# PROJECT SEMINAR:

# COMMUNITY DETECTION IN SOCIAL NETWORKS

Presented by :
Shruti Pattajoshi
17CS01053

# OUTLINE

1. What is a community?
2. What is community detection? And why ?
3. Visualization of community networks
4. Different types of community networks

# OUTLINE

5. Clustering methodologies
6. Community evaluation
7. Applications in real life
8. Challenges and Conclusion

# Let’s discuss !

1. What is a community?
2. What is community detection? And why ?
3. Visualization of community networks
4. Different types of community networks

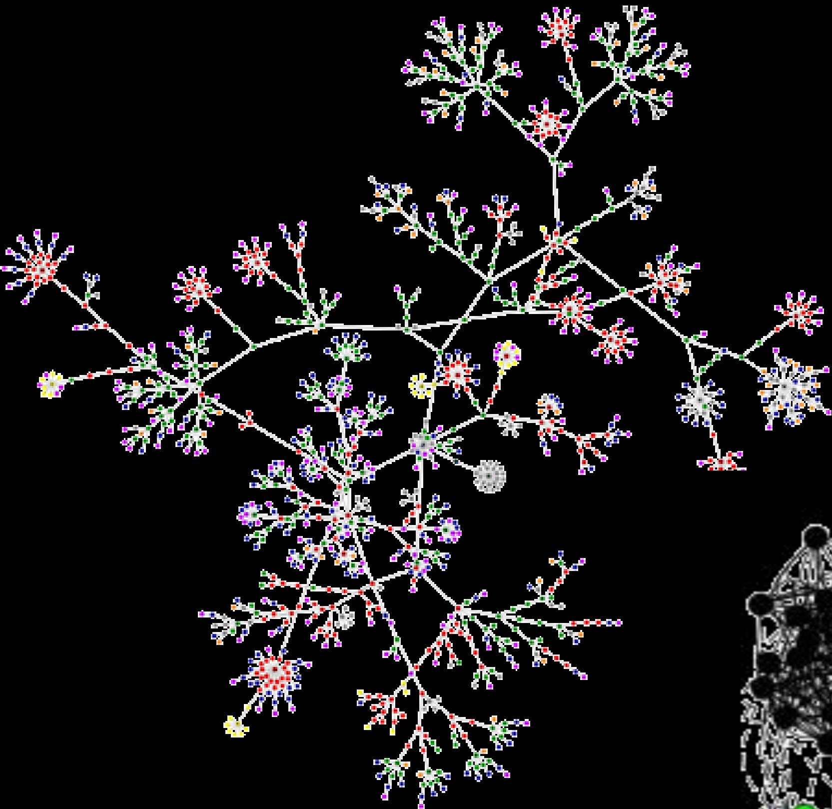
# What is a 'Community'?

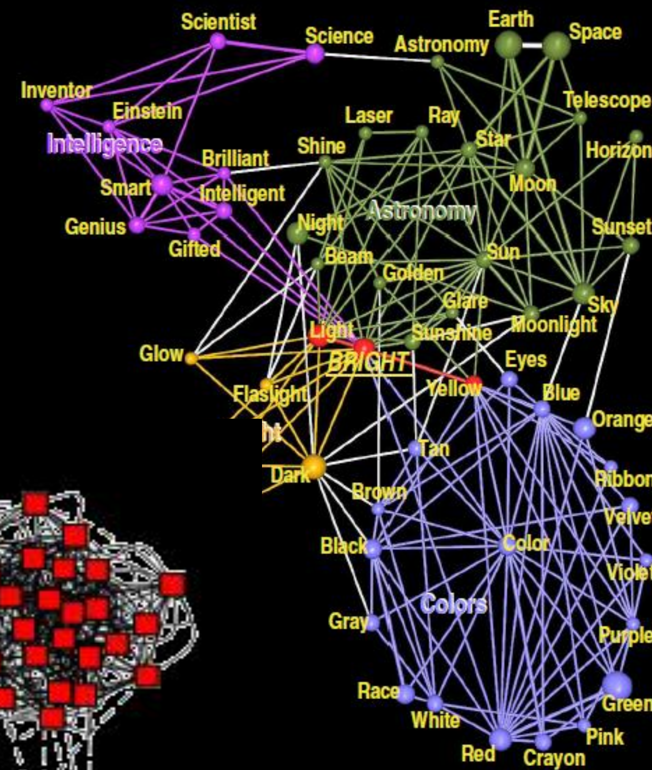
A *community* in a network is a subset of nodes that share common or similar characteristics, based on which they are grouped.

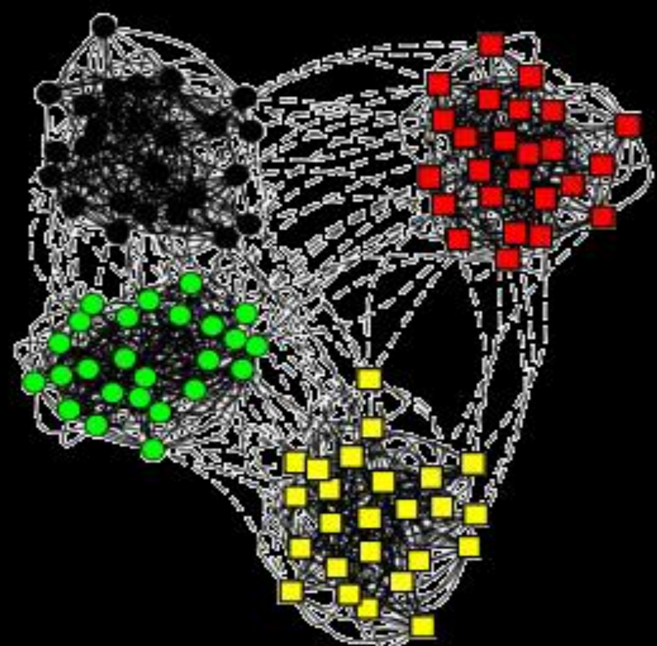
Circle of friends in social media

On an email network, group of Emails following similar patterns /domain.

Group of WebPages on closely related topics





Informally, a community C is a subset of nodes of V such that there are **more edges inside the community than edges linking vertices of C** with the rest of the graph

# OUTLINE

1. What is a community?
2. What is community detection? And why ?
3. Visualization of community networks
4. Different types of community networks

# What is ‘Community Detection’?

- Community Detection is all about partitioning , grouping, clustering or finding cohesive subgroups in a network based on common interests.
- Community detection makes sense only on **sparse graphs**


- The amount of research since 2002 in this area is massive
- Based on its usefulness, community detection became one of the most prominent directions of research in network science.
- It is one of the common analysis tools in understanding networks

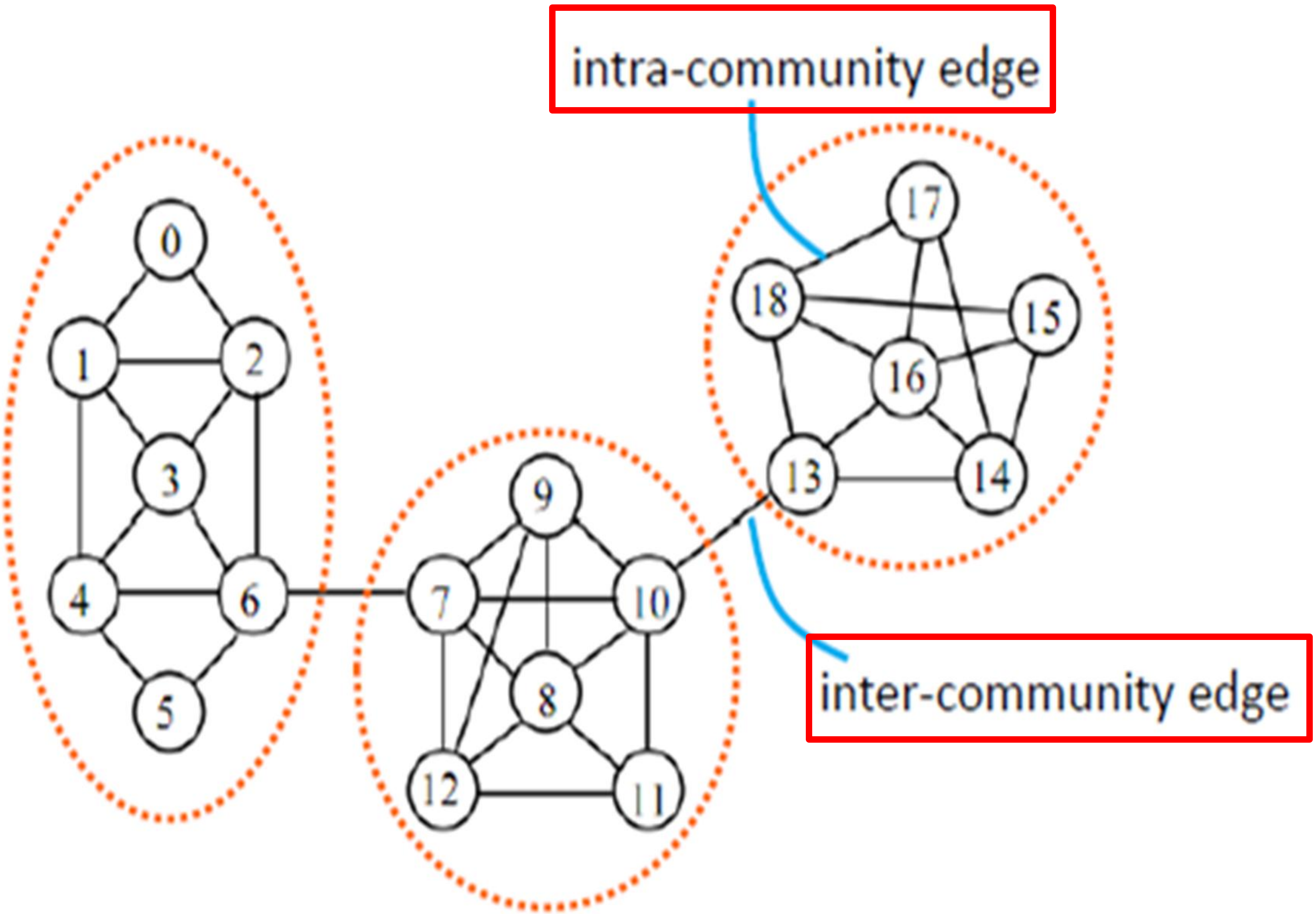
# Why ‘Community Detection’?

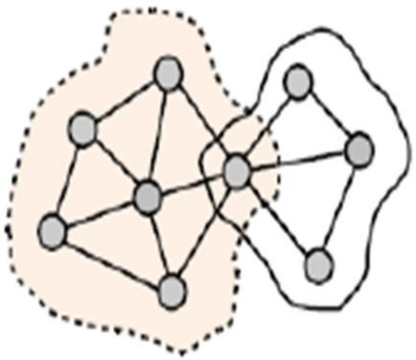
- Identify web clients with similar interest and geographically near each other
- Identify customer with similar interests (purchasing history)
- Difficult to meet friends in the physical world, but much easier to find friend online with similar interests
- Easy-to-use social media allows people to extend their social life in unprecedented ways

# OUTLINE

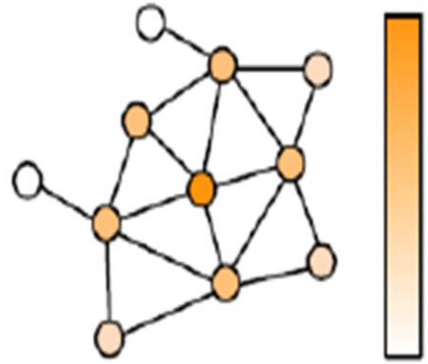
1. What is a community?
2. What is community detection? And why ?
3. **Visualization of community networks**
4. Different types of community networks

# Community Attributes :

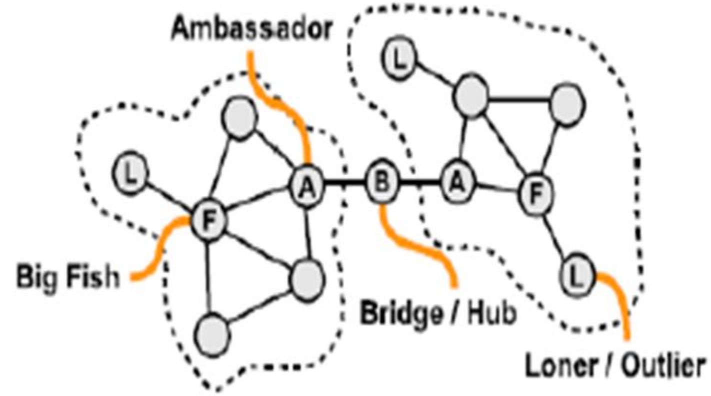




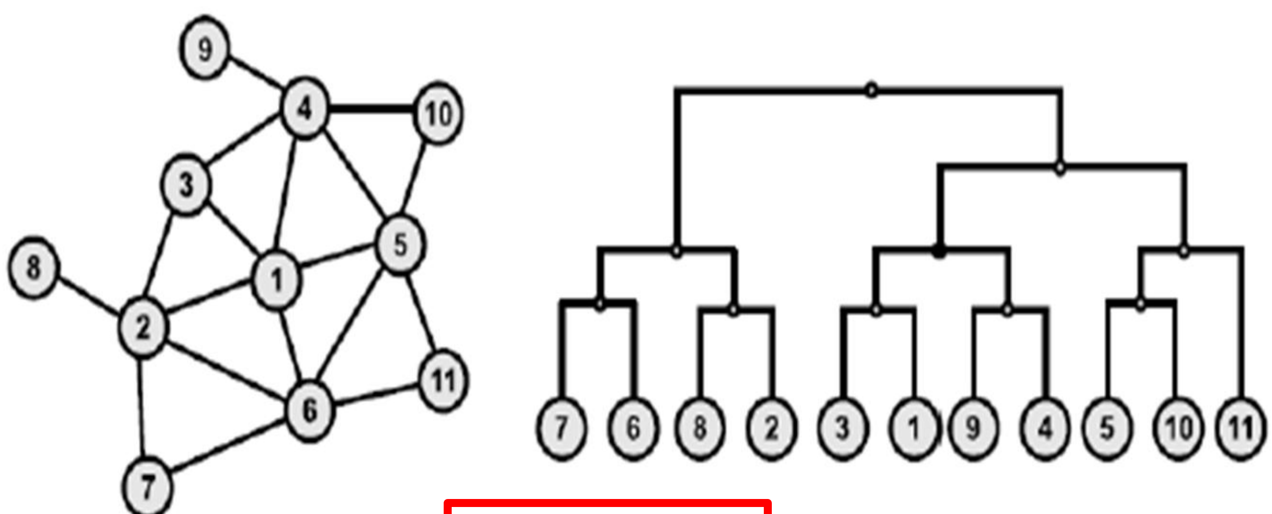
overlap



weighted participation

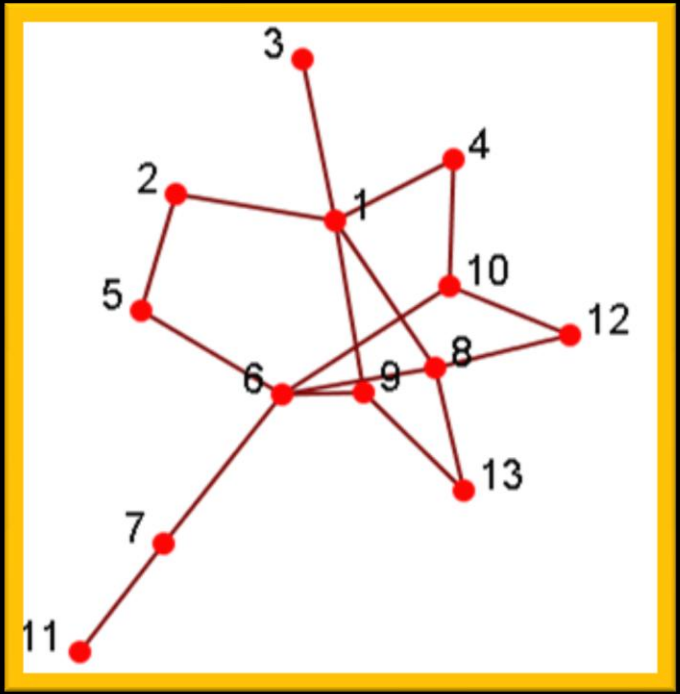


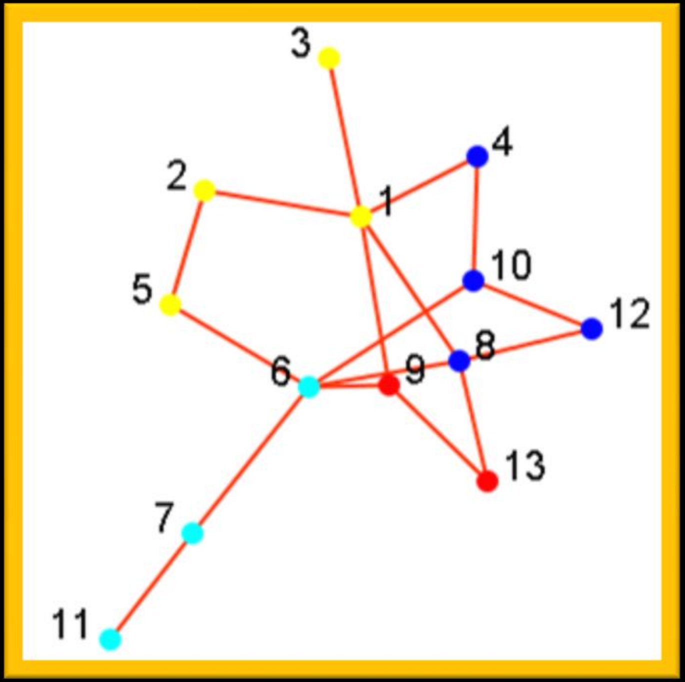

roles



hierarchy

# Visualization after grouping:




**4 Groups:**
**{1,2,3,5}**
**{4,8,10,12}**
**{6,7,11}**
**{9,13}**

**(Nodes colored by Community Membership)**

# OUTLINE

5. Clustering methodologies

6. Community evaluation

7. Applications in real life

8. Challenges and Conclusion

# Types of communities and their detection methods :

## Non-Overlapping

1. **Louvain Method**
2. **Girvan-Newman algorithm**
3. **Modularity maximization**

(One vertex may belong to only one group)

## Over-lapping

1. **Clique Percolation Method (CPM)**

(One vertex may belong to more than one Group)

# Communities Detection Methods:

1. Girvan-Newman algorithm
2. Louvain Method
3. Minimum-cut method
4. Clique Percolation Method

# Edge Betweenness

The "edge betweenness" of an edge can be defines as :
**The number of shortest paths between pairs of nodes that run along it.**

If there is more than one shortest path between a pair of nodes, each path is assigned equal weight such that the total weight of all of the paths is equal to unity.

**The edges connecting communities will have high edge betweenness**

Instead of trying to construct a measure which tells us which edges are most central to communities, we focus instead on those edges which are least central

If a network contains communities or groups that are only loosely connected by a few inter-group edges, then all shortest paths between different communities must go along one of these few edges

# INTRODUCTION

**A divisive algorithm based on "edge-betweenness"**

Focuses on edges that are most 'between' the communities and communities are constructed progressively by removing these edges from the original graph.

# TIME-COMPLEXITY

The worst-case time complexity of the edge-betweenness algorithm is **$O(m^2 * n)$** and is ***$O(n^3)$*** for sparse graphs, where *m* denotes the number of edges, and *n* is the number of vertices.

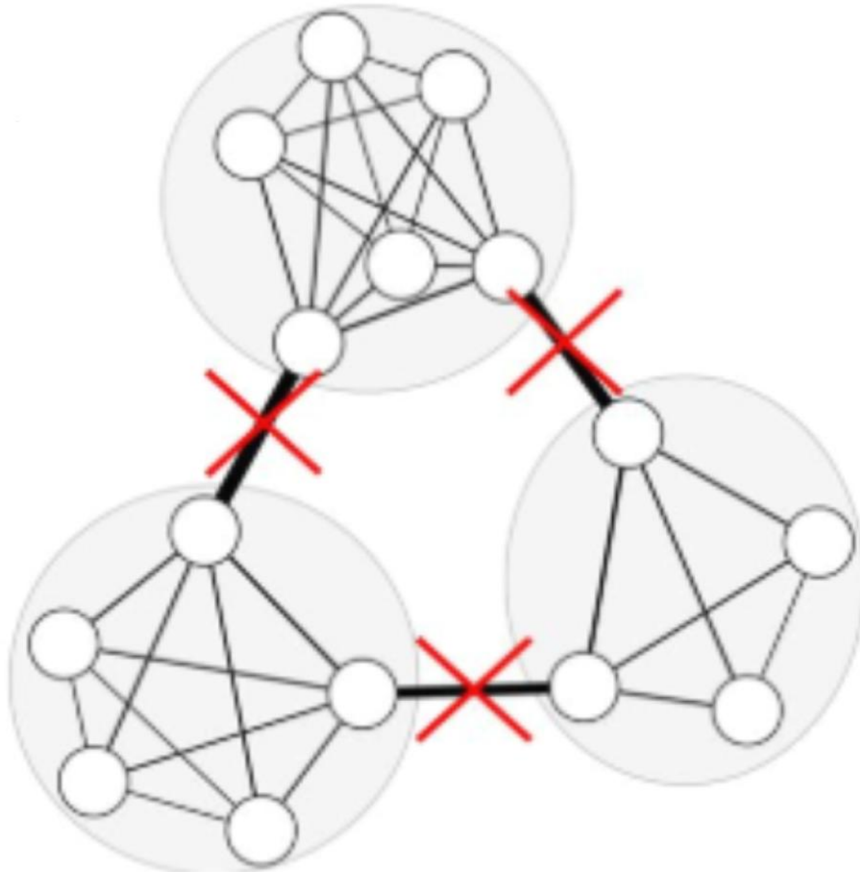
# ALGORITHM

Basic principle:
>Compute betweenness centrality for each edge.
>Remove edge with highest score.
>Re-compute all scores.
>Repeat 2nd step.

# APPLICATIONS-DEV.

Improve precision by the use of different between-ness measures or reduce complexity, e.g. by sampling or local computations.

Real Life Datasets/graphs:
- Social network in Zachary karate club



Partitioning approaches

Example: Girvan-Newman
(*edge-betweenness*)

# Communities Detection Methods:

1. Girvan-Newman algorithm
2. Modularity maximization
3. Louvain Method
4. Clique Percolation Method

# What is 'Modularity'?

- A graph can be **split into communities in numerous ways**, i.e. for each graph there are many possible community structures.

- In the simple case, a community structure is defined as a graph partition into a **set of node sets $C = \{C_i\}$**.

- To provide a measure of the quality of a community structure, we make use of modularity. Its used to gauge the goodness of the modules obtained from the community detection algorithms with **high modularity corresponding to a better community structure.**

- In **a random graph** (ER model), we expect that any possible partition would lead to $Q = 0$.

- Typically, in **non-random graphs** modularity takes values between **0.3 and 0.7.**

# Modulatiry Computation:

Modularity value

Degrees of nodes-pair

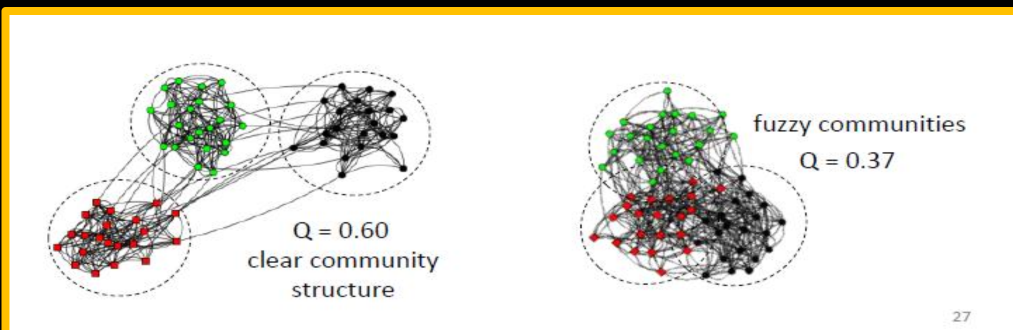
$$Q = \frac{1}{2m} \sum_{ij} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j).$$

# edges

Graph adjacency matrix

Probability of an edge if degrees are set and edges placed in random

In-same-cluster indicator variable



# INTRODUCTION

Methods seeks for a community structure that maximizes the value of modularity.

Merge nodes trying in each merging step to maximize the graph modularity (Newman, 2004).

# ALGORITHM

Initially, each node belongs to its own community (**N nodes -N communities**)
Visit each node in a order
To each node, **assign the community of their neighbor** as long as this leads to an **increase in modularity**.
This step is repeated many times until a local modularity maximum is found.

# TIME-COMPLEXITY

Leads to a hierarchical structure.
Complexity in a sparse graph: **O(n2)**

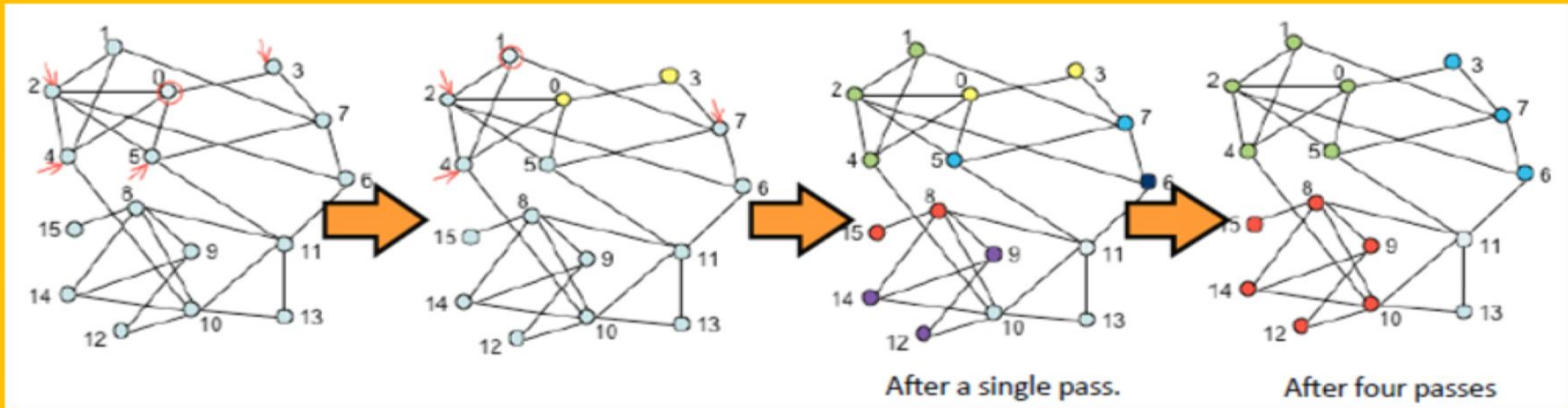
Use of appropriate data structures (max-heaps) can lead to complexity reduction (Clauset et al., 2004)
**O(nlog2n)**

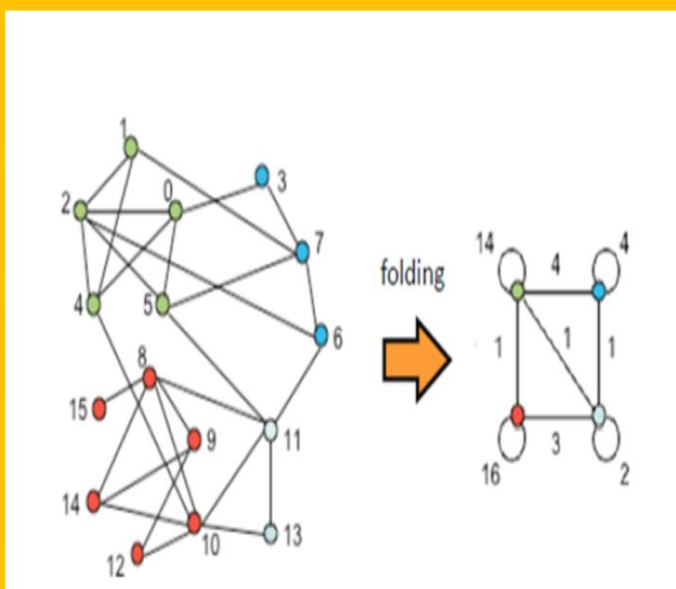
# ADVANTAGES

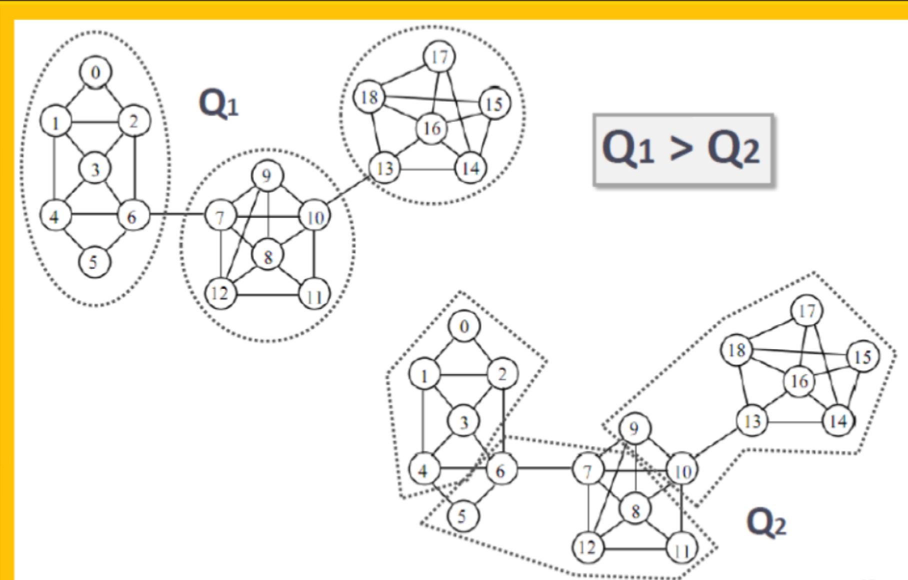
Folding: Create new graph in which nodes correspond to the communities detected in the previous step.

- Edge weights between community nodes are defined by the number of inter-community edges.
- Folding ensures rapid decrease in the number of nodes that need to be examined and thus enables large-scale application of the method.

# Modularity Maximization



# Communities Detection Methods:

1. Girvan-Newman algorithm
2. Modularity maximization
3. Louvain Method
4. Clique Percolation Method

# INTRODUCTION

Blondel et al.35 designed an iterative two-phase algorithm known as the *Louvain method.*

**Goal**: Optimize modularity → theoretically results in the best possible grouping of the nodes of a given network

# ALGORITHM

– Find small communities by optimizing modularity locally on all nodes,
– Then each small community is grouped into one node
– Then the first step is repeated

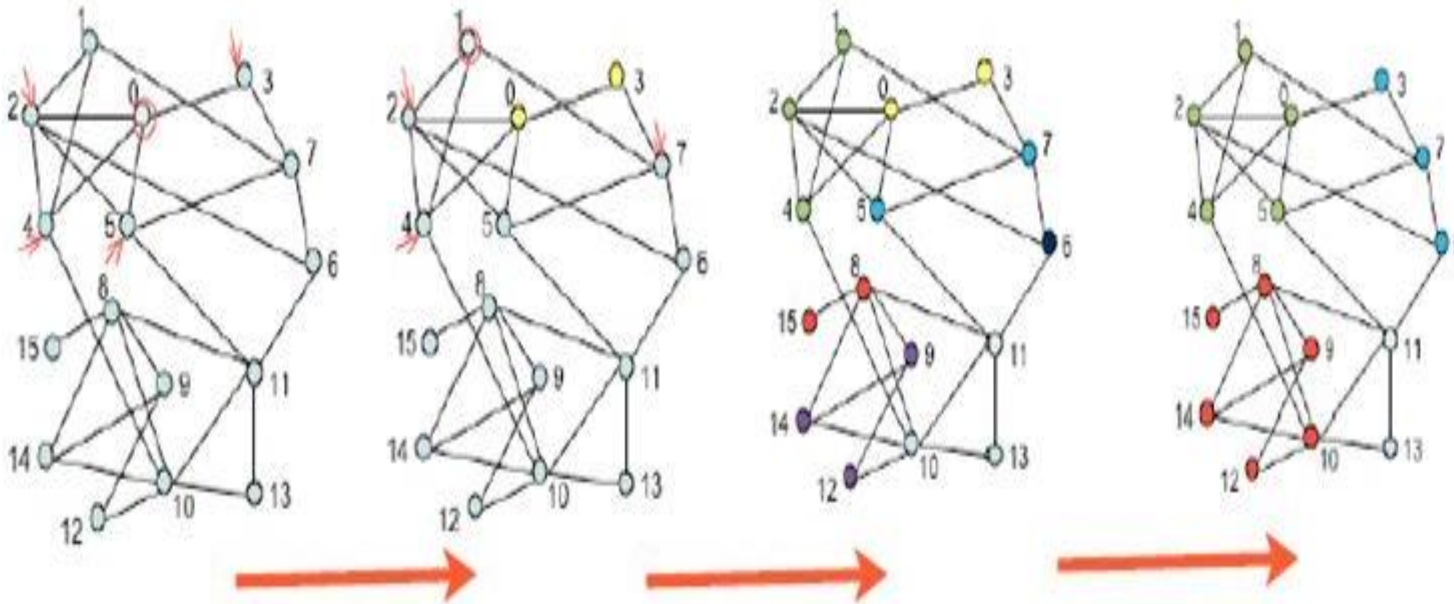
# TIME-COMPLEXITY

The algorithm improves the time complexity of the GN algorithm. It has a linear run time of ***O(m)***.

Simple, efficient and easy-to-implement (implemented in NetworkX, Matlab, C++, and Gephi)

# ADVANTAGES

- The method unveils hierarchies of communities and allows to zoom within communities to discover sub-communities, sub-sub-communities, etc.
- It is today one of the most widely used method for detecting communities in large networks.



Node 0 moves to the community of Node 3

After N nodes have been considered

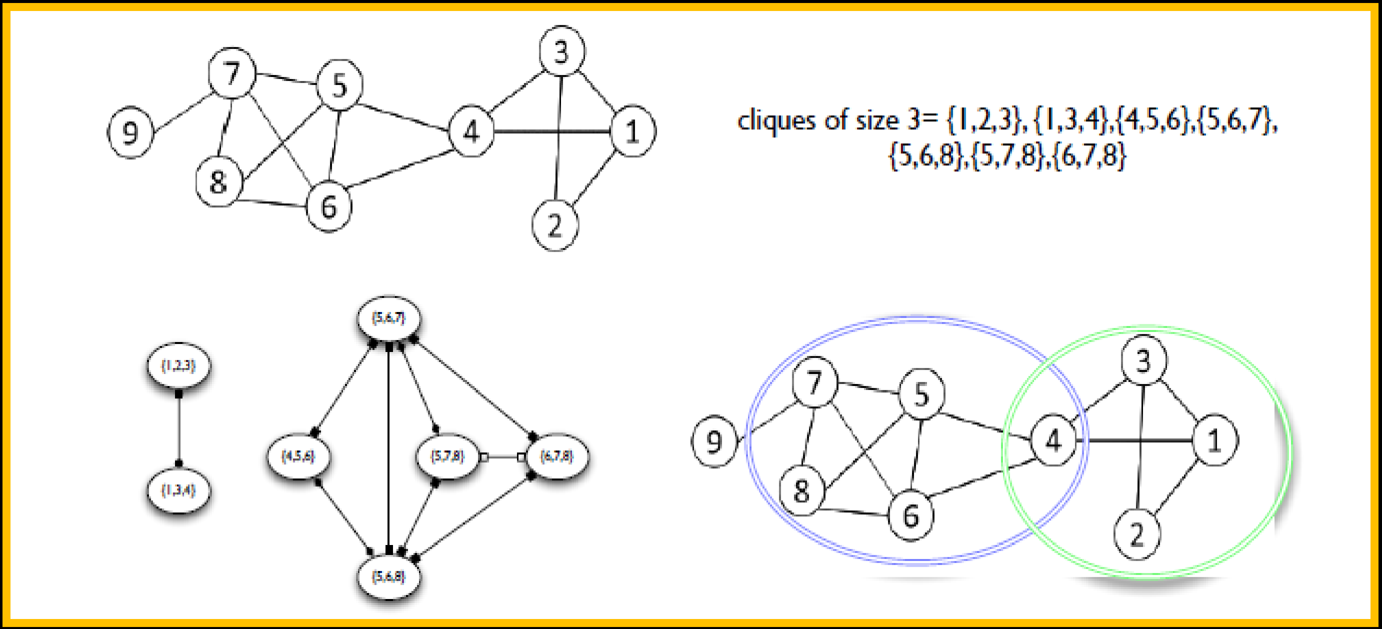
After each nodes has been considered 4 times

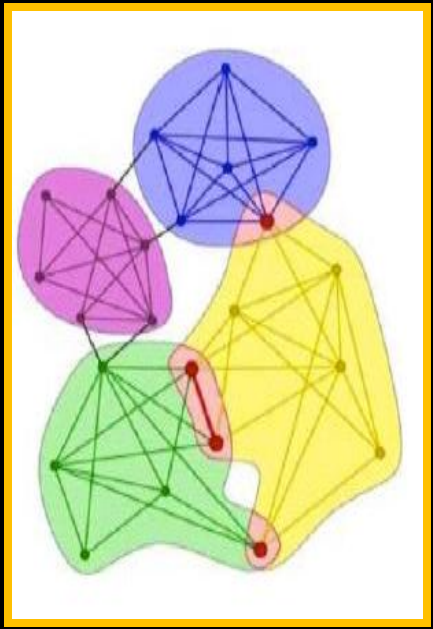
# Communities Detection Methods:

1. Girvan-Newman algorithm
2. Modularity maximization
3. Louvain Method
4. Clique Percolation Method

# Defining Clique

**Clique**: A maximum complete sub-graph in which all nodes are adjacent to each other



It is a NP-hard problem to find the maximum clique in a network

Normally use cliques as a core or a seed to find larger communities

# INTRODUCTION

The most popular technique to discover overlapping communities is the **Clique Percolation Method (CPM)**

- The internal edges of a community are likely to form cliques due to their high density
- It is unlikely that inter-community edges form cliques

# ALGORITHM

Given a parameter k:

1.Find all the cliques of size k

2.Construct a clique graph. 2 cliques are adjacent if they share k-1 vertices

3.Each connect component of the clique graph forms a community

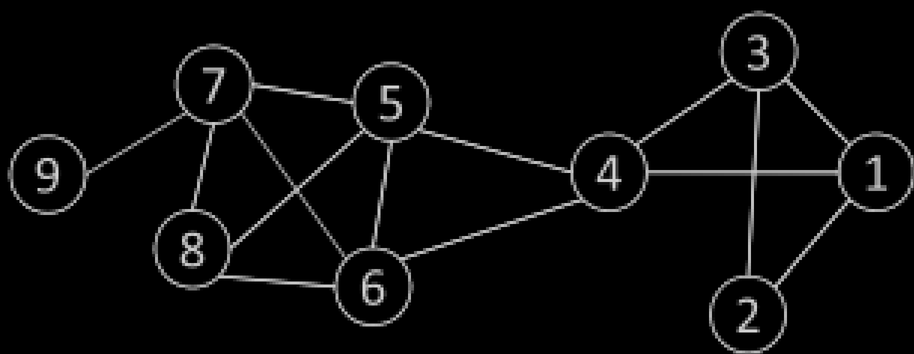
# TIME-COMPLEXITY

CPM has a run time of $O(\exp(n))$.

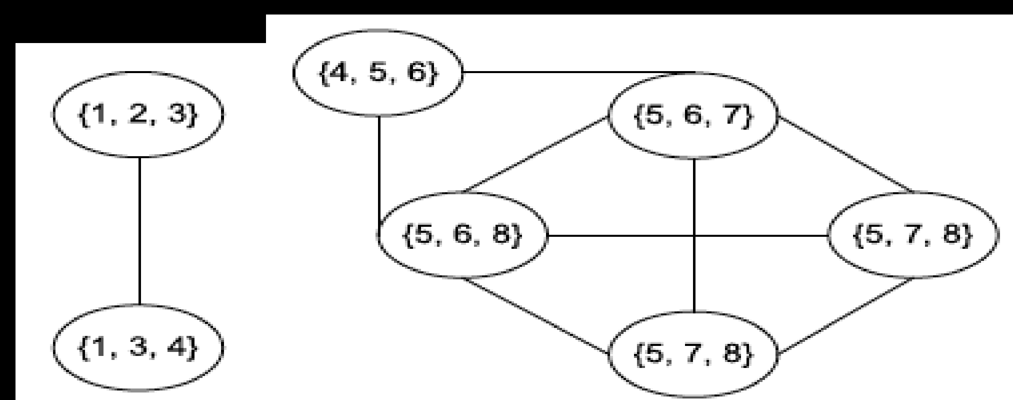
# LIMITATIONS

The CPM proposed by Palla et al.74 could not discover the hierarchical structure along with the overlapping attribute. This limitation was overcome through the method proposed by Lancichinetti et al.75

- Suppose we sample a sub-network with nodes {1-9} and find a clique {1, 2, 3} of size 3
- In order to find a clique >3, remove all nodes with degree <=3-1=2
  - Remove nodes 2 and 9
  - Remove nodes 1 and 3
  - Remove node 4




**Cliques of size 3:**
{1, 2, 3}, {1, 3, 4}, {4, 5, 6}, {5, 6, 7}, {5, 6, 8}, {5, 7, 8}, {6, 7, 8}




**Communities:**
{1, 2, 3, 4}
{4, 5, 6, 7, 8}

# Let's discuss!

5. Clustering methodologies
6. **Community evaluation**
7. Applications in real life
8. Challenges and Conclusion

## WITH GROUND TRUTH

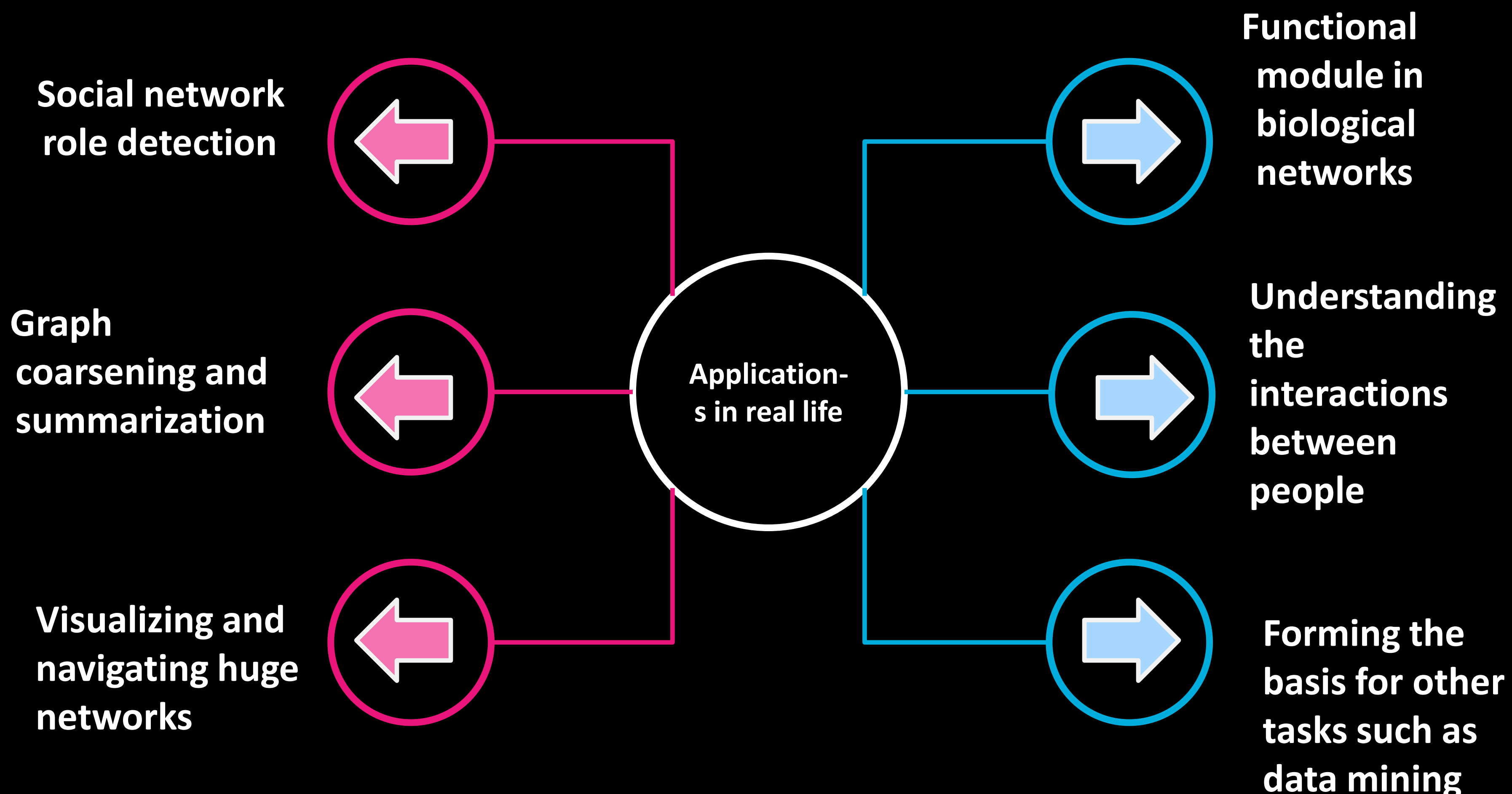
- Compare the partition provided by the algorithm with the **ground truth**
- We assume that the community membership for each vertex is known
- Subjective discussion/evaluation of results.

## WITHOUT GROUND TRUTH

- Extract communities from a (training)network
- Evaluate the quality of the community structure on a network constructed from a different date or based on a related type of interaction
- Quantitative evaluation functions like modularity, link prediction are used

# Let’s discuss!

5. Clustering methodologies
6. Community evaluation
7. **Applications in real life**
8. Challenges and Conclusion


Applications in real life
Social network role detection
Graph coarsening and summarization
Visualizing and navigating huge networks
Functional module in biological networks
Understanding the interactions between people
Forming the basis for other tasks such as data mining

# Let’s discuss!

5. Clustering methodologies
6. Community evaluation
7. Applications in real life
8. **Challenges and Conclusion**

CHALLENGES FACED

Concepts of “cluster”, “community” are not quantitively well defined Interpretation/Evaluation of networks

Community Detection Techniques only work when the graphs are sparse

Few Clustering problems are NP-Hard Problems so approaching/using them might stand expensive

# Conclusion

01 A valuable tool for understanding structure in massive networks

02 The optimal method depends on applications, networks, computational resources etc.

03 Other lines of research include Communities in directed networks and overlapping ones.

04 It also includes Community evolution and group profiling and interpretation