International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 6 Issue 12|| December 2017 || PP. 54-59

Some Observations on Community Detection Algorithms in Social Networks

Dipika Singh[1], Rakhi Garg[2]
[1]*(Department of Computer Science, BHU, Varanasi, Uttar Pradesh, India)*
[2]*(Computer Science Section, MMV, BHU, Varanasi, Uttar Pradesh, India)*
Corresponding Author: Dipika Singh

Abstract: *This paper focuses on community detection algorithms in social networks. We have given a brief background about graph, social networks and community detection. The popular algorithms used for community detection are categorized into three broad categories, namely graph partition based, clustering based and overlapping community detection. Some of the most frequently used algorithms are tabulated with their advantages and disadvantages. The primary goal of the paper is to provide a summarize survey of recent updates in community detection research, which will provide base for new researchers to work in this direction.*

Keywords*: community detection, graph partition, clustering, overlapping community detection, fuzzy community detection.*

Date of Submission: 15-12-2017 Date of acceptance: 03-01-2018

I. Introduction

With the advances in technology, use of Internet and Social networking sites are becoming inseparable part of our daily life [1]. The reason for this is people are so busy with their own life , they have little or no interactions with friends who are far off. Social sites have provided easily accessible platform for people to communicate and share their views easily and efficiently with no extra cost.

Social networking sites can be considered as graphs with nodes representing individuals, and edges representing their interactions. A community is formed if we can partition nodes into disjoint or overlapping sets such that edges in a set are more than the other by a considerable amount. Community structures can be hierarchical also.

Community detection is the process of discovering clusters or cohesive groups. It is very important aspect of social media mining. It has a variety of applications. For example, we can recommend a specific product to a group of people based on community [2], political results can be anticipated, awareness about disease can be send etc.

Community detection does not have a universal definition which can be treated as exact definition [3]. As a result there are no proper guidelines to compare and judge different algorithms. This benefits us by having freedom of discovering and applying diverse techniques to the problem. But the drawback of lack of standard definition is that lot of noise is introduced which slows down the progress.

In this paper we discuss the algorithms of community detection. Section 2 gives background of social networks, community detection, community detection methods, datasets used for community detection. In Section 3 we have categorized community detection algorithms into three categories i.e. graph partitioning algorithms, clustering algorithms, and overlapping community detection algorithms. Important algorithms in each category are discussed. At last Section 4 concludes the paper.

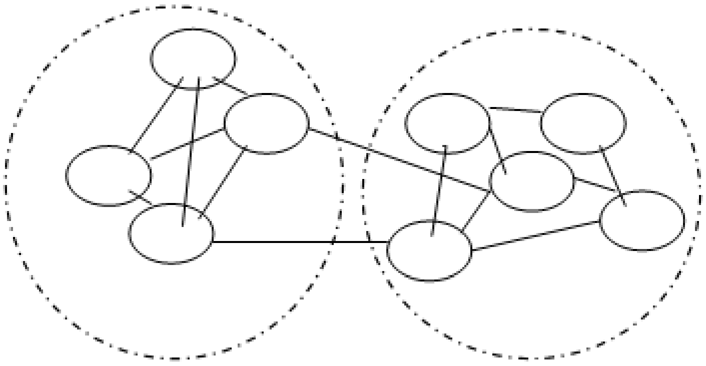
II. Background

2.1 Social Network

Social network can be considered as a set of nodes and edges i.e. a graph [1]. Nodes represent individuals or entities and edges represent their interaction. It can be represented by adjacency matrix with entry 1 for a_{ij} , where i and j entities have interaction between them ,otherwise it is 0. These social networks graph can be studied and analyzed to discover important results.

2.2 Community

Community is formed with such nodes which have more edges between them than the edges connected to other nodes [3].In other words intra-community edge is more than inter community edge as shown in figure 1.People who interact with each other more often than others form a community. The community can be implicit or explicit. Similarity or distance measures are used to calculate the similarity between entities. *Overlapping communities* also exist as entities may belong to one or more set of communities at the same time, for example an individual might be enrolled in a college and karate class at the same time. So he will belong to both communities.



Community 1 Community 2

Figure1: Community formation

2.3 Community Detection Methods

Graph partitioning and clustering are common methods used for community detection. As communities are nothing but graph having more inner connections than outer. So similarity between nodes is calculated to form a community.

2.3.1 Graph partitioning

As the name suggest, its concept is to partition the graph into smaller parts based on certain characteristics [4]. Cut size is crucial for partition. Cut stands for partition of nodes of graph into disjoint sets. Cut size means number of edges in each cut component. If removal of any edge divides the graph into two partitions then it is called multicut. For graph partitioning, number of required components and size of components should be known. It is not practically possible, so it is not efficient method of community detection.

2.3.2 Clustering

Clustering deals with formation of clusters, which are formed by grouping of nodes with similarity between them [3, 5].Two common techniques of clustering are hierarchical clustering and partitioning. Hierarchical clustering is also known as leveling; it deals with formation of a hierarchy of clusters. It has two approaches i.e. agglomerative and divisive. Agglomerative follows bottom up approach. Nodes are agglomerated with similar nodes to form clusters. Divisive approach is opposite of hierarchical where clusters are broken into smaller clusters.

In Partitioning clustering, such as K-means Clustering, there is an initial partition and instances are relocated across clusters. All likely partitions are considered and evaluated for achievement of optimality. Its drawback is that it is time consuming and becomes practically infeasible some times.

2.4 Dataset for Community Detection

Datasets which are generally used for community detection are classified as real data set and benchmark dataset [6]. Real datasets are collected from the real interaction of some groups .Some of the real datasets are Zachary Karate Club, Dolphin Dataset, American College Football Network etc.

Benchmark datasets are artificially designed datasets [7] .GN Benchmark and LFR Benchmark are most popular benchmark datasets. GN Benchmark was given by Girvan Newman, consist of 128 vertices and 4 communities are formed. LFR Benchmark was given by Lancichinetti, Fortunato and Radicchi.

Apart from above methods, datasets can also be obtained from Social networking websites by implementing proper techniques. The data thus obtained requires refinement. Robust algorithms are required when these datasets are used as quantity and variation of data is very large. Figure 2 shows the difference between Zachary Karate club dataset and Facebook dataset.

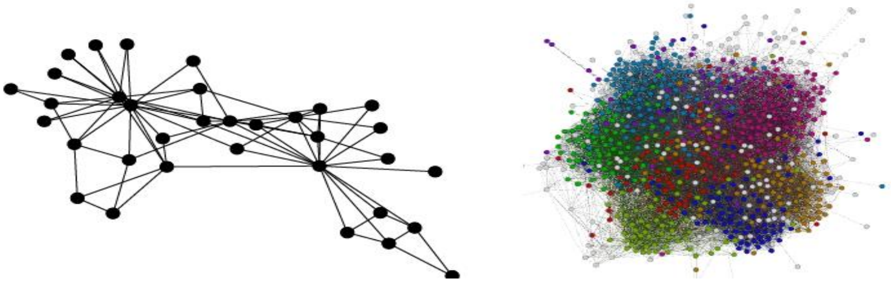


Figure 2: Zachary Karate club network and Facebook friends network

III. Algorithms For Community Detection

Research of communities started in early 1970s. Many algorithms are proposed for community detection and some of them are updated and modified also to meet the new challenges. Important algorithms are discussed in this section.

3.1 Graph Partitioning Algorithms

Kernighan-Lin algorithm [8] is oldest and most popular graph partitioning algorithm and today also it is used with other algorithms. Basic concept is to partitions the nodes of the graph based on cost on edges into subsets of predefined sizes, minimizing the total cost. It is very fast with complexity O (n^2log n). As the number of clusters increase, the run-time and storage cost also increases rapidly.

Spectral bisection method is also very popular approach in graph partitioning. It uses Laplacian Matrix spectrum concept. It is fast and partitions formed are good.

3.2 Clustering based Algorithms

The most important algorithm in community detection was given by Girvan Newman [9]. The algorithm uses divisive approach of hierarchical clustering. It uses the concept of edge betweeenness, which gives criteria to find the edges which connect two communities. These edges are removed to form the community. Time complexity of algorithm is $O(m^2\ n)$ and is $O(n^3)$ for sparse graphs, here m is number of edges and n is the number of vertices. Chen et al [10] has done the extension of GN algorithm for partitioning weighted graphs to obtain functional modules in the yeast proteome network. Rattigan et al [11] proposed reduced the computational complexity of GN algorithm by using the indexing methods. They gave the concept of strong and weak communities. *Radicchi* et al [12] used GN algorithm and proposed the concept of 'strong' and 'weak' communities. Divisive edge removal step of GN algorithm is implemented by using *edge clustering coefficient .Its running time* is $O(\mathrm{m}^4\mathrm{n}^2)$ and $O(n^2)$ for sparse graphs. Parallel version of GN algorithm is proposed by *Moon et al* for handling large amount of data. To implement it they have used MapReduce model (Apache Hadoop) and GraphChi.

Modularity was introduced by Newman. It is defined as Q= $\sum$ (e_{ii} - a_i^2). Here e_{ii} = edges from group i to j, a_i = edges from/to group i in random network. If intra community edges are not better than random Q=0 and for ideal condition Q=1.Generally expected values range from 0.3 to 0.7. Modularity optimization is used as basis for various algorithms of community detection.

Newman modified their algorithm [13] to optimize modularity. The changed modularity is calculated as $\Delta Q=eij+eji-2aiaj=2(eij-aiaj)$ are calculated in constant time. So this method is faster in execution than old GN algorithm. The run time of the algorithm for sparse algorithms is $O(n^2)$ and for others it is $O((m+n)n)$.

Clauset et al [14] applied greedy optimization of modularity for detecting communities in large networks.

Blondel et al [15] designed Louvain method, an iterative two phase algorithm. In first phase, nodes are kept into distinct communities and then the modularity gain by shifting a node i from one community to another is calculated. If positive modularity gain is obtained, the node is shifted to a new community. In second phase nodes are formed from all the communities found in earlier phase and the weight of links is calculated. The time complexity is linear O(m) which improves complexity of GN algorithm.

A popular graph flow simulation algorithm is Markov Clustering Algorithm [16], is used to detect clusters in a graph. Expansion and inflation process alternately takes place in this method.It performs random walk in the graph. If there are 'n' number of nodes and 'k' resources then worst case time complexity is given by O (nk^2). Nikolaev et al [17] and Steinhaeuser [18] implemented Markov chain concept in their algorithm and obtained runtime of O (n^2log n).

3.3 Algorithms for overlapping community detection

3.3.1 Clique Percolation Method (CPM)

It is assumed that communities are overlapping set of cliques. Clique is a fully connected subgraph. In this method, all cliques of size k are identified in the network. These cliques are treated as vertex and a new graph is formed such that. There is an edge between two vertex if and only if their corresponding clique share k-1 members. Now the connected components determine which cliques constitute communities. It works good for dense graphs also. Palla et al. [19] implemented CPM in CFinder algorithm. It works well in many applications with polynomial time complexity. For large networks, sometimes it fails to terminate.

Farkas et al. [20] extended the method for weighted graphs. It is known as CPMw. A threshold is defined for subgraph intensity in weighted graph. All those cliques which have intensity smaller than threshold are rejected. Rest of them form the community. Kumpula et al [21] gave Sequential Clique Percolation Method (SCP).In this all values of k are not considered for clique formation. The size of clique communities is fixed. It is faster than CPM

3.3.2 Line Graph and Link Partitioning

The concept used in this technique is that in place of nodes, links should be partitioned for community formation. The overlapping of node occurs if links connected to it are from more than one cluster.

Ahn et al [22] used this concept, in this hierarchical clustering is used for link partitioning. If there are two edges e_{ik} and e_{jk} incident on vertex k, similarity between them is calculated by Jaccard index

$S(e_{ik}, e_{jk}) = |N_i \cap N_j| / |N_i \cup N_j|$

Here N_i and N_j are used to denote neighborhood of I and j nodes respectively including them. After this single linkage hierarchical clustering is applied to form link dendogram. We obtain link community by cutting the dendogram at some threshold. The time complexity of this algorithm is O (nk^2_{max}), where k_{max} is maximum node degree in the network.

Evans [23] gave the concept of line graph for weighted graphs, whose nodes are the links of the original graph. After this another algorithm for disjoint community can be applied. Partitioning the node of a line graph leads to partitioning of edge of the original graph.

Wu et al. proposed a post processing procedure, CDAEO for determining the extent of overlapping. Evans [24] modified his Line graph method to clique graph, where cliques of a given order are represented as nodes in a weighted graph.

3.3.3Local Expansion and Optimization

These algorithms are based on growing a natural community or a partial community. A local benefit function is used for characterizing the quality of community. Baumes et al.[25] gave a solution of two steps. In the first step a RankRemoval algorithm is used for ranking the nodes based on some predefined criteria. Then highly ranked nodes are repeatedly removed till small and disjoint clusters are obtained. These act as seed for next step. In the second step following local density function is used

$$f = \frac{w_{in}^{c}}{w_{in}^{c} + w_{out}^{c}}$$

Here w_{in}^{c} *and* w_{out}^{c} are internal and external weights respectively of the community c.

Iterative Scan (IS), tries to improve above function by adding and removing nodes in the core.

The running time complexity in worst case is O (n^2). Sometimes this algorithm produces disconnected components. So it is modified by Kelley [26] in form of CIS i.e. Connected Iterative Scan. In this connectivity is checked after each iteration. A new fitness function is used in CIS which considers edge probability e_p also.

$$f = \frac{w_{in}^{c}}{w_{in}^{c} + w_{out}^{c}} + \lambda e_p$$

Here λ is a parameter which controls the behavior of algorithm in sparse areas of the network.

Lancichinetti et al. [27] gave LFM based on fitness function

$$f(c) = \frac{k_{in}^{c}}{(k_{in}^{c} + k_{out}^{c})^{\alpha}}$$

Here $k_{in}^{c}\ and\ k_{out}^{c}$ are internal and external degree of community c. α is used as resolution factor for controlling the size of community.This fitness function is optimized to get a local maximal and till then community is expanded from a random seed node. When one community formation is complete, LFM make selection of a random node which is not assigned to any community and the previous process is repeated for new community. The parameter α is very significant in LFM approach. The worst-case complexity is $O(n^2)$.

3.3.4 Fuzzy Detection Technique

The previous approaches are binary in nature i.e. a node may belong to a community or does not belong, the strength of their belonging is not considered. In fuzzy approach, the strength of association of a node to a community is given importance. In this approach, for each node, a membership vector, or belonging factor , is calculated. The value of membership vector can be calculated or directly provided as parameter.

Nepusz [28] presented the idea that overlapping community detection is a nonlinear constrained optimization problem. Following objective function is used to minimize

$$f = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(s_{\tilde{ij}} - s_{ij}\right)^2$$

Here w_{ij} is predefined weight, $s_{\tilde{ij}}$ is similarity between i and j before applying fuzzy logic, s_{ij} is calculated by following formula.

$$s_{ij} = \sum_c a_{ic} a_{jc}$$

Here a_{ic} denotes the fuzzy membership of node i to community c.

Table 1: Advantage and shortcomings of popular community detection algorithms

Algorithms	**Advantage**	**Shortcomings**
K means Clustering [3]	Implementation is easy and satisfactory performance.	Number of clusters are required to be known in advance. Achieves local optimum and terminates in several cases
Hierarchical Clustering [3]	Advance knowledge of number of clusters not required	No knowledge of where to cut dendogram tree. Results depend on heuristic/merging quality
Girvan Newman Algorithm [9]	Advance knowledge of number of clusters not required	No knowledge of where to cut dendogram tree. It is slow
Newman's Fast Algorithm [13]	Advance knowledge of number of clusters not required Faster than previous GN algorithm	Unlike greedy algorithms there is no guarantee of good result theoretically
Benchmark graph partition [3]	Very easy with less complexity level.	A common benchmark cannot be decided for all.
Spectral Clustering [3]	Performs well for complex network and gives good result.	Conflict in the objective and not efficient.
Kerninghan Lin [8]	Fast	Number of clusters are required to be known in advance.
Infomap [3]	It considers weight and direction	Structural characteristics are only considered.
Clique Percolation [19]	Overlapping community can be detected	Performance is good only if there are fully connected subgraphs.
LINK Algorithm [29]	Overlapping community can be detected	No knowledge of where to cut dendogram tree
COPRA [29]	Overlapping community can be detected	It is highly uncertain.
Clauset's Algorithm [14]	It is efficient and easy	Size of clusters is required to be known in advance.
Label Propagation Algorithm [29]	Overlapping community can be detected, efficient, time complexity is low.	One community can be detected.
Local Node Expansion [29]	Efficient and accurate, niche targeting	Structural characteristics are considered only.
Local Optimization [29]	Overlapping community can be detected, can work well also for directed, weighted and dynamic networks.	Results can be less accurate than other methods.

IV. Conclusion

In this paper, we have discussed community detection and important algorithms developed in this area.There has been many research in this topic and some methods are very effective and valuable.However complexity of networks is growing each day due to growth of Internet. So there is tremendous scope of future research in community detection. The major problems which are yet to be resolved are speed and accuracy. Most of the current algorithms perform well on benchmark data but do not have same performance for real world dataset. We are living in the generation of big data, so efficient algorithms are required to speed up the process and decrease overhead. Quality of community detection is still a field to be explored .New approaches are needed to improve the accuracy of detected community.

References

[1]. Reza Zafarani, Mohammad Ali Abbasi , Huan Liu , "Social Media Mining : An Introduction",2014,Book,Cambridge University Press New York.

[2]. J. Kamahara,T. Asakawa,S.Shimojo,H.Miyahara, "A community based recommendation systemto reveal unexpected results", Multimedia Modelling Conference, 2005,Proceedings of the 11th International ,IEEE.

[3]. Santo Fortunato, "Community Detection in Graphs",2010,Physics Reports 486,page 75 – 174

[4]. M.E.J. Newman, "Community Detection and Graph Partitioning" ,2013,arXiv 1305:4974[cs.SI], Europhys. Lett.103,28003.

[5]. Fragkiskos D.Malliaros, Michalis Vazirgiannis, "Clustering and Community Detection in Directed Networks: A Survey",Physics Reports, Volume 533,Issue 4,30 December 2013,pages 95- 142.

[6]. Leto Peel, Daniel B. Larremore, Aaron Clauset, "The groundtruth about metadata and Community Detection in networks", 2017, Science Advances,Vol 3, No. 5, e1602548.

[7]. Andria Lancichinneti, Santo Fortunato, Filippo Radicchhi, "Benchmark graphs for testing community detection algorithms", 2008, arXiv:0805.4770v4 [physics.soc-ph] .

[8]. B.W. Kernighan , S. Lin, "An efficient heuristic procedure for partitioning graphs", 1970, Bell Sys. Tech. J. vol. 49, no. 2, pages 291 – 307.

[9]. M. Girvan, M.E.J. Newman, "Community structure in social and biological networks", 2002, pnas.org

[10]. Zhenping Li, Shihua Jhang, Rui Sheng Wang, Xiang Sun Zhang, Luonan Chen, "Quantitative function for Community detection",2008, American physical society,Vol. 77, Iss. 3.

[11]. M.J. Rattigan, Marc Maier, David Jensen, "Graph clustering with network structure indices", ICML '07, proceeding of 24th international conference on machine learning, pages 783- 790

[12]. Fillipo Radicchhi, Claudio Castellano, "Defining and identifying communities in a network", Proceedings of National Academy of Sciences of USA, vol 101, no. 9, pages 2658 – 2663.

[13]. M.E.J. Newman, "Modularity and community structure in network",2006, PNAS, Vol 103 no. 23, pages 8577 – 8582.

[14]. Aaron Clauset, M.E.J. Newman, Cristopher Moore, "Finding community structure in very large network" ,2004, Physical Review E 70.

[15]. Vincent D. Blondel, Jean Loup Guillaume, Renaud Lambiotte, Ettienne Lefebvre, 2008, IOP Science.

[16]. Pascal Pons, Matthieu Latapy, "Computing Communities in Large Network using Random Walk", 2006, Journal of Graph Algorithms and Applications, Vol. 10, no. 2, pages 191 – 218.

[17]. A.G. Nikolaev, R. Razib, A. Kucheriya, "On efficient use of entropy centrality for social network analysis in social network and community detection", Social Networks, 2015, Elsevier.

[18]. K. Steinhaeuser, N.V. Chawla, "Identifying and evaluating community structure in complex networks", Pattern Recognition Letters, 2010, Elsevier.

[19]. Balázs Adamcsek, Gergely Palla Illés, J. Farkas, Imre Derényi, Tamás Vicsek, "CFinder: locating cliques and overlapping modules in biological networks", bioinformatics, Volume 22, Issue 8,pages 1021- 1023.

[20]. Gergely Palla, Imre Derenyi, Illes Farkas, Tamas Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society", 2005, Nature 435, pages 814- 818.

[21]. JM Kumpula, M Kivelä, K Kaski, J Saramäki, "Sequential algorithm for fast clique percolation", Physical Review E, 2008.

[22]. Yong- Yeol Ahn, James P. Bagrow, Sune Lehmann, "Link communities reveal multiscale complexity in networks", 2010, Nature 466, pages 761-764.

[23]. T.S. Evans, R. Lambiotte, "Line Graphs, Link Partitions and Overlapping Communities", 2009, Phy. Rev. E.,arXiv:0903:2181[physics.soc.ph]

[24]. T.S. Evans "Clique Graphs and Overlapping Communities", 2010, arXiv:1009:0638[physics.soc-ph].

[25]. J.Baumes, M. Goldberg, Malik Magdon-Ismail," Efficient Identification of Overlapping Communities",2005, Intelligence and Security Informatics, pages 27- 36.

[26]. Stephen Kelley,Mark Goldberg, "Defining and Discovering Communities in Social Networks", 2011, Handbook of Optimization in Complex Networks, pages 139- 168.

[27]. A. Lancichinetti, Santo Fortunato, Janos Kertesz, "Detecting the Overlapping and hierarchical community structure in complex networks", 2009, New journal of physics, Volume 8.

[28]. T. Nepusz, A. Petroczi, L. Negyessy, F. Bazso, "Fuzzy communities and the concept of bridgeness in complex networks", 2008, Phys. Rev. E 77,016107.

[29]. J.Xie, S.Kelley, B.K. Szymanski, "Overlapping Community Detection :State of Art and Comparative Study" , 2013, ACM Comput. Surv. 45, 4, Article 43 .

International Journal of Engineering Science Invention (IJESI) is UGC approved Journal with Sl. No. 3822. Journal no. 43302.

Dipika Singh "Some Observations on Community Detection Algorithms in Social Networks." International Journal of Engineering Science Invention(IJESI), vol. 6, no. 12, 2017, pp. 54-59.