International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-3, Issue-3, May-2015

Identification of Communities from Social Networks

Seema Rani, Monica Mehrotra

***Abstract*—A social network is a social structure of people, related (directly or indirectly) to each other through a common relation or interest. Social network analysis (SNA) is the study of social networks to understand their structure and behavior. For studying structural and behavioral properties of these networks, communities are identified by grouping of individuals according to given context into subgroups. Community detection is very rich domain in social network analysis as it is useful in various domains like business, marketing, healthcare etc. Data analytic techniques such as data mining and predictive modeling are being used to gain new insights into social network analysis (SNA). This has the unique ability to play a new role in exploring the context and situations that lead to efficient and effective predictions. Identifying these social communities can bring benefit to understanding and predicting user's behaviors. This paper is an attempt to study the various approaches for community detection (CD), application area of CD and evaluation of CD algorithm. It also presents the emerging and ongoing research towards improvement in existing CD algorithms in the area of social network analysis.**

***Index Terms*— Community Detection, Evaluation of Identified Communities, Healthcare, Overlapping Community Detection, Social Network.**

I. INTRODUCTION

In recent times, user activities on web based social networks has increased enormously irrespective of time and place that generates magnanimous datasets which offers tremendous scope for both mining interesting user behavior and knowledge discovery. Social Networking now a days is considered as one of the most important feature as so many critical; activities are depended on it. In this paper, the basic concept of social networking and various terminologies related to social network are discussed. The study focuses on the concept of social network and community structure which is considered as one of the most important features of social network and also the importance of detecting these communities. In recent years, complex networks such as social networks have received great attention due to their popularity, also the need to understand their structure and their usefulness in several domains such as healthcare, education, marketing and business.

Manuscript received May 23, 2015

Seema Rani, Department of Computer Science, Jamia Millia Islamia University, New Delhi, India,+011 9811476855, (e-mail: seema7519@yahoo.com).

Monica Mehrotra, Department of Computer Science, Jamia Millia Islamia University, New Delhi, India, 9818846513, (e-mail: drmehrotra2000@gmail.com).

The community structure captures the tendency of nodes in the network to group together with other similar nodes into communities. This property has been observed in many real-world networks. Despite excessive studies of the community structure of networks, there is no consensus on a single quantitative definition for the concept of community and different studies have used different definitions. A community, also known as a cluster, is usually thought of as a group of nodes that have many connections to each other and few connections to the rest of the network. Identifying communities in a network can provide valuable information about the structural properties of the network, the interactions among nodes in the communities, and the role of the nodes in each community. Community is groups of vertices are more densely connected than to other vertices in the network. Community detected from the social network provides basic information for other tasks. Community detection methods broadly categorizes into four: Node-Centric Community, Group-Centric Community, Network-Centric Community and Hierarchy-Centric Community [20]. In node centric community each node in a group satisfies certain properties. Group centric community considers the connections within a group as a whole. The group has to satisfy certain properties without going into node-level detail. Network centric community partition the whole network into disjoint sets. Hierarchy-Centric Community constructs a hierarchical structure of communities.

II. PRELIMINARIES & RELATED TERMINOLOGIES

A. *Information graph*

Let graph $G = \{V, E\}$,where V is the network of individuals (or nodes). $V = \{v_1, v_2, \ldots v_n\}$ contains n nodes. E is a collection of links (or edges) in the network, $E = \{e_1, e_2, \ldots e_m\}$. Each node has p attributes. The collection of node V is $V_{att} = \{a_1, a_2, \ldots a_p\}$. Node attributes and links matrix of the graph can be constructed by the above information The following matrix can represent the information links:

$$\begin{pmatrix} A_{11} \ldots\ldots\ldots\ldots\ldots A_{1n} \\ \ldots\ldots A_{ij} \ldots\ldots\ldots\ldots \\ \ldots A_{ji} \ldots\ldots\ldots\ldots \\ A_{m1} \ldots\ldots\ldots\ldots\ldots A_{mn} \end{pmatrix}$$

wherein A_{ij} is connection information of node i to node j.

B. Information Gain

Information gain [1] is defined as the difference between the original demand and the new demand. The original demand refers to the expectations of the original sample classification. The new demand in the division based on classification information needs the known properties of A sample. Assume that D (the sample) has m categories, the ith i = (1,2,...,m) category accounted for the proportion of the total number of samples is p_i . Then:

$Info(D)=-\sum_{i=1}^{m} p_i log_2(p_i)$

Assumptions divided in accordance with the characteristic A to D, the D can be divided into v subsets $\{ D_1, D_2,\ldots D_v \}$ .Where D_j in the A has value a_j . According to the desired information required by this division of the A to D as follows:

$Info_A(D)=\sum_{j=1}^{v} (|D_j| / |D| . Info(D_j))$

Then the information gain of A is:

$Gain(A) = Info(D) - Info_A(D)$

Where $Info(D)$ is the original information needs, and $Info_A(D)$ is the original information needs, and $Info_A(D)$ is the new information needs.

C. Modularity

Graph $G=\{V,E\}$ contains n nodes and m edges. In the graph of no weights and no direction, Modularity Q is defined as:

$Q = \frac{1}{2m}\sum_{i,j}^{n} (A_{ij}-\frac{1}{2m}k_i k_j)\ \delta(C_i,C_j)$

Where $k_i=\sum_{j=1}^{n} A_{ij}$ is the degree of V_i . Node i belongs to Community C_i . When node i and j belong to the same community, $\delta(C_i, C_j)$= 1, otherwise the value is 0.

D. Centrality

Betweenness Centrality: Freeman [17] defined a family of measures of centrality to find the degree to which a point falls on the shortest path between others. There are two kinds of betweenness centrality of social network: vertex betweenness BC(v) centrality and edge betweenness centrality BC(e) [18].

$BC(v) = \sum_{u,w\varepsilon V, u\neq w\neq v} \sigma_{uw}(v)/ \sigma_{uw}$

$BC(e) = \sum_{u,w\varepsilon V, u\neq w} \sigma_{uw}(e)/ \sigma_{uw}$

Where $\sigma_{uw}(v)$ is the total number of shortest path between pairs of vertices u,w ε V that pass through vertex v; σ_{uw} is the total number of shortest path between u and w; $\sigma_{uw}(e)$ is the total number of shortest path between pairs of vertices u, w ε V that pass through edge e and σ_{uw} is the total number of shortest path between u and w.

Closeness Centrality: Closeness centrality measures how close a vertex is to all other vertices in the graph by measuring how many steps are required to access every other vertex from a given vertex [19]. The group closeness centrality is defined as the normalized inverse sum of distances from the group to all nodes outside the group. The closeness centrality of a vertex $v\in V$ is defined as the inverse of the sum of distances from v to all other vertices.

$CC(v)= 1/ \sum_{t\in V\setminus v} d_G(v,t)$

Where $d_G(v,t)$ is the length of a shortest directed path from v to t.

Degree Centrality: Number of links incident upon a node or number of people attached to each person denoted as $C_D(v)=deg(v)$.

E. Graph Density

The density of a graph [3] defined as the ratio of the number of edges to maximal number of edges. The opposite a graph with a few edges will be a sparse graph. The difference between sparse and dense graphs depends on the context. For undirected simple graphs, the graph density is defined as: D=2E / n (n - 1) Where E denotes to the number of existence edges in the graph and denotes to the number of vertices (nodes).

F. Clustering Coefficient

Given a graph G = (V, E) and a vertex vεV, the clustering coefficient [3] of v denoted by C(v) is defined as the number of directed links that exist between the nodes neighbors, divided by the number of possible directed links that could exist between the nodes neighbors.

C(v)=num_of_pairs_of_neigbors_connected_by_edges/num_of_pairs_of_neigbors

The clustering coefficient of a graph C(G) is the average clustering coefficient of all its vertices.

$C(G) =\sum_{v\varepsilon V} C(v)/ |V|$.

G. Normalized Mutual Information

Normalized mutual information (NMI) is used to measure the differences of detected communities and real communities. Assumed C_o is the real community structure. C_e is derived from the algorithm. NMI is defined as:

$NMI(C_o , C_e) = H(C_o) + H(C_e)- H(C_o , C_e) / \sqrt{H(Co)H(Ce)}$

Where $H(C)$ is the Shannon information entropy of C .When C_e and C_o is exactly the same $NMI(C_o , C_e)$ =1 when C_e and C_o is completely different, $NMI(C_o , C_e)$ = 0.Greater the value of NMI, the result is closer to the real community.

III. Related Work

A wide variety of community detection algorithms, also known as clustering algorithms, have been proposed to identify the communities in a network. Since different community detection algorithms use different definitions of a community, they yield different communities. Many traditional community detection methods are borrowed or inspired from graph clustering algorithms. Partitioning the nodes in a network into a predetermined number of disjoint communities is one of the traditional methods for identifying communities. However, since the community structure of real-world networks are not usually known, making assumptions about the number of communities or the size of the communities are not realistic. Moreover, many real-world networks have a hierarchical structure where meaningful communities at different scales can exist and such community structures cannot be captured by partitioning algorithms. Therefore, another group of community detection algorithms have been introduced which can identify hierarchical communities. Hierarchical clustering techniques can be divided into agglomerative and divisive methods. Agglomerative algorithms use a bottom-up approach where clusters are iteratively merged. Divisive algorithms use a top-down approach where the

clusters are iteratively split. Overall, using hierarchical algorithms allow us to choose the suitable level of hierarchy and study the communities at that level of hierarchy. In many real-world networks, nodes can naturally belong to multiple communities, therefore the communities can overlap. In social networks, an individual can belong to a community of family members, to a community of friends, and to a community of colleagues. In an information network, a web page can cover topics that are associated with different communities. Traditional community detection algorithms fail to uncover the community overlaps. Not being able to identify community overlaps in networks with naturally overlapping communities' means missing valuable information about the structure of the network. Therefore, overlapping community detection algorithms have gained a lot of attention. Overlapping communities can be identified using different approaches. One of these approaches is based on partitioning the edges of a network into communities rather than partitioning the nodes.

Bing Kong et al [2] proposed a new dynamic algorithm based on the modularity given by Newman and Girvan (NG modularity for short). Detection of network community is an important basis for disclosing the relationship between network structure and functions [4]. The dynamic property of the algorithm means number of communities should be changed gradually to make the community detected by the algorithm in the end is the maximum one among all largest modularities corresponding to different Community numbers. This process is not a hierarchical clustering which avoids the problem of retention of mistakenly merged clusters to the next step.

Yuan Huang et al [1] presented an algorithm that realizes community detection of the social network by combining the link and node attribute. The Anh Dang et al [8] defined attribute modularity, combined with Newman proposed modularity [7] in the community detection. In addition, their presented a KNN graph detection algorithm, but the algorithm is limited by the value of *K* size.

Ramasuri Narayanam et al [6] introduced a notion called signature of a social network and propose an efficient approach to compute it. The signature of a social network is essentially a sparse subgraph of the original social network such that it succinctly captures key information contained within the data sources (both linked and interaction data). The signature of a social network need not be unique. The value behind computing such a signature stems from the fact that once computed, any subsequent SNA (e.g. community detection, influence propagation, etc.) becomes much faster while not compromising much with quality. The concept of importance weights of the edges has been the guiding principle for us behind the idea of signature of a social network. Used four different measures - modularity, precision, recall, and F Measure to offer a comparison between communities detected from original graph versus signature graph.

Many algorithms are proposed to solve the problem of community detection in social network some of them used clustering techniques as in [12], [13]. The density-based spatial clustering of applications with noise (DB SCAN) algorithm [3] which was previously proposed by Ester et al. in [9] is used in social network analysis where the network members are classified to seeds or to core members of the groups. By eliminating the outliers the dataset will be noise free to deal with it.

The identification of overlapping community is a crucial task. Existing methods present a high complexity as the size of the network increases. Zeineb Dhouioui et al [10] present a new method allowing overlapping community detection based on the principle of edge betweenness. Such an algorithm confront mainly the following challenges, firstly the scalability that's means to deal with large networks such as real-world networks and secondly tolerating the case of overlapping.

Social networks are benefic in several domains even in healthcare[10]. Actually, individuals in such healthcare community interact; individuals can be doctors, patients or nurses [14]. Healthcare requires discussions, cooperation and interactions between members to share information or advices and experiences. Moreover the use of social networks ensures feedbacks for example sharing a diet experience can be motivating for many members in such a community. Data analytic techniques such as data mining and predictive modelling are being used to gain new insights into health care costs, performance and quality of care. In this context, social network analysis (SNA) has the unique ability to play a new role in exploring the context and situations that lead to efficient and effective healthcare [5]. Describe our SNA based approach (applied to health insurance claims) to understand the nature of collaboration among doctors treating hospital inpatients and explore the impact of collaboration on cost and quality of care.

There are many community detection algorithms for discovering communities in networks, but very few deal with networks that change structure [11]. The SCAN (Structural Clustering Algorithm for Networks) algorithm is one of these algorithms that detect communities in static networks. There are many community detection algorithms in use today, ranging from label propagation [15] to density analysis [16]. Many of these algorithms are designed to discover communities in static networks and do not scale well. To make SCAN more effective for the dynamic social networks that are continually changing their structure, Nathan Aston et al [11] proposed the algorithm DSCAN (Dynamic SCAN) which improves SCAN to allow it to update a local structure in less time than it would to run SCAN on the entire network.

IV. DISCUSSION AND CONCLUSION

Use Searching for and detecting communities automatically in large-scale complicated network is helpful for finding new knowledge and phenomenon, and is

significant for understanding social network structure and analyzing social network features [2]. Complex networks such as social networks have received great attention due to their popularity, also the need to understand their structure and their usefulness in several domains such as healthcare [10].Networks today include millions of nodes and billions of edges and are continually changing their structure [11].The presence of both variety and volume in these datasets pose new challenges, and thereby opportunities for the field of social network analysis (SNA) [6]. Much research work is going on in the direction of improvement of existing community detection algorithms. As Community identification in Social Networks as wide applicability in various domains like business, marketing, healthcare etc. requires greater attention in the research domain.

REFERENCES

[1] Yuan Huang, Wei Hou, Xiaowei Li & Shaomei Li. An Effective Community Detection Algorithm of the Social Networks. *International Conference on Information Science and Technology*, 978-1-4673-2764-0/13 ©2013 IEEE, 824-827.

[2] Bing Kong, Hongmei Chen, Weiyi Liu & Lihua Zhou. A Dynamic Algorithm for Community Detection in Social Networks. *World Congress on Intelligent Control and Automation*, 978-1-4673-1398-8/12 ©2012 IEEE, 350-354.

[3] Yomna M. ElBarawy , Ramadan F. Mohamed & Neveen I. Ghali. Improving Social Network Community Detection Using DBSCAN Algorithm. 978-1-4799-2806-4/14 ©2014 IEEE.

[4] L. Danon, A.D. Guilera, J. Duch, A. Arenas. Comparing community structure identification[J]. *Journal of Statistical Mechanics*, 2005, p09008.

[5] Fei Wang, Uma Srinivasan, Shahadat Uddin & Sanjay Chawla. Application of Network Analysis on Healthcare. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 978-1-4799-5877-1/14 ©2014 IEEE. 596-603.

[6] Ramasuri Narayanam, Dinesh Garg & Hemank Lamba,IBM Research India. Discovering Signature of Social Networks with Application to Community Detection. 978-1-4799-3635-9/14 ©2014 IEEE.

[7] A. Clauset, M.E.J. Newman & C. Moore. Finding Community Structure in very large networks [J]. *Physical Review E, vol 69, 06613*, 2004.

[8] T.A. Dang, E. Viennet. Community Detection based on structural and Attribute Similarities. *The Sixth International Conference on Digital Society*. 2012.

[9] M. Ester, H. Kriegel, J. Sander, and X. Xu. A Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (KDD-96), pp.226-231, 1996.

[10] Zeineb Dhouioui & Jalel Akaichi. Overlapping Community Detection in Social Networks. *IEEE International conference on Bioinformatics and Biomedicine*,978-1-4799-1310-7/13©2013, 17-23.

[11] Aston, N. & Hu, W. Community Detection in Dynamic Social Networks. *Scientific Research Communications and Network, 6*, 124-136. http://dx.doi.org/10.4236/cn.2014.62015. 2014.

[12] M. Newman. Detecting Community Structure in Networks. *The European Physical Journal B- Condensed Matter and Complex systems*, Vol. 38(2), pp. 321330, 2004.

[13] Y. Wang. An Improved complex Network Community Detection Algorithm Based on K-Means. *Advances in Intelligent and Soft Computing*, Vol.160, pp. 243-248, 2012.

[14] J. G. Anderson. Evaluation in Health Informatics: Social Network Analysis. pp. 189-204, 2005.

[15] Fortunato, S. Community Detection in Graphs. *Physics Reports*, 486, 75-174. http://dx.doi.org/10.1016/j.physrep.2009.11.002. 2010

[16] Xu, X., Yuruk, N., Feng, Z. and Schweiger, T. SCAN: A Structural Clustering Algorithm for Networks. KDD'07. ACM, 824-833. http://dx.doi.org/10.1145/1281192.1281280. 2007.

[17] Freeman, L. C., Centrality in Social Networks I: Conceptual Clarification. *Social Networks*, 1, 215-239, 1979.

[18] S. Narayanan. The betweenness Centrality of Biological Networks. M.Sc thesis, Virginia Polytechnic Institute and State University, Virginia, 2005.

[19] D. Gomez, J. Figueira, and A. Eusebio. Modeling Centrality Measures in Social Network Analysis Using Bi-Criteria Network Flow Optimization Problems. *European Journal of Operation Research*, Vol 226(2), pp. 354-365, 2013.

[20] Lei Tang and Hoan Liu, *Community Detection and Mining in Social Media*, Morgan & Claypool Publishers.

Seema Rani is pursuing Ph.D Computer Science from Jamia Millia Islamia University. She has completed her M.Tech in 2009 from GGSIP University. Her research interest includes Data Mining, Algorithms, and Social Network Analysis.

Monica Mehrotra presently working as Assitant Professor in Jamia Millia Islamia University. She has completed her Ph.D in 2007 from Jamia Millia Islamia University. She has over eighteen years of teaching experience. Her research interest area includes Data Mining, Information Retrieval, Software Engineering and Neural Networks.

.