AXIOMATIC HIERARCHICAL CLUSTERING GIVEN INTERVALS OF METRIC DISTANCES

Weiyu Huang *Alejandro Ribeiro*

Department of Electrical and Systems Engineering, University of Pennsylvania

ABSTRACT

This paper examines metric spaces in which the distance between any pair of nodes is given by an interval. The goal is to investigate methods for hierarchical clustering, i.e., a family of nested partitions indexed by a connectivity parameter, deduced from the underlying distance intervals of the metric spaces. Our construction is based on designing admissible methods abiding to the axioms of value and transformation. Two admissible methods are constructed and are shown to provide upper and lower bounds in the space of all admissible methods. Practical implications are explored by clustering moving points via snapshots. The proposed clustering methods succeed in identifying underlying clustering structures via the maximum and minimum distances in all snapshots.

***Index Terms*—** Clustering, hierarchical clustering, axiomatic clustering, network theory, network science, metric spaces.

1. INTRODUCTION

We often encounter datasets representing points in a metric space but in which the computation of exact distances between points is intractable. When this happens it is customary to resort to tractable lower and upper bounds. This is the case when, e.g., the points themselves represent individual unlabeled networks. The space of networks can be endowed with a metric that is computationally intractable because unlabeled networks are invariant to permutations [2, 3]. However, upper bounds are readily available by looking at specific permutations and lower bounds can be computed using homological features [4, 5]. In this paper we study hierarchical clustering methods for problems of this form. I.e., we want to hierarchically cluster points in a metric space in which the distances between pairs of points are only known to belong to some interval.

The approach we take is axiomatic in nature and builds on the increasingly strong theoretical understanding of clustering methods [6–13]. Our particular interest here is in hierarchical clustering where instead of a single partition, we search for a family of partitions indexed by a connectivity parameter, e.g., [14–16]. It has been proved in [11] that single linkage [15, Ch. 4] is the unique hierarchical clustering method that abides to three reasonable axioms. These results were later extended to asymmetric networks not necessarily metric and the number of of axioms required for unicity results reduced to only two [11, 13]. In the case of metric spaces the two properties that are imposed as axioms in [13] can be intuitively stated as: (A1) The nodes in a network with two nodes are clustered at the resolution specified by their distance. (A2) A network that is uniformly dominated by another should have clusters that are also uniformly dominated. Property (A1) is dubbed the Axiom of Value and property (A2) the Axiom of Transformation. The goal of this paper is to extend the axiomatic construction of hierarchical clustering in [11, 13] for clustering based on distance intervals. Clustering methods that attempt to take uncertainty into consideration include the construction of models to replicate the properties of uncertainties in the data [17–19] as well as the consideration of multiple observations of points given in a Euclidean space [20–23]. Our work differs in that we investigate situations where the only available information are the upper and lower bounds of the actual metric distances. This can be considered as a more crude observation and a generalization of the approaches in [17–23].

Supported by NSF CCF-1217963 and AFOSR MURI FA9550-10-1-0567. Previous results appeared in [1].

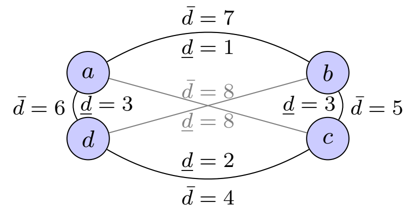


Fig. 1. An example of metric space where distances between pairs of nodes are given in lower and upper bounds.

This paper aims to continue the axiomatic hierarchical clustering previously explored in [11, 13] for clustering of metric spaces in which distances are given by intervals, and to impose desired properties that one expects for rational methods. With these properties, we proceed to characterize the space of methods that are admissible with respect to them and apply the methods to cluster moving points via their snapshots.

2. PRELIMINARIES

We consider a metric space M_X to be a pair (X, d_X) where X is a finite set of nodes and $d_X : X \times X \to \mathbb{R}_+$ is a metric distance. In specific, $d_X(x, x')$ between nodes x and x' is assumed to be nonnegative, is symmetric such that $d_X(x, x') = d_X(x', x)$, and is 0 if and only if the nodes coincide with $x = x'$; d_X also satisfies triangle inequality with $d_X(x, x'') \leq d_X(x, x') + d_X(x', x'')$ for any triplets $x, x', x'' \in X$. The interest of study in this paper is not on the metric space M_X, but in situations where observation of $d_X(x, x')$ is only given in an interval. Formally, we consider I_X as the triplet $(X, \bar{d}_X, \underline{d}_X)$ where given a pair of nodes $x, x' \in X$, we have the relationship $0 < \underline{d}_X(x, x') \leq d_X(x, x') \leq \bar{d}_X(x, x')$. The bounds $\underline{d}(x, x')$ and $\bar{d}(x, x')$ between nodes $x, x' \in X$ are nonnegative for all pairs and 0 if and only if $x = x'$; moreover, they are symmetric, i.e. $\underline{d}(x, x') = \underline{d}(x', x)$ and similarly for $\bar{d}(x, x')$. However, they may not necessarily satisfy the triangle inequality. We define $\mathcal{I}$ as the set of all metric spaces where the actual distance is observed in a confidence interval. Entities in $\mathcal{I}$ may have different node sets X as well as different distance lower or upper bounds.

An example metric space with distance given by intervals is shown in Fig. 1. The smallest nontrivial case with nodes p and q and distance bounds $\underline{d}(p, q) = \underline{d}$ and $\bar{d}(p, q) = \bar{d} \geq \underline{d} > 0$ is described in Fig. 2. We define the two-node space $\Delta_2(\underline{d}, \bar{d})$ with bounds $\underline{d}$ and $\bar{d}$ as

$$\Delta_2(\underline{d}, \bar{d}) := (\{p, q\}, \underline{d}, \bar{d}). \tag{1}$$

A clustering of the set X denotes a partition P_X of X — a collection of sets $P_X = \{S_1, \ldots, S_J\}$ with $S_i \cap S_j = \emptyset$ for any $i \neq j$ covering X, $\cup_{j=1}^{J} S_j = X$. The sets $S_1, \ldots, S_J$ are named the clusters of P_X. An equivalence relation $\sim$ on X is a binary relation such that for all $x, x', x'' \in X$ we have that (1) $x \sim x$, (2) $x \sim x'$ if and only if $x' \sim x$, and (3) $x \sim x'$ and $x' \sim x''$ would imply $x \sim x''$. A partition $P_X = \{S_1, \ldots, S_J\}$ of X always induces and is induced by an equivalence relation $\sim_{P_X}$ on X where for all $x, x' \in X$ we have that $x \sim_{P_X} x'$ if and only if x and x' is cluttered to the same set S_j for some j.

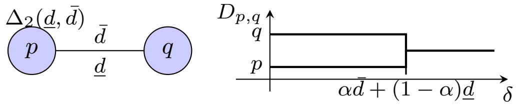


Fig. 2. Two-node space $\Delta_2(\bar{d}, \underline{d})$ and the Axiom of Value: nodes are clustered at the combination of the distance bounds.

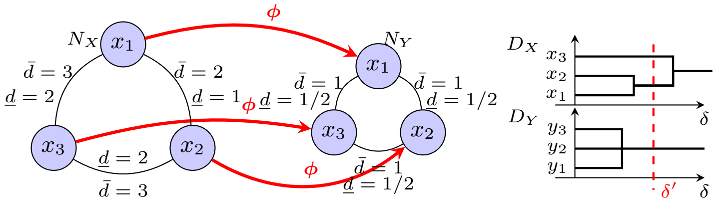


Fig. 3. Axiom of Transformation. If I_X can be mapped to I_Y using a α-distance-reducing map ϕ, then for every δ nodes clustered together in $D_X(\delta)$ must also be clustered in $D_Y(\delta)$.

In this paper we focus on hierarchical clustering methods [11, 13]. The output of hierarchical clustering methods is not a single partition P_X but a nested collection D_X of partitions $D_X(\delta)$ of X indexed by the resolution parameter $\delta \geq 0$. In the language of equivalence relation, for a given D_X, we say that two nodes x and x' are equivalent at resolution δ with notation $x \sim_{D_X(\delta)} x'$ if and only if nodes x and x' are in the same cluster of $D_X(\delta)$. The nested collection $\mathcal{D}_X$ is named a *dendrogram*.

Dendrograms are difficult to analyze. A more convenient representation is acquired when dendrograms are identified with finite ultrametric spaces. An ultrametric on the space X is a metric $u_X : X \times X \to \mathbb{R}_+$ satisfying the stronger triangle inequality such that for any points $x, x', x'' \in X$, $u_X(x, x'')$, $u_X(x, x')$, and $u_X(x', x'')$ abide to

$$u_X(x, x'') \leq \max\left(u_X(x, x'), u_X(x', x'')\right). \tag{2}$$

We investigate ultrametrics because it preserves the structure of dendrograms [11]. Given a dendrogram, its corresponding ultrametric is $u_X(x, x') := \min\left\{\delta > 0 : x \sim_{D_X(\delta)} x'\right\}$; given an ultrametric, its associated equivalence relationship $\sim_{u_X(\delta)}$ is $x \sim_{u_X(\delta)} x' \Leftrightarrow u_X(x, x') \leq \delta$. From this equivalence, we consider hierarchical clustering method $\mathcal{H}$ as a map $\mathcal{H} : \mathcal{I} \to \mathcal{U}$ to the space of ultrametrics.

The notions of chain and chain costs are substantial in the development. Given a metric space with distance intervals $(X, \underline{d}_X, \bar{d}_X)$ and a pair $x, x' \in X$, a *chain* from x to x' is an ordered sequence of nodes in X, i.e. $[x = x_0, x_1, \ldots, x_{l-1}, x_l = x']$, which begins with x and ends at x'. We denote $C(x, x')$ as one such chain and say $C(x, x')$ connects x to x'. Given two chains $C(x, x')$ and $C(x', x'')$ such that the end point x' of the first chain is the same as the starting point of the second, we define the *concatenated chain* $C(x, x') \uplus C(x', x)$ as $[x = x_0, x_1, \ldots, x_{l-1}, x_l = x' = x'_0, x'_1, \ldots, x'_{l'} = x'']$. For $C(x, x')$, we define its *upper cost* and *lower cost* as

$$\max_{i | x_i \in C(x,x')} \bar{d}_X(x_i, x_{i+1}), \qquad \max_{i | x_i \in C(x,x')} \underline{d}_X(x_i, x_{i+1}), \tag{3}$$

The *minimum upper chain cost* $\bar{c}(x, x')$ and the *minimum lower chain cost* $\underline{c}(x, x')$ between a pair x and x' is then defined respectively as the minimum upper and lower cost among all chains connecting x to x',

$$\bar{c}_X(x, x') := \min_{C(x,x')} \max_{i | x_i \in C(x,x')} \bar{d}_X(x_i, x_{i+1}), \tag{4}$$

$$\underline{c}_X(x, x') := \min_{C(x,x')} \max_{i | x_i \in C(x,x')} \underline{d}_X(x_i, x_{i+1}). \tag{5}$$

The minimum upper and lower chain costs are different in general, however they are equal in the degenerate case where distance bounds coincide with $\underline{d}_X(x, x') = \bar{d}_X(x, x') := d_X(x, x')$ for any $x, x' \in X$. In this case, the minimum cost $\bar{c}_X(x, x') = \underline{c}_X(x, x')$ are important in the construction of the single linkage [11]. In specific, single linkage ultrametric $u_X^{\mathrm{SL}}(x, x')$ between x and x' is

$$u_X^{\mathrm{SL}}(x, x') = \min_{C(x,x')} \max_{i | x_i \in C(x,x')} d_X(x_i, x_{i+1}). \tag{6}$$

It can be seen that $\bar{c}_X$ is the result of applying single linkage towards the node set X equipped with dissimilarity $\bar{d}_X$ despite the fact that $\bar{d}_X$ may not be a valid metric; similar result holds for $\underline{c}_X$. In the degenerative case where distance lower bounds and upper bounds coincide, it is equivalent to consider metric spaces (X, d_X). It has been shown [11] that single linkage is the unique hierarchical clustering method fulfilling axioms (A1) and (A2) discussed in Section 3. In the case when the dissimilarity $d_X(x, x')$ is unknown but given in an interval $[\underline{d}_X(x, x'), \bar{d}_X(x, x')]$ instead, the space of methods satisfying axioms (A1) and (A2) and their analogous ones becomes richer, as we explain throughout the paper.

3. AXIOMS OF VALUE AND TRANSFORMATION

To study hierarchical clustering methods on metric spaces where dissimilarities are given in distance intervals, we translate resonable intuitions into the axioms of value and transformation, described in this section. We say a hierarchical clustering method $\mathcal{H}$ is *admissible* if and only if it satisfies both the the axioms of transformation and value.

The Axiom of Value is achieved by considering the two-node space $\Delta_2(\underline{d}, \bar{d})$. In the degenerate special case where $\underline{d} = \bar{d} := d(p, q)$, it is apparent that the resolution at which nodes p and q are first clustered together should be $d(p, q)$. In general scenarios where the dissimilarity $d(p, q)$ is given in an interval $[\underline{d}, \bar{d}]$ with $\underline{d} < \bar{d}$, we say that nodes p and q form a single cluster first at resolution $\delta := \alpha\bar{d} + (1-\alpha)\underline{d}$, the convex combination of the upper and lower bounds $\bar{d}$ and $\underline{d}$. Property of hierarchical clustering then indicates nodes p and q are clustered together at any resolution $\delta \geq \alpha\bar{d} + (1-\alpha)\underline{d}$. The parameter α controls the level of confidence in examining the distance intervals. A higher value of α implies a more conservative consideration, where in the extreme case with $\alpha = 1$, nodes p and q are clustered together at the distance upper bound $\bar{d}$. We formalize this intuition as next.

(A1) Axiom of Value. Given $0 \leq \alpha \leq 1$, the ultrametric output $(\{p, q\}, u_{p,q}) = \mathcal{H}(\Delta_2(\underline{d}, \bar{d}))$ resulted from applying $\mathcal{H}$ upon the two-node space $\Delta_2(\underline{d}, \bar{d})$ satisfies that $u_{p,q}(p, q) = \alpha\bar{d} + (1-\alpha)\underline{d}$.

The second requirement on the space of desired methods $\mathcal{H}$ formalizes the intuition for the behavior of $\mathcal{H}$ when considering a transformation on the distance bounds on the underlying space X; see Fig. 3. Consider $I_X = (X, \underline{d}_X, \bar{d}_X)$ and $I_Y = (Y, \underline{d}_Y, \bar{d}_Y)$ and denote $D_X = \mathcal{H}(X, \underline{d}_X, \bar{d}_X)$ and $D_Y = \mathcal{H}(Y, \underline{d}_Y, \bar{d}_Y)$ as the corresponding dendrogram outputs. If we can map all the nodes of the triplet $(X, \underline{d}_X, \bar{d}_X)$ into nodes of $(Y, \underline{d}_Y, \bar{d}_Y)$ such that the combination of lower and upper bounds for any pair of nodes is not increased, we expect the latter metric distance intervals to be more clustered than the former one at any given resolution. Intuitively, nodes in I_Y are less dissimilar with respect to each other, and therefore at any resolution δ in the respective dendrograms, we expect that for nodes that are clustered in I_X, their corresponding nodes in Y are also clustered in I_Y. To formalize this intuition, we introduce the following notion that given $I_X = (X, \underline{d}_X, \bar{d}_X)$, $I_Y = (Y, \underline{d}_Y, \bar{d}_Y)$, and a value $0 \leq \alpha \leq 1$, the map $\phi : X \to Y$ is called *α-distance-reducing* if for any $x, x' \in X$, it holds that

$$\hat{d}_X(x, x') \geq \hat{d}_Y(\phi(x), \phi(x')), \quad \hat{c}_X(x, x') \geq \hat{c}_Y(\phi(x), \phi(x')), \tag{7}$$

where we define $\hat{d}_X(x, x') := \alpha\bar{d}_X(x, x') + (1-\alpha)\underline{d}_X(x, x')$ and similarly $\hat{c}_X(x, x') := \alpha\bar{c}_X(x, x') + (1-\alpha)\underline{c}_X(x, x')$. A mapping is α-distance-reducing if both the combinations of distance bounds and chain costs are non-increasing. In the degenerate case where distance lower and upper bounds coincide, $u_X^{\mathrm{SL}}(x, x') := \bar{c}_X(x, x') = \underline{c}_X(x, x')$ is the output of applying single linkage upon the metric space. Therefore the first inequality in (7) becomes identical as $d_X(x, x') \geq$

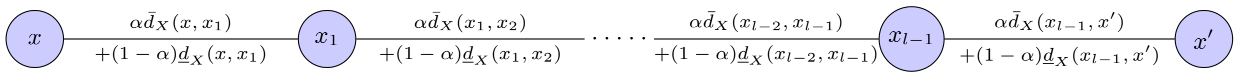


Fig. 4. Combine-and-cluster clustering. Nodes x and x' are clustered together at resolution δ if there exists a chain such that the maximum convex combination of distance bounds $\hat{d}_X(x_i, x_{i+1})$ is no greater than δ [cf. (9)].

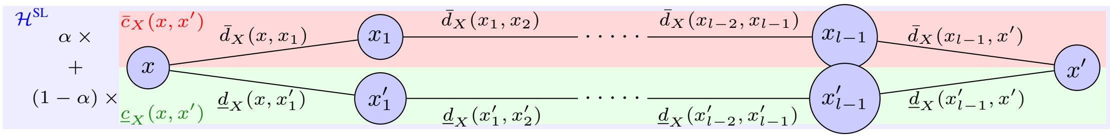


Fig. 5. Cluster-and-combine clustering. Nodes x and x' are clustered together at resolution δ if there exists a chain such that the maximum convex combination $\hat{c}_X(x,x')$ of minimum upper and lower chain costs is no greater than δ [cf. (10)].

$d_Y(\phi(x), \phi(x'))$, from which $c_X(x,x') \geq c_Y(\phi(x), \phi(x'))$ follows directly. In general cases, $\hat{c}_X(x,x') \geq \hat{c}_Y(\phi(x), \phi(x'))$ does not follow from $\hat{d}_X(x,x') \geq \hat{d}_Y(\phi(x), \phi(x'))$. The Axiom of Transformation is a formal statement of the intinction.

(A2) Axiom of Transformation. Consider $I_X = (X, \underline{d}_X, \bar{d}_X)$ and $I_Y = (Y, \underline{d}_Y, \bar{d}_Y)$ and a given α-distance-reducing map $\phi: X \to Y$. For any pair of nodes $x, x' \in X$, the output ultrametrics $u_X = \mathcal{H}(X, \underline{d}_X, \bar{d}_X)$ and $u_Y = \mathcal{H}(Y, \underline{d}_Y, \bar{d}_Y)$ satisfy $u_X(x,x') \geq u_Y(\phi(x), \phi(x'))$.

In summary, Axiom (A1) specifies our tendency in believing lower or upper bounds. Axiom (A2) states that if we reduce both the distance bounds, clusters may be combined but cannot be separated. These axioms are an adaption of the axioms proposed in [11] for the degenerate case of $\underline{d}_X = \bar{d}_X$, and the axioms in [13] for asymmetric networks.

In the degenerate case where distance lower and upper bounds coincide, another intuitive idea in clustering is that no clusters should be formed at resolutions smaller than the smallest dissimilarity in the metric space. To generalize this idea to scenarios where the bounds differ, we define α-separation $s_X^\alpha(x,x')$ between two different nodes $x, x' \in X$ in a metric space with distances given by intervals $(X, \underline{d}_X, \bar{d}_X)$ as

$$s_X^\alpha(x,x') = \alpha\bar{c}_X(x,x') + (1-\alpha)\underline{c}_X(x,x'). \tag{8}$$

The *α-separation* for $(X, \underline{d}_X, \bar{d}_X)$ is then defined as $\text{sep}^\alpha(X, \underline{d}_X, \bar{d}_X) := \min_{x \neq x'} s_X^\alpha(x,x')$. In the degenerate case we would have $\text{sep}^\alpha(X, \underline{d}_X, \bar{d}_X) = \text{sep}(X, d_X)$ for any α. Following the notion of separation, for resolutions $0 \leq \delta < \text{sep}^\alpha(X, \underline{d}_X, \bar{d}_X)$, no nodes should be clustered together. This implies that we must have $u_X(x,x') \geq \text{sep}^\alpha(X, \underline{d}_X, \bar{d}_X)$ for any pair of different nodes $x \neq x' \in X$ as we state next.

(P1) Property of Minimum Separation. For $(X, \underline{d}_X, \bar{d}_X)$, the output ultrametric $(X, u_X) = \mathcal{H}(X, \underline{d}_X, \bar{d}_X)$ of the hierarchical clustering method $\mathcal{H}$ needs to satisfy that the ultrametric $u_X(x,x')$ between any pair x and x' cannot be smaller than the α-separation $\text{sep}^\alpha(X, \underline{d}_X, \bar{d}_X)$.

Notice that if we apply (P1) onto the two-node space $\Delta_2(\underline{d}, \bar{d})$, we must have $u_{p,q}(p,q) \geq \alpha\bar{d} + (1-\alpha)\underline{d}$, which means that (P1) and (A1) are compatible. We can therefore construct two axiomatic formulations where admissible methods are required to satisfy the (A2) as well as (P1), or (A2) as well as (A1). As we demonstrate as next that (P1) is implied by (A2) and (A1). Therefore, the two formulations are equivalent.

Theorem 1 *If a hierarchical clustering method satisfies the Axiom of Value (A1) and Axiom of Transformation (A2), it satisfies the Property of Minimum Separation (P1).*

Proof: We refer readers to [24] for details of proofs in the paper. ■

4. ADMISSIBLE ULTRAMETRICS

Consider a specific metric space with distances given by intervals $I_X = (X, \underline{d}_X, \bar{d}_X) \in \mathcal{I}$. Given a value $0 \leq \alpha \leq 1$, one particular clustering method satisfying axioms (A1) and (A2) can be established by examining the *α-combined* dissimilarity $\hat{d}_X(x,x')$ for all pair of nodes $x, x' \in X$. Though $\hat{d}_X$ does not necessarily satisfy the triangle inequality as the original metric distance d_X, it is symmetric; therefore the α-combined dissimilarity effectively reduces the problem to clustering of symmetric data, a case where the single linkage method defined in (6) is shown to abide to axioms analogous to (A1) and (A2) [11]. Based on this observation, we define the *combine-and-cluster* method $\mathcal{H}^{\text{CO}}$ with output $(X, u_X^{\text{CO}}) = \mathcal{H}^{\text{CO}}(X, A_X)$ between a pair x and x' as

$$u_X^{\text{CO}}(x,x') := \min_{C(x,x')} \max_{i | x_i \in C(x,x')} \hat{d}_X(x_i, x_{i+1}). \tag{9}$$

An illustration of the combine-and-cluster clustering method is shown in Fig. 4. For a given pair x and x', we look for chains $C(x,x')$ connecting them. For the chain we examine each of its link, connecting say x_i with x_{i+1}, and investigate the convex combination of the distance bounds, i.e. $\hat{d}_X(x_i, x_{i+1}) = \alpha\bar{d}_X(x_i, x_{i+1}) + (1-\alpha)\underline{d}_X(x_i, x_{i+1})$. The maximum value across all links in this chain is then recorded. The combine-and-cluster ultrametric $u_X^{\text{CO}}(x,x')$ between points x and x' is the minimum of this value across all possible chains connecting x and x'.

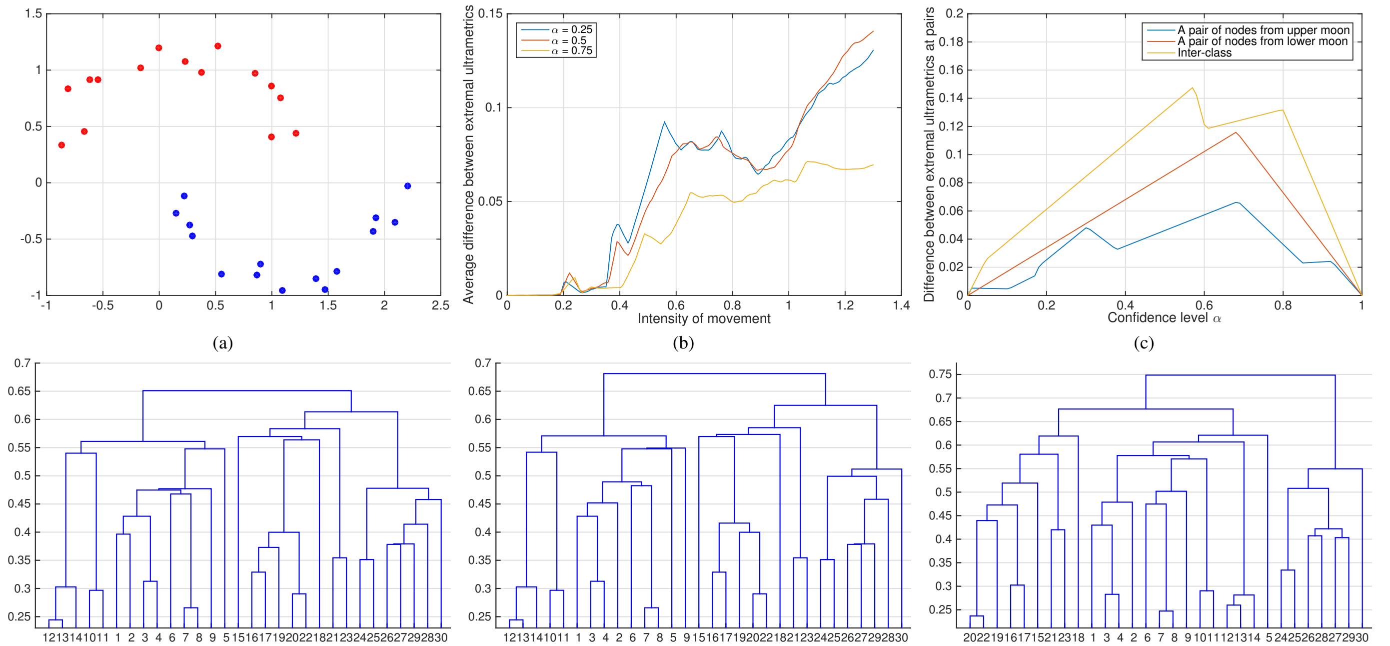
In combine-and-cluster clustering, nodes x and x' belong to the same cluster at resolution δ whenever we can find a single chain such that the maximum convex combination of distance bounds is no greater than δ. In *cluster-and-combine* clustering, we switch the order of operations and investigate chains, potentially different, connecting x and x', with one chain focusing on the distance upper bounds and the other chain examining the distance lower bounds, before combining the upper and lower estimations. To state this definition regarding ultrametrics, consider $I_X = (X, \underline{d}_X, \bar{d}_X)$ and $0 \leq \alpha \leq 1$. We define the cluster-and-combine method $\mathcal{H}^{\text{CL}}$ with output $(X, u_X^{\text{CL}}) = \mathcal{H}^{\text{CL}}(X, \underline{d}_X, \bar{d}_X)$ as

$$u_X^{\text{CL}}(x,x') := \min_{C(x,x')} \max_{i | x_i \in C(x,x')} \hat{c}_X(x_i, x_{i+1})), \tag{10}$$

An illustration of the cluster-and-combine clustering method is described in Fig. 5. For any pair of nodes, we consider the convex combination $\hat{c}_X(x,x')$ of minimum chain costs. The output of the cluster-and-combine clustering method is the result by applying single linkage $\mathcal{H}^{\text{SL}}$ [cf. (6)] onto the convex combination $\hat{c}_X(x,x')$. The single linkage is applied towards $\hat{c}_X(x,x')$ because convex combination of ultrametrics is a metric but not necessarily an ultrametric. We demonstrate that the output u_X^{CO} and u_X^{CL} are valid ultrametrics and the methods $\mathcal{H}^{\text{CL}}$ and $\mathcal{H}^{\text{CO}}$ abide to axioms (A1) and (A2).

Proposition 1 *The combine-and-cluster method $\mathcal{H}^{CO}$ and cluster-and-combine method $\mathcal{H}^{CL}$ is valid and admissible: u_X^{CO} defined by (9) and u_X^{CL} defined by (10) are ultrametrics for all $I_X = (X, \underline{d}_X, \bar{d}_X)$; moreover, the methods $\mathcal{H}^{CO}$ and $\mathcal{H}^{CL}$ satisfy axioms (A1) and (A2).*

Given that we have constructed two admissible methods satisfying axioms (A1)-(A2), it is natural to ask whether these two constructions are special with respect to other satisfying methods. We prove the important characterization that any method $\mathcal{H}$ satisfying axioms (A1)-(A2) yields ultrametrics that lie between u_X^{CL} and u_X^{CO}.



(a) (b) (c)

(d) Cluster-and-combine dendrogram (e) Combine-and-cluster dendrogram (f) Benchmark: dendrogram via mean distance

Fig. 6. (a) Initial position of points, which correspond to two half moons. (b) Relationship between the average difference $u_X^{\text{CO}}(x, x') - u_X^{\text{CL}}(x, x')$ of the two extremal clustering methods across all pairs of nodes $x \neq x' \in X$ and the intensity of movement σ^2. The difference in $u_X^{\text{CO}}(x, x') - u_X^{\text{CL}}(x, x')$ increases but is significantly smaller comparing to $\bar{d}_X(x, x') - \underline{d}_X(x, x')$. (c) Relationship between the difference $u_X^{\text{CO}}(x, x') - u_X^{\text{CL}}(x, x')$ at three different pairs of points and the confidence level α. Resulting dendrograms of (d) cluster-and-combine method and (e) combine-and-cluster method. (f) Benchmark dendrogram: single linkage applied upon the mean distance $\tilde{d}_X(x, x') = \frac{1}{T} \sum_{t=1}^{T} d_X^t(x, x')$.

Theorem 2 *Consider an admissible method $\mathcal{H}$ satisfying (A1)-(A2). Given $I_X = (X, \bar{d}_X, \underline{d}_X)$ and $0 \leq \alpha \leq 1$, denote $(X, u_X) = \mathcal{H}(I_X)$ the output of applying $\mathcal{H}$ onto I_X. Then for any pair of nodes $x, x' \in X$,*

$$u_X^{CL}(x, x') \leq u_X(x, x') \leq u_X^{CO}(x, x'). \tag{11}$$

5. CLUSTERING OF MOVING POINTS BY SNAPSHOTS

We consider the clustering of n moving points in a two-dimensional plane with the initial coordinate of the i-th point represented by $\mathbf{p}_i^0 \in \mathbb{R}^2$. Points are moving in the plane and we have T snapshots with $\mathbf{p}_i^t \in \mathbb{R}^2$ denoting the coordinate of the i-th point at the t-th snapshot. We assume that the movements are completely random and therefore model the observation as $\mathbf{p}_i^t := \mathbf{p}_i^{t-1} + \boldsymbol{\epsilon}$ for any i and any time point $1 \leq t \leq T$, where $\boldsymbol{\epsilon} \in \mathbb{R}^2$ is a two-dimensional independent zero-mean Gaussian random variable with covariance matrix $\sigma^2\mathbf{I}$. Having no knowledge about the starting coordinates, we would like to evaluate clustering based on observations $\{\mathbf{p}_i^t\}_{i=1,\ldots,n,t=1,\ldots,T}$. To do so, we consider the node set X where $x_i \in X$ denotes the i-th point $\mathbf{p}_i$, and use $d_X^t(x_i, x_j) = \|\mathbf{p}_i^t - \mathbf{p}_j^t\|_2$ to represent the distance between the i-th and the j-th points at the t-th snapshot. Then we define metric space with distances given by intervals $(X, \underline{d}_X, \bar{d}_X)$ such that given a pair of nodes $x_i \neq x_j$, we set the distance lower bound $\underline{d}_X(x_i, x_j) = \min_{1 \leq t \leq T} d_X^t(x_i, x_j)$ as the minimum distance between the pair at all snapshots. Similarly, we define $\bar{d}_X(x_i, x_j) = \max_{1 \leq t \leq T} d_X^t(x_i, x_j)$. Clustering methods are applied upon $(X, \underline{d}_X, \bar{d}_X)$.

As an example, we consider $n = 30$ points whose initial coordinates form two half moons (Fig. 6 (a)), and investigate $T = 10$ snapshots of these moving points. We apply cluster-and-combine clustering $\mathcal{H}^{\text{CL}}$ and combine-and-cluster clustering $\mathcal{H}^{\text{CO}}$ onto the distance bounds $(X, \underline{d}_X, \bar{d}_X)$. The average difference between the output ultrametrics $u_X^{\text{CO}}(x, x') - u_X^{\text{CL}}(x, x')$ across all pairs of nodes $x \neq x' \in X$ with respect to the intensity of movement, i.e. the variance σ^2 of $\boldsymbol{\epsilon}$, is displayed in Fig. 6 (b) at three different α. The average difference generally increases with the intensity, however, does not increase significantly. As a comparison, the average difference between the distance bounds $\bar{d}_X(x, x') - \underline{d}_X(x, x')$ is 0.6353; even at relatively intense movement with $\sigma^2 \geq 1.2$, the average difference between the ultrametrics $u_X^{\text{CO}}(x, x') - u_X^{\text{CL}}(x, x')$ is less than 20% of the difference between the input distance bounds $\bar{d}_X(x, x') - \underline{d}_X(x, x')$. The relationship between the difference $u_X^{\text{CO}}(x, x') - u_X^{\text{CL}}(x, x')$ at three different pairs and α is plotted in Fig. 6 (c) where σ^2 is set to 0.9. The difference $u_X^{\text{CO}}(x, x') - u_X^{\text{CL}}(x, x')$ is not very high. Combining this with Theorem 2, the outputs of all admissible hierarchical clustering methods do not differ by much.

Finally, Fig. 6 (d) and (e) show the output dendrograms of cluster-and-combine and combine-and-cluster methods, respectively. The variance parameter σ^2 of movement $\boldsymbol{\epsilon}$ is set as 0.4 and the confidence level α as 0.5. Nodes 1 to 15 correspond to points in the upper moon regarding their initial coordinates; nodes 16 to 30 correspond to points in the lower moon. It can be seen from Fig. 6 (d) and (e) that (i) both $\mathcal{H}^{\text{CL}}$ and $\mathcal{H}^{\text{CO}}$ yield the desired output (two clusters correspond to two half moons), with only one point (15) gets misclassified, (ii) limited difference exists between the two dendrograms, and (iii) points closer in their initial positions (e.g. points 1 to 9, 10 to 14, 15 to 23, and 24 to 30) tend to be clustered together at low resolutions. As a benchmark, we consider the mean distance between any pair of nodes $\tilde{d}_X(x, x') = \frac{1}{T} \sum_{t=1}^{T} d_X^t(x, x')$ and apply single linkage upon $(X, \tilde{d}_X)$. Fig. 6 (f) shows the resulting dendrogram, which fails to identify the clusters correctly.

6. CONCLUSIONS

We developed a theory for hierarchically clustering metric spaces with distances given by intervals. We begin by identifying reasonable axioms. Two admissible methods were constructed and were proved as upper and lower bounds for all admissible methods. We explored the practical usefulness by clustering moving points via snapshots. The proposed methods succeeded in identifying the underlying clustering structures.

7. REFERENCES

[1] W. Huang and A. Ribeiro, “Axiomatic hierarchical clustering for intervals of metric distances,” in *Proc. Global Conf. Signal Info. Process. (GlobalSIP)*, Washington DC, 2016, p. (to appear).

[2] ——, “Metrics in the space of high order networks,” *Signal Processing, IEEE Transactions on*, vol. 64, no. 3, pp. 615–629, February 2016.

[3] ——, “Metrics in the space of high order proximity networks,” in *Proc. Int. Conf. Accoustics Speech Signal Process.*, Brisbane, Australia, 2015, pp. 4135–4139.

[4] ——, “Persistent homology lower bounds on high order network distances,” *IEEE Trans. Signal Process.*, vol. 64, no. 3, pp. 615–629, January 2017.

[5] ——, “Persistent homology lower bounds on network distances,” in *Proc. Int. Conf. Accoustics Speech Signal Process.*, Shanghai, China, 2016, pp. 4845–4849.

[6] U. Von Luxburg and S. Ben-David, “Towards a statistical theory of clustering,” in *Pascal work. on stat. & opt. of clust.*, 2005, pp. 20–26.

[7] S. Ben-David, U. Von Luxburg, and D. Pál, “A sober look at clustering stability,” in *Learning theory*. Springer, 2006, pp. 5–19.

[8] I. Guyon, U. Von Luxburg, and R. C. Williamson, “Clustering: Science or art,” in *NIPS 2009 workshop on clustering theory*, 2009, pp. 1–11.

[9] S. Ben-David and M. Ackerman, “Measures of clustering quality: A working set of axioms for clustering,” in *Advances in neural information processing systems*, 2009, pp. 121–128.

[10] R. B. Zadeh and S. Ben-David, “A uniqueness theorem for clustering,” in *Proc. of the twenty-fifth conference on uncertainty in artificial intelligence*, 2009, pp. 639–646.

[11] G. Carlsson and F. Mémoli, “Characterization, stability and convergence of hierarchical clustering methods,” *J. Mach. Learn. Res*, vol. 11, pp. 1425–1470, 2010.

[12] ——, “Classifying clustering schemes,” *Found. Comput. Math.*, vol. 13, no. 2, pp. 221–252, 2013.

[13] G. Carlsson, F. Memoli, A. Ribeiro, and S. Segarra, “Axiomatic construction of hierarchical clustering in asymmetric networks,” *Trans. Inf. Theory*, vol. (submitted.), 2015.

[14] G. N. Lance and W. T. Williams, “A general theory of classificatory sorting strategies ii. clustering systems,” *The computer J.*, vol. 10, no. 3, pp. 271–277, 1967.

[15] A. K. Jain and R. C. Dubes, *Algorithms For Clustering Data*. Prentice-Hall, Inc., 1988.

[16] Y. Zhao, G. Karypis, and U. Fayyad, “Hierarchical clustering algorithms for document datasets,” *Data mining and knowledge discovery*, vol. 10, no. 2, pp. 141–168, 2005.

[17] G. Cormode and A. McGregor, “Approximation algorithms for clustering uncertain data,” in *Proc. of the 27th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2008, pp. 191–200.

[18] F. Gullo, G. Ponti, A. Tagarelli, and S. Greco, “A hierarchical algorithm for clustering uncertain data via an information-theoretic approach,” in *Data Mining, 2008. ICDM’08. Eighth IEEE Int. Conf. on*, 2008, pp. 821–826.

[19] E. Schubert, A. Koos, T. Emrich, A. Züfle, K. A. Schmid, and A. Zimek, “A framework for clustering uncertain data,” *Proc. of the VLDB Endowment*, vol. 8, no. 12, pp. 1976–1979, 2015.

[20] M. Chavent and Y. Lechevallier, “Dynamical clustering of interval data: Optimization of an adequacy criterion based on hausdorff distance,” in *Classification, Clustering, and Data Analysis*. Springer, 2002, pp. 53–60.

[21] R. M. de Souza and F. d. A. De Carvalho, “Clustering of interval data based on city–block distances,” *Pattern Recognition Let.*, vol. 25, no. 3, pp. 353–365, 2004.

[22] P. Yu, L. Qinghua, and P. Xiyuan, “Uck-means: A customized k-means for clustering uncertain measurement data,” in *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth Int. Conf. on*, vol. 2, 2011, pp. 1196–1200.

[23] B. Jiang, J. Pei, Y. Tao, and X. Lin, “Clustering uncertain data based on probability distribution similarity,” *Knowledge and Data Engineering, IEEE Trans. on*, vol. 25, no. 4, pp. 751–763, 2013.

[24] W. Huang and A. Ribeiro, “Hierarchical clustering given confidence intervals of metric distances,” *Signal Processing, IEEE Transactions on*, vol. (submitted), October 2016. [Online]. Available: https://arxiv.org/abs/1610.04274