Overlapping Clustering of Network Data Using Cut Metrics

Fernando Gama, Santiago Segarra & Alejandro Ribeiro
Dept. of Electrical and Systems Engineering
University of Pennsylvania
fgama@seas.upenn.edu

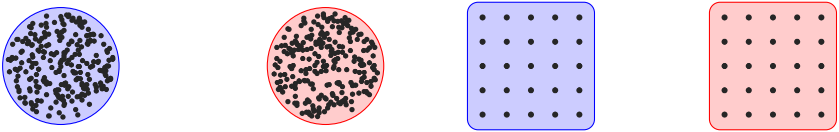
ICASSP, March 24, 2016

Clustering

- Find node partition such that nodes within partition are similar
  $\Rightarrow$ Ill defined: What is similar? Why a partition?
- Similarity entails inherent notion of scale $\Rightarrow$ Hierarchical clustering
- All scales are important $\Rightarrow$ Nested cluster family indexed by scale
- Datasets are very rarely separable into clean partitions
- Some points are. Others could be members of multiple “partitions”

Clustering

- Find node partition such that nodes within partition are similar
  $\Rightarrow$ Ill defined: What is similar? Why a partition?
- Similarity entails inherent notion of scale $\Rightarrow$ Hierarchical clustering
- All scales are important $\Rightarrow$ Nested cluster family indexed by scale
- Datasets are very rarely separable into clean partitions
- Some points are. Others could be members of multiple "partitions"

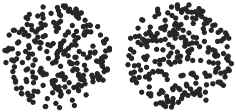


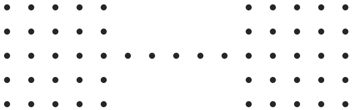
Clustering

- Find node partition such that nodes within partition are similar
  $\Rightarrow$ Ill defined: What is similar? Why a partition?

- Similarity entails inherent notion of scale $\Rightarrow$ Hierarchical clustering
- All scales are important $\Rightarrow$ Nested cluster family indexed by scale

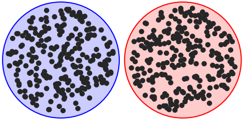
- Datasets are very rarely separable into clean partitions
- Some points are. Others could be members of multiple "partitions"

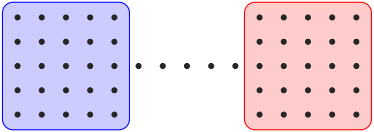




Clustering

- Find node partition such that nodes within partition are similar
  $\Rightarrow$ Ill defined: What is similar? Why a partition?
- Similarity entails inherent notion of scale $\Rightarrow$ Hierarchical clustering
- All scales are important $\Rightarrow$ Nested cluster family indexed by scale
- Datasets are very rarely separable into clean partitions
- Some points are. Others could be members of multiple "partitions"



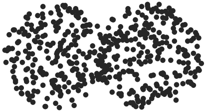


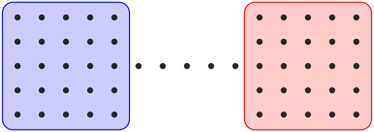
Clustering

- Find node partition such that nodes within partition are similar
  $\Rightarrow$ Ill defined: What is similar? Why a partition?

- Similarity entails inherent notion of scale $\Rightarrow$ Hierarchical clustering
- All scales are important $\Rightarrow$ Nested cluster family indexed by scale

- Datasets are very rarely separable into clean partitions
- Some points are. Others could be members of multiple "partitions"



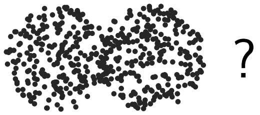


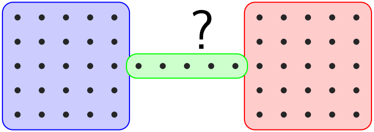
Clustering

- Find node partition such that nodes within partition are similar
  $\Rightarrow$ Ill defined: What is similar? Why a partition?

- Similarity entails inherent notion of scale $\Rightarrow$ Hierarchical clustering
- All scales are important $\Rightarrow$ Nested cluster family indexed by scale

- Datasets are very rarely separable into clean partitions
- Some points are. Others could be members of multiple "partitions"



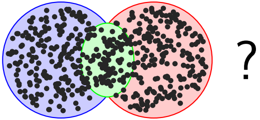


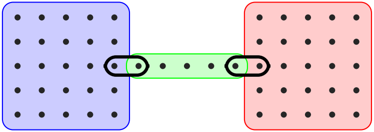
Clustering

- Find node partition such that nodes within partition are similar
  $\Rightarrow$ Ill defined: What is similar? Why a partition?

- Similarity entails inherent notion of scale $\Rightarrow$ Hierarchical clustering
- All scales are important $\Rightarrow$ Nested cluster family indexed by scale

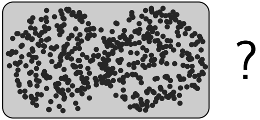
- Datasets are very rarely separable into clean partitions
- Some points are. Others could be members of multiple "partitions"

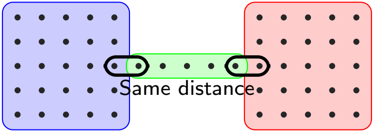




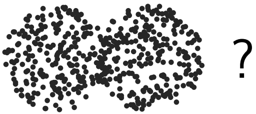
Clustering

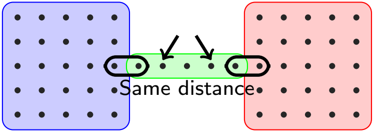
- Find node partition such that nodes within partition are similar
  $\Rightarrow$ Ill defined: What is similar? Why a partition?
- Similarity entails inherent notion of scale $\Rightarrow$ Hierarchical clustering
- All scales are important $\Rightarrow$ Nested cluster family indexed by scale
- Datasets are very rarely separable into clean partitions
- Some points are. Others could be members of multiple "partitions"





- Find node partition such that nodes within partition are similar
  - $\Rightarrow$ Ill defined: What is similar? Why a partition?
- Similarity entails inherent notion of scale $\Rightarrow$ Hierarchical clustering
- All scales are important $\Rightarrow$ Nested cluster family indexed by scale
- Datasets are very rarely separable into clean partitions
- Some points are. Others could be members of multiple “partitions”



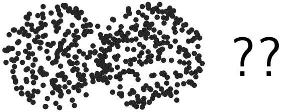


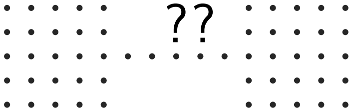
Clustering

- Find node partition such that nodes within partition are similar
  $\Rightarrow$ Ill defined: What is similar? Why a partition?

- Similarity entails inherent notion of scale $\Rightarrow$ Hierarchical clustering
- All scales are important $\Rightarrow$ Nested cluster family indexed by scale

- Datasets are very rarely separable into clean partitions
- Some points are. Others could be members of multiple "partitions"





- Allow classification of some elements into multiple partitions

Hierarchical overlapping clustering

- Points in multiple partitions? $\Rightarrow$ Coverings instead of partitions
- Scale also a problem $\Rightarrow$ Nested family of coverings indexed by scale
- We know that in clustering
  $\Rightarrow$ Equivalences $\Rightarrow$ Partitions $\Rightarrow$ Nested partitions $\Rightarrow$ Ultrametrics
- We will see that in overlapping clustering
  $\Rightarrow$ Tolerances $\Rightarrow$ Coverings $\Rightarrow$ Nested coverings $\Rightarrow$ Cut metrics
- Obtain cut metrics as linear combinations of ultrametrics
- Overlapping clustering is not just an interesting curiosity
- Badly written numbers $\Rightarrow$ Classify in two clusters to avoid mistakes
- Shakespeare's plays, Fletcher's play, and Henry VIII

Clustering (non-hierarchical)

- Network $N = (X, A_X)$ with nodes X and dissimilarities A_X
- Clusters = Partitions = Nonintersecting subsets that cover space X

$$P_X = \{B_1, \ldots, B_m\}, \qquad \bigcup_{i=1}^{m} B_i = X, \quad B_i \cap B_j = \emptyset$$

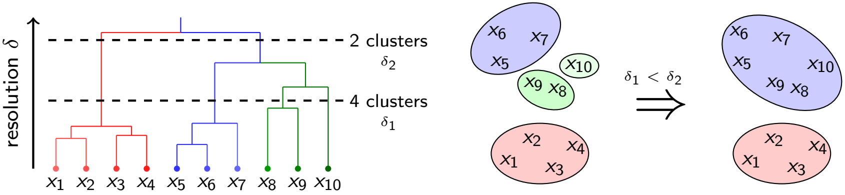
- Equivalence relation: Reflexive ($x \sim x$). Symmetric ($x \sim x' \Leftrightarrow x' \sim x$).
  $\Rightarrow$ Transitive $\Rightarrow x \sim x', x' \sim x'' \Rightarrow x \sim x''$
- A partition is defined by an equivalence relation (converse true as well)
- A partition appears the moment we adopt an equivalence relation

Hierarchical clustering

- Dendrogram $D_X = \{D_X(\delta), \delta \geq 0\}$: collection of partitions at scale δ
- Partitions $D_X(\delta)$ are nested $\Rightarrow \delta \leq \delta'$, $x \sim_\delta x' \Rightarrow x \sim_{\delta'} x'$
- Once two nodes are deemed similar, they stay clustered

- Dendrograms D_X are equivalent to ultrametrics u_X
- u_X: Metric that satisfies the strong triangle inequality

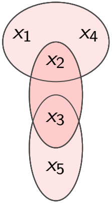
  $\Rightarrow u_X(x, x'') \leq \max\{u_X(x, x'), u_X(x', x'')\}$



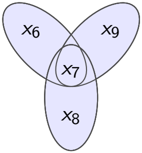
Overlapping clustering

- Clusters = Coverings = Possibly intersecting subsets that cover space X

$$Q_X = \{C_1, \ldots, C_m\}, \quad \bigcup_{i=1}^{m} C_i = X, \quad C_i \cap C_j = C_{ij}$$



- C_{ij} need not be the emptyset $\emptyset$
- Tolerance relation:
  - $\Rightarrow$ Reflexive $(x \leftrightarrow x)$
  - $\Rightarrow$ Symmetric $(x \leftrightarrow x' \Leftrightarrow x' \leftrightarrow x)$
  - $\Rightarrow$ Not transitive



- Tolerance relations induce coverings

Hierarchical overlapping clustering

Theorem

If, for each $\delta \geq 0$, the covering $K_X(\delta)$ is induced by the tolerance relation obtained from a cut metric

$$c_X(x, x') \leq \delta \Rightarrow x \leftrightarrow_\delta x'$$

Then the collection of coverings $K_X = \{K_X(\delta), \delta \geq 0\}$ is nested.

- Coverings $K_X(\delta)$ are nested $\Rightarrow \delta \leq \delta'$, $x \leftrightarrow_\delta x' \Rightarrow x \leftrightarrow_{\delta'} x'$
- Once two nodes become related, it cannot be undone
- Cut metric: Similar role to ultrametrics in building equivalence relations
- Nested collection of coverings: Analogous to dendrograms

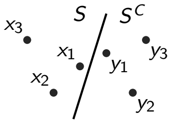
- First, define a cut semimetric $\delta_S(x, x')$ of a subset S of the node set

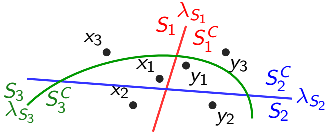
$$\delta_S(x, x') = \mathbb{I}\left\{S \cap \{x, x'\} \neq \emptyset\right\} \mathbb{I}\left\{S^C \cap \{x, x'\} \neq \emptyset\right\}$$

- Cuts the node set in two: unit distance for nodes in opposite sides
- Define cut metric: c_X. Conic combination of cut semimetrics

$$c_X(x, x') = \sum_{S \subseteq X} \lambda_S \delta_S(x, x') \ , \ \lambda_S \geq 0$$

- All possible subsets $S \subseteq X$, each one with different weight λ_S





Construction of cut metrics

Theorem

A convex combination of ultrametrics results in a cut metric

$$c_X(x, x') = \sum k_i u_{X,i}(x, x') \ , \ \sum k_i = 1 \ , \ k_i \geq 0$$

- We know how to obtain ultrametrics $\Rightarrow$ Hierarchical clustering $\mathcal{H}$
- Dithering: Perturb the dissimilarity function with random noise
- Get ultrametric $\tilde{u}_X(x, x')$ of perturbed network by applying $\mathcal{H}$
- Get cut metric combining the ultrametrics

$$c_X(x, x') = \mathbb{E}[\tilde{u}_X(x, x')]$$

Parallelism

Method	Hierarchical non-Overlapping: $\mathcal{H}$	Overlapping: $\mathcal{O}$
Metric	Ultrametric: $u_X(x, x')$	Cut Metric: $c_X(x, x')$
Relation	Equivalence: $\sim$	Tolerance: $\leftrightarrow$
Grouping	Partition: $P_X = \{B_i\}$	Covering: $Q_X = \{C_i\}$
Hierarchy	Dendrogram: D_X	Nested Covering: K_X

- Hierarchical Non-Overlapping clustering
  $\Rightarrow$ Equivalences $\Rightarrow$ Partitions $\Rightarrow$ Dendrograms $\Rightarrow$ Ultrametrics
- Hierarchical Overlapping clustering
  $\Rightarrow$ Tolerances $\Rightarrow$ Coverings $\Rightarrow$ Nested coverings $\Rightarrow$ Cut metrics

Overlapping

- Overlapping function $f_{\mathrm{ol}} : \mathbb{R}_+ \to \mathbb{N}_0$
- Helps in selecting relevant resolutions δ to observe
- For each δ, counts the number of overlapping nodes

$$f_{\mathrm{ol}}(\delta) = \sum_{k=1}^{n} \mathbb{I}\left\{C_i \cap C_j = \{x_k\}, i \neq j,\ i,j = 1, \ldots, m(\delta)\right\}$$

- We use f_{ol} to define clusterability of a dataset
- $f_{\mathrm{ol}}(\delta) = 0$ for some meaningful δ $\Rightarrow$ no overlap $\Rightarrow$ partition

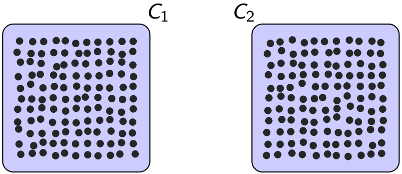
  $\Rightarrow$ Cannot be $\delta = 0$ $\Rightarrow$ $f_{\mathrm{ol}}(0) = 0$ but all nodes separated

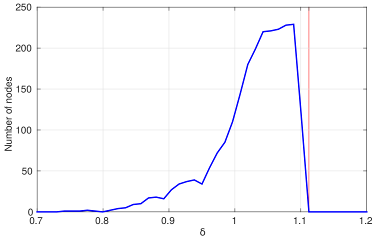
  $\Rightarrow$ Cannot be large δ $\Rightarrow$ $f_{\mathrm{ol}}(\delta) = 0$ but all nodes together

- In general, we are interested in coverings with small overlap

Example: two clouds

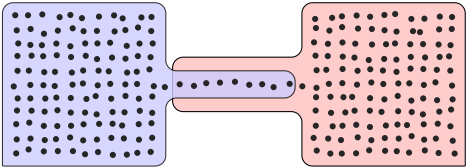
- Setting: $d = 1, D = 13$. Dissimilarity: distance between points
- This dataset has two evident clusters
- Dithering: 100 realizations.
- Gaussian noise of power: $10^{-1} \times$min distance
- Hierarchical non-overlapping clustering $\mathcal{H}$: single linkage
- Overlapping function, $\delta = 1.11$ $\Rightarrow$ Similar to d

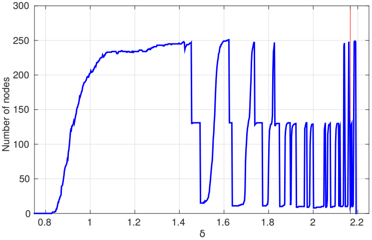




Example: dumbbell network

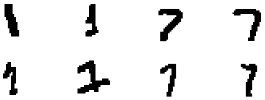
- Setting: $d = 1$. Dissimilarity: distance between points
- There are no two clear clusters
- Dithering: 100 realizations.
- Gaussian noise of power: $1\times$min distance
- $\mathcal{H}$: single linkage $\Rightarrow$ ultrametric
- Overlapping function, $\delta = 2.17$



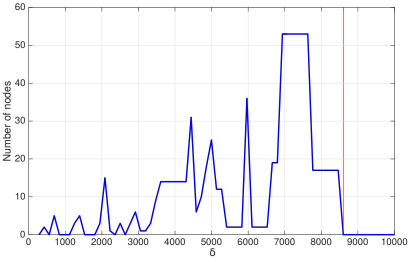


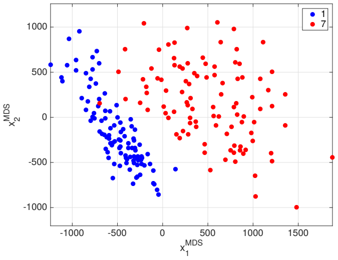
Two digit classification

- Digits: 1, 7. 100 each
- 20 PCA components
- $\mathcal{H}$: Ward $\Rightarrow$ ultrametric
- Dithering: 100 realizations



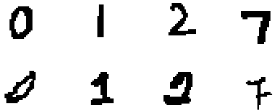
- Gaussian noise of power: $10^{-2} \times$min PCA distance
- Output: $\{1(\times 100), 7\}$, $\{7(\times 99)\}$ $\Rightarrow$ 0.5% error rate



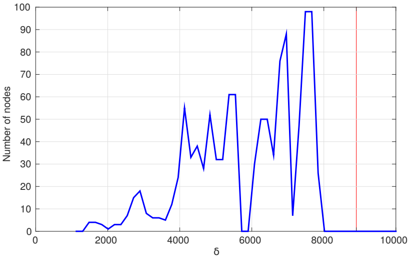


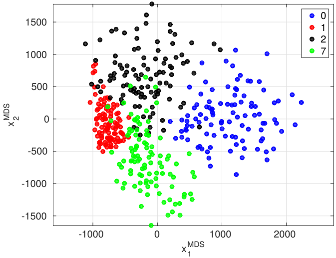
Four digit classification

- Digits: 0, 1, 2, 7. 100 each
- 20 PCA components
- $\mathcal{H}$: Ward $\Rightarrow$ ultrametric
- Dithering: 100 realizations



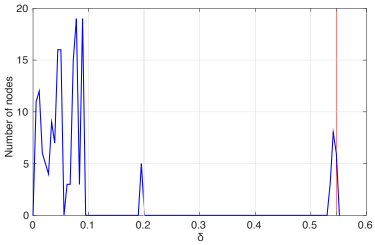
- Gaussian noise of power: $5 \cdot 10^{-3} \times$min PCA distance
- Output: $\{0(\times 100), 2(\times 3)\}$, $\{1(\times 99), 7(\times 2)\}$, $\{1, 2(\times 86), 7(\times 3)\}$, $\{2(\times 11), 7(\times 95)\}$ $\Rightarrow$ 5% error rate

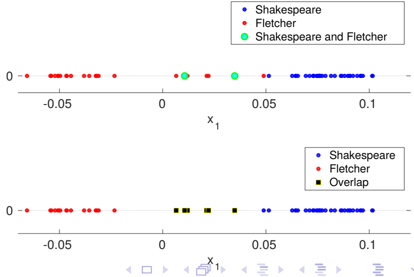




Shakespeare and Fletcher

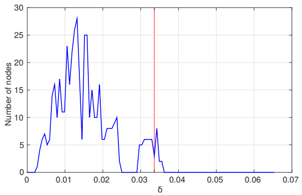
- Word adjacency networks $\Rightarrow$ Author profiles
  $\Rightarrow$ Classify plays by author $\Rightarrow$ Identify co-authored plays
- Dissimilarity: Distance from play to profile
- $\mathcal{H}$: Ward. Dithering: 100 realizations
- Gaussian noise of power: $4\times$min distance
- Overlap: 2 co-authored plays and 4 Fletcher plays
  {S (x 33), F (x1), F (x 4), S&F (x2)}; {F (x 16), F (x4), S&F (x 2)}



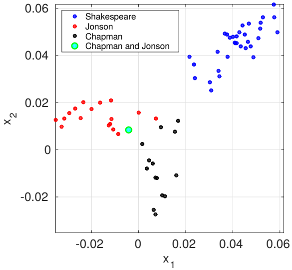


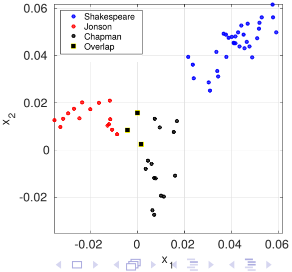
Shakespeare, Chapman and Jonson

- $\mathcal{H}$: average $\Rightarrow$ ultrametric
- Gaussian noise
- Noise power: $1\times$min distance
- Dithering: 100 realizations
- Output:



- Shakespeare plays classified correctly: {S (x33)}
- Overlap: {J (x16), C&J}; {J, C (x13), J, C}; {J, C, C&J}





Conclusions

- Hierarchical: Collection of groups. Levels of similarity
- Overlapping: Allow nodes to belong to more than one cluster
- Achieved through the use of cut metrics to get nested coverings
- Get cut metrics from ultrametrics through dithering
- Identify nodes that have traits of more than one group
- Applicable to data that is not partitionable
- Definition of overlapping function $\Rightarrow$ Notion of clusterability
- Synthetic examples $\Rightarrow$ General intuition and properties
- Handwritten digit classification $\Rightarrow$ Partitionable dataset
- Authorship Attribution $\Rightarrow$ Co-authored plays