<u>Introduction</u>

Increased popularity of social networks

Increased number of users

In short amount of time.

Billions/Millions of users are accessing the networks.

At a instant

Thousands of users are getting added.

Generating big volumes of data.

Images

Videos

Text

Sound

People are getting

Added

Generate **BIG DATA**

Blogs
Social Networking Sites
NewsGroups
Chat Rooms

Massive data is available.

Is stored at node level.

Identify Patterns
To get
Knowledge
Predictions
Decision-making.

Objective: Mining such social networks for patterns.

Social Network is represented as a Graph.
Task:

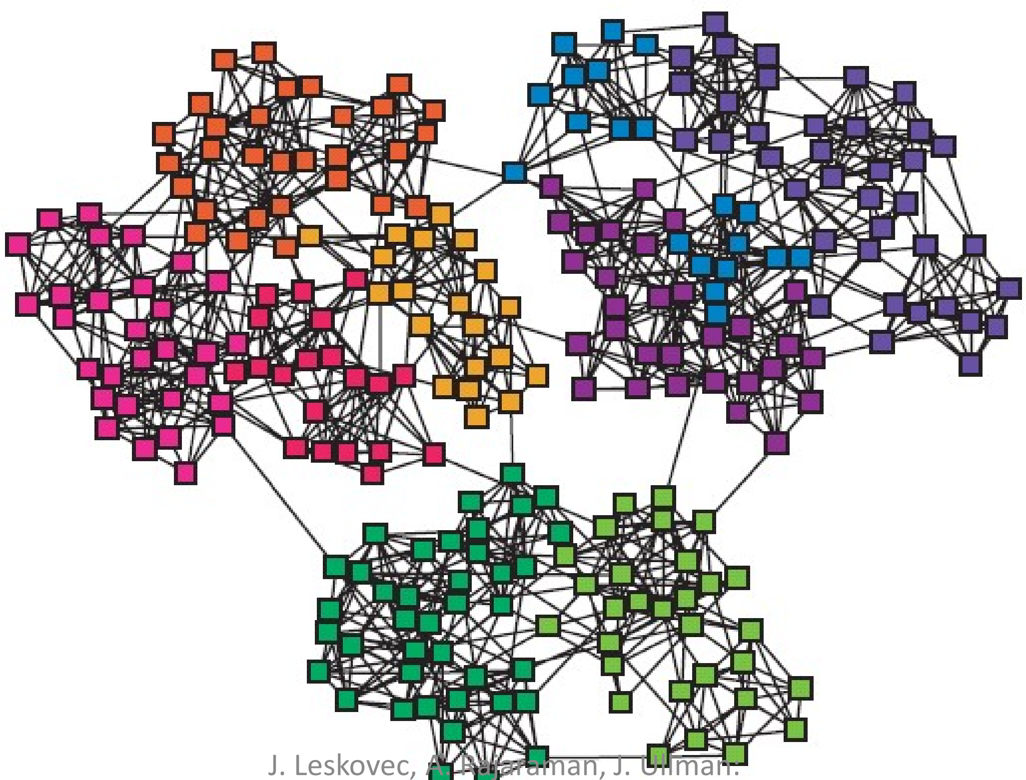
1. Clustering techniques to **identify communities** in a social network scenario.
2. Community detection which identifies **dense subgraphs** from social network graphs.

LEADS TO STANDARD GRAPH ALGORITHMS

1. **SimRank** algorithm provides a way to discover similarities among the nodes of a graph.
2. **Triangle counting** as a way to measure the connectedness of a community.

Networks & Communities

- **We often think of networks being organized into modules, cluster, communities.**
- **Goal is to find densely connected clusters.**



Applications of Social network Mining

1. **Viral Marketing applications :**

    explores how individuals get influenced by the buying habits of others.
    aims to optimise positive word-of-mouth effect among customers.
    identify strong communities and influential nodes.

IT CAN SPEND MONEY ON MAKETING TO AN INDIVIDUAL WHO HAS MANY CONNECTIONS.

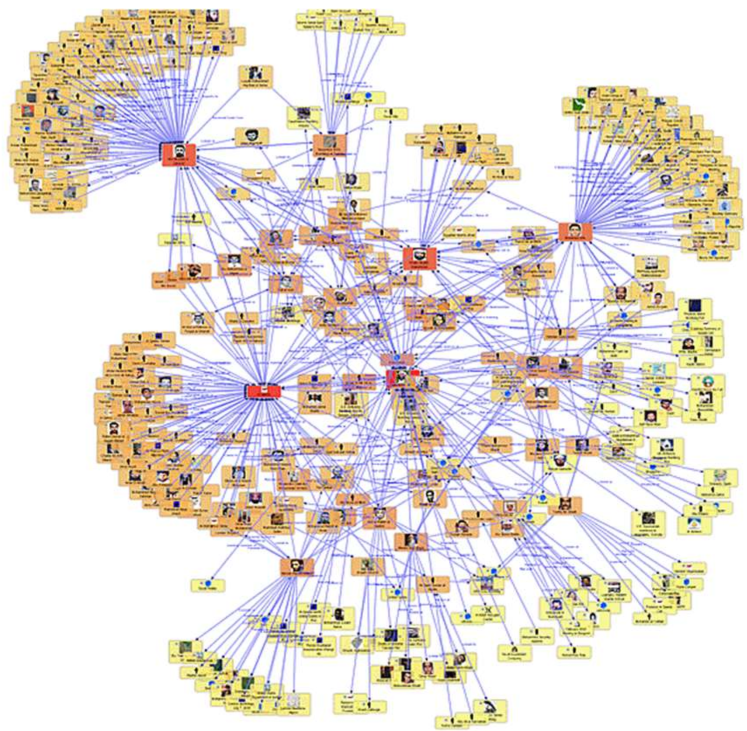
2. **Personalised Recommendation:**

    Grouping together customers who have similar buying profiles.
    Community discovery in mobile ad-hoc networks can enable message routing and posting.

3. **Applications like:**

    Data aggregation and mining
    Network propagation mining
    Network modeling and sampling
    User attribute and behaviour analysis
    Community maintained resource support.
    Location based interaction analysis
    Social sharing and filtering
    Recommendation Systems
    Customer interactions and analysis
    Targeted marketing.

Social Networks as a Graph



Social Network === Large Graph

Node = Object [represent one person or
group of persons
or Organizations or document or
computers]

Links = relationships / interactions between People, groups, organisations, computers, info/knowledge processing entities.

LinkedIn

User profiles: Registered User

Connections: real-world professional
relationship between people.

FIND A PERSON

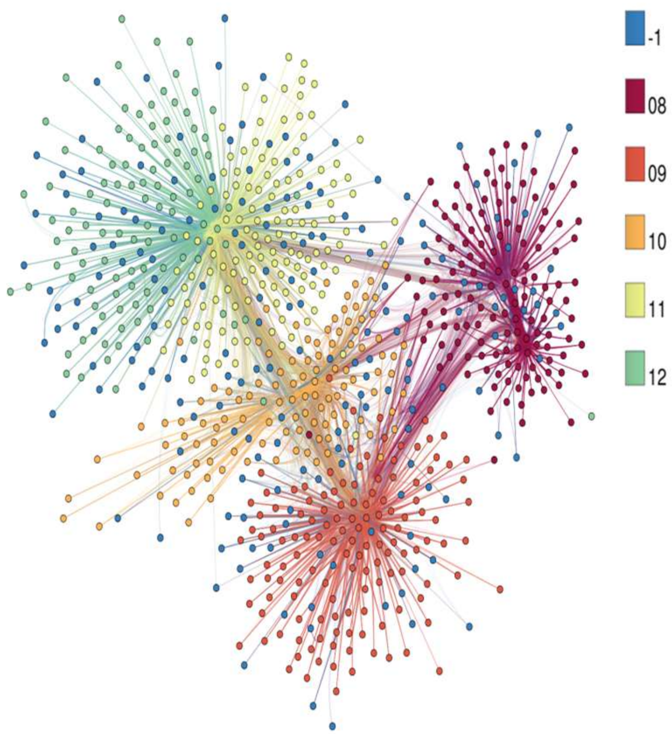
Finding a path in the network.

Company --- Targeted Advertising

Wants to find out most influential person due to which others will get influenced.

IDENTIFY THE NODES WITH HIGH OUT-DEGREE IN THE GRAPH REPRESENTING SOCIAL NETWORK.

Social Networks as a Graph



- <u>**Heterogeneous and Multi-relational dataset.**</u>

  [standard model of graph : node sets are same type]
- Nodes and Edges can have attributes.
- Objects may have class labels.
- **Facebook** : connect entities thru relationship called as **Friends**
- **LinkedIn** :connect entities thru relationship called as **Endorse**
- Relationship need not be yes or no in some SN
- Relationship can be a degree.

[Degree is represented by labeling edges.
Edges can be one-directional/bi directional/need not be binary.]

E.g. degree of endorsement of a skill as novice, expert.. It can also be a real number.

- <u>SN have a property called as **non-randomness** called as **locality.**</u>

Locality = property of SN

<u>Nodes and edges tend to cluster in communities.</u>

Difficult to formalize.

<u>**Relationships tend to cluster.**</u>

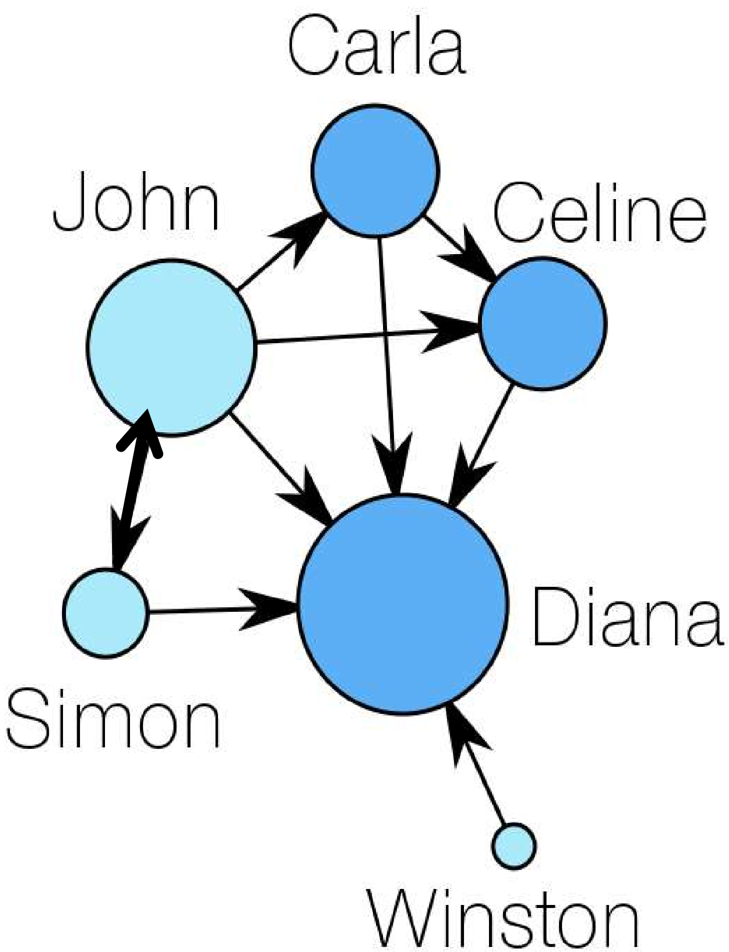
If A is related to B and C

Higher probability than average the B and C are related.

<u>**Most relationship in real world tend to cluster around a small set of individuals.**</u>

Follow Relationship in Twitter.

John follows Carla, Celine, Diana and Simon
Diana follows nobody.
John and Simon follow each other.



Nodes

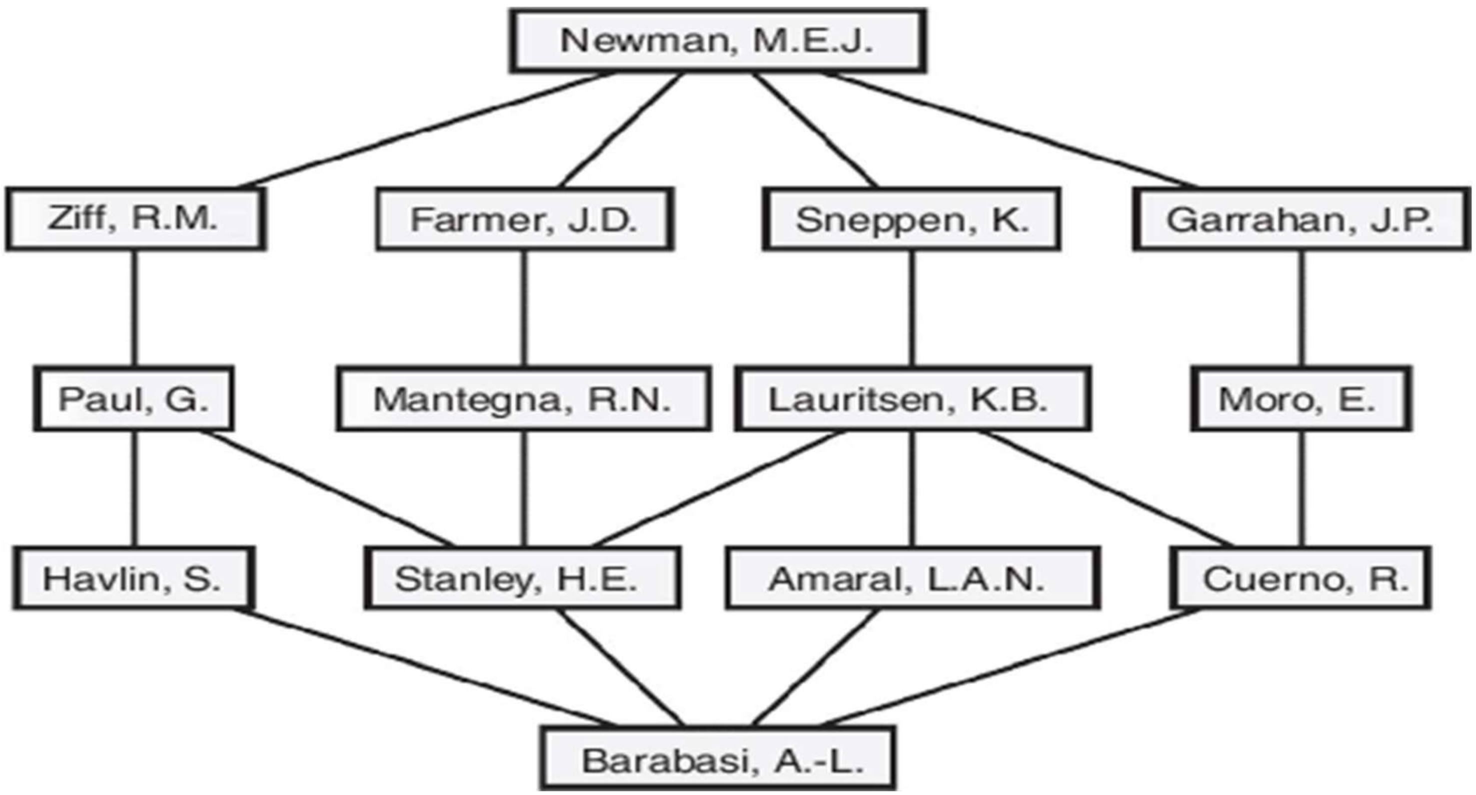
Id,Label,Attribute
1,John,1
2,Carla,2
3,Simon,1
4,Celine,2
5,Winston,1
6,Diana,2

Edges

Source,Target
1,2
1,3
1,4
1,6
2,4
2,6
3,6
4,6
5,6

CO-Authorship relationship in DBLP

Edge can be labelled via number where these authors have co-authored together.



Types of Social Networks

http://snap.stanford.edu/data //large list of social network examples.

1. **Collaboration graph:** displays interactions among entities.
   Co-Authorships among scientists
   Co-Appearance in movies by actors and actresses
   WikiPedia editors updated same article.
   NBA graph in sports indicating two player played together in a team

2. **Who-Talks-to-Whom Graphs:** Communication network.
   Microsoft's instant messenger (IM) graph.
   Edges between A and B is A talks to B
   Enron Email communication network
   nodes = email addresses j to j directed edge indicates I sent at least 1 email to j
   Phone conversation records.
   Call graphs: node = phone number and edge is the conversation between two numbers

Types of Social Networks

3. **Information Linkage Graphs:**

    Snapshot of WEB

        Nodes = web pages and Directed Edge = link from one page to other

        Millions of personal pages on social networking sites linked to other web pages of their interest.

    Product co-purchasing products

        Nodes = products Edges= commonly co-purchased products.

    Internet Networks

    Road Networks

    WikiPedia/Flickr/Reddit

4. **Heterogeneous Social Networks:** Heterogeneous nodes and links.

    Multimode network and relationships

    Product nodes ---------- customer nodes

    Amazon

        Customer , Product, Distributor

    Movie

        k-partite Graph: k disjoint set of nodes with no edges between the nodes of same set.

        movie, role, studio, distributor, genre, award, country

Clustering of Social Graphs

Discovering Communities is a fundamental requirement in SN apps.

- Community allows us to discover groups of interacting objects and relations between them.
- Community = collection of individuals with dense relationship patterns within a group and sparse link outside the group.
- Target marketing can benefit by identifying clusters of shoppers and targeting a campaign wholly customized for them.
- Summary of network, easy to visualize and understand.
- Simple method to find communities is use clustering technique on social graph.
- K-means and Hierarchical clustering can not be used in extended social graphs.
- Popular Graph Clustering Technique will be Girvan-Newman Algorithm.

Applying Standard Clustering techniques

1. **Cluster similar data points in a social network graph:**

need to find similarity measure.

if it is a labeled graph then that value = similarity measure.

e.g. DBLP: Edge labeled as number of papers co-authored.

But most of the graphs are unlabeled.

1. if edge is present then SimM = 1 else SimM=0 [or dist=1 or 1.5]

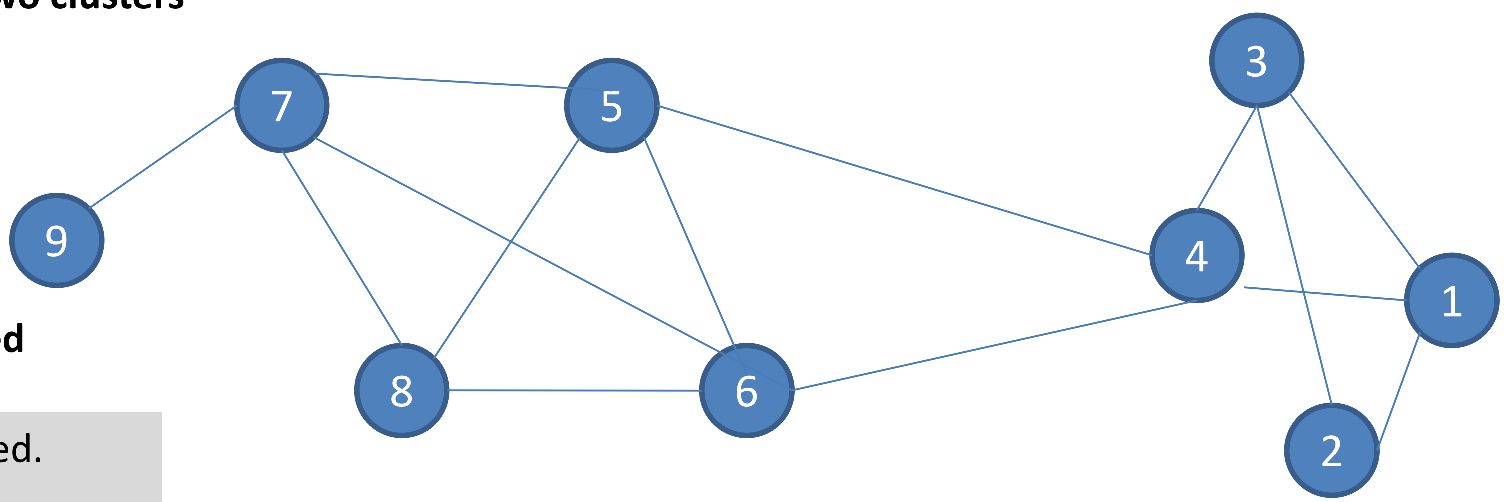
problem = it does not satisfy triangular in equality property.

i.e. dist(AB) +dist(BC) < dist(AC)

because the notion of distance does not have any meaning in social network.

{ 1,2,3,4}
{5,6,7,8} are two clusters

{1,2,3}
{4,5,6}
{5,6,7}
{6,7,8}
Clusters formed



9 is left isolated.
If **hierarchical** clustering is used then {1-8} may have a single large cluster.

K- means : would land into wrong clusters.
If initial centroids are 4 and 8 then 5 and 6 are equidistant
So they would be wrongly clustered instead they should be clustered to 8

Betweenness Measure of Graph Clustering

Extract dense subgraph from the social graph. (DISJOINT COMMUNITIES)

One dense subgraph may be connected to other dense subgraph by minimum set of edges. identify these edges and remove them to get dense subgraphs.

Edge betweenness: Edges that are likely to connect different dense regions of graph have higher betweenness scores than other edges in the communities.

The edge e in the graph, edge betweenness of e is defined as the number of shortest paths between all the node pairs (vi, vj) in the graph such that the shortest path between vi and vj passes through e.

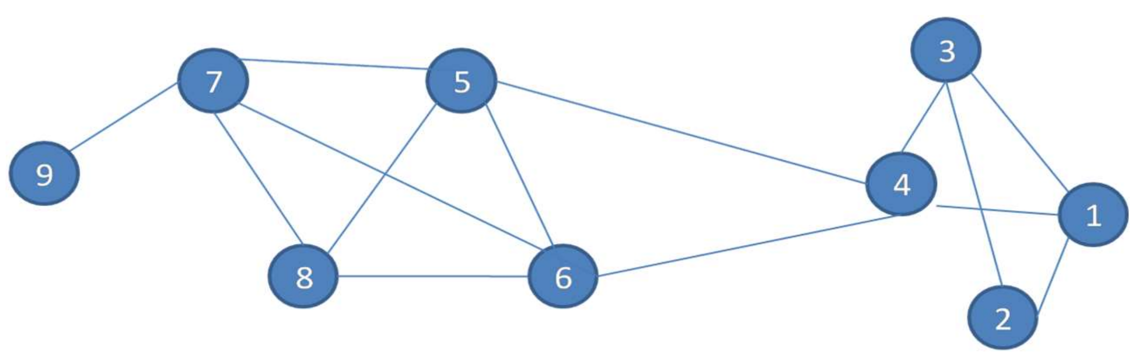
EB(1,2) = 4 (=6/2 + 1) from 2 shortest path to all the nodes 4,5,6,7,8,9 and e(1,2) = shortest path

EB(4,5) = from 9,7,8 to 4,3,2,1 either thru e(4,5) or e(4,6) = 12/2 = 6 + 5 to 1,2,3,4 = 4
thus 6 + 4 = 10

EB(4,6) = 10 EB(5,7) = 6 and so on.

1,2,3,4,5,6,7,8,9}
Remove {4,5} and {4,6}

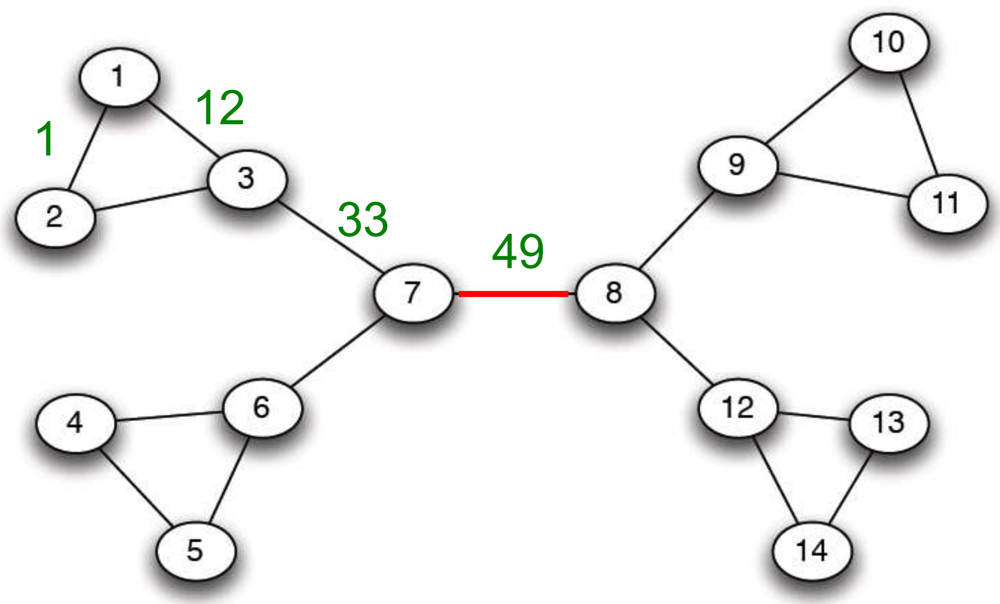
{1,2,3,4} and { 5,6,7,8,9}
remove (4,5)
{5,6,7,8} and {9}



Girvan & Newman: betweenness clustering

- Algorithm
  - compute the betweenness of all edges
  - while (betweenness of any edge > threshold):
    - remove edge with highest betweenness
    - recalculate betweenness
- Betweenness needs to be recalculated at each step
  - removal of an edge can impact the betweenness of another edge
  - very expensive: all pairs shortest path – $O(N^3)$
  - may need to repeat up to N times
  - does not scale to more than a few hundred nodes, even with the fastest algorithms

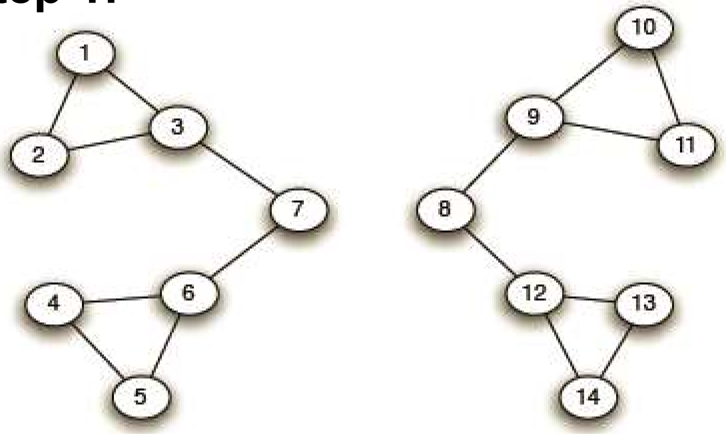
Girvan-Newman: Example



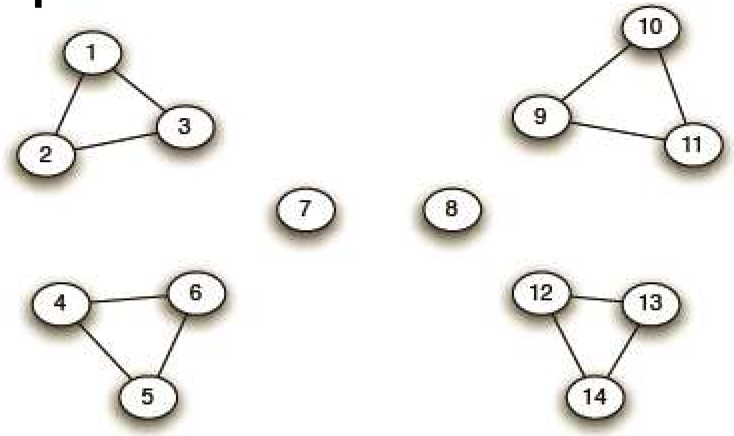
Need to re-compute betweenness at every step

Girvan-Newman: Example

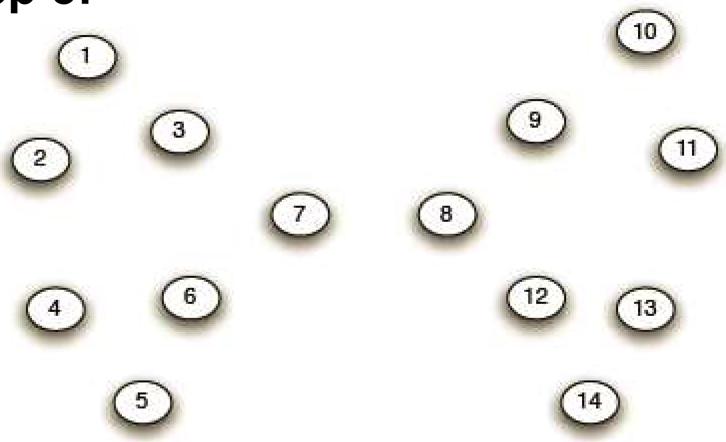
Step 1:



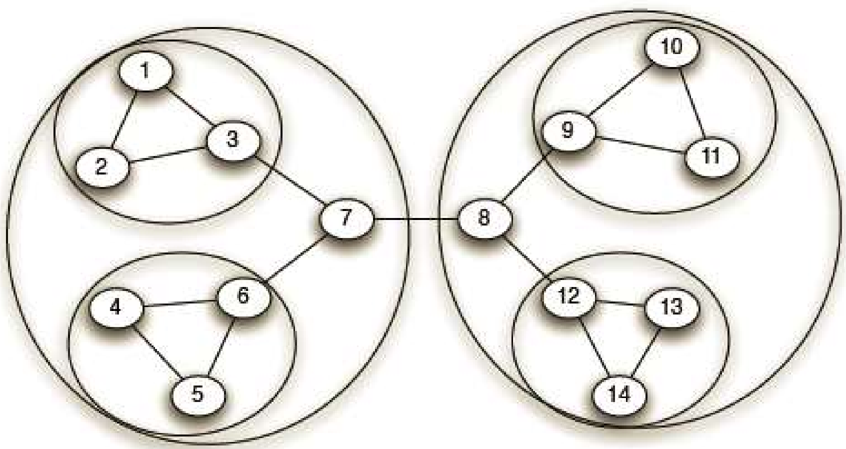
Step 2:



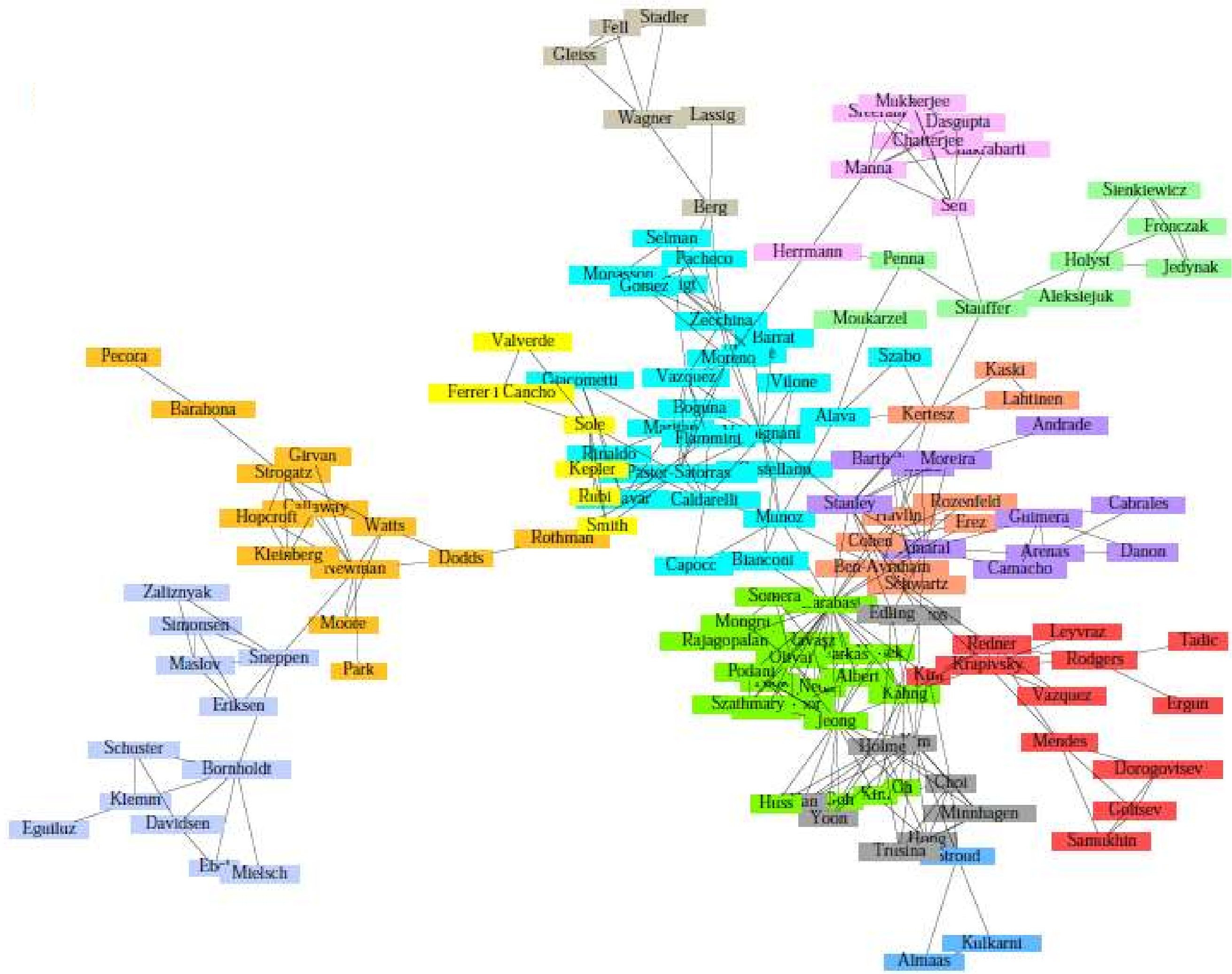
Step 3:



Hierarchical network decomposition:

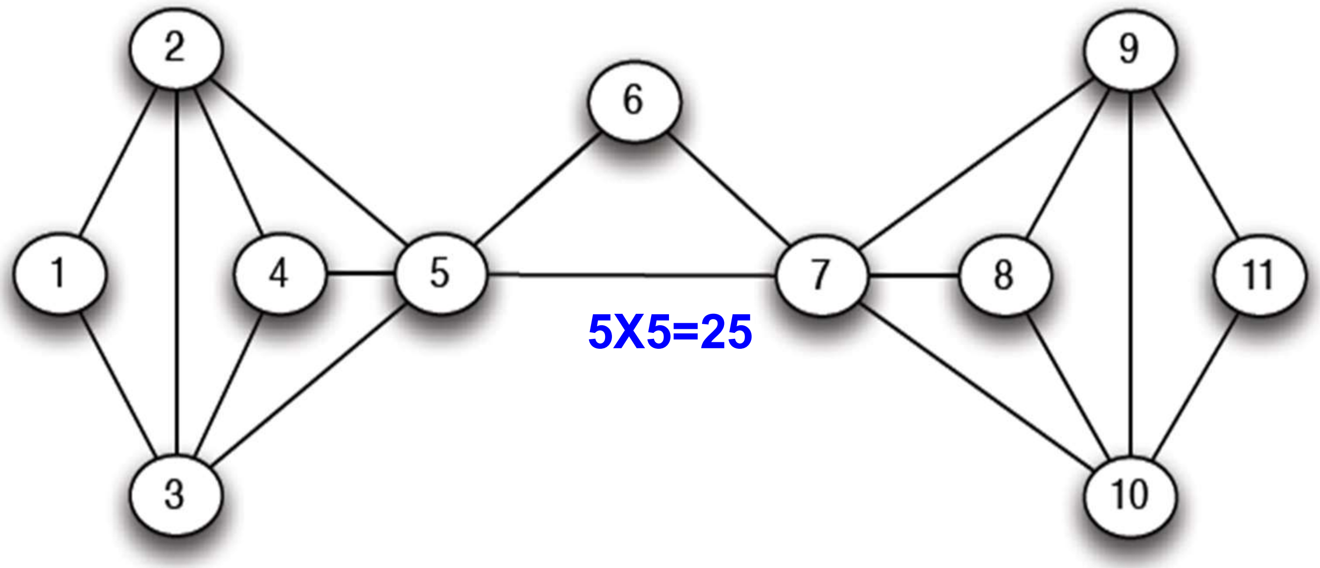


Girvan-Newman: Results

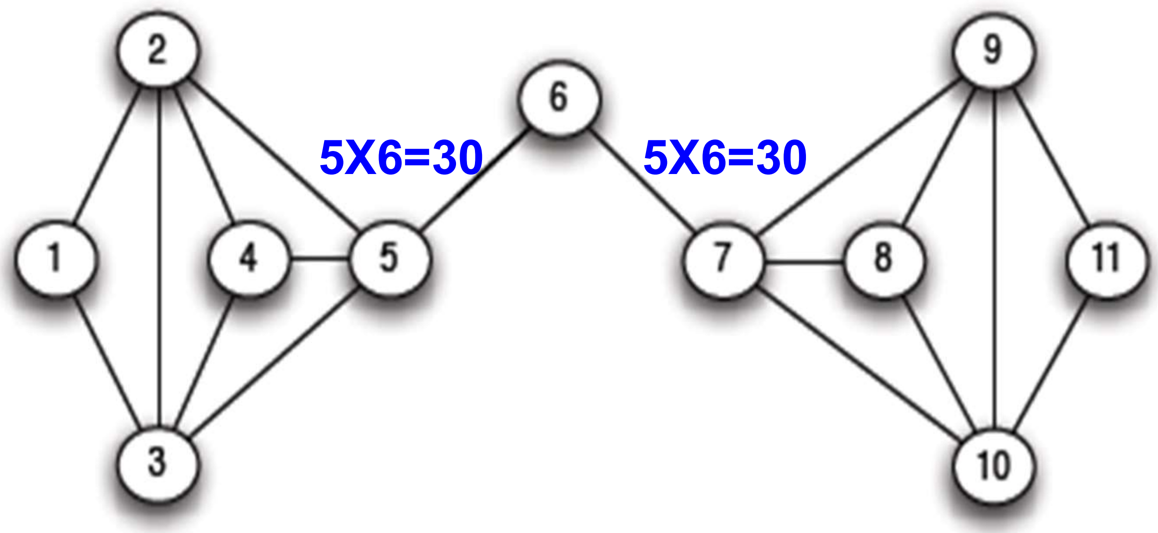


Communities in physics collaborations

Another example

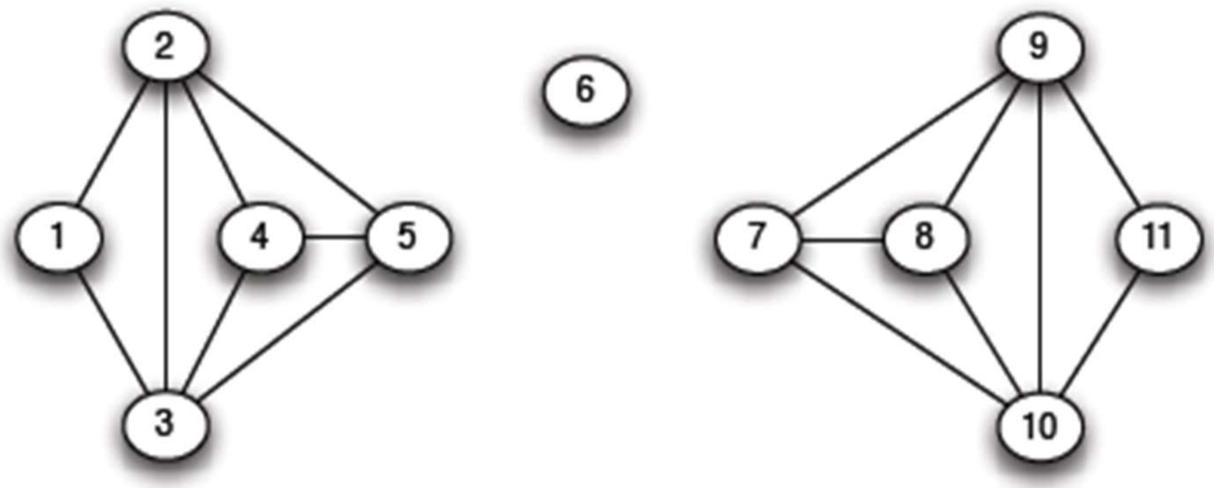


Another example



(a) *Step 1*

Another example



(b) *Step 2*

Direct discovery of communities in a Social Graph.

Classification of Communities:

- Disjoint communities
  - Girvan Nirmann algorithm

- Overlapped communities
  - In SN it is possible that the individual may be a part of different communities at a time.
  - Twitter and facebook
  - Clique Percolation Method (CPM) – find clique in a graph.

Clique Percolation Method (CPM)

Eugene Lim

Contents

- What is CPM?
- Algorithm
- Analysis
- Conclusion

What is CPM?

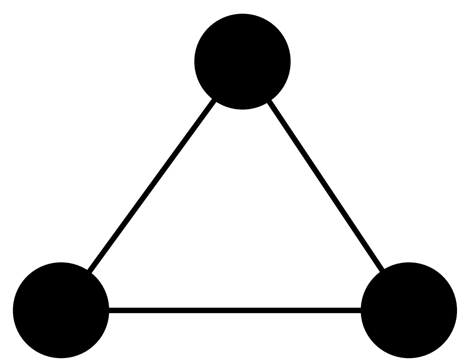
- Method to find **overlapping** communities
- Finding all cliques of a given size = NP-hard problem
- Based on concept:
  - internal edges of community likely to form cliques
  - Intercommunity edges unlikely to form cliques

Clique

- Clique: Complete graph
- k-clique: Complete graph with k vertices

Clique

- Clique: Complete graph
- k-clique: Complete graph with k vertices

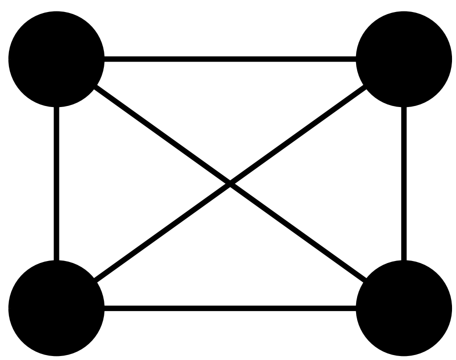


3-clique

Clique

- Clique: Complete graph

- k-clique: Complete graph with k vertices

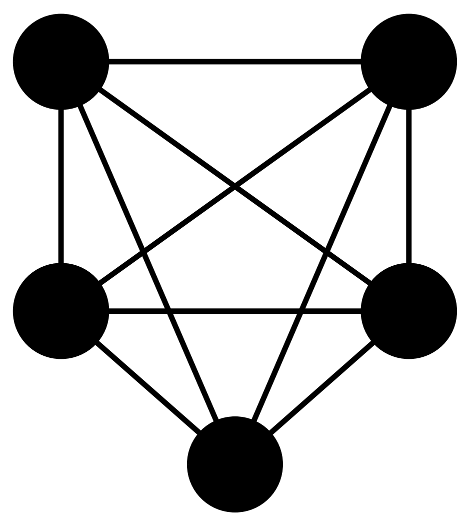


4-clique

Clique

- Clique: Complete graph

- k-clique: Complete graph with k vertices



5-clique

k-Clique Communities

- **Adjacent k-cliques**

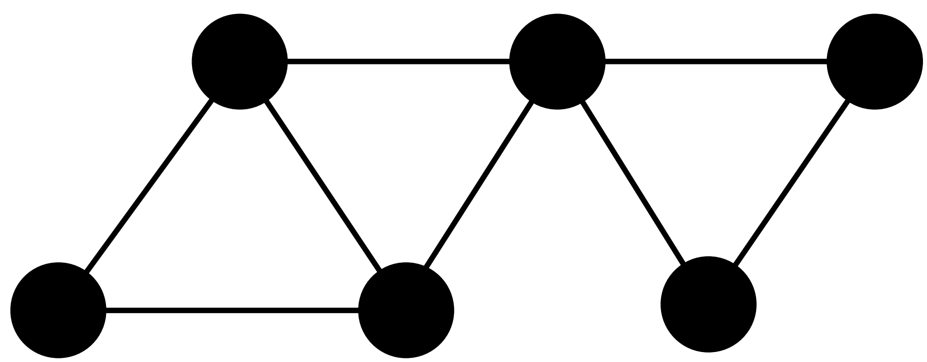
  Two k-cliques are adjacent when they share **k-1** nodes

k-Clique Communities

- **Adjacent k-cliques**

  Two k-cliques are adjacent when they share **k-1** nodes

k = 3

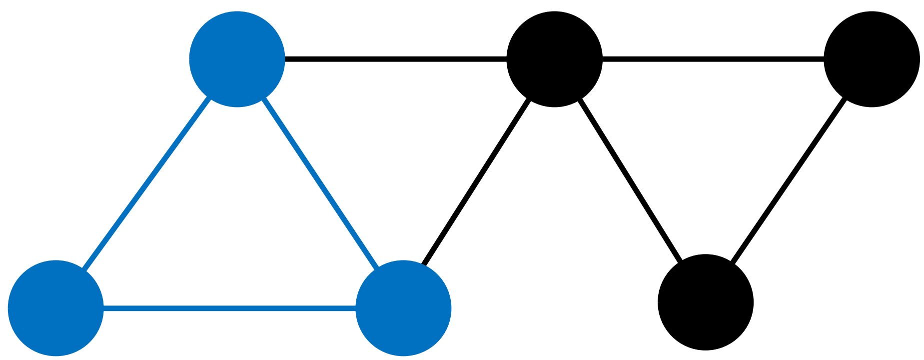


k-Clique Communities

- **Adjacent k-cliques**

Two k-cliques are adjacent when they share **k-1** nodes

k = 3

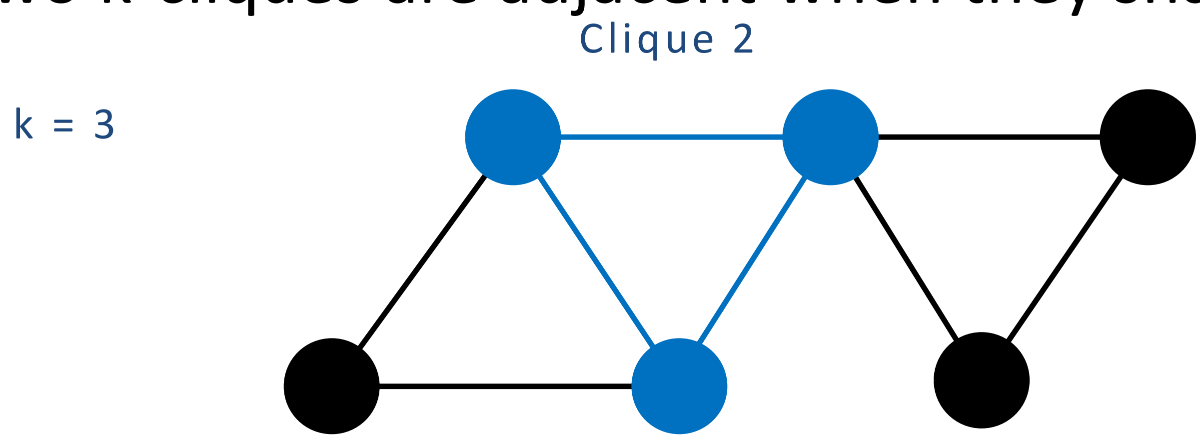


Clique 1

k-Clique Communities

- **Adjacent k-cliques**

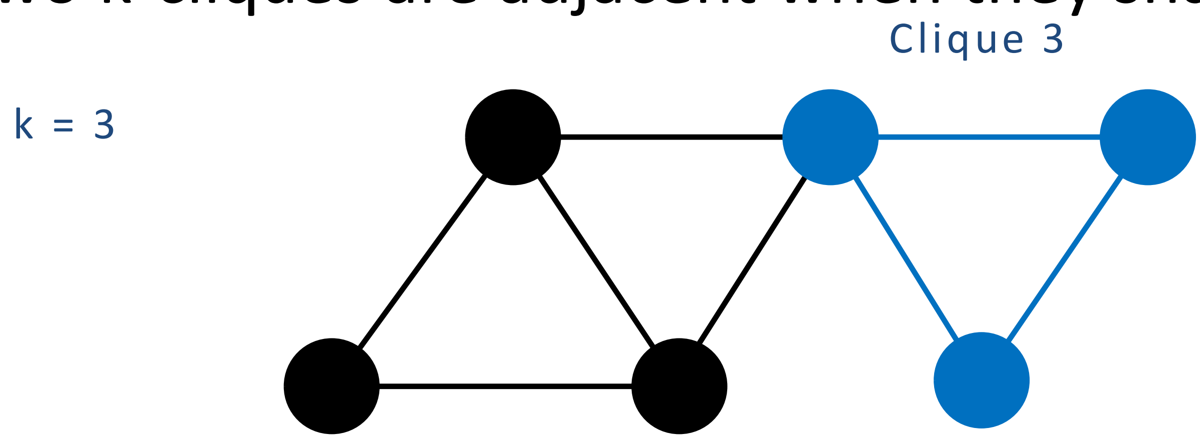
Two k-cliques are adjacent when they share **k-1** nodes



k-Clique Communities

- **Adjacent k-cliques**

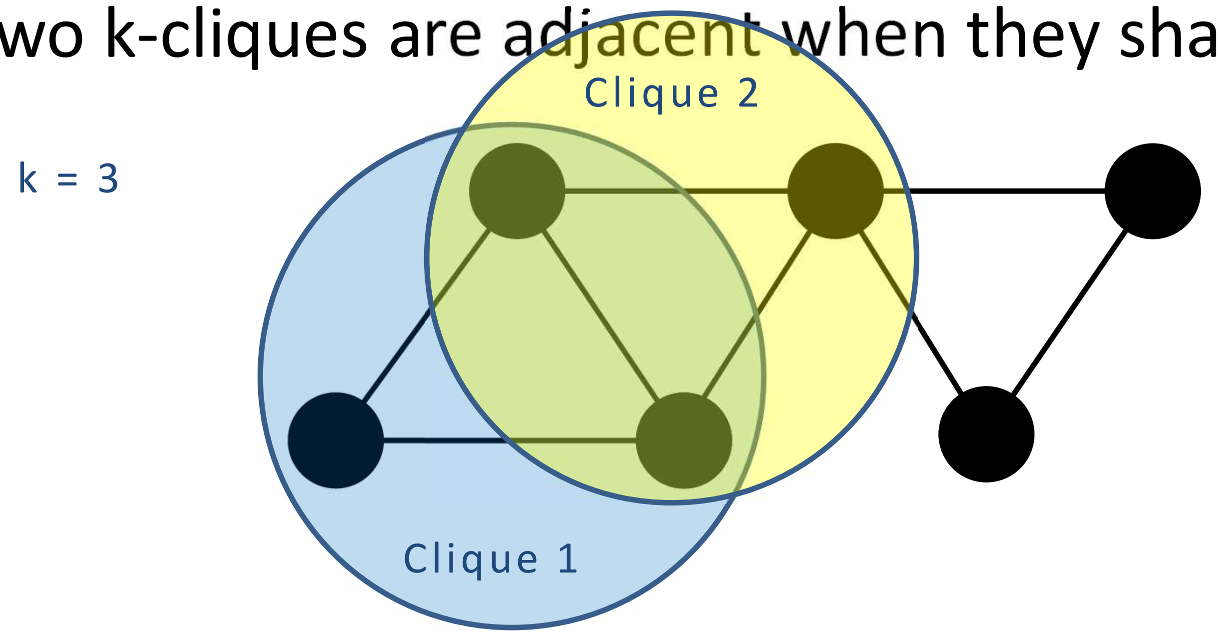
Two k-cliques are adjacent when they share **k-1** nodes



k-Clique Communities

- **Adjacent k-cliques**

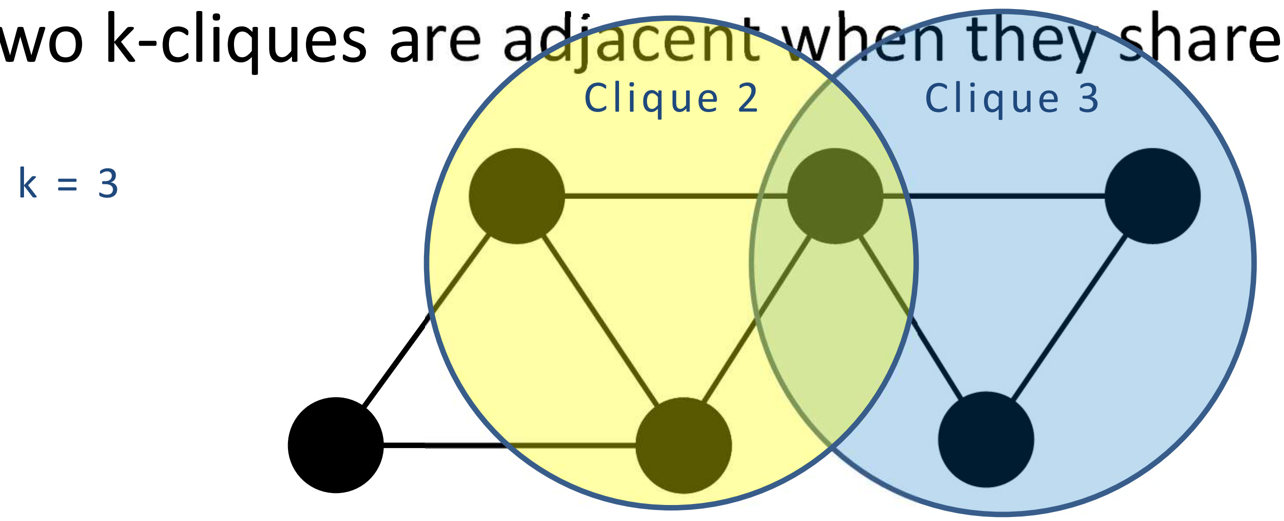
Two k-cliques are adjacent when they share **k-1** nodes



k-Clique Communities

- **Adjacent k-cliques**

  Two k-cliques are adjacent when they share **k-1** nodes



k-Clique Communities

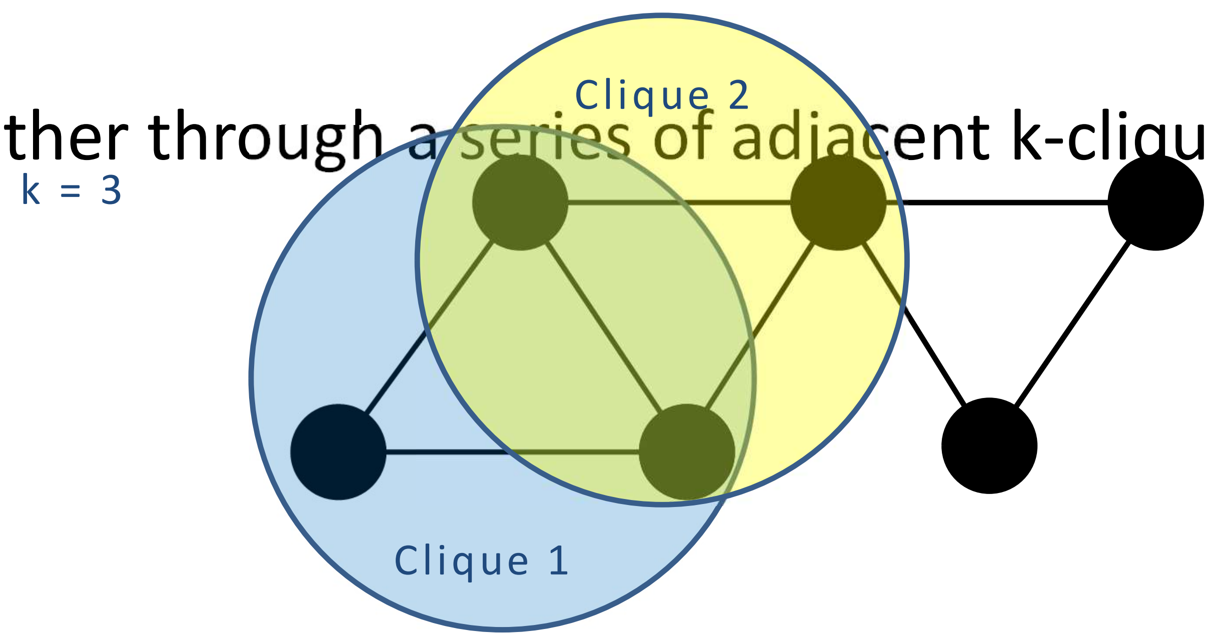
- **k-clique community**

  Union of all k-cliques that can be reached from each other through a series of adjacent k-cliques

k-Clique Communities

- **k-clique community**

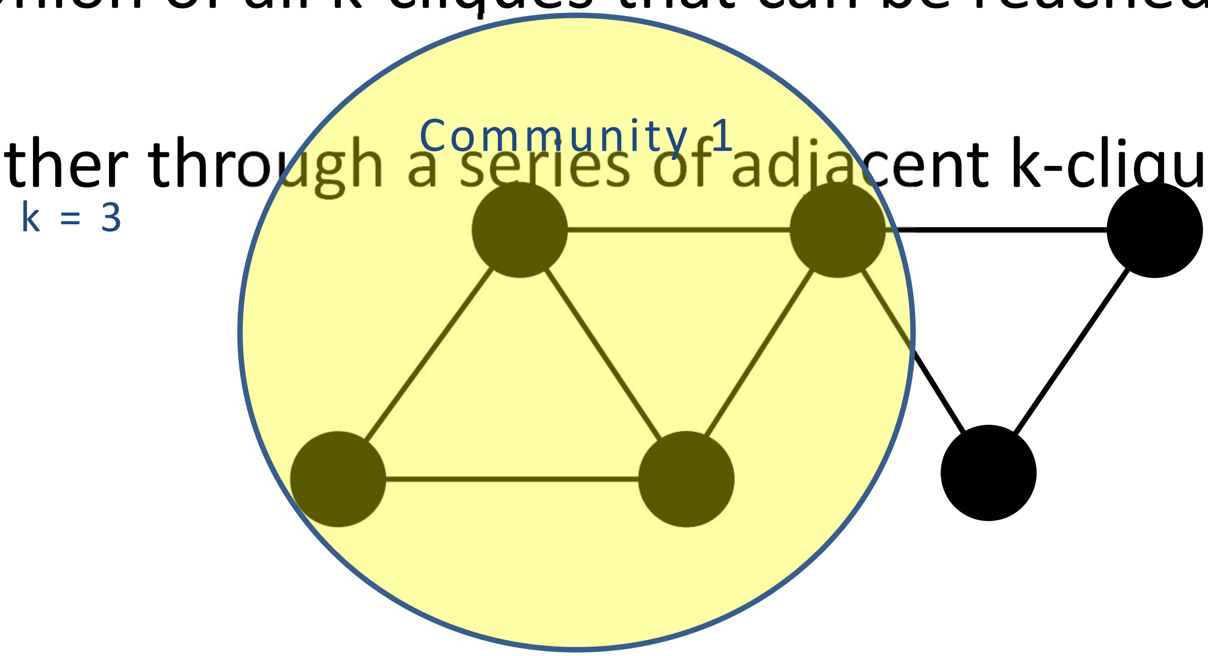
  Union of all k-cliques that can be reached from each other through a series of adjacent k-cliques



k-Clique Communities

- **k-clique community**

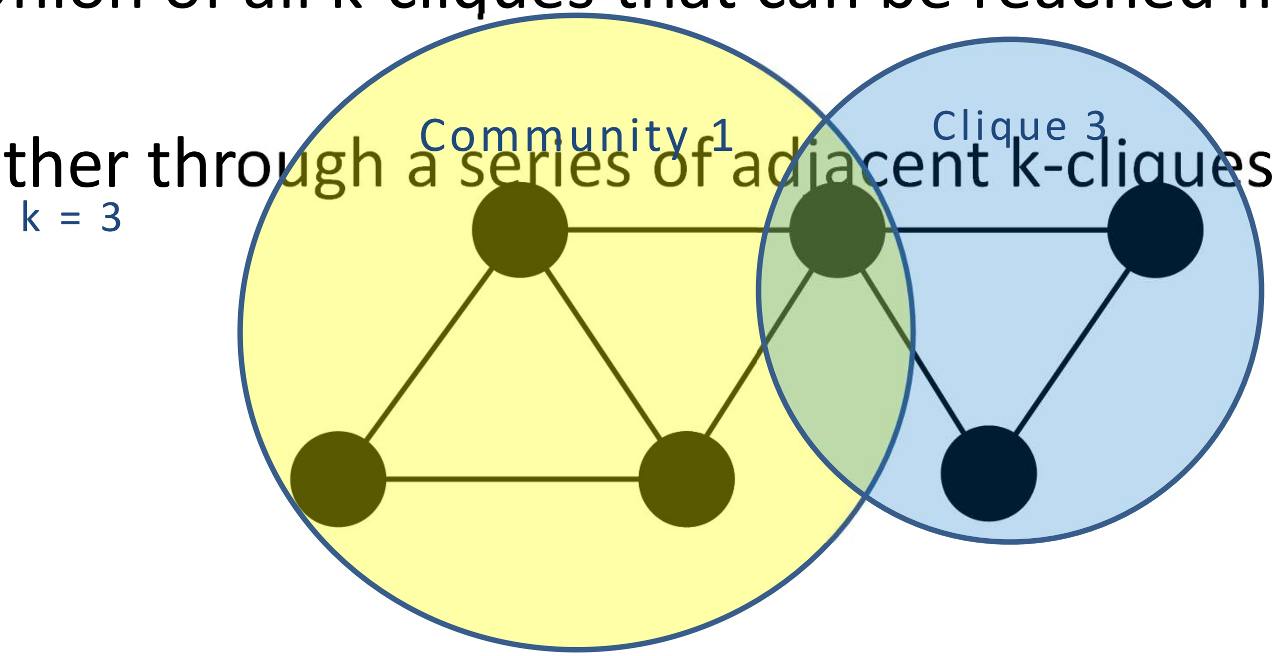
  Union of all k-cliques that can be reached from each other through a series of adjacent k-cliques



k-Clique Communities

- **k-clique community**

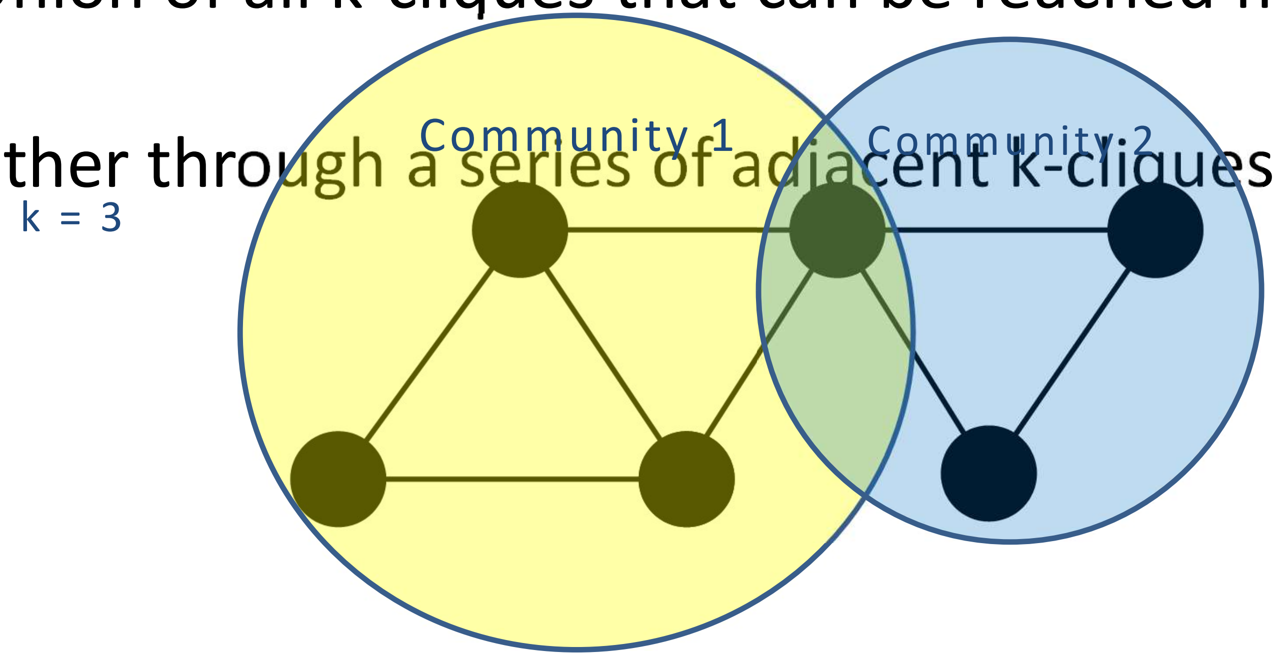
  Union of all k-cliques that can be reached from each other through a series of adjacent k-cliques



k-Clique Communities

- **k-clique community**

  Union of all k-cliques that can be reached from each other through a series of adjacent k-cliques



SimRank

SimRank = Measure of node similarity between nodes of same type.

Movie Database : Check similarity of two actors or two movies.

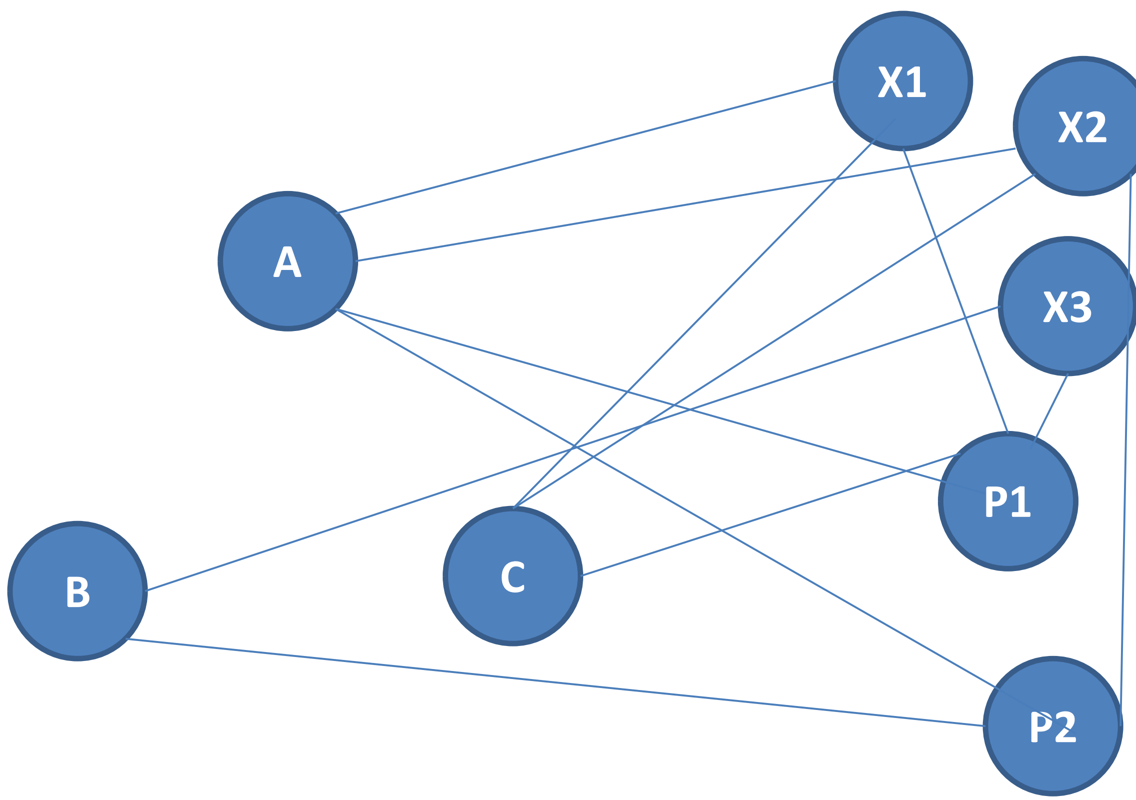
Applicable in any domain where relationships between objects are represented.

It gives a numeric value to similarity of structural context in which objects occur based on their relationships with other objects.

TRI-Partitle Graph:

A, C : Actors

X1, X2, X3 :Movies



If walker starts at X1..
he may reach at A or C or P1

If reaches at A
He may reach
X1 or X2 or P1or P2

If reaches at C
He may reach
X1 or X2 or P1

So From X1 there are chances to reach to X2 than X3. Computations similar to that of PageRank

Counting Triangle in Social Graph

<u>Identify small connected communities and count their occurrence.</u>

Sub-Graph = triangle (3-clique)

Most Commonly occurring pattern found in SN Graph.

- tendency of similar individuals to form group.
- Transitive nature of relationships A-B and B-D then A-D
- The basic structure of a graph is a triangle or 3-clique.
- Counting Triangle is an important aspect in SN analysis.

- **Why to count triangles?**
- **Clustering Coefficient:** degree to which a node's neighbors are themselves neighbors.
- **= No. of closed triplets in nodes neighborhood/ Total number of triplets in the neighborhood**

- **High Clustering Coefficient =** Closely connected communities.
- **Such communities are interested for the apps**
- Targeted Marketing Social recommendation

- **Low Clustering Coefficient =** Structural hole
- **A vertex that is well connected to many communities that are not otherwise connected to each other**
- **Such vertices can act as a bridge between communities.**
- Apps need to find influential nodes who can propagate info.

Counting Triangle Approach

Brute Force Method: Check every group if three triangles and check if they form a triangle or not?

Complexity = O(n3) times where n = number of vertices

High Computation cost : check if 3 v form triangle or to determine it does not form triangle

Smarter approach

- List all two-edge paths that are formed in the graph.
- Such paths of size 2 are called as **wedges**
- For every edge of (x, y) check if (x,y,z) forms a triangle .. If so add 1 to count of triangles. Repeat process.
- Analysis = $O(\sum_{v \in V} \text{degree}^{2}{}_{v})$