International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-1, January 2016
ISSN: 2395-3470
www.ijseas.com

Survey of Community Detection Algorithms to Identify the Best Community in Real-Time Networks

Dhanya Sudhakaran[1], Shini Renjith[2]
[1] PG Scholar, Department of CSE,
Sree Buddha College of Engineering, Alappuzha, India
dhanyasudhakaran1991@gmail.com

[2] Assistant Professor, Department of CSE,
Sree Buddha College of Engineering, Alappuzha, India
shinirenjith@gmail.com

Abstract
Community detection is a common problem in graph data analytics. It consists of finding groups of densely connected nodes with few connections to nodes outside of the group. In particular, identifying communities in large-scale networks is an important task in many scientific domains. Community detection algorithms are used to study the structural properties of real-world networks. In this review, we evaluated some of the traditional algorithms for overlapping and disjoint community detection on large-scale real-world networks to identify the best community in real- time networks.

Keywords: *Community detection, Graph Partitioning, Overlapping Community, Label Propagation.*

1. Introduction

Detecting clusters or communities in large real-world graphs such as large social or information networks is a problem of considerable interest. Community detection is a common area in graph data computations and data mining computations [1] [2]. It consists of finding groups of densely connected nodes with few connections to nodes outside of the group. Networks can be either multi-dimensional networks or uni-dimensional networks. Multi-dimensional networks are networks with multiple kind of relations. Examples of multi-dimensional networks are social networks, genetic networks, co-citation networks. Each node in a network is an item corresponding to a dimension or entity in a network and each edge indicates a relationship between two nodes. Figure 1 shows a network having three communities. In social networks, finding a community structure means finding a group of users who interact on different entities like tags, photos, comments or stories. In case of a co-citation network, community structure represents a group of authors who interact on publication information such as titles, abstracts, keywords etc. Detecting communities is of great importance in sociology, biology, computer science etc. In particular, identifying communities in large-scale networks is an important task in many scientific domains. Large-scale networks with thousands to millions of nodes are common across many scientific domains. Finding community structures from this networks are of particular interest. Identifying communities in a large-scale network is a complex task because there exists many definitions of community and intractability of the community detection algorithms. The community detection problem has many widespread applications and has hence proven to be very important. This survey reviews about the different community detection algorithms and methods for finding the best community in a network. The best community implies one with less amount of noisy interactions among the networks. The rest of the paper is organized as follows. Some of the surveys done in the area of community detection are presented in section 2. A brief study of the community detection algorithms and approaches are presented in section 3. Results and discussions are presented in section 4. The concluding remarks are given in section 5.

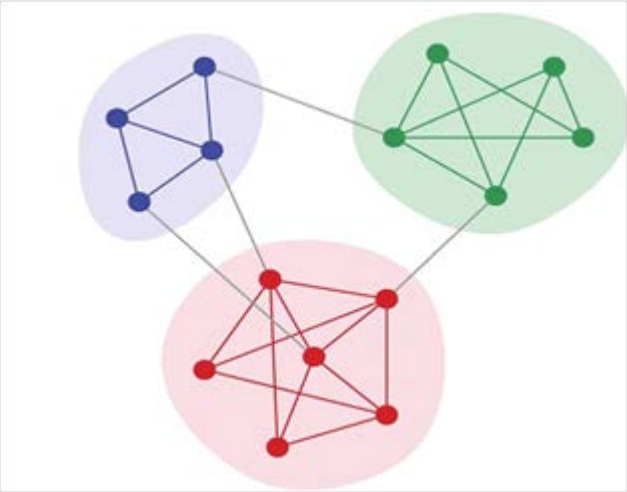

Fig 1. Network with three communities

2. Related Works

In recent years, several surveys have been published in the area of community detection. Moradi and Olovsson et al. [3] experimentally evaluated the qualitative performance of several community detection algorithms using large-scale E-mail networks. They compared the quality of the algorithms with respect to a number of structural quality functions. Leskovec and Lang et al. [4] presented an empirical comparison of algorithms for network community detection. They evaluate several common objective functions that are used to formalize the notion of a network community and examined several different classes of approximation algorithms that aim to optimize such objective functions. Lancichinetti and Fortunato [5] provided a serious assessment of the goodness of community detection algorithms. Malliaros and Vazirgiannis [6] focused on community detection algorithms for directed graphs, and suggested a methodology-based taxonomy to classify the different algorithmic approaches.

3. Community Detection

A community could be loosely described as a collection of vertices within a graph that are densely connected amongst themselves while being loosely connected to the rest of the graph. Bagrow and Bolt [7] proposed a local method for detecting communities. They proposed an algorithm which consists of a shell 'l' spreading outward from a starting vertex. As the starting vertex's nearest neighbors and next nearest neighbors etc. are visited by the shell 'l', two quantities are computed: emerging degree and total emerging degree. Algorithm works by expanding the shell outward from some starting vertex 'j' and comparing the total emerging degree change to some threshold value. When the 'l' shell ceases to grow, all vertices covered by shells of a depth <= l are listed as members of vertex j's community. Ruan and Zhang [8] proposed a quantitative measure called modularity to assess the quality of community structures. Modularity means the measure of fraction of edges falling within communities subtracted by what one would expect if the edges are randomly placed. It provides a good quality measure to compare different community structures. A larger modularity value means stronger community structures. Newman [9], Duch and Arenas [10] proposed an algorithm by optimizing the modularity measure.

A graph partition method based on min-max clustering principle was proposed by Ding and Zha et al. [11]. The principle states that the similarity or association between two subgraphs is minimized, while the similarity or association within each subgraph is maximized. Luo and Wang et al. [12] proposed a framework to identify modules within a biological network. Networks are divided into sub-networks and the identification of modules is based on their topology. For this, the concept of edge-betweenness was used. Edge-betweenness is the number of shortest path between all pairs of vertices that run through the edge. Edges between modules tend to have shortest paths through them than do edges inside modules and thus have higher betweenness values. The deletion of edges with high betweenness can separate the network, while keeping the modules structure in the network intact. Sun and Castro et al. [13] proposed a framework, MetaFac that extracts community structures from social media networks. Mehler and Skicna [14] presented a general method for network community expansion from seed set of members. It is achieved by assigning a score to all entities in the network and selecting the highest scoring outside vertex to join the community. Some of the scoring criteria in order to rank the selection are neighbor count, juxta position count,

neighbor ratio, juxta position ratio, binomial probability. The essential function of the community expansion method is to identify the most promising next member to be added to the community. Some representative community detection methods [15] such as latent space models, block model approximation, spectral clustering and modularity maximization.

Adaptive algorithms were developed for detecting community structures in dynamic social networks. Quick Community Adaptation (QCA) [16] is an adaptive modularity based method for identifying and tracing community structure in dynamic social networks. Modularity based approaches are used for finding community structures in very large networks. Modularity is a property of a network and a specified proposed division of that network into communities. Clauset and Newman [17] proposed an algorithm based on the modularity property of the network. Chikhi and Rothenburger [18] proposed a probabilistic approach known as Smoothed Probabilistic Community Explorer (SPCE), a generative model for community structure identification. SPCE provides several advantages. It finds coherent and overlapping community structures. It takes as input only the number of communities to identify and not their size. It detects communities in directed and undirected networks. It provides a two-view community structure in directed networks and is able to analyze weighted and unweighted networks. Usha and Reka et al. [19] proposed a localized community detection algorithm based on label propagation. For finding overlapping communities in large networks label propagation method [20] can be used. Chang and Yi-Hsu et al. [21] also developed a general probabilistic framework for detecting community structure. Key idea of generalization is to characterize a network by a bivariate distribution that specifies the probability of the two vertices appearing at both ends of a randomly selected path in the graph. Riedy and Bader et al. [22] presented a greedy agglomerative algorithm that grows a community around a given small seed set. Starting from a set of seed vertices the algorithm pull adjacent vertices into the community to maximize modularity. Random walk process can be used to compute communities in large networks. Such an algorithm known as walktrap was proposed by Pons and Latapy [23]. Improved community detection algorithm based on random walk by taking into account node attribute information was proposed by Daxiang and Sun et al. [24] Random walk process can also be used for detecting community structures for undirected graphs [25]. For finding overlapping communities Ball and Karrer et al. [26] proposed a method based on a statistical approach using generative network models. Algorithms named RaRe (Rank Removal) and IS (Iterative Scan) were proposed by Baumes and Goldberg et al. in [27]. IS iteratively constructs clusters and RaRe attempts to identify high ranking nodes and remove them from graph, in order to disconnect the graph into smaller connected components.

A visual data mining approach to find overlapping communities in networks was proposed by Chen and Osman et al. The proposed algorithm was known as ONDOCS (Ordering Nodes to Detect Overlapping Community Structure), helps the user to make appropriate parameter selections by observing initial data visualizations and finds and extracts overlapping community structures from the network. A game theoretic framework to address the community detection based on the structures of the social networks was proposed [28] to find the overlapping communities. Community formation s formulated as a strategic game known as community formation game. A modification to Cluster Overlap Newman- Girman Algorithm (CONGA) was proposed by Gregory [29] known as CONGA Optimized (CONGO). Overlapping communities were detected by using the concept of split-betweenness. A two-phase method of overlapping communities was proposed in [30] known as Peacock Algorithm. In the first phase, a network is transformed to a new one by splitting vertices using the idea of split betweenness. In the second phase, the transformed network is processed by a disjoint community detection algorithm. This approach had the potential to convert any disjoint community detection algorithm into an overlapping community detection algorithm.

4. Results and Discussions

Table 1. Comparative Study of Some Algorithms

Algorithm	Overlapping Communities	Directed Graph	Weighted Graph	Input Parameters
LPA	No	Yes	Yes	O(m)
Fastgreedy	No	No	Yes	$O(n\log^2 n)$
Walktrap	No	No	No	$O(mn^2)$
ONDOCS	Yes	No	No	O(nlog n)

From the survey it can be inferred that community structure identification at the early times were applicable only in the case of uni-dimensional networks. In literature, many methods have been proposed to extract community structures from uni-dimensional networks. However these methods may not be used to yield good performance for multi-dimensional networks. For analyzing large networks such as social networks where the user changes are constantly changing and co-evolving considering the uni-dimensionality may be critical. Thus multiple dimensions have to be considered. All the different s for community detection vary differently in case of accuracy, efficiency and their complexity. Table 1 shows a comparative study of some algorithms implemented on graphs (networks) with n nodes and m vertices, based on whether they could be implemented on directed graph or weighted graph.

5. Conclusion

Community detection algorithms are widely used to study the structural and topological properties of real-world networks. In this review, we have evaluated some of the community detection approaches for overlapping and disjoint community detection on large-scale real- world networks. 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.

References

[1] Shini Renjith, C Anjali, “Fitness function in genetic algorithm based information retrieval: A Survey”, ICMIC13, December 2013, pp. 80-86.

[2] Dhanya Sudhakaran, Shini Renjith, “Phase based resource aware scheduler with job profiling for MapReduce”, IJLTET, vol 6, Issue 2, Nov 2015, pp.92-96.

[3] F Moradi, T Olovsson, P Tsigas, “An of community detection algorithms on large-scale email traffic”, in: SEA. Berlin/Heidelberg: Springer; 2012; 283–294.

[4] J Leskovec, K.J Lang, M.W Mahoney,” Empirical comparison of algorithms for network community detection”, CoRR, abs/1004.3539, 2010.

[5] A Lancichinetti, S Fortunato, “Community detection algorithms: a comparative analysis”, Phys Rev E 2009, 80:056117.

[6] F.D Malliaros, M Vazirgiannis, “Clustering and community detection in directed networks: a survey”, CoRR, abs/1308.0971, 2013.

[7] J. Bagrow and E. Bolt, “Local Method for Detecting Communities,” Physical Rev. E, vol. 72, no. 4, p. 046108, 2005.

[8] J. Ruan and W. Zhang, “An Efficient Spectral Algorithm for Network Community Discovery and Its Applications to Biological and Social Networks,” Proc. Seventh IEEE Int’l Conf. Data Mining (ICDM ’07), pp. 643-648, Jan. 2007.

[9] M. Newman, “The Structure and Function of Complex Networks,” SIAM Rev., vol. 45, no. 2, pp. 167-256, 2003.

[10] Jordi Duch, Alex Arenas, “Community detection in complex networks using extremal optimization”, Phys, Rev E72, 027104, August 2005.

[11] C.H.Q. Ding, X. He, H. Zha, and M. Gu and H.D. Simon, “A Min- Max Cut Algorithm for Graph Partitioning and Data Clustering,” Proc. IEEE Int’l Conf. Data Mining, pp. 107-114, 2001.

[12] F. Luo, J.Z. Wang, and E. Promislow, “Exploring Local Community Structures in Large Networks,” Web Intelligence and Agent Systems, vol. 6, no. 4, pp. 387-400, 2008.

[13] Y. Lin, J. Sun, P. Castro, R. Konuru, H. Sundaram, and A. Kelliher, “Metafac: Community Discovery via Relational Hypergraph Factorization,” Proc. 15th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD ’09), pp. 527-536, 2009.

[14] A. Mehler and S. Skiena, “Expanding Network Communities from Representative Examples,” ACM Trans. Knowledge Discovery from Data, vol. 3, no. 2, article 7, 2009.

[15] 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.

[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] Aaron Clauset, M. E. J. Newman, and Cristopher Moore, "Finding community structure in very large networks", Phy Rev E70, vol.6, December 2008.
[18] Nacim Fateh Chikhi, Bernard Rothenburger, Nathalie Aussenac-Gilles, "Community Structure Identification: A Probabilistic Approach", conference paper, IEEE explore January 2010.
[19] Usha Nandini Raghavan, R eka Albert, and Soundar Kumara, "Near linear time algorithm to detect community structures in large-scale networks", Physical Review E, 76(3):036106, 2007.
[20] Steve Gregory, "Finding overlapping communities in networks by label propagation", New journal of Physics, October 2010.
[21] Cheng-Shang Chang, Chin-Yi Hsu, Jay Cheng, and Duan-Shin Lee, "A General Probabilistic Framework for Detecting Community Structure in Networks"
[22] Jason Riedy, A. David Bader Karl, Jiang Pushkar Pande, Richa Sharma," Detecting Communities from Given Seeds in Social Networks", February 22, 2011.
[23] Pascal Pons and Matthieu Latapy, "Computing communities in large networks using random walks", ISCIS, Springer, vol 3733, 2005, pp.284-293.
[24] 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.
[25] Xiaoming Liu, Yadong Zhou, Chengchen Hu, Xiaohong Guan, Junyuan Leng, "Detecting Community Structure for Undirected Big Graphs Based on Random Walks", www'14 Proceedings of 23rd international conference on WWW, 2014, pp.1151-1156.
[26] Brian Ball, Brian Karrer, M.E.J Newman, "An efficient and principled method for detecting communities in networks", April 2011.
[27] Jeffrey Baumes, Mark Goldberg, Mukkai Krishnamoorthy, "Finding communities by clustering a graph into overlapping subgraphs," IADIS International Conference on Applied Computing 2005.
[28] Wei Chen, Zhenming Liu, Xiaorui Sun, Yajun Wang, "A game-theoretic framework to identify overlapping communities in social networks", Data Min Knowl Disc (2010) 21:224–240.
[29] S. Gregory, "A fast algorithm to find overlapping communities in networks," in PKDD, 2008, pp. 408–423.
[30] S. Gregory, "An algorithm to find overlapping community structure in networks," in PKDD, 2007, pp. 91–102.

Ms. Dhanya Sudhakaran is an M. Tech Post Graduate Scholar specialized in Computer Science and Engineering. She has pursued B.Tech in Computer Science and Engineering in 2013 under the University of Kerala. She has published a paper in the area of data mining and MapReduce. Her areas of interest include data mining, Big data Analytics, Cryptographic Security. Her research area is community detection in multi-dimensional networks. She is a member of Association for Computing Machinery (ACM) and Computer Society of India (CSI).

Ms. Shini Renjith has M.Tech in Computer Science from University of Kerala (2014) and B.Tech in Computer Science and Engineering from Cochin University of Science and Technology, CUSAT (2004). Also she is currently pursuing her PhD from CUSAT and is working as an Assistant Professor at Computer Science and Engineering Department in Sree Buddha College of Engineering, Alappuzha. Prior to this she has worked at College of Engineering, Munnar and College of Engineering, Thiruvananthapuram. She is qualified in UGC NET, has won the Best Paper Award at International Conference on Mobility in Computing [ICMiC13] and has presented and published 6 papers in various international conferences and journals. Her current areas of interests include big data analysis, Information filtering and computer networks. She is an active member of Association for Computing Machinery (ACM) and IEEE.