

On the Midpath Tree Conjecture: A Counter-Example

Rahul Shah*

Martin Farach-Colton†

Introduction. Clustering data based on pairwise distances is a fundamental problem. Weighted trees can be used to represent hierarchical clusters. Multidimensional Scaling (MDS), in some formulations, is the problem of clustering data based on preserving their *relative* (rather than absolute) distances. We call this an *ordinal clustering*.

A particular instance of this problem has been considered in the algorithmic computational biology community [KW, KHM]: Given a (total or partial) order on the pairwise distances between points, give a weighted tree on those points so that the pairwise pathlengths between points in the tree satisfies this order. This is, therefore, simply the MDS problem for hierarchical clustering, and we will refer to it as the HMDS problem.

It has been conjectured [Kan] that there is a polynomial time algorithm to solve the HMDS problem, both while preserving the total order of the points, as well as while relaxing the order to certain types of partial orders. This algorithm is called the *Midpath Tree Algorithm*. While it clearly runs in polynomial time, it was not known to always produce the correct output for HMDS. In this short paper, we show that it does not, in fact, always produce the correct output.

Previous work includes an algorithm described by Kearney et al in [KHM] to construct an *unweighted* tree, if it exists, which realizes the total order on pairwise distances. Clearly, in many cases an unweighted tree may not exist for an order, while a weighted tree does exist, so solving the unweighted case sheds little light on the weighted HMDS problem. Kannan and Warnow [KW] solved a similar problem to realize certain types of partial orders and posed the weighted case as an open problem. They specifically worked with a partial order constructed from triplets of data points which we call a *triangle order*. A triangle order is a partial order on distances so that that distances within each triplet of points are totally ordered.

Before presenting the Midpath Tree Algorithm and its counter-example, we present some preliminary observations.

Realizability and LP's. Define d_{T_w} to be the distance metric of tree T under a non-negative weight function w , where $d_{T_w}(s, t)$ is sum of weights of edges on the unique path $P_T(s, t)$ from leaf s to leaf t in T .

A partial order P on pairwise distances is said to be *realizable* as tree T if, for some weight function w on the edges of T , we get $d_{T_w}(a, b) < d_{T_w}(c, d)$ whenever $d(a, b) <_P d(c, d)$ ¹.

Given a tree T , we can determine in polynomial time via linear programming whether or not the partial (or total) order P can be realized as T by checking the feasibility of the order constraints

$$d_{T_x}(c, d) - d_{T_x}(a, b) \geq \delta \text{ if } d(a, b) <_P d(c, d) \quad (1)$$

$$x \geq 0 \quad (2)$$

where $\delta > 0$ is a constant.

Contractions and Expansions. A *contraction* of a tree T at the edge pq is the tree that results from removing edge pq from T and identifying vertices p and q . An *expansion* of T at a vertex v is the inverse operation of contraction. A tree T' is called a contraction of T if T' is obtained by the contraction of T at some edge, or if T' is a contraction of some contraction of T . T' is called an expansion of T if T is a contraction of T' .

Midpath Trees. Given a tree T which realizes a triangle (or total) order Δ on pairwise distances between points in set S , along with a weight function w , consider a function $m : S \times S \rightarrow E(T)$ which maps each pair of leaves a, b to the edge $m(a, b)$ on which the midpoint of the weighted path $P_T(a, b)$ falls. Now, contract all the (non-leaf) edges in T which do not have any midpoints falling on them to obtain an unweighted tree which we call the *midpath tree* T_Δ . The *midpath function* m for T_Δ is borrowed from T . For any pair $a, b \in S$, let $T_\Delta^{a/b}$ and $T_\Delta^{b/a}$ be the connected components of $T_\Delta - m(a, b)$ containing a and b respectively. Then, $c \in T_\Delta^{a/b} \Leftrightarrow d(a, c) <_\Delta d(b, c)$. The midpoint edge $m(a, b)$ gives a bipartition of points in S based on whether they are closer to a or to b . This means the midpath tree T_Δ and midpath function m (weight function not required) can be used to represent a unique triangle (*not* total) order. The midpath tree is a minimal tree on which midpath function satisfying such a bipartition property can be defined.

Clearly, the existence of the midpath tree is a prerequisite for the existence of a tree T which realizes the

*Dept. of CS, Rutgers University, sharahul@paul.rutgers.edu

†Google Inc., martin@google.com

¹For the sake of simplicity and conciseness, we shall only consider orders with strict inequalities

