COMMUNITY DETECTION SUB-GRAPH IN SOCIAL NETWORK

SUDIP SAHA

JADAVPUR UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

KOLKATA-700032, INDIA

COMMUNITY DETECTION SUBGRAPH IN SOCIAL NETWORK

Thesis submitted in partial fulfillment of the requirements for the degree of

Master Of Computer Application

In

Computer Science and Engineering

Department

By

Sudip Saha

Examination Roll no-MCA196001
Registration No-137309 of 2016-2017
Class Roll No- 001610503001

Under The Supervision of

Prof. Nirmalya chowdhury

JADAVPUR UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

TO WHOM IT MAY CONCERN

*I hereby recommend that the project entitled "**COMMUNITY DETECTION SUBGRAPH IN SOCIAL NETWORK**" prepared under my supervision and guidance at **JADAVPUR UNIVERSITY**, Kolkata by* **SUDIP SAHA***(Reg. No. **137309** of **2016 – 17**, Class Roll No. **001610503001**), may be accepted in partial fulfillment for the degree of **MASTER OF COMPUTER APPLICATION** in the **FACULTY OF ENGINEERING AND TECHNOLOGY**, **JADAVPUR UNIVERSITY**, during the academic year 2018-2019. I wish him every success in life.*

Prof.(Dr.) Mahantapas Kundu
Head of The Department
Department of Computer Science and engineering
Jadavpur University, Kolkata-32

Prof.(Dr.) Nirmalya Chowdhury
Project Supervisor
Department of Computer Science and engineering
Jadavpur University, Kolkata-32

Prof. (Dr.) Chiranjib Bhattacharjee
Dean, Faculty council of Engineering. & Technology
Jadavpur University, Kolkata – 700032.

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC PROJECT

This is to certify that the work in the project entitled **Community Detection Sub-graph in Social Network** by **Sudip Saha** is a record of an original research work carried out by him under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of Master of Computer Application in the department of Computer Science and Engineering, Jadavpur University, kolkata-32. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

Name- SUDIP SAHA
Project title- COOMUNITY DETECTION SUB-GRAPH IN SOCIAL NETWORK
Roll No- 001610503001
Signature With Date-

Acknowledgments

I am grateful to many people who have helped towards shaping this project. It gives me immense pleasure and satisfaction to express my heart-felt gratitude to my guide, Prof.(Dr.) Nirmalya Chowdhury, for accepting me as his project student and providing me with excellent guidance and constant encouragement throughout my project duration. He devoted his valuable time towards discussions, and overall viewpoints and insights which went far beyond the narrow domain of work and helped me embark on new ideas. I am very much grateful to him for his invaluable suggestions, able guidance, during this period and above all constant encouragement throughout my work.

I would like to thank Mr. Ritam Sarkar for valuable support and suggestions regarding this project.

I would like to express my sincere thanks to all my teachers for providing sound knowledge base and cooperation.

I would like to thank all my classmates and friends for helping me in one or other way.

I must acknowledge the academic resources that I have got from Jadavpur University. I would like to thank administrative and technical staff members of the Department who have been kind enough to advise and help in their respective roles.

Last, but not the least, I would like to dedicate this thesis to my family, for their love, patience, and understanding.

SUDIP SAHA
EXAM ROLL- MCA196001

JADAVPUR UNIVERSITY
FACULTY OF ENGINEERING AND TECHNOLOGY

CERTIFICATE OF APPROVAL

The forgoing project is hereby accepted as a credible study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the project only for the purpose for which it is submitted.

FINAL EXAMINATION FOR EVALUATION OF PROJECT

1. _____

2._____

(Signature of Examiners)

INDEX

Abstract

Community detection in a social network is an emerging issue in the study of network system as it helps to realize the overall network structure in depth. Communities are the natural partition of network nodes into subgroups where nodes within the subgroup are densely connected but between the subgroups connections are sparser. Real world networks, including social networks have been found to partition themselves naturally into communities. A member of a social network can be part of more than one group or community. As a member of a social network can be overlapped between more than one group, overlapping community detection technique need to be considered in order to identify the overlapping nodes. This topic of research has many applications in various fields like biology, social sciences, physics etc.

In literature, most of the proposed community detection approaches are able to detect only disjoint communities. Recently few algorithms has been emerged which are capable of discovering overlapping communities. In this work two different types of algorithms have been proposed which efficiently detect overlapping communities. A novel approach has been introduced which overcomes the short-falls of clique percolation method(CPM), an overlapping community detection algorithm mostly used in this area. Another algorithm which is based on Genetic Algorithm is also used to discover overlapping communities. Modularity measure is generally used to determine the quality of communities for the particular network. The Quality of the communities detected by the algorithms is measured by several different overlapping modularity measures. Standard real world networks used as benchmark for community detection, have been used to judge the algorithms.

Introduction

Real world complex systems can be represented in the form of networks. To understand the in-depth structure and detail function of those systems, it is important to study and analyze the networks. A trivial property of these networks is community structure obtained by partitioning the network into several groups, within which connection between nodes are more dense than the rest of the network. The sets of This type of grouping is commonly referred as communities, but also known as clusters, cohesive groups, or modules as there is no globally accepted unique definition. The concept of community detection is related to graph partitioning in some way; though it is very much dissimilar from graph partitioning. In case of graph partitioning, number of groups and the approximate size of those groups are known priory and the task is usually to divide the network into these many numbers of disjoint sub-graphs of almost same size, irrespective of whether a partition even exists. But in case of community detection, it is not known that how many communities are present in the network and it is not at all mandatory for them to be of same size. The community detection approach assumes that most of real world networks, divide naturally into groups of nodes (community) with dense connections internally and sparser connections between groups, and the experimenter's job is only to detect these already formed groups . The number of partitions and size of them are settled by the network itself and not set by the experimenter. So community detection is the technique which aims at discovering natural divisions of (social) networks into groups based on strength of connection between vertices.

Basically, community can be subdivided into two types; disjoint communities and overlapping communities. In disjoint communities nodes can be part of only a single community,

but in overlapping communities partitions are not necessarily disjoint. There could be nodes that belong to more than one community .

A social network is a collection of finite set of members (nodes) which can be a single person, a group, an organization; and relations (edges) among them may represent friendship, influence, affection or conversely, dislike, conflict or many other similar entities. In a social network, a community could be a group of people with common interest or location. Generally in any social network a person may be part of more than one different group or community, like a person can be part of his/her professional group and simultaneously can be part of his/her family group indicating overlap between the professional and family group. So for social networks, overlapping community detection technique should be considered over disjoint community detection technique.

Literature survey

Community detection is a stimulating field of research. There are various community detection algorithms available but most of the algorithms are able to detect disjoint communities only. As overlapping community detection is comparatively new approach less algorithms are present for this approach. Some of these work are described below. In 2004, Newman proposed a disjoint community detection algorithm. Communities are found using edge betweenness. In this work first modularity measure is introduced. In the year of 2005, Clauset et al. proposed another community detection algorithm based on a local modularity measure proposed by them. At each step the algorithm adds a node into a partial community and update the neighbors of that community. It discovers only disjoint communities. Palla et al. presented the first overlapping community detection algorithm in 2005. In this approach communities are identified based on the adjacent cliques.

This algorithm allows a node to be part of more than one community, resulting overlapping community structure. In 2006, Newman proposed another community detection algorithm . The algorithm works by using eigenvector of matrices. Here he has proposed a new quality function, modularity matrix which has been used to detect the community structure. This algorithm results disjoint communities. Nepsuz et al. in 2008 proposed a new approach to evaluate modularity for overlapping communities. Lancichinetti et al. in 2009 have proposed an algorithm which works based on local optimization of a fitness function. This algorithm finds over-lapping communities by maximizing the fitness value. Shen et al., 2009 proposed another overlapping community detection approach. Overlapping communities are detected based on maximal cliques. An overlapping modularity measure is also proposed here based on number of maximal cliques. Gregory, 2009 , pro-posed a two phase method for overlapping community detection. In the first phase of this method, the network is transformed into a new one by splitting the nodes using split betweenness. In the second phase of the method, a disjoint community detection algorithm has been applied to process the transformed network. Ahn et al., 2010 presented another overlapping community detection algorithm based on link partition. Using hierarchical clustering links are partitioned to link dendogram. Overlapping communities are detected by cutting this dendogram at some threshold point. Here one more modularity measure partition density is introduced. Chen et al., 2010 proposed another algorithm to detect overlap-ping communities in weighted network. It detects overlapping communities using a local algorithm which works by expanding a partial community which is started from a special single node. They have introduced another overlapping modularity measure. Lazar et al. 2010 proposed a overlapping modularity measure for overlapping communities based on difference between inward and outward edges. Nguyen et al., 2011 proposed a two-phase framework which detect the over-lapping community structure in a dynamic network by quickly and

adaptively updating the network structure only based on its history without re-computing from scratch. Coscia et al., 2012 presented an overlapping community discovery algorithm. Each node vote for the communities it sees surrounding it using a label propagation algorithm and finally, the local communities are merged into a global collection. Tooth et al., 2013 studied various modularity measures by applying them on community structures obtained using clique percolation method.

Those are some of graph theory based overlapping community detection approaches. Few author has attempted to detect overlapping communities using genetic algorithm by maximizing the modularity value. Some of these type of approach are described here. Clara Pizzuti, 2008 proposed an genetic algorithm based disjoint community detection approach which uses node based clustering. A simplified objective function is also proposed here. Clara Pizzuti, 2009 has proposed first genetic algorithm based overlapping communities detection algorithm. It uses edge-clustering approach instead of node clustering. They have used line graph concept to achieve that. Cai et al., 2011 proposed another genetic algorithm based approach to detect overlapping communities. The algorithm first finds the link communities by optimizing objective function partition density, and then map the link communities to overlapping node communities. They have used the concept of bridge node to adjust the node membership of overlapped nodes. Dickinson et al., 2013 also attempted to detect overlapping communities using genetic algorithm. Overlapping communities are detected using two different approach Label Rank algorithm and genetic algorithm. For genetic algorithm edge based clustering has been used and Modularity by Shen et al. is used as the objective function.

STATEMENT OF THE PROBLEM

Zachary's Karate Club Network

The method proposed in this chapter is evaluated on this network. Initially with k = 4, CPM detects three communities which are (1, 2, 3, 4, 8, 14), (24, 30, 33,34) and (9, 31, 33, 34) as represented in Figure 1 with different colors. Node 33 and 34 are overlapped between last two communities. Total 22 nodes are not included in any community, though they are connected to the network i.e., their friends are part of some community. Proposed method includes these nodes into initial communities on the basis of their belonging coefficient as discussed in Algorithm 2. The updated communities are (1, 2, 3, 4, 8, 14, 5, 6, 7, 10, 11, 12, 13, 18, 20, 22, 17), (24, 33, 34, 30, 10, 15, 16, 19, 21, 23, 26, 27, 28, 29, 32, 25) and (33, 34, 9, 31, 10, 15, 16, 19, 21, 23, 29). Here it can be observed that second and third detected communities are very much similar as they share most of the members. Similarity between these two communities is found as 0.81. So these two communities are integrated into a single community as (9, 10, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34). Finally according to the proposed method, two communities are found in the network which are (1, 2, 3, 4, 8, 14, 5, 6, 7, 10, 11, 12, 13, 18, 20, 22, 17) and (9, 10, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34). This indicates the actual partition in the network [25]. Details of community structure for both CPM and ECPM is given in Table 1. The cover detected by ECPM is presented in Figure 2. With k = 3, three communities have been found initially. Node 1 and 32 are found as overlapped nodes. This community structure does not include node 10 and 12 into any community. Proposed method covers these two nodes by including node 10 and 12 to initial communities. Node 10 included as overlapped

between two communities as it shares one connection to both the communities. Proposed method results three communities with three overlapped nodes. Detailed outcome of both the algorithm, CPM and ECPM is given in Table 1.

Table 1: Community Structure Details of Karate Club Network

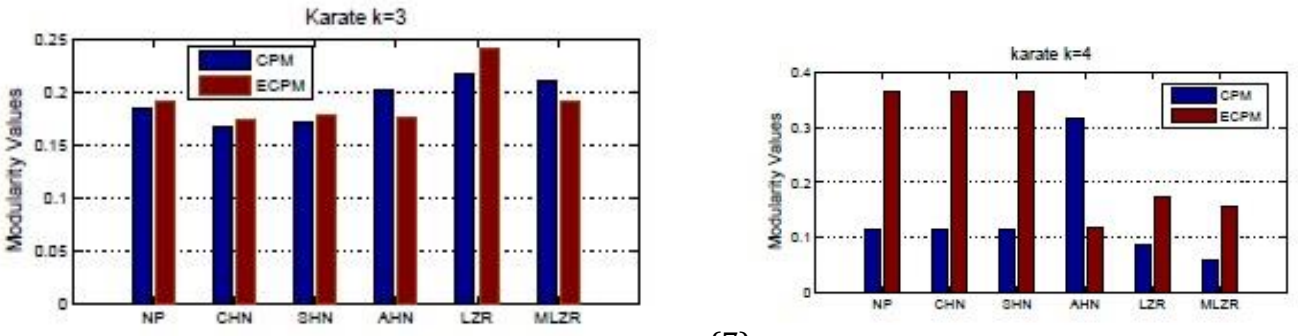
k	**CN(%)**		**UNC**		|C|		**OV**	
	CPM	ECPM	CPM	ECPM	CPM	ECPM	CPM	ECPM
3	94 %	100 %	2	0	3	3	2	3
4	35 %	100 %	22	0	3	2	2	1

CN: % of nodes covered; UNC: Number of nodes uncovered; |C|: Number of communities

OV: Number of overlapped nodes; ECPM: Proposed algorithm (Extended CPM)

The details of community structures detected by CPM and ECPM for Karate network for k value 3 and 4 are summarized and compared in Table 1. It is visible that proposed method ECPM covers all nodes in connected network resulting 100% node coverage i.e, 0 uncovered node.

Figure 3.4: Modularity values for community structures detected using CPM and ECPM in karate club network.



As modularity defines the quality of the detected community structure, modularity is measured for each community structures. As there are no universal modularity measure, more than one overlapping modularity measures are considered to assess the quality of the covers. Here six overlapping modularity measures have been considered. For both k = 3 and k = 4, overlapping modularity measures have been computed for detected community structures using CPM and ECPM. Modularity values are shown in Figure 3.4. Meaning of NP, CHN, SHN, AHN, LZR, MLZR is defined in Table 2.1. For both the cases with k value as 3 and 4, modularity values for covers, detected using ECPM are greater than the modularity values for CPM cover. It indicates ECPM results better quality detection

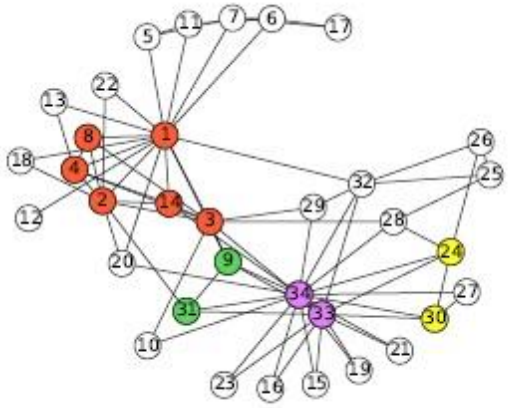


Figure 1: Communities detected using CPM in Karate Club Network with k = 4. Node colored white are uncovered that is not part of any community. Node 33 and 34 are overlapped.

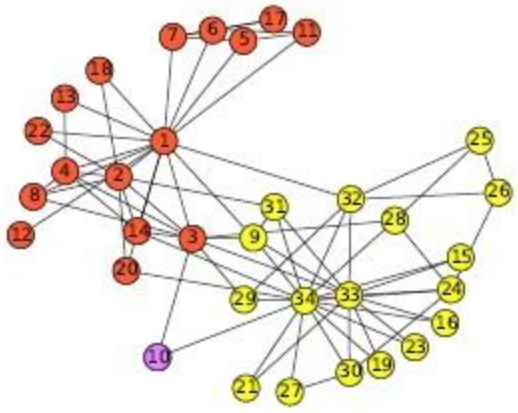


Figure 2: Communities detected using proposed method in Karate Club Network with k = 4. Node 10 is overlapped between two communities.

these fundamental concepts are related to community detection are described which have been used throughout the thesis.

➢ Network :

A network represents the actors and relations among them. A network is generally presented as a graph G(V, E), where V is set of n nodes and E is set of m edges. Actors in the network are called as nodes, vertex and relations can be called as edge, arc, link, tie, bond or connections.

The nodes and edges in the network may have different properties depending on the concerned system. For example edges or connections can be weighted, directed or undirected and nodes can have a size, color or many other attributes generally termed as labels. For example, in a social network, one might need to consider the gender or age of the considered population. In the corresponding network, this translates into labels associated to each node. Links labels are more common, and generally take the form of numerical values called weights, expressing the strength of the corresponding to relationships. Another important link attribute is the direction, which introduces a distinction between the connected nodes: one is the tail, the other is the head, and the link represents an asymmetrical relationship from the tail towards the head. Depending on the links, a network is said to be (un)weighted and (un)directed. In this work networks, with attribute-less nodes and unweighted, undirected links has been considered.

➢ Adjacency Matrix :

The information about a network can be represented by some mathematical expressions, vectors, matrices etc. One of the most common way of network representation is adjacency matrix. It describes how the nodes are connected.

$$A_{ij} = \begin{cases} 1, & \text{if a link exists between node i and node j} \\ 0, & \text{otherwise} \end{cases}$$

If the network is weighted then the weights are used in place of binary values.

➢ Degree of Nodes :

Degree of a node is the number of connections, a node is having with other nodes. In another words, number of neighbors a node is directly associated with. For a directed network a node has two type of degree which are in-degree and out-degree. These are distinguished by the direction of links. In-degree of a node is the number of incoming links where out-degree is the count of outgoing links. Nodes with no links that is of zero degree are called as isolated nodes .

➢ Average Degree :

Average degree of a network is the mean degree processed over all of the nodes of the network. It depends on the considered system, the number of links and nodes. The average degree of an undirected, unweighted network can be calculated as

$$d_{avg} = \frac{1}{n} \sum_{i} d_i$$

Here, d_{avg} is average degree of network, d_i symbolizes the i^{th} node's degree and n is the total node number.

➢ Degree of Distribution :

A significant property of real world network is power law degree distribution. The degree distribution, p(k) of a considered network is defined as the fraction of nodes in the network having degree of k. If there are n nodes in total in a network and n_k of them have degree k, then $p(k) = n_k/n$. The networks which follows power-law degree distribution is known as scale free network. It has been seen that degree distribution of real world network follows power law .

➢ Small World :

If the average path length between two nodes is small then the considered network is having small world property. The famous experiment of Milgram's named as, "six degrees of separation" presents the idea, that any two person in earth is at most six steps away from each other. A chain of, "a friend of a friend" can be made which will be able to connect any two person in six steps or less distance. The empirical studies are shown that many real networks has small world property.

➢ Component :

A component is a sub-network in which any node is reachable from any other node by a walk. Putting concisely, it is a maximal connected subgraph. For an undirected network, a component is a set of connected nodes with no links with other nodes from the same network. But for directed networks, it is less straightforward. A component is said to be strongly connected if there is a directed walk between each pair of nodes It is called weakly connected if there is at least an undirected walk between each pair of nodes. A network with only one component is said to be connected. An isolated node (i.e. a node with a degree zero) is a component of its own.

➢ Community:

A community is a group of actors or nodes, and connections or links between these nodes where nodes are clustered into tightly knit groups with high density of within-groups edges and low density of between-group edges . Communities have been studied in diverse domains which results several names to refer them like modules, partitions, clusters or cohesive subgroups. In Figure 1, nodes 1, 2 and 3 combinedly represents a community. Again Nodes 3, 4 and 5 represents another community.

Basically communities can be subdivided into two types; disjoint communities and overlapping communities. In disjoint communities nodes can be part of only single community but in overlapping communities partitions are not bounded to be disjoint. There could be nodes that belong to multiple communities .

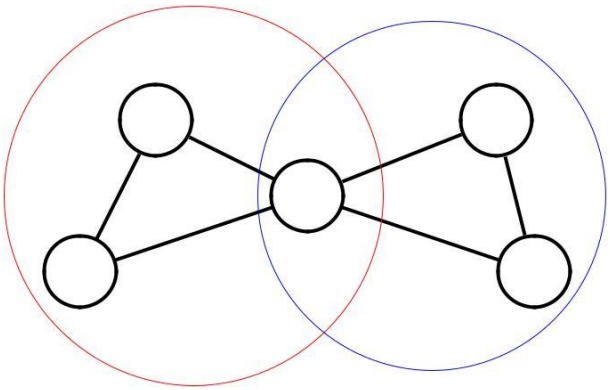


Figure 1: Community Structure in a network

➢ Community Structure :

Set of communities discovered in a network combinedly referred as the community structure or cover, represented as C = {c_1, c_2, c_3, ..., c_k}. Here C is the cover or community structure and c_1, c_2, c_k are communities. The size of cover represented by |C| indicate number of communities in the cover. For example in Figure 1 communities are c_1={1, 2, 3} and c_2={3, 4, 5} and cover |C|= {c_1, c_2}.

➢ Overlapped and Non-Overlapped node :

An overlapped node is shared by more than one community. Type of community structure where some nodes are overlapped is known as overlapping community structure. For example in Figure 1 node 3 is an overlapped node.

An non-overlapped node belongs to only one community. Type of community structure where all nodes are non-overlapped is known as non-overlapping or disjoint community structure.

➢ Modularity Measure :

Actual community structure or cover in a network is not always fixed. It corresponds to the arrangement of edges. Modularity is a quantitative measure of the quality of the community structure or inversely it defines how good the founded community structure is . It can be either positive or negative, where positive value indicates possible presence of community structure.

Presently most commonly used modularity measure for overlapping community structure is Normalized Mutual Information (NMI) . The problem with this modularity measure is on ground truth about the network which needs to be known priori. But it is nearly not possible for real world networks. Recently few modularity measures for overlapping community structure have been developed which does not require the ground truth information . In most of the papers researchers have considered a single modularity measure to rate the community structure. Every modularity measure has its own strength and weakness. It will not be fair enough to declare a single modularity measure as ideal one; as there is no such universal modularity. In this work a number of modularity measures have been considered for quality measure.

➢ Modularity by Girvan and Newman :

The concept of modularity is proposed by Girvan and Newman . According to them a good community division in network is not that where there are fewer edges between the communities; it is where there are more edges within communities than between the communities . Modularity is defined as the difference between number of edges falling inside the communities of the network and the number of edges expected to fall inside the communities in an equivalent random network. The reason of this definition is that number of connections inside the community should be larger than the what is expected for a random network. So the algorithm works by comparing the fraction of connections or edges inside the communities to the expected fraction of edges in a random graph where the degrees of individual nodes are equal to the degrees of nodes in the original network.

In a random graph having equal number of edges, m of original network, the probability for having a connection between two nodes i and j with degrees d_i and d_j respectively is $d_i d_j / 4m^2$. Accordingly, the expected fraction of edges inside a community c is $(d_c/2m)^2$, where $d_c = \sum_{i \in c} d_i$ is the sum of node degrees in community c.

As modularity is measured by comparing fraction of edges inside the community with expected fraction of edges in random model, the modularity contribution by community c is given in equation (1)

$$Q_c = \frac{l_c}{m} - \left(\frac{d_c}{2m}\right)^2 \qquad (1)$$

where, m is total number of edges in the network and l_c is total number of edges within the community c. So l_c/m is fraction of edges or connections in community c. Based on the above, the modularity of the cover can be computed as the sum of contribution for all communities, as mentioned below:

$$Q = \sum_{c \in C} Q_c = \sum_{c \in C} \left[\frac{l_c}{m} - \left(\frac{d_c}{2m} \right)^2 \right] \quad (2)$$

Equation (2) can also be written in a different form with summation over the individual nodes as:

$$Q = \frac{1}{2m} \sum_{c \in C} \sum_{ij} \left(A_{ij} - \frac{d_i d_j}{2m} \right) \delta_{ic} \delta_{jc} \quad (3)$$

where, A_{ij} stands for the corresponding value in adjacency matrix. δ_{ic} indicates whether the node i belongs to community c or not. $\delta_{ic} = 1$, if node i is part of community c and $\delta_{ic} = 0$, when it is not.

The modularity measure suggested by Newman and Girvan may be considered as appropriate for disjoint communities as, the value of δ_{ic} in Eq. (3) can only be 1 or 0, which indicates a node can be part of a single community or none. But in case of overlapping communities a node may belong to more than one community. So δ_{ic} needs to be evaluated in a way so that it can measure how much fraction of node i is dedicated to community c. Recently few modularity measures are proposed which takes this overlapping case into consideration.

➢ Clique and k-clique:

Clique in a graph is a subset of nodes where each pair of node is connected through an edge, that is a complete sub-graph. Finding all cliques with a given size in a graph, is an NP-hard problem . The notation, k-clique used in the community detection technique is completely different from the k-distance clique in graph theory. Here k-clique indicates size of the clique i.e., the clique consist of k nodes e.g. a 3-clique indicate a complete sub-graphs having 3 nodes . Figure 1 shows an example network having 3-clique and 4-clique. Six 3-cliques are (1, 2, 3), (1, 2, 8), (2, 4, 5), (2, 4, 6), (2, 5, 6) and (4, 5, 6) and one 4-clique is (2, 4, 5, 6).

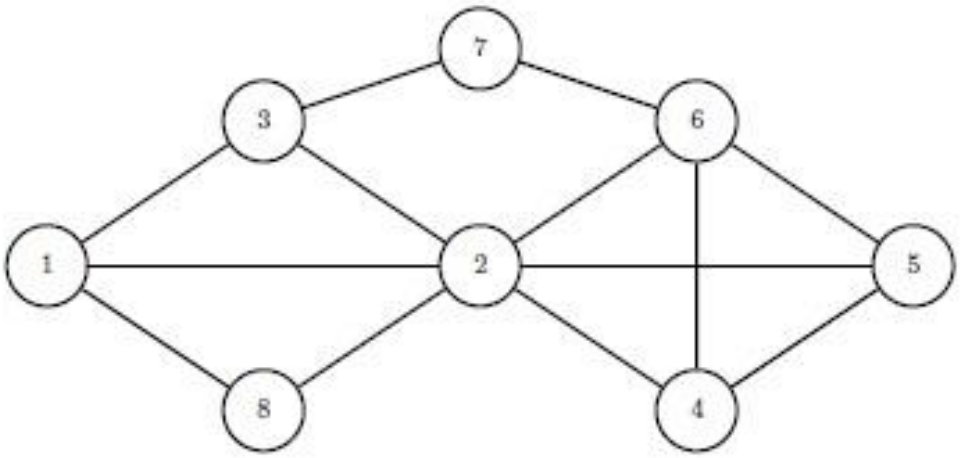


Figure 1: An example of Network

➢ Clique Percolation Method (CPM) :

CPM is a clique based overlapping community detection algorithm used mostly in this field. As the connection between the nodes within the community is dense, it is obvious that edges within a community form cliques (complete subgraph) due to their high density. But it is unlikely that edges between communities i.e., inter-community edges form cliques.

The assumption based on which clique percolation method works is that a community comprises of overlapping sets of fully connected sub graphs. So this algorithm detect communities by searching for adjacent cliques. It begins by exploring all the k-cliques (clique of size k) in the network. When all the k-cliques have been found a new graph commonly referred as clique-graph is constructed where each vertex represents a k-clique. Two nodes in this clique-graph is connected or adjacent if they share (k − 1) members. Each connected component in the clique-graph represents a community . The overall process of the approach is described below in Algorithm 1.

- **Algorithm 1** Clique Percolation Method

Input: The network, G and the clique size, k

Output: Community structure, C

Step 1: All k-cliques present in the network G are identified.

Step 2: A new network, referred as clique-graph, G_C is formed where each node represents an identified clique and two nodes (clique) in the network, G_C are connected by an edge, if they share k − 1 members.

Step 3: Connected components in G_C are identified.

Step 4: Each connected component in G_C represents a community. Set of communities forms the identified community structure for the network, G

For example, the network shown in Figure 3.1 have six 3-cliques which repre-sents individual node in clique graph. Six cliques are,

a:(1, 2, 3); b:(1, 2, 8); c:(2, 4, 5); d:(2, 4, 6); e:(2, 5, 6); f :(4, 5, 6)

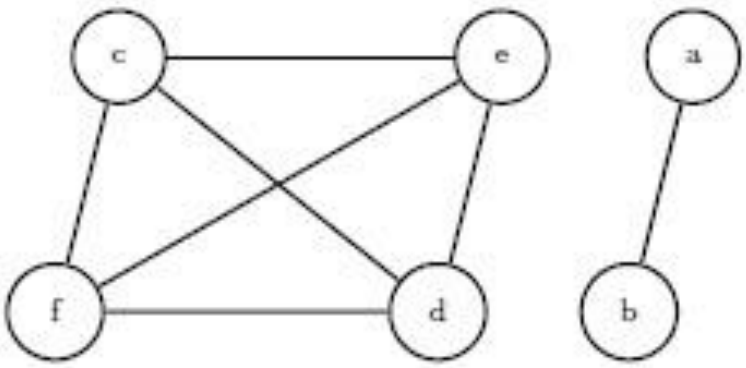


Figure: Corresponding Clique Graph of example network

If two cliques share minimum of 2 nodes (as k=3 here) then they are connected by an edge. Clique a and b share two nodes(node 1 and 2). So these two clique nodes will be connected through an edge. In the same way other cliques also make connection with each other to form the clique graph, G_C as shown in Figure. Connected components in G_C are (a, b) and (c, d, e, f). Connected components represent communities. So, in this case two connected components correspond to two communities which are,

c_1 : (1, 2, 3, 8)

c_2 : (2, 4, 5, 6)

The community structure or cover is C = {c_1, c_2} and node 2 is overlapped between these two communities. Node 7 is not included to any community as it is not a part of any 3-cliques.

➢ Community Similarity:

Community similarity defines how much similar are two communities. There are various ways to measure similarity between two entities. In this paper Jacquard index has been considered as similarity measure. If c_i and c_j are two communities in community structure C then similarity between these two communities is defined as

(4)

$$s = \frac{|c_i \cap c_j|}{|c_i \cap c_j| + |c_i - c_j| + |c_j - c_i|}$$

$|c_i \cap c_j|$ indicate number of common members i.e., members present in both the communities. $|c_i - c_j|$ and $|c_j - c_i|$ indicate the numbers of members present only in one or the other community.

➢ Proposed Method: Extended Clique Percolation Method (ECPM) :

The k-clique method only considers the fully connected subgraphs of size k. It may happen that a node is associated to a community by some edges, but it does not form any clique of size k. So CPM algorithm leads to a community structure which may not include many member of the network, though they are linked with some communities. The proposed approach aims to cover all the connected members of the network by including them to at least one community.

➢ The approach is based on fuzzy assignments of nodes. A node can take part in multiple communities and is associated with each community by some fraction.

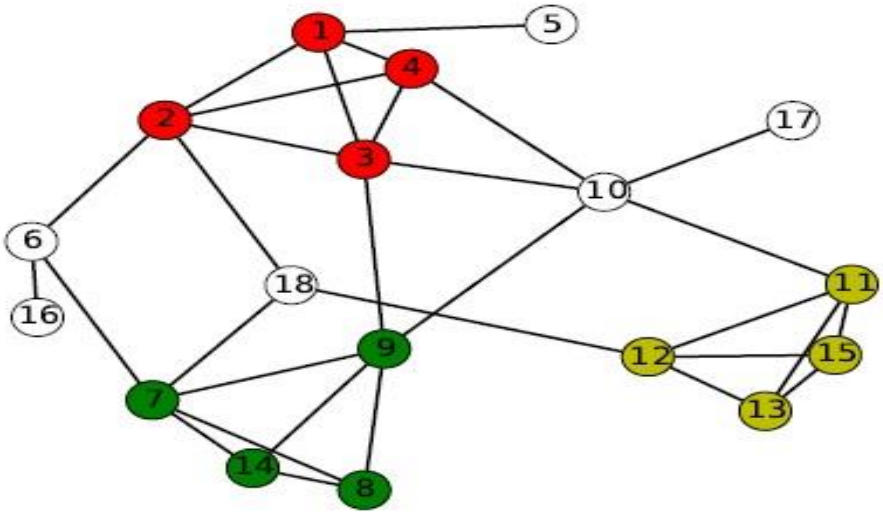


Figure: An example of network showing three initial communities detected by CPM with k=4. Communities are represented by different colors.

This fraction is the belonging co-efficient. The idea is a node can participate in several communities and is associated with each communities with some value. It is assumed that each node can have maximum attachment value of 1 and mini-mum attachment value of 0. And it can be attached with a community with an attachment value between 0 to 1. If the node is not part of any community then attachment values of this node to all the communities is 0. If the node is part of few communities then sum of all the attachment values need to be 1. So if there are two communities in a cover and a node is part of both the communities then, a node can have attachment values to two communities as 0.2 and 0.8 as (0.8+0.2=1) but can not be 0.3 and 0.8 (as 0.3+0.8 6= 1). Here this attachment value is described as belonging co-efficient. So one more constraint for belonging co-efficient is

$$\sum_{c \in C} B(i,c) = 1 \qquad (5)$$

where c indicates the communities in cover C and i indicate a particular node.

The method proposed here comprises of three major components which are,

a. **Finding the initial community structure:** The initial communities are detected for networks using CPM for the given value of k, the clique size.

b. **Updating the initial communities:** If initial community structure includes all the members in the network into some community then there is no need of this step. When some nodes are not included in any of the initial communities because of not forming a clique of size k in the network, this step includes those left out nodes into at least one community. Belonging coefficient is used to decide the deserving membership of left out nodes.

C. **Merging similar communities:** When all nodes are included into some community, it may be possible that two or more communities are very much similar. If two or more communities are similar more than a threshold then these communities are merged into a single one. Similarity is measured using Eq. 4.

The overall algorithm has been described in Algorithm 2 where, G is the considered network, k is the clique-size and t is the threshold value for community similarity.

Algorithm 2 Extended Clique Percolation Method

Require: The network G, the clique size, k and the similarity threshold, t

Ensure: Community structure or cover, C

- C ←Compute initial cover using CPM for given k.
- |C|←Number of initial communities.
- {**Updating communities by adding left out nodes**}
- L_f ← Compute list of nodes which does not belong to any initial community.
- **while** L_f 6= ∅ **do**
- **for each** i in L_f **do**
- ∀c ∈ C, Find belonging coefficient , B(i, c).
- Find the maximum B(i, c).
- **if** maximum B(i, c) = 0 **then**
- N = 0
- **else**
- N ← Count the number of B(i, c) with maximum value.
- **end if**

- **if** N = 1 **then**
- Include node, i to the community, c for which B(i, c) is maximum and remove i from L_f . {both L_f and C is updated.}
- **else if** N > 1 **then**
- Include i to all those communities for which B(i, c) is maximum and remove i from L_f . {i is overlapped between multiple communities.}
- **else**
- do nothing
- **end if**
- **end for**
- **end while**
- {**Merging of similar communities**}
- **for** each two communities ∈ C **do**
- Find similarity.
- **if** similarity ≥ threshold, t **then**
- Merge these two communities and update the cover.
- **end if**
- **end for**
- **return** C

For example, in the network shown in Figure 3.3, three communities found initially by CPM with k = 4 are

c_1 : (1, 2, 3, 4)
c_2 : (7, 8, 9, 14)
c_3 : (11, 12, 13, 15)

with Cover, C = {c_1, c_2, c_3} Left out nodes are L_f = (5, 6, 10, 16, 17, 18)

Node	$B(i,c)$		
	c_1	c_2	c_3
5	1	0	0
6	0.5	0.5	0
10	0.5	0.25	0.25
16	0	0	0
17	0	0	0
18	0.33	0.33	0.33

Table 1 : Belonging Coefficients to initial communities

Belonging coefficient of nodes in L_f to three initial communities are given in Table 3.1. Node 5 having maximum B(5, c) with c = c_1 will be included to community c_1. Node 6 has B(6, c_1) = 0.5 and as well as B(6, c_2) = 0.5. So node 6 will be included to both c_1 and c_2. Node 10 shares it's maximum to community c_1, B(10, c_1) = 0.5. So node 10 is included to c_1. In the same way node 18 is included to all the three communities c_1, c_2 and c_3 as all three communities having same belonging coefficient of 0.33. Node 16 and 17 are not included to any community in this step as they are having belonging coefficient 0 to all the three communities. In next step the communities are updated as

c'_1 : (1, 2, 3, 4, 5, 6, 10, 18)
c'_2 : (6, 7, 8, 9, 14, 18)
c'_3 : (11, 12, 13, 15, 18)
and left out nodes are L_f = (16, 17)

Node	$B(i,c)$		
	c'_1	c'_2	c'_3
16	0.5	0.5	0
17	1	0	0

Table 2: Belonging Coefficients to updated communities

In next step again, belonging coefficients are computed for nodes ∈ L_f as given in Table 3.2. So node 16 is included to both community c'_1 and c'_2 and node 17 is included to community c'_1 resulting communities c''_1 : (1, 2, 3, 4, 5, 6, 10, 16, 17, 18)
c''_2 : (6, 7, 8, 9, 14, 16, 18)
c''_3 : (11, 12, 13, 15, 18)

The final cover, C = {c''_1, c''_2, c''_3}, with overlapped nodes as 6, 16 and 18. As all the nodes are covered the algorithm stops here. This algorithm ensures that all connected nodes will be included to at least one community.

➢ Implementation :

Preprocessing of Network :

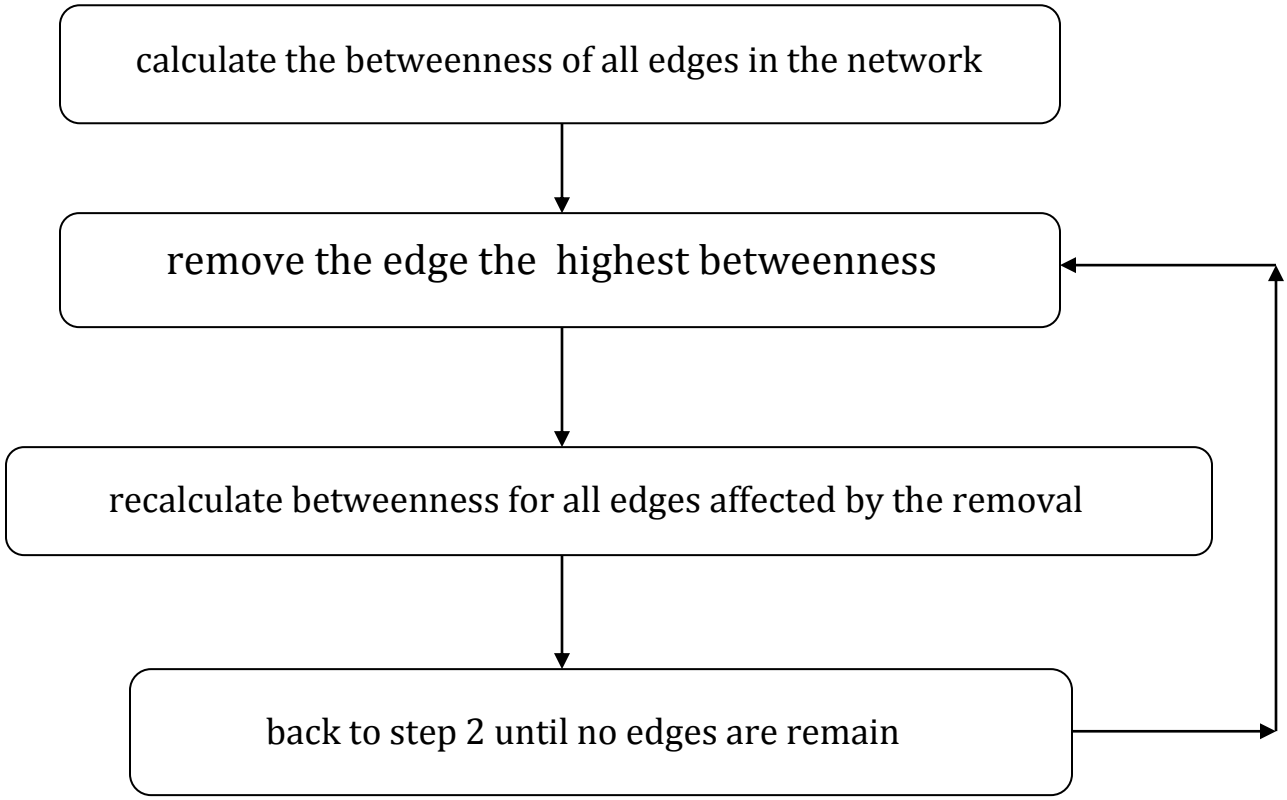
In any real-world network there may have some isolated nodes that is members having no connection. Surely these nodes are not going to be part of any community as they are not connected to any one else in the network. In the proposed method all left out nodes have been considered to include in community structure. For isolated nodes belonging coefficient will always be zero and they will not be included to any community. The algorithm described in Algorithm 2 will stop only when there will be no left out nodes. The algorithm will never stop if isolated nodes are considered as these nodes will always be in left out list. So isolated nodes are removed from the network before searching for communities in the network.

Implementation :

Communities are detected for networks using the proposed approach for different value of k, the clique size. Minimum permitted value for k is 3. In literature it is found that generally value of k ranges from 3 to 6 [5, 20].

Here k value is taken as 3 and 4. Here threshold value, t is taken as 0.8 that is two communities are similar to the extent of 80% will be merged. NetworkX package and Pyplot library of Python language has been used for network manipulation. implementation of algorithms and visualization.

➢ Girvan and Newman Algorithmic Flowchart :



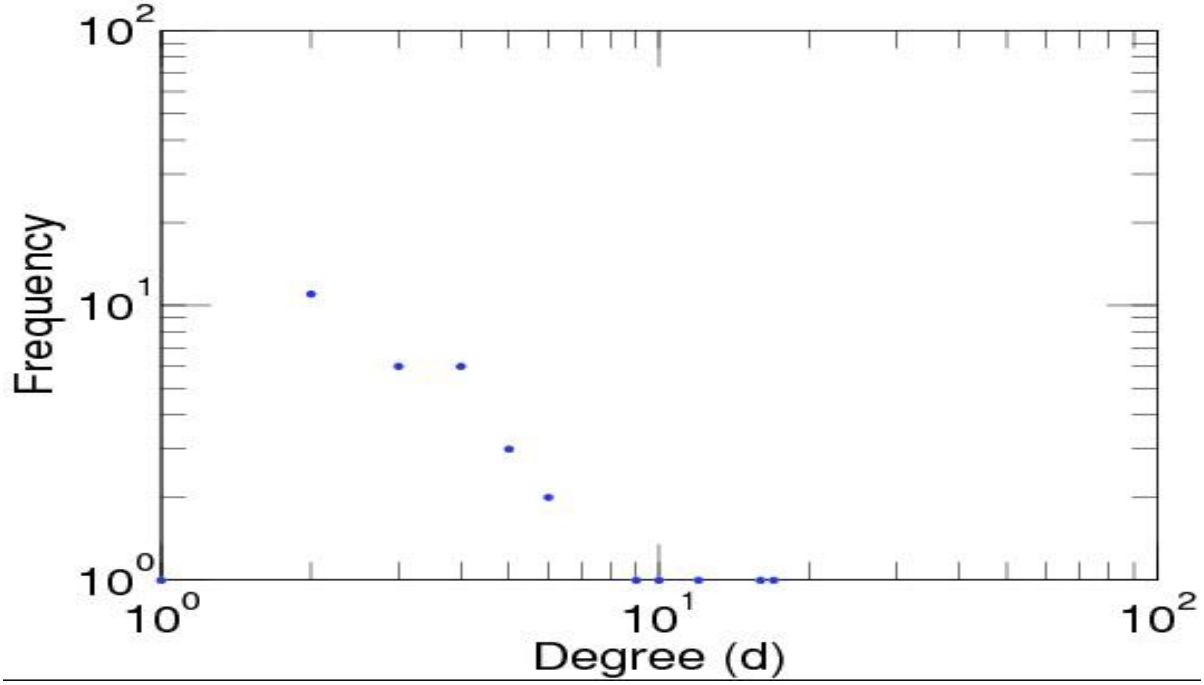
➢ Experimental Result :

Zachary karate club network- The data was collected from the members of a university karate club by Wayne Zachary in 1977. Each node represents a member of the club, and each edge represents a tie between two members of the club. The network is undirected. An often discussed problem using this dataset is to find the two groups of people into which the karate club split after an argument between two teachers.

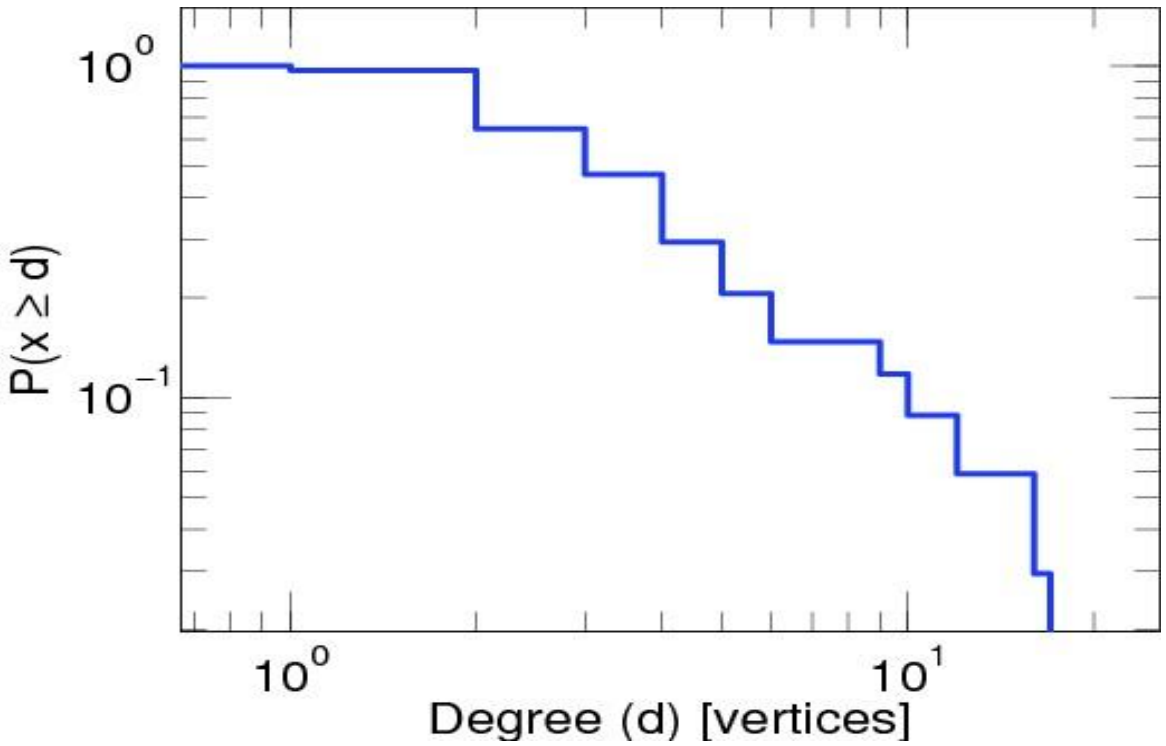
Network Info

Code	**ZA**
Category	HumanSocial
Data source	http://vlado.fmf.uni-lj.si/pub/networks/data/ucinet/ucidata.htm#zachary
Vertex type	Member
Edge type	Tie
Format	Undirected
Edge weights	Unweighted
Size	34 vertices (members)
Volume	78 edges (ties)
Average degree	4.5882 edges / vertex
Fill	0.13904 edges / vertex^2
Maximum degree	17 edges
Size of LCC	34 vertices (network is connected)
Wedge count	528
Claw count	1,764
Triangle count	45
Square count	154
4-tour count	3,500
Power law exponent (estimated) with d_{min}	2.1610 (d_{min} = 2)
Gini coefficient	38.5%
Relative edge distribution entropy	92.5%
Assortativity	–0.47561
Clustering coefficient	25.6%
Diameter	5 edges
90-percentile effective diameter	3.44 edges
Mean shortest path length	2.44 edges
Spectral norm	6.7257
Algebraic connectivity	0.46853

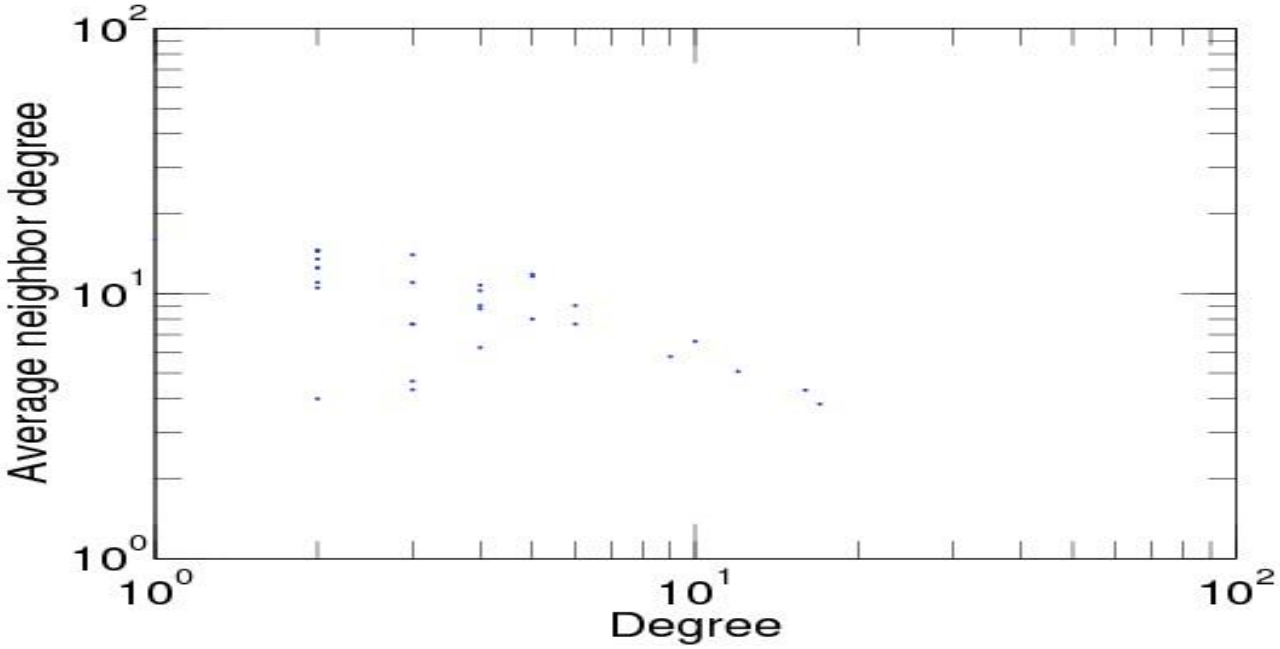
➢ Degree Distribution :



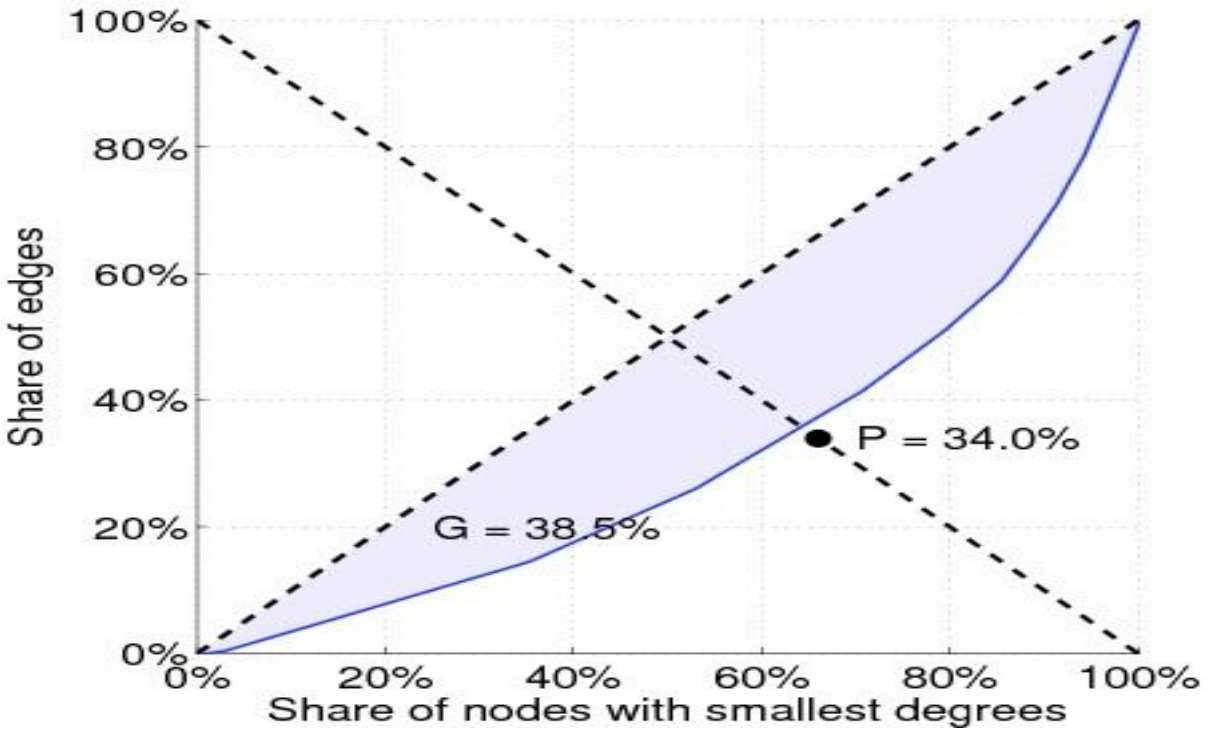
➢ Cumulative Degree distribution :



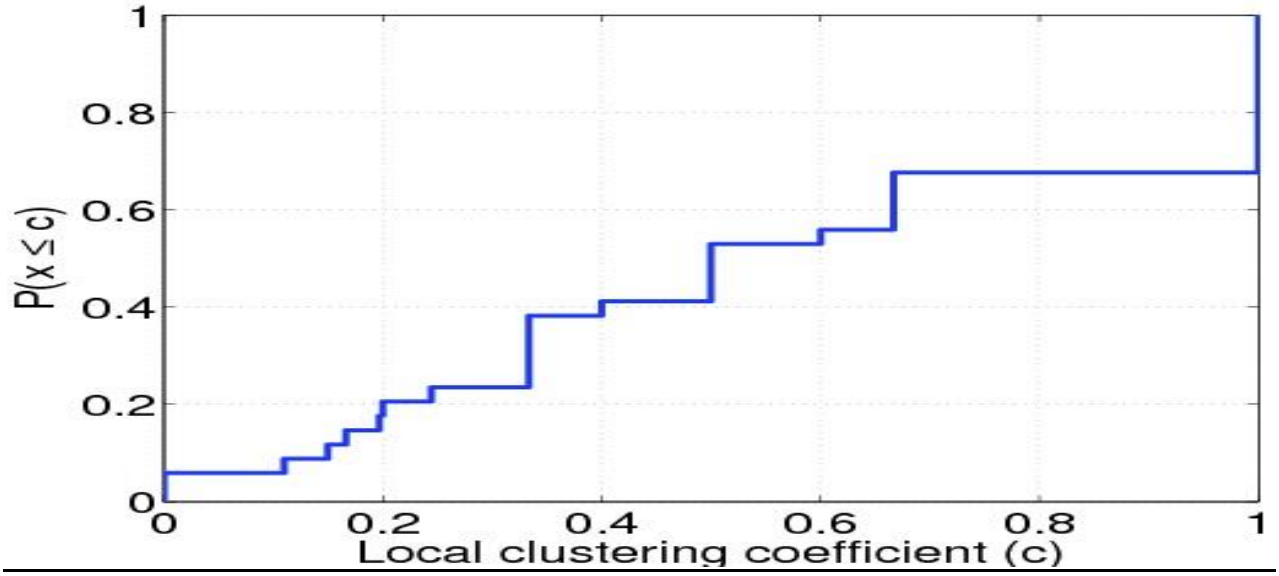
➢ Assortativity Plot :



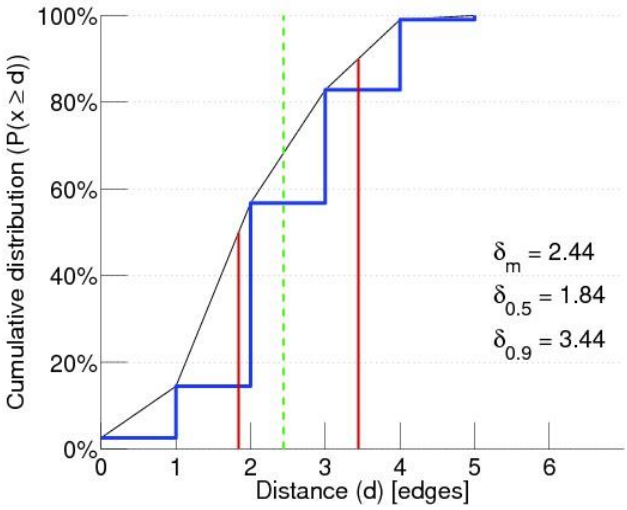
➢ Lorenz Curve :

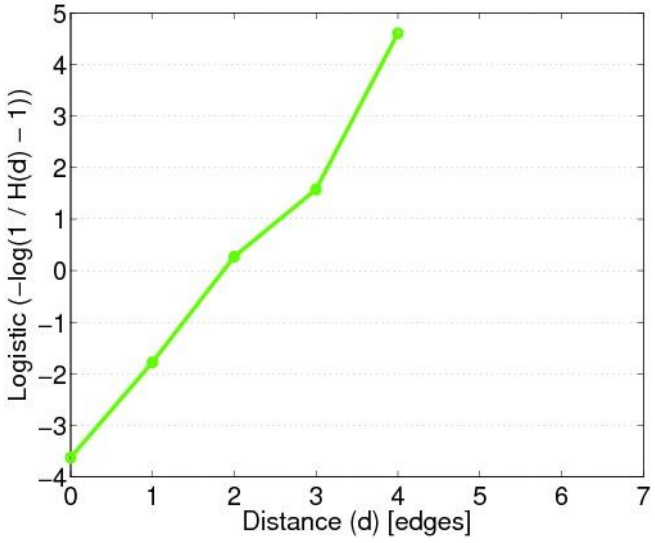


➢ Clustering Coefficient Distribution :

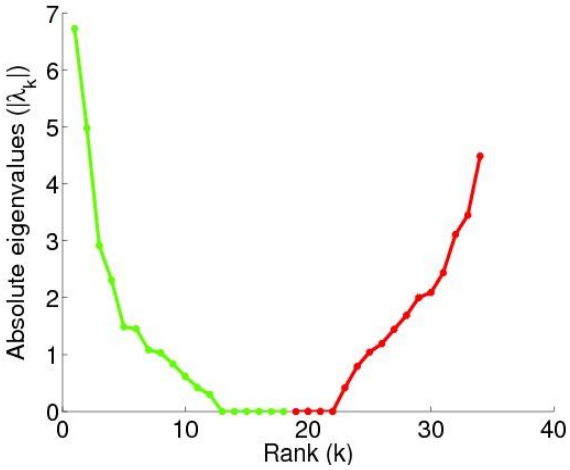


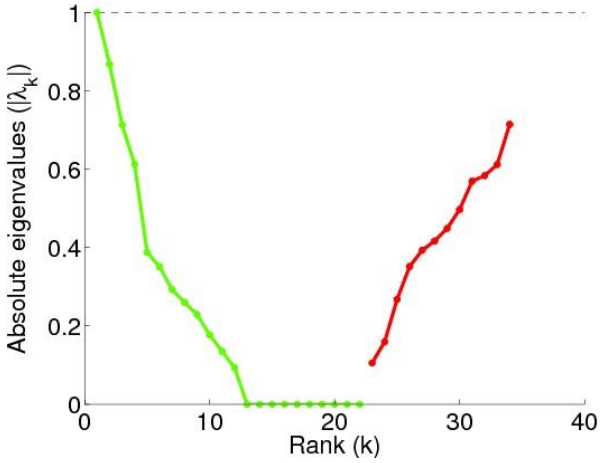
➢ Distance Distribution :

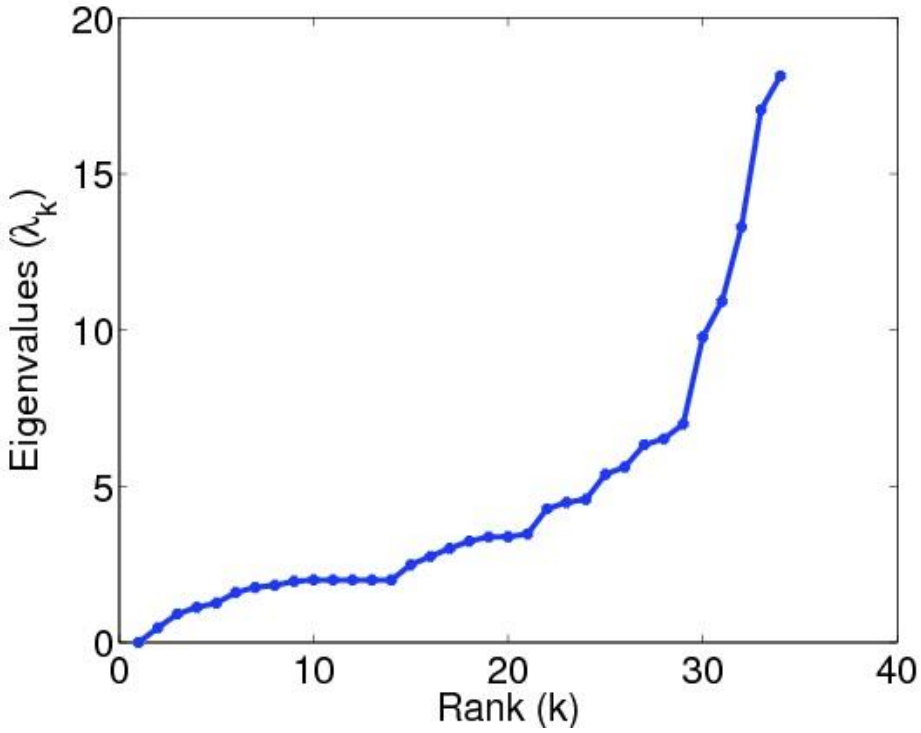




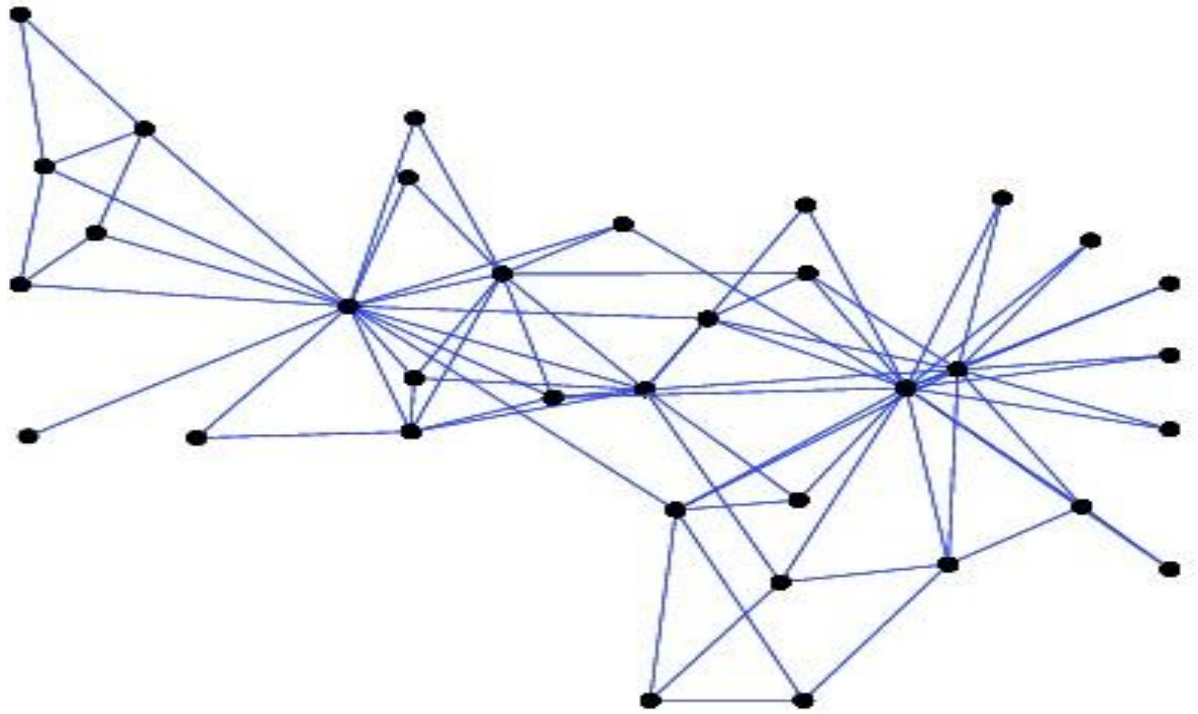
➢ Top-k eigen values :

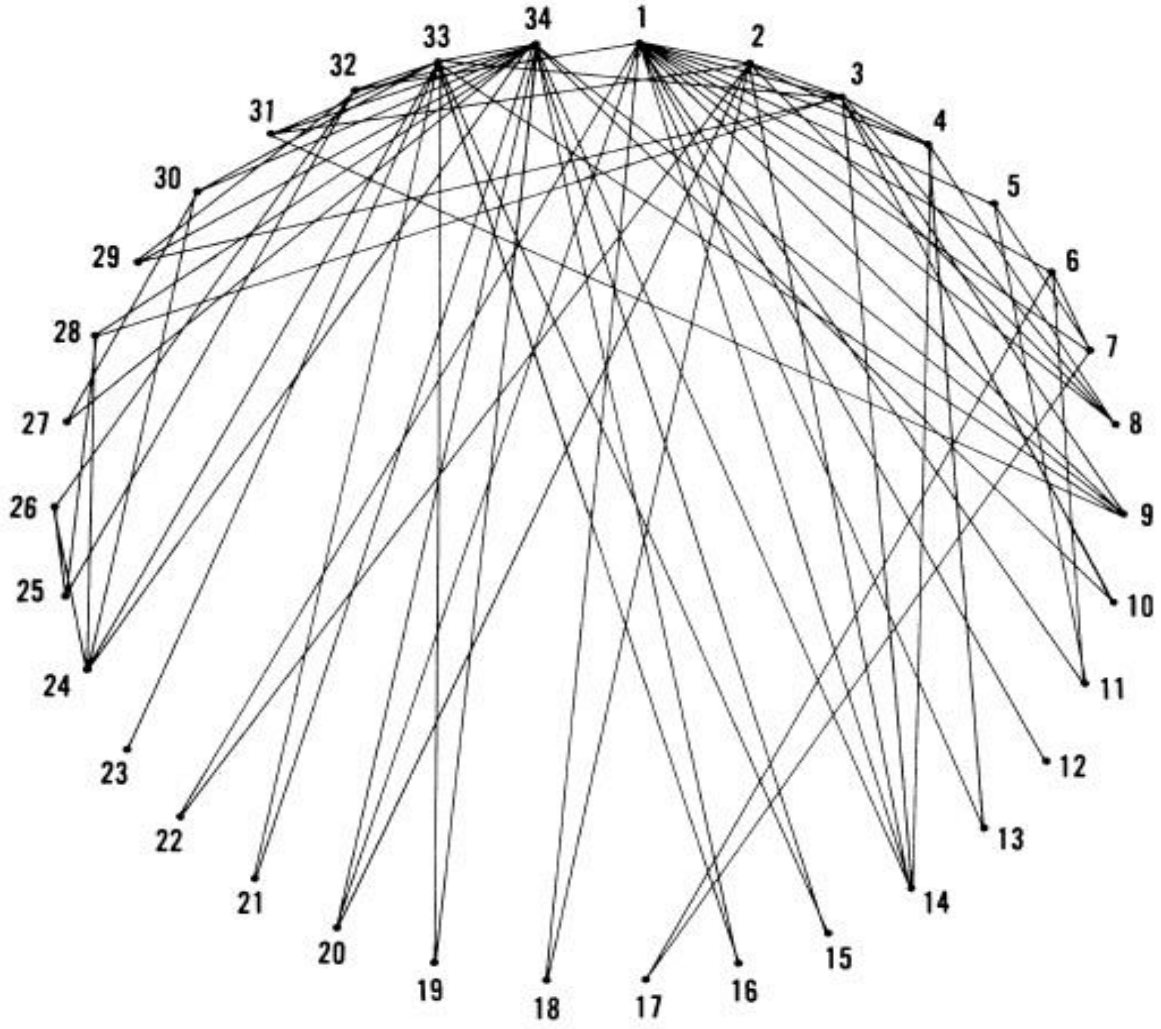






- **Layout :**



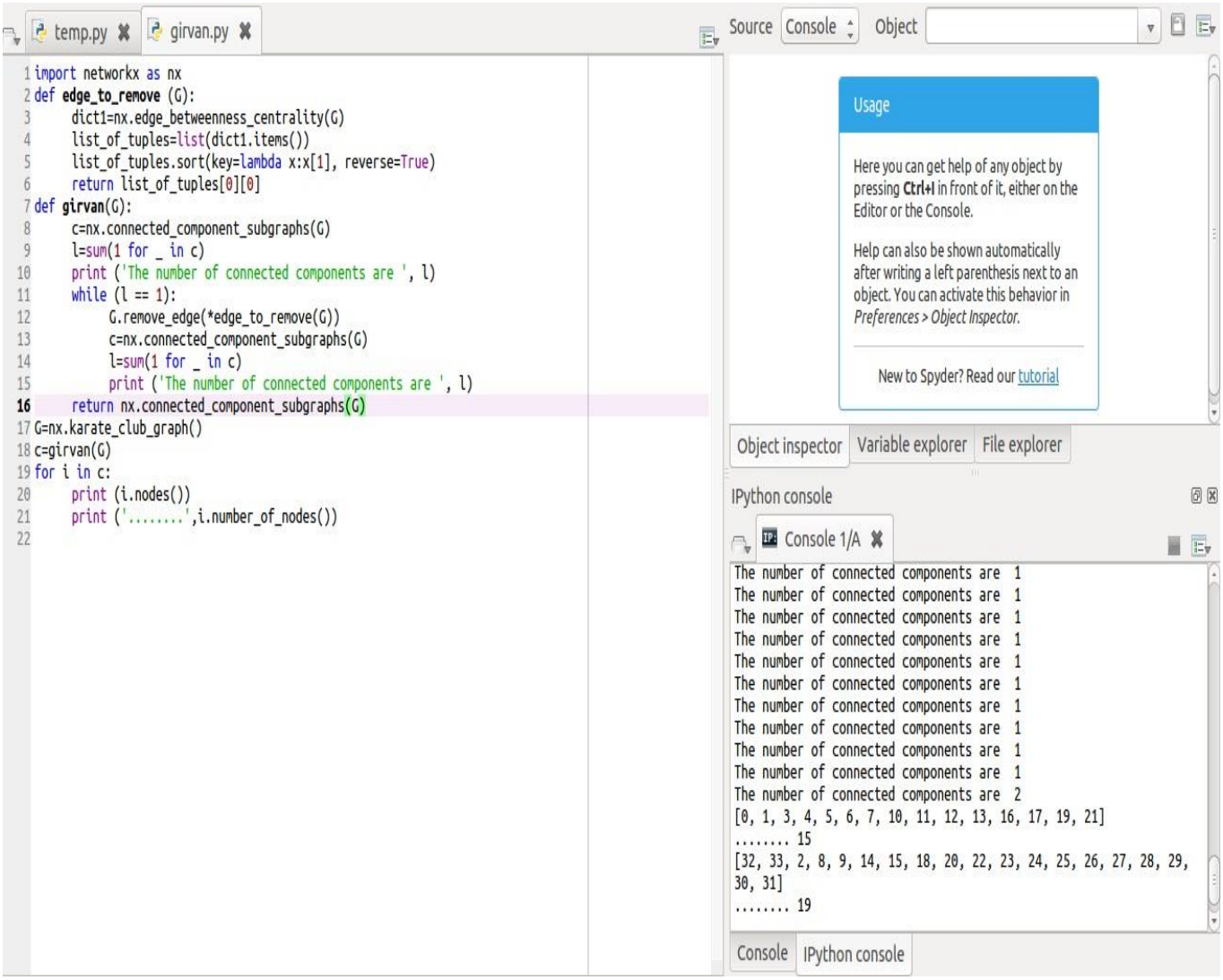


This figure 1 is the graphic representation of the social relationships among the 34 individuals in the karate club. A line is drawn between two points when the two individuals being represented consistently interacted in contexts outside those of karate classes, workouts, and club meetings. Each such line drawn is referred to as an edge.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0
2	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0
3	1	1	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0
4	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	
10	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
11	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
17	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
20	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
22	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
28	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
29	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
31	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
32	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
33	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1	0	1	0	1	1	0	0	0	0	0	1	1	1	0	1
34	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0	1	1	0	0	1	1	1	1	1	1	1	0

This is the matrix representation of the graph shown in Figure 1. The rows and columns represent individuals in the club. An entry, determined by a row/column pair, is valued at 1 if an edge was drawn in Figure 1 between the two individuals represented by the row and column. The entry is valued at 0 otherwise. For example, the entry in the fifth row, seventh column, is a 1, indicating there is an edge existing between individuals 5 and 7. Notice that the matrix is symmetrical--i.e., the seventh row fifth column is also valued at 1, since it identifies the same edge. Because, by definition, no individual can interact with himself, zeroes occupy the main diagonal (from row/column 1 to row/column 34). Later, this matrix will be referred to by the symbol E.

- **Output :**




```
import networkx as nx
def edge_to_remove (G):
    dict1=nx.edge_betweenness_centrality(G)
    list_of_tuples=list(dict1.items())
    list_of_tuples.sort(key=lambda x:x[1], reverse=True)
    return list_of_tuples[0][0]
def girvan(G):
    c=nx.connected_component_subgraphs(G)
    l=sum(1 for _ in c)
    print ('The number of connected components are ', l)
    while (l == 1):
        G.remove_edge(*edge_to_remove(G))
        c=nx.connected_component_subgraphs(G)
        l=sum(1 for _ in c)
        print ('The number of connected components are ', l)
    return nx.connected_component_subgraphs(G)
G=nx.karate_club_graph()
c=girvan(G)
for i in c:
    print (i.nodes())
    print ('........',i.number_of_nodes())
```

```
The number of connected components are  1
The number of connected components are  1
The number of connected components are  1
The number of connected components are  1
The number of connected components are  1
The number of connected components are  1
The number of connected components are  1
The number of connected components are  1
The number of connected components are  1
The number of connected components are  1
The number of connected components are  2
[0, 1, 3, 4, 5, 6, 7, 10, 11, 12, 13, 16, 17, 19, 21]
........ 15
[32, 33, 2, 8, 9, 14, 15, 18, 20, 22, 23, 24, 25, 26, 27, 28, 29,
30, 31]
........ 19
```

➢ Conclusion and Future Work

Conclusion :

In this project, I have presented an organized view on subgroup and community detection on attributed graphs. Specifically, we described subgroup discovery for compositional network analysis concerning properties of the actors, with extensions to the analysis of complex target concepts like correlations between a set of variables, or dense subgraphs . Then, this directly extends to community detection on attributed graphs. Here, I started with an introduction of methods for community detection, continuing on methods for mining overlapping communities, to approaches that target descriptions leveraging structural and compositional attribute information. In particular, we summarized the Girvan and Newman algorithm that combines subgroup discovery and community detection, resulting in a description-oriented approach for community detection. Here it is concluded by this project is that subgroup discovery and community detection enable the identification of sub group at different levels and dimensions such as compositional dimension, structural and compositional dimension and providing explicit description. These are all can be combined for obtaining descriptive community patterns according to standard community quality function. Efficient tools are require for detection and analyze.

Future Work :

community detection is still a challenge. Though there are several proposed methods, but most of them take a huge amount of processing time. So emphasis should be given to effective algorithms which will be able to detect communities in a huge social network in allowable time. In this work only unweighted and undirected networks has been taken into consideration. In future weighted and directed networks are needed to be considered for community detection. Now days almost all social networks are dynamic that is some members are joining and some are leaving every moment. So it will be great if communities can be detected in dynamic networks. This project can take challenges about using ubiquitous and social data like heterogeneous data and complex network. It can work on integration of multiples network and temporal information. It can support for integration and analysis. The necessary thing is that efficient methods and tools for the mining of such data.

➢ Reference :

- Lancichinetti, S. Fortunato, and J. Kert´esz, "Detecting the overlapping and hierarchical community structure in complex networks," New Journal of Physics, vol. 11, no. 3, p. 033015, 2009.
- H. Shen, X. Cheng, K. Cai, and M.-B. Hu, "Detect overlapping and hierar-chical community structure in networks," Physica A: Statistical Mechanics and its Applications, vol. 388, no. 8, pp. 1706–1712, 2009
- M. E. Newman, "Finding community structure in networks using the eigen-vectors of matrices," Physical review E, vol. 74, no. 3, p. 036104, 2006
- M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," Physical review E, vol. 69, no. 2, p. 026113, 2004
- M. E. Newman, "Modularity and community structure in networks," Proceed-ings of the National Academy of Sciences, vol. 103, no. 23, pp. 8577–8582, 2006.
- D. E. Knuth, The Stanford GraphBase: a platform for combinatorial comput-ing, vol. 37. Addison-Wesley Reading, 1993.
- D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations," Behavioral Ecology and Socio-biology, vol. 54, no. 4, pp. 396–405, 2003.
- W. Zachary, "An information flow modelfor conflict and fission in small groups1," Journal of anthropological research, vol. 33, no. 4, pp. 452–473, 1977.
- B. Dickinson, B. Valyou, and W. Hu, "A genetic algorithm for identifying overlapping communities in social networks using an optimized search space," Social Networking, vol. 2, p. 193, 2013.
- Y. Cai, C. Shi, Y. Dong, Q. Ke, and B. Wu, "A novel genetic algorithm for overlapping community detection," in Advanced Data Mining and Applica-tions, pp. 97–108, Springer, 2011.

- J. Travers and S. Milgram, “An experimental study of the small world prob-lem,” Sociometry, vol. 32, no. 4, pp. 425–443, 1969.
- C. Pizzuti, “Overlapped community detection in complex networks.,” in GECCO, vol. 9, pp. 859–866, 2009.

M. Scholz, “Node similarity as a basic principle behind connectivity in com-plex networks,” arXiv preprint arXiv:1010.0803, 2010.

- C. Pizzuti, “Ga-net: A genetic algorithm for community detection in social networks,” in Parallel Problem Solving from Nature–PPSN X, pp. 1081–1090, Springer, 2008
- S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, “Complex networks: Structure and dynamics,” Physics reports, vol. 424, no. 4, pp. 175– 308, 2006.
- B. T´oth, T. Vicsek, and G. Palla, “Overlapping modularity at the critical point of k-clique percolation,” Journal of statistical physics, vol. 151, no. 3-4, pp. 689–706, 2013.
- M. E. Newman, “The structure and function of complex networks,” SIAM review, vol. 45, no. 2, pp. 167–256, 2003.
- M. Coscia, G. Rossetti, F. Giannotti, and D. Pedreschi, “Demon: a local-first discovery method for overlapping communities,” in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 615–623, ACM, 2012.
- S. Engen, V. Grøtan, and B.-E. Sæther, “Estimating similarity of communi-ties: a parametric approach to spatio-temporal analysis of species diversity,” Ecography, vol. 34, no. 2, pp. 220–231, 2011.
- L´az´ar, D. Abel, and T. Vicsek, “Modularity measure of networks with over-lapping communities,” EPL (Europhysics Letters), vol. 90, no. 1, p. 18001, 2010.

- V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri, “Extending the defini-tion of modularity to directed graphs with overlapping communities,” Journal of Statistical Mechanics: Theory and Experiment, vol. 2009, no. 03, p. P03024, 2009.
- D. Chen, M. Shang, Z. Lv, and Y. Fu, “Detecting overlapping communities of weighted networks via a local algorithm,” Physica A: Statistical Mechanics and its Applications, vol. 389, no. 19, pp. 4177–4187, 2010.
- http://konect.uni-koblenz.de/networks/ucidata-zachary
- M. Van Steen, Graph Theory and Complex Networks: An Introduction. Maarten van Steen, 2010.
- Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, “Link communities reveal multi-scale complexity in networks,” Nature, vol. 466, no. 7307, pp. 761–764, 2010.