CSE8803/CX4803 Machine Learning in Computational Biology

Lecture 10: Clustering II

Xiuwei Zhang

School of Computational Science and Engineering

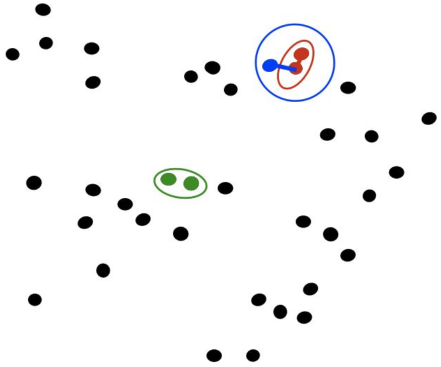
Course logistics

- Final exam: take home and timed
  - 24 hour time window: May 4, 10:00am to May 5, 10:00am
  - 4 hour duration including the time to scan and submit your solutions to Canavs quizzes. You must successfully submit your solutions to Canvas or it is considered as late submission.
- Paper presentation comments
  - Each student in the audience can submit optional comments on paper presentations
  - Some comments will be considered when evaluating the presentation
  - Reasonable and meaningful comments on the presentation are also counted towards your "participation" points. You can also submit insights on the paper (that is not discussed during the presentation) and that can also be considered towards participation points.

Hierarchical clustering

Idea: agglomerative clustering

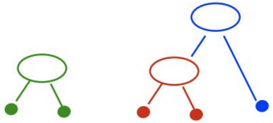
- First merge very similar instances
- Incrementally build larger clusters out of smaller clusters



Algorithm: Maintain a set of clusters

- Initially, each instance in its own cluster
- Repeat:
    - Pick the two closest clusters
    - Merge them into a new cluster
    - Stop when there's only one cluster left

How to define?



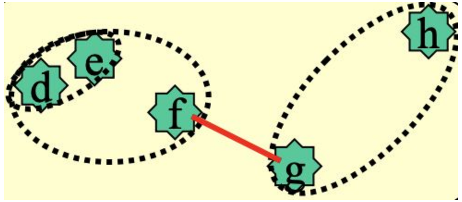
Output:

Produces not one clustering, but a family of clusterings represented by a dendrogram

Hierarchical clustering: define “closest” for clusters

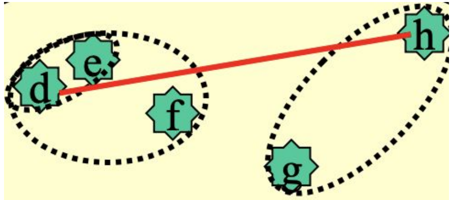
Closest pair
(single-link)

$$CD(X,Y)=\min_{x \in X,\ y \in Y} D(x,y)$$



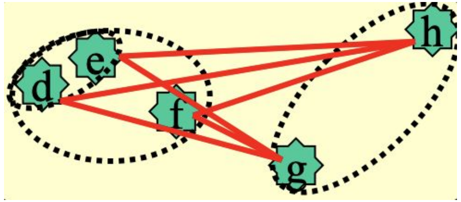
Furthest pair
(complete-link)

$$CD(X,Y)=\max_{x \in X,\ y \in Y} D(x,y)$$



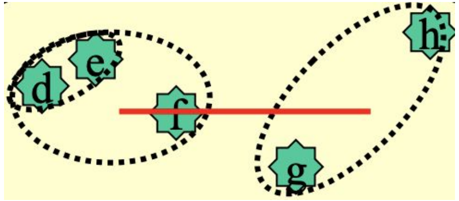
average-link

$$CD(X,Y)=\mathrm{avg}_{x \in X,\ y \in Y} D(x,y)$$



centroids method

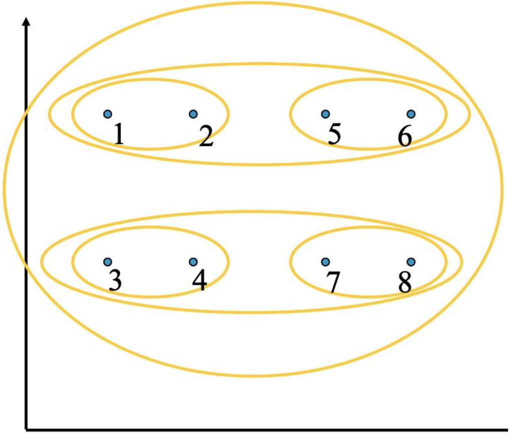
$$CD(X,Y)=D(\ \mathrm{avg}(X)\ ,\ \mathrm{avg}(Y)\ )$$



Hierarchical clustering: define “closest” for clusters

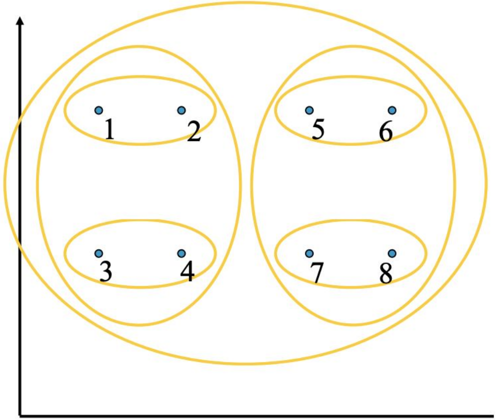
Cluster distance affects both results and runtime

Closest pair
(single-link clustering)



Potentially long and skinny clusters

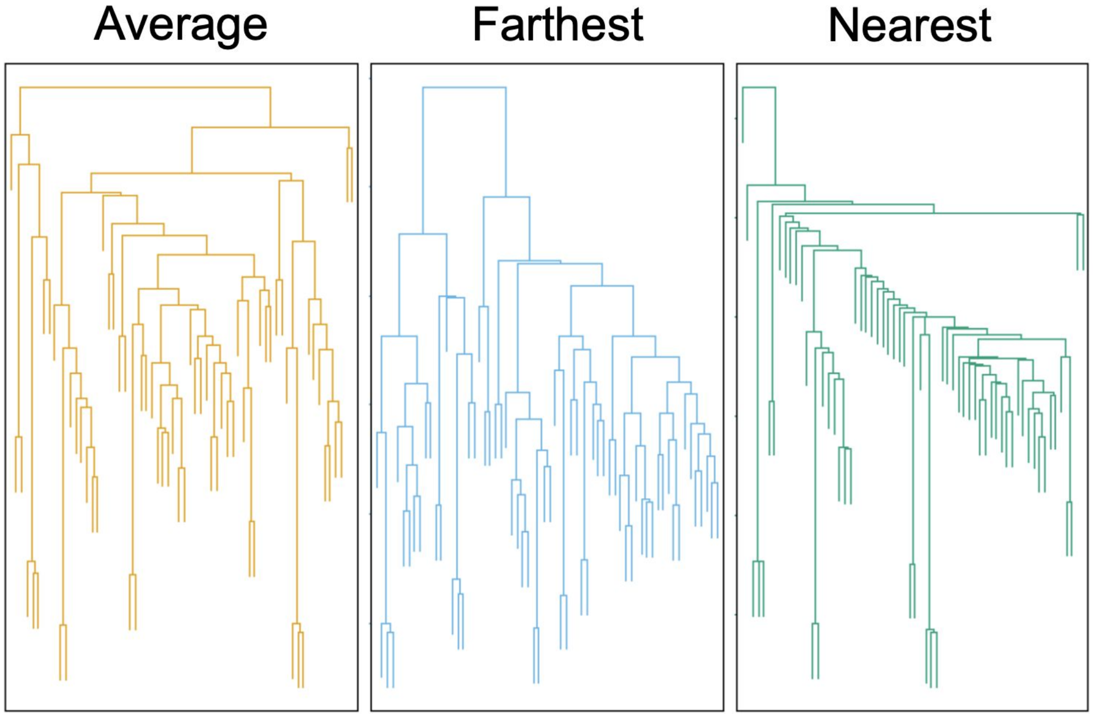
Farthest pair
(complete-link clustering)



tight clusters

Figure credit: Thorsten Joachims

Hierarchical clustering: define “closest” for clusters



Average

Farthest

Nearest

Centroid

fast

Robust to noise
Widely used

Figure credit: David Sontag

Hierarchical clustering: distance measures

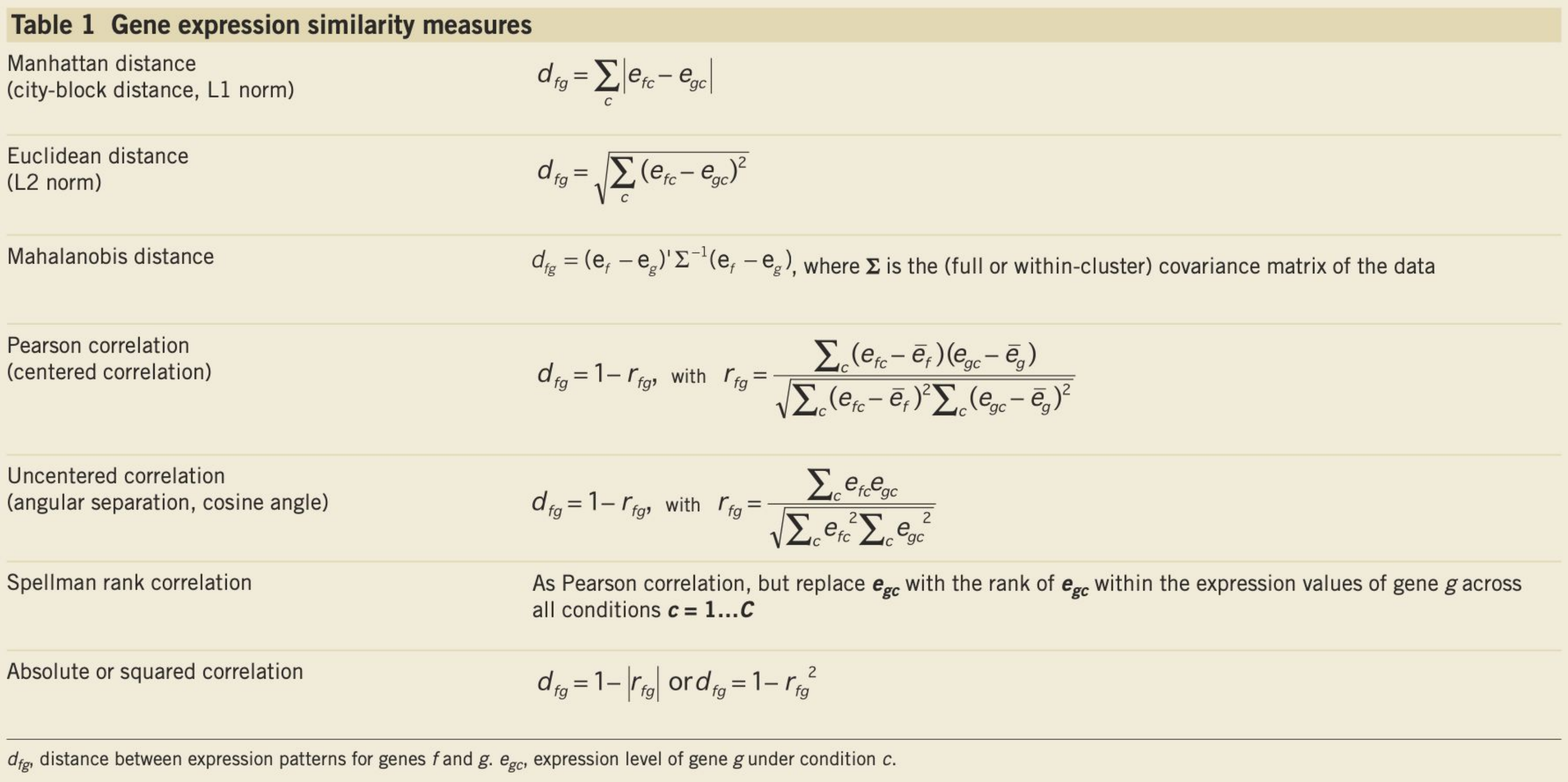


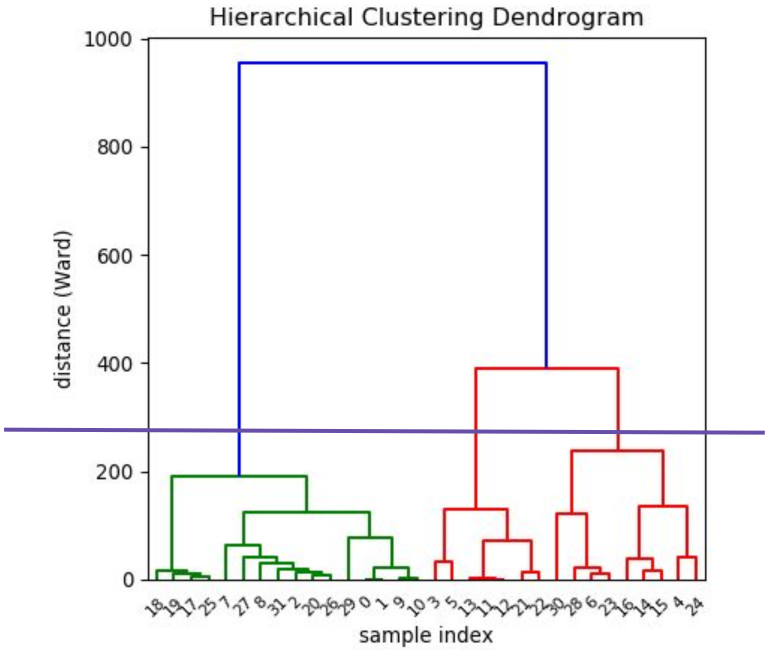
Table 1 Gene expression similarity measures

Manhattan distance (city-block distance, L1 norm)	$d_{fg} = \sum_c \lvert e_{fc} - e_{gc} \rvert$
Euclidean distance (L2 norm)	$d_{fg} = \sqrt{\sum_c (e_{fc} - e_{gc})^2}$
Mahalanobis distance	$d_{fg} = (\mathrm{e}_f - \mathrm{e}_g)'\Sigma^{-1}(\mathrm{e}_f - \mathrm{e}_g)$, where **Σ** is the (full or within-cluster) covariance matrix of the data
Pearson correlation (centered correlation)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c (e_{fc} - \bar{e}_f)(e_{gc} - \bar{e}_g)}{\sqrt{\sum_c (e_{fc} - \bar{e}_f)^2 \sum_c (e_{gc} - \bar{e}_g)^2}}$
Uncentered correlation (angular separation, cosine angle)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c e_{fc} e_{gc}}{\sqrt{\sum_c e_{fc}^2 \sum_c e_{gc}^2}}$
Spellman rank correlation	As Pearson correlation, but replace $\boldsymbol{e_{gc}}$ with the rank of $\boldsymbol{e_{gc}}$ within the expression values of gene *g* across all conditions $\boldsymbol{c = 1 \ldots C}$
Absolute or squared correlation	$d_{fg} = 1 - \lvert r_{fg} \rvert$ or $d_{fg} = 1 - r_{fg}^2$

d_{fg}, distance between expression patterns for genes *f* and *g*. e_{gc}, expression level of gene *g* under condition *c*.

https://www.nature.com/articles/nbt1205-1499

Hierarchical clustering: obtaining the clusters



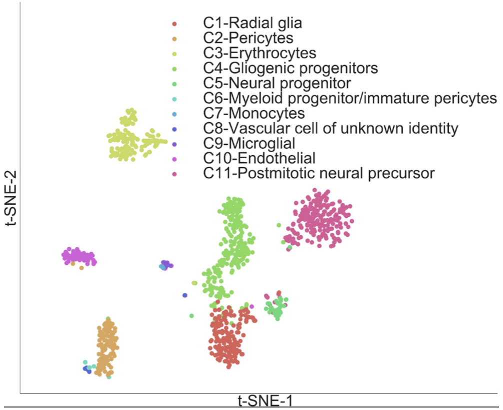
Hierarchical clustering: Summary

- No need to specify the number of clusters in advance.
- Hierarchical structure maps nicely onto human intuition for some domains
- They do not scale well: where n is the number of total objects.
- Like any heuristic search algorithms, local optima are a problem.
- Interpretation of results can be subjective

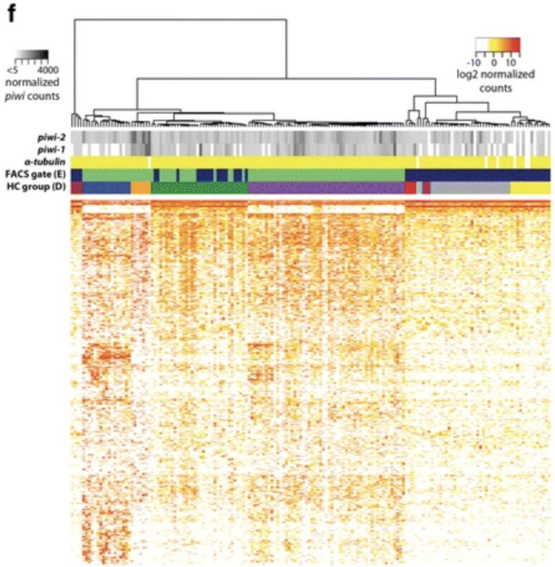
Clustering algorithms

Clustering - unsupervised learning methods

- K-means (and fuzzy k-means)
- Hierarchical clustering
- Spectral clustering
- Louvain, Leiden clustering



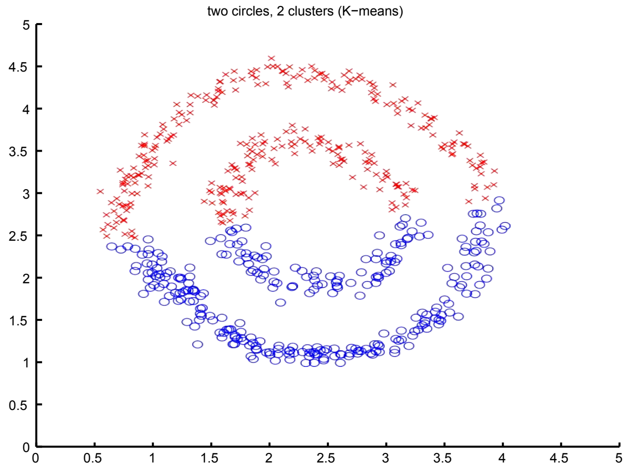
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007040



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0937-9/figures/1

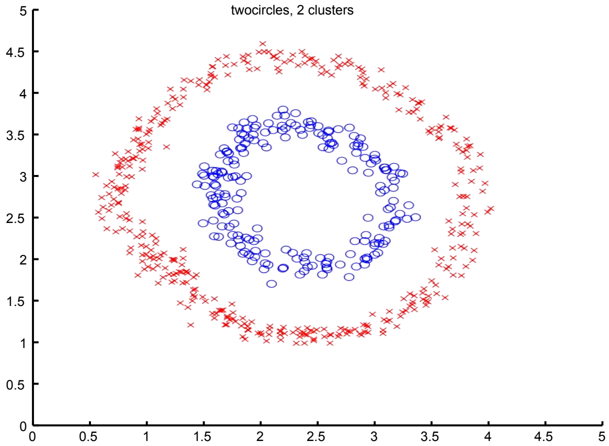
Spectral clustering

K-means



(i)

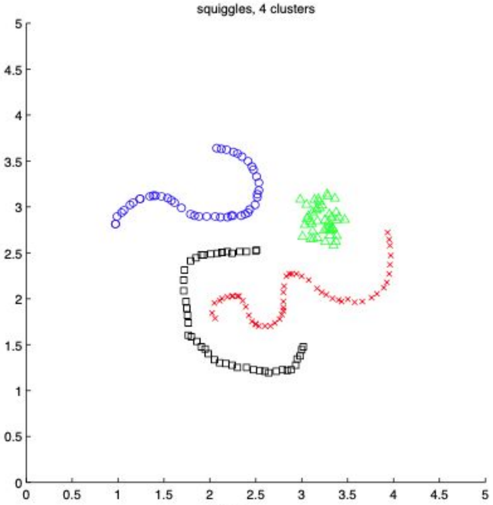
Spectral clustering



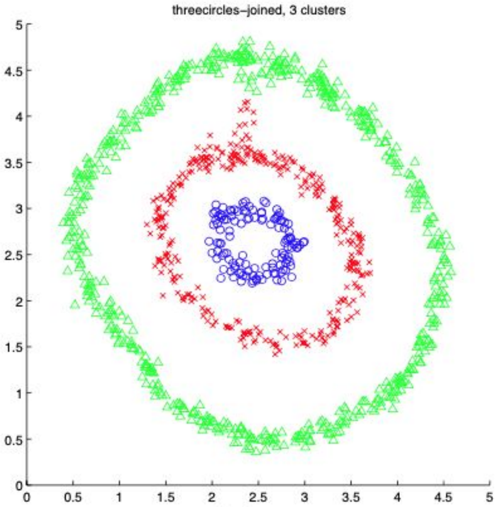
(e)

[Shi & Malik ‘00; Ng, Jordan, Weiss NIPS ‘01]

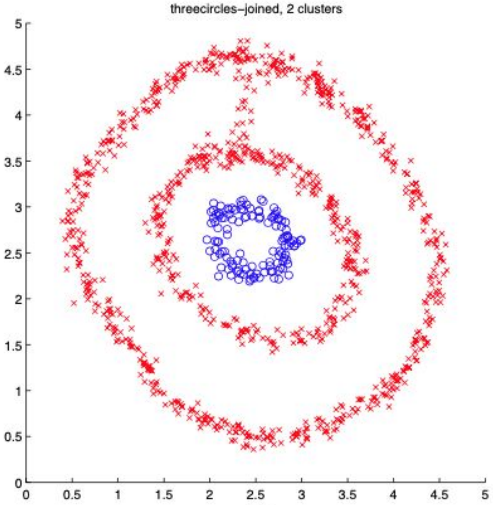
Spectral clustering


squiggles, 4 clusters

(d)


threecircles–joined, 3 clusters

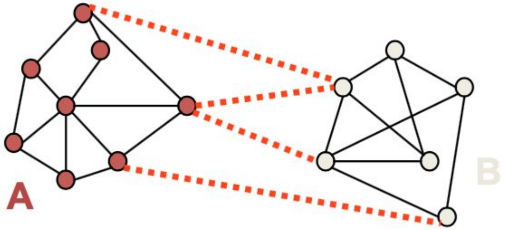
(g)


threecircles–joined, 2 clusters

(f)

Spectral clustering

Idea: group points based on links in a graph

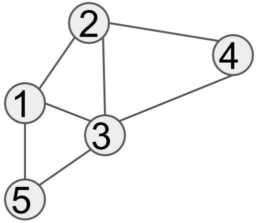


To construct the graph:

- Fully connected weighted graph
- K nearest neighbor (kNN) graph, where for each vertex, it's connected to its k nearest neighbors.

Question: degree of nodes in a kNN graph?

k=2

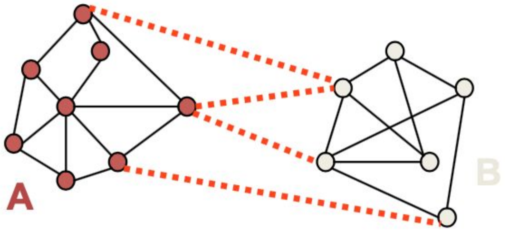


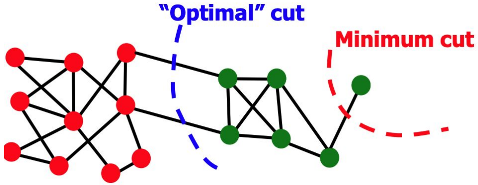
Spectral clustering

The *min-cut* problem

Partition graph into two sets A and B such that weight of edges connecting vertices in A to vertices in B is minimum.

For an unweighted graph, find a *cut* with the smallest number of edges.





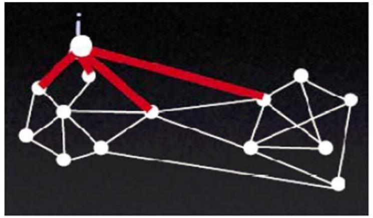
Problem with the min-cut criterion for clustering:

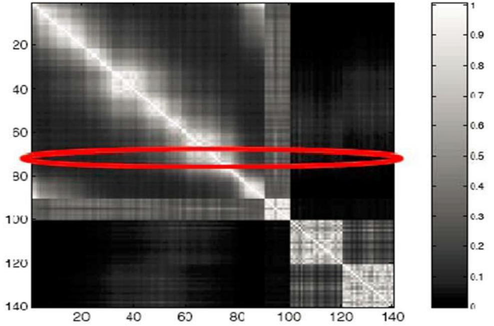
- Only considers external cluster connections
- Does not consider internal cluster connectivity

Spectral clustering: graph terminologies

- Degree of nodes

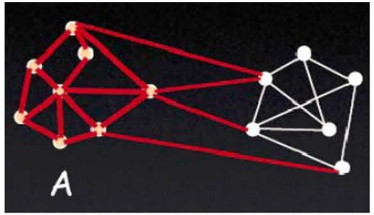
$$d_i = \sum_j w_{i,j}$$

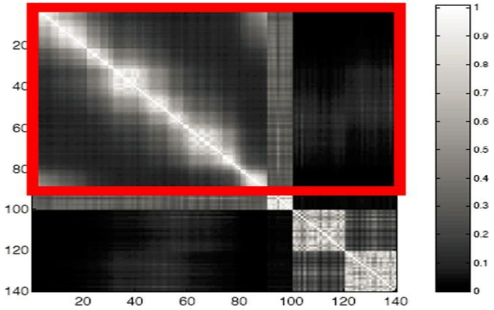




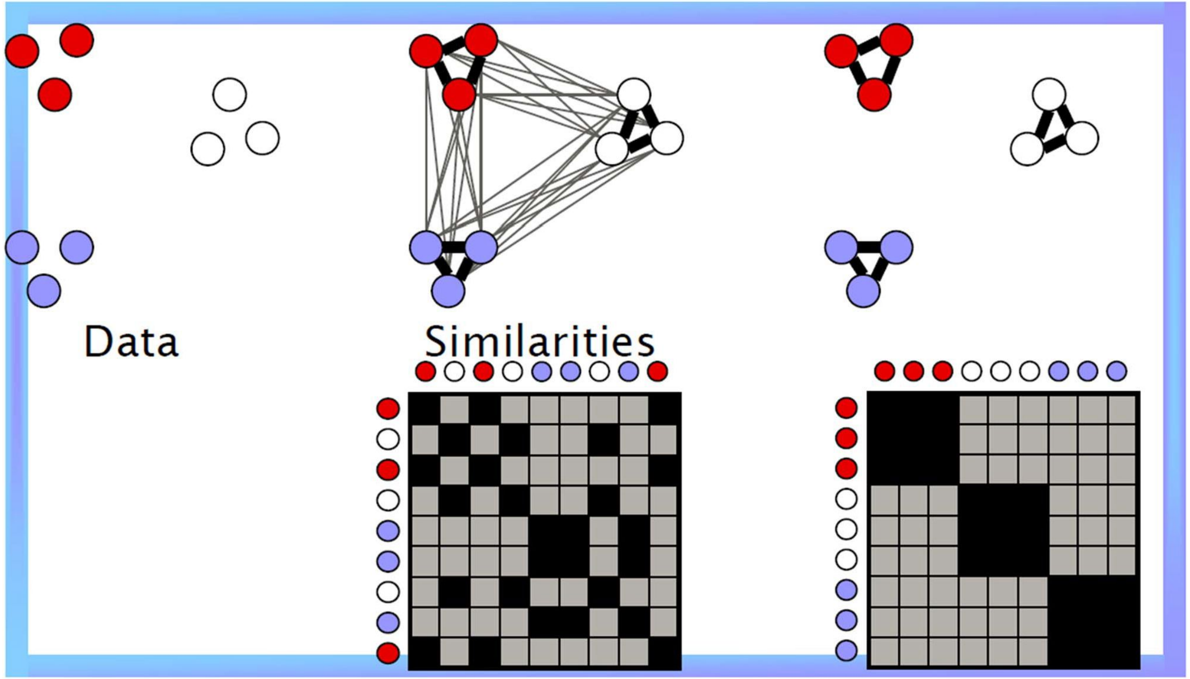
- Volume of a set

$$vol(A) = \sum_{i \in A} d_i, A \subseteq V$$





Spectral Clustering: graph partitioning



Spectral clustering: objective

Objective of min-cut: minimize $\text{cut}(A,\ B) = \sum_{i \in A, j \in B} w_{i,j}$

Objective of spectral clustering: minimize normalized cut

$$\text{Ncut}(A,\ B) = \text{cut}(A, B) \left(\frac{1}{\text{vol}(A)} + \frac{1}{\text{vol}(B)} \right)$$

Which represents connectivity between groups relative to the density of each group

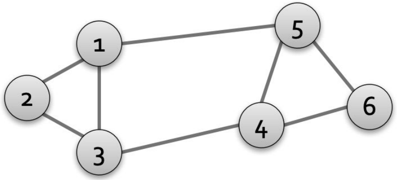
Advantage: produces more balanced clusters

Spectral clustering: objective

Objective of spectral clustering: minimize normalized cut

$$\mathrm{Ncut}(A,\ B) = \mathrm{cut}(A, B)\left(\frac{1}{\mathrm{vol}(A)} + \frac{1}{\mathrm{vol}(B)}\right)$$

Define D, which is a diagonal matrix with the degree of every node: $D(i,i) = \sum_j W(i,j)$



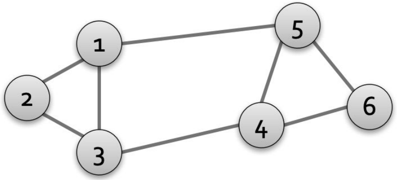
	1	2	3	4	5	6
1	3	0	0	0	0	0
2	0	2	0	0	0	0
3	0	0	3	0	0	0
4	0	0	0	3	0	0
5	0	0	0	0	3	0
6	0	0	0	0	0	2

Spectral clustering: objective

Objective of spectral clustering: minimize normalized cut

$$\mathrm{Ncut}(A,\ B) = \mathrm{cut}(A, B)\left(\frac{1}{\mathrm{vol}(A)} + \frac{1}{\mathrm{vol}(B)}\right)$$

Define L = D - W (graph Laplacian)



	1	2	3	4	5	6
1	3	0	0	0	0	0
2	0	2	0	0	0	0
3	0	0	3	0	0	0
4	0	0	0	3	0	0
5	0	0	0	0	3	0
6	0	0	0	0	0	2

	1	2	3	4	5	6
1	3	-1	-1	0	-1	0
2	-1	2	-1	0	0	0
3	-1	-1	3	-1	0	0
4	0	0	-1	3	-1	-1
5	-1	0	0	-1	3	-1
6	0	0	0	-1	-1	2

The laplacian matrix L is positive semi-definite

- All eigenvalues are ≥ 0
- $x^T L x = \sum_{ij} L_{ij} x_i x_j \geq 0$, for every *x*
- $L = N^T \cdot N$

Spectral clustering: steps

Objective of spectral clustering: minimize normalized cut

$$\mathrm{Ncut}(A,\ B) = \mathrm{cut}(A, B)\left(\frac{1}{\mathrm{vol}(A)} + \frac{1}{\mathrm{vol}(B)}\right)$$

Steps:

1. Calculate D, $D(i,i) = \sum_j W(i,j)$
2. Calculate L, L = D - W, and L', L'= $D^{-1/2}LD^{-1/2}$
3. Calculate eigenvalues and eigenvectors of L'. Sort the eigenvalues from low to high values.
4. Take the corresponding first k eigenvectors
5. Perform k-means on these eigenvectors and obtain clusters

Spectral clustering: solving the objective

Objective of spectral clustering: minimize normalized cut

$$\mathrm{Ncut}(A,\ B) = \mathrm{cut}(A, B)\left(\frac{1}{\mathrm{vol}(A)} + \frac{1}{\mathrm{vol}(B)}\right)$$

define $f = \left[f_1, f_2, \ldots, f_n\right]^T$ where $f_i = \begin{cases} \dfrac{1}{\mathrm{vol}(A)} & \text{if } i \in A \\ -\dfrac{1}{\mathrm{vol}(B)} & \text{if } i \in B \end{cases}$

f is determined by the cluster assignment of nodes!

$$\mathbf{f}^T\mathbf{D}\mathbf{f} = \sum_i d_i \mathbf{f}_i^2 = \sum_{i\in A} \frac{d_i}{\mathrm{vol}(A)^2} + \sum_{j\in B} \frac{d_j}{\mathrm{vol}(B)^2} = \frac{1}{\mathrm{vol}(A)} + \frac{1}{\mathrm{vol}(B)}$$

D is a diagonal matrix

Divide all nodes into set A and set B

$\sum_{i\in A} d_i = \mathrm{vol}(A)$

Spectral clustering: solving the objective

Objective of spectral clustering: minimize normalized cut

$$\mathrm{Ncut}(A,\ B) = \mathrm{cut}(A, B)\left(\frac{1}{\mathrm{vol}(A)} + \frac{1}{\mathrm{vol}(B)}\right)$$

define $f = \left[f_1, f_2, \ldots, f_n\right]^T$ where $f_i = \begin{cases} \dfrac{1}{\mathrm{vol}(A)} & \text{if } i \in A \\ -\dfrac{1}{\mathrm{vol}(B)} & \text{if } i \in B \end{cases}$

$$\mathbf{f}^T\mathbf{L}\mathbf{f} = \sum_{ij} w_{ij}(\mathbf{f}_i - \mathbf{f}_j)^2 = \sum_{i\in A, j\in B} w_{ij}\left(\frac{1}{\mathrm{vol}(A)} + \frac{1}{\mathrm{vol}(B)}\right)^2$$

L=D-W
w_{ii}=0

If i,j in the same cluster, then f_i-f_j=0

Spectral clustering: solving the objective

Objective of spectral clustering: minimize normalized cut

$$\begin{aligned}\mathrm{Ncut}(A,\ B) &= \mathrm{cut}(A,B)\left(\frac{1}{\mathrm{vol}(A)}+\frac{1}{\mathrm{vol}(B)}\right)\\ &= \sum_{i\in A, j\in B} w_{ij}\left(\frac{1}{\mathrm{vol}(A)}+\frac{1}{\mathrm{vol}(B)}\right)\end{aligned}$$

$$\mathrm{Ncut}(A,\ B) = \frac{\mathbf{f}^T\mathbf{L}\mathbf{f}}{\mathbf{f}^T\mathbf{D}\mathbf{f}}$$

$$\mathbf{f}^T\mathbf{L}\mathbf{f} = \sum_{ij} w_{ij}(\mathbf{f}_i - \mathbf{f}_j)^2 = \boxed{\sum_{i\in A, j\in B} w_{ij}\left(\frac{1}{\mathrm{vol}(A)}+\frac{1}{\mathrm{vol}(B)}\right)^2}$$

$$\mathbf{f}^T\mathbf{D}\mathbf{f} = \sum_{j} d_i\mathbf{f}_i^2 = \sum_{i\in A}\frac{d_i}{\mathrm{vol}(A)^2} + \sum_{j\in B}\frac{d_i}{\mathrm{vol}(B)^2} = \boxed{\frac{1}{\mathrm{vol}(A)}+\frac{1}{\mathrm{vol}(B)}}$$

Spectral clustering: solving the objective

Objective of spectral clustering: minimize normalized cut

$$\text{min } \mathrm{Ncut}(A,\ B) = \min_{\mathbf{f}} \frac{\mathbf{f}^T\mathbf{L}\mathbf{f}}{\mathbf{f}^T\mathbf{D}\mathbf{f}}$$

$f^T D1=0$, that is, $\Sigma_i d_i\, f_i = 0$

$$f_i = \begin{cases} \dfrac{1}{\mathrm{vol}(A)} & \text{if } i \in A \\ -\dfrac{1}{\mathrm{vol}(B)} & \text{if } i \in B \end{cases}$$

$$\Longrightarrow \quad \min_f f^T L f \quad \text{s.t. } f^T D f = 1$$

With relaxation on **f** from discrete to continuous space

Once this is solved, we can obtain cluster labels by the sign of f_i ($f_i > 0$ or $f_i < 0$)

Spectral clustering: solving the objective

Objective of spectral clustering: minimize normalized cut

$$\min \operatorname{Ncut}(A, B) = \min_{\mathbf{f}} \frac{\mathbf{f}^T \mathbf{L} \mathbf{f}}{\mathbf{f}^T \mathbf{D} \mathbf{f}}$$

$f^T D 1 = 0$, that is, $\Sigma_i d_i f_i = 0$

$$f_i = \begin{cases} \frac{1}{\operatorname{vol}(A)} & \text{if } i \in A \\ -\frac{1}{\operatorname{vol}(B)} & \text{if } i \in B \end{cases}$$

⇒ $$\min_f f^T L f \quad \text{s.t. } f^T D f = 1$$

Now let $u = D^{1/2} f$, then $u^T = f^T D^{1/2}$, thus the constraint becomes:

$$f^T D^{\frac{1}{2}} D^{\frac{1}{2}} f = 1$$

$$u^T u = 1$$

The objective thus becomes

$$f^T L f = f^T D^{1/2} D^{-1/2} L D^{-1/2} D^{1/2} f$$

$$= u^T D^{-1/2} L D^{-1/2} u$$

Let $L' = D^{-1/2} L D^{-1/2}$, the optimization problem becomes:

$$\min_u u^T L' u$$

$$\text{s.t. } u^T u = 1$$

Spectral clustering: solving the objective

Objective of spectral clustering: minimize normalized cut

$$\boxed{\begin{array}{l} \min_u \; u^T L' \, u \\ \text{s.t. } u^T u = 1 \end{array}}$$

Looks familiar?

Objective function:

$$\max_{w:\|w\|\le 1} \mathbf{w}^T \hat{\mathbf{R}} \mathbf{w}$$

PCA proof

The eigen decomposition of L' will give us the solution.
As we are minimizing, we will sort eigenvalues from small to large.
However, can we use the smallest eigenvalue?

Observation. L has an eigenvalue (smallest) which is 0, corresponding to an eigenvector with all 1s.

$$(D - W)\,\vec{1} = D\vec{1} - W\vec{1} = \vec{0}$$

$$L\vec{1} = 0 \cdot \vec{1}$$

What about L'?

$$\begin{aligned} & L' D^{1/2}\vec{1} \\ &= D^{-1/2} L D^{-1/2} D^{1/2}\vec{1} \\ &= D^{-1/2}\,\vec{0} \\ &= 0 \cdot \left(D^{1/2}\vec{1}\right) \end{aligned}$$

L' also has a 0 eigenvalue and the corresponding eigenvector is $D^{1/2}\vec{1}$

Spectral clustering: solving the objective

Objective of spectral clustering: minimize normalized cut

$$\boxed{\begin{aligned} &\min_u \; u^T L' \, u \\ &\text{s.t. } u^T u = 1 \end{aligned}}$$

Looks familiar?

Objective function:

$$\max_{w:\|w\|\leq 1} \mathbf{w}^T \hat{\mathbf{R}} \mathbf{w}$$

PCA proof

The eigen decomposition of L' will give us the solution.

As we are minimizing, we will sort eigenvalues from small to large.

However, can we use the smallest eigenvalue?

What about L'?

$$\begin{aligned} & L' D^{1/2} \vec{1} \\ = \; & D^{-1/2} L D^{-1/2} D^{1/2} \vec{1} \\ = \; & D^{-1/2} \vec{0} \\ = \; & 0 \cdot \left(D^{1/2} \vec{1} \right) \end{aligned}$$

L' also has a 0 eigenvalue and the corresponding eigenvector is $D^{1/2}\vec{1}$

If we take this eigenvalue, u= $D^{1/2}\vec{1}$

What is f?

u=$D^{1/2}$f

f=$D^{-1/2}$u=$D^{-1/2}D^{1/2}$ **1** = **1**

This f is not informative for us to find cluster assignment for the nodes!

Spectral clustering: solving the objective

Objective of spectral clustering: minimize normalized cut

$$\boxed{\begin{aligned}&\min_u \; u^T L' \, u \\ &\text{s.t. } u^T u = 1\end{aligned}}$$

Looks familiar?

Objective function:

$$\max_{w:\|w\|\leq 1} \mathbf{w}^T \hat{\mathbf{R}} \mathbf{w}$$

PCA proof

The eigen decomposition of L' will give us the solution.
As we are minimizing, we will sort eigenvalues from small to large.
However, can we use the smallest eigenvalue?

If we take this eigenvalue, u= $D^{1/2}\vec{1}$
What is f?

u= $D^{1/2}f$
f= $D^{-1/2}u = D^{-1/2}D^{1/2}\,\mathbf{1} = \mathbf{1}$

This f is not informative for us to find cluster assignment for the nodes!

We will use the **2nd smallest** eigenvalue of L' and its corresponding eigenvector to obtain u and f, and then the cluster assignment.

Spectral clustering: partition a graph into 2 clusters

Objective of spectral clustering: minimize normalized cut

$$\begin{aligned}\mathrm{Ncut}(A, B) &= \mathrm{cut}(A, B)\left(\frac{1}{\mathrm{vol}(A)} + \frac{1}{\mathrm{vol}(B)}\right) \\ &= \sum_{i\in A, j\in B} w_{ij}\left(\frac{1}{\mathrm{vol}(A)} + \frac{1}{\mathrm{vol}(B)}\right)\end{aligned}$$

$$\min_f f^T L f \quad \mathrm{s.t.}\ f^T D f = 1$$

$$\begin{aligned}&\min_u u^T L' u \\ &\mathrm{s.t.}\ u^T u = 1\end{aligned}$$

Solution:

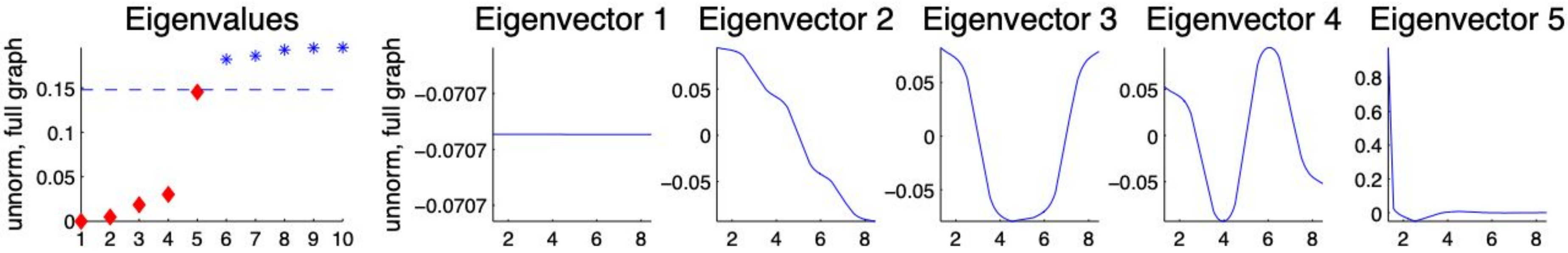
1. Calculate the eigenvalues and eigenvectors of L’
2. u = the 2nd smallest eigenvalue
3. f = $D^{-1/2}$u
4. Apply a threshold on f to determine the cluster assignment of nodes.

Other versions of spectral clustering:
Perform eigen-decomposition on L or other versions of L.

From 2-way partition to k-way partition

Cluster multiple eigenvectors

- Build a reduced space from multiple eigenvectors.
- (its like doing dimension reduction then k-means)



- 1st Eigenvector is the all ones vector ***1*** (if graph is connected)
- 2nd Eigenvector thresholded at 0 separates first two clusters from last two
- k-means clustering of the 4 eigenvectors identifies all clusters

Figure from Ulrike von Luxburg 2007

Spectral clustering: steps

Objective of spectral clustering: minimize normalized cut

$$\mathrm{Ncut}(A,\ B) = \mathrm{cut}(A, B)\left(\frac{1}{\mathrm{vol}(A)} + \frac{1}{\mathrm{vol}(B)}\right)$$

Steps:

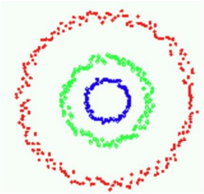
1. Calculate D, $D(i,i) = \sum_j W(i,j)$
2. Calculate L, L = D - W, and L’, L’= $D^{-1/2}LD^{-1/2}$
3. Calculate eigenvalues and eigenvectors of L’. Sort the eigenvalues from low to high values.
4. Take the corresponding first k eigenvectors and obtain $U = \begin{bmatrix} u_1 & u_2 & \ldots & u_k \end{bmatrix}_{\mathrm{nxk}}$
5. Perform k-means on these eigenvectors and obtain clusters

There are different normalized forms of L

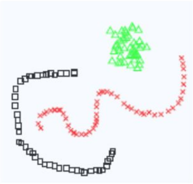
Normalize each row of U

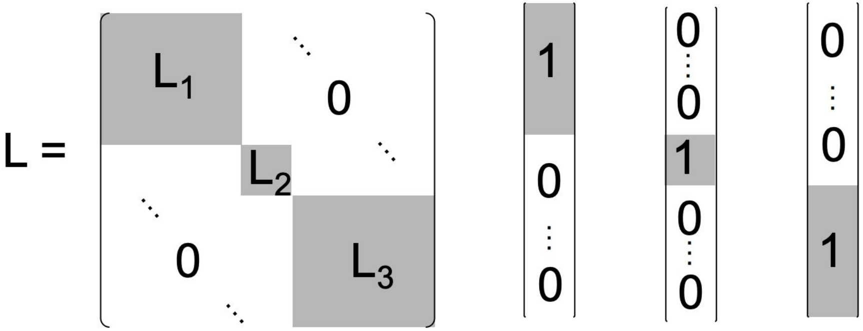
Spectral clustering: more fun facts

- If graph is connected, first Laplacian evec is constant (all 1s)
- If graph is disconnected (k connected components), Laplacian is block diagonal and first k Laplacian evecs are:



OR





First three eigenvectors

First 3 eigenvalues are 0s!

Slide credit: Aarti Singh

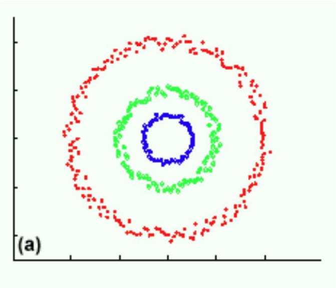
Spectral clustering: why it works

Steps:

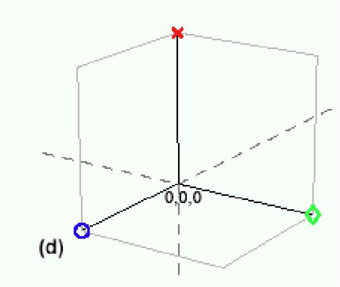
1. Calculate D, $D(i,i) = \sum_j W(i,j)$
2. Calculate L, L = D - W
3. Calculate eigenvalues and eigenvectors of L. Sort the eigenvalues from low to high values.
4. Take the corresponding first k eigenvectors
5. Perform k-means on these eigenvectors and obtain clusters

Non-linear embedding

Original data

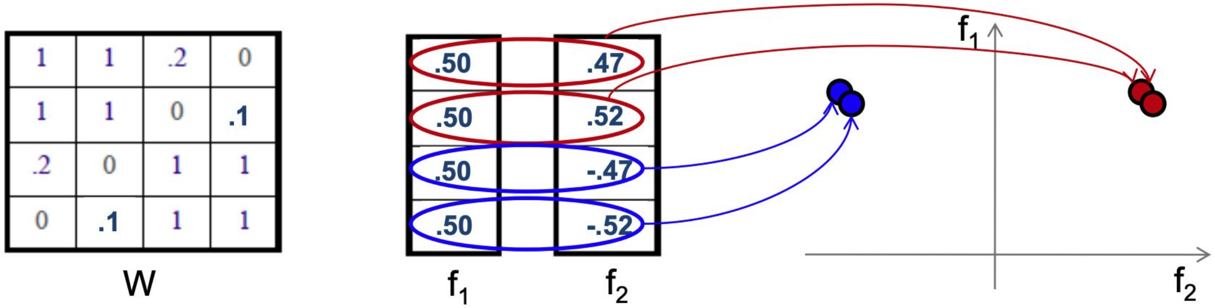


Projected data



Spectral clustering: why it works

Can put data points into blocks using eigenvectors:

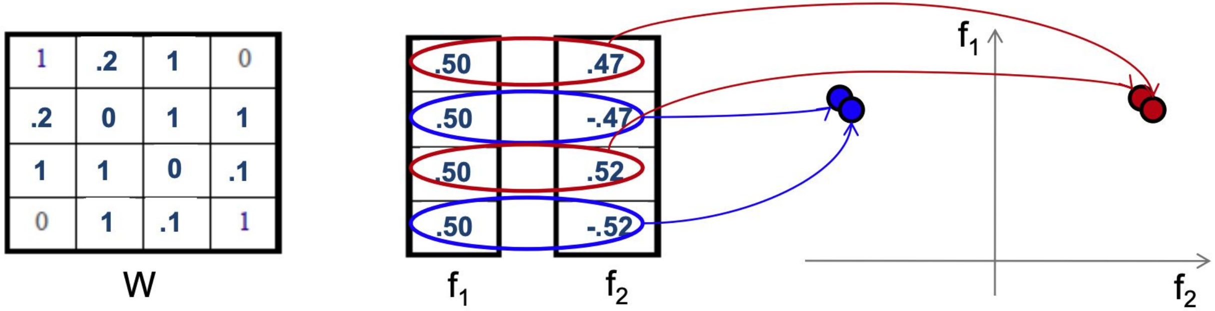


1	1	.2	0
1	1	0	.1
.2	0	1	1
0	.1	1	1

f_1	f_2
.50	.47
.50	.52
.50	-.47
.50	-.52

Non-linear embedding

Embedding is same regardless of data ordering:



1	.2	1	0
.2	0	1	1
1	1	0	.1
0	1	.1	1

f_1	f_2
.50	.47
.50	-.47
.50	.52
.50	-.52

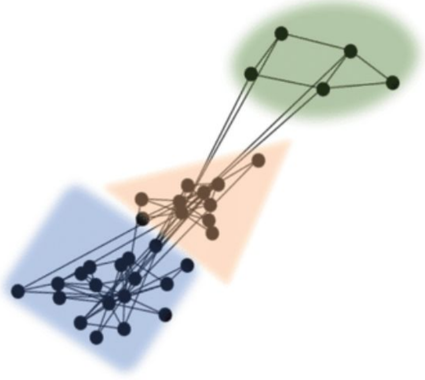
Slide credit: Aarti Singh

Graph based clustering methods

Spectral clustering

Louvain/Leiden clustering methods:

Optimize *Modularity*



References

Jianbo Shi and J. Malik, "Normalized cuts and image segmentation," Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, PR, USA, 1997, pp. 731-737, doi: 10.1109/CVPR.1997.609407.

Barton, T., Bruna, T. & Kordik, P. Chameleon 2: An Improved Graph-Based Clustering Algorithm. *ACM Trans. Knowl. Discov. Data* **13**, 1–27 (2019)