A Survey on Community Detection in Social Network using Genetic Algorithm

[1]Vishakha V. Patel, [2]Kamal Sutaria
[1]PG Student, [2]Assistant Professor,
[1] Department of computer engineering,
[2]V.V.P. Engineering College, Rajkot, Gujarat, India

***Abstract* - Community detection has been an important research topic in complex networks.Community detection can be viewed as an optimization problem. Complex network can be easily represented in graph. Graph based data mining, is the extraction of novel and useful knowledge from a graph representation of data .There have been many algorithms proposed so far to detect community structures in complex networks, where most of the algorithms are not suitable for very large networks because of their time complexity. Genetic Algorithm is an effective optimization technique to solve the community detection problem in Social network. Social network is one type of Complex network. Social networks are defined as a set of actors and relationships that represent the interactions between entities in the network. To detect a Community in social network is main challenge.**

***Key words* - Community detection, Social network, Graph mining, Genetic algorithm**

I. INTRODUCTION

Community structure identification has created a great interest among physics and computer society who are focusing on the properties of complex networks like the Internet, social networks, citation networks, food networks, e-mail networks and biochemical networks. A complex network is a representation of a complex system from real life in terms of nodes and edges, where a node is an individual member in the system and an edge is a link between nodes according to a relation in the system [1].

Community structure, which is a property of complex networks, can be described as the gathering of vertices into groups such that there is a higher density of edges within groups than between them [2]. From the definition, the nodes in a community should have more intra-community connections rather than inter-community connections. There has been many methods and algorithms proposed so far to reveal the underlying community structure in complex networks [3].

II. II. COMMUNITY DETECTION

A network is said to have community structure if the nodes of the network can be easily grouped into sets of nodes. Network divides naturally into groups of nodes with dense connections internally and sparser connections between groups[4].

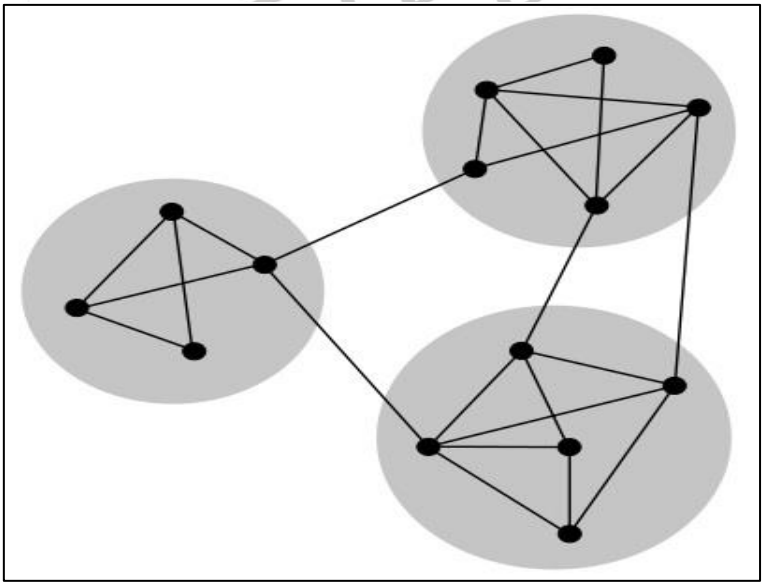

Fig 1 Community detection [4]

Community detection methods can be divided into 4 categories: [4]

- **Node-Centric Community:** Each Node in a group satisfies certain properties
- **Group-Centric Community:** Consider the connections within a group as a whole.
- **Network-Centric Community:** Partition the whole network into several disjoint sets

- **Hierarchy-Centric Community:** Construct a hierarchical structure of communities

III. SOCIAL NETWORK

Many real-world complex systems can be represented as complex networks. Networks could be modelled as graphs, where nodes represent the objects and edges represent the interactions among these objects [5].Social networks are defined as a set of actors and relationships that represent the interactions between entities in the network. A social network is a social structure of people, related (directly or indirectly) to each other through a common relation or interest [6]. Types of Social network are below [10]:

1. **Signed social network -** Signed network include both positive and negative links. In such networks, positive edge represent positive relationships, such as friendship and negative edge represent negative relations, such as hostile relations.
2. **Static social network -** Graph partitioning: The problem of graph partitioning is dividing the vertices into a number of groups with pre-defined size.
3. **Dynamic social network -** In many social networks, activities and interactions between entities change over time. Many methods have been presented with the assumption of a static network structure leading to a lack of consideration of the evolution and dynamics of networks.
4. **Directed social network -** Generally, identifying communities in non-directional networks have been conducted. However, the networks in a number of fields of application are directed. Such as networks of web pages, research papers and Twitter users.

Two types of groups in Social media [7]:
Explicit Groups: Formed by user subscriptions
Implicit Groups: Implicitly formed by social interactions
Social network analysis (SNA) is the study of social networks to understand their structure and behavior. SNA is based on an assumption of the importance of relationship among interacting units. Analyzing large networks, it gives observation that large real-world networks, independently of the domain, satisfy a number of statistical regularities [7].

IV. GRAPH MINING

Graphs are being increasingly used to model a wide range of scientific data. The primary goal of data mining is to extract statistically significant and useful knowledge from data. The data of interest can take many forms: vectors, tables, texts, images, and so on. Structured data and semi-structured data are naturally suited to graph representations. Graph mining is an important research area within the domain of data mining [8].

A graph is a set of nodes and links (vertices and edges), where the nodes and links can have arbitrary labels, and the links can be directed or undirected. Mining graph data is called graph based data mining, is the extraction of novel and useful knowledge from a graph representation of data. Graph Mining is essentially the problem of discovering repetitive subgraphs occurring in the input graphs. Graphs become increasingly important in modelling complicated structures, such as circuits, images, chemical compounds, protein structures, biological networks, social networks, the Web, workflows, and XML documents[9]

V. GENETIC ALGORITHM

Genetic algorithm first proposed in is an optimization method in artificial intelligence. It is a practical method especially when the solution space of a problem is very large and an exhaustive search for the exact solution is impractical. Each member in solution set, which is called a chromosome, represents a possible solution to the problem and the algorithm tries to find the best fitting solution member. In order to improve the quality of the solution members the algorithm uses genetic operations on possible solution members for a predefined number of iterations. The algorithm randomly initializes the chromosomes at the beginning. Then for a number of iterations, it uses a fitness function to assign a fitness value to each solution member, which shows how good a solution member is to solve the problem[12].

1. **Cross-over:** Cross-over operation in genetic algorithm is done by selecting two chromosomes according to their fitness values. Then a cross-over point in chromosome is selected, all the genes after that selection point is exchanged between chromosomes [12].
2. **Mutations:** After performing a number of cross-overs, perform mutation to some number of randomly selected chromosomes. In mutation function, a node is placed into a random community in the network[12].

Optimization

Optimization is the mechanism by which one finds the maximum or minimum value of a function or process. This mechanism is used in fields such as physics, chemistry, economics, and engineering where the goal is to maximize efficiency, production or some other measure. Optimization consists in trying variations on an initial concept and using the information gained to improve on the idea. A computer is the perfect tool for optimization as long as the idea or variable influencing the idea can be input in electronic format. Feed the computer some data and out comes the solution [13].

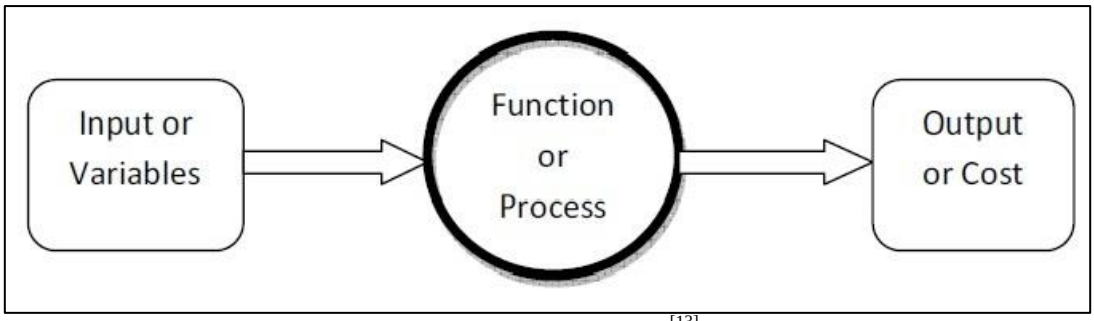

Fig 2.5.2 Optimization Process [13]

VI. ALGORITHMS FOR FINDING COMMUNITIES [11,14]

(1) Minimum-cut method

One of the oldest algorithms for dividing networks into parts is the minimum-cut method. In this, the network is divided into a predetermined number of parts, usually of approximately the same size, chosen such that the number of edges between groups is minimized. The method works well in many of the applications for which it was originally intended but is less than ideal for finding community structure in general networks and it will find only a fixed number of them.

(2) Hierarchical clustering

In this method one defines a similarity measure quantifying some (usually topological) type of similarity between node pairs. Commonly used measures include the cosine similarity, the Jaccard index, and the Hamming distance between rows of the adjacency matrix. Then one groups similar nodes into communities according to this measure. The two simplest being **single-linkage clustering**, in which two groups are considered separate communities if and only if all pairs of nodes in different groups have similarity lower than a given threshold, and **complete linkage clustering**, in which all nodes within every group have similarity greater than a threshold.

(3) Girvan–Newman algorithm

This algorithm identifies edges in a network that lie between communities and then removes them, leaving behind just the communities themselves. The identification is performed by employing the graph-theoretic measure betweenness, which assigns a number to each edge which is large if the edge lies "between" many pairs of nodes.

The Girvan–Newman algorithm returns results of reasonable quality and is popular because it has been implemented in a number of standard software packages. But it also runs slowly, taking time $O(m^2n)$ on a network of n vertices and m edges, making it impractical for networks of more than a few thousand nodes.

The steps of the algorithm are [11]:

1. Computation of the centrality for all edges;
2. Removal of edge with largest centrality: in case of ties with other edges, one of them is picked at random;
3. Recalculation of centralities on the running graph;
4. Iteration of the cycle from step 2.

(4) Modularity maximization

One of the most widely used methods for community detection is modularity maximization. Modularity is a benefit function that measures the quality of a particular division of a network into communities. The modularity maximization method detects communities by searching over possible divisions of a network for one or more that have particularly high modularity. Since exhaustive search over all possible divisions is usually intractable, practical algorithms are based on approximate optimization methods such as greedy algorithms, simulated annealing, or spectral optimization The currently best modularity maximization algorithm is an iterative ensemble algorithm. Modularity optimization often fails to detect clusters smaller than some scale, depending on the size of the network (resolution limit). Several algorithms able to find fairly good approximations of the modularity maximum [11]:

1. **Greedy techniques:**

   The first algorithm devised to maximize modularity was a greedy method of Newman (Newman, 2004b). It is an agglomerative hierarchical clustering method. This greedy optimization of modularity tends to form quickly large communities at the expenses of small ones, which often yields poor values of the modularity maxima.

2. **Simulated annealing:**

   Simulated annealing is a probabilistic procedure for global optimization used in different fields and problems. It consists in performing an exploration of the space of possible states, looking for the global optimum of a function F, say its maximum.

3. **External optimization:**

   External optimization (EO) is a heuristic search procedure proposed by Boettcher and Percus (Boettcher and Percus, 2001), in order to achieve an accuracy comparable with simulated annealing, but with a substantial gain in computer time. It is based on the optimization of local variables, expressing the contribution of each unit of the system to the global function at study.

VII. CONCLUSION

In this paper, we have presented a basic idea of 1) Community detection 2) Social network 3) Genetic algorithm and 4) algorithms for finding communities.

VIII. ACKNOWLEDGMENT

Special thanks to V.V.P. engineering college to support this research. I am grateful to prof. Kamal Sutaria for helping in discussion.

REFERENCES

[1] S.N. Dorogovtsev, J. F. F. Mendes, "Evolution of networks", *Advances in Physics*, 6th March 2001
[2] Clauset, A., Newman, M.E.J., and Moore, C. "Finding community structure in very large networks", *Physical Review E*, 70:066111, 2004.
[3] Duch, J., Arenas, A. "Community detection in complex networks using extremal optimization", Pre-print condmat/0501368, 2005.
[4] Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining: Concepts and Techniques, 3rd Edn, Published by Morgan Kaufmann ,USA, 2012
[5] JingjingMa, Jie Liu, WenpingMa, Maoguo Gong, and Licheng Jiao, "Decomposition-Based Multiobjective Evolutionary Algorithm for Community Detection in Dynamic Social Networks", Hindawi Publishing Corporation the Scientific World Journal, Article ID 402345, Volume 2014,
[6] Jaideep Srivastava, Muhammad A. Ahmad, Nishith Pathak, and David Kuo-Wei Hsu, "Data Mining Based Social Network Analysis from Online Behaviour", SIAM Conference on Data Mining, 2008
[7] Stanley Wasserman and Katherine Faust, Social Network Analysis: Methods and Applications, Cambridge University Press,1994
[8] Chuntao Jiang, Frans Coenen, and Michele Zito, "A Survey of Frequent Subgraph Mining Algorithms", The Knowledge Engineering Review, Vol. 00:0, Cambridge University Press, 2004, 1-31.
[9] Diane J. Cook, Lawrence B. Holder, Mining Graph Data, Published by John Wiley & Sons,New Jersey 2007
[10] Maryam pourkazemi and Mohammadreza Keyvanpour, "A survey on community detection methods based on the nature of social networks", 3rd International Conference on Computer and Knowledge Engineering (ICCKE), IEEE, 2013
[11] Santo Fortunato, "Community detection in graphs", Complex Networks and Systems Lagrange Laboratory, ISI Foundation, Viale S. Severo 65, 10133, Torino,I-ITALY,2010
[12] Mursel Tasgin, Haluk Bingol. "Community Detection in Complex Networks using Genetic Algorithm", PACS numbers: 89.75.Fb, 89.20.Ff, 02.60.Gf
[13] S. RAJASEKARAN and G. A. VIJAYALAKSHMI PAI, NEURAL NETWORKS, FUZZY LOGIC AND GENETIC ALGORITHM: SYNTHESIS AND APPLICATION, PHI Learning Private Limited, New Delhi, 2003
[14] http://en.wikipedia.org/wiki/Community_structure