IJRSET JULY 2017 Volume 4, Issue 7 Pages: 46-53

ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

A STUDY AND ANALYSIS OF VARIOUS COMMUNITY DETECTION TECHNIQUES IN LARGE AND COMPLEX NETWORKS

[1] Mrs. SARADHA, [2] Dr. P. ARUL,
[1] Research Scholar, [2] Assistant Professor,
[1] Bharathiar University, [2] Department of Info. Technology,
[1] Coimbatore. [2] Govt Arts College,
[2] Trichy.

ABSTRACT: Real world complex networks such as social networks, biological networks usually exhibit in homogeneity, resulting in densely interconnected nodes, communities, which play an important functional role in the original system. Analyzing such communities in large networks has rapidly become one of the major topics in complex networks. Complex systems are composed of a large number of interacting elements such that the system as a whole exhibits emergent properties not obvious from the properties of its individual parts. Complex networks describe a wide range of systems in nature and society. To understand complex networks, it is crucial to investigate their community structure. Detecting such communities in large networks has rapidly become one of the focal topics in the science of complex networks. The challenge in community detection is to define what constitutes a community in such a way that this definition not only yields meaningful communities but also allows for sufficiently fast algorithmic implementation to find them.

In particular, identifying communities in large-complex networks is an important task in many scientific domains. In this review, we evaluated state-of-the-art and traditional algorithms for overlapping and disjoint community detection on large-scale real-world networks with known ground-truth communities. In this paper, we study a focused review of different motivations that underpin community detection. This problem-driven classification is useful in applied network science, where it is important to select an appropriate algorithm for the various purposes.

Keywords: - Community, detection, networks, algorithm, complex, large scale.

1. INTRODUCTION

Complex networks describe a wide range of systems in nature and society [1–3]. Frequently cited examples include the Internet in which routers and computers are connected by physical links, and collaboration networks in which researchers are linked by coauthoring. To understand the formation, evolution, and function of complex networks, it is crucial to investigate their community structure, not only for uncovering the relations between internal structure and functions, but also for practical applications in many disciplines such as biology and sociology. Complex network is a structure made up of nodes, representing entities, and links or edges, representing relationships of interactions between entities. Complex systems in various domains may be

modeled as complex network, such as the Internet, World Wide Web, Biological networks, Communication networks and Social networks. Most of the complex networks are generally sparse in global but dense in local, which can be described as the nodes within the group which have higher density of edges, while nodes among groups have lower density of edges. Those groups are called communities which act as a key element to reveal the hidden features of a given network.

The study of complex system investigates how relationships between parts of a system give rise to the collective behaviours of a system and how the system interacts and forms relationships with its environment. The key problem of complex systems is the difficulty with their formal modeling. Complex systems are defined on the basis of different perspectives in different research contexts and they can be represented as network. The modern science of networks is an active field of research within the interdisciplinary science of complex systems. Since complex networks have many interconnected components, it has become an important topic in the study of complex systems. In mathematics, the study of networks starts with graph theory. Pure graph theory deals mostly with regular and abstract constructions which have little in common with real networks.

This paper presents the study to complex networks and community structures and lists the various characteristics of complex networks. It describes the important features of communities in complex networks and analyses the necessity of identifying communities in complex networks. In some kinds of complex networks, new edges continually appear while old edges do not disappear, resulting in a large network. For example, citation networks are growing as new papers cite existing papers. To efficiently process these kinds of networks, we desire a community detection algorithm that will be able to process a network (1) without recomputing whole network after every new edge/node and (2) without the need of whole network structure available at each update.

2. PROPERTIES OF COMPLEX NETWORKS

In order to understand the functionality of a complex system which can be represented as complex network, the properties of the network should be characterized. A decade of research in network analysis has revealed a number of common properties of complex real-world networks. When complex networks were studied first, most important basic topological properties such as degree distributions, clustering and the small-world effect were focused. But, when the focus of network research turned towards functionality and dynamics of networks in the last decade, communities were identified as an important property of complex networks. Some of the important properties of complex networks are listed below.

1. Path

A basic characteristic of a network is the average distance between all pairs of nodes or the maximum distance. A path in a network is simply a chain of links forming a connection between two nodes.

2. Degree Distribution

An important quality which is based on the network's topology is the degree of the nodes. The degree is the basic property measuring the number of neighbours of a single node i.e. the number of edges connected to a particular node.

3. Clustering coefficient

In complex networks, a clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together.

4. Small-world effect

The tendency for individual elements in a large system to be separated from any other element in the system by only a few steps is called small-world effect. The small-world effect is found in many real-world phenomena like food chains, the connectivity of the internet, networks of brain neurons and social networks.

5. Communities

A common topological feature among all kinds of networks is community structure. Networks, in nature, possess a remarkable

amount of structure. The identification of high order structures reveals the functional organization of the networks.

3. PROBLEM STATEMENT

Detecting communities is of prime importance in sociology, biology and computer science disciplines where systems are often represented as graphs. This problem is very hard and not yet satisfactorily solved, despite the huge effort of a large interdisciplinary community of scientists working on it over the past one and half decades.

Besides this, several other challenges have been encountered during the analysis of community structure in large networks, some of which are as follows:

- Most community detection algorithms are based on optimizing a combinatorial parameter. This optimization is generally non-deterministic, thus merely changing the vertex order can alter the vertex-to-community assignments.
- Modularity is a widely accepted metric for measuring the quality of community structure identified by various community detection algorithms.
- For each detected community an effort is made to interpret it as a "real" community by identifying a common property or external attribute shared by all the members of the community.
- Although there is a large volume of research on community detection, systematic post-hoc analysis of the communities, which can unfold interesting characteristic properties of various real systems, is missing in the literature.

Given this scenario, it is clear that we need to develop a better understanding of community structure in various types of large networks. The goal of our research is to study different aspects of community analysis in complex networks that mainly focus on two major directions – (i) identification of realistic communities in different large networks and (ii) leveraging such community structure for developing various applications.

4. LITERATURE REVIEW

Formally, a network is a collection of nodes and links connecting pairs of nodes. The links and nodes may be physical entities like routers and optical fibers of the Internet, or they may represent more abstract relations like networks of word synonyms. The fundamental problem is to exactly define what constitutes a community and how such structures can be efficiently identified in large real networks.

Many researchers have been conducted in order to understand the nature of ground truth communities in real-world networks as well as ones identified by community detection algorithms over a broad range of networks. Although the notion of community is not straight forward, these researches provide essential information so that one can study several qualities of communities as well as their characteristics.

Guimer`a et al. proposed a methodology that allows one to extract and display information about node roles in complex networks. Specifically, the role of a node in a network partition can be defined by its value of within-module connectivity and its participation into inter-cluster connections. Our work here is based on a similar method of illustration, but instead of analyzing roles of nodes in a network partition, we conduct a community-level analysis to expose the nature of communities that constitute the network.

Schaub, Michael T. et al. proposed, Community detection, the decomposition of a graph into essential building blocks, has been a core research topic in network science over the past years. Since a precise notion of what constitutes a community has remained evasive, community detection algorithms have often been compared on benchmark graphs with a particular form of assortative community structure and classified based on the mathematical techniques they employ. However, this comparison can be misleading because apparent similarities in their mathematical machinery can disguise different goals and reasons for why we want to employ community detection in the first place. Here we provide a focused review of these different motivations that underpin

community detection. This problem-driven classification is useful in applied network science, where it is important to select an appropriate algorithm for the given purpose. Moreover, highlighting the different facets of community detection also delineates the many lines of research and points out open directions and avenues for future research.

Gregori et al. proposed the analysis of real-world complex networks has been the focus of recent research. Detecting communities helps in uncovering their structural and functional organization. Valuable insight can be obtained by analyzing the dense, overlapping, and highly interwoven k-clique communities. The novel method has an unbounded, user configurable, and input-independent maximum degree of parallelism, and hence is able to make full use of computational resources. Theoretical tight upper bounds on its worst case time and space complexities are given as well. Experiments on real-world networks such as the Internet and the World Wide Web confirmed the almost optimal use of parallelism (i.e., a linear speedup).

Mahajan & Kaur et. al. proposed to identifying strongly associated clusters in large complex networks has received an increased amount of interest since the past decade. The problem of community detection in complex networks is an NP complete problem that necessitates the clustering of a network into communities of compactly linked nodes in such a manner that the interconnection between the nodes is found to be denser than the intra-connection between the communities. In this paper, different approaches given by the authors in the field of community detection have been described with each methodology being classified according to algorithm type, along with the comparative analysis of these approaches on the basis of NMI and Modularity for four real world networks.

Estrada et. al. proposed to use four different quality criteria for detecting the best clustering and compare the new approach with the Girvan–Newman algorithm for the analysis of two "classical" networks: karate club and bottlenose dolphins. Finally, we analyze the more challenging case of homogeneous networks with community structure, for which the Girvan–Newman completely fails in detecting any clustering. The N-ComBa K-means approach performs very well in these situations and we applied it to detect the community structure in an international trade network of miscellaneous manufactures of metal having these characteristics.

5. COMMUNITY DETECTION METHODS

Among the various properties used to study complex networks, communities has become one of the most important property. Finding community structures in networks is an interesting step towards understanding the complex systems they represent. One of the most prominent features of social and biological networks is the presence of communities i.e. the organization of vertices in modules, with a high level of connectivity inside the modules and low connectivity among modules. Communities provide an insight into not only structural organization of networks, but also functional behavior of various real world systems.

There are many definitions of community detection. There are two types of community definitions; local and global. Local definitions focus on the subgraph under study but neglect the rest of the graph. On the other hand, in the case of global definitions, communities are defined with respect to the graph as a whole. Communities are important for understanding the homogeneous node groupings and identifying the leaders in the group or connectors of different groups. Communities help to have a compact and understandable description of a complex network as a whole. The community structure of a network can also act as a powerful visual representation of a complex system.

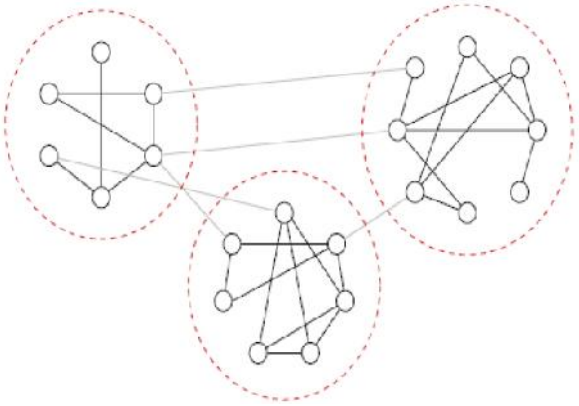

Figure 4.1 - Example of Networks with Communities

The research on community detection in complex network is classified into two categories. The first category includes research on disjoint communities which deal with nodes that belong to only one community. The second category involves finding the overlapping communities in the network where nodes belong to more than one community. When networks have structures, where a clear assignment of a node to a community is not possible or not desirable, then, the necessity of overlapping community detection algorithms arises.

Generally, community definitions can be done globally, locally or based on vertex similarity.

4.1. Global Methods

Global methods utilize the whole network structure for defining the communities. When clusters are essential parts of the graph, which cannot be taken apart without seriously affecting the functioning of the system, only global definitions suit for the community. In the literature, there are many global criteria to identify communities which are indirect definitions i.e. some global property of the graph is used in the algorithm that produces communities at the end. Based on the assumption that a graph offers a community structure if it is not a random graph, then many global criteria are used to identify communities. In the global definitions, a global criterion is associated with the graph which is used to compute communities. This global criterion purely depends on the chosen algorithm.

A random graph that matches some structural properties of the original graph can be said as a null model. To find whether a graph exhibits community structure or not, null model is used as a term of comparison.

4.2. Local Methods

Definitions based purely on local network structure have gained more popularity in the current

decade. Local definitions study the inner structure of the remaining part of the graph independently. The study of local structures is preferable for large complex networks where each node does not depend on most of its peers. Particularly, a user in social network does not have any idea about how large the network is, but form topical communities based only on partial information. Communities can also be defined by a fitness measure. The fitness measure expresses to which extent a subgraph satisfies a given property related to its cohesion. The larger the fitness, the community is more definite. Communities are defined based on quality functions also. Quality functions give an estimate of the goodness of a graph partition.

Another important measure of interest for defining community is the relative density which is defined as the ratio between the internal and total degree of subgraph. The methods that define communities through global objective functions like modularity may fail to discover the ring communities.

6. VARIOUS COMMUNITY DETECTION ALGORITHMS

Community detection algorithms can be classified into different techniques depending upon various features. These features can specify constraints for input data and can improve the power of the results or facilitate the process of community detection.

Though, there are various properties desirable for community detection approach, listed the following as important features:

i) Parameter free: -

An algorithm should be able to make explicit knowledge that is hidden inside the data without getting any external information from the user regarding the data or the problem. An efficient community detection algorithm should attempt to detect communities without getting any external parameter from the user or try to minimize the number of parameters.

ii) Multi-Dimensional Input: -

If a complex network contains a number of different kinds of relations established between the nodes of the network then it is said to be multi-dimensional. When dealing with multi-dimensions, the notion of community changes and the algorithms that are designed for simple static networks cannot handle the multiple dimensions of the network relations.

iii) Incremental: -

An incremental algorithm should be able to provide an output without an exhaustive search of the entire input. As time evolves, the network structure may change by addition of nodes or removal of nodes.

iv) Multipartite: -

In general, networks consisting of different types of nodes with edges running only between unlike types are called multipartite networks, of which the bipartite graph is a special case.

The tremendous growth of real-world networks has forced the research community to develop scalable approaches that can be applied to complex networks with several millions of nodes and billions of edges. Many algorithms have been proposed to deal with community structure detection based on the principles such as hierarchical clustering, graph clustering, optimization methods, spectral partitioning of the network and many more. Depending on the criteria selection, one algorithm can belong to more than one category.

Given the various choices in defining a community, it is natural that a large number of methods and related algorithms have been proposed over the years using a variety of techniques. Each algorithm has a view about the relation that exists between communities in the network. Depending upon the view of the researcher about the relation between communities, community detections algorithms can be classified into variety of techniques.

6.1. Divisive Algorithms: -

Inter community edges are the edges which go from a node belonging to one community to another node which belongs to a different community. Divisive algorithm identifies such edges and removes them one by one. The algorithm can either have a stopping criterion or remove all the edges and then construct a dendrogram using the order of removal of edges.

6.2. Agglomeration Algorithms: -

In agglomeration algorithms, each node in the network is assumed to be of individual community i.e., we will have as many communities as the number of nodes in the network. Every community is merged with the neighboring communities based on a criterion. The merging of communities is stopped once the stopping criteria are reached.

6.3. Random Walk: -

In random walk based methods, similarity between vertices are calculated based on the probability of a random walker choosing that path. This probability will generally be high for vertices which are closer than the ones farther apart. The similarity score is then used to find communities by either divisive or agglomeration technique.

6.4. Spectral Methods: -

Spectral methods transform the adjacency matrix of the network into a suitable form. Then, it uses the values of the Eigen vectors of this matrix to find communities.

Generally, community detection methods can be broadly classified into two main categories namely disjoint and overlapping community detection methods. If a network has overlapping communities, a disjoint algorithm cannot find them; conversely, if communities are known to be disjoint, a disjoint algorithm will generally perform better than an overlapping algorithm. To obtain best results for a given network, it

is important to use the right kind of algorithm. Depending on the nature of the network evolution, community detection algorithms can be classified as static and dynamic community detection.

The important problems in community detection algorithms are: 1. How to know that the produced communities by an algorithm are good ones? 2. Whether the community reveals the correct structure? There should be some methods to say that the communities produced from the algorithm are the best communities for a given network.

Many community detection algorithms have been developed from various disciplines such as physics, biology, applied mathematics, computer and social sciences. An algorithm cannot be justified as correct without testing it. Real-world networks and synthetic networks are essential for testing a community detection algorithm.

CONCLUSIONS

In this review, we empirically evaluated several state-of-the-art community detection algorithms for overlapping and disjoint community detection on large-scale real-world networks. The algorithms were evaluated by measuring the structural properties of their identified communities, as well as their performance with respect to the known ground-truth communities. There are many classes of algorithms for detecting overlapping communities. Identification of the best community among the network based on the current scenario is a big challenge. In this paper, we have reviewed a collection of community detection algorithms, including variants specifically designed for complex networks that have previously been used to cluster complex networks. Modularity based algorithms suffer from a well-known resolution limit but the best-performing algorithm for large networks, the random-walks based Infomap, cannot be applied to a bipartite network directly. Overlapping 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 large and complex network in allowable time.

REFERENCES

[1]. B. Amiri, L. Hossain, J. W. Crawford and R. T. Wigand, “Community Detection in Complex Networks: Multi–objective Enhanced Firefly Algorithm,” Science Direct, Knowledge Based Systems, Elsevier, vol.46, 2013.

[2]. L. Ma, M. Gong, J. Liu, Q. Cai and L. jiao, “Multi-level learning based memetic algorithm for community detection,” Science Direct, Applied Soft Computing, Elsevier, vol.19, pp.121-133, 2014.

[3]. Chen M, Nguyen T, Szymanski BK (2015) A new metric for quality of network community structure. arXiv:150704308

[4]. Browet A, Hendrickx JM, Sarlette A (2016) Incompatibility boundaries for properties of community partitions. arXiv:160300621.
https://arxiv.org/abs/1603.00621

[5]. Bickel PJ, Sarkar P (2016) Hypothesis testing for automated community detection in networks. J R Stat Soci Series B (StatMethodol) 78(1):253–273

[6]. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature **466**(7307), 761–764 (2010).

[7]. Barber, M.J.: Modularity and community detection in bipartite networks. Phys. Rev. E **76**(6), 066102 (2007)

[8].Liu, X., Murata, T.: An efficient algorithm for optimizing bipartite modularity in bipartite networks. JACIII **14**, 408–415 (2010)

[9]. Steinhaeuser K, Chawla NV, Ganguly AR. Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science. Stat Anal Data Mining 2011, 4:497–511.

[10]. Coscia M, Giannotti F, Pedreschi D. A classification for community discovery methods in complex networks. CoRR, abs/1206.3552, 2012.

[11]. Malliaros FD, Vazirgiannis M. Clustering and communitydetection in

directed networks: a survey. CoRR, abs/1308.0971, 2013.
[12]. F Moradi, T Olovsson, P Tsigas, “An of community detection algorithms on large-scale email traffic”, in: SEA. Berlin/Heidelberg: Springer; 2012; 283–294.
[13]. Lei Tang, Xufei Wang, Huan Liu, “Community Detection in Multi-Dimensional Networks”, Technical Report, TR-10-006, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287, 2010.
[14]. F.D Malliaros, M Vazirgiannis, “Clustering and community detection in directed networks: a survey”, CoRR, abs/1308.0971, 2013.
[15]. J Leskovec, K.J Lang, M.W Mahoney,” Empirical comparison of algorithms for network community detection”, CoRR, abs/1004.3539, 2010.
[16]. Nam P. Nguyen, Thang N. Dinh, Ying Xuan, My T. Thai, “Adaptive Algorithms for Detecting Community Structure in Dynamic Social Networks”, IEEE infocom, 2011.
[17]. Daxiang Ji, Yuqing Sun and Demin Li, “ Improved Random Walk Based Community Detection Algorithm”, International Journal of Multimedia and Ubiquitous Engineering Vol.9, No.5 2014, pp.131-142.
[18]. Brian Ball, Brian Karrer, M.E.J Newman, “An efficient and principled method for detecting communities in networks”, April 2011.
[19]. F. Altunbey, "Overlapping Community Detection in Social Networks Using Parliamentary Optimization Algorithm," IJCNA, pp. 12-19, 2015.
[20]. W. Zhan, "Identifying overlapping communities in networks using evolutionary method," Physica A, pp. 182-192, 2016.