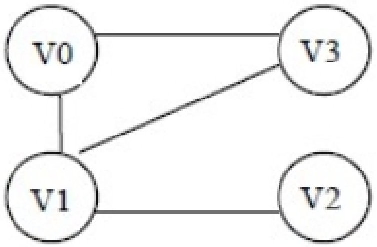
Graph Cluster

1. Graph theory
2. Markov chains
3. the different definitions of clusters
4. cluster properties
5. Measures for identifying clusters

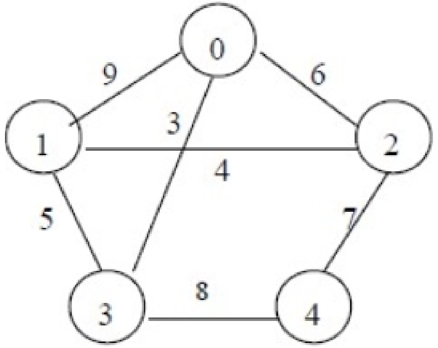
1. Graph theory

1. $G = (V, E), \ |V| = n, \ |E| = m$

2. Adjacency matrix A



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



$$A = \begin{pmatrix} \infty & 9 & 6 & 3 & \infty \\ 9 & \infty & 4 & 5 & \infty \\ 6 & 4 & \infty & \infty & 7 \\ 3 & 5 & \infty & \infty & 8 \\ \infty & \infty & 7 & 8 & \infty \end{pmatrix}$$

3. Degree matrix D

$$\mathbf{D} = \begin{pmatrix} \deg(v_1) & 0 & 0 & \dots & 0 & 0 \\ 0 & \deg(v_2) & 0 & \dots & 0 & 0 \\ 0 & 0 & \deg(v_3) & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \deg(v_{n-1}) & 0 \\ 0 & 0 & 0 & \dots & 0 & \deg(v_n) \end{pmatrix}$$

4. Density

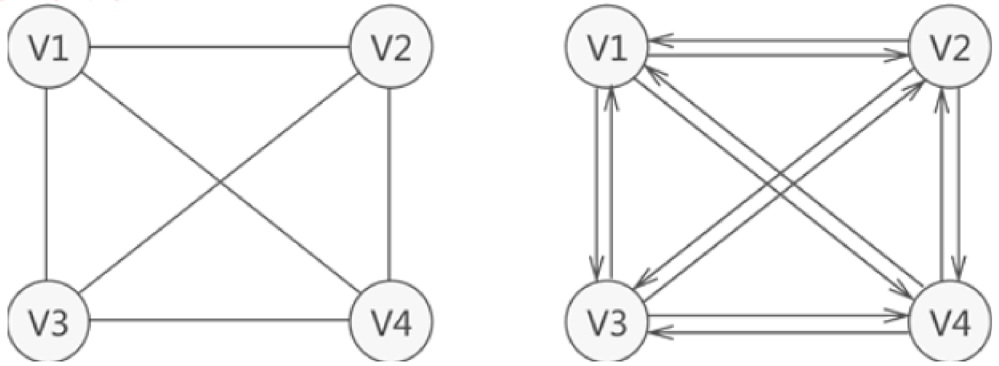
$$\delta(G) = \frac{m}{\binom{n}{2}} \qquad \delta'(G) = \frac{m}{n} \qquad \delta'_{max}(G) = \max_{S \subset V} \frac{|E(S)|}{|S|}$$

$\binom{n}{2}$ = n(n-1)/2 *or* n(n-1)

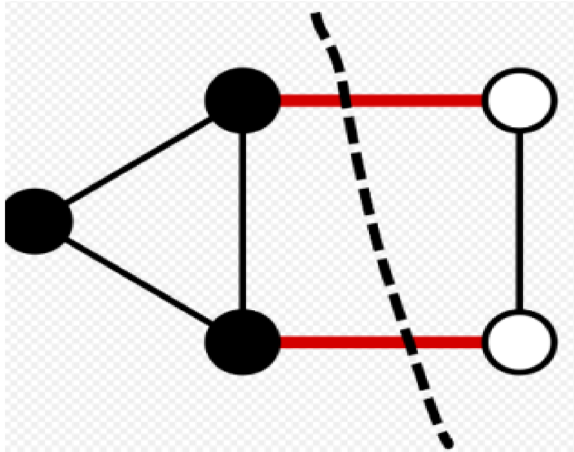
(S is a subgraph)

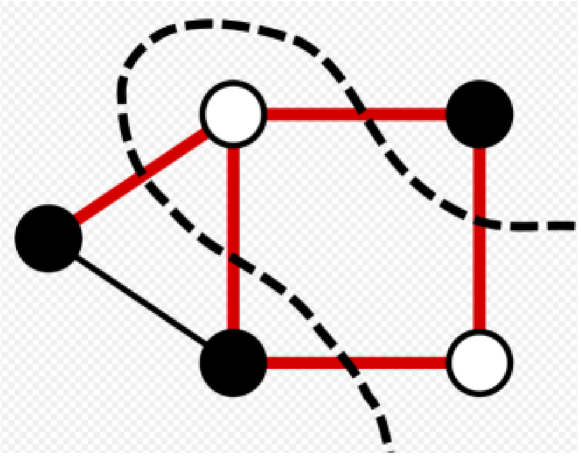
For n ∈ {0, 1}, we set δ (G) = 0

A graph of density one is called complete



5. Cut (a graph G = (V, E) into two disjoint nonempty sets S and V\S)





$c(S, V\backslash S) = |\{\{v, u\} \in E \mid u \in S, v \in V\backslash S\}|$ (cut size)

$\deg(S) = \Sigma_{v \in S} \deg(v)$

6. If such a path exists, v and u are connected. The path is simple if no vertex is repeated.

$$\{v, v_1\}, \{v_1, v_2\}, \ldots, \{v_{k-1}, v_k\}, \{v_k, u\}$$

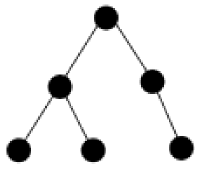
7. the shortest path = $SUM_{min} (\{v, v_1\},\ldots,\{v_{k-1}, v_k\},\{v_k, u\})$

8. A graph is connected if there exist paths between all pairs of vertices

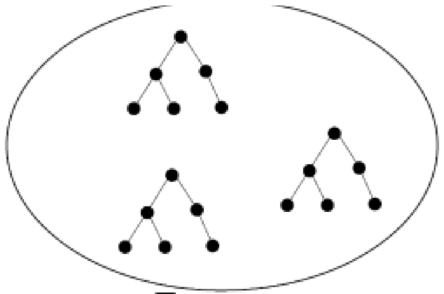
9. A cycle is a simple path that begins and ends at the same vertex

A acyclic graph is call a forest

A connected forest is called a tree

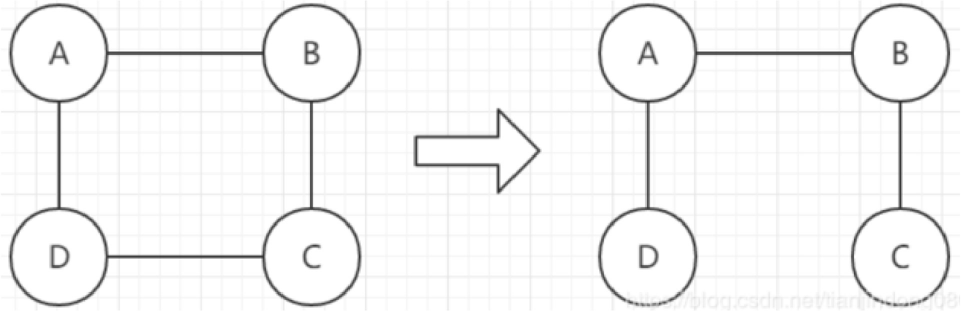


Tree

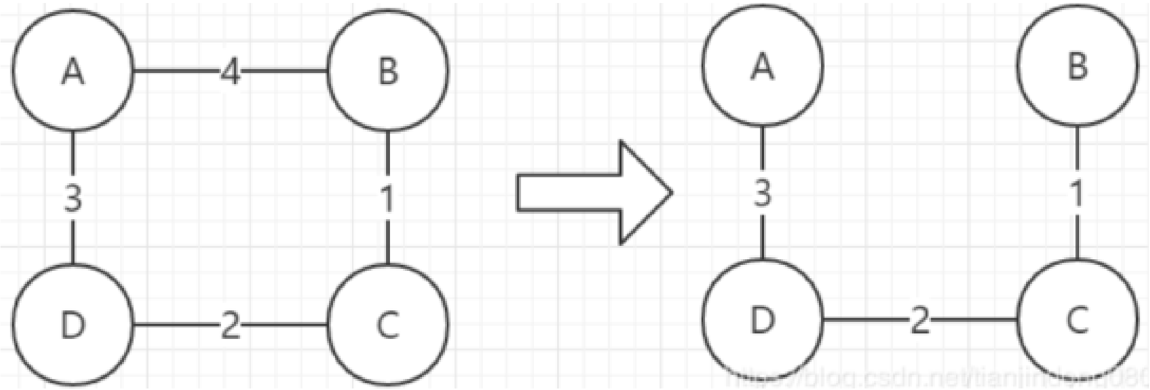


Forest

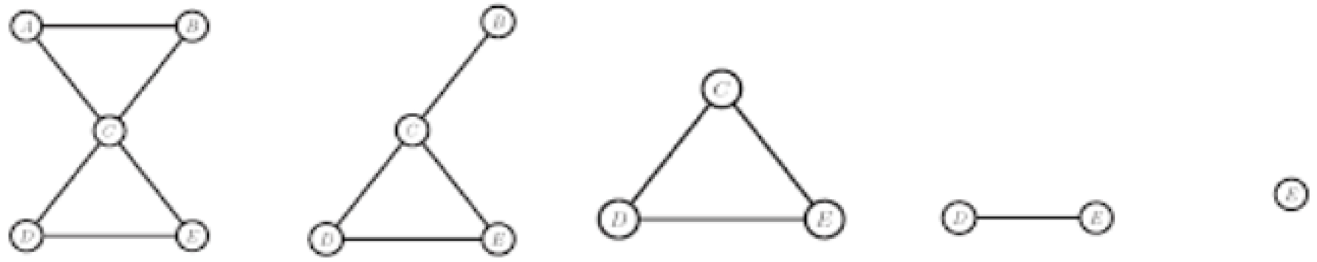
10. A connected acyclic subgraph that includes all vertices is called a spanning tree



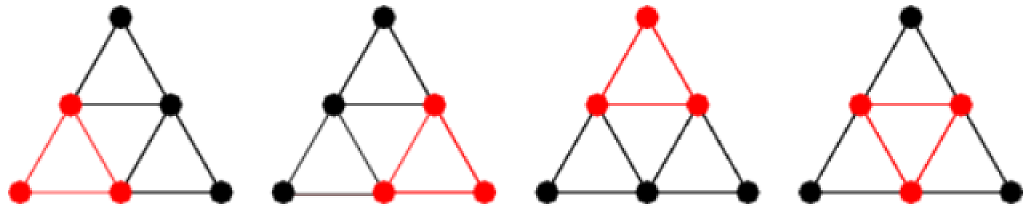
11. If the edges are assigned weights, the spanning tree with smallest total weight is called the minimum spanning tree.



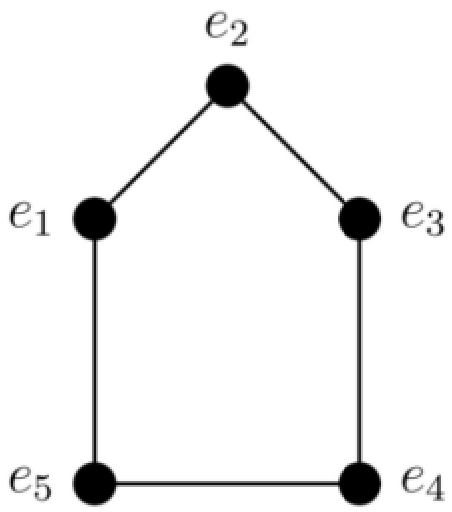
12. An induced subgraph of a graph G = (V, E) is the graph with the vertex set S ⊆ V with an edge set E(S) that includes all such edges {v, u} in E with both of the vertices v and u included in the set S

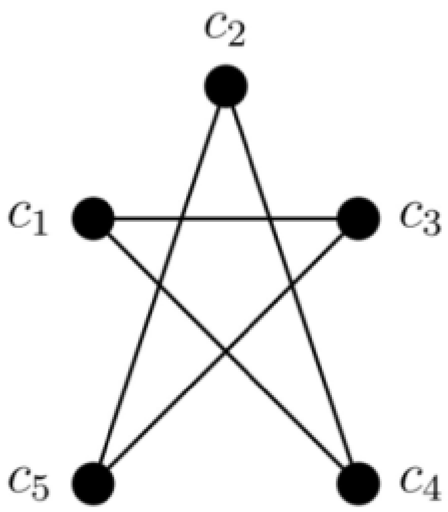


13. An induced subgraph that is a complete graph is called a clique (maximal clique)



14. Two graphs $G_i = (V_i, E_i)$ and $G_j = (V_j, E_j)$ are isomorphic if there exists a bijective (one-to-one) mapping $f : V_i \rightarrow V_j$ and $\{v, w\} \in Ei$ if and only if $\{f(v), f(w)\} \in Ej$





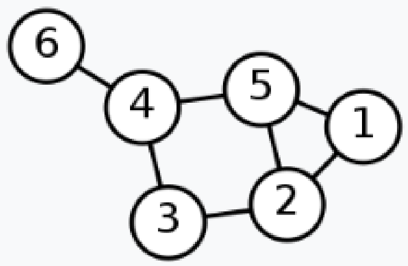
If G_i has some properties, G_j has some.

15. Graphs that have the same spectrum are called cospectral

16. The spectrum of a graph G = (V, E) is defined as the list of eigenvalues of the adjacency matrix

17. Laplacian matrix $L = D - A_G$

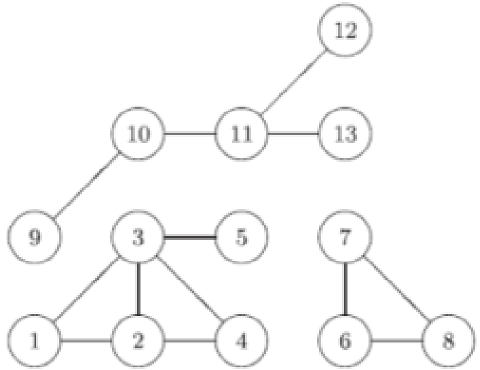
$$L_{i,j} := \begin{cases} \deg(v_i) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise} \end{cases}$$

标记图	度矩阵	邻接矩阵	拉普拉斯矩阵
	$\begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}$

Why Laplacian? Zero eigenvalue of Laplacian means a cluster

(Adjacency can not do) (Laplacian has at least one 0 eigenvalue)

Three zero eigenvalues of graph mean it has three clusters



	v_1	v_2	v_3
1	0	0	1
2	0	0	1
3	0	0	1
4	0	0	1
5	0	0	1
6	0	1	0
7	0	1	0
8	0	1	0
9	1	0	0
10	1	0	0
11	1	0	0
12	1	0	0
13	1	0	0

18. Normalized Laplacian

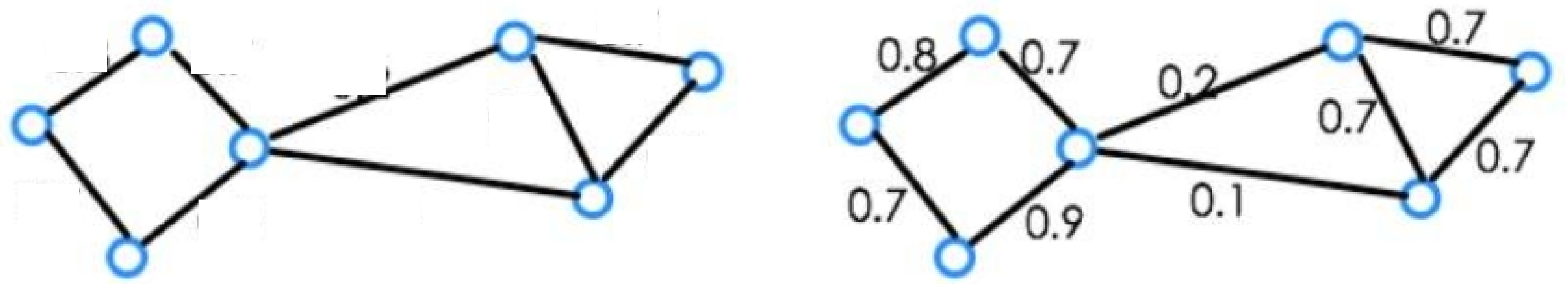
eigenvalues∈[0， 2], smallest eigenvalue is always zero

$$\mathcal{L} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}_G\mathbf{D}^{-\frac{1}{2}}$$

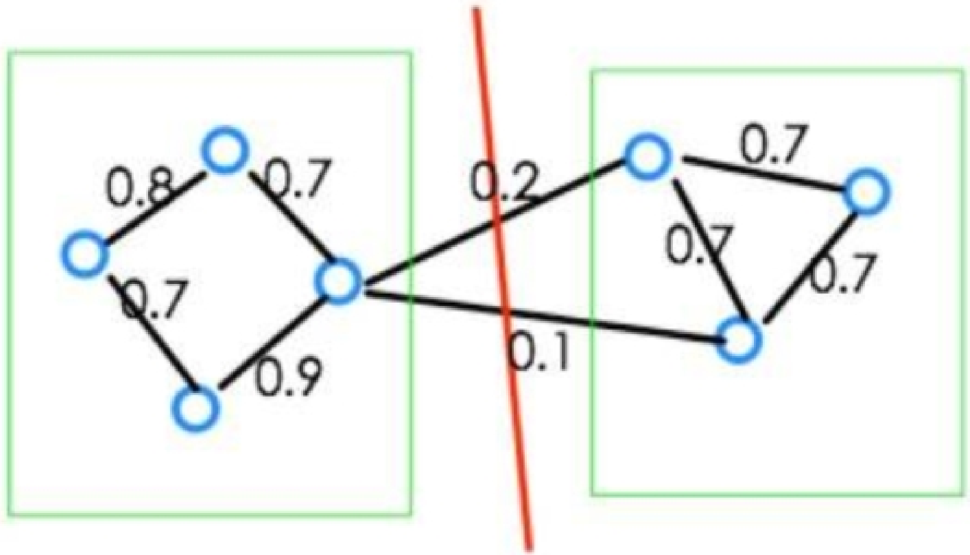
$$\mathcal{L}_{uv} = \begin{cases} 1, & \text{if } u = v \text{ and } \deg(v) > 0, \\ -\dfrac{1}{\sqrt{\deg(u)\cdot\deg(v)}}, & \text{if } u \in \Gamma(v), \\ 0, & \text{otherwise.} \end{cases}$$

$$L = \begin{bmatrix} 1 & \frac{-1}{\sqrt{1\cdot 4}} & 0 & 0 & 0 \\ \frac{-1}{\sqrt{4\cdot 1}} & 1 & \frac{-1}{\sqrt{4\cdot 2}} & \frac{-1}{\sqrt{4\cdot 3}} & \frac{-1}{\sqrt{4\cdot 2}} \\ 0 & \frac{-1}{\sqrt{2\cdot 4}} & 1 & \frac{-1}{\sqrt{2\cdot 3}} & 0 \\ 0 & \frac{-1}{\sqrt{3\cdot 4}} & \frac{-1}{\sqrt{3\cdot 2}} & 1 & \frac{-1}{\sqrt{3\cdot 2}} \\ 0 & \frac{-1}{\sqrt{2\cdot 4}} & 0 & \frac{-1}{\sqrt{2\cdot 3}} & 1 \end{bmatrix}$$

19. spectral clustering reconstruct similarity graph with unweighted undirected graph by the similarity of vertexes



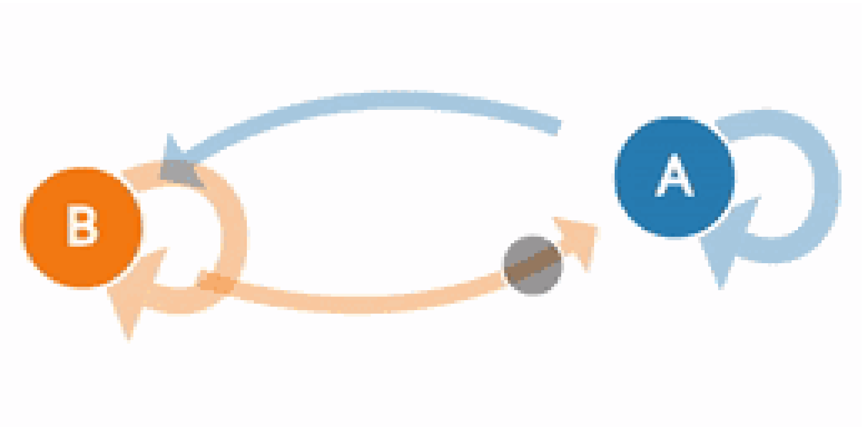
20. Minimum cut

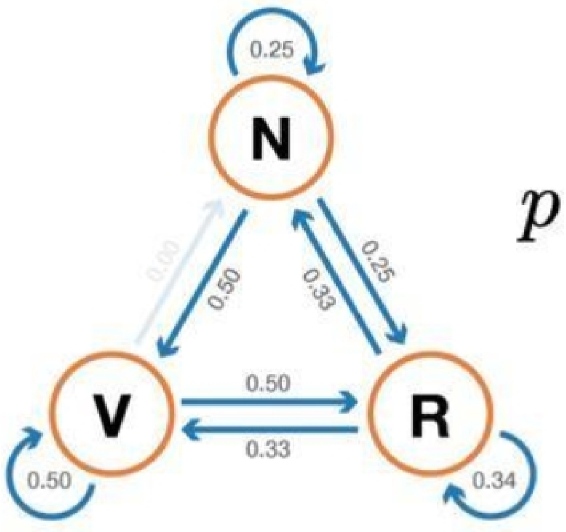


2. Markov chains

1. A Markov chain is a stochastic process in which future states only depend on the **current state**, not the past. (No Memory)

2. The probabilities for moving to another state from current state form the transition matrix of the Markov chain.



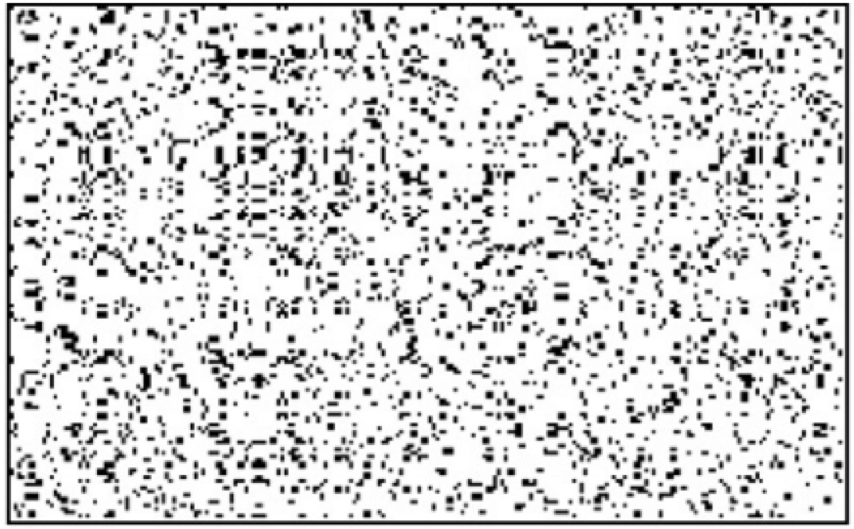


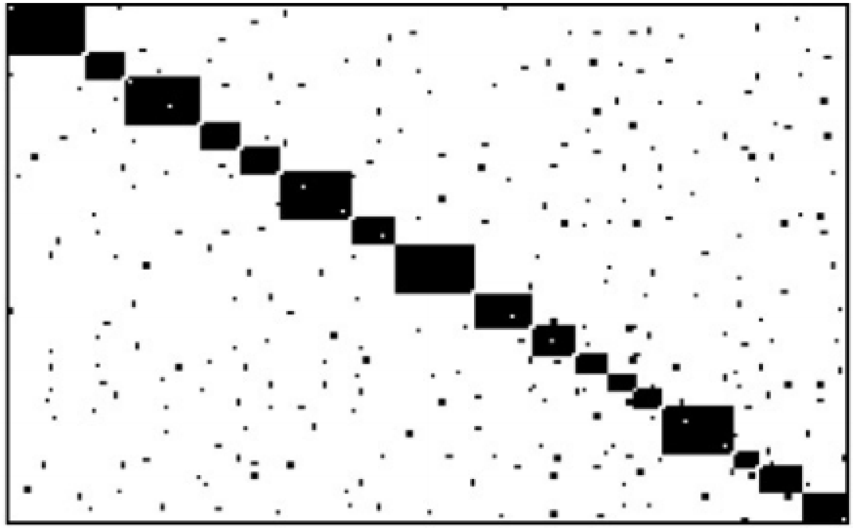
$$p = \begin{pmatrix} 0.25 & 0.50 & 0.25 \\ 0.00 & 0.50 & 0.50 \\ 0.33 & 0.33 & 0.34 \end{pmatrix}$$

3. the different definitions of clusters

1. a graph with $n = 210$ vertices and $m = 1505$ edges

(Matrix diagonalization)

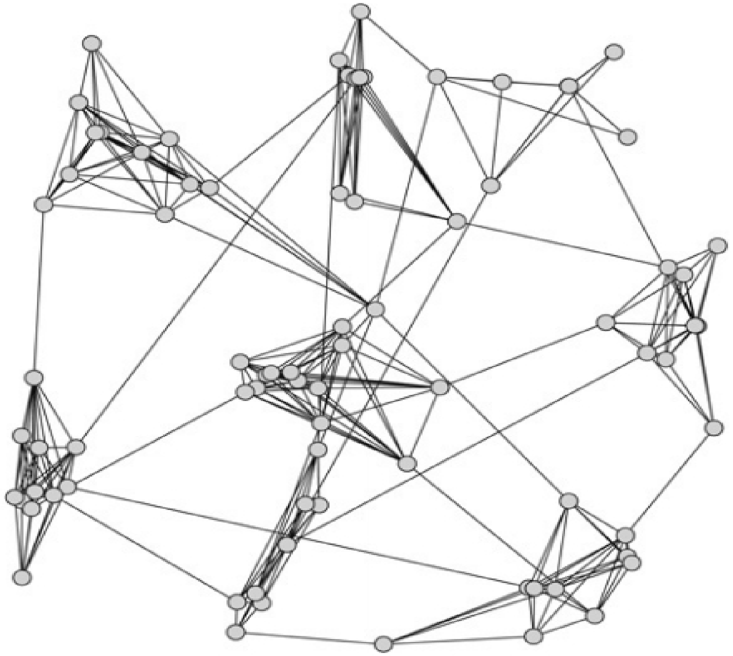




2. Generation models

a. **uniform random graph**
(With n vertices, each of the n(n-1)/2 possible edges)
(each pair of vertices independently degree distribution is Poissonian)
(construction uniformly， No dense clusters)

b. **relaxed caveman structure**
(linking together a ring of small complete graphs called caves)
(social network)



Planted l-partition model

(A generalization of the uniform random graph, especially designed to produce clusters)

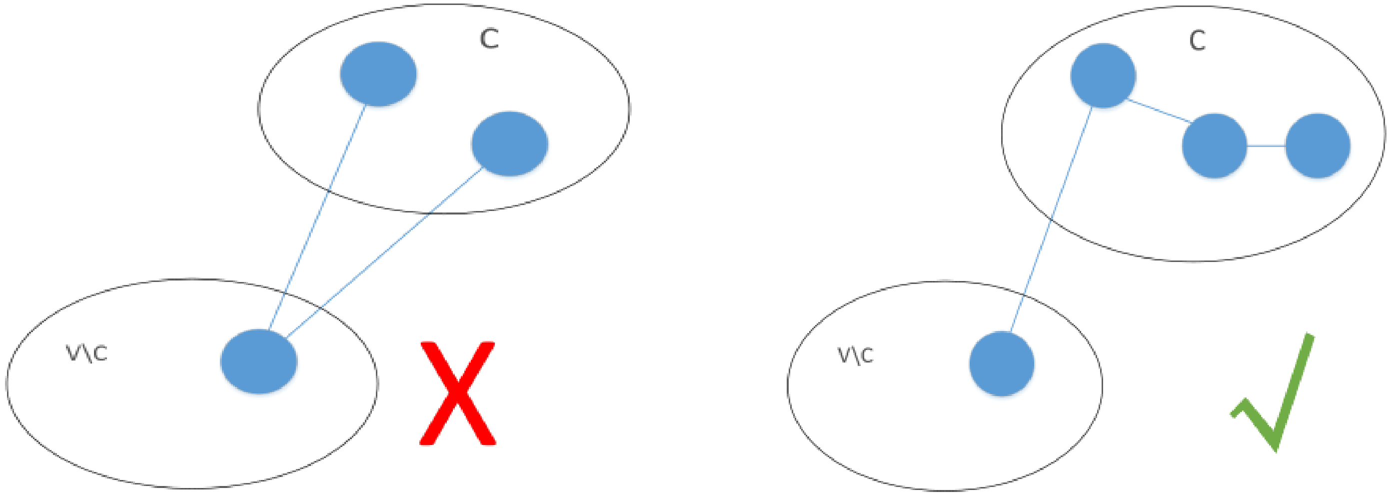
(n= l*k vertices, partitioned into l groups each with k vertices)

(each pair of vertices that are in the same group share an edge with the higher probability p, whereas each pair of vertices in different groups shares an edge with the lower probability r)

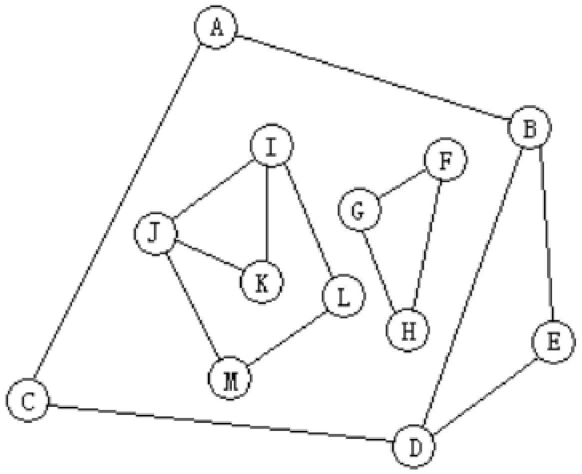
4. cluster properties

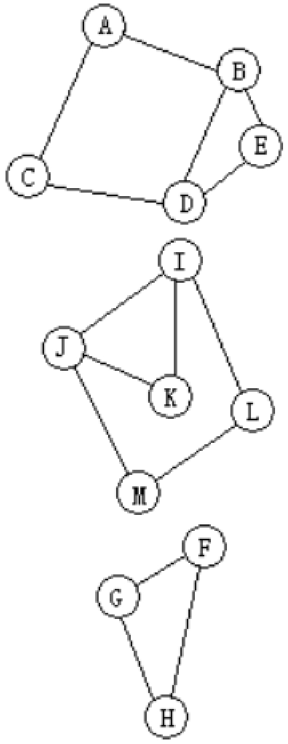
1. A cluster should be at least a connected subgraph. Preferably more paths (dense) within the subgraph

2. If a vertex u cannot be reached from a vertex v, they should not be grouped in the same cluster. Two vertices v and u in C also need to be connected by a path that only visits vertices included in C



3. when clustering a disconnected graph with known components, the clustering should usually be conducted on each component separately, unless some global restriction on the resulting clusters is imposed.



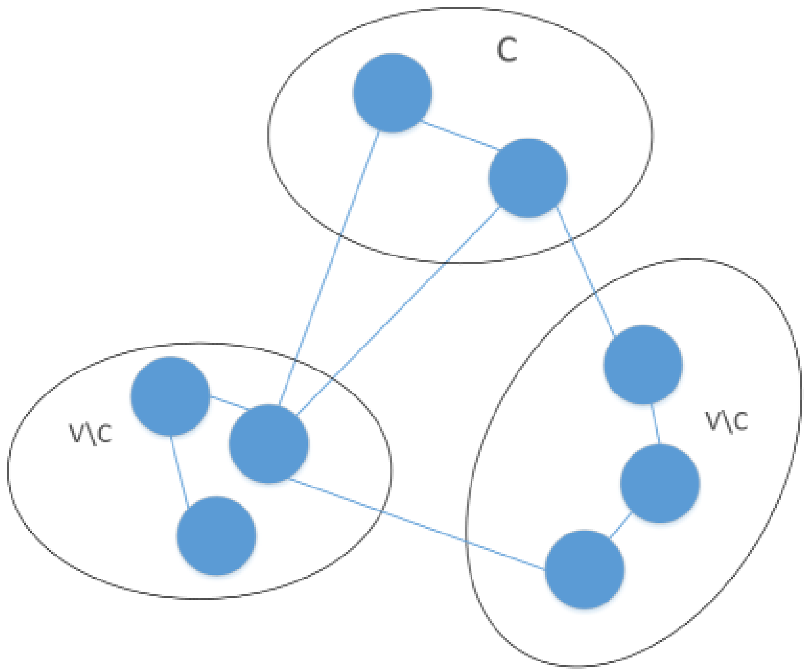


4. We classify the edges incident on $v \in C$ into two groups: internal edges that connect v to other vertices also in C, and external edges that connect v to vertices that are not included in the cluster C. (degext (v) = 0 implies that C containing v could be a good cluster)

$$\deg_{\text{int}}(v, \mathcal{C}) = |\Gamma(v) \cap \mathcal{C}|$$

$$\deg_{\text{ext}}(v, \mathcal{C}) = |\Gamma(v) \cap (V \backslash \mathcal{C})|$$

$$\deg(v) = \deg_{\text{int}}(v, \mathcal{C}) + \deg_{\text{ext}}(v, \mathcal{C})$$



5. the internal or intra-cluster density

$$\delta_{\text{int}}(\mathcal{C}) = \frac{|\{\{v,u\} \mid v \in \mathcal{C}, u \in \mathcal{C}\}|}{|\mathcal{C}|\,(|\mathcal{C}|-1)}.$$

The intercluster density of a graph G

$$\delta_{\text{int}}(G \mid \mathcal{C}_1, \ldots, \mathcal{C}_k) = \frac{1}{k}\sum_{i=1}^{k} \delta_{\text{int}}(\mathcal{C}_i).$$

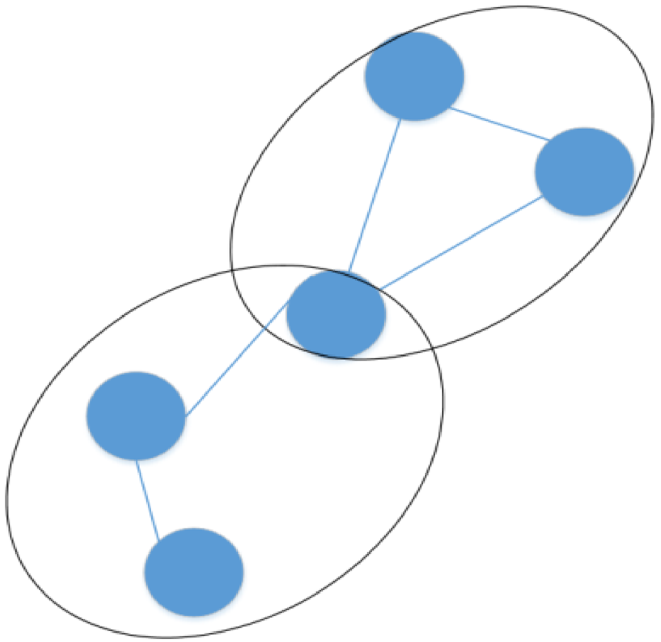
The external or inter-cluster density

$$\delta_{\text{ext}}(G \mid \mathcal{C}_1, \ldots, \mathcal{C}_k) = \frac{\left|\left\{\{v,u\} \mid v \in \mathcal{C}_i, u \in \mathcal{C}_j, i \neq j\right\}\right|}{n(n-1) - \sum_{\ell=1}^{k} (|\mathcal{C}_\ell|\,(|\mathcal{C}_\ell|-1))}.$$

6. the internal density of a good clustering should be notably higher than the density of the graph δ (G) and the intercluster density of the clustering should be lower than the graph density

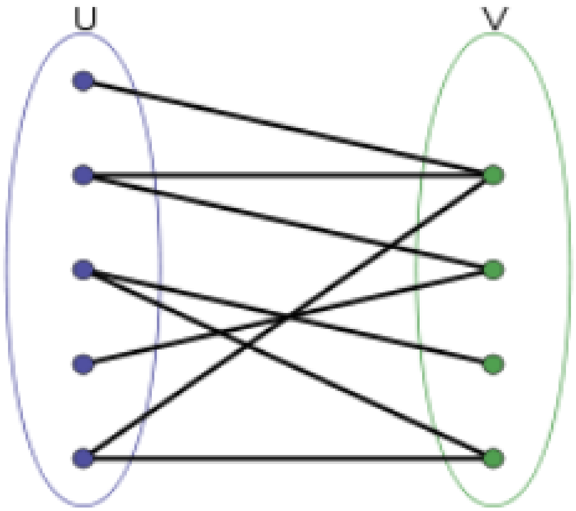
7. the loosest possible definition of a graph cluster is that of a connected component, and the strictest definition is that each cluster should be a maximal clique

8. It is not always clear whether each vertex should be assigned fully to a cluster or could it instead have different "levels of membership" in several clusters?



9. A fuzzy graph allows nodes to be in multiple clusters

10. Bipartite graphs



customers books

cluster situation:

grouping the customers by the types of books they purchase

grouping books purchased by the same people

cluster method:

the overlap of the neighbourhoods the one side of the graph reflects the similarity of the vertices of the other side and vice versa

5. Measures for identifying clusters

1. How to identify a good cluster?

i. compute some values for the vertices and then classify the vertices into clusters based on the values obtained

ii. compute a **fitness measure** over the set of possible clusters and then choose among the set of cluster candidates those that optimize the measure used

2. Vertex similarity (vertex-based)

i. Distance-based measures
(Compare the internal properties of the node, such as the author of the book, content, etc.)

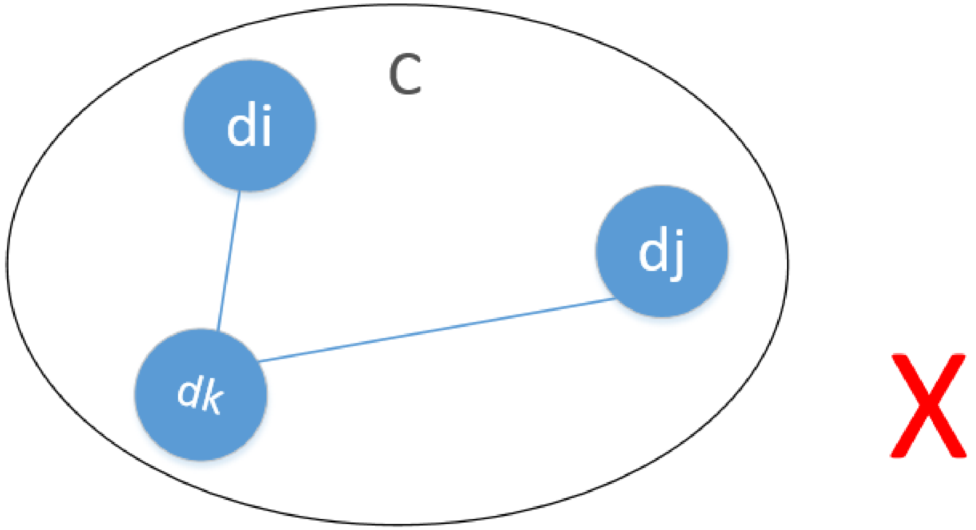
ii. Adjacency-based measures
(lack additional internal properties)
(Compare node extrinsic properties, such as whether the user owned by the book is the same)

iii. Connectivity measures
(depend on the path of the vertices)

i. Distance-based measures

a. The distance from a datum to itself is zero: dist (di, di) = 0
b. The distances are symmetrical: dist (di, dj) = dist (dj, di)
c. The triangle inequality holds:

$$\text{dist}(di, dj) \leq \text{dist}(di, dk) + \text{dist}(dk, dj)$$



For Euclidean

a. the Euclidean distance

$$\text{dist}\,(d_i d_j) = \sum_{k=1}^{n} \sqrt{(d_{i,k} - d_{j,k})^2}$$

b. the L2 norm, the Manhattan distance

$$\text{dist}\,(d_i d_j) = \max_{k \in [1,n]} |d_{i,k} - d_{j,k}|$$

c. the L1 norm

$$\text{dist}\,(d_i d_j) = \sum_{k=1}^{n} |d_{i,k} - d_{j,k}|$$

For unEuclidean

vector representations of textual data (document D_j, datum d_j)

a. cosine similarity (angle in [0, π))

$$\rho(d_i, d_j) = \arccos \frac{d_i \cdot d_j}{\sqrt{\sum_{k=1}^{n} (d_{i,k}^2)} \sqrt{\sum_{k=1}^{n} (d_{j,k}^2)}}.$$

b. the Jaccard coefficient

$$\rho(A, B) = \frac{|A| \cap |B|}{|A| \cup |B|}.$$

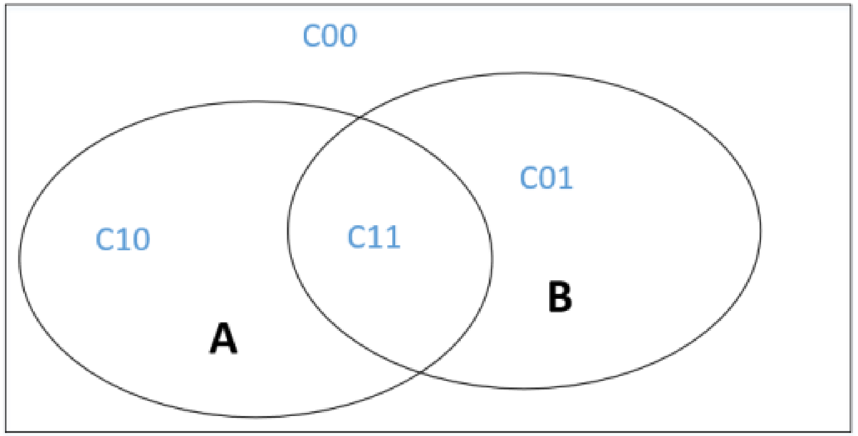
$$\rho(A, B) = \frac{C_{1,1}}{C_{0,1} + C_{1,0} + C_{1,1}}$$

Jaccard distance = 1-ρ(A,B)

$$\text{dist}\,(A, B) = \frac{C_{1,0} + C_{0,1}}{C_{0,1} + C_{1,0} + C_{1,1}}.$$

c. the Tanimoto coefficient

$$\rho(A, B) = \frac{A \cdot B}{\sqrt{\sum_{k=1}^{n} a_1} + \sqrt{\sum_{k=1}^{n} b_1} - A \cdot B}$$



ii. Adjacency-based measures

a. the overlap of their neighbourhoods ([0, 1])

$$\omega(v, w) = \frac{|\Gamma(v) \cap \Gamma(w)|}{|\Gamma(v) \cup \Gamma(w)|}$$

b. Pearson correlation (Expand on cosine similarity) ([-1, 1])

$$\frac{n\left(\sum_{k=1}^{n}(c_{i,k}c_{j,k})\right) - \deg(v_i)\deg(v_j)}{\sqrt{\deg(v_i)\deg(v_j)(n-\deg(v_i))\left(n-\deg(v_j)\right)}}.$$

iii. Connectivity measures

a good cluster

1) be highly connected to each other in the same cluster

2) if they are at least connected by a short path, it is not absolutely necessary that two included vertices v and u are connected by a direct edge

threshold the path length

a. all vertices in a cluster must be at distance at most k from each other

b. set the threshold k by the diameter of the input graph which is the maximum distance over all pairs of nodes

3. fitness measures (cluster-based)

i. Density measures (dense)

$$\delta_{int}(\mathcal{C}) = \frac{|\{u,v\} \mid u\in\mathcal{C}, v\in\mathcal{C}|}{|\mathcal{C}|(|\mathcal{C}|-1)}$$

ii. Cut-based measures (sparse)

$$\deg_{\text{int}}(\mathcal{C}) = |\{\{v,u\} \in E \mid v, u \in \mathcal{C}\}|$$

$$\deg_{\text{ext}}(\mathcal{C}) = |\{\{v,u\} \in E \mid v \in \mathcal{C}, u \in V\setminus\mathcal{C}\}|$$

$$= \text{cut}\ (\mathcal{C}, V\setminus\mathcal{C})$$

(independence measures)

$$\rho(\mathcal{C}) = \frac{\deg_{\text{int}}(\mathcal{C})}{\deg_{\text{int}}(\mathcal{C}) + \deg_{\text{ext}}(\mathcal{C})}$$

$$= \frac{\sum_{v\in\mathcal{C}} \deg_{\text{int}}(v,\mathcal{C})}{\sum_{v\in\mathcal{C}} \deg_{\text{int}}(v,\mathcal{C}) + 2\deg_{\text{ext}}(v,\mathcal{C})}$$

END

THANKS