Social Media Analytics: Graph Essentials

Lecture:7

Garima Jindal

BITS Pilani
Pilani Campus

Homophily

Movy

locality

conference

people in a park

Homophily in the Society

- Tendency of individuals to associate and bond with similar others
- Similar nodes tend to attract each other, and dissimilar nodes tend to get away from each other
- Causes formation of a community structure in a social network

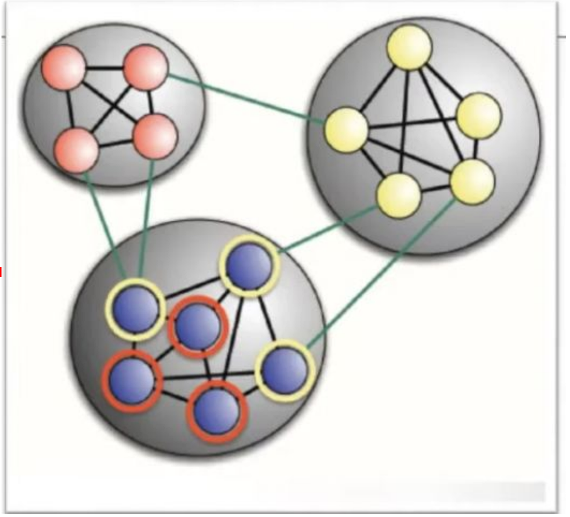
- Homophily occurs against a number of categories:
  - Age
  - Sex and Gender
  - Class: Education, occupation, and Social
  - Religion, Race, and Ethnicity
  - Interests
  - Organizational role, etc.

Communities in a Network

http://bit.ly/3jZIs6O

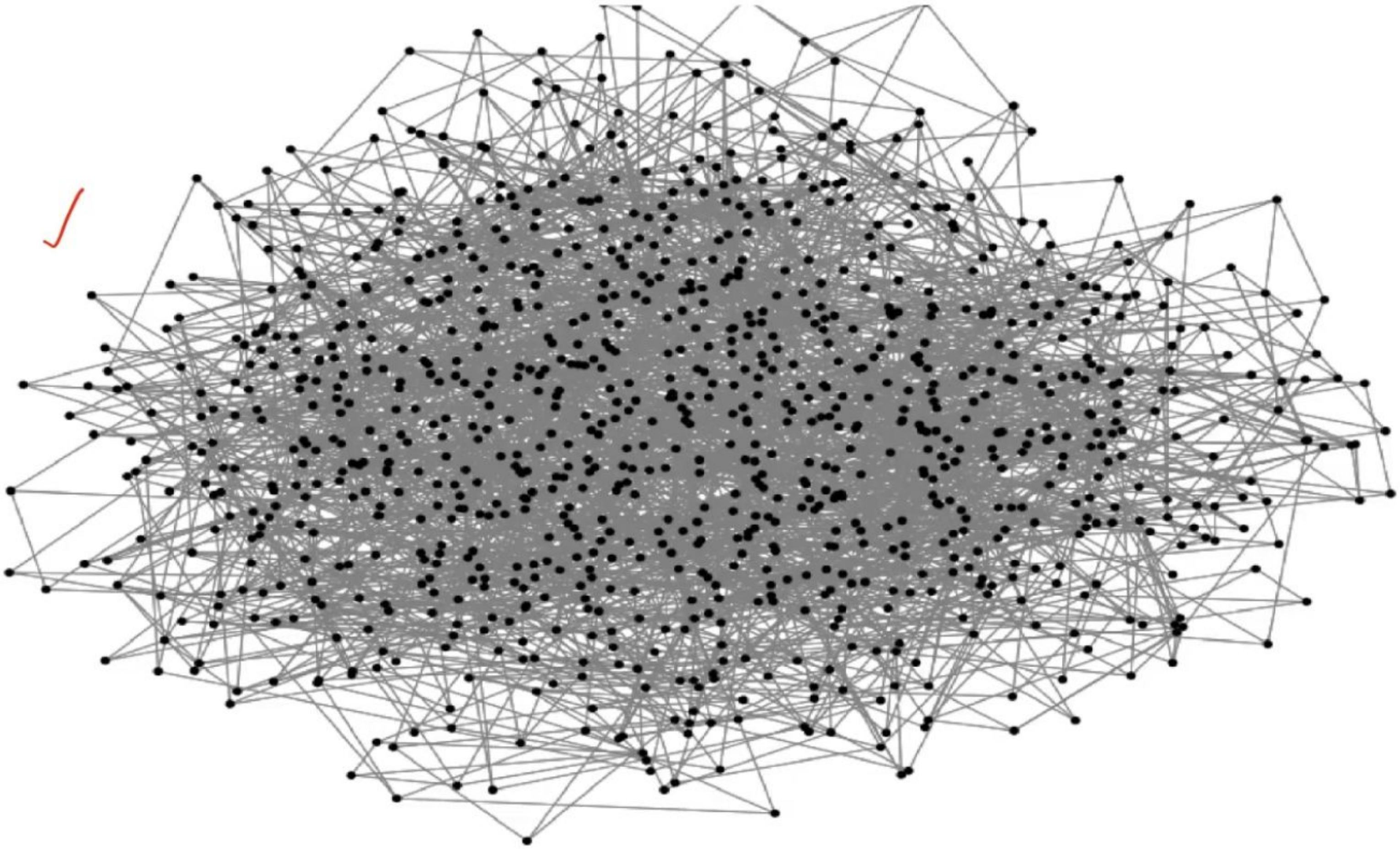
- Identifying communities gives an insight about the inherent network structure
- Community detection is *an ill-defined* problem
  - What we mean by a 'community' is often not concrete
  - Often hard to reliably define a ground-truth annotation for communities
  - No standard measure to asses the performance
- Diverse approaches to the problem depending on how we define a community structure in the network

Community Detection in Networks: Applications

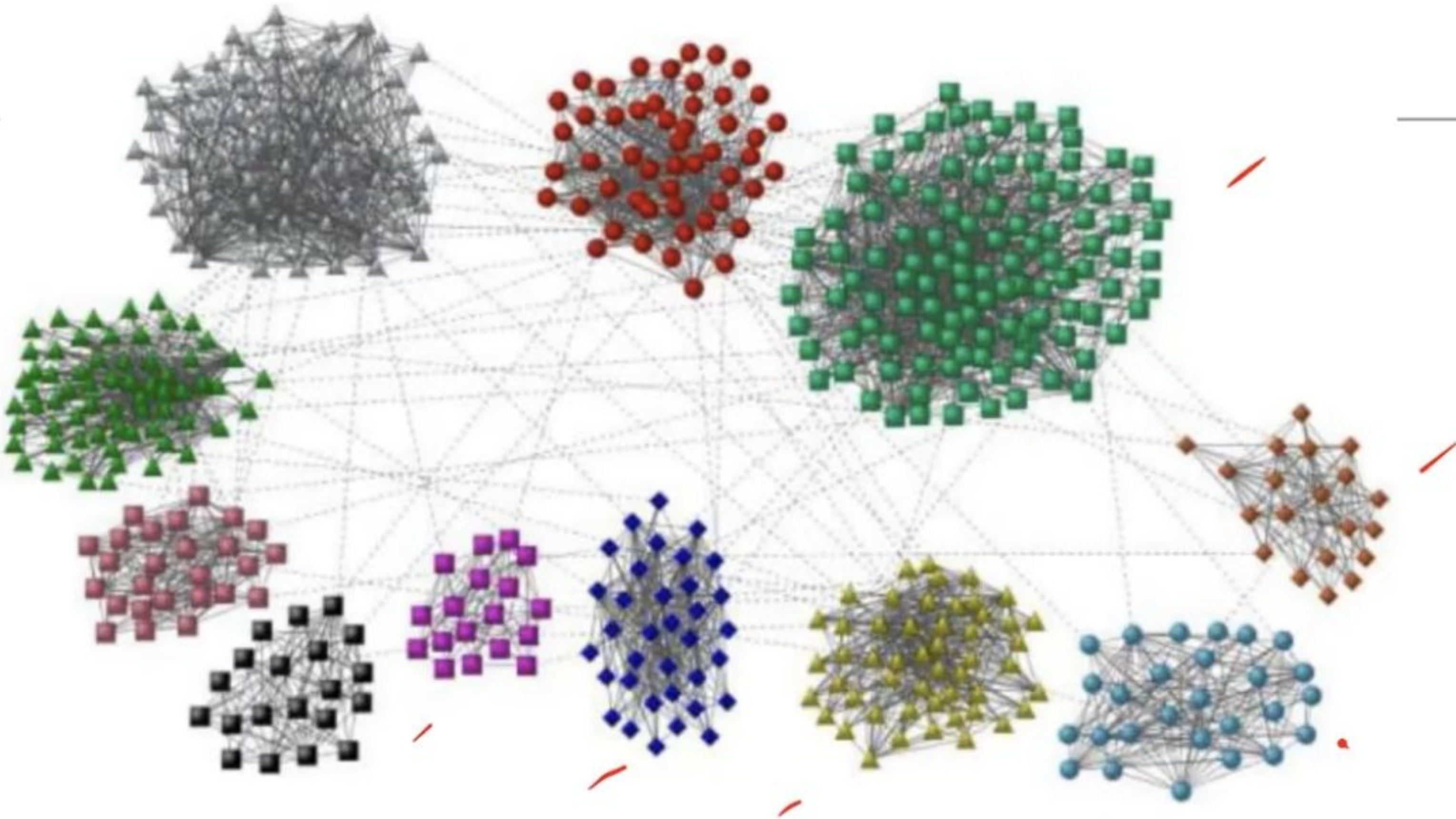


- Performance enhancement of the similarity-based link prediction algorithms
- Improving recommendation quality in Recommender systems by separating like-minded people
- Controlling information diffusion within a network by identifying community memberships
- Designing better marketing strategy by identifying position of the target group within the network
- Restricting epidemic propagation by suitably isolating and immunizing the vulnerable population
- Better anomaly detection in nodes, especially in evolving networks
- Studying evolution of communities
- Applications in criminology and detecting terrorist groups

The Network



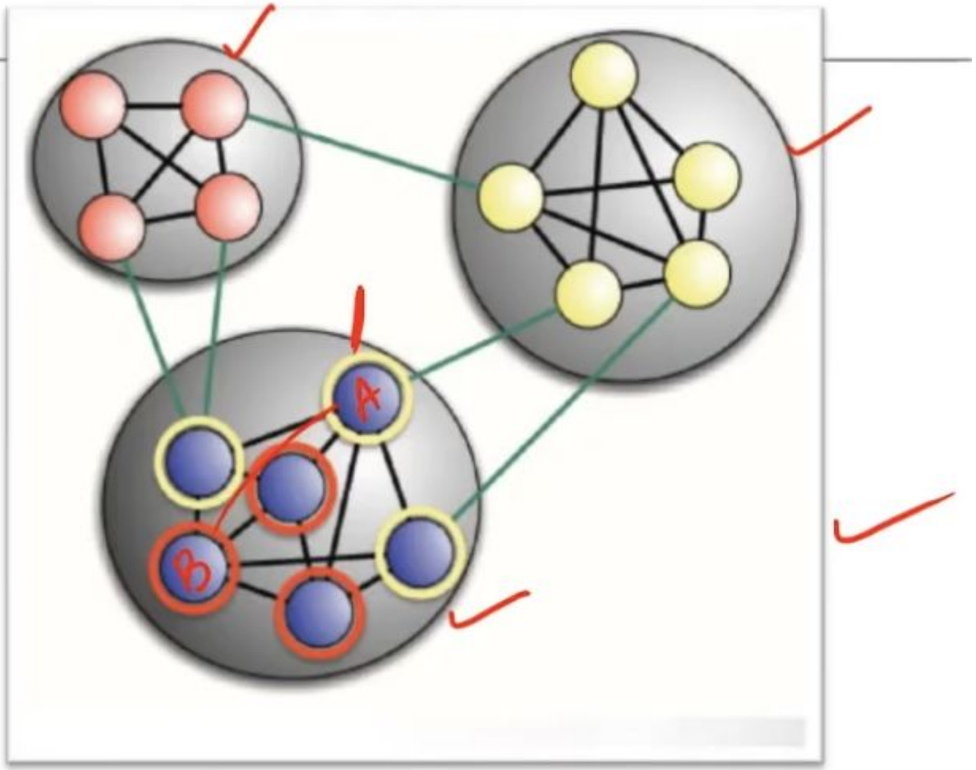
The Community Structure



Community Structure

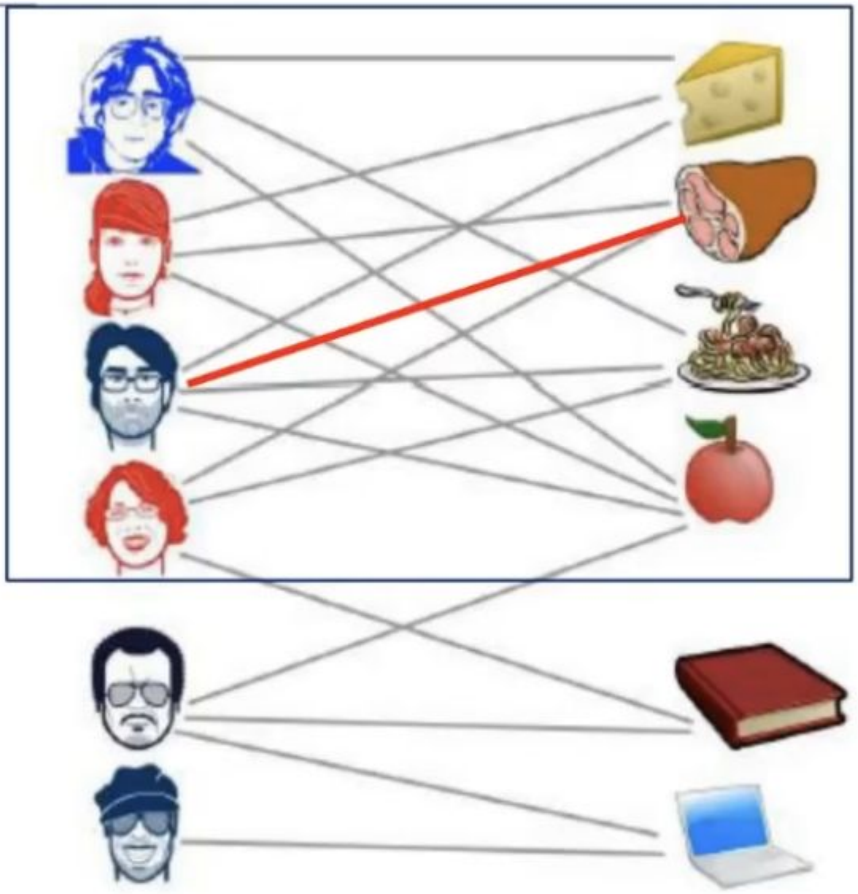
Theoretical reasons

- Organization
- Node features
- Node classification
- Missing links

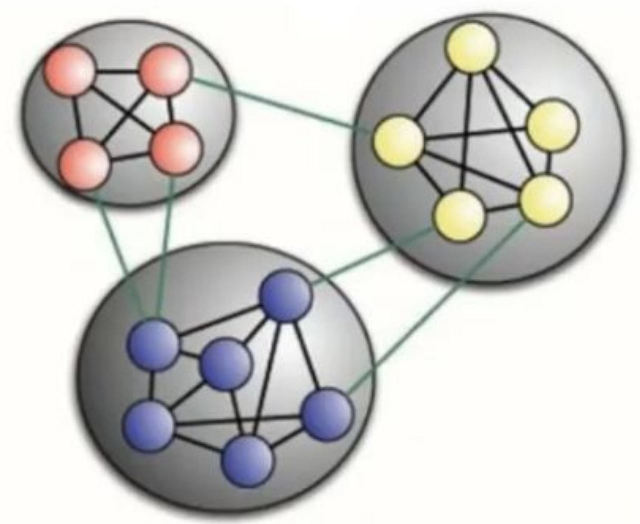


Community Detection

Practical Reasons: Recommendation Systems



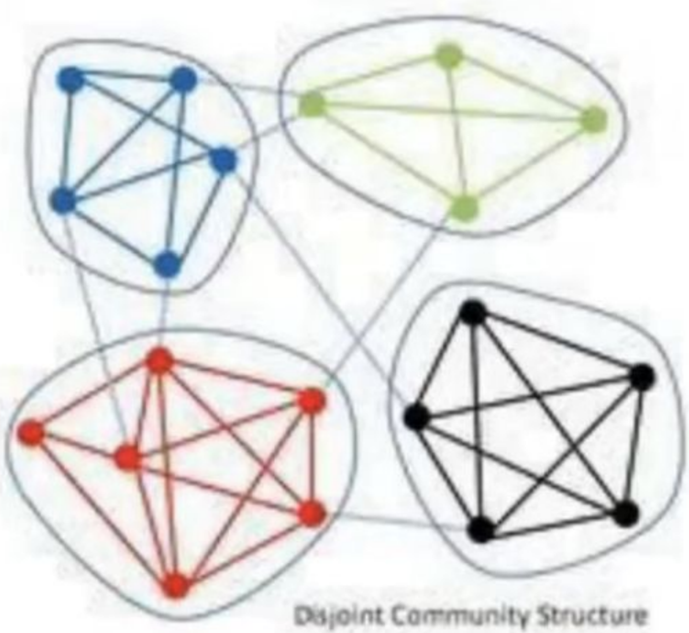
Difficult Problem !!



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1								1		1	1			1	1
2						1		1						1	
3	1											1			1
4				1						1		1			1
5							1	1					1	1	
6		1						1							1
7			1						1			1	1		
8		1				1	1							1	
9			1				1					1	1		
10		1						1						1	1
11							1			1		1	1		
12			1				1		1						1
13				1		1		1				1			
14		1				1	1	1							
15	1			1	1							1			

	6	2	8	14	5	7	13	11	9	12	10	4	1	3	15
6		1	1	1											
2	1		1	1							1				
8	1	1		1		1									
14	1	1	1			1									
5			1	1		1	1			1					
7					1		1	1	1						
13					1	1		1		1					
11						1	1		1	1					
9						1		1		1			1		
12					1		1	1	1		1				
10		1										1	1	1	1
4									1		1		1		1
1								1		1	1			1	1
3										1			1		1
15										1	1	1		1	

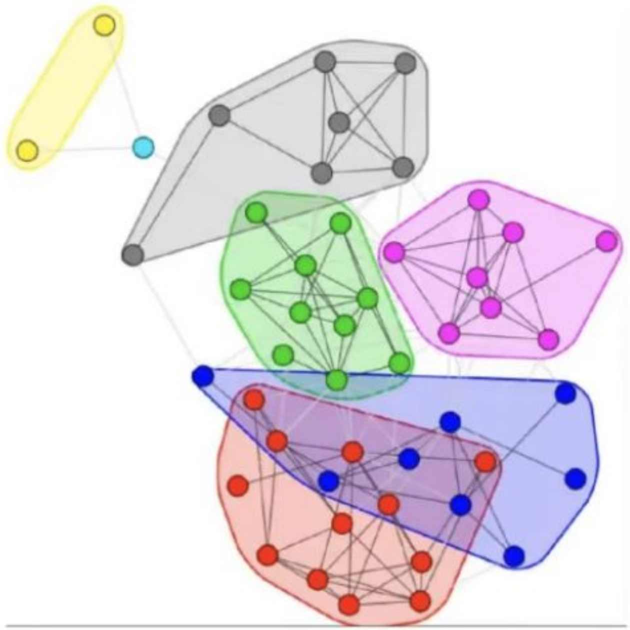
Types of Communities: Disjoint Communities



http://optnetsci.cise.ufl.edu/research/disjoint-overlapping-communities/

- Also referred to as flat communities
- Each node in the network can belong to at most one community
- Differs from disconnected components:
  - nodes in two different communities can still have connecting edges
  - referred to as bridges
- Example: Full-time employees of an organization

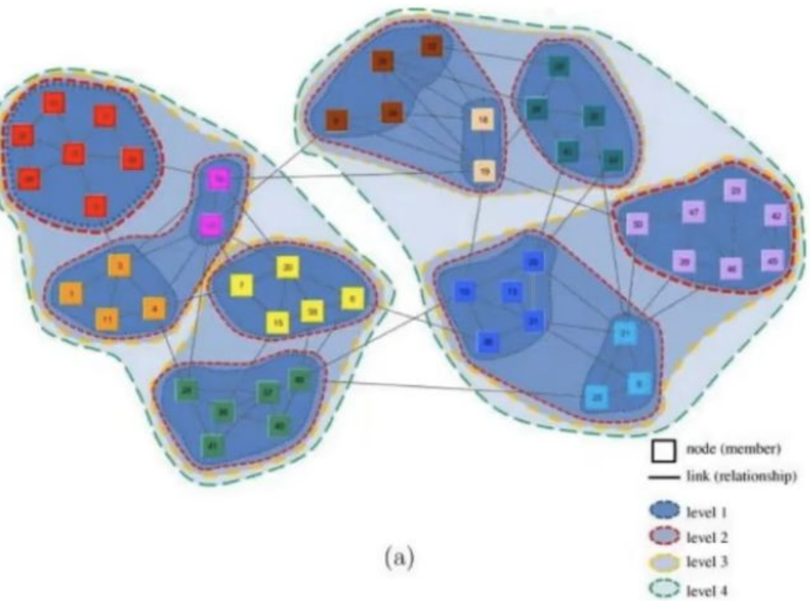
Types of Communities: Overlapping Communities



https://stackoverflow.com/questions/51102350/python-remove-overlapping-communities-in-igraph-plot

- Members can belong to more than one community at a time
- Communities can even share edges
- Realistic and generic community structure
- Harder to find than flat communities
- Example: Various groups in social networks

Types of Communities: Hierarchical Communities



(a)

https://www.sciencedirect.com/science/article/abs/pii/S0020025514011463

- ❑Outcome of merging two or more flat or overlapping communities in a network
- ❑Can be linked to other hierarchical, overlapping, or flat communities
- ❑Example: various city-level communities merged to form a state-level community

Types of Communities: Local Communities

https://www.digitaltrends.com/features/the-history-of-social-networking/

- Shows a community structure from local perspective without focusing on global structure
- Example: citation network formed by research groups inside a university

Community Detection Methods: A Taxonomy

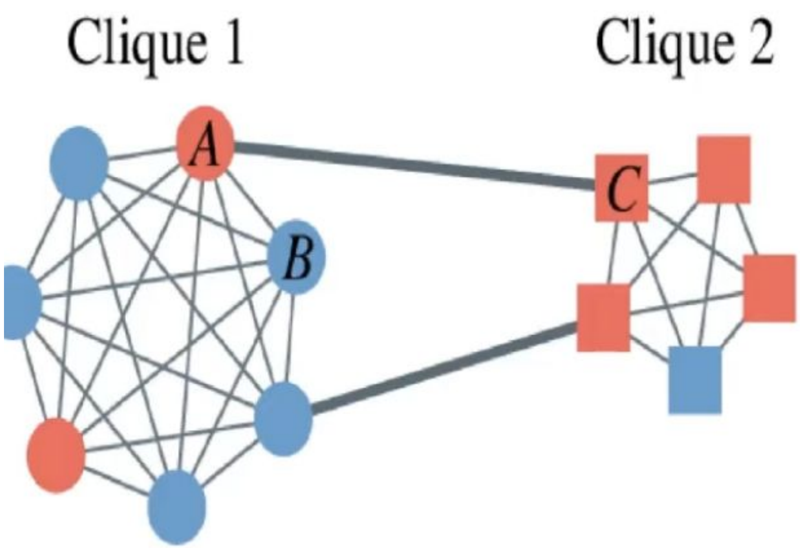
- Community Detection Methods
  - Disjoint Community Detection
    - Node-centric Measures
    - Modularity
    - Information Theoretic Measures
    - Permanence
    - Dynamic Algorithm
  - Overlapping Community Detection
    - Clique Percolation Method
    - Link Partition
    - BigClam
    - Dynamics GenPerm
    - Local Expansion
    - Statistical Inference Methods
  - Local Community Detection

Pull up for precise seeking

Node-centric Community Detection

- Use the property of the nodes to find community structure in the network
- Exploits node-centric features in a number of ways:
  - Complete Mutuality
    - Cliques
  - Reachability of Members
    - K-cliques
    - K-clan
    - K-club
  - Node Degree
    - K-plex
    - K-core

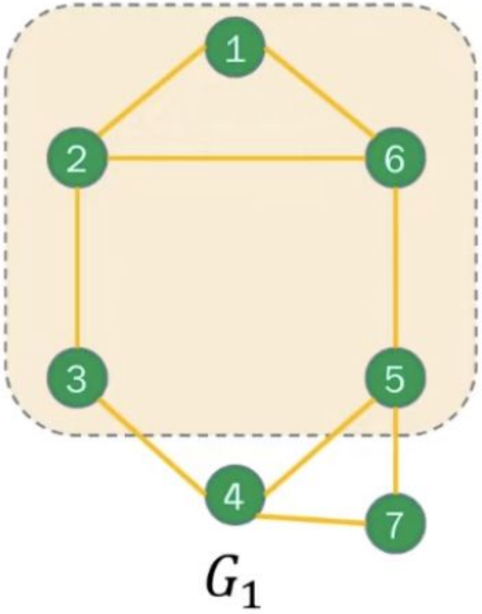
Node-centric Community Detection: Finding Cliques



https://www.researchgate.net/figure/Illustrative-example-of-a-small-two-clique-network-In-this-example-clique-1-is-a_fig1_337025822

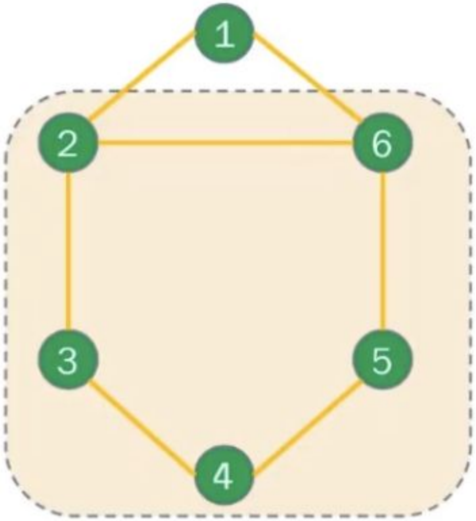
- A subgraph of a graph is a clique if every vertex-pair in the subgraph are adjacent
- Has diameter of 1
- Can be considered as communities
- A couple of problems with this approach
  - Finding cliques from a network is NP-complete
  - Constraints on cliques are too strict a requirement
  - Large cliques are not present in social networks usually

Node-centric Community Detection: K-Cliques



- The maximal subset of vertices of the network such that, for any two nodes belonging to this subset, the shortest distance between them is less than or equal to K
- 1-clique is normal clique
- The nodes {1,2,3,5,6} forms a 2-clique in the network G_1
- 2-cliques are known as known as friend of a friend in social network analysis
- Issue:
  - A node not present in K-clique can contribute in formation of the shortest distance in it!!

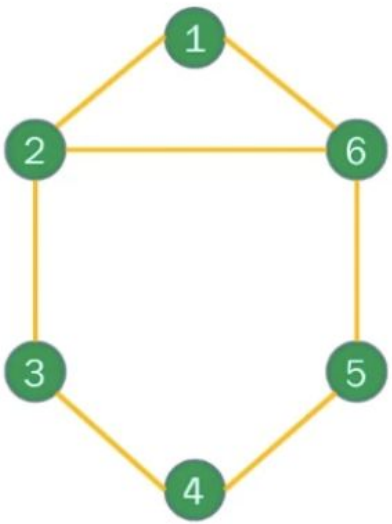
Node-centric Community Detection: K-clan



G_2

- A stricter version of K-clique
- Only the nodes present in the set under inspection are used to create the subgraph in which the distance between any two nodes should be less than or equal to K
- In the network G_1,
- The nodes {1,2,3,4,5,6} forms a 2-clique, but it is not a 2-clan
- The nodes {2,3,4,5,6} forms a 2-clan in the network G_2
- Maximality condition of K-clique also persists in K-clan

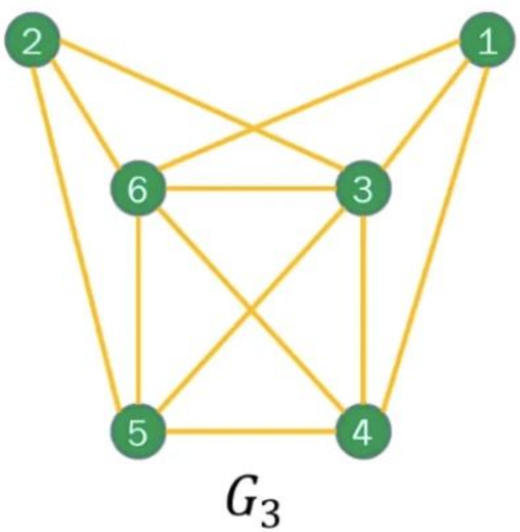
Node-centric Community Detection: K-club



G_2

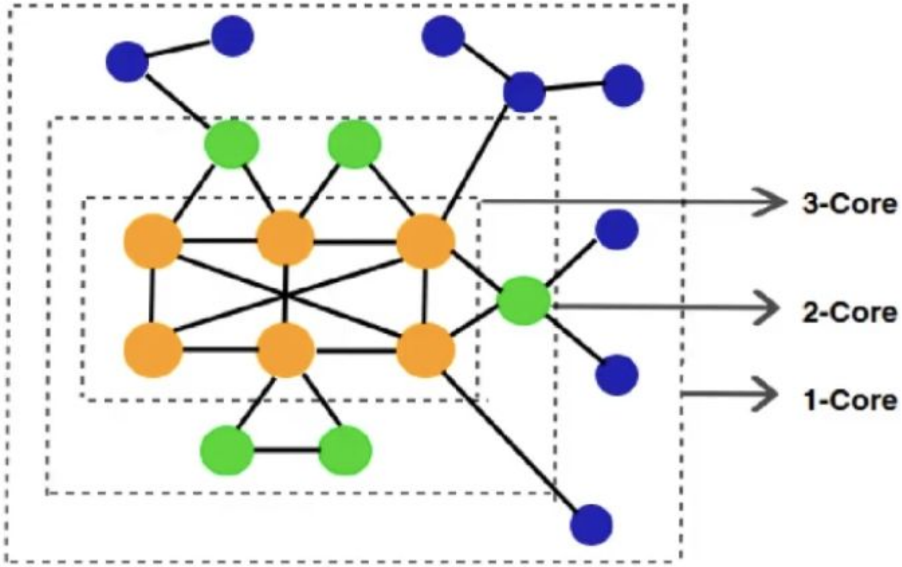
- K-club is a K-clan minus the maximality condition
- {2, 3, 4}, {3, 4, 5}, {4, 5, 6}, {5, 6, 2}, and {6, 2, 3} in G_2 are all 2-clubs
- Every K-clan is a K-club as well as a K-clique
- Challenges:
  - These algorithms are still computationally expensive for large K
  - Deciding appropriate K is difficult

Node-centric Community Detection: K-plex



- A subset of vertices S in a graph is a K-plex if every vertex of the induced subgraph $G[S]$ has degree at least $|S| - K$
- A measure based on the degree of the nodes
- In the network G_3,
- The subset {3,4,5,6} is a 1-plex, i.e., a regular clique
- The subset {1,3,4,5,6} is a 2-plex, but not a 1-plex
- The subset {1,2,3,4,5,6} is a 3-plex, but not a 2-plex

Node-centric Community Detection: K-core



- Another degree-centric measure
- A subgraph G' of a graph G in which each node has degree greater than or equal to K
- K+1 core subgraph can be created from the current K core subgraph by recursively removing nodes of degree K.
- This above should be repeated until there is no node of degree K in the current subgraph.
- Issues:
  - Checking whether a given network is K-core or K-plex is computationally easy
  - Finding maximal K-core/K-plex is NP-complete!!

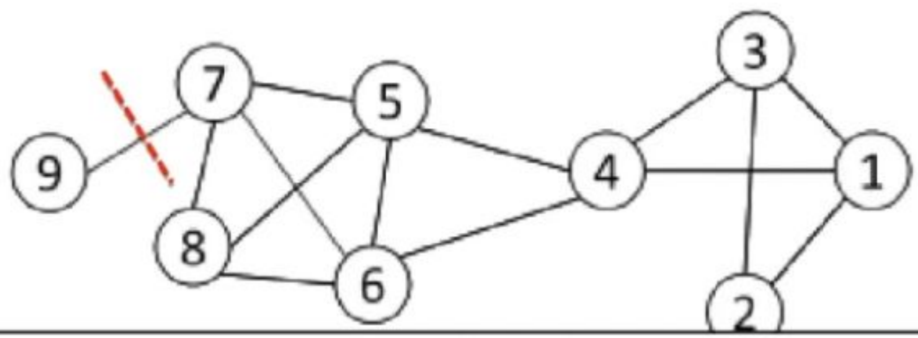
Cut ✓

- Most interactions are within group whereas interactions between groups are few.

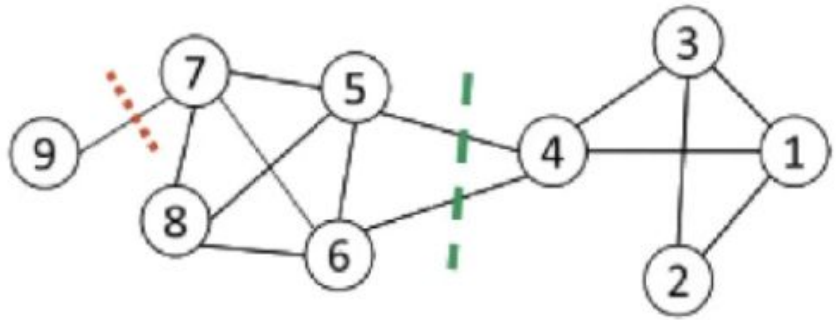
- Community detection → Minimum cut problem

Cut: A partition of vertices of a graph into two disjoint sets

Minimum cut problem: find a graph partition such that the number of edges between the two sets is minimized



Ratio Cut & Normalized Cut



- Minimum cut often returns an **imbalanced** partition, with one set being a singleton, e.g. node 9

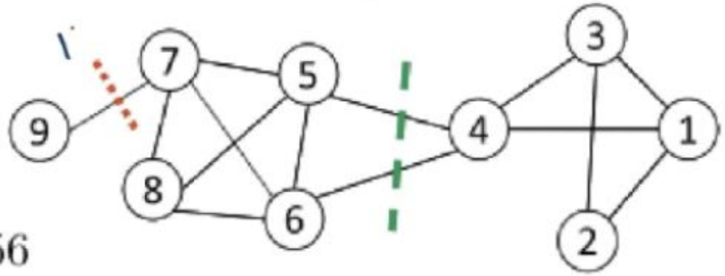
Change the objective function to consider community size

$$\text{Ratio Cut}(\pi) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(C_i, \bar{C}_i)}{|C_i|},$$

$$\text{Normalized Cut}(\pi) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(C_i, \bar{C}_i)}{vol(C_i)}$$

C_i: a community
$|C_i|$: number of nodes in C_i
$vol(C_i)$: sum of degrees in C_i

Ratio Cut & Normalized Cut Example



For partition in red: π_1

$$\text{Ratio Cut}(\pi_1) = \frac{1}{2}\left(\frac{1}{1} + \frac{1}{8}\right) = 9/16 = 0.56$$

$$\text{Normalized Cut}(\pi_1) = \frac{1}{2}\left(\frac{1}{1} + \frac{1}{27}\right) = 14/27 = 0.52$$

For partition in green: π_2

$$\text{Ratio Cut}(\pi_2) = \frac{1}{2}\left(\frac{2}{4} + \frac{2}{5}\right) = 9/20 = 0.45 < \text{Ratio Cut}(\pi_1)$$

$$\text{Normalized Cut}(\pi_2) = \frac{1}{2}\left(\frac{2}{12} + \frac{2}{16}\right) = 7/48 = 0.15 < \text{Normalized Cut}(\pi_1)$$

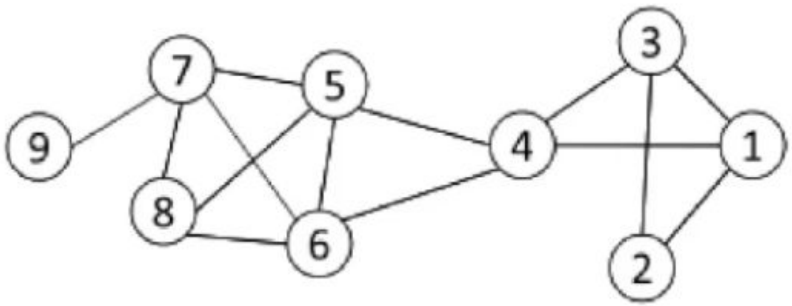
Both ratio cut and normalized cut prefer a **balanced** partition.

Edge Betweenness

Girvan & Newman, PNAS, 2002

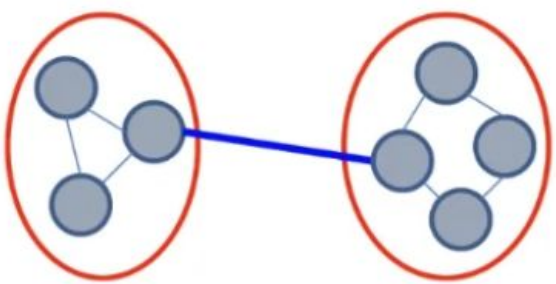
The strength of a tie can be measured by edge betweenness

Edge betweenness: the number of shortest paths that pass along with the edge

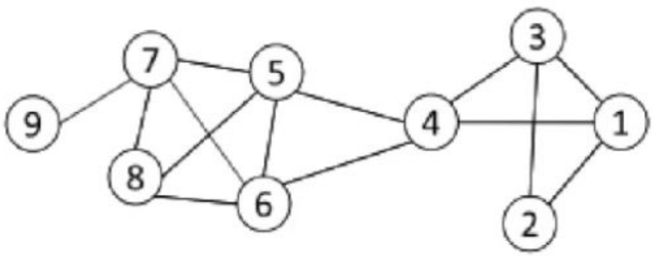


The edge betweenness of **e(1, 2)** is 4 (=6/2 + 1), as all the shortest paths from 2 to {4, 5, 6, 7, 8, 9} have to either pass e(1, 2) or e(2, 3), and e(1,2) is the shortest path between 1 and 2.

- The edge with higher betweenness tends to be the **bridge** between two communities.



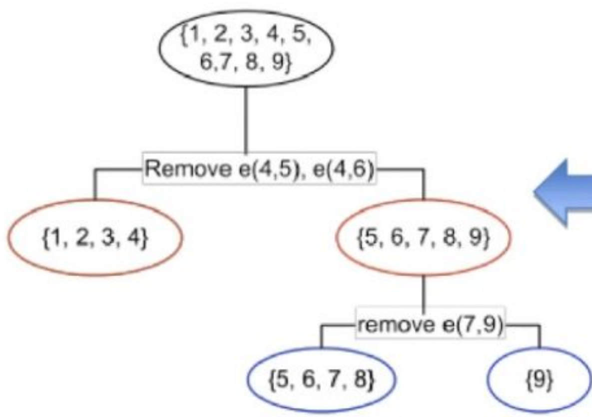
Divisive Clustering based on Edge Betweenness



Initial betweenness value

Table : Edge Betweenness

	1	2	3	4	5	6	7	8	9
1	0	4	1	9	0	0	0	0	0
2	4	0	4	0	0	0	0	0	0
3	1	4	0	9	0	0	0	0	0
4	9	0	9	0	10	10	0	0	0
5	0	0	0	10	0	1	6	3	0
6	0	0	0	10	1	0	6	3	0
7	0	0	0	0	6	6	0	2	8
8	0	0	0	0	3	3	2	0	0
9	0	0	0	0	0	0	8	0	0



After remove e(4,5), the betweenness of e(4, 6) becomes 20, which is the highest;

After remove e(4,6), the edge e(7,9) has the highest betweenness value 4, and should be removed.

Idea: progressively removing edges with the highest betweenness

Community Detection: Modularity

- ❑Node-centric methods discussed so far are not very useful when the network is large
- ❑Modularity comes from the word 'module'
- ❑a network-centric metric to determine the quality of a community structure
- ❑Based on the principle of comparison between
  - ❑the actual number of edges in a subgraph and its expected number of edges
  - ❑the expected number of edges is calculated by assuming a null model
- ❑In the null model,
  - ❑each vertex is randomly connected to other vertices irrespective of the community structure
  - ❑However, some of the structural properties are preserved
  - ❑ One popular structural property is the degree distribution

Questions?

BITS Pilani
Pilani Campus