IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-10, Oct. 2018 E-ISSN: 2347-2693

A Comprehensive Analysis of Dis-Joint Community Detection Algorithms for Massive Datasets

Kamal Sutaira[1*], Kalpesh Wandra[2], C. K. Bhensdadia [3]

[1,2] Dept. of Computer Engineering, C. U. Shah University, Wadhwan City, Gujarat, India
[3]Dept. of Computer Engineering, Dharamsinh Desai University, Nadiad, Gujarat, India

[*]*Corresponding Author: kamal.sutaria@gmail.com, Tel.: +91-94282-32881*

Available online at: www.ijcseonline.org

Accepted: 10/Oct/2018, Published: 31/Oct/2018

***Abstract*—** With the growth of Internet and computer knowledge, more and more persons connect socially. People communicate with each other and express their views on social media, which may form a complex network of association. Entities in the social networks create a "relation structure" through several connections which produces a huge amount of information. This "relation structure" is the group or community that we are interested in research. Community detection is very imperative to disclose 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 essential field of social networks analysis, a large number of the academic literature suggested several approaches to community detection. In this paper, we first describe the concepts of the social network, community, community detection and criterions of community quality. Then we classify the methods of community detection into the following categories. And at last, we summarize and discuss these methods as well as the potential future directions of community detection.

***Keywords*—** Social Network Analysis, Community Detection, Graph Data, Massive Datasets, Disjoint Community Detection

I. INTRODUCTION

In today's scenario, social media is an emerging field for many researchers. In social media the data generated through the user side is enormous. To maintain the user-generated data, there are many mining tasks are present in social media mining. There are many social networking sites where the user makes their community from their interest. As it is known that social media is a big virtual world in that many users have their profile, and they are connected to different types of groups. To see the behavior of the user it needs to understand the background of the user. It is not that easy in the social network to identify the action of the single-use. Therefore it is required to perform community detection in the social network. Many researchers had done a lot of work in this field of the social network

Social media mining is a process of representing extracting and analyzing actionable patterns from social media data. Social media shatters the boundaries between the real world and the virtual world. We can now integrate social theories with computational methods to study how individuals interact and form communities. The uniqueness of social media data is for novel data mining techniques that can effectively handle user-generated Content with the vibrant social relation. There are much-emerging Research areas in social media mining. The most known research area of social media mining is community detection.

Community detection is a process of detecting communities form in social media by ground truth given from social media data. Here we are doing community detection based on influence. In community detection data points are defined as actors in social media and similarity between actors are determined based on the interest these user shares. In social networking sites, the only fraction of user gets influenced by other users. Community detection has received attention in all kinds of networks, such as social network, biological network and the World Wide Web. Now we discuss social forces through which users or nodes are connected in social communities.

In this work, we organized as follows. Section I gives Introduction. Section II Social Media Mining. Section III Community Detection in SNA. Section IV offers background study and literature review for SNA. Section V presents the parameters to evaluate communities, and the last section contain a conclusion.

II. SOCIAL MEDIA MINING

Social media mining is a process of visualizing, evaluating and extracting useful patterns over the Social network [1].

Through Social media mining, they have integrated social theories with computational methods. Social media mining defines basic principle and concepts for investigating the massive amount of social media data. In this mining, they have discussed different disciplines such as computer science, data mining, social media, machine learning, etc.
For social media mining they have encompasses the tools to formally represent, model, measure and extract meaningful pattern for large social media networks. Social media sites generate user data which is different from traditional attribute-values of data for Hellenic data mining. The data which is produced from social sites are noisy, distributed, not in proper structure and frequent. All the characteristics of social media data pose challenges for data mining task and for that new techniques and algorithm have to be developed. Following are examples for communities for well-known datasets [2].

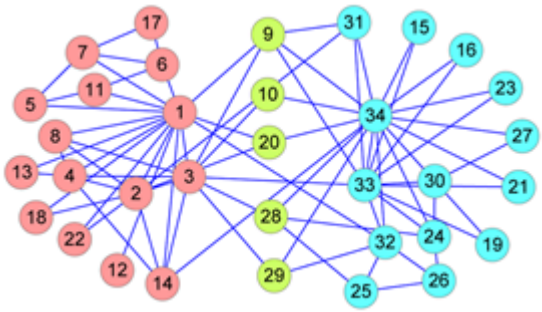

Fig: 1 Community Structure of Zachary Karate Club

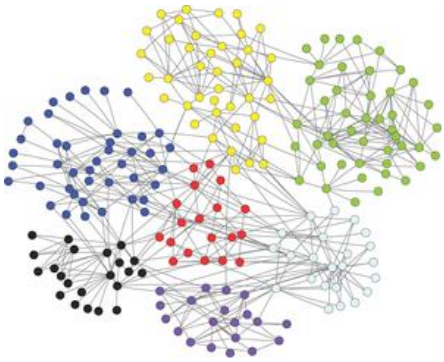

Fig: 2 Community Structure of Random Network

III. COMMUNITY DETECTION IN SNA

It is a process of detecting communities in the social network. Community detection is essential in social media, due to many reasons. First, users create the group by their interest. There is two type of communities; explicit communities and implicit communities. It means users need to subscribe personally. For example, many social networks provide some communities, which are predefined so the user may join or may not. An implicit community means the user has not subscribed personally. So from given network, the aim is to identify the various communities from user's behavior or interest.

Next section contains the review of current literature on various algorithms for unfolding the communities and other similar problems. Initially, the multiple efforts done on identifying disjoint communities in several networks are discussed and later numerous metrics used to evaluate the community structures.

IV. BACK GROUND STUDY AND LITERATURE REVIEW

A varied range of community detection techniques has been recently derived to unfold disjoint groups or communities from the steady network. Following Some of the review, papers to see, i.e., Fortunato,[2] lancichinetti, Fortunato et al.[3] and harenberg.[4] Later these algorithms could be roughly split into the following categories.

- **Traditional Methods of Community Detection**

Community Detection using graph partition: The issue of graph partitioning contains isolating the nodes into numerous clusters of predefined size, such that quantity of edges lying between the groups is as possible as the minimum. The number of edges available between the communities is known as cut size. In community detection so many algorithms that could perform a better job, moreover, its results aren't essentially decent.[5,6] Graphs could also be partitioned by minimizing parameters which can be related to the cut size, like conductance,[7] normalized cut[8] and ratio cut.? Another well-known technique may be the spectral bisection method that will be on the basis of the assets of the spectral range of the Laplacian matrix.[9] These algorithms of graph partitioning are bad to revealing the community because initially, it requires as an input how many partition/groups and sometimes its sizes too.

Spectral clustering: This section contains all the approaches and techniques that partition the group of nodes into the cluster. Specifically, the objects might be data/points in vector space, or the vertices of any graphs by means of the eigenvectors of matrices and other matrices based on it. Spectral clustering includes an alteration of the first group of nodes into some points in space, and whose data are elements of eigenvectors. The collection of data is then grouped via some well-known standard techniques, clustering through k-means. The very initial method has been derived by Donath and Hoffmann [10] for spectral clustering. There are three standard types of the spectral clustering, the first one is unnormalized spectral clustering by Shi and Malik;[8] and the other two techniques are proposed by Ng et al. [11] On the other hand, the Nadler and Galun [12] claimed that the restrictions of such techniques for a particular example cannot group datasets which are of different size and density.

Clustering using Partition: This type of approach assumes that the number of group or cluster is predefined, k. the data are stable in a metric space, so each and every node is considered as a poia nt and a distance is defined among the data points in the space. The distance is a measure of the variation among nodes. The aim is split up the data points in a number of k clusters in order to minimize/ maximize a cost function depends on distances between points and/or from points to centroids. Some functions include minimum k-clustering, k-center, k-median, k-clustering sum. The widely used partitioning method in the literature is k-means clustering. [13] Later on an improved version of k-means clustering to networks have been proposed by Hlaoui and Wang. [14,15] However the drawback of this approach is just like the graph partitioning algorithms i.e. initially specified the total number of clusters as well as the technique is unable to derive it.

Hierarchical clustering: The real-world graphs mainly have a hierarchical structure where we can see different levels of a combination of nodes. One combination of such nodes can be called a group. In this type of situation, one small cluster is included in the large one and that is ultimately part of the main superior cluster. In this type of situation, we have to use hierarchical clustering algorithms. [16] The hierarchical clustering algorithms can uncover the multilevel structure of the graph. These algorithms can be divided into two parts: Divisive (top-down) algorithms and Agglomerative (bottom-up) algorithms. This approach has the benefit that it doesn't need a prior understanding of the size and number of the groups or clusters. Further to this, it does not offer the technique to pick the partition that can better constitute the community of the graph.

- **Divisive Algorithms for Community Detection**

The divisive algorithms can detect the edges which connect the nodes of different communities and then removing them, so as the cluster gets disconnected from each other. Girvan and Newman has proposed the standard algorithm for the same. They have used edge betweenness as a parameter to select the edges. Tylor proposed the modification of the method which reduced the time for community detection and made it faster. Rattigan et al. [17] too, proposed the modification over the basic Girvan Newman method to further make the community detection calculation faster. In that, they have used an approximation of the edge betweenness values using network structure index that included a couple of nodes along with the distance measure. The modified version of the GN algorithm is proposed by Tyler et al. to improve the calculation speed for community detection. [18]

- **Modularity-based Algorithms for Community Detection**

Modularity is first presented by Girvan and Newman, [19] the most used and best known quality parameter to check the community structure. The basic foundation of modularity is the fact that the real graph or the random graph does not contain a specific cluster pattern. Due to this, the strength of the cluster cannot be measured directly. It can be measured by the contrast among the original density of the edges within a community and the density you might have a much in the community if the vertices of the graph were involved irrespective of community building. Modularity can be written as follows:

$$Q = \frac{1}{2M} \Sigma_i^j [A_{ij} - \frac{K_i K_j}{2M}] \delta(C_i, C_j)$$

Where the summation is calculated over all pairs of nodes, A is the adjacency matrix of an input network/graph, m is the total number of edges of the graph, ki indicates the degree of the node named i, the value of the delta function is 1 if the node i and node j are in the same group; otherwise the function value is 0. From this it can be clear that higher the value of the modularity is, higher the quality of the partition is. Newman [20] proposed the first ever algorithm to increase the modularity using the greedy method. The greedy methods always work by choosing the locally best solution rather than the global one and thus will work faster. In the method suggested by Newman, [20] the set of nodes are merged if the modularity increases by doing so.

- **Dynamic Algorithms**

Right here this phase describes strategies the use of procedures running at the graph, concentrating on spin-spin interactions, random walks and synchronization. Random walks [21] also can be handy to unfold the groups from the network. If a graph includes a sturdy community shape, a random walker spends pretty a long time inside a network because of the excessive density of inner edges and next range of paths that would be observed. Zhou [22] used random walks to describe a distance among pairs of nodes: the distance dij among I and j is the common range of edges that a random walker has to move to attain j starting from i.

- **Statistical Inference based Community Detection**

Statistical inference aims at deducing properties of datasets, beginning with some observation and model hypotheses. For the graph data, the model, primarily based on hypotheses on how nodes are connected to each other, has to in shape the real graph. Maximum of the methods adopted Bayesian inference, [23] in which the best fit is acquired through the maximization of a probability. Newman and Leicht [24] designed a method primarily based on a combination model and the expectation-maximization technique. The main negative point is required high memory desires for these methods.

- **Mixed Methods**

Right here it comprises some strategies that don't suitable in the earlier classes. Raghavan et al.[25] developed a label

propagation, which is considered as an easy and rapid technique for detecting communities from network. The key benefit of the technique is the truth that it generally does not need any informative data on the number and the size of the clusters. It generally does not need any parameter, either.

V. COMMUNITY EVALUATION PARAMETERS

Another important aspect of community detection is to evaluate the detected community structure. If we know the specific community structure of a network, it will be easier to judge the detected communities simply by comparing them with the specific community structure. Moreover, all the time, collecting the specific ground-truth community structure is tough, and therefore we depend on the basic property of the community structure.

o Ground-truth Based Metrics for Evaluating Communities
Evaluating the quality of a detected communities is nontrivial, and extending evaluation measures for disjoint communities is hardly straightforward. In this section, we discuss a few of the popxular evaluation metrics which are generally used to compare the detected communities with the ground-truth communities.

1. In ***purity***, we assume that the majority of a community represents the community. Hence, we use the label of the majority of the community against the label of each member of the community to evaluate the algorithm.

$$Purity = \frac{1}{N}\sum_{i=1}^{k} \max_{j} |C_i \cap L_j|,$$

It is an important to observe that the purity is not a symmetric measure. Therefore, the typical approach is to take the harmonic mean of PU(X,Y) and PU(Y,X). The upper limit is 1, it corresponds to a perfect match between the groups. The lower limit is 0 and indicates the totally mismatch among the groups.

2. The ***Rand Index*** [26] is a way of relating non-overlapping groups that is based on pairs of the nodes being grouped. Two resolutions are said to agree on a pair of nodes if they each put both nodes into the same group or each into different groups. This can be formalized as follow:

$$RandIndex = \frac{(a+b)}{N}$$

Where N is the range of pairs of objects, a is the wide variety of instances solutions agree on putting a couple within the equal group and d is the wide variety of instances answer agree on setting a couple with the dissimilar group.

3. ***Conductance*** has been also widely used for measuring the detected community of a given network. For instance Leskovec et al. [27] presented the concept of network community profile plot to measure the quality of a 'good' community as a characteristic of community size in a graph. The conductance of a set of vertices is the ratio of edges leaving to the total edges.

$$\emptyset(S) = \frac{cut\ (S)}{min\big(vol\ (S), vol\ (\bar{S})\big)}$$

They have got used conductance to measure the goodness of a group and analyze a huge range of groups of various size in real-world social networks.

4. Density: Another important parameter to measure the quality of detected communities of network is density (D). Density can be defined as follow:

$$D = \frac{2E}{V * V - 1}$$

Where E is the total number of edges and V is the total number of nodes of a given network. The maximum number of edges of a graph is (V‡V−1) 2 so the maximum value of D is 1, and the minimum value for D is 0.

VI. CONCLUSION

In this work, several state-of-the-art community detection algorithms for disjoint community are analyzed. The analysis can be applied in a dynamic environment for communities using machine learning techniques. Edge weights have a major role in determining the strength of node in a community. Few researches are made considering this edge weight as a key role in community detection in the social network field. Community detection algorithms are widely used to study the structural and topological properties of real-world networks. Here, we have evaluated some of the community detection approaches for 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] J. Bruhn, "The sociology of community connections," Springer Science+Business Media B.V., 2011.
[2] S. Fortunato, "Community detection in graphs," Physics Reports, vol. 486(3-5), p. 75 – 174, 2010.
[3] A. Lancichinetti and S. Fortunato, "Community detection algorithms: A comparative analysis," Phys. Rev. E, 2009.

[4] L. G. S. R. J. H. R. S. K. P. S. Harenberg, G. Bello and N. Samatova, "Community detection in large-scale networks: a survey and empirical evaluation," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 6(6), p. 426–439, 2014.
[5] S. L. B.W. Kernighan, "An efficient heuristic procedure for partitioning graphs," Bell System Technical Journal, vol. 49(2), pp. 291–307, 1970.
[6] A. Pothen, "Graph partitioning algorithms with applications to scientific computing," Technical report.
[7] B. Bollob´as, "Modern graph theory," Graduate texts in mathematics, 1998.
[8] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern Anal. Mach. Intel., vol. 22(8), p. 888–905, 2000.
[9] E. R. Barnes, "An algorithm for partitioning the nodes of a graph," Technical Report RC 08690, 1981.
[10] W. E. Donath and A. J. Hoffman, "Lower bounds for the partitioning of graphs," IBM J. Res. Dev., vol. 17(5), p. 420–425, 1973.
[11] A. Y. Ng and Y. Weiss, "On spectral clustering: Analysis and an algorithm," In Advances in Neural Information Processing Systems, vol. 14, p. 849–856, 2001.
[12] B. Nadler and M. Galun, "Fundamental limitations of spectral clustering methods," Advances in Neural Information Processing Systems, vol. 19, 2007.
[13] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," In L. M. L. Cam and J. Neyman, editors, Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297, 1967.
[14] J. C. Bezdek, "Pattern recognition with fuzzy objective function algorithms," 1981.
[15] A. Hlaoui and S. Wang, "Median graph computation for graph clustering," Soft Computing, vol. 10(1), pp. 47–53, 2006.
[16] R. T. T. Hastie and J. Friedman, "The elements of statistical learning," Springer Series in Statistics, 2001.
[17] M. M. M. J. Rattigan and D. Jensen, "Graph clustering with network structure indices," In Proceedings of the 24th International Conference on Machine Learning, ICML '07, New York, NY, USA, ACM, pp. 783–790, 2007.
[18] D. M. W. J. R. Tyler and B. A. Huberman, "E-mail as a spectroscopy: Automated discovery of community structure within organizations," In M. Huysman, E. Wenger, and V. Wulfs, editors, Proceedings of the First Iternational Conference on Communities and Technologies., 2003.
[19] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," Physical Review Edition, vol. 69, 2004.
[20] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," Phys. Rev. Edition, vol. 69:066133, 2004.
[21] B. Hughes, "Random walks and random environments: Random walks," Number v. 1 in Oxford science publications, 1995.
[22] H. Zhou, "Distance, dissimilarity index, and network community structure," Physical Review Edition, vol. 67(6):061901, 2003.
[23] R. Winkler, "An introduction to bayesian inference and decision," International series in decision processes, 1972.
[24] M. E. J. Newman and E. A. Leicht, "Mixture models and exploratory analysis in networks," Proceedings of the National Academy of Sciences, vol. 104(23), p. 9564–9569, 2007.
[25] R. A. U. N. Raghavan and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," Physics review Edition, vol. 76:036106, 2007.
[26] W. Rand, "Objective criteria for the evaluation of clustering methods," Journal of the American Statistical Association, vol. 66(336):, pp. 846–850, 1971.
[27] A. D. J. Leskovec, K. J. Lang and M. W. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," Internet Mathematics, vol. 6(1), p. 29–123, 2009.

Authors Profile

Kamal Sutaria, pursed M.E. in Computer Engineering from Dharamsinh Desai University, Nadiad, India in 2010 and currently pursuing Ph. D in Computer Engineering in the field of Social Network Analysis from C. U. Shah University.

Prof. (Dr.) Kalpesh H. Wandra, is currently working as a principal and professor in Gujarat Maritime Board, Rajula, Gujarat. He has done his Ph.D. in Computer Engineering. He has done his Master in Electrical Engineering. He has more than 15 years of academic experience in the field of Computer Engineering. He has written more than 30 research papers in various fields.

Prof. (Dr.) C. K. Bhensdadia, is currently working as a Head and Professor in Department of Computer Engineering, Dharamsinh Desai University, Nadiad since 2004 and is one of the 15 members of All India Board of Information Technology Education and is the only member from the state of Gujarat. He plays key role in curriculum design of various universities. He has contributed significantly towards projects for rural development for Department of Information Technology, Ministry of Communication and IT, Government of India and has also worked for consultancy project of a European country. He is with the department as a faculty since 1990 and was a student from 1986-90, in the first B.Tech. Computer engineering batch of the university. He has done his Ph.D. in the field of Computer Engineering. He has Guided more than 50 M.Tech. Dissertation in various areas.

