *Bulletin of Mathematical Biology 81*(2), 568–597.

Zairis, S., H. Khiabanian, A. J. Blumberg, and R. Rabadan (2014). Moduli spaces of phylogenetic trees describing tumor evolutionary patterns. In *International Conference on Brain Informatics and Health*, pp. 528–539. Springer.

Zairis, S., H. Khiabanian, A. J. Blumberg, and R. Rabadán (2016). Genomic Data Analysis in Tree Spaces. *arXiv:1607.07503*.