# Geometry of the Space of Phylogenetic Trees

Louis J. Billera

*Department of Mathematics, Malott Hall, Cornell University, Ithaca, NY 14853*
E-mail: billera@math.cornell.edu

and

Susan P. Holmes

*INRA, Montpellier, France and Department of Statistics, Stanford University, Stanford, CA 94305*
E-mail: susan@stat.stanford.edu

and

Karen Vogtmann

*Department of Mathematics, Malott Hall, Cornell University, Ithaca, NY 14853*
E-mail: vogtmann@math.cornell.edu

We consider a continuous space which models the set of all phylogenetic trees having a fixed set of leaves. This space has a natural metric of non positive curvature, giving a way of measuring distance between phylogenetic trees and providing some procedures for averaging or combining several trees whose leaves are identical. This geometry also shows which trees appear within a fixed distance of a given tree and enables construction of convex hulls of a set of trees.

This geometric model of tree space provides a setting in which questions that have been posed by biologists and statisticians over the last decade can be approached in a systematic fashion. For example, it provides a justification for disregarding portions of a collection of trees that agree, thus simplifying the space in which comparisons are to be made.

Mathematics Subject Classification: 92D15, 92B10, 05C05, 62P10.

Keywords: Phylogenetic trees, semi-labeled trees, associahedron, CAT(0) space, consensus, bootstrap.

This work was supported, in part, by NSF grants DMS9800910, DMS9973891 and DMS 9971607.

## MOTIVATION

Trees have been used extensively in biology and other fields to graphically represent various types of hierarchical relationships, including evolutionary relationships between species, divergent patterns between subpopulations and evolutionary relationships between genes. These trees are generally rooted and semi-labeled, *i.e.*, they descend from a single node called the root, bifurcate at lower nodes and end at terminal nodes, called tips or leaves; the leaves are labeled by the names of the species, subpopulations or genes being studied. In biological studies the latter are called operational taxonomic units (OTU's).

Traditionally, trees were inferred form morphological similarities among the OTU's. To build an evolutionary species tree, or *phylogenetic tree*, two species which shared the most characteristics were classified as 'siblings' and assumed to share a common ancestor which is not the ancestor of any other species. Such 'siblings' are said to be *homologous*, and it is this basic homology which has been of interest to biologists for a very long time. In Figure 1 we reproduce a tree from Haeckel (1866) which represents an attempt at depicting the relationships between all living organisms.

Over the last few decades, biologists have been building trees based on DNA sequences from certain parts of the genome. This has led to remarkable advances in the study of homology. Examples of the kinds of issues on which new light has been shed include the origin of diseases such as AIDS (Krushkal and Li (1998)) and the most deadly form of malaria (Escalante and Ayala (1995)), and connections between tribal groups such as those raised by the African tribe whose oral tradition holds that the tribe is descended from Jewish priests (DNA analysis does indicate such a relation).

In spite of the successes of DNA analysis, a great deal of uncertainty remains about precise relationships between the tips or leaves of the tree. Uncertainty about which branching order is the correct one is sometimes represented by filling out the tree as in Figure 2 to cover several possible binary trees and exclude others which biologists are sure are impossible.

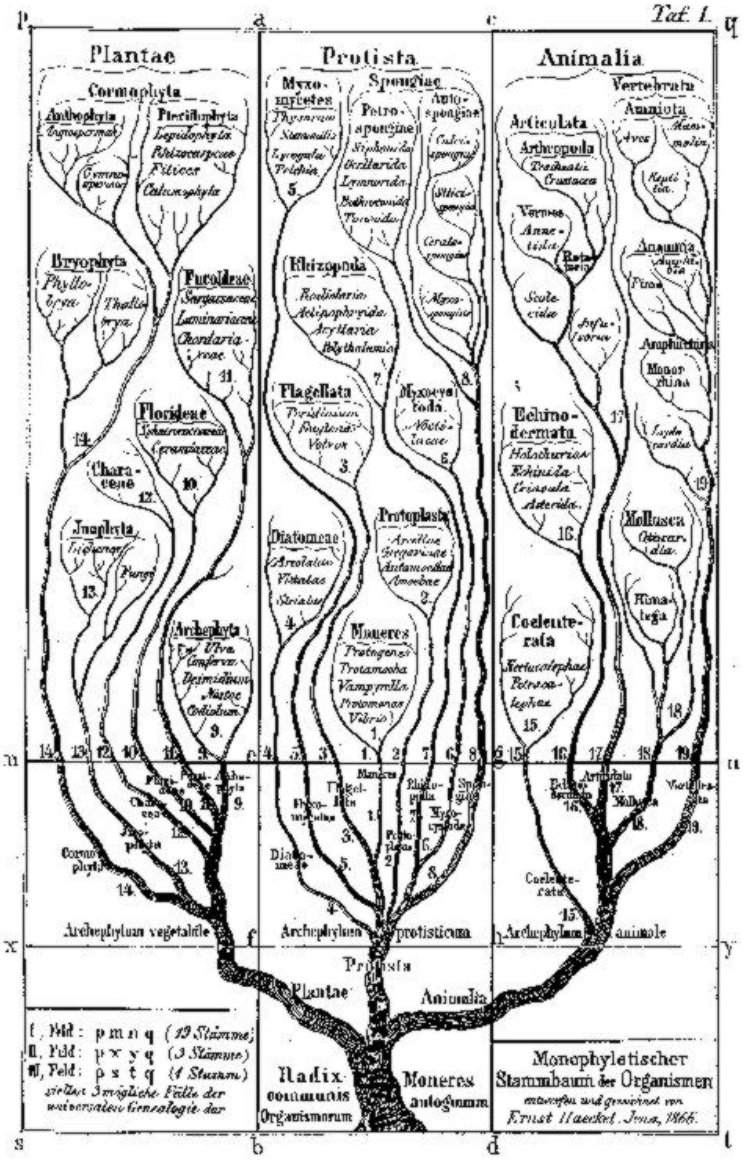


*Figure 1: Haeckel's tree with 3 branches*

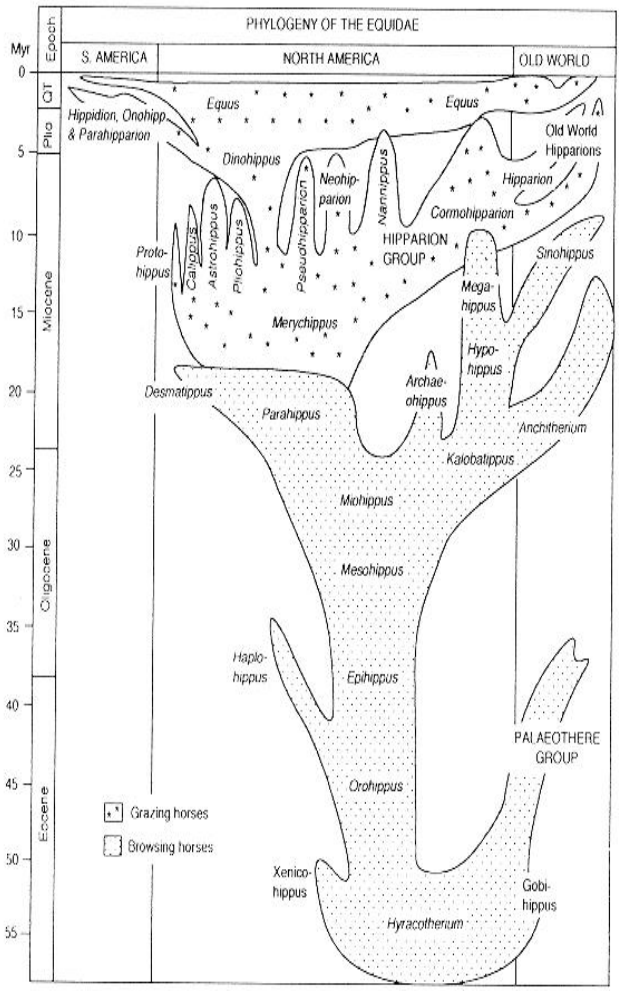


*Figure 2: Equus tree from (MacFadden, 1985)*

For example Figure 2 from MacFadden (1985) implicitly rules out the possibility of *Sinohippus* and *Protohippus* being homologous; however it also allows for indetermination of the branching order of *Neohipparion*, *Pseudohipparion* and *Cormohipparion*. In this paper we propose a geometric model which parameterizes the set of trees with a fixed set of OTU's; in this model, uncertainty can be represented by coloring in the portions of the space corresponding to possible trees.

One reason for uncertainty about the true phylogenetic tree is that different choices for DNA sequences (usually the choice of a single gene or coding region) often point to different trees, each of which is called a 'gene-tree' (Doyle, 1992). Finding the best way of combining the information contained in numerous different gene-trees for the same set of species remains an open problem in contemporary biology. Several methods have been proposed to solve this combination problem. One proposal is to treat the data from different genes as if they came from a single gene. For example, Brooks (1981) has suggested building all the different trees and then coding the tree data into binary columns, combining them and finding the best tree for the combined columns. Other proposed methods use some specified set of combination rules such as majority rule, strict consensus or Bayesian combination. A difficulty with combining data from different genes into a single, larger data set arises from differences in the mutation rates in different genes. Another interesting effect is that in simulation studies, where the true tree topology is known in advance, investigators have observed that a more accurate tree is obtained by subdividing the data into many different sequences and then averaging by some method than by agglomerating all of the sequences and then building a single tree with the merged data. Perturbing the simulated data by bootstrap resampling and then averaging also produces a tree which is closer to the known original tree (Berry and Gascuel, 1996). This points to the importance of understanding the rules used to average trees. None of the proposed consensus rules has previously been studied in a geometric context. Details of their comparison in the geometric context introduced in this paper will be explained in Billera et al. (2001).

Uncertainty about the true phylogenetic tree arises also from problems of statistical stability. The classical tree-building algorithms attempt to find a single tree consistent with the data. The question of how sure one is that the tree is correct is thus also a statistical one: the tree becomes an unknown parameter that the various procedures are trying to estimate. Would a small change in the data resulting from a sequencing or an alignment error result in a change of choice of the resulting tree? This is currently studied by using bootstrapping as a perturbation tool (Felsenstein, 1983), but in fact this can be interpreted as a problem in the estimation process. This problem has inspired certain authors (see Efron et al. (1996) and Zharkikh and Li (1995)) to imagine partitioning a space of trees into regions, each labeled by a different binary tree. When a data set is associated to a point in this space, the question of the resulting tree's stability can be translated into a question about how close the point is to the boundary between different regions. The question was raised in Zharkikh and Li (1995) as to how many regions are within a certain range of a given point. The current paper attempts to give the intuitive arguments presented in the above cited

papers a rigorous geometric interpretation. In particular, since our space of trees has a metric, this allows a "Voronoi" decomposition into nearest-neighbor regions, that is, regions consisting of those trees closest to each of a fixed finite set of trees (see Edelsbrunner (1987)).

One more reason for uncertainty about the true phylogenetic tree involves the tree-building process. The first problem encountered by taxonomists who build phylogenetic trees using any of the several methods available is the complexity of the underlying optimization problem. There are

$$(2n-3)!! = (2n-3) \times (2n-5) \times \ldots 3 = \frac{(2n-2)!}{2^{n-1}(n-1)!}$$

rooted binary semi-labeled trees with $n$ leaves (Schröder, 1870). The problem of computing the *best* tree for a certain data set is NP complete for two of the most common methods, the maximum likelihood methods and the parsimony methods (Foulds and Graham, 1982). As a consequence biologists have to use approximate optimization algorithms that use random starting points and certain random moves between trees. The resulting trees thus vary from run to run. The geometric model we introduce in this paper allows one to compare these trees in a quantitative way. Such comparisons could be useful in contexts such as those discussed in Lin and Gerstein (2000).

Biologists use a range of methods to construct trees from DNA sequences, each of which results in a tree with branch lengths. At one end of the spectrum lie the *parametric* models, such as the maximum likelihood method. In this method, a probability is given for each possible base change in a DNA sequence, and the tree that maximizes the likelihood under this model is the one chosen as the best estimate. Many biologists believe that as more data becomes available the mutation rates will be known with better accuracy and parametric models will be better justified. The geometric model of tree space presented in this paper enables one to represent the maximum likelihood tree as a point in a space of trees with branch lengths; it should then be possible to define isocontour regions around the estimated tree to build the desired confidence regions.

In a parametric model, the data are approximated by points in a very low-dimensional manifold, thereby losing much of the information contained in the original data. The Jukes-Cantor model, for instance, uses an $n$-dimensional parameterization of the data corresponding to trees with $n$ leaves. To get a rough idea of this, imagine asserting that the data points lie on an ellipse and then choosing the two parameters of the ellipse so as to minimize the sum of the distances from the points to the ellipse. The ellipse is parameterized by two numbers, and represents the parametric model that biologists will try to fit the data to.

At the other end of the spectrum of tree-building methods lie the *non-parametric* models, such as the parsimony representation. A nonparametric approach could simply interpolate between points; as the number of points increases the number of descriptive parameters increases. A more sophisticated nonparametric approach would propose a smooth curve minimizing the distance to the points. Thus nonparametric methods are also said to be *infinite dimensional*. For instance, in the parsimony model, the tree is defined to be the minimal Steiner tree compatible with the observed distances between the OTU's, the branch lengths then represent numbers of mutation events.

In between these two extremes lie the distance-based methods, which are *semi-parametric* models, in which the mutation model is parametric with very few parameters (usually between one and four) and the tree building procedure is non-parametric. See Holmes (1999) for a detailed comparison of these three estimation paradigms.

Each method of producing trees from data results in trees with branch lengths, but these branch lengths have different meanings in different methods. The choice of which procedure is used to produce trees will not affect the geometric representation of the space of trees as we propose it here, but only the interpretation of points in the space.

A brief summary of the paper follows. In §1, we describe two preliminary attempts to obtain a geometric setting for the study of trees, each closely related to a convex polytope (the matching polytope and the associahedron). In §2 we give an explicit construction of the space of trees $\mathcal{T}_n$, and in §3 we give some of its basic combinatorial properties. While $\mathcal{T}_n$ is not a manifold, the underlying combinatorial properties of trees help expose some of its structure. In §4 we study the geometric properties of $\mathcal{T}_n$ such as curvature (the CAT(0) property), geodesics and centroids. We also discuss ways to introduce probability measures on this space in order to find a geometric setting for the statistical study of tree data. We conclude in §5 with a discussion of some of the questions that arise when considering such data.

## 1. TWO PRELIMINARY ATTEMPTS

In Diaconis and Holmes (1998), trees were coded as "matchings" on a complete graph. These matchings allow trees to be identified with the vertices of a convex polytope, called the *matching polytope* (see Lovasz and Plummer (1985)). A shortcoming of this matching representation is that a small move on the matching polytope may have either a very small or a very large effect on the tree, as it interchanges two nodes which may be either far from or close to the root. This asymmetry in the matching

representation is not present in the geometric representation presented in this paper.

There is another convex polytope, called the *associahedron* (see Lee (1989) or Stasheff (1963)) whose vertices can be identified with the set of planar rooted binary trees with $n$ leaves in a fixed order or, equivalently, with the set of triangulations of an $(n+1)$-gon. The associahedron for $n = 4$ is a pentagon, and is illustrated in Figure 3; the triangulations are indicated by dotted lines and the corresponding binary trees are drawn with solid lines. Two vertices of the associahedron are adjacent if the corresponding triangulations differ by "rotating" a single interior edge $e$, *i.e.*, removing $e$ to form a quadrilateral in the interior of the $(n+1)$-gon and then replacing $e$ by the opposite diagonal of the quadrilateral. The corresponding trees are also said to be linked by *rotation* (see Figure 15).

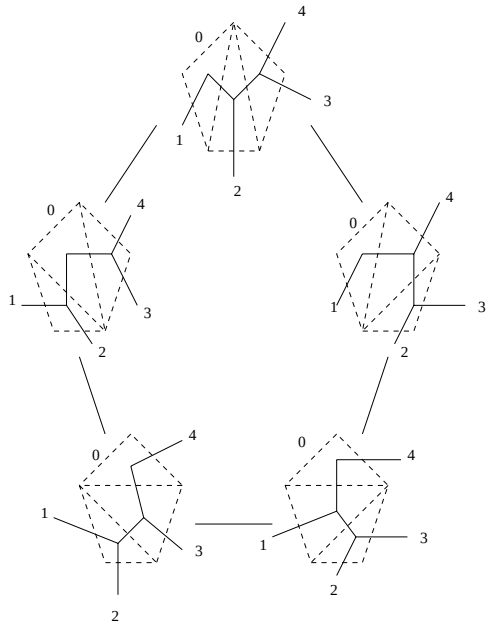


*Figure 3: Associahedron in the case $n = 4$*

By "gluing" associahedra together, one can construct a space of planar labeled trees with $n$ leaves, where each associahedron corresponds to a different ordering of the labels. This space has appeared in several different contexts (Davis et al., 1998; Devadoss, 1999; Kapranov, 1993), and is denoted $\overline{M}_{0,n+1}$. The space $\overline{M}_{0,5}$ is tiled with 12 pentagons, corresponding to all possible permutations of the leaves up to complete reversal. Each space $\overline{M}_{0,n+1}$ has a dual tiling by $(n-3)$-dimensional cubes. The dual tiling of $\overline{M}_{0,5}$, by squares, is illustrated in Figure 4; in the dual tiling, the

12 pentagons become 12 vertices of degree 5. The shaded region shows a single tile of the tiling by associahedra.

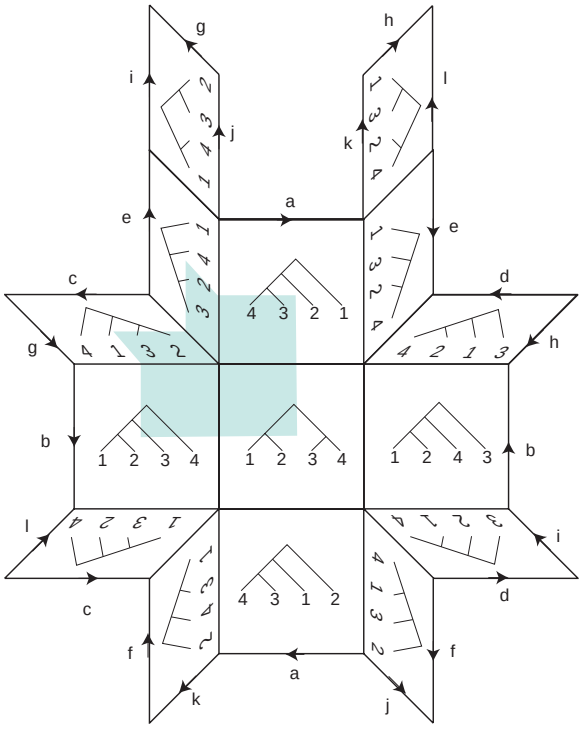


*Figure 4: Cubical tiling of $\overline{M}_{0,5}$, where the arrows indicate oriented identifications.*

A problem with the above representation is that we are interested in the abstract combinatorial information contained in the tree, which does not depend on how the tree is embedded in the plane. The space of trees as described in this paper is in fact a quotient of $\overline{M}_{0,n+1}$, but a direct construction seems easier to visualize. One should be able to view this space as the subset of the cone of all metrics on a fixed finite set consisting of those metrics that are derived from trees. See, for example, Böcker and Dress (1998) for the relation between trees and metrics.

## 2. CONSTRUCTION OF THE SPACE OF TREES

In this section, we describe a geometric model for tree space, in which each point represents a rooted semi-labeled tree with $n$ leaves and positive branch lengths on all interior edges. In general one moves around in the space by varying the branch lengths of the trees, but when a branch length

reaches 0 some degeneration or uncertainty occurs which can be resolved in one of several ways, each of which leads to a new tree.

We now proceed to formally define the space. The term *n-tree* will mean a tree (*i.e.*, a connected graph with no circuits), with a distinguished vertex, called the *root*, and $n$ vertices of degree 1, called *leaves*, that are labeled from 1 to $n$. Although we are primarily interested in binary trees (*i.e.*, trees in which the root has degree 2 and all other vertices have degree 1 or 3), in order to interpolate between these we will also need to consider trees whose vertices have larger degree. Perversely, mathematicians usually put the root at the top when drawing a picture of a tree, so that the tree "grows downward" from its root (see Figure 5).

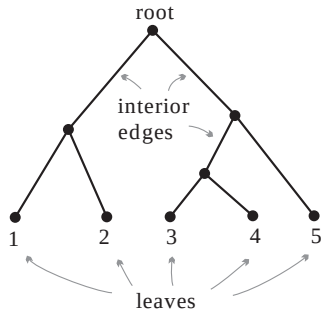


*Figure 5: A semi-labeled binary tree*

For technical reasons, it will often be convenient to "hang each tree up by its root," *i.e.*, to place an edge directly above the root of every tree, with the corresponding leaf labeled with 0. Note that there are several ways of drawing a diagram of the same tree, depending on how it is embedded in the plane. For example, the three pictures in Figure 6 represent the same tree.

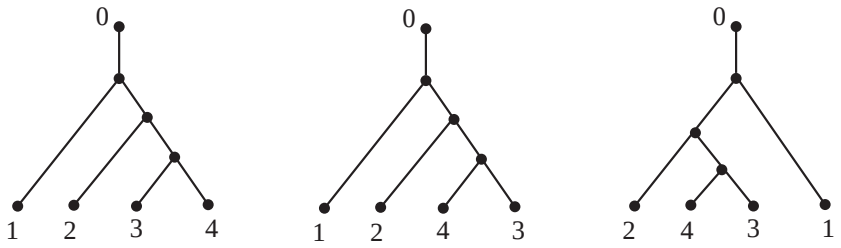


*Figure 6: Three pictures of the same tree*

On the other hand, two trees that have exactly the same combinatorial structure but whose leaves are labeled differently are considered different (see Figure 7). The number of different binary trees on $n$ leaves is equal

to $(2n-3)!!$. In contrast, the number of different *unlabeled* trees with $n$ leaves is the Catalan number $C_{n-1} = \frac{1}{n}\binom{2(n-1)}{n-1}$. For example, there are 15 different binary trees with 4 leaves. If we do not restrict ourselves to binary trees, the enumeration can be done through an exponential generating function (Stanley, 1999, p. 14).

The problem of enumerating labeled trees is Schröder's fourth problem Schröder (1870). Stanley (1999, p.14) finds that there is no analytical formula. The solution to Exercise 5.40 (page 133) in Stanley (1999) gives references and a discussion.

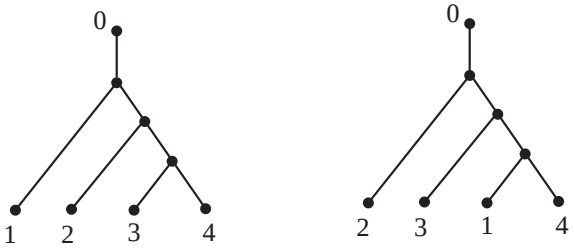


*Figure 7: Different trees*

A *metric $n$-tree* is an $n$-tree with lengths greater than 0 on all of its interior edges. (An edge of an $n$-tree is called *interior* if it is not connected to a leaf.) In what follows, the term "tree" will mean a metric $n$-tree, unless otherwise specified.

One could also consider trees with positive lengths on *all* edges, including those leading to leaves. However, the effect of this on tree space is simply to take the product with an $n$-dimensional Euclidean space. Since this does not significantly affect the geometry of the space, we will ignore this, knowing that it is possible to add this information at any later point that we wish.

Now consider a tree $T$, with interior edges $e_1, \ldots, e_r$ of lengths $l_1, \ldots, l_r$ respectively. If $T$ is binary, then $r = n-2$; otherwise $r < n-2$. The vector $(l_1, \ldots, l_r)$ specifies a point in the positive open orthant $(0, \infty)^r$. To each other point in this orthant, we associate the unique metric $n$-tree which is combinatorially the same as $T$ but has different edge lengths, specified by the coordinates of that point. Points on the boundary of the orthant, *i.e.*, length vectors with at least one coordinate equal to zero, correspond to metric $n$-trees which are obtained from $T$ by shrinking some interior edges of $T$ to 0; thus each point in the orthant $[0, \infty)^r$ corresponds to a unique metric $n$-tree (see Figure 8).

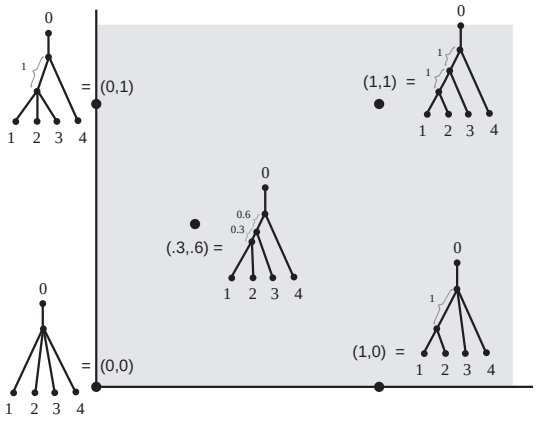


*Figure 8: The 2-dimensional quadrant corresponding to a metric 4-tree*

An $n$-tree has the maximal possible number of interior edges (namely $n-2$), and thus determines the largest possible dimensional orthant, when it is a binary tree; in this case the orthant is $(n-2)$-dimensional. The orthant corresponding to each tree which is not binary appears as a boundary face of the orthants corresponding to at least three binary trees; in particular the origin of each orthant corresponds to the (unique) tree with no interior edges. We construct the space $\mathcal{T}_n$ by taking one $(n-2)$-dimensional orthant for each of the $(2n-3)!! = (2n-3)\cdot(2n-5)\cdots 5\cdot 3\cdot 1$ possible binary trees, and gluing them together along their common faces.

For $n = 3$ there are three binary trees, each with 1 interior edge. Each tree thus determines a 1-dimensional "orthant," *i.e.*, a ray from the origin. The three rays are identified at their origins (see Figure 9).

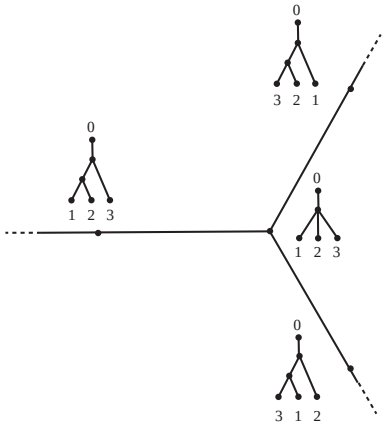


*Figure 9: $\mathcal{T}_3$*

For $n = 4$ there are 15 binary trees, so that the space $\mathcal{T}_4$ consists of 15 two-dimensional quadrants which all share a common origin. Each boundary ray appears in exactly 3 of the quadrants as in Figure 10. Note that a horizontal slice of this figure forms a copy of $\mathcal{T}_3$ embedded in $\mathcal{T}_4$. In general, $\mathcal{T}_n$ contains many embedded copies of $\mathcal{T}_k$ for $k < n$.

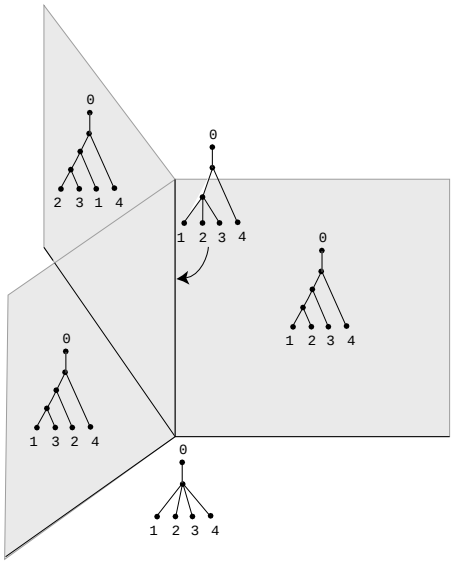


*Figure 10: Three quadrants sharing a common boundary ray in $\mathcal{T}_4$*

All 15 quadrants for $n = 4$ share the same origin. If we take the diagonal line segment $x + y = 1$ in each quadrant, we obtain a graph with an edge for each quadrant and a trivalent vertex for each boundary ray (see Figure 11). This graph is called the *link of the origin.*

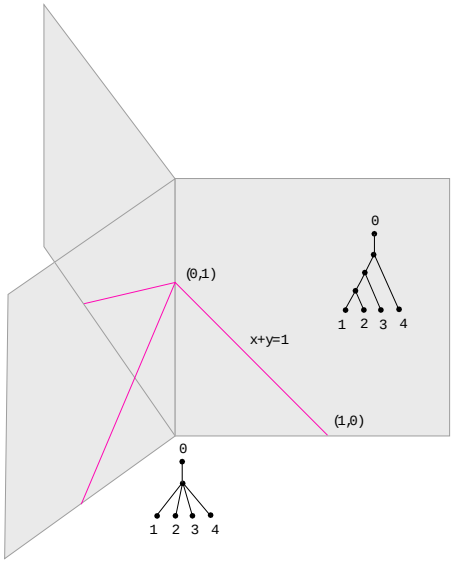


*Figure 11: Constructing the link of the origin in $\mathcal{T}_4$*

Figure 12 shows another portion of the link which forms a pentagon embedded in its ambient quadrants.

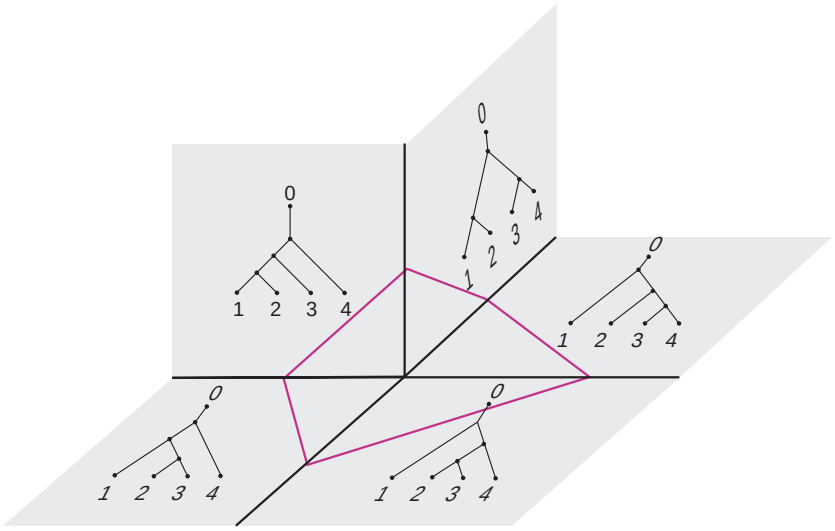


*Figure 12: A pentagon in the link*

The entire link of the origin is shown in Figure 13, without the ambient quadrants. The entire space $\mathcal{T}_4$ is an infinite cone on this graph, with cone point the origin. It is interesting to note that the link of the origin in

this case is a well-known graph, called the *Peterson graph*. The Peterson graph has no planar embedding, and the space $\mathcal{T}_4$ cannot be embedded in 3-dimensional space.

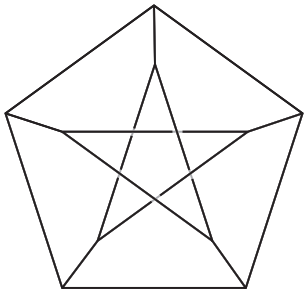

*Figure 13: Link of the origin in $\mathcal{T}_4$*

One can visualize $\mathcal{T}_4$ as being obtained from the space pictured in Figure 14 by gluing together edges with the same label. We note that the figure does not look metrically correct, since each triangle should be a right triangle with right angle at the origin; also, each triangle should extend forever in the direction away from the origin.

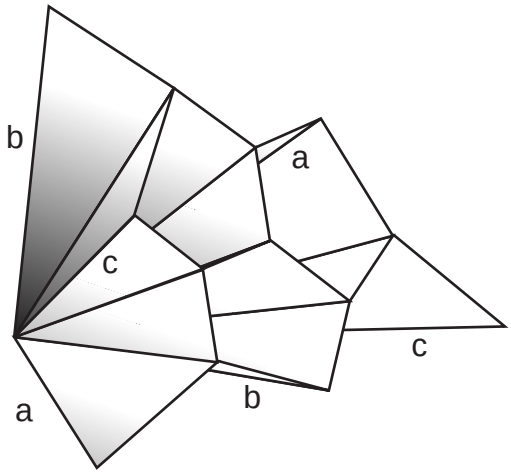


*Figure 14: $\mathcal{T}_4$*

## 3. COMBINATORICS OF THE SPACE OF TREES

In this section we consider certain combinatorial aspects of the space of trees, and in particular relations to combinatorial structures which have been studied in other contexts. The combinatorial properties of the link of

the origin of this space will be useful in the study of its geometry in the following section.

### 3.1. Relation to the associahedron and moduli spaces

We observe that the link of the origin in the space $\mathcal{T}_4$ is a graph whose shortest circuit has length 5.

Figure 12 above showed a length 5 circuit in this graph, embedded in the appropriate quadrants of $\mathcal{T}_4$. This pentagon is easily identified with the boundary of the dual polytope of the associahedron on 4 letters (see Figure 3). This is a general phenomenon.

The link of the origin $L_n$ is defined for all values of $n$, as the union of the sets of points in each orthant with coordinate sum equal to 1. Since the set of such points in a single orthant forms a simplex, $L_n$ has the structure of a simplicial complex of dimension $n-3$, with one $k$-simplex for every tree with $k+1$ interior edges.

PROPOSITION 3.1. *The dual of the associahedron on $n$ letters is embedded in $\mathcal{T}_n$; its boundary is a subcomplex of the link $L_n$.*

**Proof:** The associahedron parameterizes the set of planar rooted trees with $n$ leaves in a fixed order. ∎

If we restrict the branch lengths to be bounded by some constant $C > 0$, then the resulting subspace of $\mathcal{T}_n$ is a quotient of the manifold $\overline{M}_{0,n+1}$ defined in section 1. Points of $\overline{M}_{0,n+1}$ can be interpreted as rooted *planar* trees with branch lengths between 0 and $C$, modulo a certain equivalence relation, given as follows: a rooted planar tree has a natural left-to-right ordering on the edges descending from each vertex; if the edge above a vertex $P$ has length $C$, then reversing all orderings at $P$ and at all vertices below $P$ produces an equivalent tree. The manifold $\overline{M}_{0,n+1}$ has been studied by mathematicians in a variety of different guises (moduli space of stable $(n+1)$-pointed curves, minimal blow-up of the projective braid arrangement, cyclic operad of mosaics). See for example Davis et al. (1998); Devadoss (1999); Kapranov (1993); the latter especially gives some background references.

### 3.2. Combinatorics of the link of the origin

An alternate description of the link $L_n$ can be given in terms of partitions of the set $\{0, 1, \ldots, n\}$ of leaves (recall that we have attached a leaf labeled 0 to the root). The correspondence between partitions and trees hinges on the observation that each interior edge of a tree partitions the leaves into two sets, each with at least two elements (such a partition is called *thick*). Different edges of the same tree give compatible partitions,

where two partitions $\{X, Y\}$ and $\{X', Y'\}$ of $\{0, 1, \ldots, n\}$ are defined to be *compatible* if one of the subsets

$$X \cap X' \qquad X \cap Y' \qquad X' \cap Y \qquad Y \cap Y'$$

is empty. The link $L_n$ can now be identified with the simplicial complex whose $k$-simplices are sets of $k+1$ pairwise compatible thick partitions of $\{0, 1, \ldots, n\}$. In this guise, $L_n$ is studied in Vogtmann (1990), where it is shown that $L_n$ has the homotopy type of a wedge of $(n-1)!$ spheres of dimension $(n-3)$ (in fact, $L_n$ is Cohen-Macaulay); see also Robinson and Whitehouse (1996). Each of these spheres corresponds to the boundary of an associahedron embedded in $\mathcal{T}_n$.

### 3.3. Tree rotations

Combinatorialists sometimes measure the distance between binary trees by counting the number of rotations needed to change one tree to another. Here a *rotation* is a move which collapses an interior edge to zero, then expands the resulting degree 4 vertex into an edge and two degree 3 vertices in a new way (see Figure 15). This move is known to the biologists as a nearest neighbor interchange (NNI) Waterman and Smith (1978).

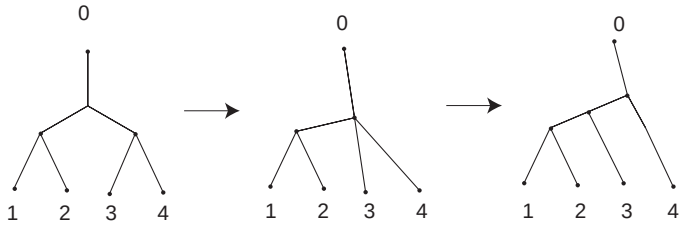


*Figure 15: Rotation*

In the link $L_n$ as we have defined it, each maximal simplex corresponds to a binary tree, and two maximal simplices share a codimension 1 face if and only if the corresponding trees differ by a rotation move. In Sleator et al. (1992) it is shown that the maximal rotation distance between two trees on $n$ leaves is $O(n \log n)$, while the maximal rotation distance between two trees contained in the same associahedron is exactly $2n-6$ (see Sleator et al. (1988)). These results give an indication of the size of our space of trees.

## 4. GEOMETRY OF THE SPACE OF TREES

By the geometry of the space we mean its metric, as opposed to combinatorial, properties. The space of trees comes equipped with a natural distance function, due to the fact that it is made up of standard Euclidean

orthants. The distance between any two points in the same orthant is simply the usual Euclidean distance. If two points are in different orthants, we can join them by a sequence of straight segments, with each segment lying in a single orthant; we can then measure the length of the path by adding up the lengths of the segments. We define the distance between the two points to be the minimum of the lengths of such "segmented" paths joining the two points. A segmented path giving the smallest distance between two points is called a *geodesic*.

### 4.1. Non-positive curvature

A metric space $X$ is said to have *non-positive curvature* if triangles in $X$ are "at least as thin" as Euclidean triangles (see Figure 16). More precisely, $X$ is said to be *CAT(0)* if the following is true: given any three points $a, b$ and $c$ in $X$, with distances $d_1 = d(b, c), d_2 = d(a, c)$ and $d_3 = d(a, b)$, form a "comparison triangle" in the Euclidean plane with vertices $a', b'$ and $c'$ with side lengths $d_1 = d(b', c'), d_2 = d(a', c')$ and $d_3 = d(a', b')$. If $x$ is a point on the geodesic from $a$ to $b$, at distance $d$ from $a$, find the corresponding point $x'$ on the straight line from $a'$ to $b'$ at distance $d$ from $a'$. Then $d(x, c) \leq d(x', c')$.

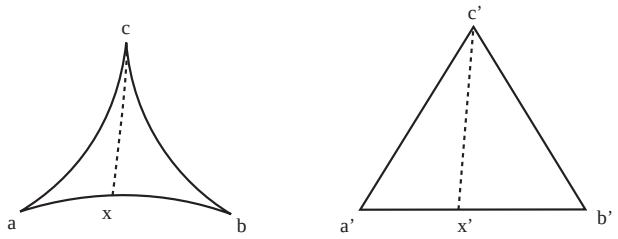


*Figure 16: Comparison triangle*

The following lemma shows that the natural metric on $\mathcal{T}_n$ has non-positive curvature. This key property of $\mathcal{T}_n$ has many important consequences, including uniqueness of geodesic paths and existence and uniqueness of various types of centroids.

LEMMA 4.1. *$\mathcal{T}_n$ is a CAT(0) space.*

*Proof.* We first subdivide each orthant into the unit cubes having integral vertices. The space $\mathcal{T}_n$ is then a cubical complex. A theorem of Gromov (1987) states that a cubical complex is CAT(0) if and only if the link of every vertex is a *flag* complex, *i.e.*, a simplicial complex in which a simplex belongs to the complex if and only if its entire 1-skeleton does. (In particular, if all the edges of a triangle are in the complex then so is the triangle; if all edges of a tetrahedron are in the complex, then so is the

tetrahedron, and so on). Note that the link $L_n$ of the origin, defined in the previous section, is such a complex, since simplices are defined by pairwise compatibility of partitions.

Let $v$ be an arbitrary vertex of the cube complex, which lies in the interior of a (unique) orthant of dimension $k$. This orthant corresponds to a tree with $k$ interior edges, and thus to a set $S$ of $k$ pairwise compatible partitions of $\{0, \ldots, n\}$. If $k$ is maximal, *i.e.*, $k = n-2$, the link of $v$ is a triangulated sphere, which we think of as the $k$-fold suspension of the empty set. In general, the link of $v$ is the $k$-fold suspension of the subcomplex of $L_n$ spanned by all partitions compatible with $S$. Since this itself is a flag complex, and since the suspension of a flag complex is again flag, this completes the proof. ∎

Alternatively, $\mathcal{T}_n$ is the 0-cone on the link $L_n$ (for definition, see Bridson and Haefliger (1999, I.5)). Since $L_n$ is a flag complex, it is is CAT(1) by Gromov's theorem (Bridson and Haefliger (1999, 5.18,p. 211)). A theorem of Berestowski (Bridson and Haefliger (1999, 3.14, p. 188)) then implies that $\mathcal{T}_n$ is CAT(0).

In the case $n = 4$, the flag condition says that the links of all vertices are graphs with no triangles; note that, for example, the smallest circuit in the link of the origin has length 5. The fact that the set of unlabeled trees forms a flag complex was noted in (Billera et al., 1999).

### 4.2. Geodesics

Since the tree space $\mathcal{T}_n$ is CAT(0), it follows by Gromov (1987) that there is a unique shortest path connecting any two points of $\mathcal{T}_n$, called the *geodesic*. In this section we characterize geodesics and show how to find them. Once the geodesic is found, its length gives the distance between the two trees.

There is an obvious path between any two trees $T$ and $T'$ in $\mathcal{T}_n$, obtained by connecting $T$ to the origin by a straight line segment, then connecting the origin to $T'$ by another straight line segment; we will call this path the *cone path* from $T$ to $T'$. The cone path may or may not be a geodesic, depending on the "angle" it makes at the origin $T_0$. One makes this precise as follows.

We have described the link of the origin in $\mathcal{T}_n$ as the union of "flat" simplices, consisting of all points in each orthant with coordinate sum equal to one. We could just as well have considered each simplex as the intersection of the unit sphere with the appropriate orthant, *i.e.*, the set of points such that the sum of the squares of the coordinates is equal to one. This new metric on simplices extends to a natural metric on the entire link $L_n$, in which each simplex is a right-angled spherical simplex with all edges of length $\pi/2$. Each tree $T$ of $\mathcal{T}_n$ lies on a unique ray from the origin. The intersection of this ray with $L_n$ is called the *projection* of $T$ onto $L_n$, and

is denoted $t(T)$. The *angle* between $T$ and $T'$, denoted $\angle(T, T')$, is defined to be the distance between $t(T)$ and $t(T')$ in the spherical metric on $L_n$.

Standard CAT(0) theory (see Bridson and Haefliger (1999), 5.6-5.10) tells us that the cone path is a geodesic if and only if the angle between $T$ and $T'$ is at least $\pi$. If $\angle(T, T') < \pi$, the geodesic $g$ from $T$ to $T'$ projects to the unique geodesic $\gamma$ from $t(T)$ to $t(T')$ in $L_n$; furthermore, if we know $\gamma$, we can reconstruct $g$.

Another standard notion we will need in this section is that of the *development* of a geodesic in a spherical complex (see Bridson and Haefliger (1999, p. 104), where the development of a geodesic is more generally defined). Let $v$ be a vertex of $L_n$, and let $\gamma$ be a geodesic in $L_n$ starting at a point $t$ in the interior of a simplex in the star of $v$. If $\gamma$ intersects a simplex $\sigma$ in an arc of positive length, we say that $\gamma$ *traverses* $\sigma$. Let $\sigma_1, \sigma_2, \ldots$ be the sequence of simplices which $\gamma$ traverses. For each $i$, $\gamma$ intersects the common face $\sigma_i \cap \sigma_{i+1}$ in a single point, which we will call $t_i$. If $t_1 \neq v$, take the totally geodesic surface in $\sigma_1$ containing $t, v$ and $t_1$; this surface is a spherical triangle $\tau_1$, with the distance of $v$ to the other vertices equal to $\pi/2$ (if we think of $\sigma_1$ as lying in the unit sphere in a Euclidean orthant, this surface is the intersection of $\sigma_1$ with the three-dimensional subspace containing these three points). We embed this triangle as a triangle $\bar{\tau}_1$ in $S^2$ with the image $\bar{v}$ of $v$ at the north pole (see Figure 17).

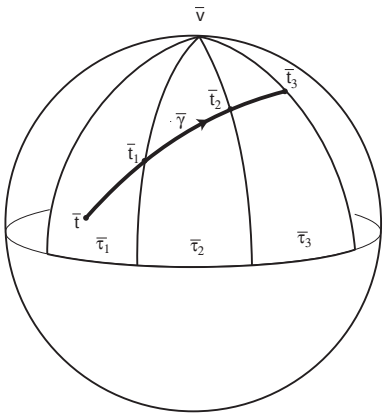


*Figure 17: Development of $\gamma$ on $S^2$*

If $\gamma$ exits $\sigma_1$ via a face with a vertex at $v$, we next take the totally geodesic surface in $\sigma_2$ containing $v$, $t_1$ and $t_2$; again this is a spherical triangle $\tau_2$ which we embed in $S^2$ as a triangle $\bar{\tau}_2$ adjacent to $\bar{\tau}_1$, with the image of $v$ at the north pole. We continue to lay out triangles in $S^2$ as long as the "exit faces" of $\gamma$ have $v$ as a vertex. The image $\bar{\gamma}$ of $\gamma$ in $S^2$ is called the *development of $\gamma$ near $v$*, and is isometric to its preimage in $\gamma$. Recall

from Section 3 that each interior edge of a tree $T$ partitions the leaves of $T$ into two sets, each with at least two elements. Edges of $T$ and $T'$ are said to be the *same* if they determine the same partition of the leaf-labels, and *compatible* if the corresponding partitions are compatible. A set of partitions corresponds to the set of interior edges of a tree if and only if the partitions are pairwise compatible.

PROPOSITION 4.1. *If the cone path from $T$ to $T'$ is not a geodesic, then there are non-empty sets $E_1 \supset E_2 \supset \ldots \supset E_k$ of the edges $E(T)$ of $T$, and $F_1 \subset F_2 \subset \ldots \subset F_k$ of the edges $E(T')$ of $T'$ such that*
*(i) each element of $E_i$ is compatible with each element of $F_i$, so that $E_i \cup F_i$ form the vertices of a simplex $\sigma_i$ of $L_n$ and*
*(ii) the geodesic in $L_n$ from $t(T)$ to $t(T')$ traverses each simplex in the sequence $\sigma_1, \ldots, \sigma_k$.*

*Proof.* Since the cone path is not a geodesic, the geodesic $\gamma$ realizing the distance between $t(T)$ and $t(T')$ has length less than $\pi$.

Let $\sigma$ be the simplex of $L_n$ spanned by the edges $E(T)$. We first consider the case that $\gamma$ traverses $\sigma$. If $\gamma$ is contained in the closure of $\sigma$, the proposition is trivial. If not, $\gamma$ leaves $\sigma$ via a face corresponding to a subset of $E(T)$, which we define to be $E_1$; this face is also a face of the next simplex $\sigma_1$ which $\gamma$ traverses.

Fix any vertex $v$ in $\sigma_1$ which is not in $E_1$, and develop $\gamma$ near $v$. Since $\gamma$ has length less than $\pi$, the development $\bar{\gamma}$ remains in the northern hemisphere; this translates to the fact all simplices encountered by $\gamma$ must have $v$ as a vertex, including the simplex containing $t(T')$, *i.e.*, $v$ corresponds to an edge of $T'$. Since $v$ was an arbitrary vertex of $\sigma_1$ not in $E_1$, the set of vertices of $\sigma_1$ consists of $E_1$ plus a subset $F_1$ of edges of $T'$.

We now continue following the simplices traversed by $\gamma$, and repeat the argument to find vertex sets $E_i$ and $F_i$ as in the statement of the proposition until we arrive at the simplex containing $t(T')$.

If $\gamma$ does not traverse $\sigma$, we set $E_1 = E(T)$, and let $\sigma_1$ be the first simplex traversed by $\gamma$. We take any vertex $v$ of $\sigma_1$ which is not not in the face spanned by $E_1$ and develop $\gamma$ near $v$. We conclude that every simplex encountered by $\gamma$ has $v$ as a vertex, including the simplex containing $t(T')$. Thus all vertices of $\sigma_1$ which are not in $E_1$ are in $E(T')$, and we can set $F_1$ to be the vertices of $\sigma_1$ not in $E_1$. We now continue as before until we arrive at $t(T')$. ∎

COROLLARY 4.1. *If no edge of $T$ is compatible with any edge of $T'$, then the cone path is a geodesic.*

*Proof.* If the cone path is not a geodesic, any element of $F_1$ is compatible with any element of $E_1$, by the proposition. ▌

**Example.** The cone path may be a geodesic, even if $T$ and $T'$ do have some compatible edges. For example, let $T$ be the tree on four leaves with edges $e_1 = \{2,3|0,1,4\}$ and $e_2 = \{1,2,3|0,4\}$. and let $T'$ be the tree with edges $f_1 = \{0,1|2,3,4\}$ and $f_2 = \{0,1,2|3,4\}$. Then $e_1$ and $f_1$ are compatible. If the lengths of $e_1$ and $f_1$ are relatively large, then the geodesic from $T$ to $T'$ passes through trees with edges $\{e_1, f_1\}$. However, if the lengths of $e_1$ and $f_1$ are small, the cone path will be a geodesic (see Figure 18 ).

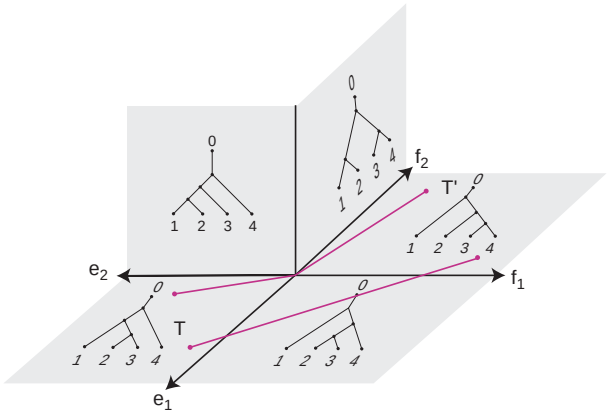


*Figure 18: Cone path may or may not be geodesic*

Proposition 4.1 allows us to give an effective procedure for finding the geodesic between binary trees $T$ and $T'$. We realize the orthants of $T$ and $T'$ as the totally negative and totally positive orthants of $(n-2)-$dimensional Euclidean space $\mathbf{R}^{n-2}$. We find all possible chains $E_i$ and $F_i$ as in the statement of the proposition, find a candidate geodesic for each chain, and compare their lengths. We carry out this procedure in Billera et al. (2001).

Suppose $E_i$ has $n_i$ elements and $F_i$ has $m_i$ elements. We order the edges of $T$ in such a way that edges in $E_i$ correspond to the first $n_i$ coordinates of $\mathbf{R}^{n-2}$ and edges in $F_i$ correspond to the last $m_i$ coordinates. Our candidate for the geodesic from $T$ to $T'$ is then a union of straight line segments in $\mathbf{R}^{n-2}$, constrained by the fact that each line segment must lie in one of the orthants whose first $n_i$ coordinates are negative, whose last $m_i$ coordinates are positive, and whose remaining coordinates are zero. We illustrate this with the following special case:

Let $T \in \mathcal{T}_n$ be a tree, and $e$ an interior edge of $T$. We denote by $|e|_T$ the branch length of $e$ in $T$.

PROPOSITION 4.2. *Let $T$ and $T'$ be binary trees with no edges in common. Suppose the edges $\{e_i\}$ of $T$ and $\{f_i\}$ of $T'$ can be ordered in such a way that $E_i = \{e_1, \ldots, e_i\}$ and $F_i = \{f_{i+1}, \ldots, f_{n-2}\}$ are compatible for all $k = 1, \ldots, n-3$. If for all $i < j$ we have $|e_i|_T|e_j|_{T'} - |e_j|_T|e_i|_{T'} > 0$, then the geodesic from $T$ to $T'$ contains trees with edge sets $E_i \cup F_i$ for all $i$, and the distance from $T$ to $T'$ is the length of the vector $(|e_1|_T + |e'_1|_{T'}, \ldots, |e_{n-2}|_T + |e'_{n-2}|_{T'})$.*

*Proof.* The compatibility conditions say that the orthants corresponding to the trees $T_i$ and $T_{i+1}$ share a codimension 1 face; in fact we may arrange that the orthant for $T_i$ is the orthant whose first $n-2-i$ coordinates are negative and whose last $i$ coordinates are positive. The tree $T$ corresponds to the point $(-|e_1|_T, \ldots, -|e_{n-2}|_T)$ and $T'$ to the point $(|e'_1|_{T'}, \ldots, |e'_{n-2}|_{T'})$; the inequalities ensure that the straight line between these two points is contained in the union of the orthants corresponding to the $T_i$, which is therefore the geodesic from $T$ to $T'$. ∎

The following corollary says that we can basically ignore edges of $T$ and $T'$ which are the same when we are computing the geodesic from $T$ to $T'$:

COROLLARY 4.2. *Let $e$ be an edge of $T$ which is also an edge of $T'$. Then every tree on the geodesic from $T$ to $T'$ has $e$ as an edge.*

*Proof.* The geometric meaning of this statement is that the union $X(e)$ of the orthants containing the ray $R(e)$ corresponding to $e$ is convex in $\mathcal{T}_n$.

Since $R(e)$ is an edge of the orthant containing $T$, the angle between $R(e)$ and $T$ is less than $\pi/2$. Since the orthants containing $T$ and $T'$ intersect in $R(e)$, the angle between $T$ and $T'$ is less than $\pi$, so that the cone path is not a geodesic. Consider the geodesic $\gamma$ in the link $L_n$ from $t(T)$ to $t(T')$. By Proposition 4.1, every edge in every simplex traversed by $\gamma$ is compatible with $e$, *i.e.*, $\gamma$ stays in the closed star of the vertex $v(e)$ corresponding to $e$. If we develop $\gamma$ near $v(e)$, it begins and ends in the open northern hemisphere, so remains in the open northern hemisphere at all times. This translates to the fact that every simplex encountered by $\gamma$ in fact has $v(e)$ as a vertex, as we wished to show. ∎

For any edge $e$, the union $X(e)$ of quadrants containing the ray $R(e)$ is a product $[0, \infty) \times C(e)$, where $C(e)$ is a cone on the flag complex of all sets of partitions which are compatible with $e$. The cone $C(e)$ is thus also a CAT(0) complex, consisting of a union of orthants sharing a common origin. The geodesic from $T$ to $T'$ projects to a geodesic in $C(e)$; in fact, if we know the geodesic from the projections of $T$ and $T'$ onto $C(e)$, we can

recover the geodesic from $T$ to $T'$ by following this geodesic while rescaling the length of $e$ linearly from $|e|_T$ to $|e|_{T'}$.

We conclude this section with an easily checked criterion which is sufficient to show that the cone path is *not* a geodesic:

**Notation.** Let $T \in \mathcal{T}_n$ be a tree, and $e$ an interior edge of $T$. The *norm* $\|T\|$ is the Euclidean length of the vector of branch lengths of edges of $T$, (*i.e.*, the square root of the sum of the squares of the branch lengths).

Let $E = \{e_1, \ldots, e_k\}$ be a set of edges of $T$; we denote by $T(E)$ the tree with edge set exactly $E$, with branch lengths inherited from $T$. We may also think of $T(E)$ as obtained from $T$ by collapsing every edge not in $E$. We denote by $T/E$ the tree obtained from $T$ by collapsing every edge *in* $E$.

PROPOSITION 4.3. *Suppose that $T$ and $T'$ have no edges in common, but that a set of edges $E = \{e_1, \ldots, e_k\}$ of $T$ is compatible with a set of edges $F = \{f_1, \ldots, f_l\}$ of $T'$, and that $\|T(E)\|\|T'(F)\| - \|T/E\|\|T'/F\| > 0$. Then the cone path is not a geodesic.*

*Proof.* Informally, the inequality ensures that we can produce a shorter path than the cone path by "cutting across" the orthant corresponding to the tree with edge set exactly $E \cup F$. Formally, we show that $\angle(T, T')$ is less than than $\pi$, and hence that the cone path is not a geodesic by Gromov's criterion (Gromov (1987)).

Since $T$ and $T(E)$ are in the same quadrant, the angles $\alpha = \angle(T, T(E))$ is at most $\pi/2$; similarly, and $\beta = \angle(T', T'(F)) \leq \pi/2$. Since $E$ and $F$ are disjoint but compatible, $\angle(T(E), T'(F)) = \pi/2$.(see Figure 19)

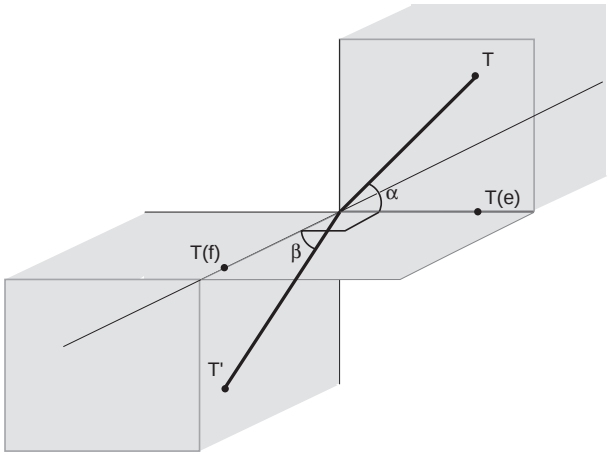


*Figure 19: The cone path from $T$ to $T'$*

The angle between $T$ and $T'$ is at most $\alpha+\pi/2+\beta$. Therefore $\angle(T,T') < \pi$ if $\alpha+\beta<\pi/2$, *i.e.*, if $\cos(\alpha+\beta)>0$. We have

$$\cos(\alpha+\beta)=\cos(\alpha)\cos(\beta)-\sin(\alpha)\sin(\beta)=\frac{\|T(E)\|}{\|T\|}\frac{\|T'(F)\|}{\|T'\|}-\frac{\|T/E\|}{\|T\|}\frac{\|T'/F\|}{\|T'\|},$$

which is positive if and only if $\|T(E)\|\|T'(F)\|-\|T/E\|\|T'/F\|>0$. ∎

### 4.3. Centroids

There are several ways of defining the center of a finite scatter of points $X$ in a CAT(0) metric space, including the center of mass, the circumcenter, and the points of maximum depth. The center of mass, defined for any probability distribution over the space, is the unique point that minimizes the expected squared distance from points in the space (see §3.2 in Jost (1997)). To define the center of mass of a finite point set, we take the uniform distribution over $X$. A clear account of this type of mean value and its properties can be found in Sturm (2000a,b). Another type of center, the circumcenter, is the center of the smallest ball enclosing the points of $X$ (see, *e.g.* Brown (1989)). The points of maximum depth (Tukey (1975)) are defined by forming the *convex hull* of $X$ (*i.e.*, the smallest set containing $X$ and containing all geodesic paths between pairs of its points), removing the extreme points (the minimal subset of $X$ having the same convex hull) and repeating until the set becomes empty.

In this section we introduce another notion of center, which we call the *centroid*. The centroid of a set of $n>2$ points is defined by iterating the operation of taking centroids of each subset of $n-1$ points, where the centroid of 2 points is defined to be the midpoint of the geodesic path that joins them. We note that for $n=3$, our construction gives the same centroid as that defined in Bruhat and Tits (1972, pp. 63-64).

It should be noted that to *compute* any of these notions of center for a finite set of points in a CAT(0) space $X$, one needs to be able to compute the geodesic paths between pairs of points, as discussed in the previous section for the space $\mathcal{T}_n$.

For $x,y\in X$, let $c(\{x,y\})$ denote the midpoint of the (unique) geodesic joining $x$ to $y$. Suppose $Y\subset X$ is a set with $n>2$ elements (some of which may be repeated), and suppose we have defined $c(W)$ for all $W\subset X$ with $|W|<n$. Then let $c^1(Y)$ denote the set $\{\ c(W)\mid W\subset Y,\ |W|=n-1\ \}$, and for $k>1$, $c^k(Y)=c^1(c^{k-1}(Y))$. Note that the sets $c^k(Y)$ all have $n$ elements (some possibly repeated).

We begin by observing that for Euclidean spaces, the sets $c^k(Y)$ can be used to find the usual centroid $c(Y)=\frac{1}{|Y|}\sum_{y\in Y}y$ of any finite set $Y$. It is straightforward to check in this case that $c(Y)=c(c^1(Y))$, and if $|Y|=n$, diam $c^1(Y)=\frac{1}{n-1}$ diam $c(Y)$. From this the following is immediate.

PROPOSITION 4.4. *If $X$ is a subset of a Euclidean space and $c(W)$ denotes the centroid of $W$, then for any finite subset $Y \subset X$, the elements in $c^k(Y)$ converge to the point $c(Y) \in X$ as $k \to \infty$.*

Our goal is to prove that the convergence in Proposition 4.4 continues to hold in an arbitrary CAT(0) space; the resulting limit point $c(Y)$ will be defined to be the centroid of the set $Y$. To do this we need to prove a general form of the convexity property that essentially defines these spaces. Suppose centroids exist for all $n$-element subsets of a CAT(0) space $X$. We say the centroid function $c(Y)$ is *convex* if whenever $Y = \{y_1, \ldots, y_n\}$ and $Y' = \{y'_1, \ldots, y'_n\}$, then $d(c(Y), c(Y')) \leq \frac{1}{n} \sum d(y_i, y'_i)$.

THEOREM 4.1. *In any CAT(0) space $X$,*

*1.centroids exist for any finite set $Y \subset X$, and*
*2.the centroid function is convex.*

*Proof.* The proof is by induction on $n = |Y|$. The case $n = 2$ is Proposition II.2.2 in Bridson and Haefliger (1999).

Suppose $n \geq 3$ and we have a convex centroid function $c(W)$ for $|W| = n-1$. Let $Y = \{y_1, \ldots, y_n\}$ and $Y_i = Y \setminus \{y_i\}$. Suppose $Y$ has diameter $D$. Then by convexity for $(n-1)$-sets, $d(c(Y_i), c(Y_j)) \leq \frac{1}{n-1}\ d(y_i, y_j) \leq \frac{1}{n-1} D$, and so diam $c^1(Y) \leq \frac{1}{n-1} D$. Thus the diameter of $c^k(Y)$ is bounded above by $\left(\frac{1}{n-1}\right)^k D$ and so goes to zero.

To show convergence, let $D_k$ denote the diameter of $c^k(Y)$, and consider the sequence $z_k \in c^k(Y)$, where $z_0 = y_1$, $z_1 = c(Y_1)$, $z_2 = c(\{c(Y_i) : i \neq 1\})$, etc. It follows by convexity for $(n-1)$-sets that $d(z_k, z_{k+1}) \leq \frac{1}{n-1} D_k$. Thus for $l \geq k$,

$$d(z_k, z_l) \leq D_k + \frac{1}{n-1}\ D_k + \left(\frac{1}{n-1}\right)^2 D_k + \cdots\ =\ \frac{n-1}{n-2}\ D_k,$$

showing that $\{z_k\}$ is a Cauchy sequence. Thus, centroids exist for $n$-sets.

To show convexity of $c(Y)$, $|Y| = n$, suppose $Y = \{y_1, \ldots, y_n\}$ and $Y' = \{y'_1, \ldots, y'_n\}$. If $Y_i = Y \setminus \{y_i\}$ and $Y'_i = Y' \setminus \{y'_i\}$, then by convexity for $(n-1)$-sets,

$$d(c(Y_i), c(Y'_i)) \leq \frac{1}{n-1} \sum_{j \neq i} d(y_j, y'_j) \tag{1}$$

for each $i$. Let $\delta_i = d(y_i, y'_i)$ and $d_0 = (\delta_1, \ldots, \delta_n)$. Then if $d_k$ is the corresponding vector of distances between elements of $c^k(Y)$ and $c^k(Y')$,

it follows from (1) that $d_k \leq B_n^k d_0$, where $B_n = \frac{1}{n-1}(J_n - I_n)$, $J_n$ is the $n \times n$ matrix of $1's$ and $I_n$ is the $n \times n$ identity matrix. Since $B_n^k \to \frac{1}{n}J$ as $k \to \infty$, it follows that $d(c(Y), c(Y')) \leq \frac{1}{n}\sum d(y_i, y_i')$ as desired. ∎

Since $\mathcal{T}_n$ is a CAT(0) space, any finite set of points has a unique centroid. If the points are all in the same orthant, *i.e.*, correspond to trees with the same combinatorial structure but possibly different branch lengths, then the centroid is the usual Euclidean centroid of the points, *i.e.*, it corresponds to the tree with the given combinatorial structure and the average of the branch lengths (see Proposition 4.4).

If two trees have all branch lengths equal to 1 but have no combinatorial structure in common, the centroid will be the origin, *i.e.*, the tree with all branch lengths equal to 0. In a large set which contains one tree which is markedly different from the others, the effect of this tree on the centroid will be negligible. An interesting effect occurs when there are duplicate trees in the set. We illustrate this by the following example: Let $T_1, T_2$ and $T_3$ be the trees illustrated in Figure 20.

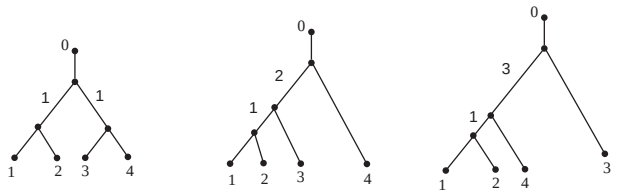


*Figure 20: Three trees*

The centroid of $\{T_1, T_2, T_3\}$ is the left tree in Figure 21, while the centroid of $\{T_1, T_1, T_2, T_2, T_3, T_3\}$ is the tree on the right. This shows a non-linear property of this definition of centroid.

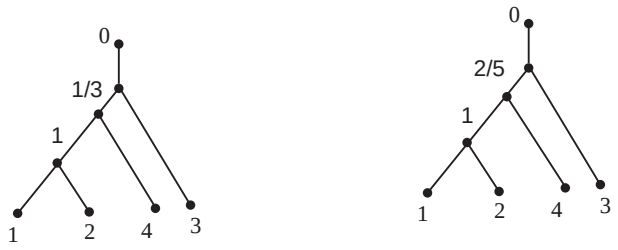


*Figure 21: Two different centroids*

This method of taking centroids provides a coherent mathematical way of forming the consensus of a set of trees. The convexity property of centroids says that given two sets $\{T_1, \ldots, T_k\}$ and $\{T_1', \ldots, T_k'\}$ of trees, the distance between the centroids will be less than or equal to the average of

the distances from $T_i$ to $T_i'$. Thus taking centroids of the two sets creates two trees which agree at least as well as the average pairwise agreement.

### 4.4. Probability Measures on Tree Space

Aldous (1996) has described several possible constructions for probability measures on combinatorial trees without branch lengths. One of the parameters he proposes to use is the *height* of the tree, which is defined as the largest number of interior edges between a leaf and the root.

There are several natural routes to complementing our geometric construction with a probability measure. They are all simplified by imposing a bound on the interior branch lengths. This has the effect of truncating each orthant to a cube.

In this section we assume the branch lengths are renormalized, to obtain unit cubes. The extension to more general compact subsets of tree space is straightforward. Here are some natural measures:

• The *base measure* called $d\tau$ puts a probability of $1/(2n-3)!!$ on each cube, while within the cube the distribution is considered uniform. Note that balls of the same radius centered at different points may have different probabilities. It is clear that for $\tau$ far from any of the boundary regions (*i.e.*, equivalently all the edges of $\tau$ sufficiently large), $d\tau$ will be proportional to the volume of a small cube around $\tau$. In this case $d\tau$ denotes the local Lebesgue measure in the cube.

If $\tau$ is a metric binary tree with exactly one small edge, then its neighborhood will meet 3 cubes. If the number of small edges is $k$ there will be at most $(2k+1)!!$ neighboring cubes for $\tau$.

For trees with $n$ leaves, the maximum volume attained is at the origin, which is contained in $2(n-3)!!$ cubes.

• If one wants to describe the simple case of a distribution concentrated around a center, then a probability distribution can be defined using the notion of distance we have developed above. This follows a Mallows' type model as developed for the symmetric group in (Mallows (1957), Diaconis (1988)) or for decision trees in Shannon and Banks (1999). In this model the central tree $\tau_0$ together with an exponential family produces a probability for any branching pattern $\tau$, defined by

$$f(\tau) = Ke^{-\lambda d(\tau,\tau_0)}d\tau.$$

The term $K$ is a normalizing constant and $\lambda$ is a concentration parameter; for $\lambda = 0$ the distribution is the base measure, and as $\lambda$ increases the measure will be more concentrated around $\tau_0$. This is a distribution concentrated around the central element $\tau_0$, which we can choose to be the centroid that we defined in section 4.3.

• Non-uniform probabilities can be constructed to agree with some information about the data; for example if one wants to be nearly sure to have a binary tree, without knowing which tree, each orthant could be given measure $1/(2n-3)!!$, but the distribution on each cube could be concentrated at the point with all branch lengths equal to 1.

Remarks:

1. In the case when the parametric maximum likelihood method has been used to determine the optimal tree, there is a natural measure on $\mathcal{T}_n$ that will result in likelihood-based confidence regions. This supposes a parametric mutation model. An example of this is provided in Billera et al. (2001).

2. Eventually our aim is to be able to map a probability on to the space so we can create isocontours determining confidence regions. Maybe a good intuitive picture is:

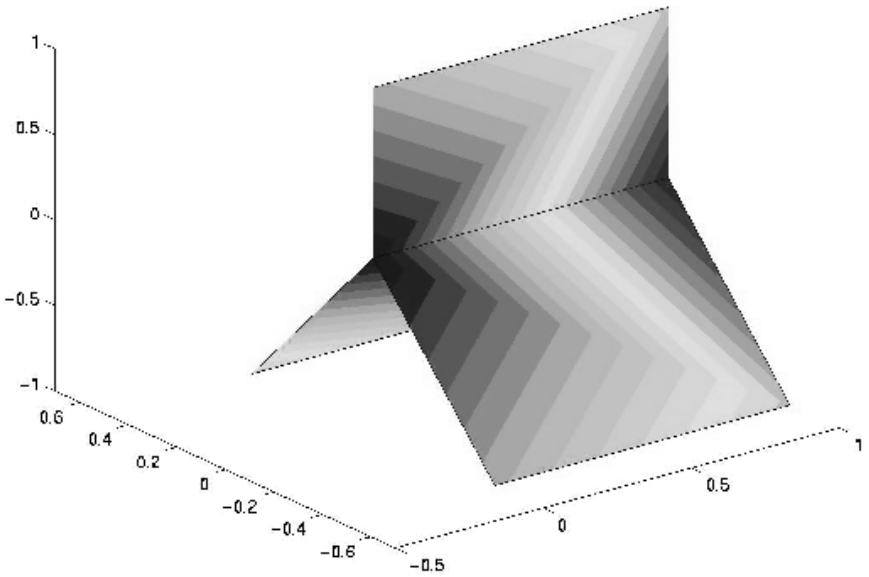


*Figure 22: A hot plot of a possible probability*

## 5. REAL DATA EXAMPLE

In this section we illustrate how the questions of averaging trees and building confidence regions in tree space come about, by examining a real data set. This data set consists of 12 mitochondrial DNA sequences, each

of length 898 bases, from 12 species of primates. This data is published in Hayasaka et al. (1988). We will use the program `dnapars`, from Felsenstein's Phylip package (available on his web site Felsenstein (1993)) to find the most parsimonious trees for these DNA sequences.

The following list shows the species names, in quotes, together with the first 80 characters of our DNA data. Data are collected by reading the DNA sequences for a specific gene occurring in all of the species. These sequences are written in rows, and the rows undergo a multiple alignment so that they have the greatest possible agreement in the columns. Here is part of the data:

```
'Lemur_catta'     AAGCTTCATAGGAGCAACCATTCTAATAATCGCACATGGCCTTACATCATCCA...
'Tarsius_syrichta'AAGTTTCATTGGAGCCACCACTCTTATAATTGCCCATGGCCTCACCTCCTCCC...
'Saimiri_sciureus'AAGCTTCACCGGCGCAATGATCCTAATAATCGCTCACGGGTTTACTTCGTCTA...
'Macaca_sylvanus' AAGCTTCTCCGGTGCAACTATCCTTATAGTTGCCCATGGACTCACCTCTTCCA...
'Macaca_fascicul.'AAGCTTCTCCGGCGCAACCACCCTTATAATCGCCCACGGGCTCACCTCTTCCA...
'Macaca_mulatta'  AAGCTTTTCTGGCGCAACCATCCTCATGATTGCTCACGGACTCACCTCTTCCA...
'Macaca_fuscata'  AAGCTTTTCCGGCGCAACCATCCTTATGATCGCTCACGGACTCACCTCTTCCA...
'Hylobates'       AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTTCCC...
'Pongo'           AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCTCCC...
'Gorilla'         AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCATCAT...
'Pan'             AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCTCAT...
'Homo_sapiens'    AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCTCAT...
```

The program `dnapars` found two different trees, each with total branch length 1163,meaning that 1163 mutations are needed to explain the DNA sequences in each tree (see Figure 23). We note that the situation of the root is unspecified *a priori*; however, it *is* known in this case to be at the Lemur branch as depicted. Simple inspection of the two trees shows that only one of its aspects seems subject to be in doubt, namely the branching between *Pan*, *Gorilla* and *Homo sapiens*.

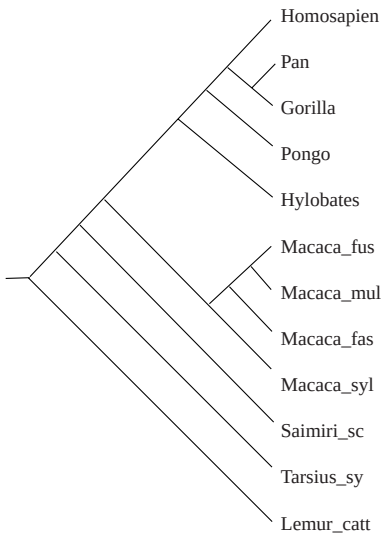


*Figure 23: First tree*

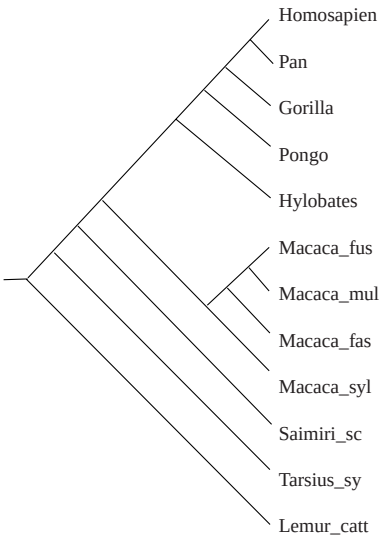


*Figure 24: Second tree*

Thus the relevant confidence statement says that we are 'sure' of all parts of the tree except for the relationship between *Homosapiens*, *Pan* and *Gorilla*. Either the first two are together or the latter two are together. Only two of the three possible subtrees with three branches are equally likely; thus a confidence region assigning equal probabilities to each of these two in a continuous way would be reasonable if no other information were available.

The fact that two different trees were produced is a result of conflicts in the data. Biologists often translate such contradictions by saying that the tree has an unresolved node and using a triple branch at this node, with `homo sapiens`, `gorilla` and `pan` all descended from a single ancestor, with no chosen two-some apparent among them. Note that the centroid of the two trees in the sense of section 4.3 in tree space is on the boundary line represented by the same unresolved tree. Thus here our notion of centroid gives a triple branch at the disputed node with `homo sapiens`, `gorilla` and `pan` all coming from a common ancestor, which is the same as the representation of uncertainty that the biologists use.

If the proportions were not fifty-fifty we would get a binary tree with non-zero edge lengths. For instance if we assigned a biological meaning to the edge lengths, such as the number of mutations along that branch, then the respective lengths given by `dnapars` on the relevant subtrees would be

those shown on the left in Figure 25. If we used the method from section 4.3 the resulting centroid tree would have the subtree on the right.

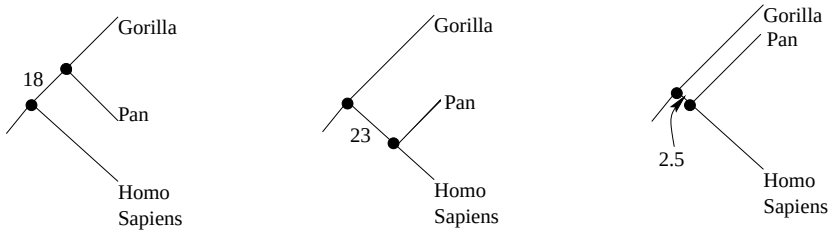


*Figure 25: Two subtrees and their centroid*

Biologists would explore the proximity to the boundary by using bootstrapping to simulate small plausible perturbations in the data. To illustrate this, we perturb the data with Phylip's `seqboot` program, thereby obtaining 100 data sets of exactly the same dimensions ($12 \times 898$). Each of these data matrices will give one or several trees.

When these trees are combined by using a majority rule consensus, (*i.e.*, which concludes that a partition is present if it is present in a majority of the trees, and labels that edge of the tree with the percentage of trees that had that particular partition), the edges are assigned a number corresponding to the frequency with which a particular partition occurred. These are interpreted by biologists as surrogate 'confidence levels' for the partitions. If a number is close to 50% this indicates a doubt as to whether the edge exists.

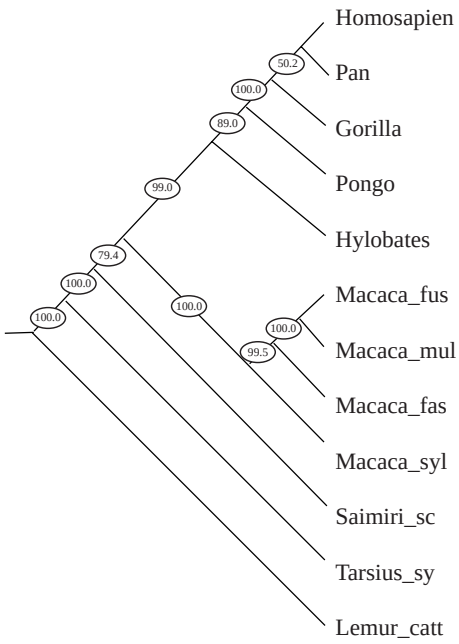


*Figure 26: Tree with confidence levels*

Such an edge-weighted tree is unsatisfactory as a summary of the perturbation analysis. A multidimensional representation would be more informative. The embedding property of the tree space will often make such a representation feasible in practice, at least approximately. In the case of Figure 26 the only notable differences that occur are on 3 edges. We can project all the trees onto a complex of three-dimensional cubes. Looking at the data this way, it is possible to further simplify since for instance the cube with the edge corresponding to the grouping of `homo sapiens` and `gorilla` does not exist. A more detailed analysis of such examples can be found in Billera et al. (2001).

## References

D. Aldous. Probability distributions on cladograms. In David Aldous and Robin Pemantle, editors, *Random discrete structures*, pages 1–18. Springer, New York, 1996.

V. Berry and O. Gascuel. Interpretation of bootstrap trees : threeshold of clade selection and induced gain. *Molecular Biology and Evolution*, 13:

999–1011, 1996.

L. Billera, C. Chan, and N. Liu. Flag complexes, labelled rooted trees, and star shellings. In B. Chazelle. J.E. Goodman and R. Pollack, editors, *Advances in discrete and computational geometry*, pages 91–102. A. M. S., Providence, RI, 1999.

L. Billera, S. Holmes, and K. Vogtmann. A geometrical perspective on the phylogenetic tree problem. Technical report, Statistics, Sequoia Hall, Stanford, CA 94305, 2001.

S. Böcker and A. W. M. Dress. Recovering symbolically dated, rooted trees from symbolic ultrametrics. *Adv. Math.*, 138(1):105–125, 1998.

M.R. Bridson and A. Haefliger. *Metric spaces of non-positive curvature*. Grundlehren der math Wiss. Springer Verlag, 1999.

D. R. Brooks. Hennig's parasitological method: A proposed solution. *Syst. Zool.*, 30:229–249, 1981.

K.S. Brown. *Buildings*. Springer Verlag, 1989.

F. Bruhat and J. Tits. Groupes réductifs sur un corps local. *Institut des Hautes Études Scientifiques*, 41:5–252, 1972.

M. Davis, T. Januszkiewicz, and R. Scott. Nonpositive curvature of blow-ups. *Selecta Math. (New Series)*, 4(4):491–547, 1998.

S. L. Devadoss. Tessellations of moduli spaces and the mosaic operad. In *Homotopy invariant algebraic structures (Baltimore, MD, 1998)*, pages 91–114. Amer. Math. Soc., Providence, RI, 1999.

P. Diaconis. *Group representations in probability and statistics*. Institute of Mathematical Statistics, Hayward, CA, 1988.

P. Diaconis and S. Holmes. Matchings and phylogenetic trees. *Proc. Natl. Acad. Sci. USA*, 95(25):14600–14602 (electronic), 1998.

J.J. Doyle. Gene trees and species trees: Molecular systematics as one-character taxonomy. *Syst. Bot.*, 17:144–163, 1992.

H. Edelsbrunner. *Algorithms in Combinatorial Geometry*. Springer Verlag, 1987.

B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA*, 93:13429–34, 1996. ISSN 1091-6490.

A. Escalante and F. Ayala. Phylogeny of the malarial genus *plasmodium* derived from rrna gene sequences. *Proc. Nat. Ac. Sciences*, 91:11371–11377, 1995.

J. Felsenstein. Statistical inference of phylogenies (with discussion). *Journal Royal Statistical Society A*, 146:246–272, 1983.

J. Felsenstein. `PHYLIP`*, (Phylogeny Inference Package) version 3.5c.* Department of Genetics, University of Washington, Seattle, version 3.5c. edition, 1993. URL `http://evolution.genetics.washington.edu/phylip.html`.

L. R. Foulds and R. L. Graham. The Steiner problem in phylogeny is NP-complete. *Adv. in Appl. Math.*, 3(1):43–49, 1982.

M. Gromov. Hyperbolic groups. In *Essays in group theory*, pages 75–263. Springer, New York, 1987.

E. Haeckel. *Generelle Morphologie der Organismen: Allgemeine Grundzuge der organischen Formen-Wissenschaft, mechanisch begrundet durch die von Charles Darwin, reformite Descendenz-Theorie*. Georg Riemer, Berlin, 1866.

K. Hayasaka, T. Gojobori, and S. Horai. Molecular phylogeny and evolution of primate mitochondrial dna. *Mol. Biol. Evol.*, 5/6:626–644, 1988.

S. Holmes. Phylogenetic trees: an overview. In *Statistics and Genetics*, number 112 in IMA, pages 81–118. Springer, New York, 1999.

J. Jost. *Nonpositive curvature: Geometric and Analytic Aspects*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, 1997.

M. M. Kapranov. The permutoassociahedron, Mac Lane's coherence theorem and asymptotic zones for the KZ equation. *J. Pure Appl. Algebra*, 85(2):119–142, 1993.

J. Krushkal and W.H. Li. Evolution of primate immunodeficiency viruses. In P. Auger and R. Jean, editors, *Advances in Mathematical Population Dynamics- Molecules, Cells and Man*, pages 5–24. World scientific, 1998.

C.W. Lee. The associahedron and triangulations of the $n$-gon. *Europ. J. Combinatorics*, 10:551–560, 1989.

J. Lin and M. Gerstein. Whole genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels,. *Genome Research*, 10(6):808–818, 2000.

L. Lovasz and M. D. Plummer. *Matching Theory*. North Holland, Amsterdam, 1985.

B.J. MacFadden. Patterns of phylogeny and rates of evolution in fossil horses: Hipparions from the Miocene and Pliocene of North America. *Paleobiology*, 1(3):245–257, 1985.

C. L. Mallows. Non-null ranking models. I. *Biometrika*, 44:114–130, 1957.

A. Robinson and S. Whitehouse. The tree representation of $\sigma_{n+1}$. *J. Pure Appl. Algebra*, 111(1-3):245–253, 1996.

E. Schröder. Vier combinatorische probleme. *Zeit. für Math. Phys.*, 15: 361–376, 1870.

W. Shannon and D. Banks. Combining classification trees using maximum likelihood estimation. *Statistics In Medicine*, 18:727–740, 1999.

D. D. Sleator, R. E. Tarjan, and W. P. Thurston. Rotation distance, triangulations, and hyperbolic geometry. *J. Amer. Math. Soc.*, 1(3): 647–681, 1988.

D. D. Sleator, R. E. Tarjan, and W. P. Thurston. Short encodings of evolving structures. *SIAM J. Discrete Mathematics*, 5(3):428–450, 1992.

R. Stanley. *Enumerative Combinatorics*. Cambridge University Press, 1999.

J.D. Stasheff. Homotopy associativity of $h$-spaces i. *Trans. Amer. Math. Soc.*, 108:275–292, 1963.

K. T. Sturm. Nonlinear markov operators, discrete heat flow, and harmonic maps between singular spaces. *Preprint from http://www-wt.iam.uni-bonn.de/homepages/sturm/sturm.html*, 2000a.

K. T. Sturm. Nonlinear martingale theory for processes with values in metric spaces of nonpositive curvature. *Preprint from http://www-wt.iam.uni-bonn.de/homepages/sturm/sturm.html*, 2000b.

J.W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, volume 2, pages 523–531, Vancouver, 1975.

K. Vogtmann. Local structure of some out($F_n$)-complexes. *Proc. Edinburgh Math. Soc. (2)*, 33(3):367–379, 1990.

M. S. Waterman and T. F. Smith. On the similarity of dendrograms. *J. Theoret. Biol.*, 73(4):789–800, 1978.

A. Zharkikh and W.H. Li. Estimation of confidence in phylogeny: The complete and partial bootstrap technique. *Mol. Phylogenet. Evol.*, 4: 44–63, 1995.

## ACKNOWLEDGMENTS

We thank Martin Bridson, Persi Diaconis, Laurent Saloff-Coste, Richard Stanley and Tom Zaslavsky for helpful discussions and an anonymous referee for constructive comments.