Tree Space and the Owens-Provan Algorithm

Jeremy Teitelbaum

University of Connecticut

January 1, 2016

Key References

BHK L. J. Billera, S. Holmes, K. Vogtmann. Geometry of the Space of Phylogenetic Trees. Advances in Applied Math. **27**, 733-767 (2001)

O M. Owen, Computing Geodesic Distance in Tree Space, preprint.

OP M. Owen, K. Provan, A fast algorithm for computing geodesic distance in tree space, IEEE/ACM Transactions on Computational Biology and Bioinformatics,**8**:1, 2–13 (2011).

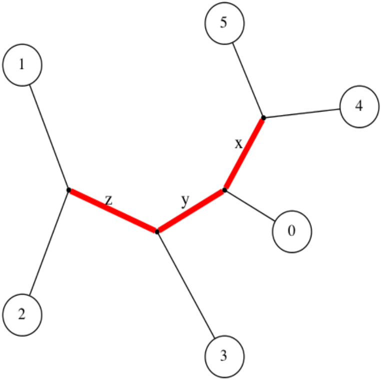
Trees

Definition

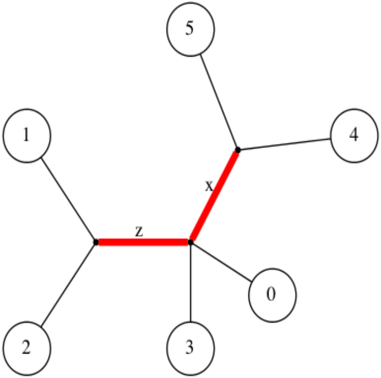
An *n*-tree is a graph with no circuits that has $n+1$ leaves (that is, vertices with only one adjacent edge).

- The leaves are labelled $0, 1, \ldots, n$ with the leaf labelled zero treated as the *root* of the tree
- In a *generic n*-tree, all of the *interior* nodes have 3 adjacent edges. There are $2n$ nodes and $2n-1$ edges. Of the edges, $n+1$ are attached to leaves and $n-2$ are *interior*.
- Every edge of the tree carries a positive length, but we will mostly ignore the lengths of the terminal edges.

Sample Trees



A Generic 5-tree

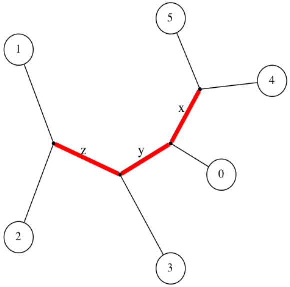


A 5-tree with a collapsed interior edge

Trees and Splits

Each edge of an *n*-tree partitions the space of leaves into two disjoint subsets A and B. We'll write $A|B$ to indicate this partition.

- The terminal edges give rise to *trivial* splits where one of A or B has only one element.
- The interior edges give rise to *non-trivial* splits.



$$x \leftrightarrow (4,5)|(0,1,2,3)$$
$$y \leftrightarrow (0,4,5)|(1,2,3)$$
$$z \leftrightarrow (1,2)|(0,3,4,5)$$

Remark

In a rooted tree, we can always choose the subset not containing the root to denote the split.

Tree topology is determined by its splits

Definition

Two splits $A|B$ and $A'|B'$ of the set $\{0, \ldots, n\}$ are called *compatible* if at least one of the four sets $A \cap B$, $A' \cap B$, $A \cap B'$, and $A' \cap B'$ is empty.

Theorem

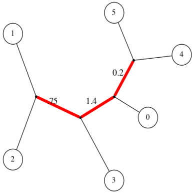
Given a set of pairwise compatible splits, there is a unique n-tree whose edges correspond to the given set of splits. [Note: it suffices to consider non-trivial splits and interior edges].

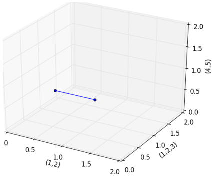
Proof.

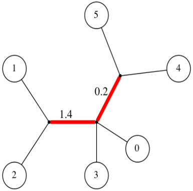
This is called the 'splits-equivalence theorem.' It is not hard to prove. You can find proofs on the web, and it is discussed in Paul Lewis's EEB5349 notes. □

Trees, Splits, and Orthants

From the splits equivalence theorem, we know that to give an *n*-tree with edge lengths is equivalent giving a non-negative valued function on the set of all splits, such that the function takes positive values on a set of pairwise compatible splits.







- We can locate a generic *n*-tree in the interior of a Euclidean orthant with the Euclidean coordinates parameterizing the length of an edge corresponding to a split.
- At the boundary of this orthant, certain interior edges of the tree collapse to zero length.

Gluing Orthants to Build TreeSpace

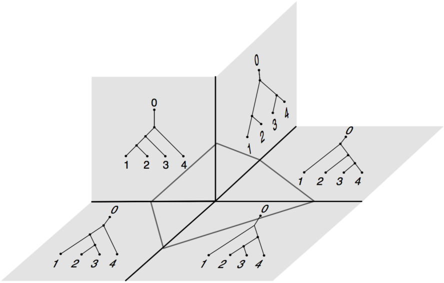


Image of part of T_4 from BHK

- Tree space T_n parameterizing n-trees is constructed by 'gluing' orthants.
- The interiors of the orthants correspond to maximal sets of pairwise compatible splits
- Orthants are glued along common faces corresponding to common subsets of compatible splits.

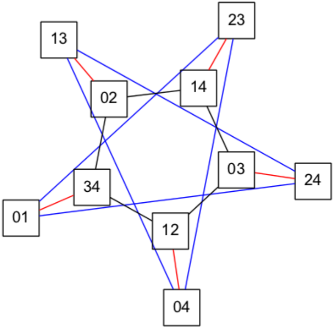
A closer look at T_4

- T_4 parameterizes trees with a total of 5 leaves and 2 internal edges. Internal edges correspond to non-trivial splits of the set $N = \{0, 1, 2, 3, 4\}$.
- A split corresponds to a choice of a two element subset of N. There are $\binom{5}{2} = 10$ such splits.
- Each tree topology arises from a compatible pair of splits. Given a split $A|B$ where A has two elements, the compatible splits are of the form $A'|B'$ where B' has three elements and contains A. This means there are $3 * 10/2 = 15$ pairs of compatible splits.
- Conclusion: T_4 has 15 2-dimensional orthants. Each orthant has two boundary rays x and y corresponding to a pair of compatible splits. Each ray, say x, meets two other orthants corresponding to the two other splits u and u' different from y and compatible with x.

The flag complex for T_4

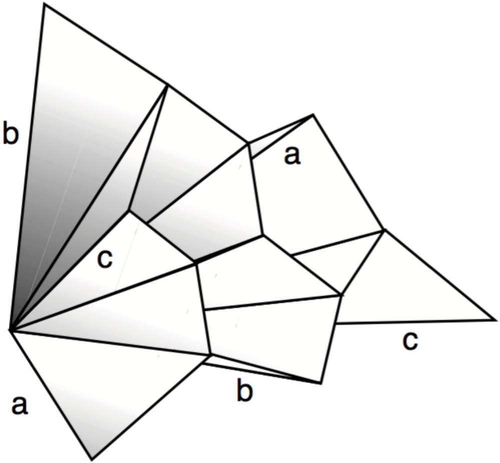
We can get a sense of the structure of T_4 by constructing a graph such that:

- the vertices correspond to the splits of $\{0, 1, 2, 3, 4\}$
- two vertices are connected by an edge if the associated splits are compatible.



(Edges correspond to 2-d orthants; vertices correspond to boundary rays.)

TreeSpace is a Cone

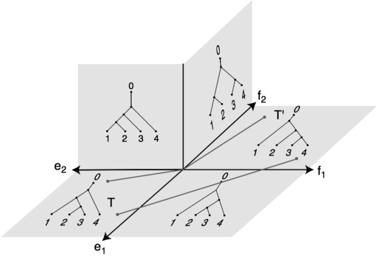


(Image from BHK)

The metric on tree space

The metric (distance function) on tree space comes from the usual Euclidean distance on the orthants.

- If two trees lie in the same orthant, the distance between them is the usual Euclidean distance.
- If two trees T and T' lie in different orthants, one can connecct them by a 'piecewise linear' path that is a straight line in each orthant it crosses. The length of such a piecewise linear path is the sum of the Euclidean lengths of its segments.



(Image from BHK)

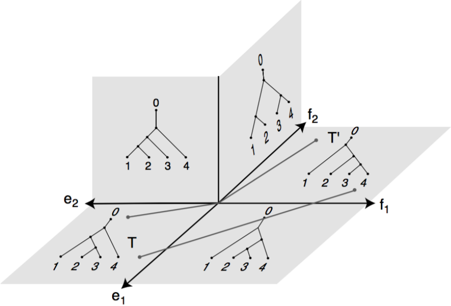
The metric, cont'd

Definition

The *distance* between T and T' is the infimum of the lengths of the piecewise linear path joining them.

Definition

Given T and T', there is a path connecting them that consists of a straight line from T to the origin, and then from the origin to T'. This is called the *cone path* and it gives an upper bound on the distance.



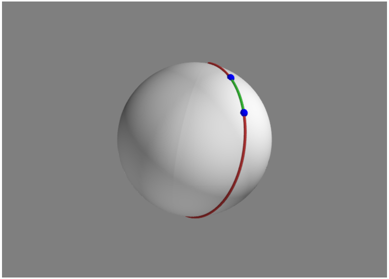
Geodesics

Definition

A locally geodesic path p between trees T and T' is a path such that, for some $\epsilon > 0$, every subpath of p of length shorter than ϵ realizes the shortest distance between its endpoints.

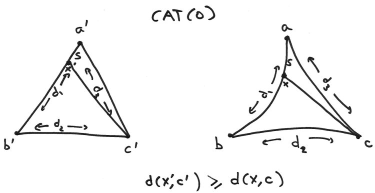
Problem

The shortest path is clearly a locally geodesic path, but in general spaces the converse isn't true. Equivalence of local geodesics and shortest paths is a global property of the space.



Global Properties of Tree Space

- Although tree space is not a manifold, it is a metric space that has the CAT(0)-property. In a CAT(0)-space, triangles with geodesic sides 'bend inwards' compared to Euclidean triangles.



- In a CAT(0) space, any two points are joined by a unique local geodesic which realizes the distance between those points.
- To establish that tree space has this property, one applies very general results of Gromov and Berestowski.

A few remarks about geodesics

- Any geodesic must be a straight line in each orthant through which it passes. All the "action" happens when crossing from one orthant to another (in other words, when some edges shrink to length zero). [This is clear].
- If an edge e appears in a tree on the geodesic from T to T', then either e is in T or e is in T'. The shortest path doesn't wander through orthants involving edges that aren't visible in the initial trees. [This is not at all obvious and follows from the general CAT(0) theory. Perhaps an elementary proof is possible?]

Combinatorial Characterization of Geodesics

Proposition

Let T and T' be two trees with no edges in common, and Suppose that p is the geodesic between them. Let $E(T)$ and $E(T')$ be the (interior) edges of T and T' respectively. Then there are partitions $\{A_1, \ldots, A_k\}$ of $E(T)$ and $\{B_1, \ldots, B_K\}$ of $E(T')$ such that:

- *for each $i > j$, A_i and B_j are compatible*
- *p passes successively through the orthants O_i, for $i = 0, \ldots, k$, where O_0 is determined by $E(T)$ and O_i for $i = 1, \ldots, k$ is determined by*

$$B_1 \cup \cdots \cup B_i \cup A_{i+1} \ldots A_k.$$

Informally: along p, one shrinks the set A_j of edges from T to zero and then grows the set B_j of edges from zero to their final length in T'.

Definition

Let A be a subset of the set $E(T)$ of edges of a tree T. Set

$$\|A\| = (\sum_{e \in A} |e|^2)^{1/2}$$

Metric Condition for Geodesics

Proposition

Suppose T and T' are trees with no common edges, and that $A_1, \ldots, A_k$ and $B_1, \ldots, B_k$ are partitions of $E(T)$ and $E(T')$ respectively such that satisfy the combinatorial condition. Suppose further that

$$\frac{\|A_1\|}{\|B_1\|} \leq \frac{\|A_2\|}{\|B_2\|} \leq \cdots \frac{\|A_k\|}{\|B_k\|}. \tag{1}$$

Then there is a path from T to T' of length

$$L = \left(\sum_{i=1}^{k} (\|A_i\| + \|B_i\|)^2 \right)^{1/2}$$

This path crosses each of the orthants $B_1 \cup B_i \cup A_{i+1} \cup A_k$ in succession and is a straight line in each such orthant.

Metric Condition (cont'd)

Consider a k-dimensional box with sides $\|A_i\| + \|B_i\|$. The diagonal of this box has length L. Starting at T:

- shrink the edge $e \in A_i$ at the rate $\lambda_e = -\frac{\|A_i\|+\|B_i\|}{\|A_i\|L}|e|$ then grow the edges $e \in B_i$ at the rate $\lambda'_e = \frac{\|A_i\|+\|B_i\|}{\|B_i\|L}|e|$.
- The edges in A_i reach length zero at the "time"

$$t_i = \frac{\|A_i\|}{\|A_i\| + \|B_i\|}L$$

  It takes an additional time

$$t'_i = \frac{\|B_i\|}{\|A_i\| + \|B_i\|}L$$

  for the edges in B_i to grow to their full length.

Metric Condition (cont'd)

- The metric condition amounts to the assertion that

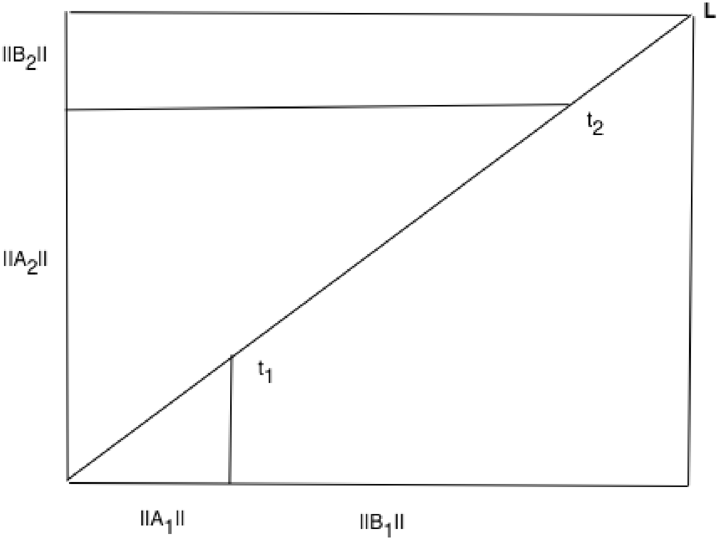
  $$t_1 \leq t_2 \leq \cdots \leq t_k.$$

  This means that the edges with non-zero lengths at any time form a compatible set and so the path does in fact stay in tree space.
- The path is in fact a straight line in each orthant since the coordinates change at constant rates.
- We have

  $$\sum_{i=1}^{j} \sum_{e \in B_i} (\lambda'_e)^2 + \sum_{i=j+1}^{k} \sum_{e \in A_i} \lambda_e^2 = 1$$

  so that this is a unit speed parameterization of the path and therefore an isometric embedding of the diagonal of the box into tree space.

Metric condition picture



Owens-Provan Characterization of Geodesics

Proposition

Suppose T and T' are trees with no edges in common and $A_1, \ldots, A_k$ and $B_1, \ldots, B_k$ are partitions of $E(T)$ and $E(T')$ satisfying the combinatorial and metric conditions. Then the corresponding path is the unique geodesic from T to T' if and only if the partitions are primitive, meaning that there is no partition C_1, C_2 and D_1, D_2 of someA_i and B_i respectively such that C_2 is compatible with D_1 and

$$\frac{\|C_1\|}{\|C_2\|} \leq \frac{\|D_1\|}{\|D_2\|}.$$

Owens-Provan Characterization (cont'd)

The idea of the proof is to study what happens at the boundary of orthants.

- If a partition exists, then one can take a "shortcut" by adding a new orthant, so the original path isn't the geodesic.
- If the path isn't the geodesic, then it must be possible to take a shortcut, which amounts to finding such a partition.

The Owens-Provan algorithm

Owens and Provan give a polynomial time algorithm to compute the geodesic from T to T'. The idea of the algorithm is:

- Reduce to the case that T and T' have no edges in common; so assume this is true.
- Start with the cone path from T to T'
- Given partitions $A_1, \dots, A_k$ and $B_1, \dots, B_k$ of the edges of T and T' satisfying the combinatorial and metric conditions, look for a partition of A_i and B_i. If you find one, refine the partitions and repeat. If not, you have constructed the geodesic.

Reduction to no common edges

Proposition (BHK)

Suppose that T and T' have an edge e in common. Let $|e|$ and $|e|'$ be the length of e in T and T' respectively. Bisect e in both T and T' and cut each tree into two, yielding trees T_0, T_1, T'_0, and T'_1 where the common edge is no longer internal. Then the distance L from T to T' is

$$L = \left(d(T_0, T'_0)^2 + d(T_1, T'_1)^2 + (|e| - |e|')^2\right)^{\frac{1}{2}}$$

Finding partitions

Given sets of edges A and B, we wish to find C_1, C_2 and D_1, D_2 so that C_2 and D_1 are compatible and so that

$$\frac{\|C_1\|}{\|C_2\|} \leq \frac{\|D_1\|}{\|D_2\|}.$$

- Construct a bipartite graph with left vertices corresponding to elements of A and right vertices corresponding to elements of B. Connect a vertex on the left to a vertex on the right whenever the corresponding edges are *incompatible*.
- Assign weight $|e|/\|A\|$ to each vertex of A and weight $|e|/\|B\|$ to each vertex of B.
- Our problem is to find an *independent set* N of vertices whose total weight is greater than one. Then $N \cap A = C_2$ and $N \cap B = D_1$.

Remark

An independent set of vertices is a set of vertices with no edges joining them.

Remark

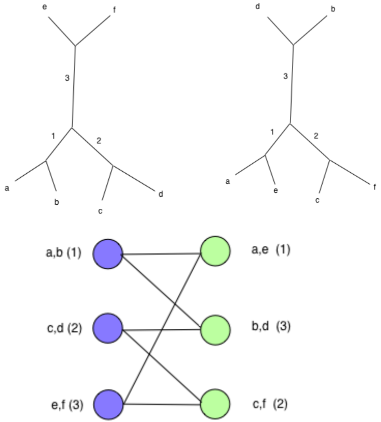
The total weight condition is:

$$\frac{\|C_2\|^2}{\|A\|^2} + \frac{\|D_1\|^2}{\|B\|^2} > 1.$$

Writing $\|A\|^2 = \|C_1\|^2 + \|C_2\|^2$ *and* $\|B\|^2 = \|D_1\|^2 + \|D_2\|^2$ *and doing some algebra yields the desired condition*

$$\frac{\|C_1\|}{\|C_2\|} \leq \frac{\|D_1\|}{\|D_2\|}.$$

Example



$$C_1 = \{(a, b), (c, d)\}, C_2 = \{(e, f)\}$$
$$D_1 = \{(b, d)\}, D_2 = \{(a, e), (c, f)\}$$

Independent sets and Vertex Covers

Definition

A vertex cover in a bipartite graph is a set of vertices V such that every edge meets V.

Lemma

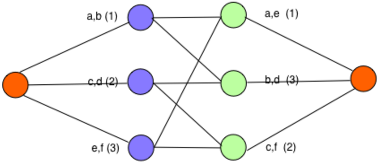
C_2 and D_1 form a maximum weight independent set in the incompatibility graph if and only if C_1 and D_2 form a minimum weight vertex cover.

Proof.

This is clear: take an edge of the graph. At least one end of it misses $C_2 \cup D_1$, so it must meet either C_1 or D_2. □

Cuts

Given a bipartite graph, construct a weighted directed graph by adding a source vertex s and edges from s to the left side of the graph, and a sink vertex t and edges from the right side of the graph to t.



A *cut* is a subset of the edges of a graph that partitions the vertices into two sets, one containing the source and the other containing the sink, and such that every path from source to sink crosses an element of the cut. *We consider only cuts that cross the 'new' edges of the graph, not the original edges.*

Vertex Covers and Cuts

Lemma

Cuts, vertex covers, and independent sets are equivalent problems.

Proof.

Given a cut, let C_2 be the set of vertices reachable from s and let D_1 be the set of vertices connected to t. Then $C_2 \cup D_1$ are an independent set and its complement $C_1 \cup D_2$ is a vertex cover. There can't be an edge from C_2 to D_1 or there would be a path from s to t. So C_2 and D_1 are independent, and therefore $C_1 \cup D_2$ is a vertex cover. □

Remark

Assign to the new edges of our graph the weight of their non-source, non-sink vertices. Then the weight of a cut is the sum of the weights of the included edges.

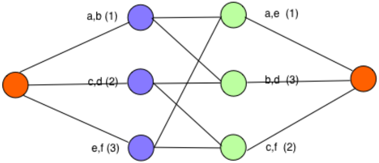
In bipartite graphs:

Finding a maximum weight independent set

Finding a minimum weight vertex cover

Finding a minimum weight cut

are equivalent problems.



Theorem

All of these graph-theoretic problems have a polynomial time solution via the max-flow/min-weight algorithm.

The Owens-Provan Algorithm

- Reduce to the case that T and T' have no edges in common; so assume this is true.
- Start with the cone path from T to T'
- Given partitions $A_1, \ldots, A_k$ and $B_1, \ldots, B_k$ of the edges of T and T' satisfying the combinatorial and metric conditions, for each $i = 1, \ldots, k$ apply the max-flow/min-cut algorithm until you find a non-trivial maximum weight independent set for A_i, B_i. If all maximum weight independent sets are trivial you are done. Otherwise, replace A_i by C_1, C_2 and B_i by D_1, D_2 and repeat this step.

Since each step in the algorithm runs in polynomial time in $E(T)$, and the loop can run at most $E(T)$ times, this algorithm runs in time polynomial in $E(T)$.

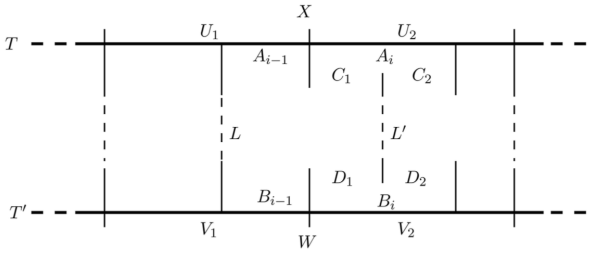
It's clear that D_1 is compatible with $C_2, A_{i+1}, \ldots A_k$ and that D_2 is compatible with $A_{i+1}, \ldots, A_k$ so the refined partition satisfies the combinatorial condition. The key element of the correctness of this algorithm is checking that the metric condition holds when the partition is refined. Since

$$\frac{\|C_1\|}{\|C_2\|} \leq \frac{\|D_1\|}{\|D_2\|}$$

is a consequence of the construction, we need to check that

$$\frac{\|A_{i-1}\|}{\|B_{i-1}\|} \leq \frac{\|C_1\|}{\|C_2\|} \text{ and } \frac{\|D_1\|}{\|D_2\|} \leq \frac{\|A_{i+1}|}{\|B_{i+1}\|}.$$

This is where we use the maximality of the weight of the independent set.



This diagram from OP is the key.

The key diagram illustrates that C_1 and C_2 arise from A_i, which in turn arose from an earlier partition $U_1 \cup U_2$, which in turn arose from an earlier partition X; with similar structure involving D_2 and D_2.

Now U_2 and V_1 are an independent set of maximal weight in X and W. On the other hand $U_2 \cup A_{i-1}$ and $V_1 \backslash B_{i-1}$ are also independent sets. Therefore

$$\frac{\|U_2\|^2}{\|X\|^2} + \frac{\|V_1\|^2}{\|W\|^2} \geq \frac{\|U_2\|^2}{\|X\|^2} + \frac{\|A_{i-1}\|^2}{\|X\|^2} + \frac{\|V_1\|^2}{\|W\|^2} - \frac{\|B_{i-1}\|^2}{\|W\|^2}$$

or

$$\frac{\|A_{i-1}\|}{\|B_{i-1}\|} \leq \frac{\|X\|}{\|W\|}$$

Similarly, $U_2 \backslash C_1$ and $V_1 \cup D_2$ are an independent set so again we have

$$\frac{\|U_2\|^2}{\|X\|^2} + \frac{\|V_1\|^2}{\|W\|^2} \geq \frac{\|U_2\|^2}{\|X\|^2} - \frac{\|C_1\|^2}{\|X\|^2} + \frac{\|V_1\|^2}{\|W\|^2} + \frac{\|D_2\|^2}{\|W\|^2}$$

or

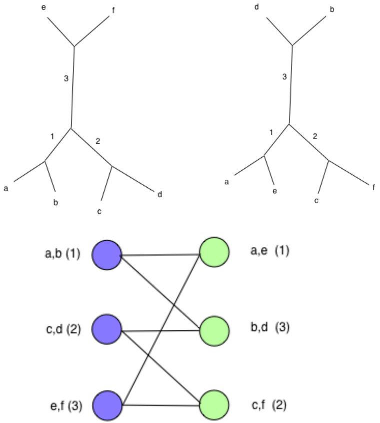
$$\frac{\|X\|}{\|W\|} \leq \frac{\|C_1\|}{\|D_2\|}$$

Looking to the right in the diagram yields a symmetric calculation giving the other necessary inequality.

Conclusion

The Owens-Provan algorithm computes a geodesic between T and T' in polynomial time.

Example



$$C_1 = \{(a,b),(c,d)\}, C_2 = \{(e,f)\}$$
$$D_1 = \{(b,d)\}, D_2 = \{(a,e),(c,f)\}$$

$$L = 3\sqrt{2} + \sqrt{10} = 7.4049$$