VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Social Network Analytics Lab

Digital Assignment -2

Report of the Assignment

Register Number: 22MCB0033

Name: Sathwik Shetty B

Course Name: Social Network Analytics Lab

Course code: MCSE618P

Topic: Community Detection Algorithm

Abstract:

Understanding the structure and dynamics of social networks requires community detection. It entails locating nodes within a network that have better connectivity between themselves than with other nodes. This project report investigates several community detection algorithms and social network analytics uses for them. We provide a thorough review of the topic by talking about the theoretical underpinnings, implementation specifics, and evaluation metrics for these algorithms. We also provide a case study that shows how community detection algorithms can be used to analyse social networks in the real world.

Introduction to Community Detection Algorithms:

1) Background and motivation

2) A description of community detection and its significance

3) Social network analytics overview

The underlying communities or clusters inside a network can be found and revealed using community detection algorithms. These algorithms divide the nodes into groups that have a higher density of connections inside them by analysing the topology and connectivity patterns of the network.

the links between the groups and the groups themselves. The following list of frequently used community detection algorithms:

1. Girvan-Newman Algorithm: This algorithm evaluates the number of shortest paths that cross through each edge in the network using the concept of edge betweenness. The highest betweenness edges are gradually removed until the network is divided into discrete communities.

2. Louvain Method: As previously mentioned, the Louvain method maximises modularity by iteratively relocating nodes between communities. It is renowned for its scalability and quickness.

3. Infomap: Infomap uses the information-theoretical idea to identify communities. In order to reduce the description of the random walker's trajectory, network navigation is treated as a random walk. Based on the best network structure compression, it assigns nodes to communities.

4. Label Propagation approach (LPA): This straightforward and effective approach assigns a distinct label to each node from the beginning. Nodes adopt the label that appears the most frequently among their neighbours when they update their labels repeatedly. This procedure is repeated until the labelling is stable.

5. Spectral Clustering: To locate communities, spectral clustering approaches make use of the eigenvectors or spectral characteristics of the network's adjacency matrix or Laplacian matrix. Traditional clustering algorithms like k-means or hierarchical clustering are then used after the network is embedded in a lower-dimensional environment.

6. Walktrap Algorithm: Walktrap investigates the hypothesis that random walks inside communities are more likely to stay inside the community than to cross community boundaries. Using their random walk trajectories as a basis, it calculates the pairwise similarity between nodes and employs hierarchical clustering to find communities.

7. Quickly ruthless algorithm By greedily merging communities based on the gain in modularity attained by the merge, the fast greedy algorithm, also known as the Clauset-Newman-Moore (CNM) method, optimises modularity. Communities are merged repeatedly until there is no more room for modularity enhancement.

These are merely a few illustrations of community detection algorithms; there are many more iterations and additions to choose from. The network's unique properties and the demands of the analysis will determine which algorithm is used. Depending on variables like network size, density, modularity, and the presence of overlapping communities, various methods may perform better or worse.

Traditional Approaches:

1) Girvan-Newman algorithm: A hierarchical algorithm based on edge-betweenness.

2) The Clauset-Newman-Moore algorithm: an algorithm for optimising modularity.

3) The edge-betweenness-based Newman-Girvan method is a dividing algorithm.

Modularity-Based Algorithms:

1) The Louvain algorithm, a quick and scalable approach for modularity optimisation.

2) Infomap algorithm: An information-theoretic measure-optimization flow-based algorithm.

3) Leading eigenvector algorithm, third option: spectral clustering strategy that maximises modularity.

4) Walktrap algorithm: An algorithm that measures network similarity based on random walks.

The Louvain Algorithm

Another well-liked approach for community detection in network analysis is the Louvain algorithm, commonly referred to as the Louvain method or Louvain algorithm. It was created by Vincent Blondel et al. in 2008 at the University of Louvain in Belgium, hence its name.

Similar to the greedy modularity algorithm, the Louvain algorithm optimises modularity in two steps. To achieve community detection, it adopts a different method.

Each node is first assigned to its own community in the algorithm's first stage. The method then goes over each node iteratively, evaluating the modularity gain as it moves each node to its neighbouring communities one at a time. If the gain in modularity is positive, the node is transferred to the community where the gain was greatest. Moving nodes across communities is kept going until no further gains in modularity can be made.

The algorithm constructs a new network in which the communities are represented as nodes in the second phase by treating each community obtained from the first stage as a single node. The sum of the weights of the original edges connecting the nodes in the appropriate communities determines the weights of the edges between the new nodes. A higher level of granularity in the detection of communities is made possible by this phase.

The first and second phases are continued until a maximum level of modularity is reached, at which point combining communities will no longer result in an increase in modularity.

Large-scale networks can benefit from the speed and scalability of the Louvain method. It has been extensively used in many different fields, including web graph analysis, biological networks, and social network analysis, among others. In several kinds of networks, the technique has shown good performance in identifying communities with a high degree of modularity.

Overlapping Community Detection:

1) Identifying overlapping communities using the clique percolation method.

2) The COPRA algorithm, which uses label propagation to identify overlapping communities.

3) Speaker-listener labelling algorithm: based on structural characteristics, detecting overlapping communities.

Greedy Modularity Algorithms

A well-liked approach for community discovery in network analysis is greedy modularity. It is founded on the idea of modularity, which evaluates how well a network is divided into communities.

The goal of community discovery is to locate groups or clusters of nodes in a network that have more connections between them than between other nodes outside the group.

Each network node in the greedy modularity method is first assigned to a distinct community. Using the increased modularity that results from the merging, it then iteratively combines communities. The method keeps merging communities until a certain level of modularity cannot be increased any longer. The detected communities in the network are represented by the resulting communities.

By comparing the actual number of edges inside communities to the anticipated number of edges if the network were randomly connected, the modularity of a community partition is determined. A better community division is indicated by a higher modularity score.

The greedy modularity algorithm has become more well-liked as a result of its effectiveness and capacity for handling massive networks. It is important to keep in mind that this technique uses heuristics and might not always locate the modularity that is optimal globally. As a result, different algorithmic iterations may result in marginally different community splits.

Overall, the greedy modularity algorithm is a commonly used method for detecting communities and has been employed in a number of industries, including social network analysis, biological networks, and recommendation engines.

Hierarchical Algorithms:

1) Hierarchical Link Clustering (HLC): Using link similarity, a hierarchy of clusters is built.

2) The iterative Markov Clustering Algorithm (MCL), which is based on random walks.

3) The BigClam algorithm, a non-negative matrix factorization-based hierarchical method.

Network Embedding-Based Approaches:

1) The DeepWalk algorithm, which trains node embeddings by means of random walks.

2) The Node2Vec technique employs biassed random walks to capture structural similarity.

3) The GraphSAGE algorithm aggregates node embeddings using a framework for neighbourhood sampling.

Evaluation Metrics:

1) Modularity, adjusted Rand index (ARI), and normalised mutual information (NMI).

2) Conductance, coverage, and F1-score for assessing communities that overlap.

3) Dendrogram-based metrics and the cophenetic correlation coefficient are examples of hierarchical clustering measures.

Modularity

Modularity is a metric used to assess how well a network is divided into communities or clusters. It measures how much a community's nodes are more closely related to one another than they are to nodes in other communities. In other words, it assesses how strong a network's community structure is.

By comparing the actual number of edges inside communities to the anticipated number of edges if the network were randomly connected, the modularity of a network partition is determined. A better community division is indicated by a higher modularity score.

Modularity (Q) is mathematically defined as: $Q = (1 / (2m)) * \Sigma[i,j] [A[i,j] - (k[i] * k[j]) / (2m)] * \delta(c[i], c[j])$ The weight of the edge between nodes i and j is represented by A[i,j] in the expression (c[i], c[j]).

• The degrees (total edge weights) of nodes i and j are represented by the letters k[i] and k[j], respectively.

• The weights of all the network's edges are added to form the number m.

• The community assignments for nodes i and j are represented by c[i] and c[j], respectively.

• If nodes i and j are in the same community, then (c[i], c[j]) equals 1, else it equals 0.

A number close to 1 suggests a strong community structure, whereas a score close to 0 or negative values suggest a poor or random community structure. The modularity score has a range of -1 to 1.

The aim of community discovery methods is frequently to maximise modularity. These algorithms seek to maximise the number of links within communities while minimising the connections between communities by optimising modularity.

It's important to remember that modularity is just one of several metrics used to gauge how well community discovery algorithms are performing, and it has its limitations. Modularity-based algorithms can experience resolution limits, which causes them to miss small or closely knit communities. Additionally, while modularity might not fully capture all facets of community structure, it may be necessary to combine it with other metrics like conductance, coverage, or silhouette scores to produce a more thorough analysis.

Applications of Community Detection:

1) Social structure analysis and social network analysis.

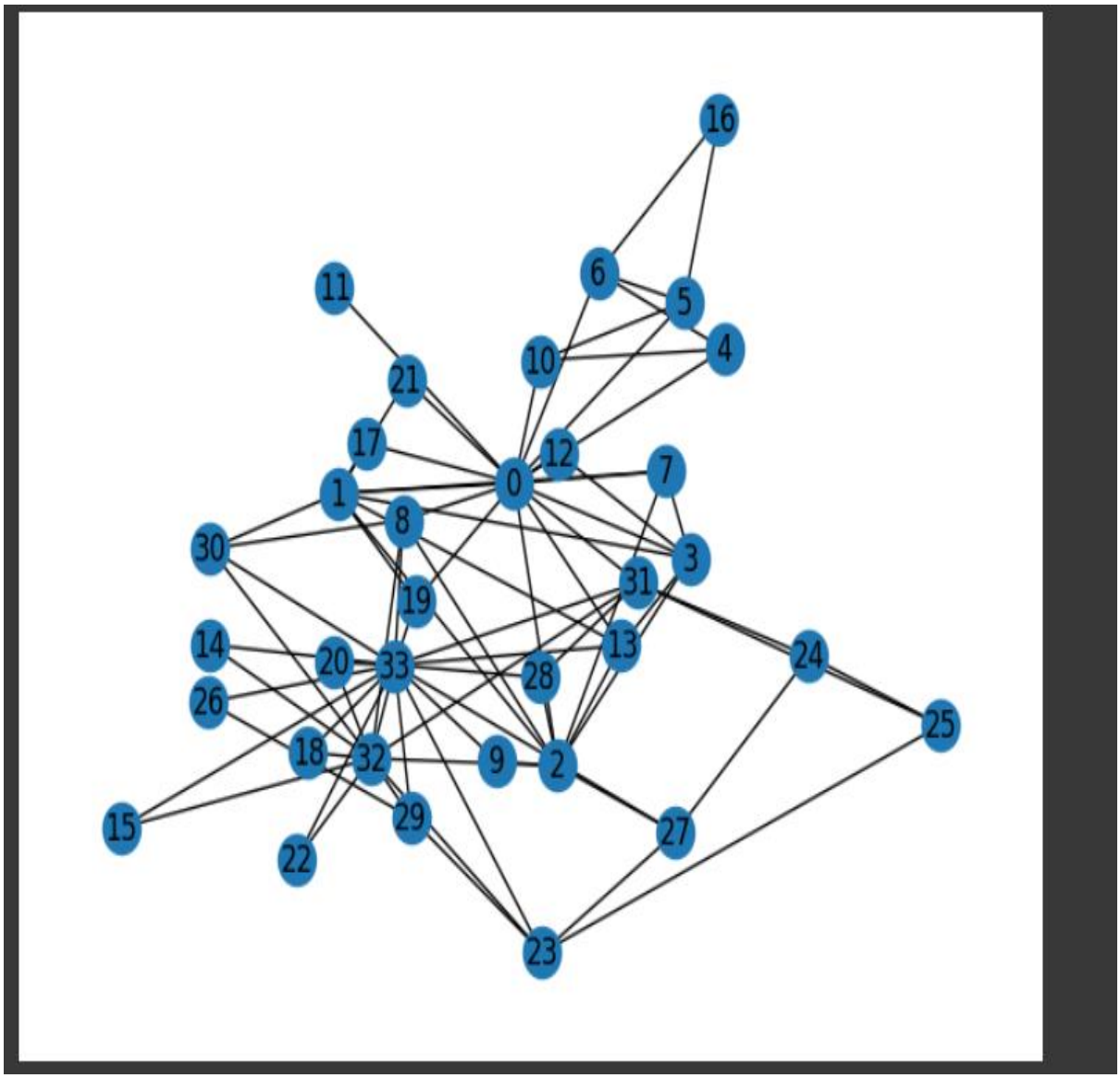
2) Personalised suggestions and recommender systems.

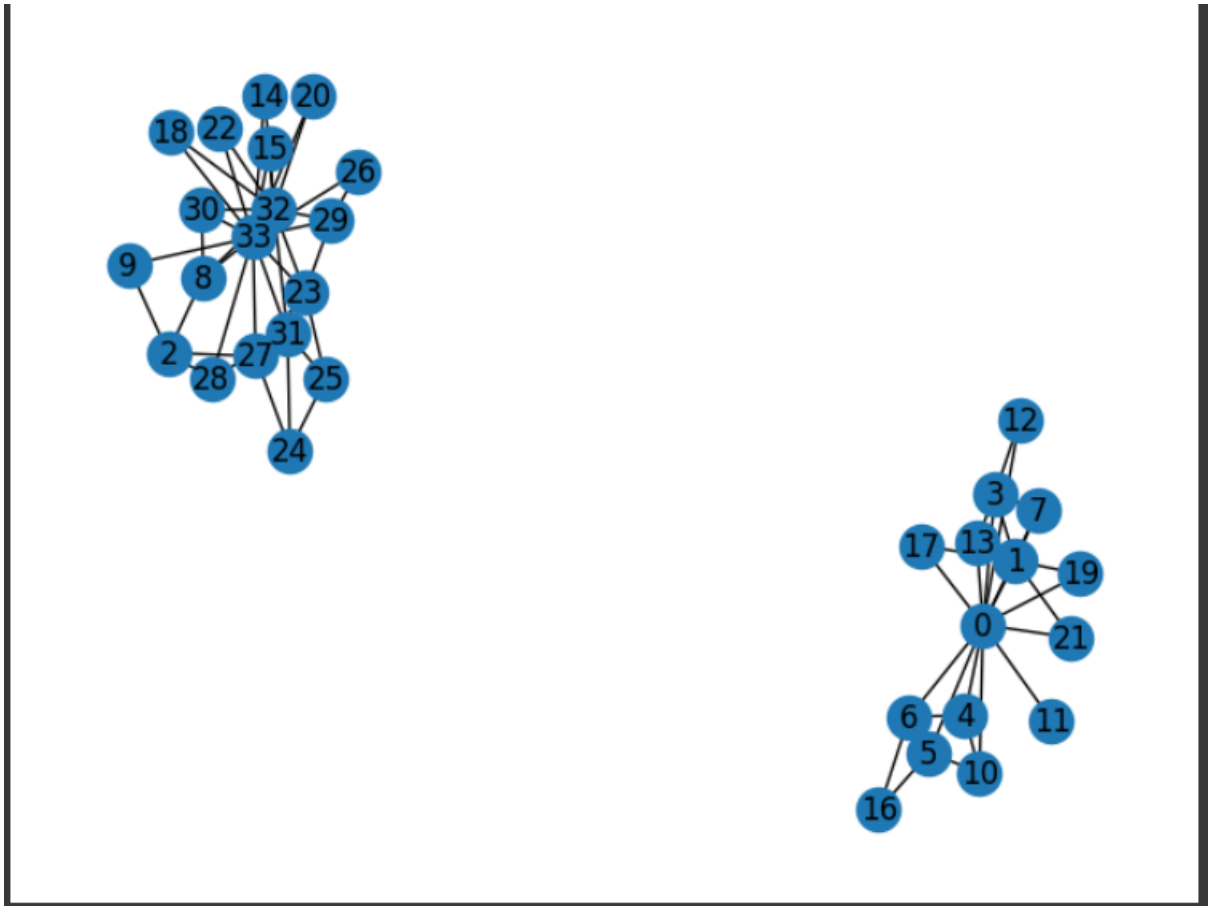
3) Viral marketing and information dissemination techniques.

4) Spotting prominent users and key opinion figures.

5. Fraud and anomaly detection in social networks.

Results:





Conclusion:

Selecting appropriate algorithms based on network characteristics and specific application requirements plays a vital role in the success of effective and accurate community detection.