EECS 445

Introduction to Machine Learning

Hierarchical Clustering

Prof. Kutty

Announcements

- HW3 is out – start early!

https://forms.gle/ffiBvNbPjHF8ghi77

clustering

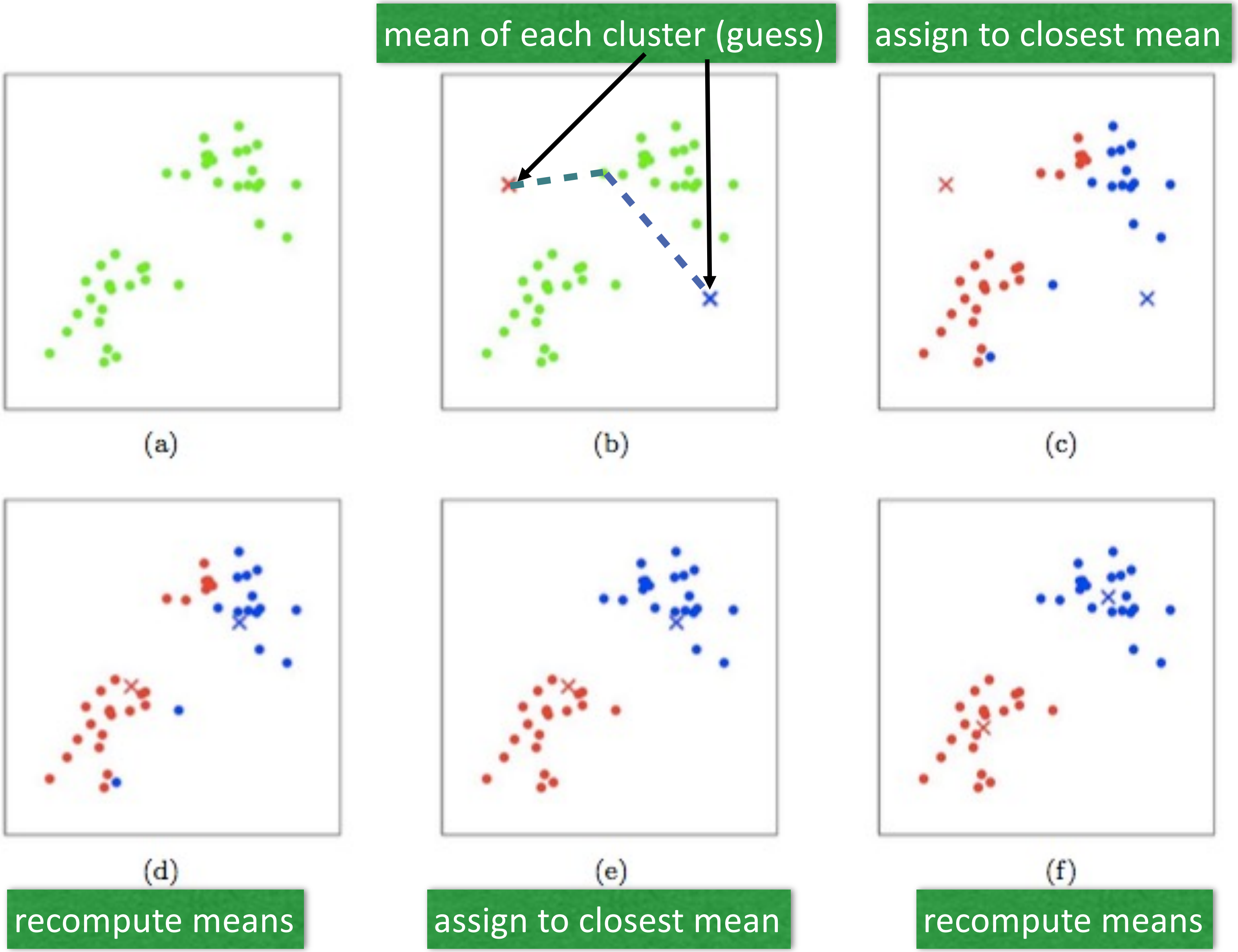
review

(hard) clustering algorithms

- k-means clustering
- spectral clustering
- agglomerative clustering

k-means Clustering

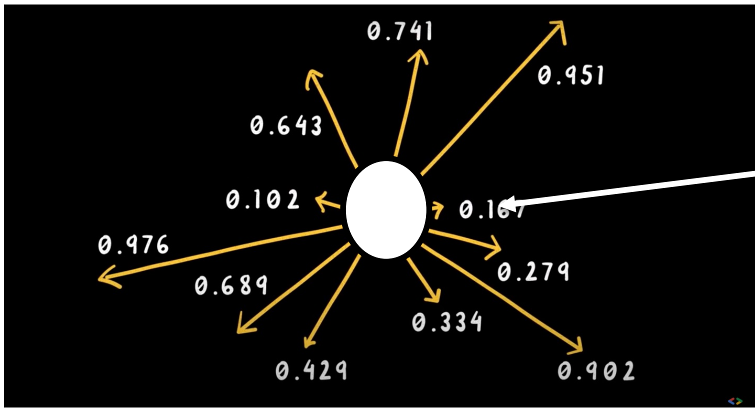
k is a hyperparameter; $k = 2$



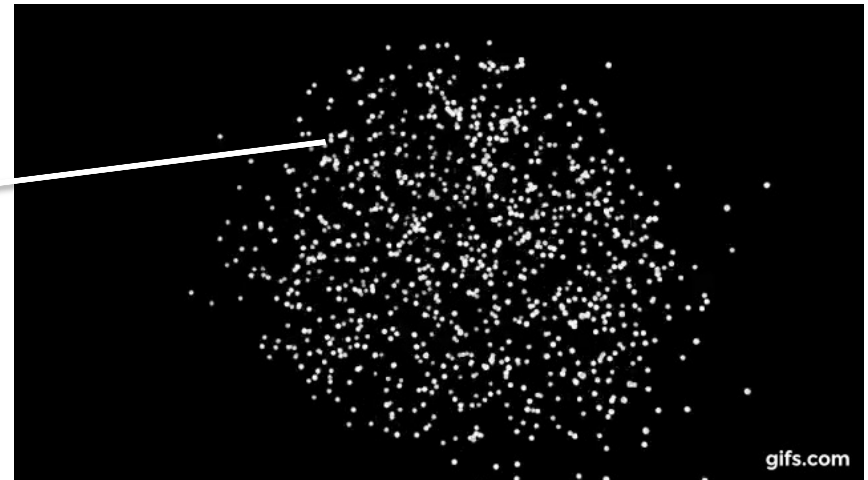
spectral clustering

review

spectral clustering: Big idea



could be high dimensional data

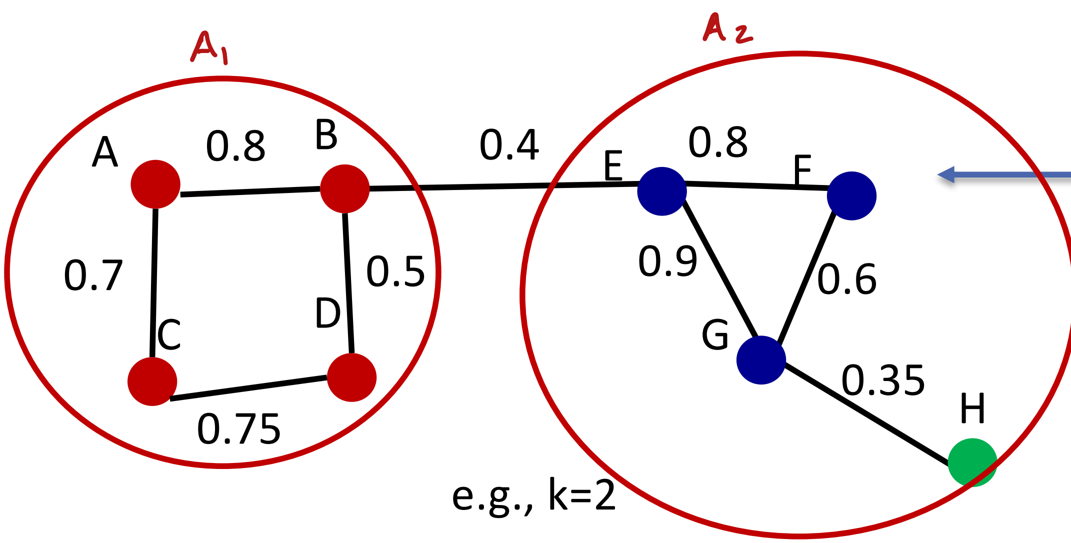


for each pair of datapoints

symmetric
$+ \; w_{ij} \geq 0$

$$w_{ij} = exp\left\{-\frac{\left\|\bar{x}^{(i)} - \bar{x}^{(j)}\right\|^2}{2\sigma^2}\right\}$$

$0 < w_{ij} \leq 1$



e.g., k=2

$$\min_{A_1,\ldots,A_k} RatioCut(A_1, \ldots, A_k)$$

$A_1 = \{A, B, C, D\}$
$A_2 = \{E, F, G\, H\}$

$A_1 \cup A_2 \cup \ldots \cup A_k = V$

$i \neq j \quad A_i \cap A_j = \emptyset$

Spectral Clustering for k partitions

given: $S_n = \{\bar{x}^{(i)}\}_{i=1}^{n}$

Input: valid similarity metric, number of clusters k

1. Build adjacency matrix $W \longrightarrow n \times n$

2. Compute graph Laplacian $L = D - W$

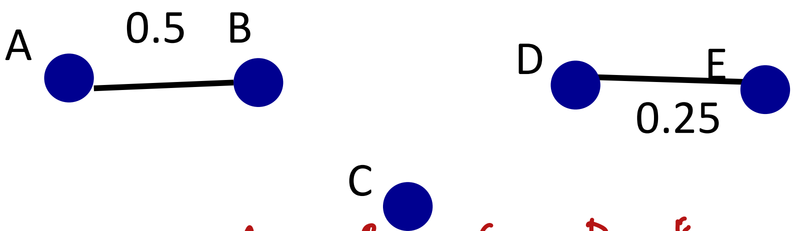
3. Compute (eigenvector, eigenvalue) pairs of L

4. Build matrix with the first k eigenvectors (corresponding to the k smallest eigenvalues) as columns interpret rows as new data points
Low dimensional embedding ($\in \mathbb{R}^k$) of the original dataset ($\in \mathbb{R}^d$)

5. Apply k-means to new data representation

Output: clusters assignments

Spectral Clustering Example #1



$$W = \begin{array}{cc} \begin{array}{ccccc} A & B & C & D & E \end{array} & \\ \begin{bmatrix} 1 & 0.5 & 0 & 0 & 0 \\ 0.5 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.25 \\ 0 & 0 & 0 & 0.25 & 1 \end{bmatrix} & \begin{array}{c} A \\ B \\ C \\ D \\ E \end{array} \end{array}$$

$$L = \begin{bmatrix} 0.5 & -0.5 & 0 & 0 & 0 \\ -0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.25 & -0.25 \\ 0 & 0 & 0 & -0.25 & 0.25 \end{bmatrix}$$

Input: similarity metric, number of clusters k

1. Build adjacency matrix W
2. Compute graph Laplacian $L = D - W$

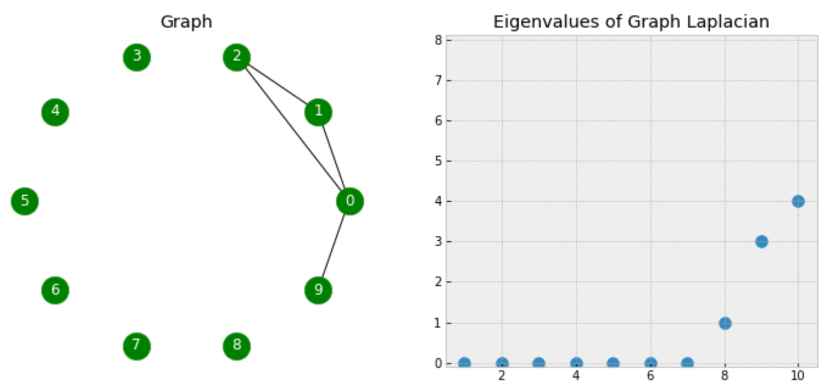
Graph Laplacian L

eigenvalue eigenvector
if $\lambda, \bar{f}$ of L
$\iff L\bar{f} = \lambda \bar{f}$

The graph Laplacian is the matrix $L = D - W$

Properties of L

- L is symmetric
- L is PSD
  - L has n non-negative real-valued eigenvalues
$$0 \leq \lambda_1 \leq \cdots \leq \lambda_n$$
- Fact: the multiplicity of the eigenvalue 0 is the number of connected components in the graph
- Examples



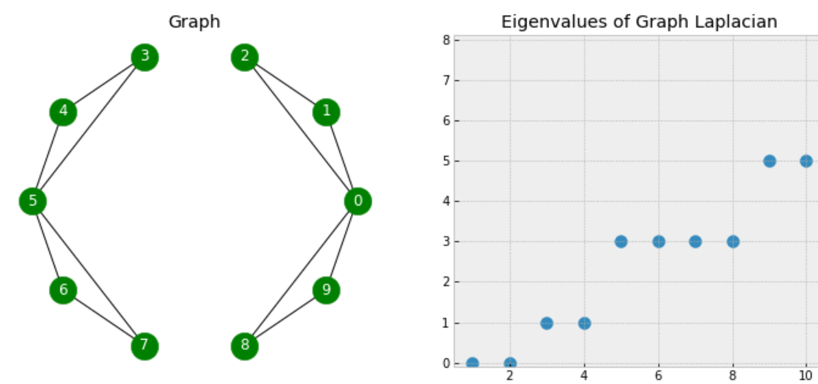
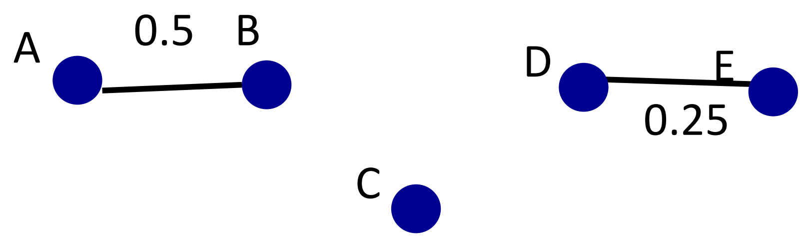


image from: https://towardsdatascience.com/spectral-clustering-aba2640c0d5b

Spectral Clustering Example #1



$$L = \begin{bmatrix} 0.5 & -0.5 & 0 & 0 & 0 \\ -0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.25 & -0.25 \\ 0 & 0 & 0 & -0.25 & 0.25 \end{bmatrix}$$

Input: similarity metric, number of clusters k

1. Build adjacency matrix W
2. Compute graph Laplacian $L = D - W$
3. Compute eigenvectors/eigenvalues of L

(eigenvector, eigenvalue) pairs

$$\begin{pmatrix} -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \\ 0 \\ 0 \\ 0 \end{pmatrix}, 0 \qquad \begin{pmatrix} 0 \\ 0 \\ 1.00 \\ 0 \\ 0 \end{pmatrix}, 0 \qquad \begin{pmatrix} 0 \\ 0 \\ 0 \\ -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix}, 0 \qquad \begin{pmatrix} 0 \\ 0 \\ 0 \\ -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}, 0.5 \qquad \begin{pmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \\ 0 \\ 0 \\ 0 \end{pmatrix}, 1$$

Spectral Clustering Example #1

Input: similarity metric, number of clusters k

1. Build adjacency matrix W

2. Compute graph Laplacian $L = D - W$

3. Compute eigenvectors of L

4. Build matrix with the first k eigenvectors (corresponding to the *k* smallest eigenvalues) as columns interpret rows as new data points

5. Apply k-means to new data representation

Output: clusters assignments

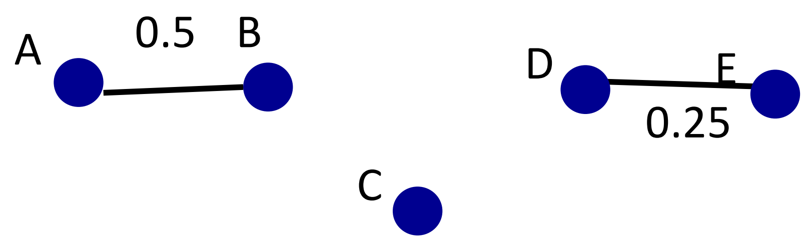
Eigenvectors

$$\begin{array}{c} A \\ B \\ C \\ D \\ E \end{array} \begin{bmatrix} -\frac{\sqrt{2}}{2} & 0 & 0 & 0 & -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & 0 & 0 & 0 & \frac{\sqrt{2}}{2} \\ 0 & 1.00 & 0 & 0 & 0 \\ 0 & 0 & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \end{bmatrix}$$

Eigenvalues

$$[0, 0, 0, 0.5, 1.0]$$

Spectral Clustering Example #1



$$L = \begin{bmatrix} 0.5 & -0.5 & 0 & 0 & 0 \\ -0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.25 & -0.25 \\ 0 & 0 & 0 & -0.25 & 0.25 \end{bmatrix}$$

Eigenvectors

$$\begin{array}{c} \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \\ \text{E} \end{array} \begin{bmatrix} -\frac{\sqrt{2}}{2} & 0 & 0 & 0 & -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & 0 & 0 & 0 & \frac{\sqrt{2}}{2} \\ 0 & 1.00 & 0 & 0 & 0 \\ 0 & 0 & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \end{bmatrix}$$

Eigenvalues

$$[0, 0, 0, 0.5, 1.0]$$

$\boxed{k = 3}$

k dimensional embedding

A = $\left[-\frac{\sqrt{2}}{2} \quad 0 \quad 0\right]^T$

B = $\left[-\frac{\sqrt{2}}{2} \quad 0 \quad 0\right]^T$

C = $\left[0 \quad 1.00 \quad 0\right]^T$

D = $\left[0 \quad 0 \quad -\frac{\sqrt{2}}{2}\right]^T$

E = $\left[0 \quad 0 \quad -\frac{\sqrt{2}}{2}\right]^T$

Run k-means clustering on this embedding

Return clusters

Spectral Clustering:

intuition for $k = 2$

Example: for $k = 2$

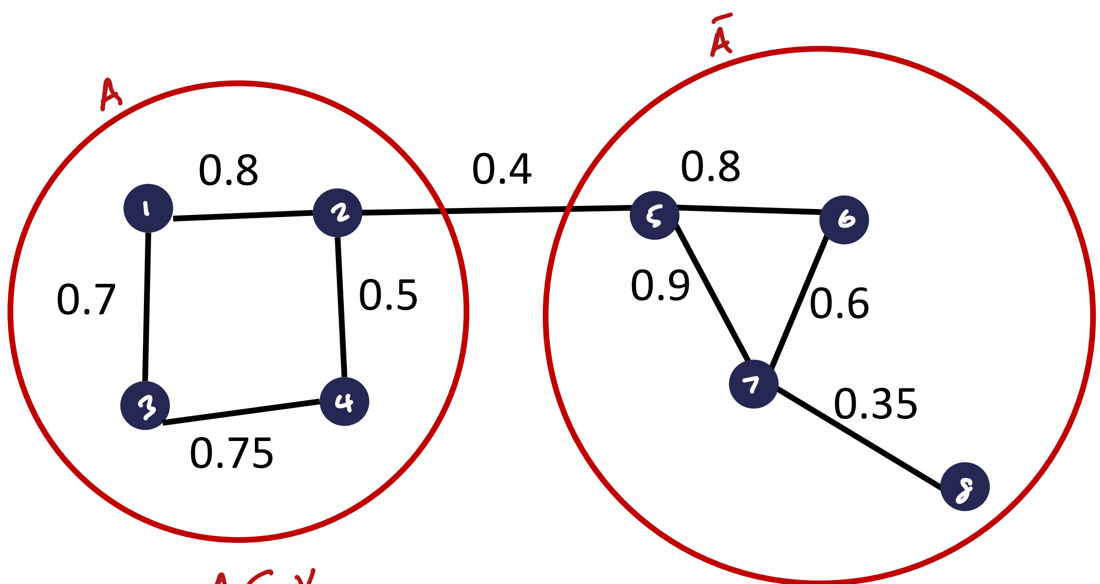
Specify a cut as an indicator vector $\bar{f} \in \mathbb{R}^n$ where for cut A and $\bar{A}$

in this example

$$f_i = \begin{cases} \sqrt{|\overline{A}|/|A|} \quad 1 & \text{if } v_i \in A \\ -\sqrt{|A|/|\overline{A}|} \quad -1 & \text{if } v_i \in \overline{A}. \end{cases}$$

$$\bar{f} = \begin{array}{c} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix} \end{array} \begin{array}{l} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \\ v_8 \end{array}$$

e.g.,



$A \subseteq V$

$A = \{v_1, v_2, v_3, v_4\}$ $\quad$ $\bar{A} = \{v_5, v_6, v_7, v_8\}$

Example: for $k = 2$

Specify a cut as an indicator vector $\bar{f} \in \mathbb{R}^n$ where for cut A and $\bar{A}$

$$f_i = \begin{cases} \sqrt{|\overline{A}|/|A|} & \text{if } v_i \in A \\ -\sqrt{|A|/|\overline{A}|} & \text{if } v_i \in \overline{A}. \end{cases}$$

Can show that $\bar{f}^T L \bar{f} \propto \ RatioCut(A, \bar{A})$

and $\sum_i f_i = 0$ and $\|\bar{f}\|^2 = n$

Example: for $k = 2$

Goal: $\min_{\bar{f}} \bar{f}^T L \bar{f}$ subject to $\sum_i f_i = 0$ and $\|\bar{f}\|^2 = n$
where $\bar{f}$ is an indicator vector as defined

NP hard

eigenvalue eigenvector
if $\lambda, \bar{f}$ of L
$\Leftrightarrow L\bar{f} = \lambda \bar{f}$

Modified Goal:

$$\min_{\bar{f} \in \mathbb{R}^n} \bar{f}^T L \bar{f} \text{ subject to } \sum_i f_i = 0 \text{ and } \|\bar{f}\|^2 = n$$

solution: eigenvector corresponding to the (second*) smallest eigenvalue of L
Why?

$\bar{f}^T L \bar{f} = \bar{f}^T \lambda \bar{f} = \lambda \underbrace{\|\bar{f}\|^2}_{\text{constant}}$

For k > 2, similar idea. Use multiple indicator vectors

* see examples

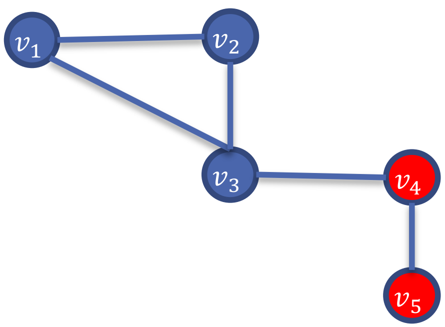
Why does this work?
Spectral Clustering for k partitions

Idea:

$$\min_{A_1,\ldots,A_k} RatioCut(A_1, \ldots, A_k)$$

$$\approx$$

$$\underbrace{\min_{\bar{f}^{(1)},\ldots,\bar{f}^{(k)}}}_{\text{indicator vectors}} \mathrm{Tr}(F^T L F) \qquad \text{s.t. } F^T F = I$$

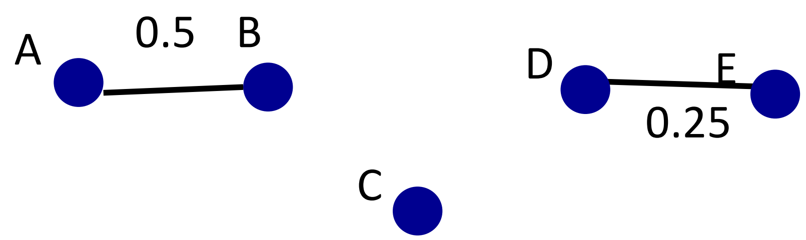


$$\approx$$

$$\underbrace{\min_{\bar{f}^{(1)},\ldots,\bar{f}^{(k)}}}_{\text{real vectors}} \mathrm{Tr}(F^T L F) \qquad \text{s.t. } F^T F = I$$

The solution to this is the first k eigenvectors of L

Spectral Clustering Example #1



$$L = \begin{bmatrix} 0.5 & -0.5 & 0 & 0 & 0 \\ -0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.25 & -0.25 \\ 0 & 0 & 0 & -0.25 & 0.25 \end{bmatrix}$$

Eigenvectors

$$\begin{array}{c} A \\ B \\ C \\ D \\ E \end{array} \begin{bmatrix} -\frac{\sqrt{2}}{2} & 0 & 0 & 0 & -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & 0 & 0 & 0 & \frac{\sqrt{2}}{2} \\ 0 & 1.00 & 0 & 0 & 0 \\ 0 & 0 & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \end{bmatrix}$$

Eigenvalues

$$[0, 0, 0, 0.5, 1.0]$$

$\boxed{k = 3}$

k dimensional embedding

$A = \left(-\frac{\sqrt{2}}{2} \quad 0 \quad 0\right)^T$

$B = \left(-\frac{\sqrt{2}}{2} \quad 0 \quad 0\right)^T$

$C = \left(0 \quad 1.00 \quad 0\right)^T$

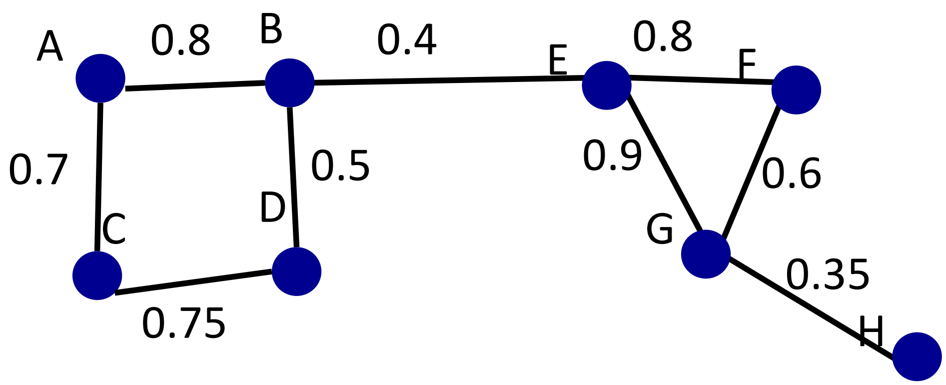
$D = \left(0 \quad 0 \quad -\frac{\sqrt{2}}{2}\right)^T$

$E = \left(0 \quad 0 \quad -\frac{\sqrt{2}}{2}\right)^T$

Run k-means clustering on this embedding

Return clusters

Spectral Clustering Example #2



Should be able to construct
W, D, L

Eigenvectors

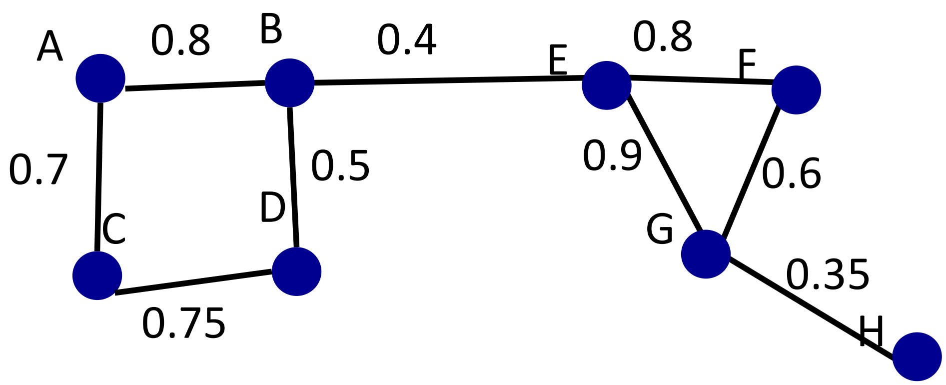
−0.3536	−0.3467	0.0704	0.6000	−0.2702	0.0594	0.4652	−0.3113
−0.3536	−0.2418	−0.0252	0.4060	0.5839	0.0496	−0.3458	0.4337
−0.3536	−0.3997	0.1334	−0.2603	−0.6097	−0.1178	−0.4406	0.2233
−0.3536	−0.3773	0.1106	−0.6251	0.4141	0.0655	0.3339	−0.2059
−0.3536	0.2188	−0.3571	0.0256	0.0968	−0.2652	−0.4338	−0.6566
−0.3536	0.2926	−0.4542	−0.1096	−0.1732	0.7125	0.0785	0.1654
−0.3536	0.3251	−0.2362	−0.0592	−0.0571	−0.6200	0.4008	0.4021
−0.3536	0.5291	0.7582	0.0227	0.0154	0.1160	−0.0583	−0.0509

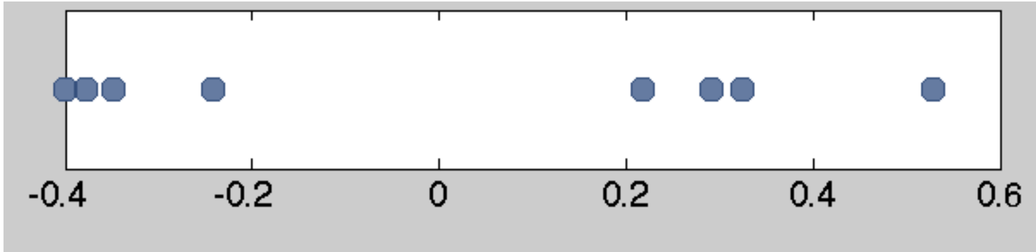
Eigenvalues

0.0000	0.1349	0.4590	1.2624	1.6492	2.2200	2.7575	3.1170

Spectral Clustering Example #2.1

k = 2





Eigenvectors

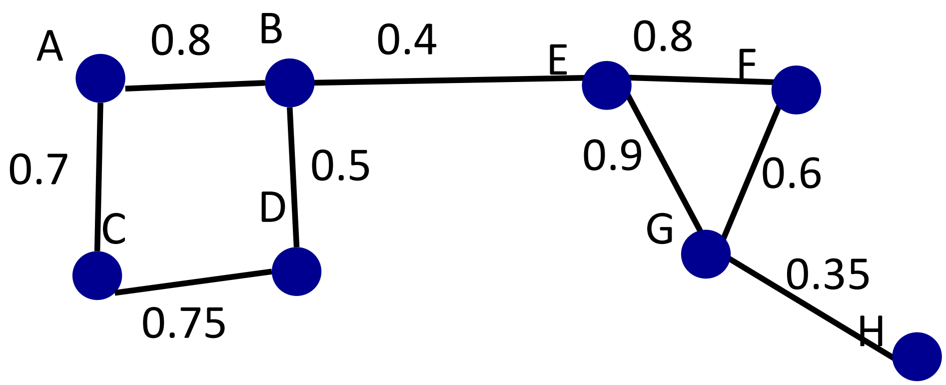
A	−0.3536	−0.3467	0.0704	0.6000	−0.2702	0.0594	0.4652	−0.3113
B	−0.3536	−0.2418	−0.0252	0.4060	0.5839	0.0496	−0.3458	0.4337
C	−0.3536	−0.3997	0.1334	−0.2603	−0.6097	−0.1178	−0.4406	0.2233
D	−0.3536	−0.3773	0.1106	−0.6251	0.4141	0.0655	0.3339	−0.2059
E	−0.3536	0.2188	−0.3571	0.0256	0.0968	−0.2652	−0.4338	−0.6566
F	−0.3536	0.2926	−0.4542	−0.1096	−0.1732	0.7125	0.0785	0.1654
G	−0.3536	0.3251	−0.2362	−0.0592	−0.0571	−0.6200	0.4008	0.4021
H	−0.3536	0.5291	0.7582	0.0227	0.0154	0.1160	−0.0583	−0.0509

Eigenvalues

0.0000	0.1349	0.4590	1.2624	1.6492	2.2200	2.7575	3.1170

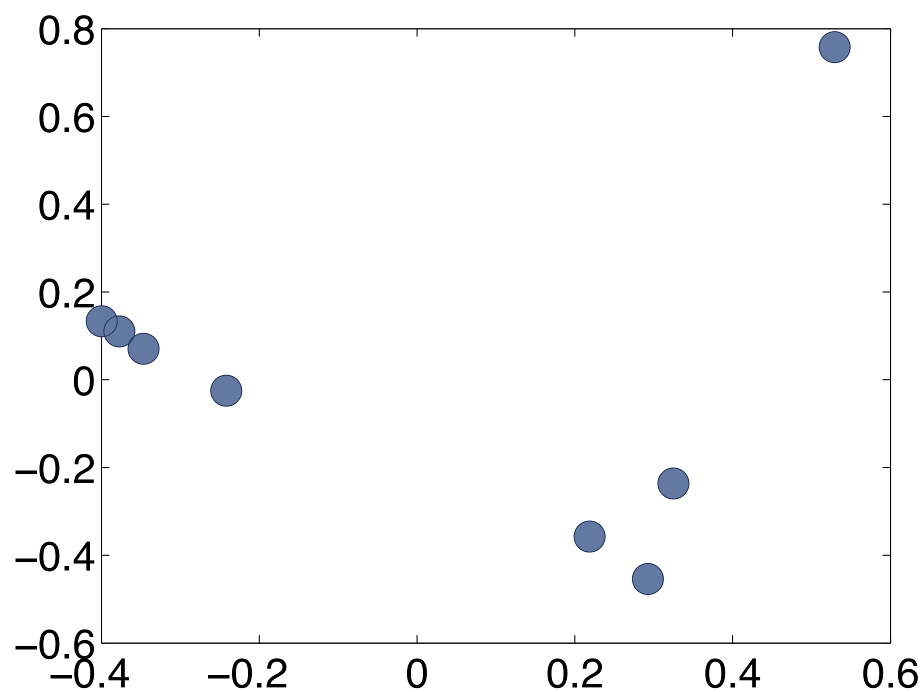
Spectral Clustering Example #2.2

k = 3



Eigenvectors

A	-0.3536	-0.3467	0.0704	0.6000
B	-0.3536	-0.2418	-0.0252	0.4060
C	-0.3536	-0.3997	0.1334	-0.2603
D	-0.3536	-0.3773	0.1106	-0.6251
E	-0.3536	0.2188	-0.3571	0.0256
F	-0.3536	0.2926	-0.4542	-0.1096
G	-0.3536	0.3251	-0.2362	-0.0592
H	-0.3536	0.5291	0.7582	0.0227



Eigenvalues

0.0000	0.1349	0.4590	1.2624	1.6492	2.2200	2.7575	3.1170

agglomerative hierarchical clustering

Clustering

Angell Hall

Mason Hall

FXB

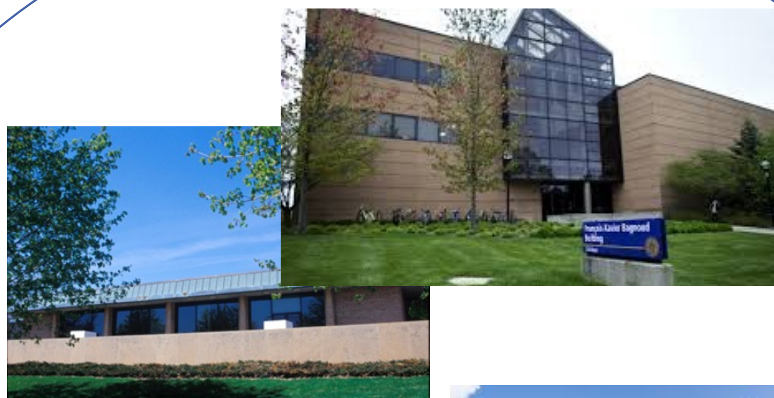
CSE

GGBL

Chrysler Center

Clustering

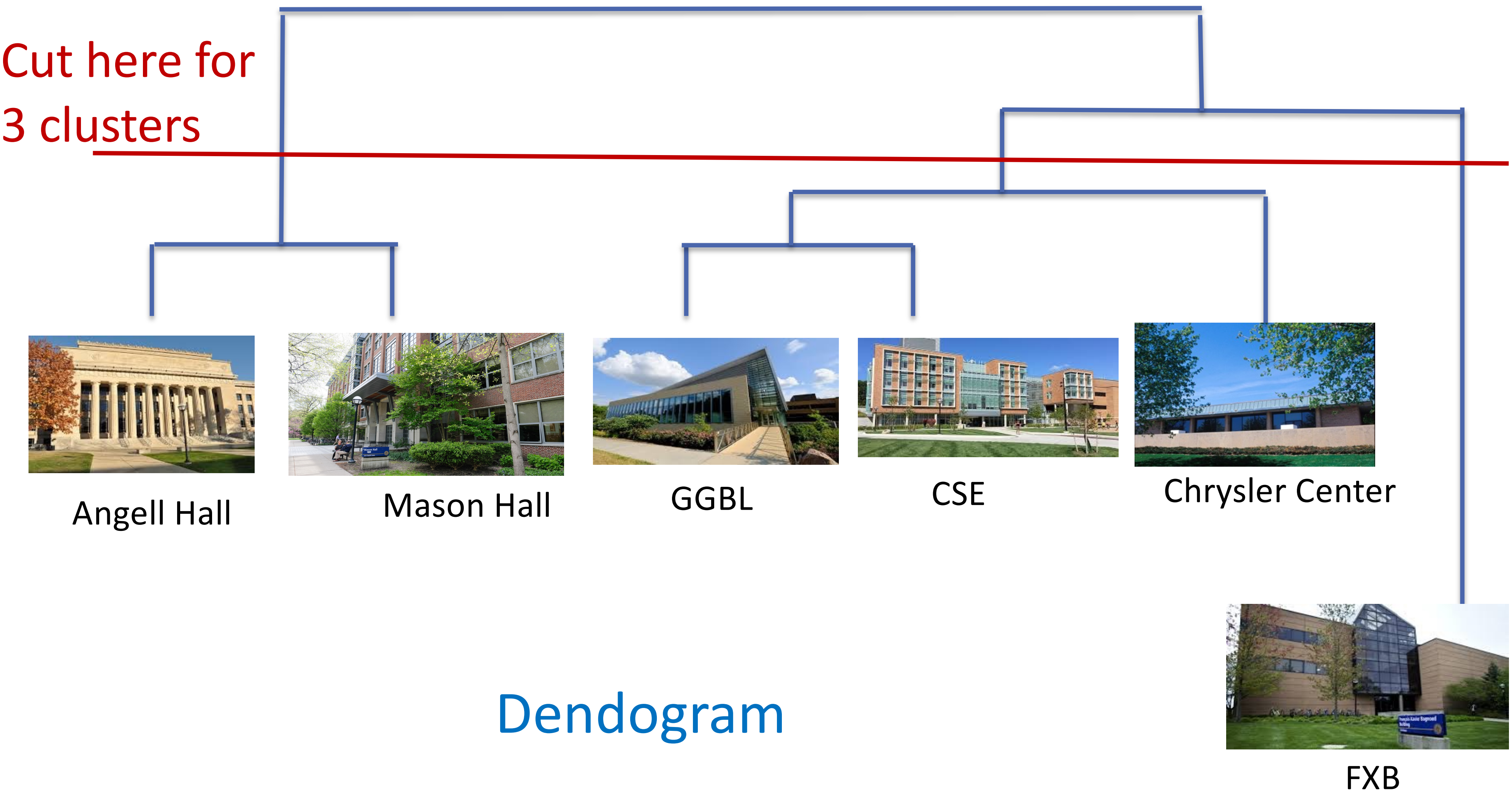
Partition



Doesn't return additional information on how the clusters relate to one another.

Hierarchical Clustering

Idea: start with each pt. in its own cluster and build a hierarchy



Agglomerative Clustering

1. Assign each pt. its own cluster

   for each point *i*, $C_i = \{\bar{x}^{(i)}\}$

2. Find the closest clusters & merge, repeat until convergence

   $\arg\min_{i,j} d(C_i, C_j)$ where *i* ≠ *j*

Notation: C_i set of all points in cluster *i*

Linkage Criteria

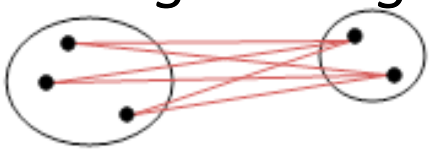
Single-linkage: $D\left(C_i, C_j\right) = \min_{\bar{x} \in C_i; \bar{x}' \in C_j} d(\bar{x}, \bar{x}')$

susceptible to outliers and *chaining phenomenon:* larger clusters are naturally biased toward having closer distances to other points and will therefore attract a successively larger number of points

Complete-linkage: $D\left(C_i, C_j\right) = \max_{\bar{x} \in C_i; \bar{x}' \in C_j} d(\bar{x}, \bar{x}')$

susceptible to outliers but tends to merge clusters so that all clusters tend to have the same diameter

Average-linkage: $D\left(C_i, C_j\right) = \frac{1}{|C_i|} \frac{1}{|C_j|} \sum_{\bar{x} \in C_i; \bar{x}' \in C_j} d(\bar{x}, \bar{x}')$

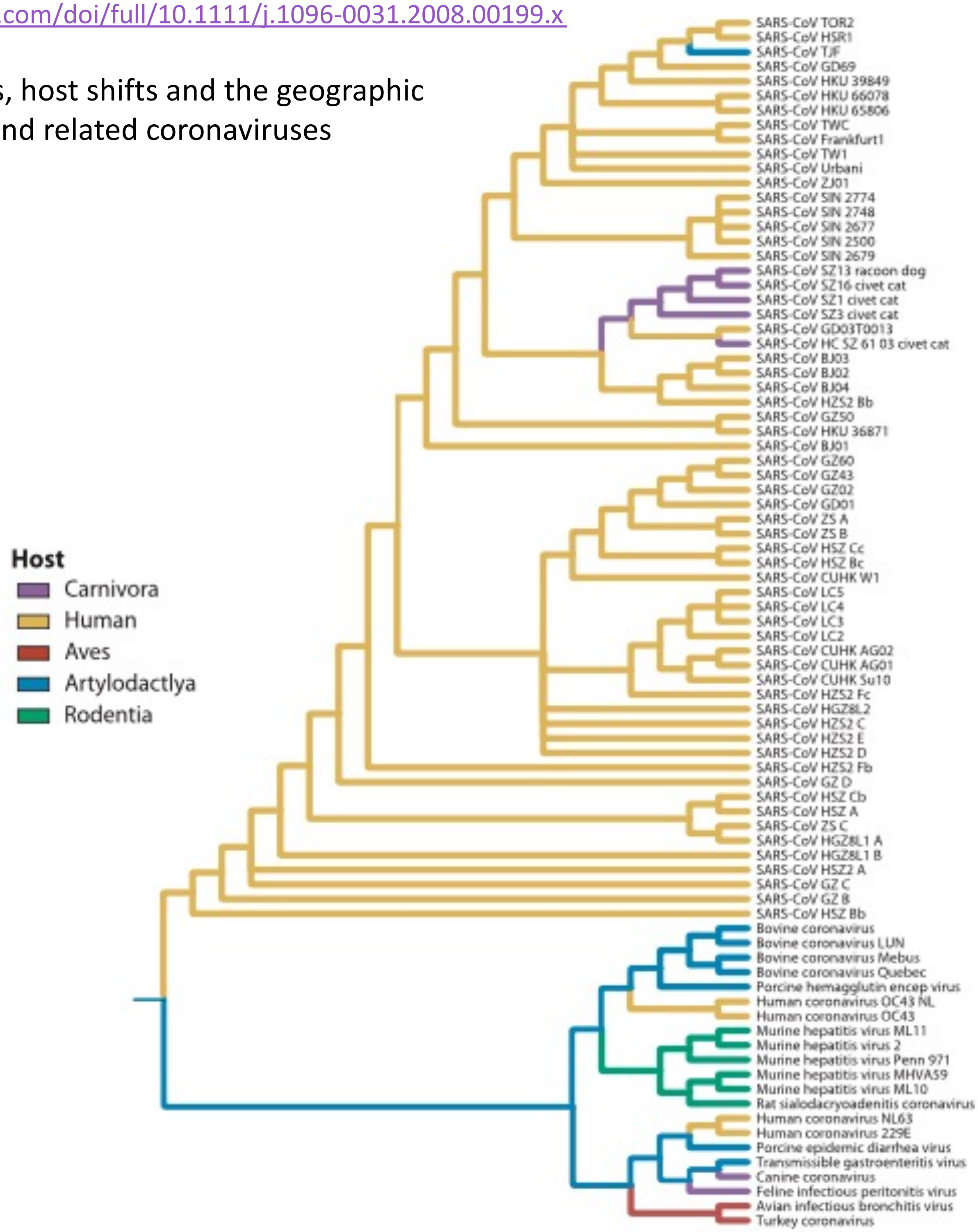


less susceptible to outliers
inefficiency concerns

Notation: C_i set of all points in cluster *i* https://onlinelibrary.wiley.com/doi/full/10.1111/j.1096-0031.2008.00199.x

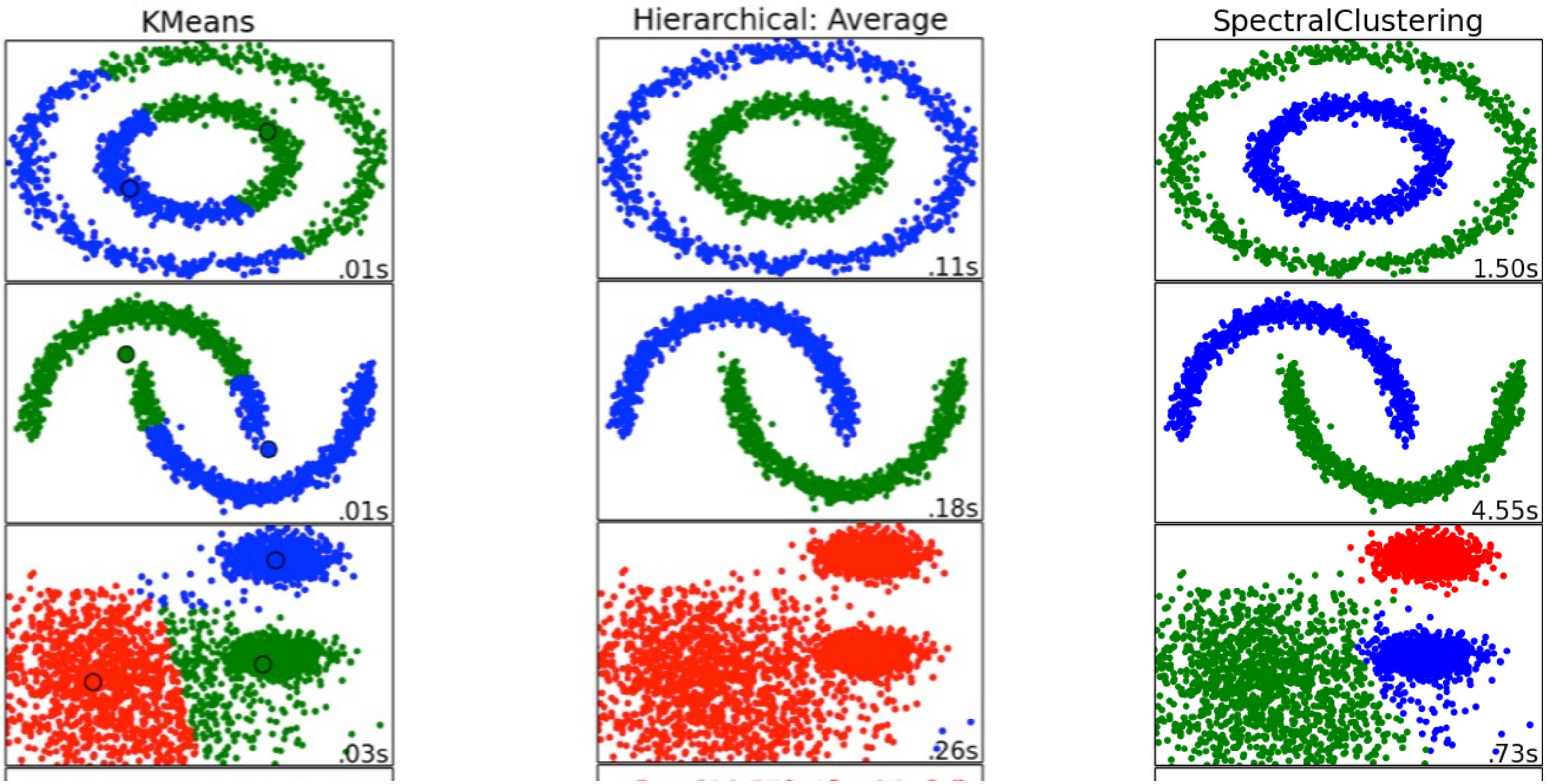
Evolution of genomes, host shifts and the geographic spread of SARS-CoV and related coronaviruses

Janies et. al



Clustering Algorithms

http://scikit-learn.org/stable/modules/clustering.html



Properties of Clustering Algorithms

K-means:
-fast, linear in all relevant quantities
-can be affected by outliers
-problems arise when clusters are different sizes, densities and shapes

Hierarchical:
-deterministic, don't need to decide # of clusters ahead of time
-only input parameters are distance measure and linkage criteria
-slow

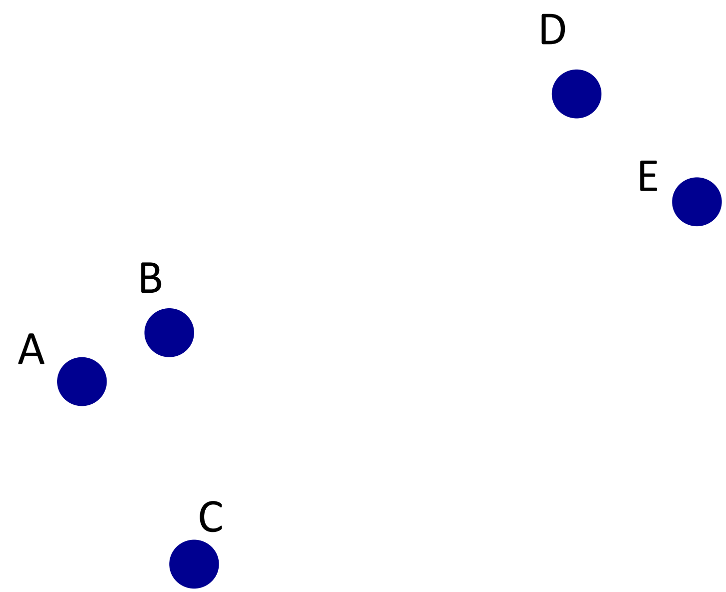
Spectral:
- high performance with smaller datasets
- not very efficient although faster approximate variants exist

  Donghui Yan, Ling Huang, and Michael I. Jordan. 2009. Fast approximate spectral clustering.

Example

- For this dataset with agglomerative clustering and single linkage criteria, the 3 clusters are

https://forms.gle/ffiBvNbPjHF8ghi77



assume to scale and euclidean distance

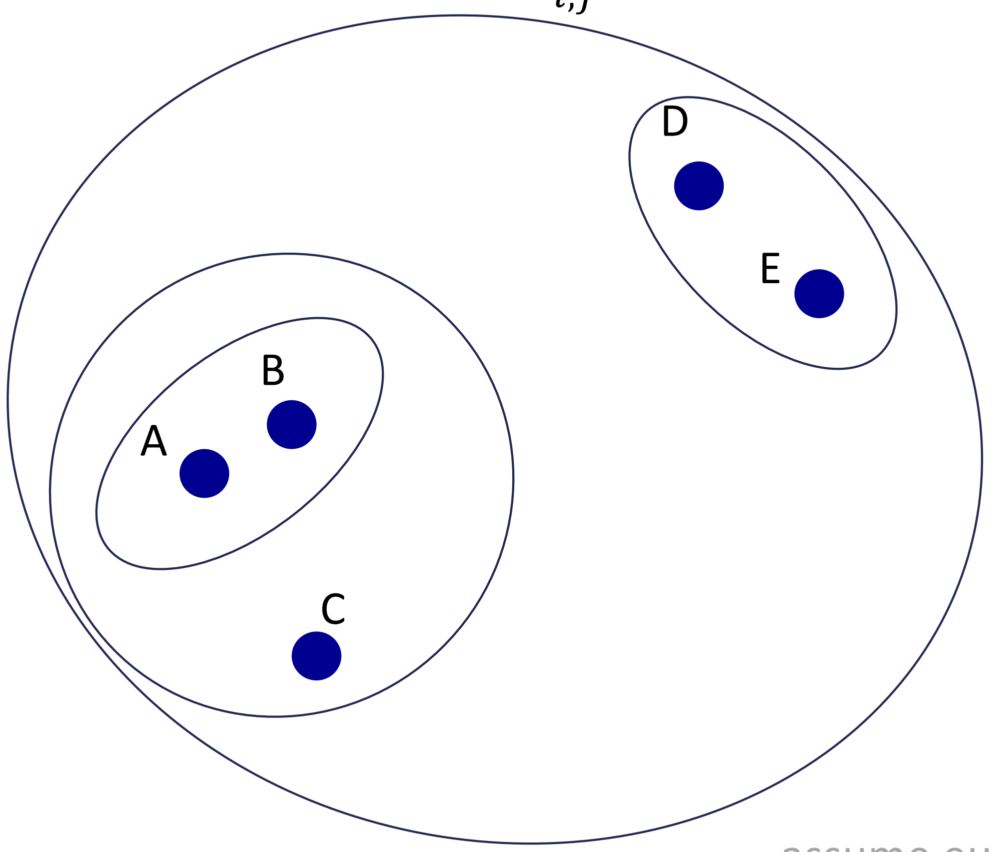
Agglomerative Clustering

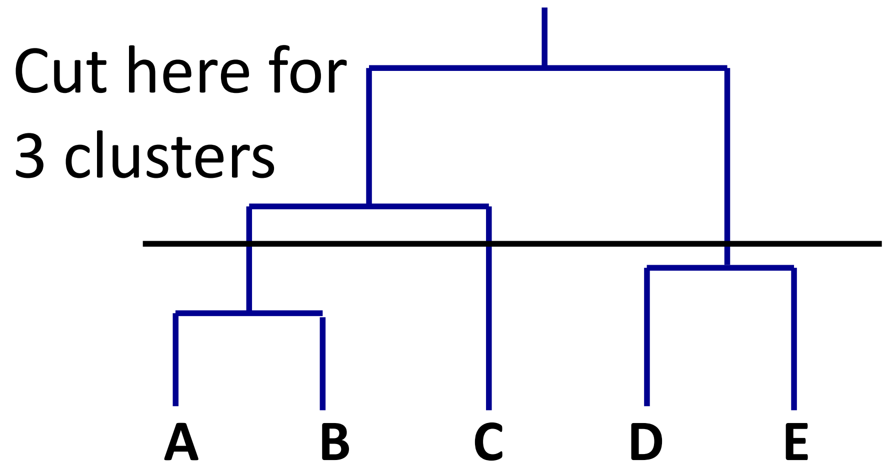
1. Assign each pt. its own cluster

   for each point *i*, $C_i = \{\bar{x}^{(i)}\}$

2. Find the closest clusters & merge, repeat until convergence

   $\arg\min_{i,j} d(C_i, C_j)$ where $i \neq j$





Dendogram

assume euclidean distance