OVERLAPPING CLUSTERING OF NETWORK DATA USING CUT METRICS

Fernando Gama, Santiago Segarra and Alejandro Ribeiro

Department of Electrical and Systems Engineering, University of Pennsylvania

ABSTRACT

We present a novel method to hierarchically cluster networked data allowing nodes to simultaneously belong to multiple clusters. Given a network, our method outputs a cut metric on the underlying node set, which can be related to data coverings at different resolutions. The cut metric is obtained by averaging a set of ultrametrics, which are themselves the output of (non-overlapping) hierarchically clustering noisy versions of the original network of interest. The resulting algorithm is illustrated in synthetic networks and is used to classify handwritten digits from the MNIST database.

Index Terms— Clustering, Network Theory, Cut Metrics, Hierarchical clustering, Covering, Dithering.

1. INTRODUCTION

Given a network, i.e. a set of nodes endowed with a pairwise dissimilarity function, the objective of clustering is to partition the node set into groups such that nodes inside one group are more similar to each other than they are to nodes outside of it [1]. The generality of such operation fosters the extended application of clustering in multiple fields of knowledge such as social network analysis [2], political sciences [3] and neuroscience [4]. Traditional methods of clustering that output only one partition of the node set are not always useful or appropriate since there might be multiple layers of interrelation within the structure of the data or particular nodes that rightfully belong to more than one category [5–7].

Hierarchical clustering methods, whose output is a dendrogram consisting of a nested set of partitions indexed by a resolution parameter [1, 8], capture multi-resolution relations in the data overcoming the first limitation of traditional clustering. Examples of these methods are UPGMA [9], Ward's method [10] or single linkage [11]. Soft or fuzzy clustering methods accommodate for the allocation of a node to multiple clusters by assigning to each node a membership degree or probability of belonging to different subsets [12–14]. An alternative solution to the second limitation of traditional methods is named *overlapping* clustering and involves non-hierarchical deterministic assignments of nodes to more than one subset [15].

In the present paper, a *hierarchical overlapping* clustering method is proposed. Our method outputs a collection of groupings that represents several layers on the structure of the data depending on the degree of relation between the nodes (hierarchical), while allowing a node to belong to more than one subset (overlapping). In order to achieve this, a systematic method for obtaining cut metrics [16, 17] is described, which are later used to obtain nested collections of coverings over the node set.

We introduce the necessary notions for the definition of our clustering method (Section 2) and then define cut metrics and explain their relation to coverings (Section 3). In Section 4 we leverage this relation to define a hierarchical overlapping clustering method based on averaging several outputs of a predetermined hierarchical (non-overlapping) clustering method applied to noisy versions of the network of interest (see Algorithm 1). Finally, in Section 5 several illustrative examples as well as the application of the proposed algorithm to classifying two digits from the MNIST database can be found.

Work in this paper is supported by NSF CCF-1217963 and AFOSR MURI FA9550-10-1-0567. Authors can be reached at {fgama,ssegarra,aribeiro}@seas.upenn.edu

2. PRELIMINARIES

Let $N = (X, A_X)$ be a network defined by a finite set of nodes X and a nonnegative dissimilarity function A_X. The dissimilarity function $A_X : X \times X \to \mathbb{R}_+$ portrays how different the nodes are and it is required to satisfy $A_X(x, x') \geq 0$ for all $x, x' \in X$ and $A_X(x, x') = 0$ if and only if $x = x'$. No assumptions are made on whether A_X is symmetric or on whether it satisfies the triangle inequality. The space of networks is denoted by $\mathcal{N}$.

We define a partition of the space X as a collection $P_X = \{B_1, \cdots, B_m\}$ of nonintersecting subsets of X that covers the whole space. That is, the elements of the partition P_X satisfy $B_i \cap B_j = \emptyset$ for all $i \neq j$ and $\cup_{i=1}^m B_i = X$. The space of partitions is denoted by $\mathcal{P}$. We further define an equivalence relation as a binary relation between elements of X such that for all triplets $x, x', x'' \in X$, we have that $x \sim x$ (reflexivity), $x \sim x'$ if and only if $x' \sim x$ (symmetry), and the relations $x \sim x'$ and $x' \sim x''$ imply the relation $x \sim x''$ (transitivity). Observe that the relation $x \sim x'$ if and only if $x, x' \in B_i$ for some i is an equivalence induced by the partition P_X. The converse, namely, that an equivalence relation defines a partition of X is also true [18, 19].

The goal of a clustering method is to partition the space X into groups of nodes that are more similar to each other than they are to the rest as indicated by the dissimilarity function A_X. This is formally specified by defining a clustering method $\mathcal{G}$ as a structure preserving map from the space of networks to the space of partitions, $\mathcal{G} : \mathcal{N} \to \mathcal{P}$. That $\mathcal{G}$ is structure preserving implies that the output partition $\mathcal{G}(N)$ is defined on the node set of N for all $N \in \mathcal{N}$. Clustering methods can be generalized to hierarchical clustering methods where the output is not a single partition but a dendrogram D_X, which is defined as a nested collection of partitions $D_X(\delta)$ indexed by a resolution parameter $\delta \geq 0$ [20]. The resolution parameter specifies what is considered sufficient proximity for the creation of the cluster. At resolution $\delta = 0$, the partition $D_X(0) = \{\{x\}, x \in X\}$ puts all nodes in separate singleton clusters. At large resolutions $\delta \geq 0$, the partition $D_X(\delta) = X$ assigns all nodes to the same cluster. For intermediate resolutions the partitions are nested in the sense that as δ increases, nodes may combine into new clusters but once they are clustered at a certain resolution they stay clustered at larger resolutions. Formally, if we use $x \sim_\delta x'$ to signify that x and x' are in the same cluster at resolution δ, it must be that $x \sim_{\delta'} x'$ for all resolutions $\delta' \geq \delta$. A hierarchical clustering method $\mathcal{H}$ is defined as a structure preserving map from the space of networks $\mathcal{N}$ to the space of dendrograms $\mathcal{D}$ [20],

$$\mathcal{H} : \mathcal{N} \to \mathcal{D} \tag{1}$$

An important limitation of clustering and hierarchical clustering methods is that nodes belong to one and only one element of a partition. However, in many practical situations there are nodes that naturally belong to more than one category. For e.g., in the dumbbell network of Fig. 3 it is intuitive that the clouds of nodes on each side should be separate clusters. However, it is not so clear that *all* of the handle should be a separate cluster as it is not unreasonable to assign its borders to the respective cloud clusters. We can overcome this issue by using coverings in lieu of partitions, tolerance relationships in lieu of equivalences, and nested collections of coverings in lieu of dendrograms; see Table 1.

Formally, we define a covering of X as a collection $C_X = \{C_1, \ldots, C_m\}$ of subsets of X such that $\cup_{i=1}^m C_i = X$ but not necessarily $C_i \cap C_j = \emptyset$ for $i \neq j$. Likewise, we define a tolerance

Method	Relation	Grouping	Nested hierarchy
Hierarchical	Equivalence	Partition	Dendrogram
Overlapping	Tolerance	Covering	Nested covering

Table 1: Parallelism between hierarchical and overlapping clustering.

relation between pairs of elements of X as a relationship that is reflexive ($x \leftrightarrow x$) and symmetric ($x \leftrightarrow x'$ implies $x' \leftrightarrow x$) but not necessarily transitive. It is immediate to realize that the relationship $x \leftrightarrow x'$ if and only if $x, x' \in C_i$ for some i is a tolerance relation induced by the covering C_X. This observation allows us to introduce the notion of a nested collection of coverings that we formally define next.

Definition 1. *Let K_X be a collection of coverings $K_X(\delta)$ indexed by the resolution parameter $\delta \geq 0$. If the covering $K_X(\delta)$ induces the tolerance relation $x \leftrightarrow_\delta x'$ (a tolerance relation $\leftrightarrow$ for each δ), then we say that K_X is nested if and only if*

Border conditions: *$K_X(0) = \{\{x\}, x \in X\}$ and there exists $\delta_{\max}$ such that for all $\delta \geq \delta_{\max}$ it holds $K_X(\delta) = \{X\}$.*

Nested coverings: *If $x \leftrightarrow_\delta x'$ at resolution δ, it must be that $x \leftrightarrow_{\delta'} x'$ for all resolutions $\delta' \geq \delta$.*

The family of all nested coverings is denoted as $\mathcal{K}$.

As per Definition 1, a nested covering is to a covering what a dendrogram is to a partition. Likewise, as coverings are generalizations of partitions, nested coverings are generalizations of dendrograms; see Table 1. The goal of this paper is to design overlapping clustering methods that we formally define as structure preserving maps

$$\mathcal{O} : \mathcal{N} \to \mathcal{K}. \tag{2}$$

Our approach to deriving overlapping clustering methods is to draw on the equivalence between ultrametrics and dendrograms and a reinterpretation of cut metrics as a nested collection of coverings as we discuss in the following section.

3. NESTED COVERINGS DERIVED FROM CUT METRICS

Dendrograms are known to be equivalent to finite ultrametric spaces [11]. An ultrametric $u : X \times X \to \mathbb{R}$ is a function that for all $x, x', x'' \in X$ satisfies $u(x, x') \geq 0$, $u(x, x') = 0$ if and only if $x = x'$, $u(x, x') = u(x', x)$, and $u(x, x'') \leq \max\{u(x, x'), u(x', x'')\}$. These three conditions imply that u is a metric that satisfies a stronger version of the triangle inequality where sides of each triangle are smaller than the maximum size of the other two instead of being smaller than their sum [21]. It is not difficult to show that the function $u : X \times X \to \mathbb{R}$ defined as

$$u(x, x') \leq \delta \iff x \sim_{\mathcal{D}_X(\delta)} x' \tag{3}$$

is an ultrametric on the finite space X [20, Theorem 1]. Thus, if we define as $\mathcal{U}$ the space of ultrametrics [20], a hierarchical clustering method $\mathcal{H}$ can be reinterpreted as a structure preserving map $\mathcal{H} : \mathcal{N} \to \mathcal{U}$.

Cut metrics are metrics that include ultrametrics as particular cases. To define a cut metric, first consider a set $S \subseteq X$; then, the cut semimetric δ_S [16, Ch. 4] associated with S assigns dissimilarity 0 to points x and x' when both belong to S or when both belong to the complement S^c and dissimilarity 1 otherwise. Thus, with $\mathbb{I}$ denoting the indicator function, the cut semimetric dissimilarity between x and x' can be written as

$$\delta_S(x, x') = \mathbb{I}\{S \cap \{x, x'\} \neq \emptyset\}\mathbb{I}\{S^c \cap \{x, x'\} \neq \emptyset\}. \tag{4}$$

This cut semimetric can also be understood as being induced by a binary classifier which partitions the node set in two (S and S^c) and assigns a unit dissimilarity to nodes belonging to different categories. If multiple binary classifiers are considered, then a cut metric c_X can be obtained as the combination of all the associated dissimilarities. That is, c_X is a metric that can be written as

$$c_X(x, x') = \sum_{S \subseteq X} \lambda_S \delta_S(x, x') \tag{5}$$

with $\lambda_S \geq 0$ and where the sum ranges over all possible subsets of X. Notice that $\lambda_{S'} = 0$ for some S' amounts to the associated binary classifier not being considered in the construction of c_X. Observe that $c_X(x, x) = 0$ and that $c_X(x, x') = c_X(x', x)$, for $x, x' \in X$. We may now define a nested collection of coverings from cut metrics.

Proposition 1. *Let X be a set of nodes and let c_X be a cut metric defined on the node set X. Let $K_X = \{K_X(\delta), \delta \geq 0\}$ be a collection of coverings $K_X(\delta)$. If, for each $\delta \geq 0$, the corresponding covering $K_X(\delta)$ is obtained from the tolerance relation given by*

$$c_X(x, x') \leq \delta \implies x \leftrightarrow_\delta x' \tag{6}$$

then K_X is a nested collection of coverings in the sense of Definition 1.

Proof: First, observe that (6) defines a genuine tolerance relation since $c_X(x, x) = 0$ implies reflexivity of $\leftrightarrow_\delta$ and $c_X(x, x') = c_X(x', x)$ implies symmetry. Second, for each δ a tolerance relation $\leftrightarrow_\delta$ is obtained using (6) and this relation is used to construct a covering by blocks $K_X(\delta)$ [22]. Finally, observe that K_X is nested since, for any $\delta < \delta'$, $c_X(x, x') \leq \delta < \delta'$ implies that $x \leftrightarrow_\delta x' \Rightarrow x \leftrightarrow_{\delta'} x'$. ■

Note that (6) is analogous to the relation between ultrametrics and equivalence relations in (3). Also, in an analogy to dendrograms, observe that for $\delta = 0$, $K_X(0) = \{\{x\}, x \in X\}$, and that there exists a $\delta_{\max}$ such that for all $\delta \geq \delta_{\max}$, $K_X(\delta) = \{X\}$. Then, for $0 < \delta < \delta_{\max}$ the resulting tolerance relations define coverings in such a way that nodes might or might not be in more than one subset, since there is no transitivity in the relation. If there exists at least one node that belongs to more than one subset, it is said that there is *overlap*. It is relevant, then, to compute the number of overlapping nodes for each value of δ. Define the *overlapping function* as

$$f_{\text{ol}}(\delta) = \sum_{k=1}^{n} \mathbb{I}\big\{x_k \leftrightarrow_\delta x_i, x_k \leftrightarrow_\delta x_j, x_i \nleftrightarrow_\delta x_j \quad \text{for } i \neq k \neq j,\ i, j = 1, \ldots, n\big\} \tag{7}$$

where the tolerance $\leftrightarrow_\delta$ is determined by a cut metric $c_X(x, x')$ as per (6). The function $f_{\text{ol}}(\delta)$ adds one for each node that is overlapping, i.e. that is in more than one subset of the covering $K_X(\delta)$. Following from the nested collection of coverings K_X it is obtained that $f_{\text{ol}}(0) = 0$ and that $f_{\text{ol}}(\delta) = 0$ for all $\delta \geq \delta_{\max}$.

The overlapping function provides essential information about the grouping structure of the network. For instance, it serves to define clusterability as follows.

Definition 2. *Let $N = (X, A_X)$ be a network, and let $c_X(x, x')$ be a cut metric defined over the set of nodes X. If there exists a δ such that $f_{ol}(\delta) = 0$, $K_X(\delta) \neq \{\{x\}, x \in X\}$ and $K_X(\delta) \neq \{X\}$, then (X, c_X) is said to be clusterable.*

In other words, the clusterability of the network under c_X is given when the resulting covering is a partition and is not the same partition as in the extreme cases $\{\{x\}, x \in X\}$ (all-separate) and $\{X\}$ (all-together). If the network is not clusterable, the overlapping function still provides valuable information about the underlying grouping structure and may help to identify those nodes that are subject to a closer scrutiny or that simply cannot be fully incorporated into one subset or another; see Section 5.

To sum up, in the same way as in hierarchical clustering, ultrametrics are used to define equivalence relations that determine partitions, in the proposed overlapping clustering method, cut metrics are used to define tolerance relations that determine coverings; see Table 1.

Remark 1. Proposition 1 can be restated for c_X being *any* symmetric dissimilarity function, however, we specialize it to cut metrics since this particular type of distance function arises when several hierarchical clustering outputs are combined; see Section 4.

input : no. of perturbations J, network N, hierarchical clustering method $\mathcal{H}$, perturbation$(\cdot)$
for i=$1{:}J$ **do**
  $\tilde{N}$=perturbation(N);
  $u_i(x,x')$=$\mathcal{H}(\tilde{N})$;
end
output: $c_X(x,x')$=avg$(u_1(x,x'),\ldots,u_J(x,x'))$

Algorithm 1: Overlapping clustering algorithm.

4. OVERLAPPING CLUSTERING ALGORITHM

In virtue of Proposition 1, cut metrics can be used to obtain nested collections of coverings. We now discuss how to obtain cut metrics that reflect the dissimilarity between the nodes of the network N in a systematic way. The first step towards addressing this issue is to consider the following proposition.

Proposition 2. *A convex combination of ultrametrics yields a cut metric.*

Proof: Given (X, d_X) where d_X is a metric, then d_X is in particular a tree metric – there exists a tree graph in which it is possible to embed the distance on its edges [16, p. 147] – if and only if it satisfies the four-point condition [23, Theorem 1]

$$d_X(x,x') + d_X(x'',x''') \leq \max\big\{d_X(x,x'') + d_X(x',x'''), \\ d_X(x,x''') + d_X(x',x'')\big\} \quad (8)$$

for all $x, x', x'', x''' \in X$. Observe that the condition that all ultrametrics satisfy, namely $u(x,x'') \leq \max\{u(x,x'), u(x',x'')\}$, is a particular case of (8). Hence, all ultrametrics are also tree metrics [16, p. 311]. Finally, note that tree metrics, as well as convex combinations of tree metrics, are ℓ_1-embeddable [16, Fact 11.1.4, 11.1.5]. Since a metric is ℓ_1-embeddable if and only if it is a cut metric [16, Proposition 4.4.2], a convex combination of ultrametrics yields a cut metric. ■

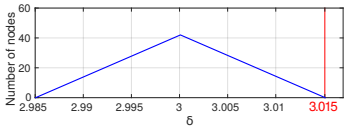
From this proposition, it is immediate that ultrametrics are particular cases of cut metrics. This also shows in the fact that equivalence relations are particular cases of tolerance relations and that partitions are particular cases of coverings; see Table 1. Also, Proposition 2 plays a key role in the generation of cut metrics as it gives a systematic way to obtain cut metrics from ultrametrics, which are readily available from the application of hierarchical clustering.

Next step is to address the problem of obtaining ultrametrics that are closely related to the network of interest. By intentionally applying random noise to the dissimilarity function A_X of a network N, a whole family of closely related networks can be obtained. Each one of these networks yields a different ultrametric when a hierarchical clustering method is applied. If all these ultrametrics are averaged, then a cut metric is obtained (cf. Proposition 2) and a systematic method for obtaining such cut metrics is readily available. This idea stems from the concept of dithering [24]. Formally, let $\tilde{N} = \{N_1, \ldots, N_J\}$ be the J networks resulting from dithering J times the dissimilarity function, $N_i = (X, \tilde{A}_i)$, $i = 1, \ldots, J$ and where $\tilde{A}_i(x,x')$ is a specific realization of a small perturbation around $A_X(x,x')$. Let $\{u_1, \ldots, u_J\}$ be the set of ultrametrics resulting from applying a predetermined clustering method $\mathcal{H}$ (cf. Section 1) to each of the networks in $\tilde{N}$. Then, in virtue of Proposition 2, build the cut metric as

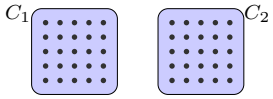
$$c_X(x,x') = \frac{1}{J}\sum_{i=1}^{J} u_i(x,x'). \quad (9)$$

Finally, use this cut metric to obtain a nested collection of coverings (Proposition 1). The algorithm is described in Algorithm 1.

Remark 2. Proposition 2 also holds for non-negative linear combinations of ultrametrics, but only a convex combination guarantees that the distance scale described by the ultrametrics is preserved in the resulting cut metric.



(a) Overlapping function



(b) Covering for $\delta = 3.0151$.

Fig. 1: Layout, overlapping function and coverings for a simple network consisting of two clouds.

Remark 3. It is important to observe that the cut metrics conform a convex cone and hence, any non-negative combination of cut metrics is still a cut metric. Also, observe that this is not true in the case of ultrametrics: a non-negative linear combination of ultrametrics is not an ultrametric [20, Section VII-B].

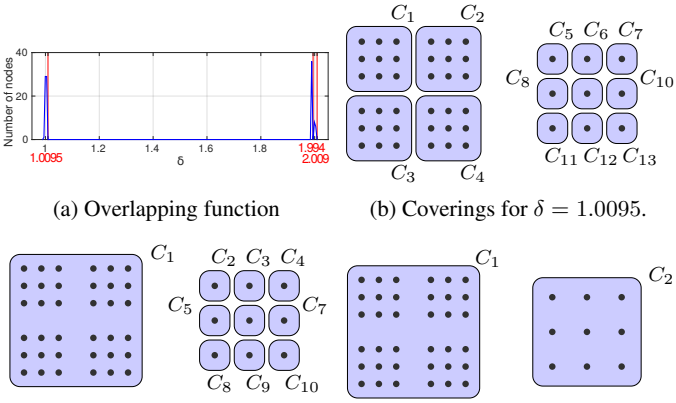
5. APPLICATIONS

The first three illustrations of Algorithm 1 are conducted on synthetic networks $N = (X, A_X)$ where X is a set of points embedded in the plane and A_X is the Euclidean distance between them. As a hierarchical clustering method $\mathcal{H}$ we apply single linkage [11] and average the clustering output of $J = 100$ different noisy realizations of N. The noisy $\tilde{N}$ are obtained by perturbing the positions of the nodes in the plane with zero-mean gaussian noise with standard deviation $\sigma = 0.01$.

Consider the simple network portrayed in Fig. 1b, where the distance between nodes in the same point cloud is $d_1 = 3$ and between clouds is $d_2 = 18$. The overlapping function is found in Fig. 1a. As expected, this network is clusterable (cf. Definition 2), i.e., a meaningful (non-overlapping) partition can be found. Indeed, for $\delta = 3.0151$, the overlapping function has a zero that does not correspond to the all-separate or the all-together coverings. In fact, the resulting covering $K_X(\delta) = \{C_1, C_2\}$ is shown in Fig. 1b and it consists of two subsets C_1 and C_2 each one containing one of the clouds. Observe that the value of δ that generates this covering is in the order of d_1.

Multiple resolution datasets. Consider now the network N with nodes as depicted in Fig. 2b,c,d, where the distance between the nodes in the four smaller clouds is $d_1 = 1$, the distance between these four clouds as well as the distance between the nodes of the fifth larger cloud is $d_2 = 2$ and the distance between the smaller clouds and the larger cloud is $d_3 = 5$. There are three values of δ for which the network is clusterable (see Fig. 2a) corresponding to the multiple meaningful resolutions that the network presents. First, for $\delta = 1.0095$ it is observed in Fig. 2b that there is one subset for each of the closer clouds and one subset for each of the nodes of the sparser cloud, this is reasonable for the value of δ that is close to d_1. For values of δ near $d_2 = 2$ we can observe two different informative partitions. For $\delta = 1.9940$, the four closer clouds have been clustered together but each of the nodes of the sparser cloud remain separate; see Fig. 2c. Finally, for $\delta = 2.0090$ as shown in Fig. 2d, the four closer clouds are all clustered together and in a separate subset the whole of the sparser cloud.

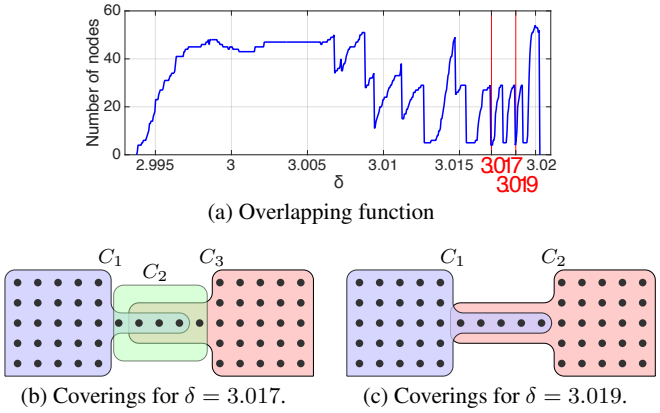
A solution to single linkage's chaining effect. Consider the network formed by the set of nodes in Fig. 3b,c where the distance between any adjacent nodes is $d = 3$. Notice that in this case, unlike the previous examples considered, the expected clustering output is not unequivocal. Indeed, one might argue, e.g., that the natural clustering of the above arrangement of points is to obtain three clusters – the two squared and the linear arrangements – with some overlap in the boundaries, or, quite differently, that the natural output should consist of two clusters centered at both squared portions and sharing the whole linear arrangement of nodes. Furthermore, if one applies single linkage clustering to the (unperturbed) network of interest, the output dendrogram consists of two different nested partitions one where every node is in a different cluster and another one where every node belongs to the same cluster.



(a) Overlapping function (b) Coverings for $\delta = 1.0095$.

(c) Coverings for $\delta = 1.9940$. (d) Coverings for $\delta = 2.0090$.

Fig. 2: Overlapping function and coverings for a multiple resolution network.



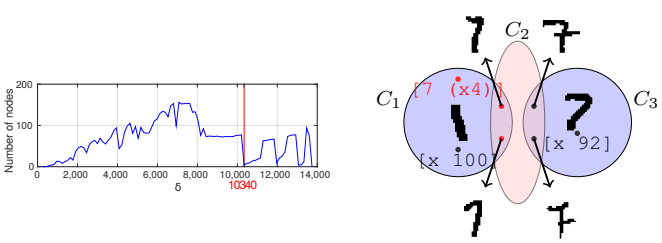
(a) Overlapping function

(b) Coverings for $\delta = 3.017$. (c) Coverings for $\delta = 3.019$.

Fig. 3: Overlapping function and coverings for a network that presents SL chaining effect.

This phenomenon is known as *chaining effect* [8] and is an undesirable concomitant of the definition of single linkage. To understand this effect, recall that the ultrametric $u^{\mathrm{SL}}(x, x')$ output by single linkage between x and x' is given by the minimum cost of a path linking these two nodes where the cost is defined as the maximum dissimilarity encountered when traversing the path [11]. Thus, for *any* pair of nodes x, x' we have that $u^{\mathrm{SL}}(x, x') = d = 3$ resulting in a global cluster formed at resolution $\delta = 3$. Our proposed approach, based on dithering, solves both aforementioned problems: the lack of a single reasonable covering and the chaining effect.

To see this, focus on the overlapping function shown in Fig. 3a. Notice that the network is not clusterable (cf. Definition 2), which aligns with the lack of an intuitive non-overlapping outcome. Moreover, the overlapping function contains several local minima and the coverings corresponding to the two smallest ones are depicted in Fig. 3b,c. These coverings correlate with our a priori notions of reasonable outputs and, by exploring the remaining local minima, additional sensible coverings are revealed. Observe also that the chaining effect vanishes when noise is added to the data points. To understand this, consider three points x, x', x'' where the first two belong to the left-most squared portion and x'' belongs to the opposite one. In the unperturbed version of the data, there are multiple paths linking x and x' of cost exactly 3, however, every path linking x and x'' with cost 3 must contain the linear portion of the dataset. Hence, when noise is added, some of the paths linking x and x' might be disrupted but others might even decrease their cost,



(a) Overlapping function. (b) Coverings for $\delta = 10340$.

Fig. 4: Overlapping function and coverings for classification of digits 1 and 7 of the MNIST digit database.

however, the disruption of the linear arrangement necessarily increases the cost of linking x and x'', thus, recognizing x'' as belonging to a different category from the other two points. At a fundamental level, by introducing an algorithm based on dithering and averaging we obtain a clustering method that depends not only on the path of minimum cost – as single linkage – but also on the density of paths of near-minimal cost.

Handwritten digit classification. The method proposed is applied to the classification of two usually hard to distinguish handwritten digits, namely 1 and 7, taken from the MNIST database [25]. For each digit, 100 black and white images of size 28×28 are obtained at random from the database and converted to a vector of size 784. Principal Component Analysis (PCA) [26] is performed on these samples by estimating the mean and covariance matrix with $5,000$ training samples of each digit. The first 20 PCA components are kept. These 200 vectors of size 20 conform the nodes of the network. The dissimilarity matrix is obtained by computing the euclidean distance between the 20-PCA vector nodes. The dithering step is repeated $J = 100$ times applying white gaussian noise of deviation σ to the positions in the PCA space. The value of σ is given by 0.05 times the minimum positive element of the dissimilarity matrix. The ultrametrics are obtained applying Ward linkage [10].

The resulting overlapping function can be found in Fig. 4a and it is observed that the network is not clusterable. The minimum of the local minima of the overlapping function is given by $\delta = 10340$. The resulting covering is portrayed in Fig. 4b. There are essentially two big covers: C_1 that contains all of the ones and 6 sevens of which 4 are considered to be classification errors; and C_3 which contains 94 sevens. Then there is a smaller overlapping covering C_2 that signals those nodes in each of the bigger covers that are hard to classify and which may require closer scrutiny. It is observed that the two digits that are also in C_1 are clearly confused with a one, also from the perspective of a human classifier. The two digits that are signaled in C_3 are those that have a seven written with an extra line, which throws off the classifier. Finally, we remark that the accuracy of the proposed method depends on the accuracy of the hierarchical method used in each dithering step.

6. CONCLUSIONS

We introduced a hierarchical overlapping clustering method for networked data that deterministically allows a node to belong to more than one cluster. Cut metrics were used to extract nested coverings of the data, and these cut metrics were obtained by averaging ultrametrics that resulted from applying a predetermined hierarchical clustering method to dithered versions of the network. The overlapping function was introduced as a tool to analyze the obtained nested collection of coverings. The proposed algorithm was applied to several illustrative examples showing that it can handle multi-resolution data and solve undesirable problems like single linkage's chaining effect, and was also used to classify handwritten digits while detecting equivocal data points.

7. REFERENCES

[1] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall Advanced Reference Series. Prentice Hall, Englewood Cliffs, NJ, 1988.

[2] C. Lu, X. Hu, and J. Park, "Exploiting the Social Tagging Network for Web Clustering," *IEEE Transactions on Systems, Man, and Cybernetics–Part A: Systems and Humans*, vol. 41, no. 5, pp. 840–852, September 2011.

[3] M. Paulus and L. Kristoufek, "Worldwide clustering of the corruption perception ," *Physica A: Statistical Mechanics and its Applications*, vol. 428, pp. 351 – 358, 2015.

[4] A. Ozdemir, M. Bolaños, E. Bernat, and A. Selin, "Hierarchical Spectral Consensus Clustering for Group Analysis of Functional Brain Networks," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 9, pp. 2158–2169, September 2015.

[5] M. A. Youssef, A. Youssef, and M. F. Younis, "Overlapping Multihop Clustering for Wireless Sensor Networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 20, no. 12, pp. 1844–1856, December 2009.

[6] P. C. H. Ma and K. C. C. Chan, "A Novel Approach for Discovering Overlapping Clusters in Gene Expression Data," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 7, pp. 1803–1809, July 2009.

[7] M. Liu, B. C. Vemuri, S.-I. Amari, and F. Nielsen, "Shape Retrieval Using Hirerarchical Total Bregman Soft Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2407–2419, December 2012.

[8] G. N. Lance and W. T. Williams, "A general theory of classificatory sorting strategies 1. Hierarchical systems," *Computer Journal*, vol. 9, no. 4, pp. 373–380, 1967.

[9] R. R. Sokal and C. D. Michener, "A Statistical Method for Evaluating Systematic Relationships," *The University of Kansas Science Bulletin*, vol. 38 (II), no. 22, pp. 1409–1438, March 1958.

[10] J. H. Ward Jr., "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, March 1963.

[11] G. E. Carlsson and F. Mémoli, "Characterization, Stability and Convergence of Hierarchical Clustering Methods," *Journal of Machine Learning Research*, vol. 11, pp. 1425–1470, April 2010.

[12] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Advanced Applications in Pattern Recognition. Plenum Press, New York, NY, 1981.

[13] A. Baraldi and P. Blonda, "A Survey for Fuzzy Clustering Algorithms for Pattern Recognition–Part I," *IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics*, vol. 29, no. 6, pp. 778–785, December 1999.

[14] A. Baraldi and P. Blonda, "A Survey for Fuzzy Clustering Algorithms for Pattern Recognition–Part II," *IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics*, vol. 29, no. 6, pp. 786–801, December 1999.

[15] G. Cleuziou, "An Extended Version of the k-means Method for Overlapping Clustering," in *19th International Conference on Pattern Recognition, 2008. ICPR 2008.*, 8-11 December 2008.

[16] M. Deza and M. Laurent, *Geometry of Cuts and Metrics*, LIENS - Ecole Normale Supérieure, Paris, France, November 1996.

[17] J. Culbertson, D. P. Guralnik, and P. F. Stiller, "Injective metrizability and the duality theory of cubings," *ArXiv e-prints*, Jan. 2015.

[18] P. R. Halmos, *Naive Set Theory*, The University Series in Undergraudate Mathematics. Van Nostrand Reinhold Company, New York, NY, 1960.

[19] E. Schechter, *Handbook of Analysis and Its Foundations*, Academic Press, San Diego, CA, 1997.

[20] G. E. Carlsson, F. Mémoli, A. Ribeiro, and S. Segarra, "Axiomatic construction of hierarchical clustering in asymmetric networks," *CoRR*, vol. abs/1301.7724, 2013.

[21] D. Burago, Y. Burago, and S. Ivanov, *A Course in Metric Geometry*, vol. 33 of *AMS Graduate Studies in Math.*, American Mathematical Society, 2001.

[22] W. Bartol, J. Miró, J. Pióro, and F. Rosselló, "On the coverings by tolerance classes," *Information Sciences*, vol. 166, no. 1-4, pp. 193–211, 2004.

[23] P. Buneman, "A Note on the Metric Properties of Trees," *Journal of Combunatorial Theory, Series B*, vol. 17, no. 1, pp. 48–50, 1974.

[24] L. Schuchman, "Dither Signals and Their Effect on Quantization Noise," *IEEE Transactions on Communication Technology*, vol. 12, no. 4, pp. 162–165, December 1964.

[25] Y. Le Cun, C. Cortes, and C. J. C. Burges, "The MNIST Database of handwritten digits," Website, http://yann.lecun.com/exdb/mnist/, 2015-08-18.

[26] J. E. Jackson, *A User's Guide to Principal Components*, John Wiley and Sons, New York, NY, 1991.