Review on Community Detection Algorithms in Social Networks

Cuijuan Wang[1], Wenzhong Tang[1], Bo Sun[2], Jing Fang[2,*], Yanyang Wang[3]
1. School of Computer Science and Technology, Beihang University, Beijing, China
2. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, China
3. School of Aeronautic Science and Engineering, Beihang University, Beijing, China
*corresponding author
Email address:fj@cert.org.cn

***Abstract*—With the development of Internet and computer science, more and more people join social networks. People communicate with each other and express their opinions on the social media, which forms a complex network relationship. Individuals in the social networks form a "relation structure" through various connections which produces a large amount of information dissemination. This "relation structure" is the community that we are going to research. Community detection is very important to reveal the structure of social networks, dig to people's views, analyze the information dissemination and grasp as well as control the public sentiment. In recent years, with community detection becoming an important field of social networks analysis, a large number of academic literatures proposed numerous methods of community detection. In this paper, we first describe the concepts of social network, community, community detection and criterions of community quality. Then we classify the methods of community detection from three classes: i) traditional algorithms of community detection; ii) algorithms of overlapping community detection; iii) algorithms of local community detection. And at last, we summarize and discuss these methods as well as the potential future directions of community detection.**

Keywords-social media; social networks; community detection

I. INTRODUCTION

With the development of Internet and computer science, more and more people start to join social networks. As a "virtual society", social network connects individuals in the real world and expand people's communication, information sharing, social activities, etc. Therefore, more and more researchers are devoting to the study of social networks analysis.

The development of graph theory has laid a solid foundation for the research of social networks, which could be described by graphs [1]. Vertex represents the entity in social networks, such as information, users and so on, while the edge represents the relation between the entities, such as the relationship of friends, information dissemination and so on[2].

Community structure is an important attribute of social networks, and community detection has a significant implication on the revealing of the social networks' structure. A large number of methods have been proposed to solve community detection problem. In this paper, section 2 introduces definitions of social network, community, community detection and criterions of community quality. In section 3, we introduce the methods of community detection from different perspectives according to the development of research of community detection: traditional algorithms; algorithms of overlapping community detection; algorithms of local community detection. Section 4 discusses the three categories. In section 5, we make a prospect of the future of community detection research.

978-1-4673-9088-0/15/$31.00 ©2015 IEEE

II. BACKGROUND

A. Social Network

There are many kinds of networks in the real word. Wang [3] defined the network as follows: "network is a set of entities with specific content and relation between entities". As a typical network, social network is developing rapidly. We define the social network as a kind of network formed by entities and relation between them, where entities could interact with each other [4]. There are many typical social networks, such as Facebook, renren, QQ, WeChat, Twitter, Sina Weibo, forum and so on. Social network is a stretch of real word on the virtual world, and its features, variation, and community structure reflect how the real world works. More and more people devote to the study of social networks.

B. Community

Community [5] is an important structure of social networks. The basic definition is that there must be more edges inside the community than edges linking vertices of the community with the rest of the graph [6]. From the perspective of graph, Newman calls the group of vertices with dense connections within group and sparser connections between group community [7]. Santo Fortunato holds that community is a set of vertices with similarity [6]. However, it is worth noting that, no definition is universally accepted, because the definition often depends on the specific system or application one has in mind [6].

C. Community Detection

Typically, community detection aims to divide a graph into different subsets, but an entity may belong to more than

one communities in reality. For instance, a researcher may participate in study of different fields, and a weibo user may have have a number of hobbies, thus might be divided into different communities according to the fields or hobbies. Therefore, overlapping community detection should be concerned about [8]. Under special circumstances, if we focus on communities of a particular user or group, our goal is to detect the communities which the target users belong to.

D. Criteria of Community Quality

We have known what a community is, but how to define a good community? In this section, we summarize the common criteria of community quality.

1). Modularity

The concept of modularity is firstly proposed by Newman [2], which means the fraction of within-community edges minus the expected value of the same quantity for randomized network. The modularity measure is defined as:

$$Q = \Sigma(e_{ii} - a_i^2) \tag{1}$$

e_{ii} is fraction of edges from group i to group j. a_i is fraction of edges from/to group i if the group by chance. If the number of within-community edges is no better than random, we will get Q=0. Values approaching Q=1, indicating networks with strong community structure. In practice, values for such networks typically fall in the range of about 0.3 to 0.7. Higher values are rare.

2). Normalized Mutual Information (NMI)

NMI [9] is often used to measure the accuracy of community detection, and the ground-truth community structure is known in advance. NMI is defined as:

$$I(A,B) = \frac{-2\Sigma_{i=1}^{C_A}\Sigma_{j=1}^{C_B} N_{ij}\log(\frac{N_{ij}N}{N_{i.}N_{.j}})}{\Sigma_{i=1}^{C_A} N_{i.}\log(\frac{N_{i.}}{N}) + \Sigma_{j=1}^{C_B} N_{.j}\log(\frac{N_{.j}}{N})} \tag{2}$$

C_A is the number of "real" communities. C_B is the number of "found" communities. The rows of N correspond to the " real" communities, and the columns correspond to the "found" communities. N_{ij} is the number of nodes in the real community i that appear in the found community j. $N_{i.}/N_{.j}$ sums over row i/ column j of matrix N_{ij}. Greater I indicates that the detected communities and the real communities are close. If I takes its maximum value of 1, the partitions are identical to the real communities.

3). Multi-criterion scores

Jure Leskovec et al. [10] consider the following metrics $f(S)$ that capture the notion of a quality of the cluster: Conductance, Expansion, Internal density, Cut Ratio, Normalized Cut, Maximum-ODF, Average-ODF, Flake-ODF. Multi-criterion scores evaluate the communities from many aspects, which help to have a better understanding of the communities. Lower value of score $f(S)$ signifies a more community-like set of nodes.

III. Algorithms of Community Detection

A large number of methods have been proposed to solve community detection problem so far. At first, some researchers proposed some traditional methods, such as clustering method, Newman's algorithm and so on. However, it is well understood that people in a social network may belong to multiple communities. So many researchers tend to study the overlapping community detection. With the development of the networks, the networks are getting much more complicated and huge. Previous methods are focused on dividing the whole graph into a number of groups which will cost much when the network is complicated and huge. So some researchers tend to study local community detection. In this section, we will introduce the algorithms of community detection. And the algorithms are categorized into three classes which reflect the development of the research of community detection.

A. Traditional Algorithms of Community Detection

People started to research communities of graph since early 1970's [6]. And a large number of classical algorithms have been proposed. And then we will introduce some typical traditional algorithms of community detection.

1). Partitional Clustering

Partitional clustering is a typical method of community detection. The algorithm assume there are k clusters in the network and the goal is to separate the points in k clusters such to maximize/minimize a given cost function based on distances between points and/or from points to centroids, i.e. suitably defined positions in space [5]. Some of the most used functions are Minimum k-clustering, k-clustering sum, k-center, k-median and so on[5]. Here, the number of clusters k is preassigned. K-means algorithm [11] is a typical partitional clustering algorithm. Partitional clustering is easy to implement and has a reasonable performance, but we need to specify the number of communities in advance.

2). Hierarchical clustering

We usually don't know the number of communities and their size. So hierarchical clustering algorithm [12] is proposed. Hierarchical clustering algorithm include agglomerative algorithms and divisive algorithms [13]. The basic idea of agglomerative algorithms is that clusters are iteratively merged bottom-up if their similarity is sufficiently high, and the typical agglomerative algorithm is betweenness clustering. Divisive algorithms' basic idea is that clusters are iteratively split top-down by removing edges connecting vertices with low similarity. We will get a dendogram tree and get communities by cutting the tree. There is no need to know the prior knowledge. But if the cutting position and merging position are not appropriate, we might get low quality communities.

3). Newman's algorithm

Girvan and Newman made great contributions to community detection and a series of classical methods were proposed [12]. Girvan and Newman proposed the concept of edge betweenness which is the number of shortest paths between all vertex pairs that run along the edge. The basic idea of this algorithm is to remove edges with the highest betweenness and then recalculate betweennesses for all edges affected by the removal. Repeat until there are no edges remain. There is no need to know the number of clusters in advance, but calculating edge betweenness spends much time and we

don't know where to cut the dendogram tree. In order to abtain better efficiency, Girvan and Newman proposed Newman's fast algorithm. In this method, a concept of modularity Q was proposed to measure the quality of communities. The basic idea of this method is to merge the pair of communities with the largest increase in Q and then get the dendogram tree. The calculation of the edge betweenness is slower than modularity, so this method is faster than G-N algorithm. These methods are focused on undirected graph, so Leicht and Newman [15] improve the original modularity which focused on directed graph later.

4). Graph Partition

Graph partition aims to divide nodes in graph into a plurality of predetermined size communities which satisfies some objective functions by removing edges.

Benchmark was used to measure community structure. The planted l-partition model [15] is the easiest recipe. In this model one "plants" a partition, consisting of a certain number of groups of nodes. Each node has a probability p_{in} of being connected to nodes of its group and a probability p_{out} of being connected to nodes of different groups. As long as $p_{in} > p_{out}$ the groups are communities[17]. The most popular version of the planted l-partition model was GN benchmark. LFR benchmark [18] generalized the GN benchmark by introducing power law distributions of degree and community size. Spectral clustering [19] comes from graph partitioning. The algorithm considers partition with min cut ratio is good partition [20], and ratio cut(R-Cut) and normalized cut(N-Cut) are widely used objective functions [21]. The algorithm needs to know the number of communities in advance, and the time complexity is high. But the spectral clustering is based on strict linear algebra and convex optimization theory and it is easy to implement [22]. Kernighan-Lin algorithm [23] is a greedy optimization algorithm. The basic idea is getting maximization of profit function Q by exchange nodes in group A and B. Kernighan-Lin algorithm is a heuristic algorithm, which can produce good results in the practical application, and the running speed is also faster. But, we need to know the size of communities in advance. Lin Yuan [24] proposed a method to try all possible sizes using Kernighan-Lin algorithm, but it's time complexity is quite high, which is infeasible for complex networks.

5). Infomap

Infomap [25] was proposed to comprehend the multipartite organization of large-scale biological and social systems. the basic idea of this method is to use the probability flow of random walks on a network as a proxy for information flows in the real system and decompose the network into modules by compressing a description of the probability flow.

B. Algorithms of Overlapping Community Detection

Most of the traditional algorithms are focused on identifying disjoint communities. However, it is well understood that people in a social network may belong to multiple communities [26]. So many researcher tend to research the algorithms of overlapping community detection. And many methods have been proposed.

1). Clique percolation

The clique percolation method (CPM) [27] is the first proposed method to identify overlapping communities. And the basic idea of CPM is regarding communities as a set of many cliques (fully connected subgraphs). It begins by finding all k-cliques(fully connected subgraphs with k vertices). Second, regard these k-cliques as nodes in the graph and two k-cliques are adjacent if they share k-1 vertices. In the new graph, the connected parts are the communities. CPMd algorithm is an improvement of CPM algorithm [28] and can used in directed graph. The CPMd algorithm made a new definition of k-clique：the directed links in k-clique is all from node with relative high out-degree to node with relative low out-degree; here cannot exist directed closed loop in k-clique. A node may exist in multiple k-cliques, so CPM and CPMd algorithms could identify overlapping communities. Besides, they may get a better result with more fully connected subgraphs.

2). LINK Algorithm

LINK algorithm is a link partitioning algorithm based on hierarchical clustering [29]. The basic idea of this algorithm is that if two links share the same node belonging to different communities, the node must be a node in the overlapping area. Treat each edge as a separated link community. And then merge the two most similar link communities until all the link communities become a single community. In literature [29], a similarity of a pair of links is computed via the Jaccard Index. Although the link partitioning for overlapping detection seems conceptually natural, there is no guarantee that it provides higher quality detection than node based detection does because these algorithms also rely on an ambiguous definition of community [26].

3). COPRA

COPRA [30] is an improvement of traditional label propagation algorithm and COPRA is a multi-label propagation algorithm. Label each vertex x with a set of pairs (c,b), where c is a community identifier and b is a belonging coefficient. And then Let the label spread on the network based on the local structure of the network. Since each node is initialized with n labels, after propagation each node may have multiple labels at the same time. Therefore, COPRA can identify overlapping communities.

C. Algorithms of Local Community Detection

With the networks becoming much more complicated and huge, it will cost great to research the community structure from the whole graph and divide the graph into several groups. So some local community detection algorithms were proposed.

1). Clauset's Algorithm

Generally according to the definition of local community measure, we optimize the local community by getting nodes into the community with high increase of local measure. The difference between various this kind of methods is the different measure of local community [6]. Clauset is the first to propose the problem of local community detection [31]. Clauset's algorithm proposed a definition of local measure. Clauset's

algorithm is simple and efficient, but needs to set community size in advance.

2). Label Propagation Algorithm

Label Propagation Algorithm(LPA) [32] is a dynamic method of community detection, and it is based on local structure of networks to identify communities. Give every vertex a unique label and update each vertex x's label by replacing it by the label used by the greatest number of neighbors until the same label tends to become associated with all members of a community. The method has a low time complexity and and it is easy to operate, but the algorithm has great uncertainty.

3). Local Node Expansion

Local node expansion is to start with a number of nodes, and then according to the specific criteria expand the nodes to get the community. One of the most popular local node expansion is seed set expansion. Joyce Jiyoung Whang et al. use the personalized PageRank for seed expansion [33]. Isabel M. Kloumann et al. [34] use the traditional PageRank algorithm for seed expansion. The algorithms compute PageRank scores, localized on seeds and then find a set of high-ranked nodes to form the community with the seed set. These two methods' basic idea comes from random walk of Infomap which make the use of the information about direction and weight of links [25]. Another kind of method is based on node centrality[35]. The method thinks hub nodes and edges that generated from hub nodes should be treated with more importance than other nodes and edges and nodes which are closely connected together should be in the same community. The algorithm introduces a centrality measurement to calculate the degree centrality of node and defines node distance based on Katz centrality *algorithm [36], and then starts from the highest centrality node until the lowest centrality node to get communities.*

4). Local Optimization

OSLOM [37] method optimises the local statistical significance of communities. OSLOM consists of three phases: First, it looks for significant clusters, until convergence; Second, it analyzes the resulting set of clusters, trying to detect their internal structure or possible unions thereof; Third, it detects the hierarchical structure of the clusters. OSLOM is the first method capable to detect clusters in networks accounting for edge directions, weights, overlapping communities and community dynamics [37].

IV. CONCLUSION

In this paper, we introduce the definition of community and criteria of community quality, besides we review a lot of methods of community detection. With the development of networks and the further research, community detection is developing. In this section, we will compare these methods in the following table.

TABLE I. COMPARISON OF ALL THE ALGORITHMS

Algorithm		***Pros***	***Cons***
Clustering	K-means	a.Easy to implement b.Reasonable performance	a.Need to specify the number of clusters in advance b.Often terminates at a local optimum
	Hierarchical clustering	a. No need to specify the number of clusters in advance	a.Don't know where to cut the dendrogram tree b.May get bad results if the merging/division heuristic is not good
Newman's algorithm	G-N algorithm	a.No need to specify the number of clusters in advance	a.Don't know where to cut the dendrogram tree; b. Slow
	Newman's fast algorithm	a.Faster than G-N algorithm; c.May get good partitions b.No need to know the number of clusters in advance	a. no theoretical guarantee compared to the greedy algorithm
Graph Partition	Benchmark	a.Easy; b.Low computational complexity	a. Difficult to agrees on a same benchmark
	Spectral Clustering	a.Yields to very good results in general b.Effective to handle complex shapes	a.Usually not efficient b.Not sure which objective is the right one to use
	Kernighan-Lin	a. fast	a.Need to know the size of the clusters in advance
infomap		a.makes use of information aboutweight and direction	a.Only consider the structural characteristics
Clique percolation		a.Can detect overlapping community	a.Suitable for networks with many full connected subgraph
LINK Algorithm		a.Can detect overlapping community	a. Don't know where to cut the dendrogram tree
COPRA		a.Can detect overlapping community	a.Has great uncertainty
Clauset's Algorithm		a. Easy; b. Efficient	a.Need to know the size of the clusters in advance
Label Propagation Algorithm		a.Low time complexity; b.Efficient c.Can detect overlapping community	a.Can detect one community
Local Node Expansion		a.high accuracy; b.niche targeting; c.Efficient	a.Only consider the structural characteristics
Local Optimization		a.Can detect overlapping community b.Can be generalized to directed graphs, weighted graphs and dynamic networks	a.may return slightly less accurate results than other methods

From table I we can draw the conclusion: i) Some traditional methods are relatively easy but often need to decide the number or the size of clusters, other traditional methods usually slow or need to know where to cut the dendrogram tree; ii) Overlapping community detection Methods can detect overlapping community, but some methods are suitable for networks with many full connected subgraph or great uncertainty. iii) Most of local methods are efficient or with low time complexity, but there also exist some shortcomings, such as can detect only one community or need to know the size of the clusters in advance. The research of the community detection has a long time which has achieved many valuable results. Different network has different method of community detection and we should choose suitable method to identify communities.

In the research of community detection, the academic community has made a lot of valuable achievements and proposed many effective methods, which characterized communities from different perspectives. However, with the development of the Internet, social networks are becoming more and more complex; there still exist many problems waiting further research.

Improvement of Speed and Accuracy: with the increasing size of the social network, the research for social networks needs to deal with more and more data. In the era of big data, we can use open source platforms to deal with massive data, but we still need more efficient algorithms to reduce the overhead. How to improve the accuracy of community recognition are what we need to deeply study.

Community Detection in Heterogeneous Network [38]: with the information and relation between information becoming more and more complex, information networks is in the direction of the isomerization. Such as a number of data sets, entities include papers, authors, publishers, title, etc. and relation includes writing, publishing, etc. which is a kind of heterogeneous networks [39]. In the research of community detection, heterogeneous networks can provide us with more comprehensive information and making full use of this information can improve the accuracy of our community detection. But there is less research on the heterogeneous network, because the heterogeneous network needs to deal with much more complex information.

REFERENCES

[1] J.F.F. Mendes, S.N. Dorogovtsev, Evolution of Networks. WWW. 2003.

[2] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 026113.

[3] X. F. Wang, Complex Networks: Topology, Dynamics and Sinchronization, International Journal of Bifurcation and Chaos, vol. 12, no. 05, pp. 885-916, 2002.

[4] C. C. Aggarwal, Social Network Data Analytics: Springer Publishing Company, Incorporated, 2011.

[5] Papadopoulos S, Kompatsiaris Y, Vakali A et al. Community detection in social media. Data Mining and Knowledge Discovery, 2011, 24（3）：515-54.

[6] Santo Fortunato. Community detection in graphs[J]. Physics Reports . 2009 (3).

[7] Michelle Girvan, M. E. J. Newman. Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99 (12) (2001) 7821－7826.

[8] M.G. Everett, S.P. Borgatti, Analyzing clique overlap, Connections 21 (1) (1998) 49－61

[9] Danon L, Diaz-Guilera A, Duch J et al. Comparing community structure identification. J Stat Mech-Theory E, 2005.

[10] Jure Leskovec, Kevin J. Lang, Michael W. Mahoney. Empirical Comparison of Algorithms for Network Community Detection. WWW . 2010.

[11] J. B. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In: L. M. L. Cam, J. Neyman, (eds.). 1967. 281-297.

[12] T. Hastie, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning, Springer, Berlin, Germany, 2001.

[13] Shangfu Gong, Wanlu Chen, Pengtao Jia. Survey on algorithms of community detection[J]. Application Research of Computers. 2013(11).

[14] Newman,M E J.Fast algorithm for detecting community structure in networks. Physical Review E Statistical Nonlinear and Soft Matter Physics . 2004.

[15] LEICHT E A, NEWMAN M J. Community structure in directed networks[J]. Physical Review Lerrers. 2008. 100(11):118703

[16] A. Condon and R. M. Karp. Algorithms for graph partitioning on the planted partition model. Random Struct. Algor., 18:116－140, 2001.

[17] Santo Fortunato, Andrea Lancichinetti. Community detection algorithms: a comparative analysis. VALUETOOLS.2009.

[18] A. Lancichinetti, S. Fortunato, F. Radicchi. Benchmark graphs for testing community detection algorithms. Phys. Rev. E, 78(4):046110, 2008.

[19] W. Donath, A. Hoffman, Lower bounds for the partitioning of graphs, IBM J. Res. Dev. 17 (5) (1973) 420－425.

[20] Bo Yang, Dayou l Liu, Jiming L, et al. Complex Network Clustering Algorithms . Journal of Software

[21] Mu Zhu. Research on the Key Technologies of Community Detection in Complex Networks. 2014(35).

[22] Von Luxburg U. A tutorial on spectral clustering [J]. Statistics and computing, 2007, 17 (4):395-416.

[23] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs[J]. Bell system technical journal, 1970, 49(2): 291-307.

[24] Lin Yuan. Research on Community Detection and Graph Partitioning. 2014(17)

[25] Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences 105(4) (2008) 1118－1123

[26] JIERUI XIE, STEPHEN KELLEY, BOLESLAW K. SZYMANSKI. Overlapping Community Detection in Networks: the State of the Art and Comparative Study. ACM Computing Surveys, vol. 45, no. 4, 2013.

[27] I. Derényi, G. Palla, T. Vicsek, Clique percolation in random networks, Phys. Rev. Lett. 94 (16) (2005) 160202.

[28] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society, Nature 435(2005) 814－818.

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

[30] Gregory S. Finding overlapping communities in networks by label propagation. New Journal of Physics, 2010, 12（10）：103018.

[31] Clauset A. Finding local community structure in networks [J]. Physical review E, 2005,72(2): 026132.

[32] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2007, 76（3 Pt 2）：036106.

[33] Joyce Jiyoung Whang, David F. Gleich, Inderjit S. Dhillon. Overlapping Community Detection Using Seed Set Expansion. CIKM'13.2013(2099-2108).

[34] Isabel M. Kloumann,Jon M. Kleinberg. Community Membership Identification from Small Seed Sets. KDD'14. 2014(1366-1375).

[35] Sorn Jarukasemratana, Murata, Xin Liu. Community Detection Algorithm based on Centrality and Node Distance in Scale-Free Networks.24th ACM Conference on Hypertext and Social Media. 2013.

[36] L. Katz. A new status index derived from sociometric analysis. Psychometrika, 18(1):39?43, Mar. 1953.

[37] Lancichinetti, A., Radicchi, F., Ramasco, J.J., Fortunato, S.: Finding statistically significant communities in networks. PLoS ONE 6(4) (April 2011) e18961+

[38] Dino Ienco,Céline Robardet,Ruggero G. Pensa,Rosa Meo. Parameter-less co-clustering for star-structured heterogeneous data[J]. Data Mining and Knowledge Discovery. 2013 (2).

[39] Wei Shen, Jiawei Han, Jianyong Wang. A probabilistic model for linking named entities in web text with heterogeneous information networks. SIGMOD. 2014(1199-1210).