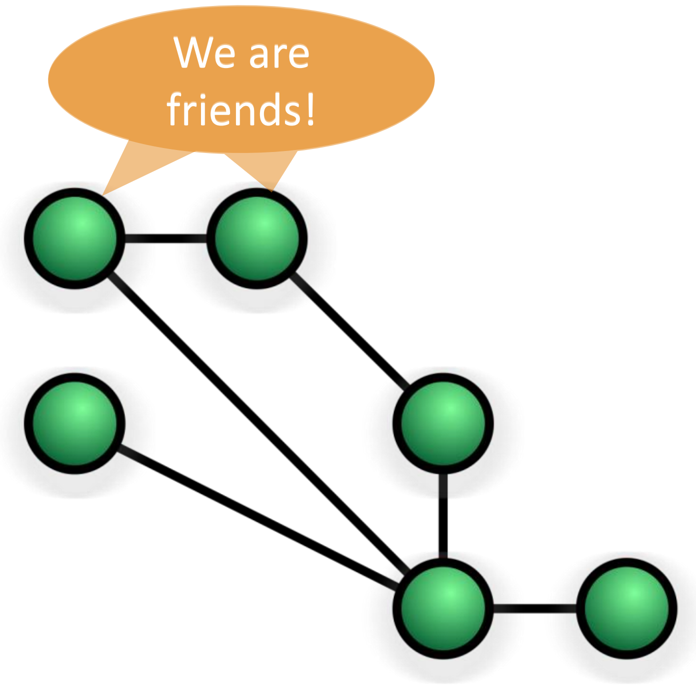
Network analysis

TOPICS OF DATA ENGINEERING

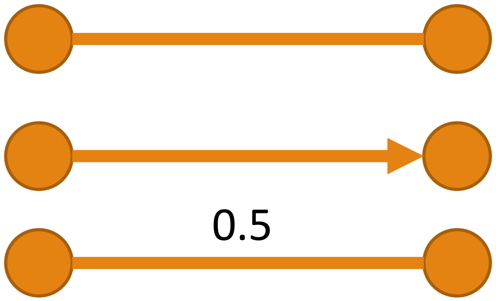
Networks appear everywhere

- Examples
  - WWW
  - Social networks
    - Friend networks
    - SNS
  - Chemical interactions in a human body



Nodes and edges

- Node (vertex)
  - A fundamental unit of a graph (network).
  - Represented as a dot
- Edge
  - Connects two nodes
- Each edge connects two nodes
  - Undirected edges
  - Directed edges
  - Weighted edges



Adjacency matrix

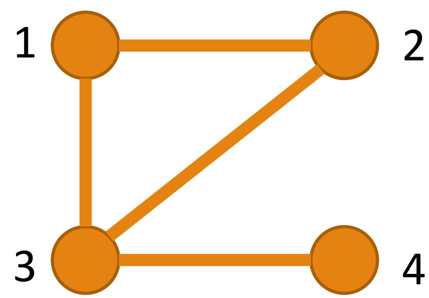
- is a representative data structure
  to describe a network structure.
- The elements are one or zero depending on connectivity.

$$A_{ij} = \begin{cases} 1 \\ 0 \end{cases}$$

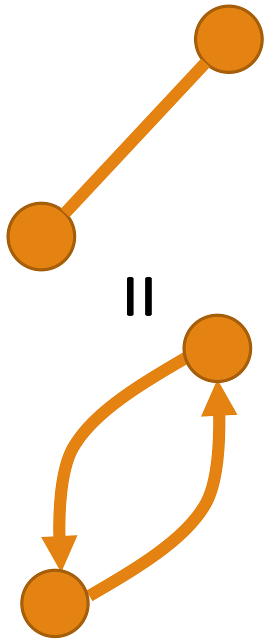
Examples of adjacency matrices

Undirected graph

$$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

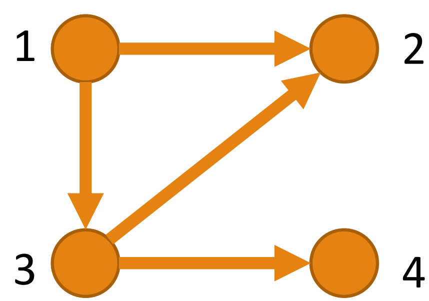


They are equivalent under



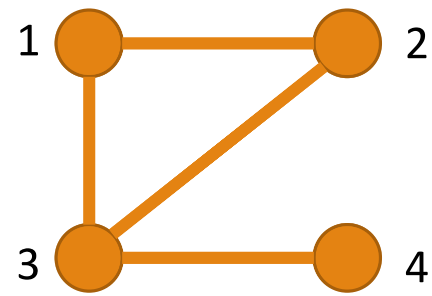
Directed graph

$$\begin{array}{cc} & \text{to} \\ \text{From} & \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

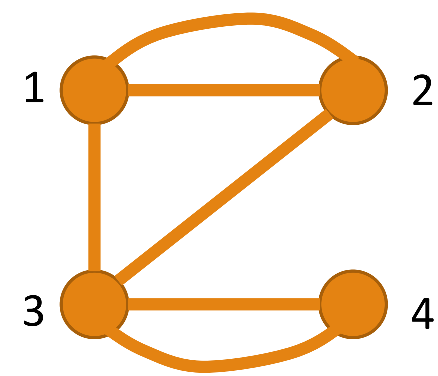


Types of graph structures

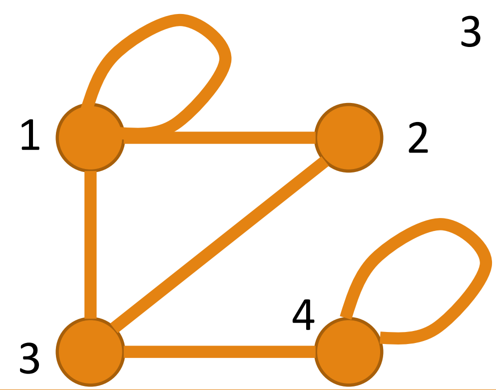
- Simple graph
  - No multiple edges
  - No self loop



- Multi graph
  - Multiple edges exist
  - No self loop



- Graph with loop
  - Self loops exist



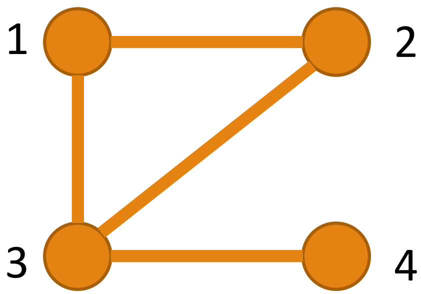
Hereafter, we focus only on a simple graph.

Degree

- The number of edges from a node

$k_1 = 2 \qquad k_2 = 2$

$k_3 = 3 \qquad k_4 = 1$



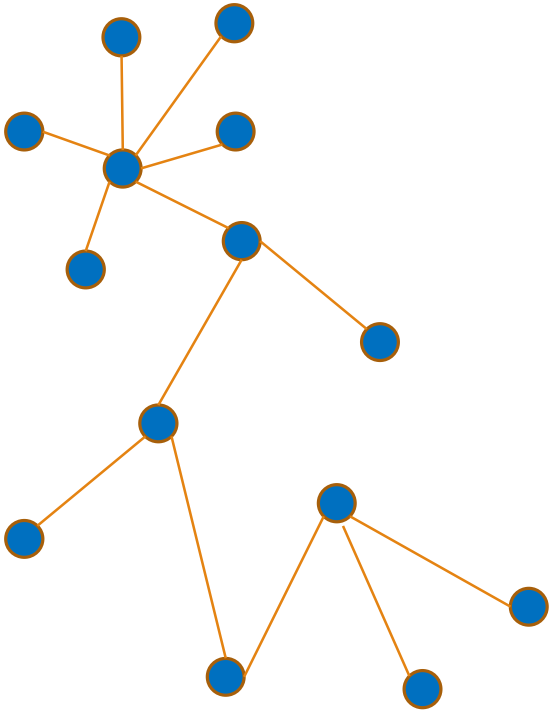
- The sum of degrees are equal to twice the number of nodes M

$$\sum_i k_i = 2 + 2 + 3 + 1 = 8 = 2M$$

Centrality

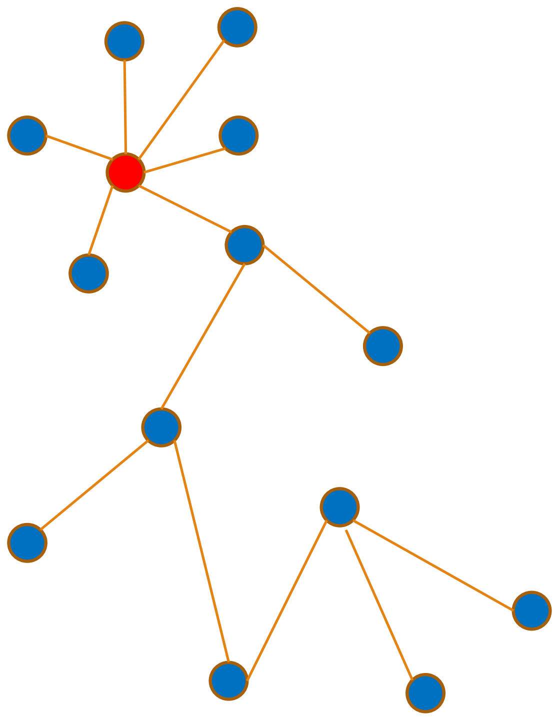
Centrality

- Node characteristics to identify importance in a graph
- Many types
  - Degree centrality
  - Betweenness centrality
  - Eigenvector centrality
  - HITS
  - Page rank centrality



Degree centrality

- Based on the idea “the person who has the most friends is important”.
- Definition
  - $C_d(i) = k_i$

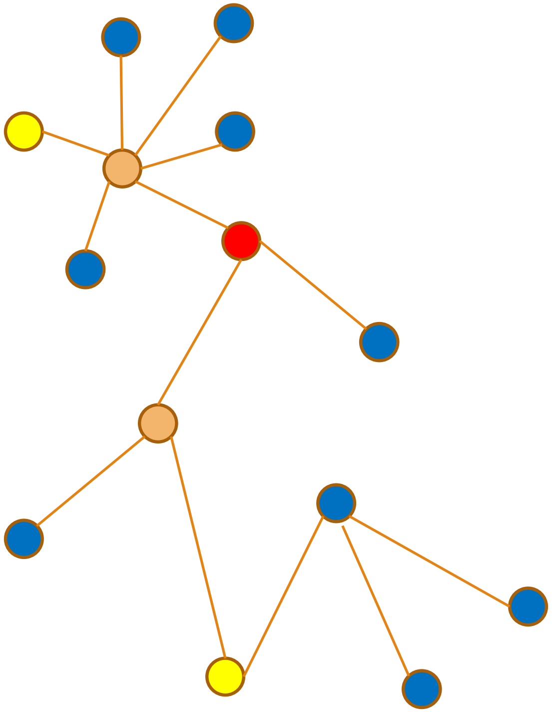


Betweenness centrality

- Based on the idea "the person who can mediate between the most pair of friends is important".
- Definition
  - $C_b(i) = \frac{\sum_{i,k} sp_{j,k}(i)}{n(n-1)/2}$

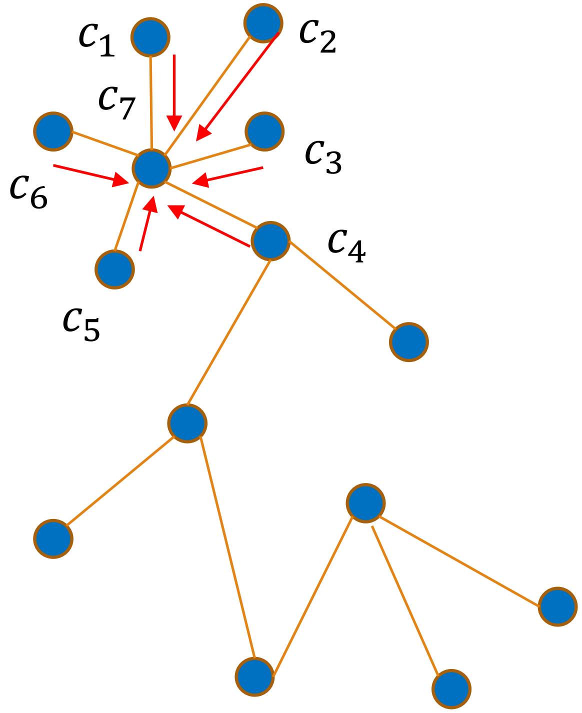
where $sp_{j,k}(i) = \begin{cases} 1 & \text{if Node } i \text{ is on the shortest path between Node } j \text{ and Node } k. \\ 0 & \text{Otherwise} \end{cases}$

$n = \#\ of\ nodes$



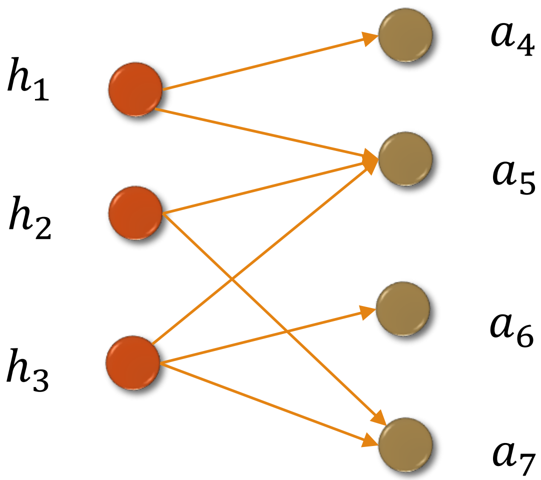
Eigenvector centrality

- Based on the idea "if important persons think a friend is important, he/she is important".
- Definition
  - $c_i^{(n+1)} = \frac{1}{\lambda} \sum_j A_{ij} c_j^{(n)}$
  - $\boldsymbol{c}^{(n+1)} = \frac{1}{\lambda} A \boldsymbol{c}^{(n)}$
  - If $\lim_{n \to \infty} \boldsymbol{c}^{(n)} = \boldsymbol{c}$, $A\boldsymbol{c} = \lambda \boldsymbol{c}$



HITS

- An extension of eigenvector centrality
- It assumes hubs and authorities



$$B = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

$$\boldsymbol{h}^{(n+1)} = \frac{1}{\lambda_1} B \boldsymbol{a}^{(n)}$$

$$\boldsymbol{a}^{(n+1)} = \frac{1}{\lambda_2} B^T \boldsymbol{h}^{(n)}$$

PageRank

- The Google algorithm of web page importance algorithm

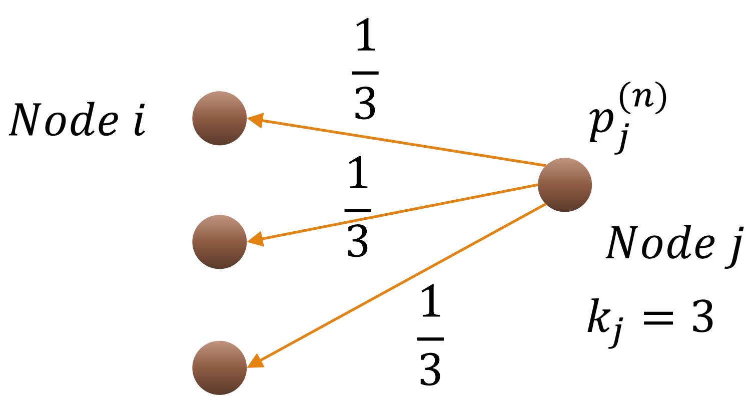
$n = \#\ of\ nodes$

- $p_i^{(n+1)} = \alpha \boxed{\sum_j \frac{A_{ij}}{k_j} p_j^{(n)}} + (1-\alpha)\boxed{\frac{1}{n}}$

The probability to see Node i

The probability after seeing Node j

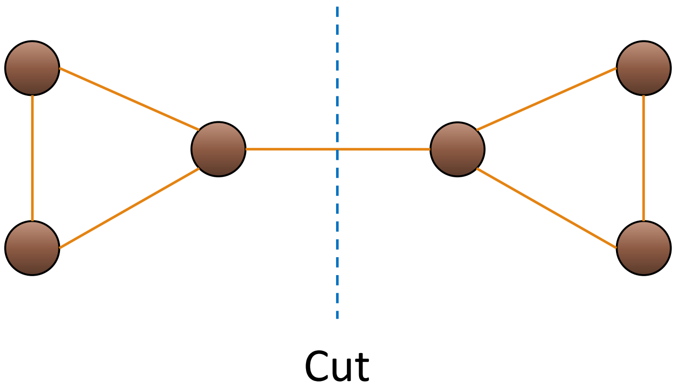
The probability to randomly see Node i



Network clustering

Cluster

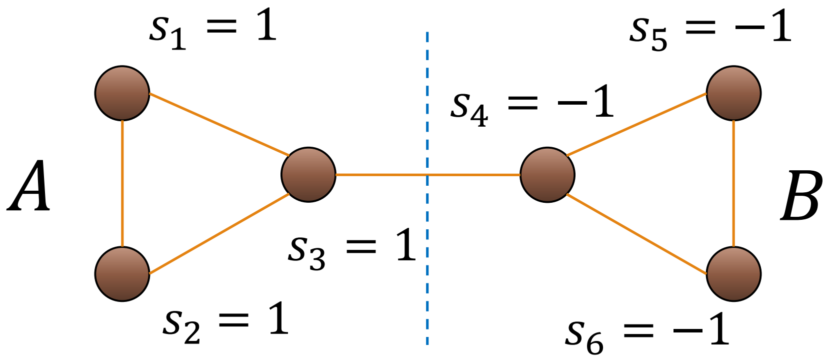
- A part of the network which is densely connected to each other
  - Cut the edges which connect clusters (MinCut)
  - Then, how should we define “densely connected” or “MinCut”?



Spectral clustering

Spectral clustering

- Separate a network into two clusters
  - easily extendable to more clusters
- Assign +1 or -1 as a cluster label to each node
- Two clusters should have a minimum bridge between them



Normalized cut – (1)

- $cut(A, B)$: the number of edges bridging Cluster A and Cluster B

$$cut(A, B) = \sum_{i \in A} \sum_{j \in B} A_{ij}$$

$$= \frac{1}{2} \cdot \frac{1}{4} \sum_{i,j} A_{ij} \left(s_i - s_j\right)^2$$

Normalized cut – (2)

- $vol(A)$: the number of edges connected to nodes in Cluster A

$$vol(A) = \sum_{i \in A} k_i = \frac{1}{2} \sum_i k_i (1 - s_i)$$

$$vol(B) = \sum_{i \in B} k_i = \frac{1}{2} \sum_j k_j (1 + s_j)$$

Normalized cut – (3)

- $ncut(A,B)$: the ratio of the number of bridges to nodes in clusters

$$ncut(A,B) = \frac{cut(A,B)}{vol(A)} + \frac{cut(A,B)}{vol(B)}$$

$$= \underbrace{\boxed{cut(A,B)}}_{\text{Part 1}} \underbrace{\boxed{\left(\frac{1}{vol(A)} + \frac{1}{vol(B)}\right)}}_{\text{Part 2}}$$

Part 2

$$\frac{2}{\sum_i k_i(1-s_i)} + \frac{2}{\sum_j k_j(1+s_j)} = 2\left(\frac{1}{2\sum_i k_i - \sum_i k_i(1+s_i)} + \frac{1}{\sum_j k_j\left(1+s_j\right)}\right)$$

$$= \frac{1}{\sum_i k_i}\left(\frac{1}{1-\frac{\sum_i k_i(1+s_i)}{2\sum_i k_i}} + \frac{1}{\frac{\sum_i k_i(1+s_i)}{2\sum_i k_i}}\right)$$

$$= \frac{1}{\sum_i k_i}\left(\frac{1}{1-r} + \frac{1}{r}\right)$$

$$= \frac{1}{\sum_i k_i}\frac{1}{r(1-r)} = \frac{1}{\sum_i k_i\left(\frac{1-r}{2}(1+s_i) - \frac{r}{2}(1-s_i)\right)^2}$$

The detail of the calculation

$$\sum_i k_i \left(\frac{(1-r)(1+s_i) - r(1-s_i)}{2}\right)^2$$

$$= \frac{1}{4}\sum_i k_i\,(1-r)^2(1+s_i)^2 - \frac{2}{4}r(1-r)\sum_i k_i\,(1+s_i)(1-s_i) + \frac{r^2}{4}\sum_i k_i\,(1-s_i)^2$$

$$= \frac{1}{2}(1-r)^2\sum_i k_i\,(1+s_i) + \frac{1}{2}r^2\sum_i k_i\,(1-s_i)$$

$$= (1-r)^2\sum_i k_i\,r + r^2\sum_i k_i\,(1-r)$$

$$= \sum_i k_i\,r(1-r)$$

Part 1

$$\frac{1}{8}\sum_{i,j} A_{ij}\left(s_i^2 - 2s_i s_j + s_j^2\right) = \frac{1}{4}\sum_{i,j}\left(A_{ij}s_i^2 - A_{ij}s_i s_j\right)$$

$$= \frac{1}{4}\sum_{i,j}\left(k_i\delta_{ij}s_i s_j - A_{ij}s_i s_j\right)$$

$$= \frac{1}{4}\sum_{i,j}\left(k_i\delta_{ij} - A_{ij}\right) s_i s_j$$

$$= \frac{1}{2}\sum_{i,j}\left(k_i\delta_{ij} - A_{ij}\right)\frac{(1-r)(1+s_i) - r(1-s_i)}{2}\,\frac{(1-r)\left(1+s_j\right) - r\left(1-s_j\right)}{2}$$

The detail of the calculation

$$\sum_{i,j}(k_i\delta_{ij} - A_{ij})\frac{(1-r)(1+s_i) - r(1-s_i)}{2}\frac{(1-r)(1+s_j) - r(1-s_j)}{2}$$

$$= \frac{1}{4}\sum_{i,j}(k_i\delta_{ij} - A_{ij})((1-r)^2(1+s_i)(1+s_j) + r^2(1-s_i)(1-s_j) - 2r(1-r)(1-s_i)(1+s_j))$$

$$= \frac{1}{2}\sum_{i,j}(k_i\delta_{ij} - A_{ij})(r^2 s_i s_j + (1-r)^2 s_i s_j + 2r(1-r) s_i s_j)$$

$$= \frac{1}{2}\sum_{i,j}(k_i\delta_{ij} - A_{ij})(1 - 2r + 2r^2 + 2r(1-r)) s_i s_j$$

$$= \frac{1}{2}\sum_{i,j}(k_i\delta_{ij} - A_{ij}) s_i s_j$$

Normalized cut – (3)

- $ncut(A,B)$: the ratio of the number of bridges to nodes in clusters

$$2 \cdot ncut(A,B) = \frac{\sum_{i,j}\left(k_i \delta_{ij} - A_{ij}\right) y_i y_j}{\sum_i k_i\, y_i^2} = \boxed{\frac{\boldsymbol{y}^T (D-A)\boldsymbol{y}}{\boldsymbol{y}^T D \boldsymbol{y}}}$$

where $y_i = \frac{(1-r)(1+s_i)-r(1-s_i)}{2}$

Relaxation of condition

- We want to find $\boldsymbol{y}$ that minimizes $\frac{\boldsymbol{y}^T(D-A)\boldsymbol{y}}{\boldsymbol{y}^T D\boldsymbol{y}}$
- Let elements in $\boldsymbol{y}$ take arbitrary real values
- Since there is a scaling symmetry, $\boldsymbol{y} \to a\boldsymbol{y}$ gives the same ncut value.
  - Take $\boldsymbol{y}$ so as to $\boldsymbol{y}^T D\boldsymbol{y} = 1$
  - Then, the problem is to minimize $\boldsymbol{y}^T(D-A)\boldsymbol{y}$ under the condition $\boldsymbol{y}^T D\boldsymbol{y} = 1$

Generalized eigenvalue problem

- Applying Lagrange multiplier, we get

$$\mathrm{L} = \boldsymbol{y}^T(D - A)\boldsymbol{y} - \lambda(\boldsymbol{y}^T D\boldsymbol{y} - 1)$$

$$\frac{1}{2}\frac{\partial \mathrm{L}}{\partial \boldsymbol{y}} = (D - A)\boldsymbol{y} - \lambda D\boldsymbol{y} = 0$$

$$(D - A)\boldsymbol{y} = \lambda D\boldsymbol{y}$$

After solving the generalized eigenvalue problem …

- Apply K-means problem

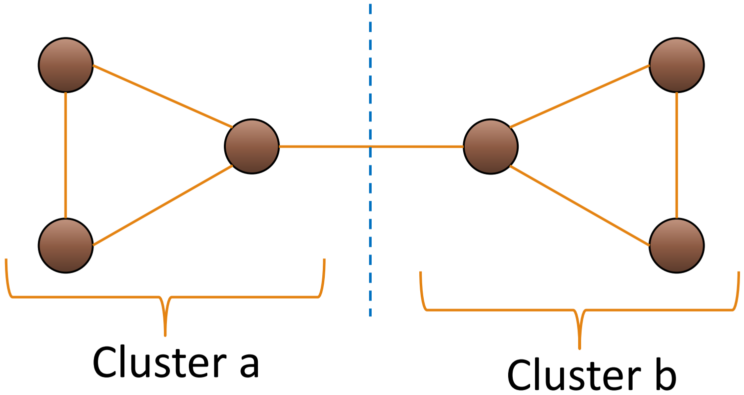
Modularity maximization

Modularity Q

- $e_{\mathrm{aa}} = \sum_{ij} \frac{A_{ij}}{2M} \delta(c_i, a) \delta(c_j, a)$
  - the probability to find edges connecting nodes inside Cluster a
- $a_{\mathrm{a}} = \sum_b e_{\mathrm{ab}} = \sum_i \frac{k_{\mathrm{i}}}{2\mathrm{M}} \delta(c_i, a)$
  - the probability of edges connecting to Cluster a
  - a_{a}^2 is a probability to find edges under the assumption that edges are randomly connected to nodes
- $Q = \sum_a (e_{aa} - a_a^2) = \frac{1}{2M} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2M}) \delta\left(c_i, c_j\right)$

An example

- The number of edges $M = 7$
- $e_{\mathrm{aa}} = \frac{3}{7}, e_{\mathrm{ab}} = \frac{1}{7}, e_{\mathrm{bb}} = \frac{3}{7}$
- $a_{\mathrm{a}} = e_{\mathrm{aa}} + e_{ab} = \frac{4}{7}, a_{\mathrm{a}}^2 = \frac{16}{49} = a_{\mathrm{b}}^2$
- $Q = e_{\mathrm{aa}} - a_{\mathrm{a}}^2 + e_{\mathrm{bb}} - a_{\mathrm{b}}^2 = 2\left(\frac{3}{7} - \frac{16}{49}\right) = \frac{10}{49}$



Resolution limit

- Modularity maximization tends to give larger clusters than expected.