Survey on Community Detection in Online Social Networks

Amit Dhumal
PG Student
Dept. of Computer Engineering
Sinhgad College of Engineering
Pune, India

Pravin Kamde
Associate Professor
Dept. of Computer Engineering
Sinhgad College of Engineering
Pune, India

ABSTRACT

The proposed survey discusses the topic of community detection in the context of online social network. Community detection helps to identify link density within network structure and to predict future missing links. In online social networks, nodes typically represent individuals and edges indicate relationships between them. Due to the complexity, dynamic nature and huge scale of network, community detection in online social networks is challenging task. In this survey various community detection methods for networks with static and dynamic nature are discussed, and results of applying them on online social network is are provided.

Keywords

Community detection, online social networks, data clustering.

1. INTRODUCTION

Social media networks such as microblogs and social networks e.g. Twitter and Facebook, are provide interactive and cheaper way for user to share ideas, exchange information and stay connected with people. Ease in using social media applications on mobile devices achieves rapid growth in social media network users and leads to generate vast amount of user generated content. Statistical study of popularity of social media networks results in following facts:

1. People spend more time on social media network sites than any other internet activity where many people remain engage with the Facebook and Twitter than any other social media network.
2. BI intelligence reported in August 2014 that there are near about 100 million daily active users of Twitter who produce 1, 40,000 tweets per second.

This large user base and their discussions produces huge amount of user generated data. Such social media data comprises rich source of information which is able to provide tremendous opportunities for companies to effectively reach out to a large number of audience.

Tracking the users topic of interest from available such huge source of information is challenging and very important task for targeting right audience in various domains e.g. Marketing, Politics etc. Also it is observed that [1] users which are connected to each other via relationship links in such social media network do not necessarily share specific interest. Thus for spreading right information to right audience it becomes important to identify group of users with common interest in such large network. Detecting community of such like minded people from large social media networks can provide benefits to applications of various domains and for suggesting like minded people to user which are still unknown to him/her.

An important practical problem in social networks is to discover communities of users based on their shared content and relationship with other users. Community in network is a pattern with dense links internally and sparse links externally. These links can be characterized by the content similarity between users, friendship between them and also other similarities in their personal data such as their location, gender, age etc. These close structures can then be used for various purposes such as targeted marketing schemes, terrorist cells.

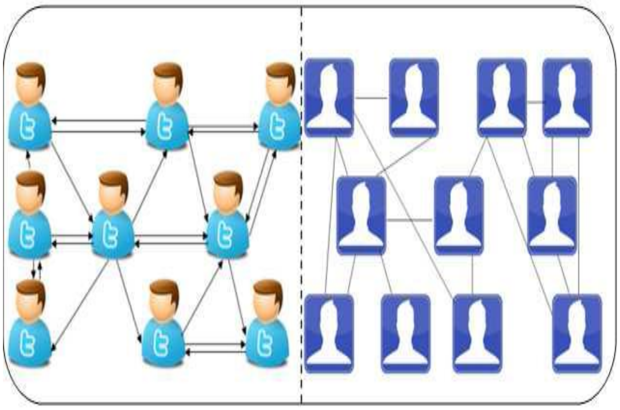

Figure 1. First section represents Follower-Followee relationship between Twitter users while second section represents Friend relationship between Facebook users

The social links of friendship is an important part of most social networks. These social links often give rise to communities in social media network. Communities in a social network might represent real social groupings, perhaps by interest or link topology to identify these communities. It will help to understand and exploit these networks more effectively. The ability to detect community structure in a network has practical applications in various domains.

Most of the existing approaches for community detection are based on link analysis and ignore the vast amount of other information available in social networks. Besides that most of the community detection algorithms divide whole network into disjoint set of nodes which are known as cohesively connected set of nodes or community. This is definitely not true in social networks. In social networks like Facebook and Twitter one user can be a part of more than one cluster. Also

Twitter has different types of links in the form of follower-following relationships, retweets, mentions and replies. Tweet contains rich source of textual information along with tweet tags e.g. hashtag and mention. Additionally, Twitter provides a lot of metadata in the form of user location, age and gender which can be used for clustering. Due to the above facts identification community of like minded people in social media network is still challenging task.

This paper is organised as: Section 2 provides fundamental information related to community detection, community criteria and reviews various categories community detection methods. Table 1 list down various parameters considered while detecting community. Section 3 reviews application of various types of community detection algorithms in social networks. Table 2 classifies applied community detection algorithm along with respective category & network used for community detection. Lastly section 4 concludes the paper.

2. FUNDAMENTALS

2.1 Community detection

Processing network as graph facilitates the understanding and analysis of network structure, such as the identification of local and global characteristics, influential nodes and the dynamics of networks. Sub graph discovery is one of the graph theoretic technique which helps to discover of hidden patterns of networks based on topology of network. Discovering sub graph of vertices which share common properties is known as “Community Discovery”. Massive set of applications in various domain is depends on community structure for various purposes.

There is no universal or exact definition of what constitutes community and the exact definition often depends on context. Community in graph is defined as a cohesive group of nodes that are connected more densely to each other than to the nodes in other communities where cohesiveness is depends upon various measures described in later sections.

Community discovery process is related with fundamental concept of data clustering as it clusters set of vertices while latter is used for clustering data points. There are two required properties of community which must need to be satisfied by sub graph:

1. Edge Density
   There must be maximum no. of edges shared within sub graph while less no. of edges outside of it.

2. Connectedness
   Path between pair of vertices of sub graph must run only through vertices of sub graph.

Any sub graph which satisfies above conditions is considered as community.

There are various types of community defined on the basis of the different parameters. Following table provides detail about community types along with definitions. Except community defined on similarity based metrics, all other types of communities satisfies the required properties of community.

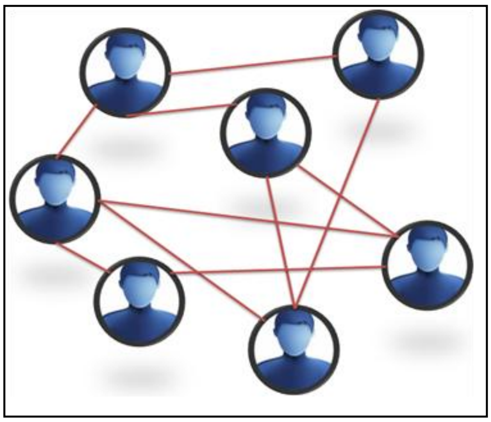

Fig 2: Sample Network

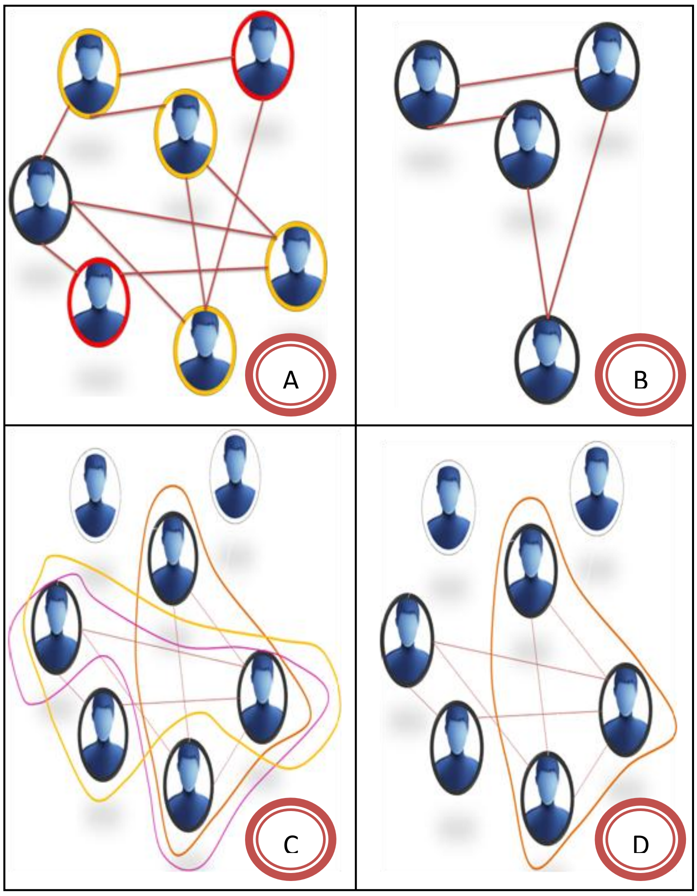

Fig 3: Different types of communities detected in Sample network using different approach: a) Vertex similarity b) Global community c) Local community (Overlapping) d) Local community (Non overlapping)

There is no exact categorization of communities are possible according to above parameters still it’s significance affects performance of community detection process and network considered for community detection. Community criteria are defined to decide type of community which has been detected given as follows:

2.1.1 Community criteria

Criterion parameters [2] [4] are defined for induced sub graph which decides whether sub graph is "Community" or not. Local and Global community relies on different set of criteria which are discussed in the following section. Provided set of parameters also helps to decide stopping condition for "Community Discovery" process.

2.1.1.1 Local community

Local community can be derived from incomplete part of whole graph of real world network as stated in previous section. Neighborhood vertices of initial source vertex plays important role in community formation. Different degree distribution of vertices of graph leads to identify set of valid vertices from neighborhood region which satisfies required properties of community. For the same purpose following properties are devised:

I. *Complete Mutuality*

   To satisfy this property, all vertices of sub graph must need to connect with each other (Mesh connection) e.g. edge is required between each vertex pair. It is one of the strict constraint which helps to detect community of "Clique" type.

II. *Reachability*

   Reachability is stated as maximum distance between two vertices of sub graph (e.g. diameter) should not exceed predefined threshold value n. It helps to identify community of vertices which are connected and available within periphery n.

   Example: n-clique, n-clan, n-club.

III. *Sub graph cohesion*

   Sub graph cohesion is achieved when each vertex of sub graph is adjacent to at most/at least k no. of vertices of sub graph. It helps to identify community predefined minimum/maximum k threshold.

   Example: k-plex, k-core.

IV. *Internal vs. External cohesion*

   Strong community is one which not only has dense connection of vertices within sub graph but also have less connection with rest of sub graph. Considering this fact detection of local community is evaluated using two factors: (internal/external) degree of vertex and sub graph. This property helps to identify strong and weak community.

   Example: LS-set.

2.1.1.2 Global community

Global community considers community structure as a property of the whole graph. It is stated in previous section that it is necessary to have complete knowledge of whole graph due to following reasons:

1. In many real world networks, sub graph of original graph comprises basic structural properties of whole graph. Only such sub graph can be identified as "Global Community" which is basically prototype model of original whole graph.
2. If complete graph is unavailable, it can lead to generate false positives for com- munity which does not satisfy basic properties.

An only single criterion is defined for the identification of global community described in the next section.

I. *Null model*

It is a random graph model which is similar in some structural features of original graph. It is used for the comparison with induced sub graph. If induced sub graph is similar to the Null model then it is considered as community.

2.2 Categories of community detection methods

Community detection in graph has been studied for a long time. Two main factor e.g. type of graph and interpretation of edge, are helps to categorize the community detection algorithm.

Methods of discovering communities in graph are mainly in divided into five major categories. Category of community detection algorithms are decided on the basis of following factors:

a) Vertex similarity

Methods are focuses on clustering vertices which have more similar feature.

b) Edge density

Methods are focuses on identifying clusters which improves intra cluster edge density rather than inter cluster edge density.

c) Distance between vertices

Methods are focuses on clustering vertices which are nearer to each other. Basic principle of clustering data points using data clustering algorithms is considered as base for same purpose.

Categories of community detection methods:

2.2.1 Graph partitioning

It is one of the earliest methods for community detection defined by Kernighan et al.[2].Graph partitioning based methods divides the vertices into predefined no. of groups where size of each group is also defined. Groups are formed in such way that minimum no. of edges should lie between different groups. Various measures are used to define good partition e.g. Minimum bisection, Spectral bisection, Max flow min cut, Normalized cut etc. Output of graph partitioning based method is "Partition" type of community. Explicit provision of total no. of clusters and size is required before community discovery process. Reasonable assumption and provision of hidden community parameters in such way does not lead to generate optimal results.

2.2.2 Hierarchical clustering

Methods of this category are relying on the fact about real world community that each community is collection of small communities at different levels. Hierarchical clustering methods are used to discover communities with multilevel vertex grouping structure. Method is based on vertex similarity parameter. There are two types of hierarchical clustering.

Table 1. Classification of communities based on various parameters

Parameters	**Community Types**	
Membership	Partition	Cover
	If communities in graph are derive in such way that vertex can be part of at most single cohesive group then such communities are called as partitions of graph.	If communities in graph are derive in such way that vertex can be part of more than single cohesive group then such communities are called as covers of graph. In general terms, it is also known as "Overlapping communities".
Autonomy	Local	Global
	Community discovery process starts with exploring & agglomerating neighbourhood vertices of query vertex and continues till community does not achieve desirable quality.	Community discovery process tries to identify such sub graphs which are similar to the structural pattern of original graph.
	Example:Clique, n-clique, n-clan, n-club, k-plex, k-core, LS-set Example:Null model	
Similarity Based	Vertex similarity	
	Community detection tries to discover such clusters which contains vertex with similar feature e.g. vertex pairs with similar distant, vertices with no. of independent paths, vertices with same neighbourhoods. Edge existence between vertices of same cluster is does not considered as necessary condition.	

Agglomerative clustering

It is bottom up approach of community detection. Initially each vertex of graph is considered as unique cluster and iterative grouping of such high similarity clusters leads to form community.

Divisive clustering

It is top down approach of community detection. Initially whole graph is considered as single cluster and iterative splitting of dissimilar clusters by removing edges leads to form community.

Fortunato et al.[4] concludes that hierarchical clustering based approach provides more reliable result than graph partitioning based method as it is capable to decide no. of cluster and cluster size based on qualitative measure. Also "Divisive clustering" is considered as base principle for various modern community detection algorithms. Even hierarchical based methods provides such benefit, use of matrix based computation creates issue of scalability.

2.2.3 Partitional clustering

Partitional clustering method [4] is based on principle of data clustering approach which divides the whole graph into predefined no. of clusters. Two parameters e.g. diameter and distance to centroid, along with their maximum or average values are used to decide cluster for vertex. Based on stated measure four subtypes are defined namely Minimum k-clustering, k-clustering sum, k-center and k-median.Lack of deciding appropriate number of cluster and computation in metric space limits the use of partitional clustering scheme for community detection.

2.2.4 Spectral clustering

According to the Porter et al.[5] spectral clustering approach makes use of spectral properties of the graph. Eigenvector value for each node in graph is calculated to plot point associated with the node in geo-dimensional space. Further partitional based clustering methods are used to cluster points. Different types of methods are available based on choice of matrix for eigenvalue computation e.g. adjacency matrix, the standard Laplacian matrix, the normalized Laplacian matrix, the modularity matrix and the correlation matrix. Due to the use of partitional based clustering approach for vertex grouping, it does not solve the issue of efficient community detection.

2.2.5 Divisive algorithm

Newman et. al [5] proposes "Divisive algorithm". Proposed algorithm tries to remove inter cluster edges which helps to identify clusters which shares less no. of vertices. Removal of edge is depend upon score assigned by quality function e.g. between-ness. Main advantage of "Divisive algorithm" that it is helps to identify cluster without provided total no. of clusters and also able to find out overlapping communities.

Various algorithms are proposed for each category. Still algorithms of hierarchical clustering and divisive algorithm category are proved more efficient due to quality function based cluster detection. Also various algorithms of these two categories are proposed for local and global community detection.

3. CLASSIFICATION OF COMMUNITY DETECTION TECHNIQUES

In this section survey of application of topology based community detection techniques is presented. Further section provides classification of those techniques along with networks on which techniques are applied.

3.1 Topology based community detection

Very first approach to study of various methods to identify interest of user and Twitter network properties is carried out. Community detection process is carried out on social graph of Twitter using local community based Clique Percolation Method. Lim et al. [7] concludes that Twitter friendship links are either reciprocal or one way and friendship probability between two users is inversely proportional to the geographic location. After identification of community, keywords from tweet associated with users of community are extracted to determine topic of interest of community.

Rather than deciding topic of interest from tweet contents, cardinality of user friendship links associated with particular user profile is considered for user interest detection. Proposed model by [11] make use of Facebook post and Twitter comments of users made for chosen Facebook and Twitter profiles which are related to the particular category of interest e.g. sports, news, politics, business etc. Provided set of chosen Facebook & Twitter user profiles are transformed into matrix where matrix cell value denotes count of users which are interested in those profiles. Further undirected graph is generated for chosen Facebook profiles and Twitter profiles separately where edge weight is identified using Jaccard weighting index. Further improved version of global community based CNM algorithm is used for discover partitions from chosen set of Facebook and Twitter profiles. Discovered partitions denote set related profiles. At the end users interested in those profiles of partitions are considered community of users with similar interest.

Lim et al [8] proposes seed centric community detection algorithm for identifying users which having large no. of followers. As it is observed that seed users are representative of category of user interest most of the time, those seed users are considered as topic of user interest. Set of candidate users are evaluated for no. of seed users which helps to identify users which are sharing no. of interest within them. Social graph is further filtered which comprises only identified set of candidate users and existed link between them for generating friendship network. Finally local community detection based Clique Percolation Method is used to identify overlapping communities from Social graph.

Correa et al [10] proposes user network model which is form on the basis of interactions carried out between users of Twitter network. Author considers RT, @ mention links more effective than existing followership link due to consideration that interactive links are representative of active and real social relationship. Directed edges form between users with associated weights where weight is calculated on the basis of frequency of interaction. While selection interaction links is depend upon topic word provided by user. Further generated graph is filtered by considering only strong links of graph whose weight is above predefined threshold value. Finally local community detection based modularity optimization algorithm is used to discover community from graph.

Targeted advertisement on OSN (Online Social Network) initiates the need of mechanism to identify right audience for particular product. Considering this fact Lim et. al [9] provides solution for identifying group of people in OSN (Online Social Network) which are sharing common interest. Proposed methodology incorporates the use of tweet tags and follower relationship with celebrity to identify users with common interests. CICD(Common Interest Community Detection) method identify user with common interest by maintaining count of follower link from user to celebrities of pre decided category, while HICD (Highly Interactive Community Detection) comprise the use of tweet tag frequency to identify interactive users which are further considered are users with common interest. After identifying group of users with common interest graph based community detection method is used to classify users into respective community. For the same purpose Clique Percolation method is used which identifies such group of k-people which are densely connected to the each other using friendship link.

Jiyang Chen et. al [12] proposed agglomerative clustering algorithm along with Max-min modularity quality measure. Proposed algorithm considers both topology of network and provided domain knowledge e.g. unrelated pair of vertices while community detection. Defined community measure considers absent links within vertices of same community which helps to identify sub graphs with number of edges within sub graphs and unrelated pairs between sub graphs having higher value than expected one.

Vertex similarity approach is proposed in Dang et al. [13] to partition graph into network considering structural and attribute equivalence. Modularity optimization scheme is used to define cluster where modularity measure consider both link strength and attribute equivalence factor. K-nearest neighbor based approach is used to agglomerate vertex into community.

Karsten Steinhaeuser et. al[14] proposes methodology which considers structural and attribute equivalence while calculating vertex similarity. Structural equivalence evaluated on the basis of degree of node, participation in triangle community and total no. of common neighbourhoods while attribute equivalence is based on value of attributes. Evaluated score are assigned as edge weight. Comprising use of attribute information of node for deciding vertex similarity allows efficient detection of community with even available little information about topology. Predefined threshold value is used to define eligibility of edge in community.

Xiao-Li Li et. al[15] proposes technique to detect community of people for future events based on available knowledge of previously attended events and topology of social network. Weighted graph is generated with inclusion of virtual links to depict relationship between people with similar interest but unknown to the each other. Vertex similarity approach is used to cluster event and to identify set of candidate people for new event.

All of the above approaches make use of different network links available in Twitter network e.g. followee-followership links, RT & mention links etc. While user specific interest identification and further devising network model based on similarity of interest for community detection is first proposed by Natarajan [16]. Author proposes LDA variant generative model for link prediction in social network which can remain helpful for identifying future friendship in social network. Unlike email network, link between users social network are not post level and most of the user's shares content of known users are two main facts which leads to consider both link topology and tweet content to model friendship network. LDA technique is used to define Link content model which predicts link between users with probability distribution which helps to identify large number of communities for different topic and different set of users. Above discussed techniques are summarized into table 2 which are classified according to the use of network structure and community detection approach.

Table 2. Classification of community detection approaches

Input		Community detection approach	Method	Output	Networks
Relationship network + User generated content	[7]	Local community detection	CPM (Clique Percolation Method)	Graph	Twitter
Relationship network	[8]				
	[9]				
Interaction network	[10]	Modularity optimization	LM based	Weighted directed graph	
User generated content	[11]		CNM	Dendogram	Facebook , Twitter
Relationship network + Domain knowledge	[12]	Agglomerative clustering	Max-min modularity based	Graph	Zacharay Karate club, Sawmill network
	[13]				DBLP
Relationship network + Node attribute values	[14]	Partitional clustering	K-nearest neighbor		DBLP, Facebook
	[15]	Threshold scheme	-		Mobile social network

4. CONCLUSION

We have presented an overview of the community detection process in online social network. The existing approaches are illustrated with the main focus input parameters which are used while performing community detection. Type of network used as input for community detection affects the use of resulting community outputs e.g. target advertising, detecting information flow in social network etc. Although satisfactory efforts have been taken in designing efficient community detection systems, there is still need of scalable community detection approaches which reflect real world community as result. Further research is still required in this field which helps to detect community by embedding use of external knowledge sources to extract more meaningful communities.

5. REFERENCES

[1] Kwan Hui Lim and Amitava Datta, Finding Twitter Communities with Common Interests using Following Links of Celebrities, MSM'12, Pages 25-32, ACM 2012

[2] Symeon Papadopoulos et al., Community detection in Social Media, in KDD 2011

[3] Kernighan, B. W., and S. Lin, An Efficient Heuristic Procedure for Partitioning Graphs, 1970, Bell System Tech. J. 49,291

[4] Santo Fortunato, Community detection in graphs, Physics Reports 486, 75-174, Arxiv 2009

[5] Mason A. Porter et al., Communities in Networks, Vol. 56, No. 9: 1082-1097, 1164-1166, Arxiv 2009

[6] M. E. J. Newman et al., Finding and evaluating community structure in networks, Phys. Rev. E, vol. 69, no. 2, p. 026113

[7] Kwan Hui Lim and Amitava Datta,Tweets beget propinquity:Detecting highly interactive communities on twitter using tweeting links,WI'12,Pages 214221,IEEE 2012

[8] Kwan Hui Lim and Amitava Datta, A Topological Approach for Detecting Twitter Communities with Common Interests, Springer 2012

[9] Java et al.,Why we Twitter: Understanding microblogging usage and communities,WebKDD/SNA-KDD 07,Pages 56-65,ACM 2007

[10] Denzil Correa et al.,iTop: interaction based topic centric community discovery on twitter, PIKM '12, Pages 51-58,ACM 2012

[11] Palsetiay et al.,User-interest based community extraction in social networks,SNAKDD'12, ACM 2012

[12] Jiyang Chen et al., Detecting Communities in Social Networks using Max-Min Modularity SDM, page 978-989. SIAM 2009

[13] Xiao-Li Li et. al, ECODE: Event-Based Community Detection from Social Networks DASFAA'11, Springer p22-37

[14] The Anh Dang and Emmanuel Viennet, Community Detection based on Structural and Attribute Similarities ICDS'12, IARIA

[15] Karsten Steinhaeuser and Nitesh V. Chawla, Community Detection in a Large Real-World Social Network, Social Computing, Behavioral Modeling, and Prediction Springer 2008, pp 168-175

[16] Nagarajan Natarajan et al.,Community detection in content-sharing social networks, ASONAM '13, Pages 82-89, ACM 2013