

A SPACE OF PHYLOGENETIC NETWORKS

SATYAN L. DEVADOSS AND SAMANTHA PETTI

ABSTRACT. A classic problem in computational biology is constructing a phylogenetic tree given a set of distances between n species. In most cases, a tree structure is too constraining. We consider a circular split network, a generalization of a tree in which multiple parallel edges signify divergence. A geometric space of such networks is introduced, forming a natural extension of the work by Billera, Holmes, and Vogtmann on tree space. We explore properties of this space, and show a natural embedding of the compactification of the real moduli space of curves within it.

1. INTRODUCTION

A classical problem in computational biology is the construction of a phylogenetic tree from a sequence alignment of n species. The main tool used to build a phylogenetic tree from this data involves computing the maximum likelihood estimate (MLE) for each of the $(2n - 5)!!$ possible trees with n leaves, where each leaf corresponds to a given species. Such a procedure is quite difficult and requires examining an exponential number of trees. One method to circumvent this problem is distance based: One can construct a distance between two species (such as Hamming distance) which records the proportion of characters where the two species differ (based on certain genetic characteristics). Such a record can be encoded by an $n \times n$ real symmetric, nonnegative matrix called a *dissimilarity matrix*. The phylogenetic problem is then to reconstruct a weighted tree (on edges) that represents this matrix.

In order to analyze tree-like data, it is necessary not just to understand individual tree structures, but also the relationships between them. Billera, Holmes, and Vogtmann laid the foundation for this process by constructing a space BHV_n of such metric trees with nonpositive curvature, making it useful for geometric methods, such as the calculation of geodesics and centroids [15].

Unfortunately, most dissimilarity matrices are not tree metrics, as data from sequence alignments often fails due to hybridization, horizontal gene transfer, recombination, or gene duplication [11]. However, under a weaker (Kalmanson) condition, one obtains a *circular split network* instead of a tree as output, a generalization of a tree in which multiple

2010 *Mathematics Subject Classification.* 52B11, 14H10, 92B10, 05E45.

Key words and phrases. phylogenetics, split networks, associahedron, moduli space.

parallel edges signify divergence. In this work, we construct a geometric space CSN_n for such networks, similar to BHV_n , and explore properties of this space. Interestingly, this space appears in the work of Hacking, Keel, and Tevelev [10, Section 8] based on the real algebraic geometry of del Pezzo surfaces and the root system D_n . Devadoss and Morava [6] constructed a smooth blowup

$$\text{real}(\overline{\mathcal{M}}_{0,n}) \longrightarrow \text{trop}(\overline{\mathcal{M}}_{0,n}) \simeq \text{BHV}_n$$

between the real points and the tropicalization [19] of the compactified moduli space of stable genus zero algebraic curves marked with distinct smooth points. In this paper, we show an embedding

$$\text{real}(\overline{\mathcal{M}}_{0,n}) \hookrightarrow \text{CSN}_n,$$

in Theorem 12 which endows this real moduli space with an inherent metric. We claim that CSN_n seems to be a canonical space of study, with natural embeddings of both BHV_n and $\overline{\mathcal{M}}_{0,n}(\mathbb{R})$ within its structure.

Section 2 provides foundational definitions of trees, splits, and dissimilarity matrices. The construction of the space of split networks, along with understanding its properties, is given in Section 3 and 4. The associahedron polytope and the real moduli space of curves is introduced in Section 5, along with the embedding map provided in Section 6.

Acknowledgments. We would like to thank Jack Morava, Lior Pachter, and Jim Stasheff for their encouragement, along with helpful conversations with Sean Keel and Megan Owen. Devadoss was partially supported by a John Templeton Foundation grant 51894. We also thank Williams College, where this work was mostly completed during summer 2015.

2. SPLITS, TREES, AND NETWORKS

2.1. We begin with some foundational definitions; for a more thorough introduction to phylogenetic trees and networks, we refer the reader to [11, 13, 18]. Throughout the paper, X denotes the finite set of n distinct species on which metrics are defined.

Definition. A *split* $S = \{A, B\}$ is a partition of X into two nonempty sets. A split is *trivial* if one set of the partition has cardinality one. A set of splits is called a *split system*.

Definition. A split system \mathcal{S} is *pairwise compatible* if for every pair of splits $S_1 = \{A_1, B_1\}$ and $S_2 = \{A_2, B_2\}$ in \mathcal{S} , at least one of the following is empty:

$$A_1 \cap A_2, \quad A_1 \cap B_2, \quad A_2 \cap B_1, \quad B_1 \cap B_2.$$

There is a canonical graph associated with a split system, called the *Buneman graph* [18]. When the split system is pairwise compatible, this graph is a tree, where the vertices are

the species X , and the edges are the splits. Figure 1 displays corresponding tree for each set of pairwise compatible splits, where each edge visually represents a split. Notice that the order in which splits are introduced is independent of the associated Buneman tree.

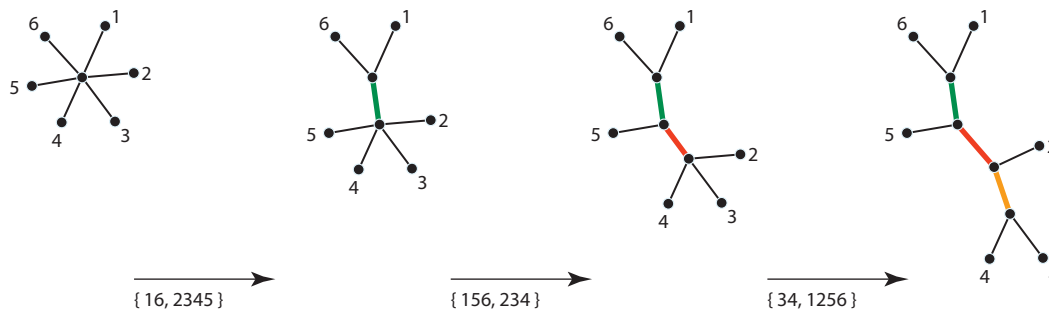


FIGURE 1. Pairwise compatible splits and the associated Buneman trees.

Thus far, we have been considering combinatorial and topological properties. However, biological data often comes in the form of distances between the species of X rather than in terms of splits, which can be encoded by an $n \times n$ real symmetric, nonnegative map

$$\delta : X \times X \longrightarrow \mathbb{R}$$

called a *dissimilarity matrix*. The phylogenetic problem is then to reconstruct an edge-weighted tree where $\delta(i, j)$ is the additive distance between species i and j , provided such a tree exists. The following gives necessary and sufficient conditions for such an existence.

Theorem 1. [18, Chapter 7] *The dissimilarity matrix δ realizes a weighted tree if and only if it satisfies the four-point condition: if for every four elements $i, j, k, l \in X$, two of the three terms below are greater than or equal to the third:*

$$\delta(i, j) + \delta(k, l), \quad \delta(i, l) + \delta(j, k), \quad \delta(i, k) + \delta(j, l).$$

Due to this result, a matrix satisfying the four-point condition is called a *tree metric*. The *neighbor-joining* algorithm [16] explicitly constructs the weighted tree from a dissimilarity matrix satisfying the four-point condition, viewed as a map between such matrices to pairwise compatible split systems.

2.2. Most data from sequence alignments often fails to satisfy the four-point condition due to hybridization, horizontal gene transfer, recombination, or gene duplication. One can relax conditions to obtain a *split network*, a generalization of a tree where each split is represented by a set of parallel edges that disconnects the graph according to the partition. Figure 2 displays an example of the construction of a split network, viewed

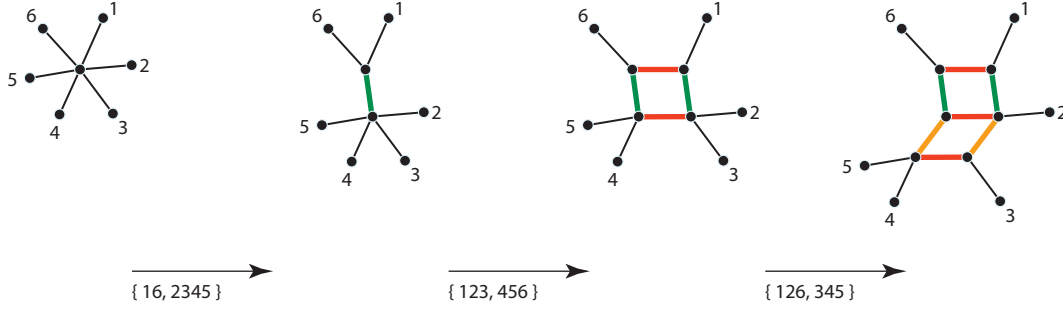


FIGURE 2. Splits and the associated split network.

as “pulling” each additional split in a different direction (and shown using distinct colors). Geometrically, the lengths of each parallel set of edges are identical.

The *convex hull* algorithm can be used to draw any split system. It works by increasing the dimensionality to add splits to a growing split network [11]. In general, for a system with k splits, the visualization of the network leads to the 1-skeleton of the k -cube.

Although any split system can be represented as a split network, not all are planar in nature. For instance, adding the split $\{14, 2356\}$ to the network of Figure 2 creates a network in 3D, the additional dimension separating $\{14\}$ from $\{2356\}$. To preserve planarity, the crucial concept needed is that of a *circular ordering* for the species set X .

Definition. A split system \mathcal{S} is *circular* with respect to some cyclic ordering $\pi = \{x_1, \dots, x_n\}$ of X if every split of \mathcal{S} is of the form $\{\{x_{i+1}, \dots, x_j\}, \{x_{j+1}, \dots, x_i\}\}$, where $x_{n+1} = x_1$.

The following establishes necessary and sufficient conditions for determining when a dissimilarity map can be realized as an edge-weighted circular split network, analogous to the four-point condition for trees of Theorem 1 above.

Theorem 2. [1] *The dissimilarity matrix δ realizes a weighted circular split system if and only if it satisfies the Kalmanson condition with respect to some circular ordering π : if for every $i < j < k < l$ in the circular ordering π , both inequalities hold:*

$$\delta(x_i, x_j) + \delta(x_k, x_l) \leq \delta(x_i, x_k) + \delta(x_j, x_l)$$

$$\delta(x_i, x_l) + \delta(x_j, x_k) \leq \delta(x_i, x_k) + \delta(x_j, x_l).$$

Due to this result, a matrix satisfying the Kalmanson condition is called a *circular decomposable metric*. Such metrics can be realized by a circular split network in which the shortest path between any two leaves contains precisely one edge from each split that separates the two leaves. Thus, the length of this path is the distance prescribed by the metric. Given a dissimilarity matrix satisfying the Kalmanson condition, the *neighbor-net*

algorithm [3] produces a circular ordering and a weighted split network compatible with that ordering.

2.3. Unlike trees, the visualization of circular split networks is not straightforward [8]. The *circular network* algorithm creates a planar split network, beginning with a star graph, and at each iteration including an additional split. Unfortunately, the structure of resultant split network depends on the order in which these splits are added, as shown by Figure 3. The top row of the figure shows the star graph with three splits $\{123, 456\}$, $\{126, 345\}$, $\{156, 234\}$ imposed on it; the bottom row shows the same three splits applied in an alternate ordering, resulting in a different diagram.

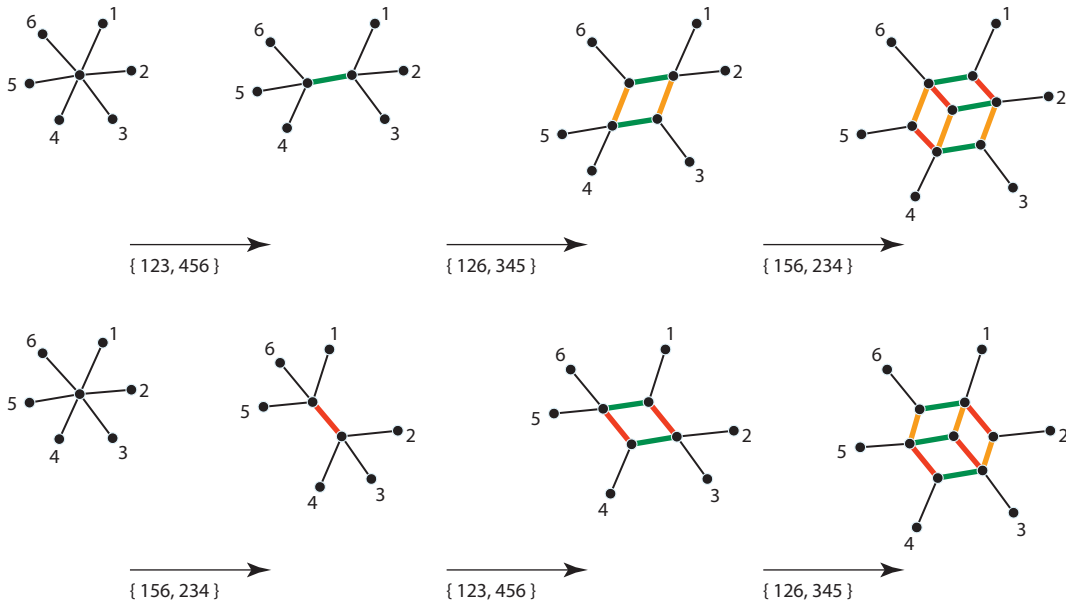


FIGURE 3. The circular split network is not commutative with respect to the order of the splits.

In order to avoid issues with commutativity of split order, we use the *dual polygonal representation* for the duration of the paper. Given a circular split system with some circular ordering π of the species, consider a regular n -gon, with the edges cyclically labeled with π . For each split, draw a diagonal partitioning the appropriate edges. Figure 4 shows the relationship between the traditionally drawn split network and its corresponding dual polygonal representations. For weighted networks, lengths of (parallel) edges are now represented by weights on the diagonals. Moreover, diagonals representing tree-like edges are *noncrossing* and network-like edges are *crossing*.

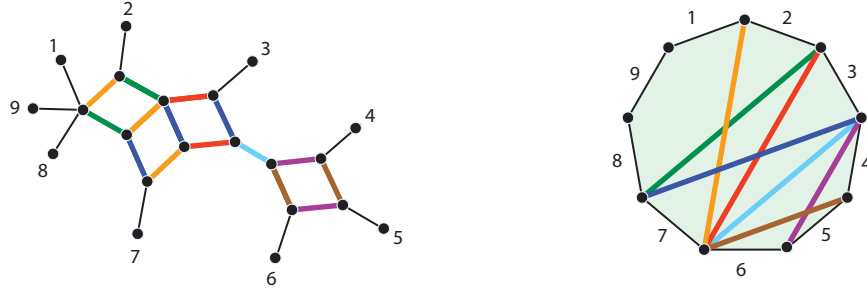


FIGURE 4. The dual polygon representations of a circular split network.

2.4. A labeled polygon with diagonals gives rise to a circular split system, where each diagonal is a split of the system. Conversely, a circular split system may be compatible with several different circular orderings, and there is a unique polygonal representation for each such ordering. A simple geometric operation allows the discovery of all such compatible orderings.

Definition. A *twist* along a noncrossing diagonal d of a labeled polygon P is obtained by ‘breaking’ P along d into two subpolygons, ‘reflecting’ one of these pieces, and ‘gluing’ them back (Figure 5).

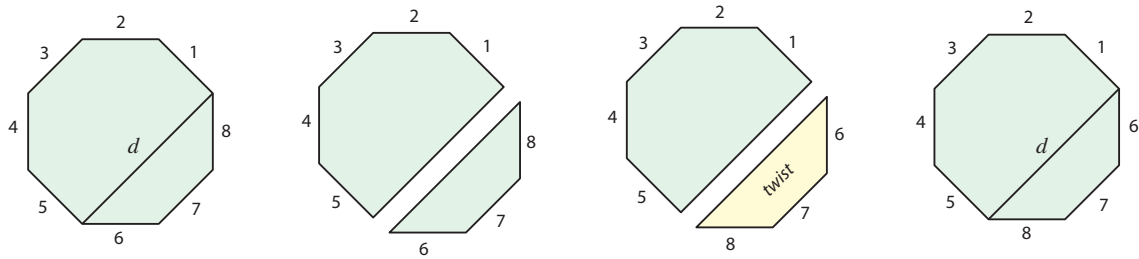


FIGURE 5. Twisting along diagonal d .

Theorem 3. *Two distinct circular orderings π_1 and π_2 are compatible with a circular split system \mathcal{S} if and only if beginning with the polygonal representation of \mathcal{S} with π_1 there exist a sequence of twists along noncrossing diagonals that results in the polygonal representation of \mathcal{S} with π_2 .*

Remark. We allow two types of noncrossing diagonals: (i) diagonal representing a split in \mathcal{S} that is noncrossing with other diagonals in the polygonal representation, or (ii) diagonal not representing a split in \mathcal{S} , but if drawn, would be noncrossing with other diagonals in the polygonal representation.

Proof. The backwards direction is immediate: twisting along a noncrossing diagonal does not change any split partitions. For the forward direction, begin with a representation of \mathcal{S} with π_1 . Let $f(x)$ be the label of the edge adjacent to x in the clockwise direction in π_2 .

- (1) If $f(x)$ is adjacent to x in π_1 in the clockwise direction, do nothing.
- (2) If $f(x)$ is adjacent to x in π_1 in the counterclockwise direction, and no diagonals exist between the two labels, draw a diagonal containing x and $f(x)$, and twist the side containing x and $f(x)$.
- (3) If $f(x)$ is adjacent to x in π_1 in the counterclockwise direction, and d is a diagonal landing between x and $f(x)$, twist along both sides of d .
- (4) Otherwise, draw diagonal d connecting the right endpoint of x with the right endpoint of $f(x)$, and twist the side containing $f(x)$. The compatibility of π_2 with \mathcal{S} guarantees d to be noncrossing.

Repeating this process with the resulting circular ordering, now letting x to be $f(x)$, produces the desired result after $n - 2$ iterations. \square

Example. The polygonal representation in Figure 6(a) is twisted along a noncrossing diagonal of type (ii) in (b) resulting in (c), an equivalent polygonal representation of the same split system. Similarly, a noncrossing diagonal of type (i) in (c) is twisted in (d) resulting in (e), another equivalent representation.

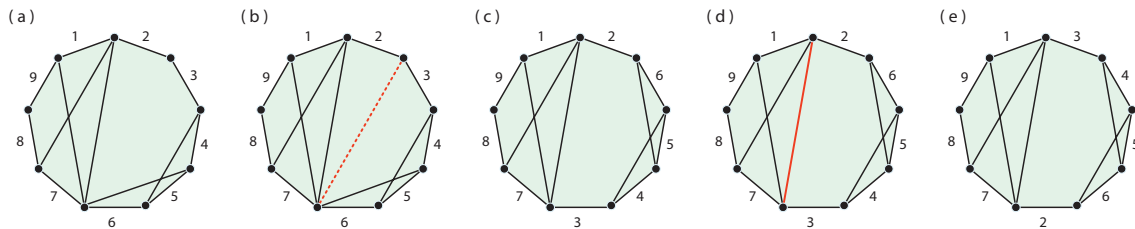


FIGURE 6. Equivalent polygonal representations due to twisting.

3. TREES AND NETWORKS

3.1. Billera, Holmes, and Vogtmann [2] constructed an elegant space BHV_n of isometry classes of metric trees with n labeled leaves.¹ Each such tree specifies a point in the orthant $[0, \infty)^{n-3}$, parametrized by the lengths of its internal edges, and thus defines coordinate charts for the space of such trees. The space BHV_n is assembled by gluing $(2n - 5)!!$ such orthants, one for each different binary tree on n leaves [7]. Two orthants of BHV_n share

¹Classically, this space is defined in terms of trees with n leaves and one root, whereas we consider unrooted trees with n leaves. Thus there is an index shift of one.

a wall if and only if their corresponding binary trees differ by a *rotation*, a move which collapses an interior edge of a binary tree, and then expands the resulting degree-four vertex into a different binary tree. Figure 7(a) shows BHV_4 consisting of three (orthant) rays glued at the origin, the degenerate tree with no internal edges.

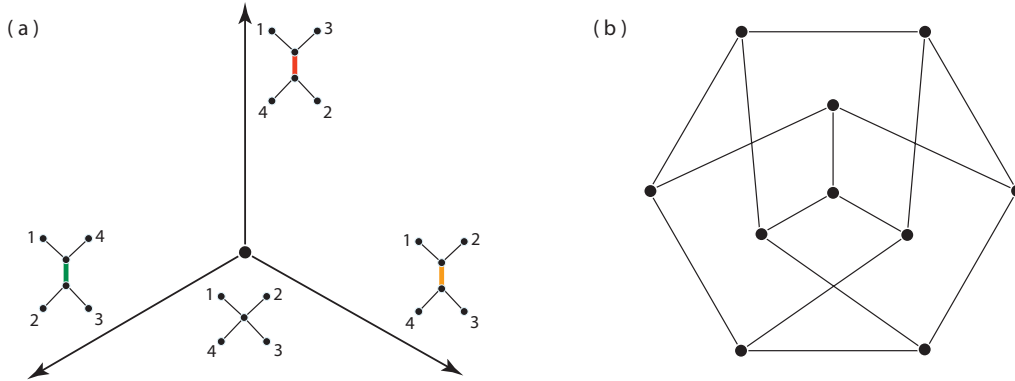


FIGURE 7. (a) Tree space BHV_4 and (b) the simplicial complex \mathcal{T}_5 .

Definition. Let \mathcal{T}_n be the subspace of BHV_n consisting of trees with internal edge lengths that sum to 1. It is a pure simplicial $(n - 3)$ -complex composed of $(2n - 3)!!$ chambers, with two adjacent chambers differing by a rotation of their underlying trees.

Indeed, BHV_n is a cone over this space (and thus contractible), where the cone-point is the degenerate tree with no internal edges. For example, BHV_5 consists of 15 quadrants $[0, \infty)^2$ glued together, and its subspace \mathcal{T}_5 is the Peterson graph with 15 edges, as displayed in Figure 7(b). Here, the 10 vertices correspond to binary trees with five leaves and two internal edges.

3.2. We now introduce a space CSN_n of isometry classes of metric circular split networks with n labeled leaves. Each network specifies a point in this network space, parametrized by the weights of its splits, or equivalently the lengths of its (parallel sets of) internal edges. The space is assembled by gluing $(n - 1)!/2$ orthants together, each of which corresponds to a unique circular ordering of the n species, up to rotation and reflection. Therefore, each orthant has dimension $n(n - 3)/2$, the maximal number of splits compatible with a given circular ordering, or the maximal number of diagonals on an n -gon. Orthants glue together along cells that represent split systems that are compatible to the orderings of their respective chambers, given by Theorem 3.

There is a natural coordinate system based on splits into which CSN_n embeds. Let

$$(3.1) \quad \delta := 2^{n-1} - n - 1,$$

enumerating networks with exactly one nontrivial split, obtained by considering half the number of partitions 2^n (compensating for double-counting) and subtracting the set of n trivial splits and 1, the empty set. Each circular split network is an element of \mathbb{R}^δ , defined by the set of its splits and parametrized by the lengths of its internal edges. However, a circular split network will have at most $n(n - 3)/2$ nonzero values in this coordinate system, based on the maximal number of diagonals on an n -gon. In summary, we have:

Proposition 4. *The simplicial fan CSN_n naturally embeds into \mathbb{R}^δ .*

Remark. Recently, Hacking, Keel, and Tevelev [10] have shown how to build this simplicial fan CSN_n canonically from the root system D_n based on real algebraic geometry of del Pezzo surfaces. Indeed, precise analogs for E_6, E_7, E_8 are also provided, based on work by Sekiguchi and Yoshida [17].

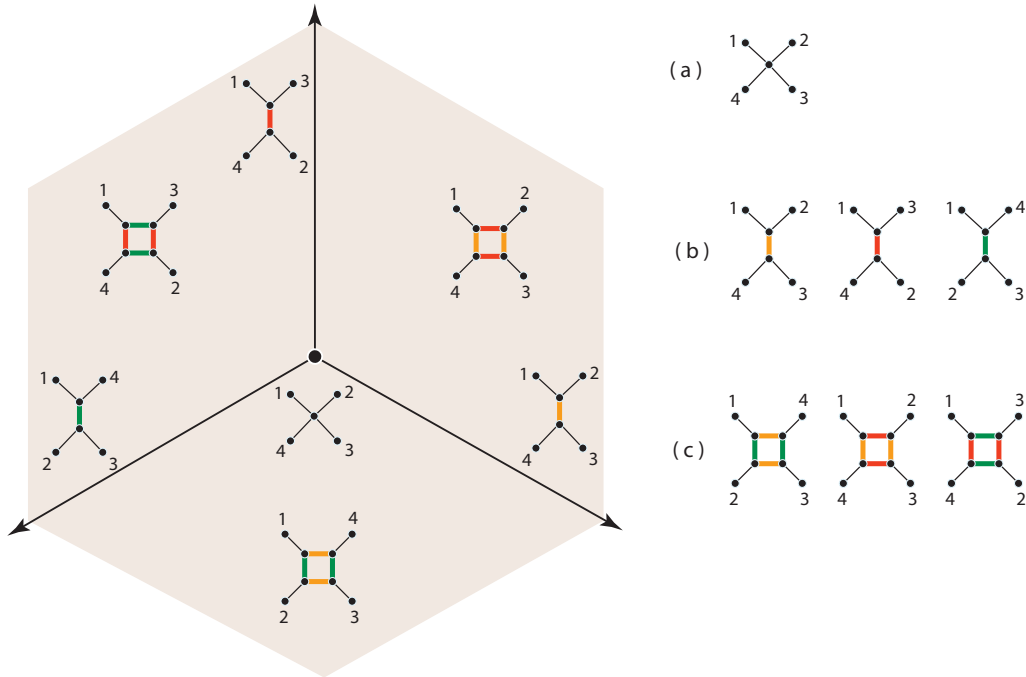


FIGURE 8. The space of circular split networks for $n = 4$.

Example. Figure 8 shows the case for CSN_4 , the space of metric circular split networks for four species. Three quadrants $\mathbb{R}_{\geq 0}^2$ tile this space (c), corresponding to the distinct circular orderings of the four labels. These quadrants glue along the three boundary rays (b), where all meet at the origin, the degenerate network with no internal edges (a). Thus CSN_4 is homeomorphic to the plane \mathbb{R}^2 . Note that BHV_4 from Figure 7(a) is a natural subspace of CSN_4 .

3.3. Indeed, BHV_n is simply the subspace of CSN_n restricted to *pairwise compatible* split systems. Another natural subspace of CSN_n is the *link of the origin* \mathfrak{S}_n , the union of the set of points in each orthant with internal edge lengths of networks that sum to 1. Since the set of such points in a single orthant forms a simplex, the following is immediate:

Proposition 5. *Network space \mathfrak{S}_n is a connected simplicial complex of dimension $(n(n-3)/2) - 1$, with one k -simplex for every labeled n -gon with $k+1$ diagonals.*

Figure 9 shows $\mathfrak{S}_4 \subset \text{CSN}_4$, the 1-dimensional simplicial complex formed by gluing three edges together, forming a triangle, parametrizing circular networks whose internal edge lengths sum to 1.

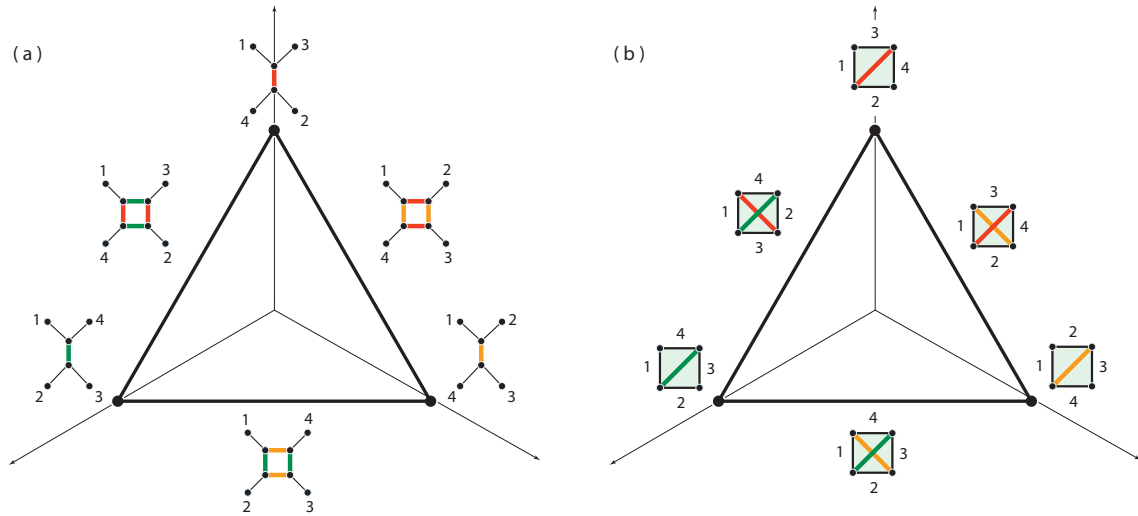


FIGURE 9. The simplicial complex (a) \mathfrak{S}_4 and (b) its polygonal representation.

It was shown in [2] that BHV_n was a $\text{CAT}(0)$ space, ensuring that any two points had a unique geodesic between them. This is not the case for networks, however, due to the underlying geometry of \mathfrak{S}_n .

Proposition 6. *The space CSN_n of circular split networks is not $\text{CAT}(0)$.*

Proof. Subdivide each orthant of CSN_n into unit cubes having integral vertices, making CSN_n into a cubical complex. By a theorem of Gromov [9], a cubical complex is $\text{CAT}(0)$ if and only if it is simply connected and the link at every vertex is a flag complex. But Figure 10 shows the link at the origin \mathfrak{S}_n having *empty triangles*, and therefore failing to be flag. Here, the three edges of the triangle exist but there is no 2-simplex (represented by a polygon with three diagonals) bounding this triangle. \square

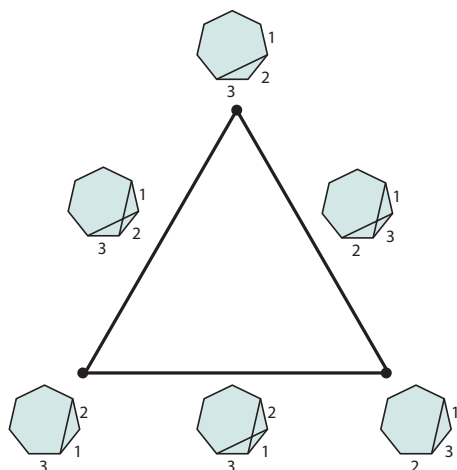


FIGURE 10. Empty triangles in \mathfrak{CSN}_n .

4. COMBINATORICS OF NETWORK SPACE

4.1. We explore some properties of the simplicial complex \mathfrak{S}_n . For $n = 4$, \mathfrak{S}_n is a triangle, with three vertices and three edges. Enumeration for larger values of n follows:

Theorem 7. *For $n > 4$, the network space \mathfrak{S}_n has $c = (n - 1)!/2$ simplicial chambers of dimension $d = n(n - 3)/2 - 1$. It has $c(d + 1)$ ridges, δ vertices and $\binom{\delta}{2}$ edges.*

Proof. The chambers are the different ways of labeling an n -gon with the maximal set of diagonals. Since there are $n(n - 3)/2$ diagonals on an n -gon, the number of ridges (codim 1 faces) corresponds to removing one of these diagonals from the maximal set. The vertices enumerate the set of networks with one split, obtained in Eq. (3.1) above. And since each pair of splits is contained in a circular split system, the enumeration of edges follows. \square

Corollary 8. *The 1-skeleton of \mathfrak{S}_n is the complete graph on δ vertices.*

Note that the number of chambers in which a k -cell of \mathfrak{S}_n resides depends on the structure of polygon corresponding to the cell, not the dimension of the cell itself. A k -cell may be part of only one chamber or it might be part of several chambers of \mathfrak{S}_n . Figure 11 (a) - (c) depict polygon corresponding to a 14-cell, 9-cell, and 10-cell, respectively, that are contained in exactly two distinct chambers of \mathfrak{S}_8 . Part (d) depicts another 10-cell belonging to a unique chamber of \mathfrak{S}_8 .

Theorem 9. *Two chambers of \mathfrak{S}_n can intersect along a face of at most dimension*

$$(n - 3)(n - 2)/2 - 1.$$

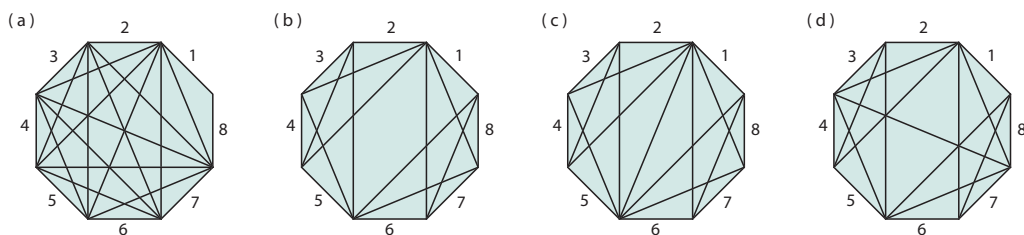


FIGURE 11. Cells (a), (b), (c) belong to two distinct chambers, whereas cell (d) to a unique chamber of \mathfrak{S}_8 .

Proof. By Theorem 3, it is necessary that there exists a possible noncrossing diagonal of the polygonal representation for two distinct circular ordering to be compatible. An n -gon can have at most $n(n-3)/2$ diagonals, and $n-3$ of them must be removed in order to make room for a noncrossing diagonal d . Indeed, there must be x vertices of one side of d and y vertices on the other such that $x+y = n-2$. It is necessary to remove xy diagonals, which is minimized when $x = n-3$ and $y = 1$, and the result follows. \square

4.2. The space \mathfrak{S}_5 is a 4-dimensional simplicial complex, composed of 12 chambers, each corresponding the unique circular orderings on five species labels. There are seven different types of cells, corresponding to split networks with different structures, as outlined in Figure 12: The rows denote dimension, polygon representation, network type, and enumeration in \mathfrak{S}_5 , respectively. There are two distinct 2-cells (orange, yellow), each bounded by two types of edges (red, blue), detailed in the left side of Figure 13.

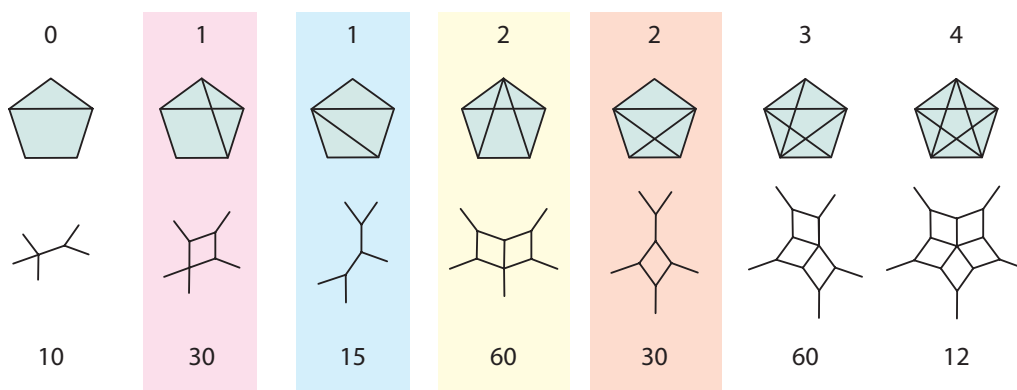


FIGURE 12. The stratification of \mathfrak{S}_5 by distinct cell types, where the number of such cells is enumerated in the last line.

The right side of Figure 13 illustrates a (4-simplex) chamber in \mathfrak{S}_5 . Of the ten triangles appearing as faces, five are orange and five are yellow; two of each color are highlighted in the illustration. Interestingly, the triangles of each color form a mobius strip within

the 4-simplex. By Theorem 9, two chambers of \mathfrak{S}_5 meet only along vertices, edges, and triangles, and not along any of the tetrahedra.

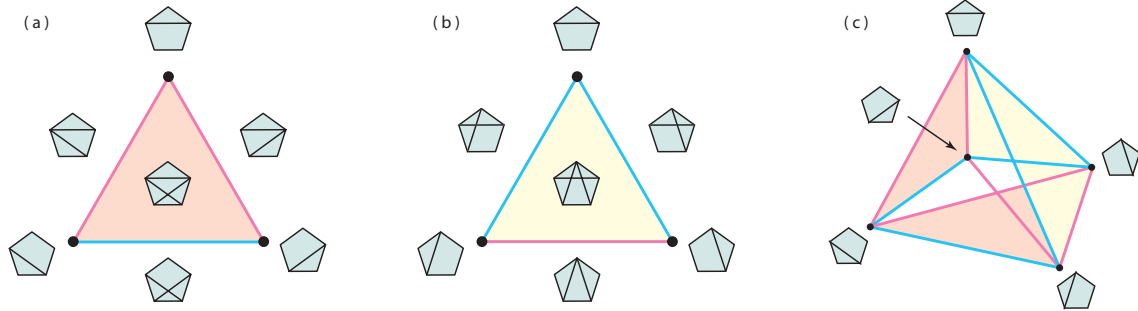


FIGURE 13. The two types of triangles along with a partial labeling of a chamber in \mathfrak{S}_5 .

The global structure of the 1-skeleton of \mathfrak{S}_5 is shown in Figure 14(a), with the coloring based on Figure 12. The red edges in (b) form the Peterson Graph, tree space \mathcal{T}_5 from Figure 7(b), and the blue edges in (c) form the skeleton of a rectified 5-cell.

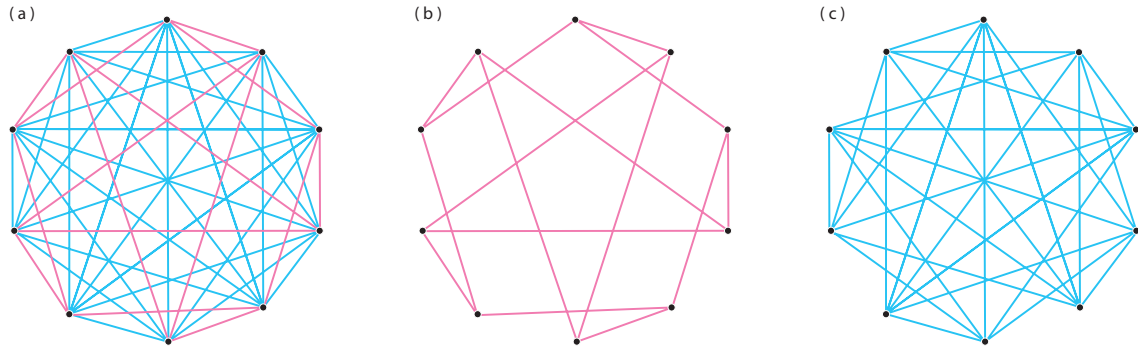


FIGURE 14. The 1-skeleton of \mathfrak{S}_5 , color-coded based on Figure 12.

5. MODULI SPACES

5.1. An elegant polytope captures the structure of the space of planar *rooted* trees:

Definition. The *associahedron* is a convex polytope of dimension $n - 2$ whose face poset is isomorphic to that of bracketings of n letters, ordered so $a \prec a'$ if a is obtained from a' by adding new brackets.

The associahedron was constructed independently by Haiman (unpublished) and Lee [12], though Stasheff had defined the underlying abstract object twenty years previously,

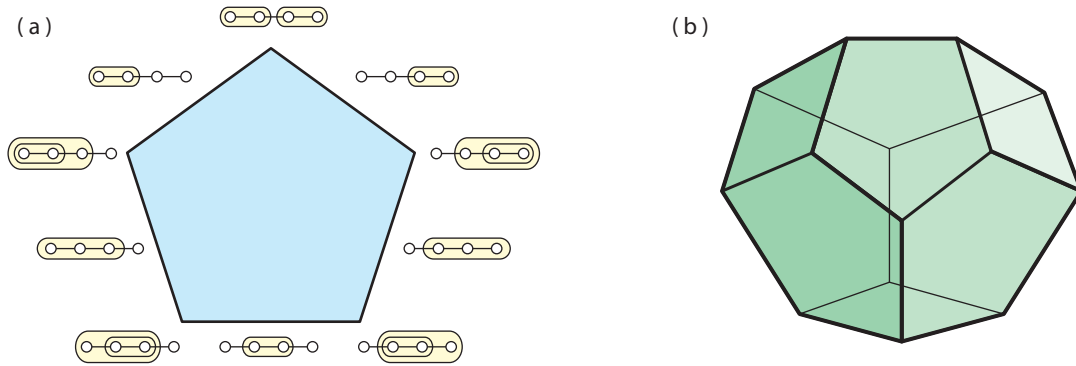


FIGURE 15. Associahedra K_4 and K_5 .

in his work on associativity in homotopy theory [21]. Figure 15(a) shows the 2D associahedron K_4 with a labeling of its faces, and (b) shows the 3D version K_5 . There are over a hundred combinatorial and geometric interpretations [20] of the Catalan number

$$(5.1) \quad C_{n-1} = \frac{1}{n} \binom{2n-2}{n-1},$$

which index the vertices of the associahedron K_n . Most important to us is the relationship between bracketings of n letters, rooted trees with n species, and polygons with noncrossing diagonals, shown in Figure 16. In particular, a codim k face of the associahedron K_{n-1} is associated to an (unlabeled) n -gon with k noncrossing diagonals.

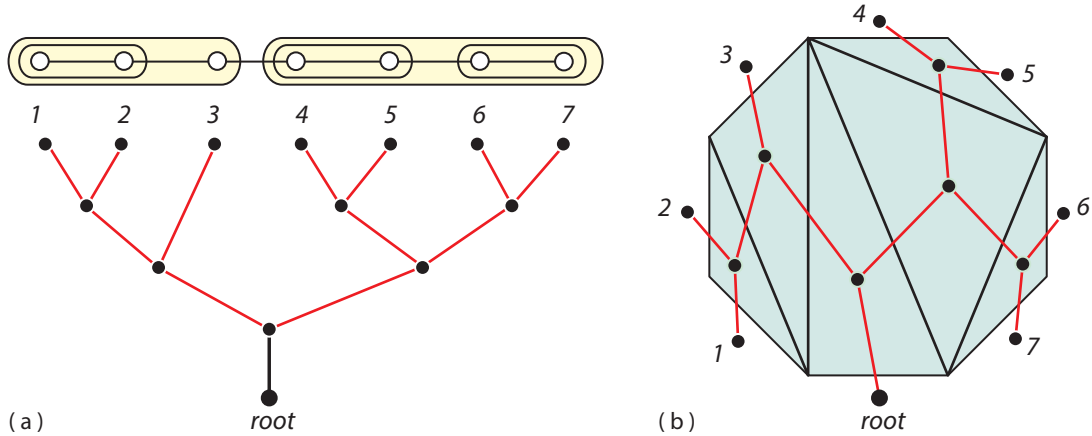


FIGURE 16. (a) Bracketings and planar trees and (b) polygons with diagonals.

5.2. The moduli problem for algebraic curves has been a central problem in mathematics since Riemann. In the 1970s it was solved over the integers \mathbb{Z} by Deligne, Mumford, Knudsen [14] and others, where a special case constructs a moduli space for *real* algebraic

curves of genus zero marked with distinct smooth points. That solution can be regarded as a good compactification $\overline{\mathcal{M}}_{0,n}(\mathbb{R})$ of the space

$$\mathcal{M}_{0,n}(\mathbb{R}) = \text{Config}^n(\mathbb{RP}^1)/\text{PGl}_2(\mathbb{R})$$

of $n + 1$ distinct particles on the real projective line.

Theorem 10. [5, Section 4] *The moduli space $\overline{\mathcal{M}}_{0,n}(\mathbb{R})$ is tiled by $(n - 1)!/2$ copies of K_{n-1} , one for each labeling of an n -gon, up to rotation and reflection. Two faces of associahedra, represented by labeled polygons with noncrossing diagonals P_1 and P_2 , are identified in $\overline{\mathcal{M}}_{0,n}(\mathbb{R})$ if twisting along certain diagonals of P_1 yields P_2 .*

The moduli space $\overline{\mathcal{M}}_{0,3}(\mathbb{R})$ is a point, and the manifold $\overline{\mathcal{M}}_{0,4}(\mathbb{R})$ is homeomorphic to a circle, with the *cross-ratio* serving as the homeomorphism; it is tiled by three K_3 line segments, glued together to form a triangle, shown in Figure 17.

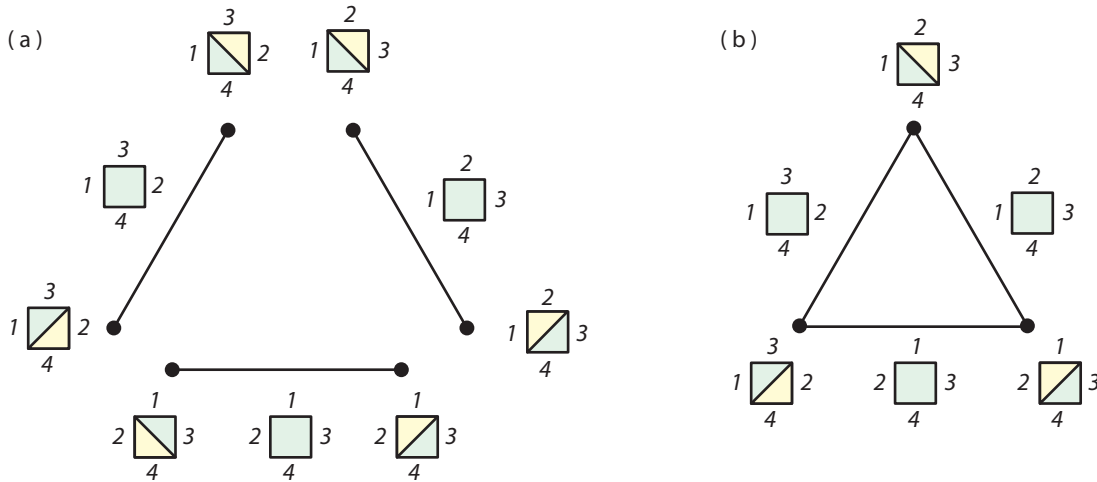


FIGURE 17. Gluing three K_3 using twists to form $\overline{\mathcal{M}}_{0,4}(\mathbb{R})$.

Example. An illustration of $\overline{\mathcal{M}}_{0,5}(\mathbb{R})$ appears in Figure 18(a), resulting in the connected sum of a torus with three real projective planes, tiled by 12 associahedra K_4 from Figure 15(a). Part (b) shows the example for $\overline{\mathcal{M}}_{0,6}(\mathbb{R})$, tiled by 60 copies of K_5 .

5.3. A natural embedding Φ of the moduli space $\overline{\mathcal{M}}_{0,n}(\mathbb{R})$ into the network space \mathfrak{S}_n is now provided. Throughout this section, we fix some labeling of the n -gon thereby choosing a chamber of $\overline{\mathcal{M}}_{0,n}(\mathbb{R})$ and \mathfrak{S}_n . The canonical coordinates for \mathfrak{S}_n is given in Proposition 11, one for each possible split of n species. For a chosen labeling however,

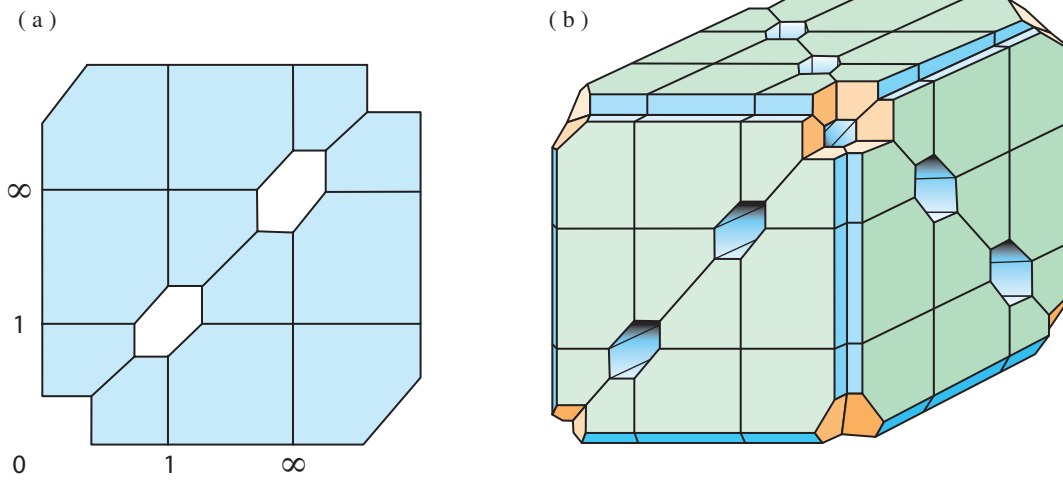


FIGURE 18. (a) $\overline{\mathcal{M}}_{0,5}(\mathbb{R})$ and (b) $\overline{\mathcal{M}}_{0,6}(\mathbb{R})$ as blowups of tori.

recall that at most $n(n-3)/2$ of these coordinates will have nonzero values, one for each diagonal of the n -gon.² Thus, the coordinate system in a chamber of \mathfrak{S}_n is

$$(5.2) \quad (x_1, x_2, \dots, x_{n(n-3)/2}),$$

a dimension x_i for each diagonal d_i of the n -gon.

Consider the set of vertices $V(K_{n-1})$ of the associahedron, each a triangulation with exactly $n-3$ diagonals. For each vertex v , assign the d -th coordinate of the map $\Phi(v)$ as

$$(5.3) \quad \Phi_d(v) := \begin{cases} 1/(n-3) & \text{if } d \text{ is a diagonal of } v \\ 0 & \text{otherwise.} \end{cases}$$

For a face f of the associahedron, let V_f be the subset of vertices of $V(K_{n-1})$ incident to f . Assign to the barycenter v_f of each face f the coordinate

$$(5.4) \quad \Phi(v_f) := \text{centroid } \{\Phi(v) \mid v \in V_f\}.$$

Remark. It follows that $\Phi(v)$ is an element of \mathfrak{S}_n , the sum of the coordinates being 1.

Example. Figure 19 illustrates this computation for several barycenters of the associahedron K_6 . The top row shows the nine polygons with one diagonal, each an axes in the coordinate system. The left column displays four different barycenters of the subdivision, along with their respective coordinates in this system.

²To be consistent with the language of associahedra, we refer to sets of splits as sets of diagonals throughout this section.

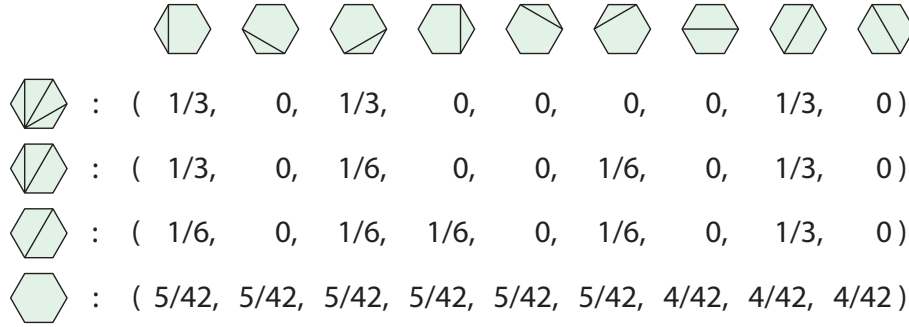


FIGURE 19. Coordinates for barycenters.

Remark. A flag F of the associahedron is a sequence $\{f_0, f_1, \dots, f_{n-3}\}$ of faces such that $f_0 \subset f_1 \subset \dots \subset f_{n-3}$. We can reinterpret this as a sequence of subsets of diagonals, where the set f_i contains exactly $n - 3 - i$ noncrossing diagonals.

Proposition 11. *For flag F of K_{n-1} , and Δ_F its associated simplex in its barycentric subdivision, the map*

$$\Phi(\Delta_F) := \text{convex hull } \{\Phi(v_{f_0}), \Phi(v_{f_1}), \dots, \Phi(v_{f_{n-3}})\}$$

is an embedding of Δ_F into a chamber of \mathfrak{S}_n , where v_{f_i} are the vertices of Δ_F , the barycenters given in Eq. (5.4). This map Φ extends to the union of all the flags of K_{n-1} , providing an embedding of the associahedron into a chamber of \mathfrak{S}_n .

The proof of this proposition is relegated to the next section. We can extend this result for all of $\overline{\mathcal{M}}_{0,n}(\mathbb{R})$.

Theorem 12. *The map Φ naturally extends to an embedding of $\overline{\mathcal{M}}_{0,n}(\mathbb{R})$ into \mathfrak{S}_n .*

Proof. We have defined the coordinates for one particular chamber (labeling of the n -gon) of \mathfrak{S}_n , as defined in Eq. (5.2). This coordinate system naturally extends to the canonical coordinates for \mathfrak{S}_n as given in Proposition 11, one for each possible split of n species, with at most $n(n - 3)/2$ of these coordinates having nonzero values. Note further that two associahedral chambers glue to form $\overline{\mathcal{M}}_{0,n}(\mathbb{R})$ by *twisting*, as given in Theorem 10, while their image in \mathfrak{S}_n under Φ glue accordingly, due to Theorem 3. □

Example. Figure 20 illustrates the embedding of $\overline{\mathcal{M}}_{0,4}(\mathbb{R})$ in \mathfrak{S}_4 , part (a) showing the polygonal labeling and (b) the coordinates of the vertices and the barycenters. Each associahedral edge sits inside a chamber of \mathfrak{S}_4 with coordinates defined by Proposition 11. The edges glue together as do the chambers of \mathfrak{S}_n , forming $\overline{\mathcal{M}}_{0,4}(\mathbb{R})$. Compare this with Figure 9(b).

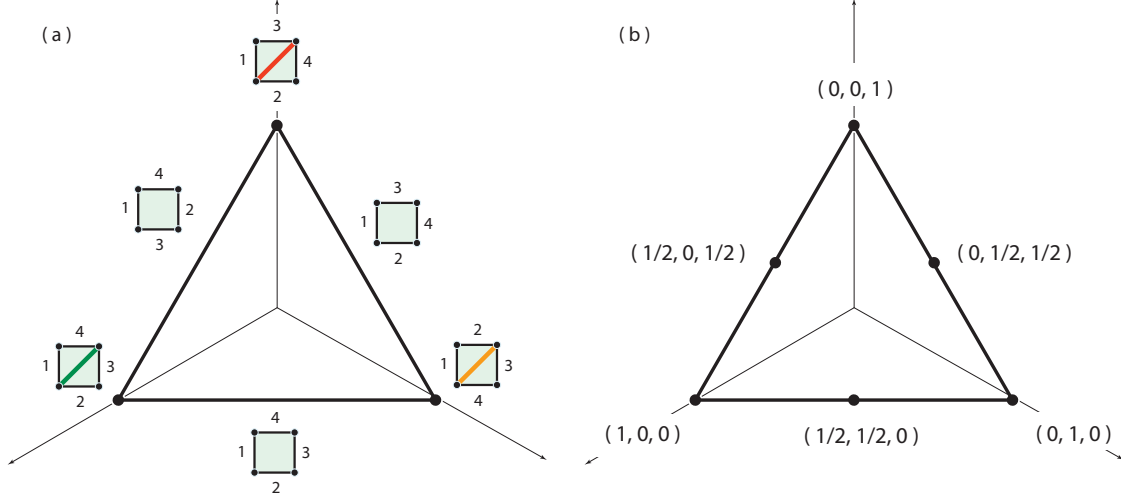


FIGURE 20. Embedding of $\overline{\mathcal{M}}_{0,4}(\mathbb{R})$ in \mathfrak{S}_4 along with coordinates.

Remark. It is interesting to wonder whether more can be said about the relationship between $\overline{\mathcal{M}}_{0,n}(\mathbb{R})$ and \mathfrak{S}_n . For example, how do the homotopy types of these two spaces compare? It was shown in [4] that $\overline{\mathcal{M}}_{0,n}(\mathbb{R})$ is *aspherical*, its homotopy properties encapsulated in its fundamental group, whereas little is known about \mathfrak{S}_n .

6. PROOF OF EMBEDDINGS

6.1. In this section, we prove Proposition 11 which states that the map Φ defined above is an embedding. We show Φ is injective and preserves dimension in Lemma 15 and Lemma 16, respectively, with embedding directly following. First, consider the following description of Φ based on the Catalan numbers C_n from Eq. (5.1).

Proposition 13. *Let f be a face of the associahedron with barycenter v_f , represented by a polygon P with noncrossing diagonals. If Φ is the map defined in Eq. (5.4), then the coordinate of $\Phi(v_f)$ associated to diagonal d is given by*

$$(6.1) \quad \Phi_d(v_f) = \begin{cases} \frac{1}{n-3} & \text{if } P \text{ contains } d \\ 0 & \text{if } d \text{ crosses } P \\ \frac{1}{n-3} \cdot \frac{C_{s-2} C_{t-2}}{C_{s+t-4}} & \text{otherwise} \end{cases}$$

where the two polygons in P on either side of d are an s -gon and t -gon.

Proof. Recall that for a diagonal d and a vertex v , the corresponding coordinate in the embedding $\Phi_d(v)$ is given in Eq. (5.3). Since $\Phi_d(v_f)$ is the average (centroid) of the set

$\{\Phi_d(v) \mid v \in V_f\}$, as shown in Eq. (5.4), we have

$$\begin{aligned}\Phi_d(v_f) &= \frac{1}{n-3} \left(\frac{\# \text{ vertices of } f \text{ that include diagonal } d}{\# \text{ vertices of } f} \right) \\ &= \frac{1}{n-3} \left(\frac{\# \text{ triangulations that include } P \text{ and } d}{\# \text{ triangulations that include } P} \right).\end{aligned}$$

The first two cases of Eq. (6.1) follow trivially. For the third case, $\Phi_d(v_f)$ depends on the way d sits inside P . Adding diagonal d divides an $(s+t-2)$ -sided subpolygon of P into an s -gon and a t -gon. Since the Catalan number C_{n-2} enumerates the triangulations of the n -gon, the result follows. \square

Lemma 14. *Let $S = \{f_0, f_1, \dots, f_m\}$ be a subflag of the associahedron K_n , and let x be a point in $\Phi(\Delta_S)$ such that*

$$x = a_0\Phi(v_{f_0}) + a_1\Phi(v_{f_1}) + \dots + a_m\Phi(v_{f_m}),$$

where $\sum a_i = 1$ and each $a_i \geq 0$. Then diagonal d is in f_k if and only if

$$(6.2) \quad x_d = \sum_{i=0}^{k-1} a_i \Phi_d(v_{f_i}) + \frac{1}{n-3} \cdot \left(1 - \sum_{i=0}^{k-1} a_i \right).$$

Proof. If diagonal $d \in f_k$, then $d \in f_i$, for all $i \geq k$. Proposition 13 then implies that $\Phi_d(v_{f_i}) = 1/(n-3)$, for all $i \geq k$, and the result follows. Conversely, Eq. (6.2) yields

$$\sum_{i=0}^{k-1} a_i \Phi_d(v_{f_i}) = \frac{1}{n-3} \cdot \sum_{i=k}^m a_i.$$

Since each $\Phi_d(v) \leq 1/(n-3)$, it follows by Proposition 13 that $\Phi_d(v_{f_i}) = 1/(n-3)$, for all $i \geq k$, and the result follows. \square

Lemma 15. *Let x be a point in the image of Φ . Then there exists a unique subflag $\{f_0, f_1, \dots, f_m\}$ and coefficients a_0, a_1, \dots, a_m , where $a_i > 0$ and $\sum a_i = 1$, such that*

$$x = a_0\Phi(v_{f_0}) + a_1\Phi(v_{f_1}) + \dots + a_m\Phi(v_{f_m}).$$

Proof. We present an algorithm that uniquely determines a_i and f_i from the coordinates of x through an iterative procedure. As the base case, by Lemma 14, f_0 is the set of diagonals d such that $x_d = 1/(n-3)$. The procedure assumes that for some j , we know f_0, f_1, \dots, f_j and a_0, a_1, \dots, a_{j-1} , and determines a_j by either

- (A) concluding f_j is the maximal set of diagonals in the subflag, or
- (B) identifying the next element f_{j+1} in the subflag.

In each iteration, every subpolygon³ of the polygonal representation of f_j is tested to obtain information that is used to determine a_j and f_{j+1} , if applicable.

Let P be the polygonal representation of f_j , and P_r be a subpolygon of P with r sides. Let D the set of diagonals in P_r that partition it into a 3-gon and an $(r-1)$ -gon. The following relates diagonals d that can be drawn within P_r to coordinate values x_d .

Claim. Diagonals in P_r are not in any set of diagonals of the subflag if and only if

$$(6.3) \quad x_d = \sum_{i=0}^{j-1} a_i \Phi_d(v_{f_i}) + \frac{1}{n-3} \cdot \frac{C_{r-1}}{C_r} \cdot \left(1 - \sum_{i=0}^{j-1} a_i\right).$$

for all diagonals d of D :

Proof. If no diagonal in P_r is contained in any diagonals of the subflag, then P contains the region P_r , for all $i \geq j$. Therefore, by Proposition 13,

$$\Phi_d(v_{f_i}) = \frac{1}{n-3} \cdot \frac{C_{r-1}}{C_r},$$

for d in D and $i \geq j$, and Eq. (6.3) follows. Conversely, suppose there is a diagonal d_* in P_r that is contained in diagonals of the subflag. Let f_k be the face containing the minimal such set of diagonals. For a diagonal d in D that intersects d_* , $\Phi_d(v_{f_i}) = 0$ for all $i \geq k$ by Proposition 13. Similarly, since P_r is in face f_i for all $j \leq i \leq k-1$, then

$$(6.4) \quad \Phi_d(v_{f_i}) = \frac{1}{n-3} \cdot \frac{C_{r-1}}{C_r}.$$

But because

$$\begin{aligned} x_d &= \sum_{i=0}^{j-1} a_i \Phi_d(p_i) + \frac{1}{n-3} \cdot \frac{C_{r-1}}{C_r} \cdot \sum_{i=j}^{k-1} a_i + 0 \cdot \sum_{i=k}^m a_i \\ &< \sum_{i=0}^{j-1} a_i \Phi_d(p_i) + \frac{1}{n-3} \cdot \frac{C_{r-1}}{C_r} \cdot \sum_{i=j}^m a_i, \end{aligned}$$

Eq. (6.3) does not hold for d , justifying the claim. \square

If for *every* subpolygon of P , diagonals in P_r are not in any set of diagonals of the subflag, then f_j is the maximal split system in the subflag. It follows that $a_j = 1 - a_0 - a_1 - \dots - a_{j-1}$, resulting in case (A) and ending the algorithm. Otherwise, there exists a diagonal in some subpolygon P_r of f_j contained in some set of diagonals of the subflag, resulting in case (B). For each such subpolygon, we compute $a_j + a_{j+1} + \dots + a_{k-1}$, where k is such that f_k is the minimal set of diagonals containing a diagonal from P_r , and we use this information to determine the diagonals of f_{j+1} and a_j .

³We say S is a subpolygon of a polygonal representation P if the boundary of S consists of edges and diagonals of P and no diagonals of the polygonal representation are drawn in the interior of S .

First we show how to compute $a_j + a_{j+1} + \dots + a_{k-1}$, where k is such that f_k is the minimal set of diagonals containing a diagonal from P_r . Let d_* be a diagonal in f_k and P_r . For all diagonals d in T , since P_r is in f_i , for all $j \leq i \leq k-1$, Eq. (6.4) holds, and it follows that

$$(6.5) \quad x_d - \sum_{i=0}^{j-1} a_i \Phi_d(v_{f_i}) = \frac{1}{n-3} \cdot \frac{C_{r-1}}{C_r} \cdot \sum_{i=j}^{k-1} a_i + \sum_{i=k}^m a_i \Phi_d(v_{f_i}).$$

For a diagonal d in T that crosses a diagonal of f_k , $\Phi_d(v_{f_i}) = 0$, for $i \geq k$, and so Eq. (6.5) becomes

$$x_d - \sum_{i=0}^{j-1} a_i \Phi_d(v_{f_i}) = \frac{1}{n-3} \cdot \frac{C_{r-1}}{C_r} \cdot \sum_{i=j}^{k-1} a_i.$$

If d does not cross diagonals in f_k , then $\Phi(v_{f_k}) > 0$ and Eq. (6.5) becomes

$$x_d - \sum_{i=0}^{j-1} a_i \Phi_d(v_{f_i}) > \frac{1}{n-3} \cdot \frac{C_{r-1}}{C_r} \cdot \sum_{i=j}^{k-1} a_i.$$

Since at least one diagonal d in T intersects f_k ,

$$\min \left\{ x_d - \sum_{i=0}^{j-1} a_i \Phi_d(v_{f_i}) \mid d \in T \right\} = \frac{1}{n-3} \cdot \frac{C_{r-1}}{C_r} \cdot \sum_{i=j}^{k-1} a_i,$$

allowing us to solve for $a_j + \dots + a_{k-1}$ by computing this minimum.

Now we use the values $a_j + \dots + a_{k-1}$ for each subpolygon to determine a_j and f_{j+1} . If P_r contains a diagonal in $f_{j+1} \setminus f_j$, the value returned by the subroutine on P_r must be a_j . But if P_r does not contain such a diagonal, the returned value is at least $a_j + a_{j+1}$. Thus, the minimum value returned over all subpolygons of f_j is a_j . Since we know a_0, a_1, \dots, a_j and f_0, f_1, \dots, f_j , use Lemma 14 to test whether this statement holds for each diagonal d , and in doing so determine precisely which diagonals are in f_{j+1} .

This process is repeated with the additional known values a_j and f_{j+1} until case (A) is reached, at which point the values of a_i and f_i are fully determined. Since at no point was there freedom in choosing values for a_i or f_i , these values are uniquely determined. \square

The following shows Φ preserving the dimension of the simplicial subdivision.

Lemma 16. *If F is a flag of the associahedron K_{n-1} , then $\Phi(\Delta_F)$ is an $(n-3)$ -simplex. More generally, if S is a subflag of K_n , then $\Phi(\Delta_S)$ is an $(|S|-1)$ -simplex.*

Proof. Let $F = \{f_0, f_1, \dots, f_{n-3}\}$ be a flag of the associahedron K_{n-1} and let $S_i = \{f_0, f_1, \dots, f_i\}$. Proceed by induction on i . Since S_0 has cardinality one, $\Phi(v_{f_0})$ is a point in \mathfrak{S}_n . Now assume $\Phi(\Delta_{S_i})$ is an i -simplex, the convex hull of $\Phi(v_{f_0}), \dots, \Phi(v_{f_i})$. In order to show $\Phi(\Delta_{S_{i+1}})$ is an $(i+1)$ -simplex, it suffices to show that $\Phi(v_{f_{i+1}})$ is not in

the convex hull of $\Phi(v_{f_0}), \dots, \Phi(v_{f_i})$. Let d be the diagonal in f_{i+1} that is not in f_i . By Lemma 14, $\Phi_d(v_{f_{i+1}}) = 1/(n-3)$ and $\Phi_d(v_{f_j}) < 1/(n-3)$, for $j < i+1$, demonstrating the claim.

And since $\Phi(\Delta_F)$ is an $(n-3)$ -simplex, arising as the convex hull of the $n-2$ points $\Phi(v_{F_0}), \dots, \Phi(v_{F_{n-3}})$, the convex hull of any subset of these points (coming from a subflag) will form a simplex with dimension one less than the cardinality of the subset. \square

REFERENCES

1. H.-J. Bandelt and A. Dress. Split decomposition: a new and useful approach to phylogenetic analysis of distance data, *Molecular Phylogenetics and Evolution* **1** (1992) 242–252.
2. L. Billera, S. Holmes, K. Vogtmann. Geometry of the space of phylogenetic trees, *Advances in Applied Mathematics* **27** (2001) 733–767.
3. D. Bryant and V. Moulton. Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks, *Molecular Biology and Evolution* **21** (2004) 255–265.
4. M. Davis, T. Januszkiewicz, R. Scott. Nonpositive curvature of blowups, *Selecta Mathematica* **4** (1998) 491–547.
5. S. Devadoss. Tessellations of moduli spaces and the mosaic operad, in *Homotopy Invariant Algebraic Structures*, Contemporary Mathematics **239** (1999) 91–114.
6. S. Devadoss and J. Morava. Navigation in tree space, preprint [arxiv:1009.3224](https://arxiv.org/abs/1009.3224).
7. P. Diaconis and S. Holmes. Matchings and phylogenetic trees, *Proceedings of the National Academy of Sciences* **95** (1998) 14600–14602.
8. P. Gambette and D. Huson. Improved layout of phylogenetic networks, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **5** (2008) 1–8.
9. M. Gromov. Hyperbolic groups, in *Essays in Group Theory*, Springer, New York, 1987.
10. P. Hacking, S. Keel, J. Tevelev. Stable pair, tropical, and log canonical compactifications of moduli spaces of del Pezzo surfaces, *Inventiones Mathematicae* **178** (2009) 173–227.
11. D. Huson, R. Rupp, C. Scornavacca. *Phylogenetic networks*, Cambridge University Press, New York, 2010.
12. C. Lee. The associahedron and triangulations of the n -gon, *European Journal of Combinatorics* **10** (1989) 551–560.
13. D. Levy and L. Pachter. The neighbor-net algorithm, *Advances in Applied Mathematics* **47** (2011) 240–258.
14. D. Mumford, J. Fogarty, F. Kirwan. *Geometric Invariant Theory*. Springer-Verlag, New York, 1994.
15. M. Owen. Computing Geodesic Distances in Tree Space, *SIAM Journal on Discrete Mathematics* **25** (2011) 1506–1529.
16. N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution* **4** (1987) 406–425.
17. J. Sekiguchi and M. Yoshida. $W(E_6)$ -action on the configuration space of six lines on the real projective plane, *Kyushu Journal of Mathematics* **51** (1997) 297–354.

18. C. Semple and M. Steel. *Phylogenetics*, Oxford University Press, London, 2003.
19. D. Speyer and B. Sturmfels. The tropical Grassmannian, *Advances in Geometry* 4 (2004) 389–411.
20. R. Stanley. *Enumerative Combinatorics, Volume 2*, Cambridge University Press, 1999.
21. J. Stasheff. Homotopy associativity of H -spaces, *Transactions of the American Mathematical Society* **108** (1963) 275–292.

S. DEVADOSS: UNIVERSITY OF SAN DIEGO, SAN DIEGO, CA 92110
E-mail address: `devadoss@sandiego.edu`

S. PETTI: GEORGIA TECH, ATLANTA, GA 30332
E-mail address: `spetti@gatech.edu`