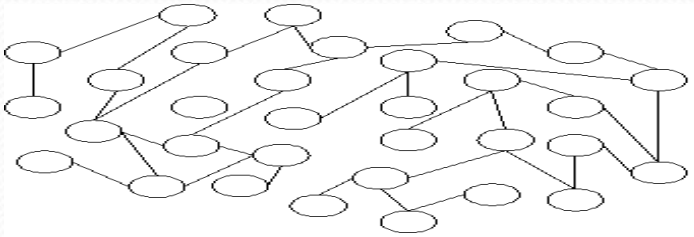
Social Network Analysis

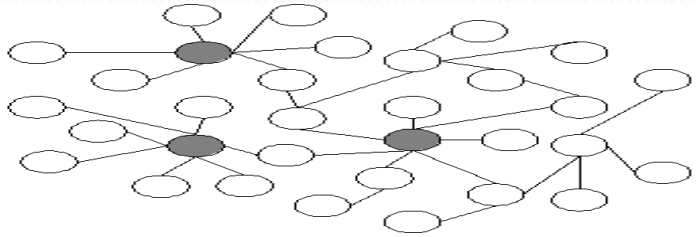
Models of Social Network Formation

- **Two popular models:**

1. **Random graph models:**
   - ✓ Links form by chance and simply governed by probabilistic rules.
   - ✓ Every pair of vertices have equal probability of having an edge between them.
   - ✓ the distribution of the number of neighbors of a vertex, or degree, is binomial, so most vertices have equal or similar degree.
   - ✓ Real networks are not always random graphs. The degree distribution is broad, with a tail that often follows a **power law:** therefore, many vertices with low degree coexist with some vertices with large degree.
   - ✓ A **scale-free** distribution is one where the frequency of degrees can be written in the form $f(d) = ad^{-b}$, for some parameters a and b, where d is the degree and f(d) is the relative frequency of nodes with degree d.



(a) Random network



(b) Scale-free network

Models of Social Network Formation

2. **Game theoretic models:**
   - ✓ Links form by choice more often.
   - ✓ Capture social and economic incentives while forming links.
   - ✓ Nodes often act strategically as link formation incurs costs in terms of cost, money and effort.

Network Formation Game

We represent the corresponding strategic network formation game with 3-tuple $< N, (Si)_{i \in N}, (ui)_{i \in N} >$ where:

1. N is the set of individuals in the network and we call them players.

2. For each $i \in N$, S_i is the set of strategies of player i. A strategy $s_i \in S_i$ of player i is the set of individuals with which player i wants to form a link.

3. For each $i \in N$, u_i is the utility of individual i and this utility depends on its neighborhood and the structure of the network.

Network Formation Game

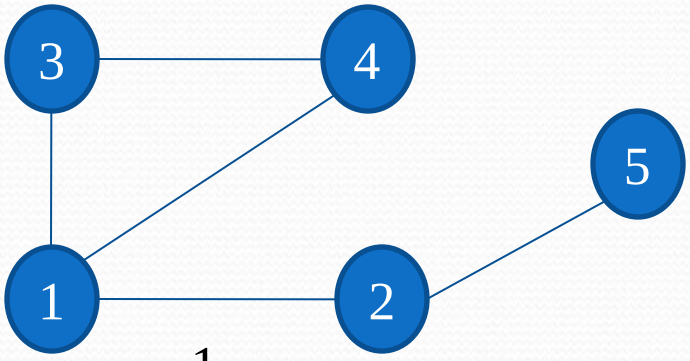
- Two fundamentally ways of modeling the formation of social contacts:

1. **Two-sided Link Formation**: A link is formed under mutual consent.
2. **One-sided Link Formation:** A link if formed under the consent of either of the individuals involved in the link formation.

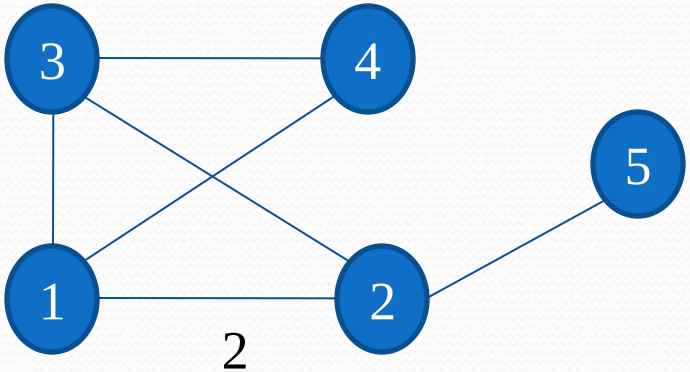
- **Network Formation Game: An Example**

  Consider N = {1, 2, 3, 4, 5} be the set of players. Assume that the strategies of the players are as follows:

  S1 = {2, 3, 4}, S2 = {1, 3, 4, 5}, S3 = {1, 4}, S4 = {1, 3}, S5 = {2}



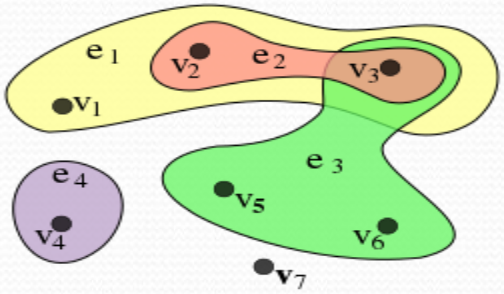
1



2

Different types of Social Networks

- Friends Networks, Email Networks, Collaboration Networks...
- Raw Social Media Networks comprise multiple types of vertices and edges, in which an edge can join any number of vertices. Mathematically they are represented by hyper graphs.

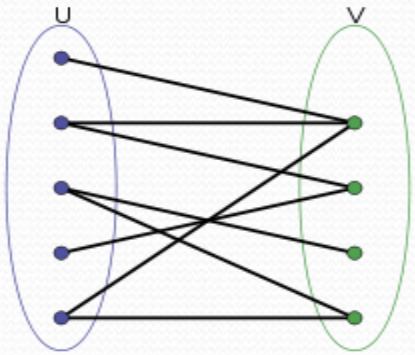


An Example of a Hyper-graph with
X = {v1, v2, v3, v4, v5, v6, v7}
E = {e1, e2, e3, e4} = {{v1, v2, v3}, {v2, v3}, {v3, v5, v6}, {v4}}

• A hyper-graph $H = \{X, E\}$ may be represented by a bipartite graph as follows: the sets X and E are the partitions of BG, and (x_1, e_1) are connected with an edge if and only if vertex x_1 is contained in edge e_1 in H.

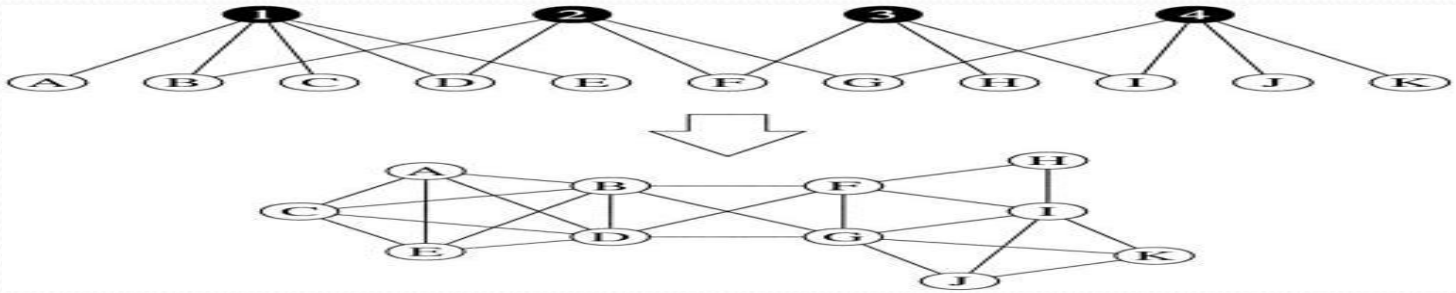
Different types of Social Networks

Hyper-graph to Bipartite Graph:



Here we can consider the set of vertices V to comprise the authors and papers of the system, i.e. V = {A,P}. Here the edge (A1, P2) means that author A1 has edited the paper P2.

These types of Social Network Data can be represented with bipartite graph.



Social Network Mining

- **Construction of graph from web data:**

Webpage www.x.com

href="www.y.com"

href = www.z.com

Webpage www.y.com

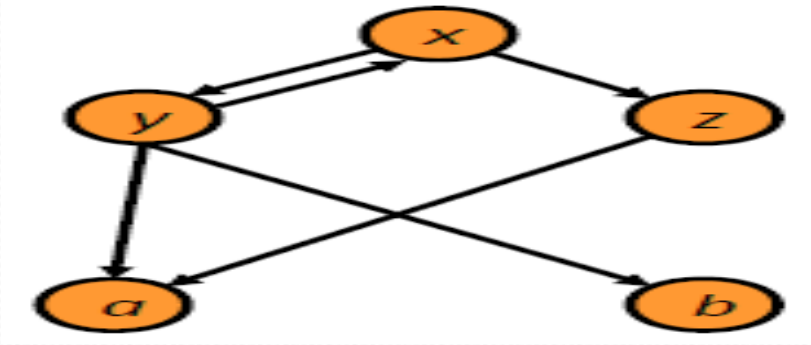
href="www.x.com"

href = "www.a.com"

href = "www.b.com"

Webpage www.z.com

href="www.a.com"

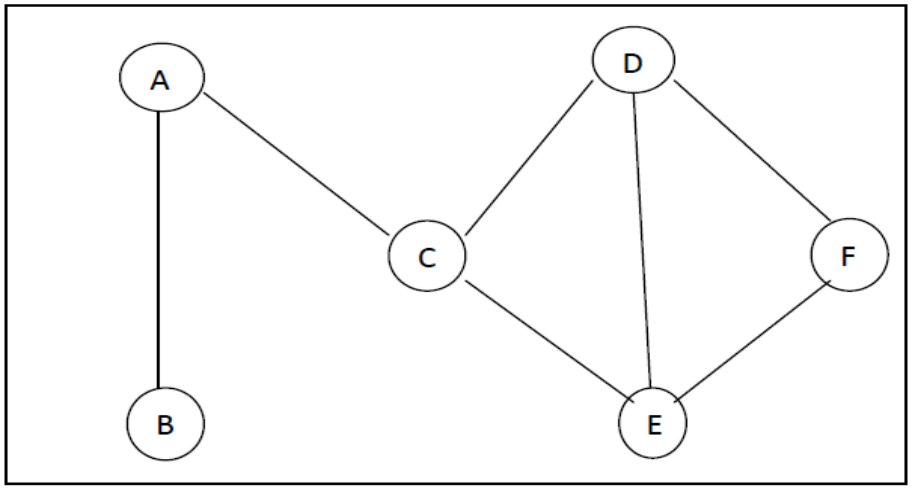


- **Classified into two categories:**

- Graph Mining
  - Finding the most important node in the network
  - Community Detection
- Text Mining
  - Opinion Mining and Sentiment Analysis

Measure node importance

- Centrality defines how important a node is within a network. It can be measured through various parameters.
- Degree Centrality: $DC(V_i) = d_i$--- for undirected graph.



A = 2, B = 1, C = 3
D = 3, E = 3, F = 2
According to Degree Centrality node rank is:
C, D, E
A, F
B

The nodes with higher in-degree is more prestigious.
The nodes with higher out-degree is more central.

} For Directed graph

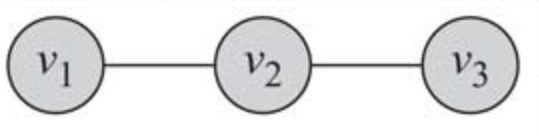
Measure node importance

- Eigenvector Centrality: tries to generalize Degree Centrality by incorporating the importance of the neighbors.
- Eigen Centrality: $\lambda Ce = ACe$ or $(A - \lambda I)Ce = 0$

  where λ = Eigen Value

  A = Adjacency matrix of the graph

  Ce = Eigen Vector



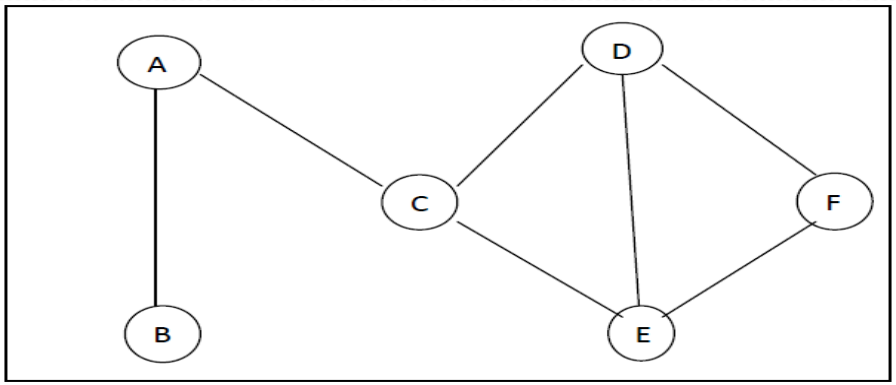
Here the Eigen values are $(-\sqrt{2}, 0, +\sqrt{2})$
For the largest Eigen value: $+\sqrt{2}$ we get three Eigen vectors: $\{1/2, +\sqrt{2}/2, 1/2\}$ which denotes that node v2 is the most central node and nodes v1 and v3 have equal centrality values.

Measure node importance

- Closeness Centrality: In closeness centrality, the intuition is that the more central nodes are, the more quickly they can reach other nodes.
- The smaller the average shortest path length, the higher the centrality for the node.
- Closeness centrality is defined as

$$Cc(v_i) = 1 / L(v_i)$$

where $l(v_i) = \frac{1}{n-1} \sum_{v_i \neq v_r} L_{i,r}$



A = 0.556 , B = 0.385, C = 0.714
D = 0.625, E = 0.625, F = 0.455
According to Closeness Centrality node rank is:
C
D, E
A
F
B
Now, A is promoted.

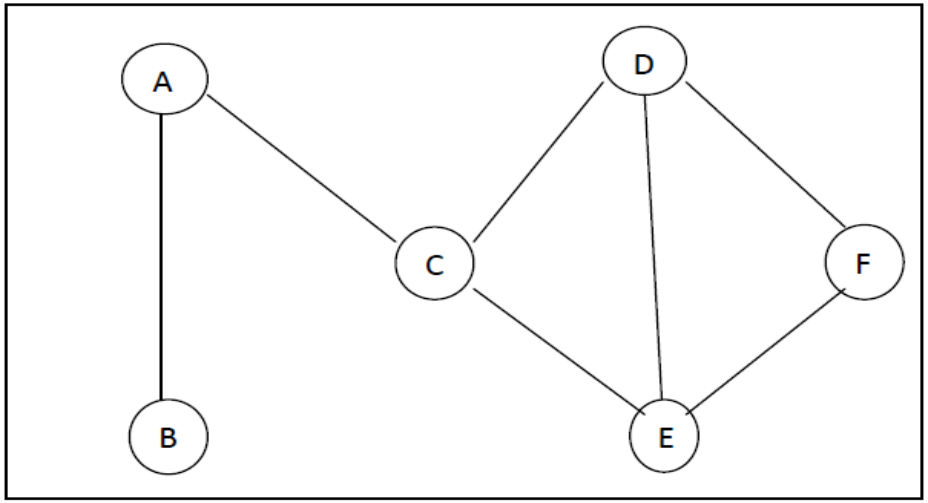
Measure node importance

- Betweenness Centrality: The Betweenness of a vertex v in a graph G = (V, E) is computed as follows:
- For each pair of vertices (s, t), compute the shortest paths between them.
- For each pair of vertices (s, t), determine the fraction of shortest paths that pass through the vertex (here, vertex v).
- Sum this fraction over all pairs of vertices (s, t).
- More compactly the Betweenness can be represented as:

$$\text{Betwenness}(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

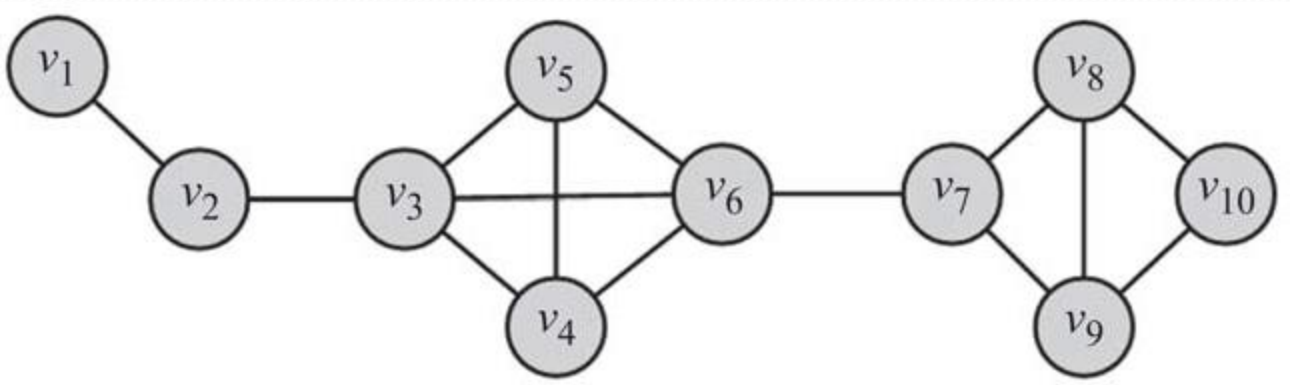
Measure node importance

- Betweenness Centrality:



A = 1/1 + 1/1 + 1/1 +2/2 = 4
B = 0
C = 1 + 1 + 1 + 1 + 1 + 1 = 6
D = 0.5 + 0.5 + 0.5 = 1.5
E = 0.5 + 0.5 + 0.5 = 1.5
F = 0
According to Betweenness Centrality node rank is:
C
A
D, E
B, F
Now, A is more important that D, E.

A Comparison between Centrality Methods



	First Node	Second Node	Third Node
Degree Centrality	V3 or v6	V6 or v3	V4, v5, v7, v8, v9
Eigenvector Centrality	v6	v3	V4 or v5
Closeness Centrality	v6	V3 or v7	V7 or v3
Betweenness Cent.	v6	v7	v3

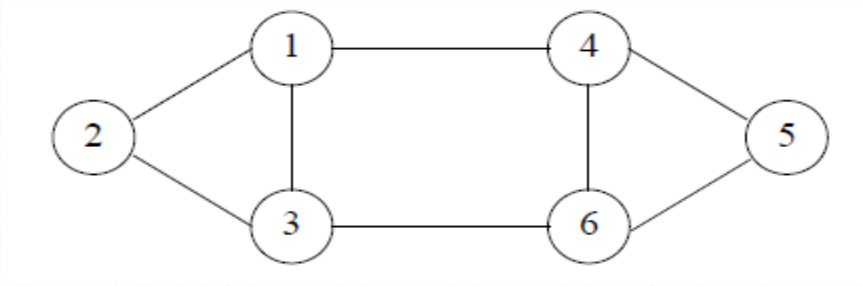
Community Detection

- The problem that community detection attempts to solve is the identification of groups of vertices that are more densely connected to each other than to the rest of the network.
- A Random Graph has no community.

Community Detection

- **Clustering of Social Network Graphs consist of several methodologies:**
- **Spectral Bisection:**

Let us consider a social network graph:



Laplacian Matrix of the above graph is:

$$\begin{bmatrix} 3 & -1 & -1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & 0 & 0 & -1 \\ -1 & 0 & 0 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & -1 & -1 & 3 \end{bmatrix}$$

Eigenvalue	0	1	3	3	4	5
Eigenvector	1	1	−5	−1	−1	−1
	1	2	4	−2	1	0
	1	1	1	3	−1	1
	1	−1	−5	−1	1	1
	1	−2	4	−2	−1	0
	1	−1	1	3	1	−1

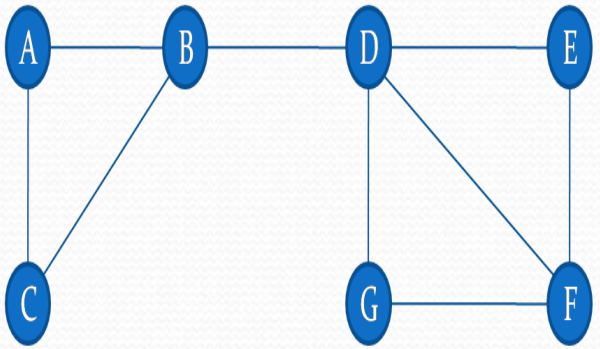
It makes the suggestion that one group should be {1, 2, 3}, and the other group should be {4, 5, 6}.

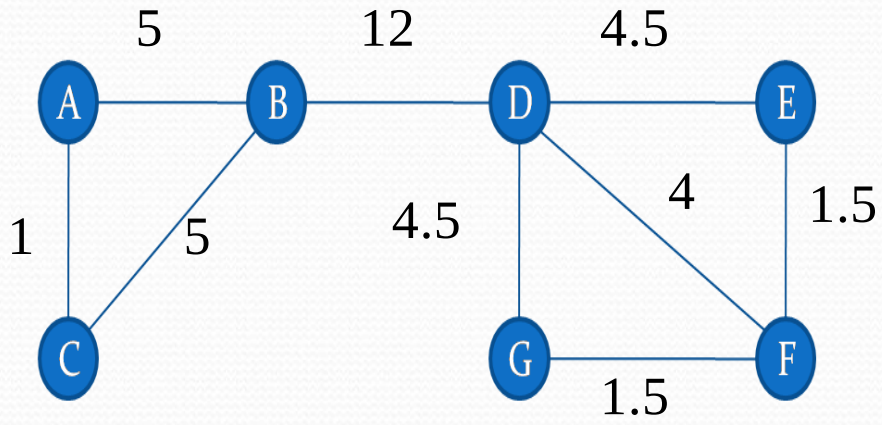
Community Detection

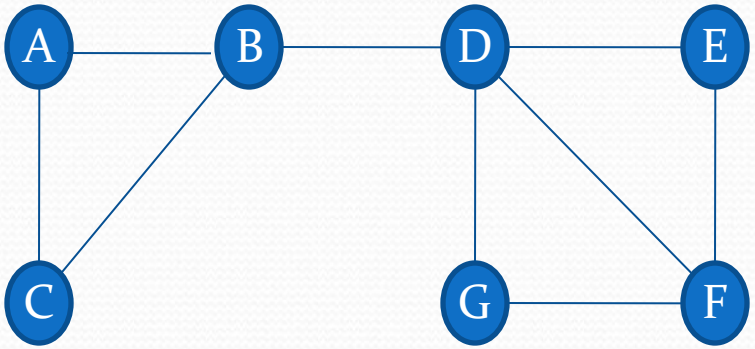
- **Agglomerative Hierarchical Clustering:**
  - Start with vertices, remove all edges.
  - Iteratively merge vertices based on similarity.
  - Compute similarity measure between vertices, no matter if they are connected or not.
  -

Community Detection

- Girvan-Newman Algorithm (Divisive):
  - Compute betweenness centrality for each edge.
  - Remove edge with highest score.
  - Re-compute all scores.
  - Repeat 2nd step.

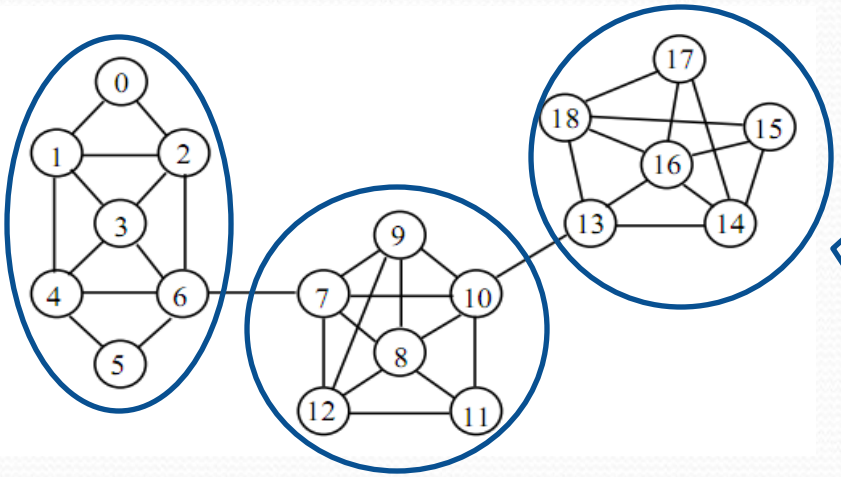




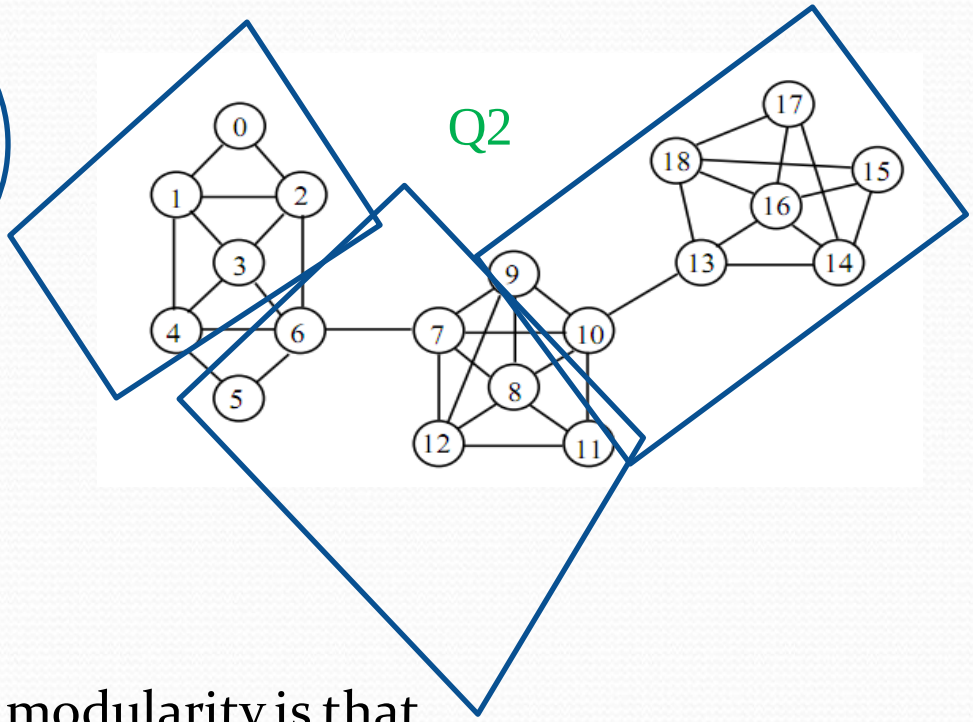


Community Detection

- Community Detection by Quality Optimization (Modularity maximization):
- Modularity indicates the quality of a given community structure.
- Let us consider two different types of community detection on same graph:



Q1



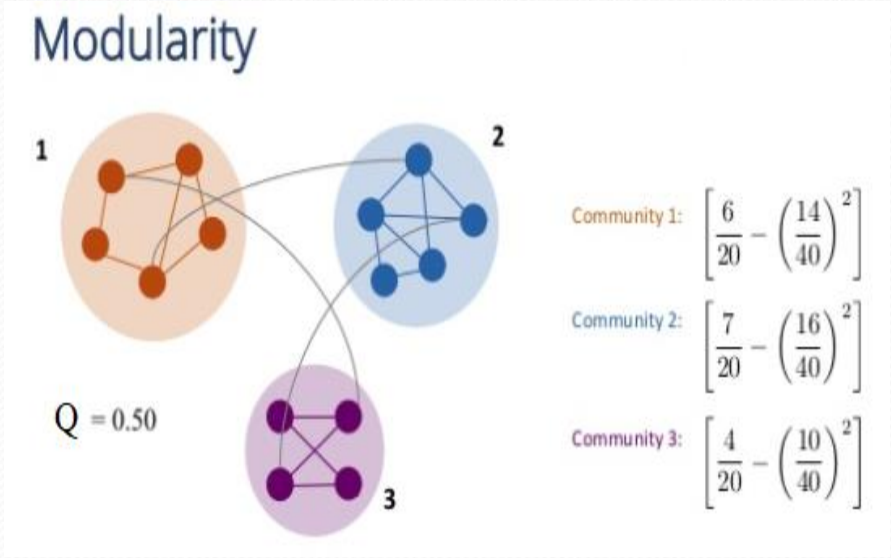
Q1 > Q2 One of the advantages of modularity is that it is independent of the number of clusters that the graph is divided into.

Modularity Maximization

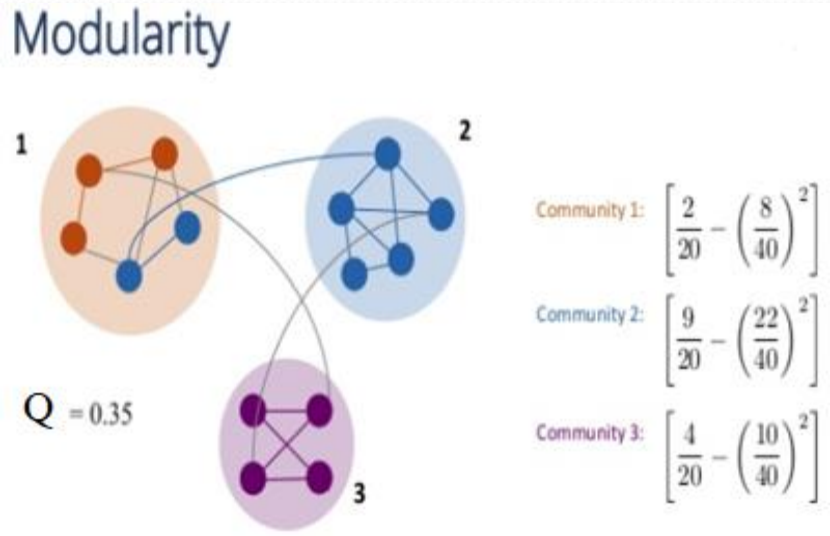
- **Idea:** Random graph not expected to have cluster structure, so possible existence of clusters is revealed by comparison between actual density of edges in a sub graph and a random sub graph.
- Modularity quantifies the community strength by comparing the fraction of edges within the community with such fraction when random connections between the nodes are made.
- Q = $\sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right]$

Where n_c is the total number of communities, L_c is the total no of edges in community c, L is the total no of edges in the graph and K_c is the total no of node degree in community c.

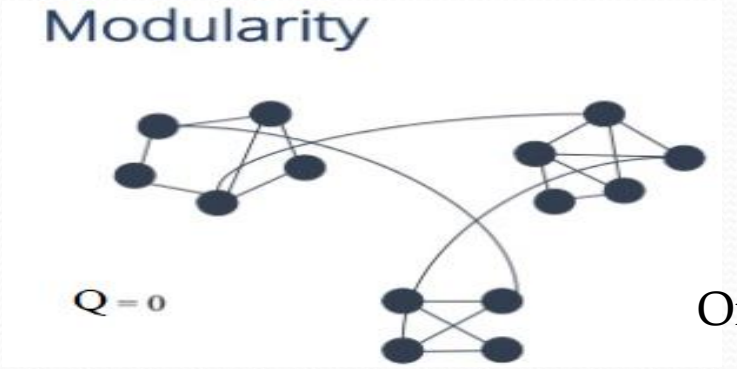
Modularity Maximization : An Example



Optimal Partition



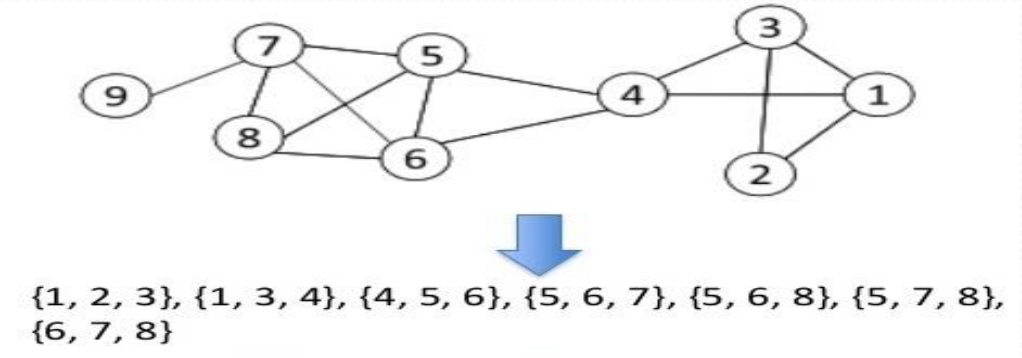
Suboptimal Partition

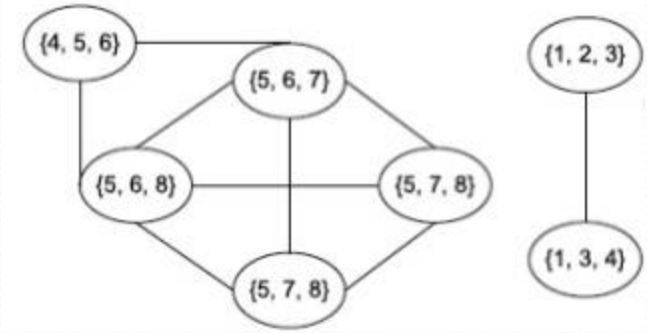


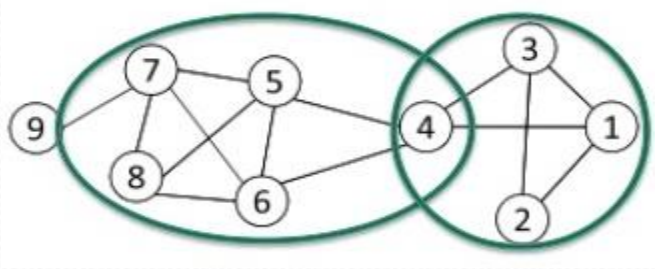
One Community

Overlapping Community Detection

- Community Detection based on Sub-Graph discovery(CPM):
- Find out all cliques of size k in the given network.
- Construct a Clique graph. Two Cliques are adjacent if they share k-1 nodes.
- Each connected components in the Clique graph form a community.







Overlapping Community Detection

- A node in the original graph is called overlapping if links connected to it are put in more than one cluster.
- Using this concept links are partitioned via hierarchical clustering of edge similarity. Given a pair of links eik and ejk incident on a node k, a similarity can be computed via the Jaccard Index defined as

$$S(eik, ejk) = \frac{|Ni \cap Nj|}{|Ni \cup Nj|}$$

- Hierarchical clustering is then used to build a link dendrogram. Cutting this dendrogram at some threshold gives link communities.