Issues and Challenges in Community Detection Algorithms

Dipika Singh,

(Department of Computer Science, BHU, Varanasi, Uttar Pradesh, India)

Rakhi Garg

(Department of Computer Science, MMV, BHU, Varanasi, Uttar Pradesh, India)

ABSTRACT

Community detection is an important approach to gain insight in the complex network structure. These networks can be biological, social, geographical or technological. Tremendous information is present in such networks. If these can be mined successfully, many important results can be inferred. Many algorithms are developed for community detection. But the very definition of community is debatable. So there are number of issues and challenges open for future research in this area. In this paper we mainly focused on four major issues related to community detection i.e. lack of precise definition, dynamic community detection, overlapping nature of community, validation of community detection algorithms and their respective challenges in brief. This paper will be beneficial for researchers to gain an insight on the hurdles that they can face in the research of community detection.

KEYWORDS

Community Detection, Dynamic Community, Overlapping Community, Social media Network.

1. INTRODUCTION

In today's era of big data, complex networks are important field of research [1]. A valuable area in the study of complex network is Community Detection. Community Detection basically views networks as graph and tries to find out nodes which are more strongly attached to each other than the others [3]. Finding such communities is beneficial for different domains of research including the fields from Biology to Social sciences [2, 4]. Studies reveal that valuable information can be obtained from community detection. This helps to detect the hidden phenomena which are not directly visible and can be used for a variety of applications such as recommendation system , automatic event detection, prediction of missing information, determining functioning of genes and proteins in a cell etc. [4, 5, 6]. In Social Media Mining, millions of data are shared every second. Community detection plays an important role to uncover the hidden patterns in these data [2].

Basically Communities' are of two types i.e. implicit or explicit [2]. Explicit communities are directly labelled by group members such as whatsapp group, facebook group [7] etc. On the other hand implicit communities do not have a proper grouping but are based on the common interest, properties or behaviour of group members. Community detection means detection of such implicit communities.

Many reviews of community detection are available [3]. But a review on major issues and challenges in this area is needed to be focussed on. We have classified the issues mainly into four types i.e. No Precise definition, Dynamic Community, Overlapping nature of Community, and Validation of Community formed as shown in Fig.1.

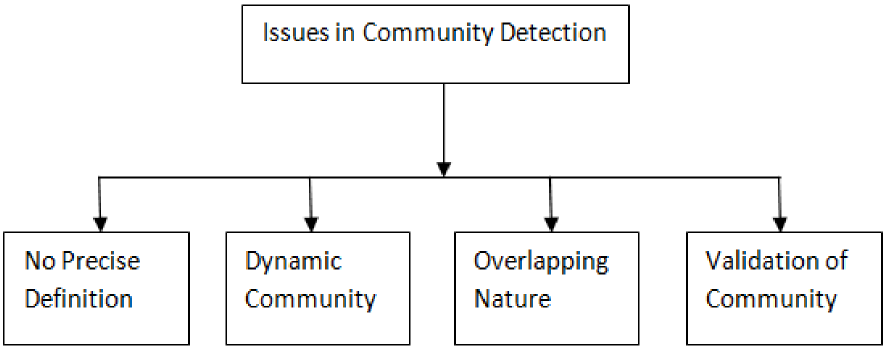

Fig.1. Issues in Community Detection Research

There is no globally accepted definition of community due to which different researchers have different perspective of community, depending on their research need [8]. Kernighan Lin considered community as parts of the graph with few ties with rest of the system[18], while some other researcher defined Community as group of vertices similar to each other [3]. Moreover majority of the algorithms of community considers community to be discrete but in reality communities are overlapping [15]. Apart from overlapping nature, communities of social media exhibit dynamically changing properties which cannot be ignored [14]. The validation issue of different algorithms of community detection is also a big hurdle in the research of community detection.

In this paper we have mainly discussed the major issues and challenges in community detection that not only helps researchers and scientists working in this area to understand the problem but also help in developing solution to it. The paper is divided into four sections. Section 1 focuses on introduction. Community Detection is discussed in Section 2. All the issues related to community detection and their respective challenges are discussed in Section 3 in brief. At last Section 4 concludes the paper.

2. COMMUNITY DETECTION

According to Kernighan –Lin algorithm ,"Communities are those parts of graph that have less ties with rest of the graph"[18]. Wasserman et al considered Community as maximal subgraph, that cannot be extended by addition of more vertices without loosing its property [3]. On the basis of fitness measure, communities are definite if fitness is larger.Vertex similarity is yet other parameter to define community, in this case communities are group of similar vertices[3].

So there is no standard definition for community [8]. But in very simple words, a community can be defined as a subgraph with more intra cluster edges than inter cluster edges. It means they are group of nodes which have more interactions among themselves than others. Fig.2 shows community in a network. This particular network consist of eleven nodes, which are divided into three communities A, B and C, consisting of five, four and three nodes respectively. The nodes are divided based on the density of edges connecting them.

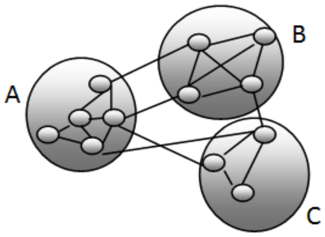

Fig.2. Community in a network

Communities can be implicit or explicit. Explicit communities are those, in which a grouping is predefined and members joining the group form a community. In this case communities are directly visible, for example whatsapp group.

Implicit communities on the other hand do not have any predefined classification. We have to analyze the activities of the individuals to form the community. Community detection is used for implicit communities only.

Community detection finds a variety of use in improving recommender system, fraud detection, bioinformatics and social science research [3, 4].

3. ISSUES AND CHALLENGES

Many algorithms are developed for community detection. Some of the major contributions in this area are listed in Table 1.

Table 1: Research in Community Detection

Author	**Research**	**Disadvantage**
Kernighan Lin [18]	Graph Partitioning algorithm for fast Community Detection	Number of communities have to be predefined
Girvan and Newman [19]	Algorithm based on edge betweenness. Iteratively remove edge with high betweenness value to get the community.	A new parameter “Modularity” has to be defined for analysing the community formed. Not suitable for overlapping community.
Tyler et al [3]	Algorithm using graph theory to discover community. Graph is split into connected components and each is observed whether it is a community If it is not a community then edges are removed ,between’s is recalculated on each removal of edge	Not suitable for overlapping communities. Not suitable for dynamic communities.
Newman [20]	Hierarchical Algorithm for community discovery from large graphs Any two communities whose join creates the largest change in modularity are merged	Depends on good modularity measure. Does not work for overlapping and dynamic Community.
Clauset et al [21]	More efficient implementation for the above algorithm using Max Heaps	The algorithm demands heavy computational resources.
Zhou et al [22]	Bayesian models is used to discover communities in email networks Also keeps into account the topics of discussion and social links	Nonlinear model require more computational resources. Not suitable for overlapping communities.
Palla et al. [23]	Clique Percolation Method (CPM) is used for locating communities It revealed four types of communities: (a) small ,stationary community (b) small ,non-stationary community (c) large ,stationary community (d) large ,non-stationary community.	More robust algorithm required. Not efficient for dynamic communities.
Albert et al. [24]	algorithm based on label propagation used for community detection	No unique solution but aggregate of many solutions is found.

The major issues and challenges in community detection that we identified are listed below:-

3.1 No Precise Definition an Issue

A general definition of community does not exist [8].This issue creates many challenges open for future research. In spite of community detection being one of the strongest fields of research in social media mining, a proper definition of this problem does not exist. Due to lack of proper definition there are different views of community by different researchers. These different views are not disjoint, they can sometime lead to common result. One view of community cannot be considered better than other.

Challenges Involved

i. Analyzing which view of community can be beneficial under what conditions [4].
ii. Selecting a particular viewpoint for community to start community detection [8].

The different views can be one of the following [8]:-

- Cut based view
- Clustering view
- Stochastic block model view
- Dynamic view

Cut based view: This perspective defines community as a collection of nodes with minimum number of links between groups without taking into account the internal structure of the group [9].

Graph partitioning algorithm by Kernighan-Lin uses this approach of the community [3]. In this algorithm user has to specify size and number of groups to be formed. The difference between the edges within the group and edge outside the group is to be optimized.

Challenges Involved

i. Internal connection of the group is not considered [3, 9].
ii. There is no way to prefer densely connected internal nodes group [3].
iii. Number of communities has to be predefined [9].

Clustering view: It aims to maximize the internal density of the group [3, 10]. The basic concept is to group the nodes in such a way that nodes within a group having frequent connections with the nodes within group and sparse connections outside group. Advantage of this approach was that number of groups need not to be predefined.

The algorithm by Newman Girvan is the most important one in this area [11]. In this algorithm divisive approach is used to remove edges with high betweenness.

Challenges Involved

i. A well defined stopping criterion has to be defined such as Modularity [12].
ii. This problem turns into optimization problem by using Modularity parameters.
iii. It is difficult to determine an optimal clustering algorithm [11, 12].

Stochastic Block Model View: This approach does not maximize internal density or minimize external links [13]. It uses the concept of structural equivalence. In this a group of such nodes is determined which connect to nodes of other communities in an equivalent way. It has several advantages. Community can be determined even from bipartite graph which was not possible in the earlier two approaches. It can also be used for forming benchmark datasets.

Challenges Involved

i. It requires more complex calculations than other approaches [13].

Dynamic View: This approach is different from other approaches. In this the structure of communities is not of prime importance rather the behavioural pattern between the nodes is important [8]. So the interest is on how short term dynamics change long time behaviour of system in a network. This is useful where communities are well defined but dynamics within are difficult to understand.

Challenges Involved

i. It is still in infancy and mainly applied to diffusion dynamics [2, 8].
ii. Research is needed to apply it to other complex systems dynamics [8].

3.2 Dynamic Nature of Communities an Issue

Most of the work in community detection is performed on a static network. These cannot be applied on Complex Network which is ever changing like Social Media [2, 14].
Dynamic communities keep on changing with time. Mathematically, a dynamic community can be denoted by an ordered pair of (nodes, periods) [14]. Here nodes means clusters at any instant of time and period represents that time period.
As dynamic communities change with time, different situations are faced in detection of such communities. These are discussed below and shown in figure 3.

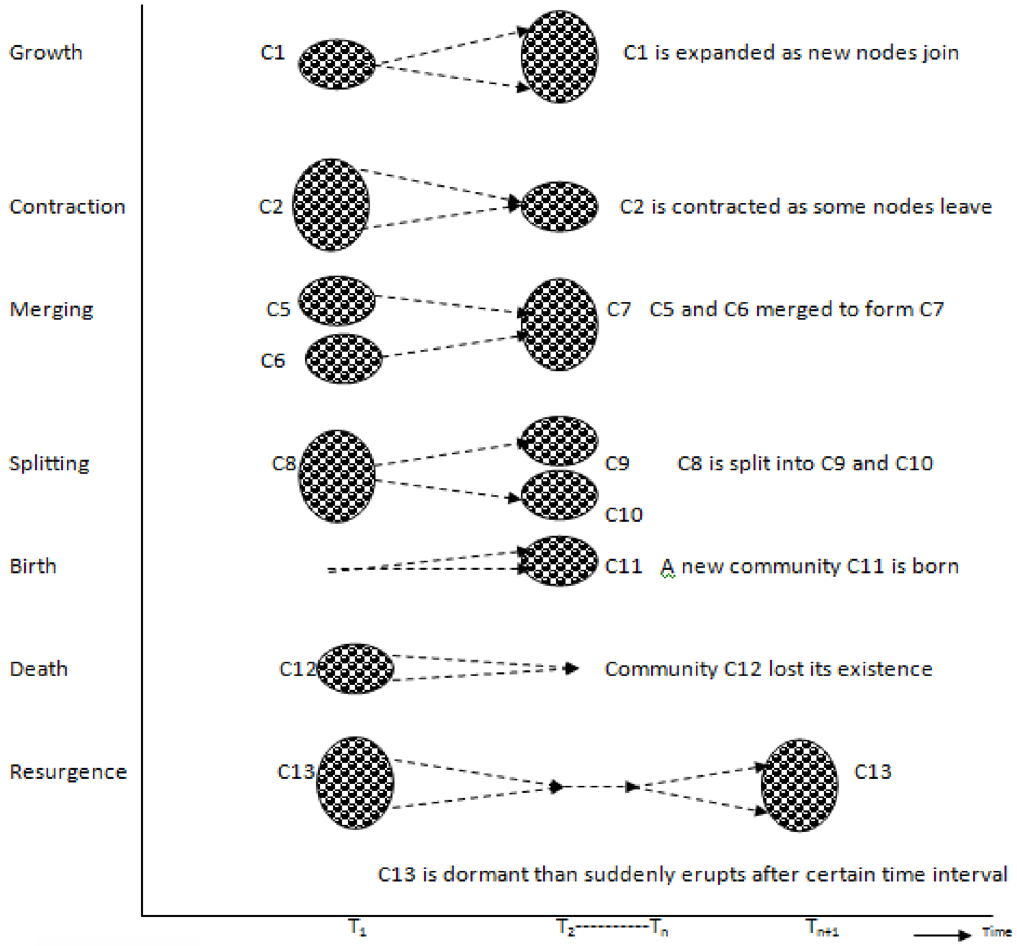
Growth
C1
C1 is expanded as new nodes join
Contraction
C2
C2 is contracted as some nodes leave
Merging
C5
C6
C7
C5 and C6 merged to form C7
Splitting
C8
C9
C10
C8 is split into C9 and C10
Birth
C11
A new community C11 is born
Death
C12
Community C12 lost its existence
Resurgence
C13
C13
C13 is dormant than suddenly erupts after certain time interval
T_1
T_2-----------T_n
T_{n+1}
Time

Fig.3. Dynamically changing communities

- **Growth**: New nodes can be included in a community with time.
- **Contraction**: Some nodes can leave the community, making the community smaller.
- **Merging**: Different communities can combine with time, resulting in a bigger merged community.
- **Splitting**: Two or more communities may be formed by splitting one community.
- **Birth**: A new community can emerge which was not existent at an earlier time interval.

- **Death**: A community can entirely disappear at any time.
- **Resurgence**: Community can be dormant for certain period and then reappear as if nothing happened.

Challenges faced in the research of dynamic communities are:-

i. Most of the existing algorithms do not consider above operations on community, as they are designed for static community [2, 3].
ii. Defining dynamic communities is more complex as they are not just nodes and edges [14].
iii. Multiple snapshots of community have to be considered and evaluated for better result.

To solve these challenges, the problem of dynamic community detection is seen as a combination of static community detection at different time stamps. Several instances of Social Media under study are plotted and traditional algorithms of static community detection can be used at each instance of graph and difference between the communities at different timestamp is studied [14].

3.3 Communities Overlap an Issue

Most of the early research in community detection assumes community to be disjoint group of densely connected nodes. But in real world a person can be member of more than one community. This also applies in Social Media. It means generally the communities are not disjoint, they are overlapping in nature [15] as shown in Fig.4. Here A, B, C and D are overlapping communities. Hence we cannot ignore to include the "Overlap" attribute of communities.

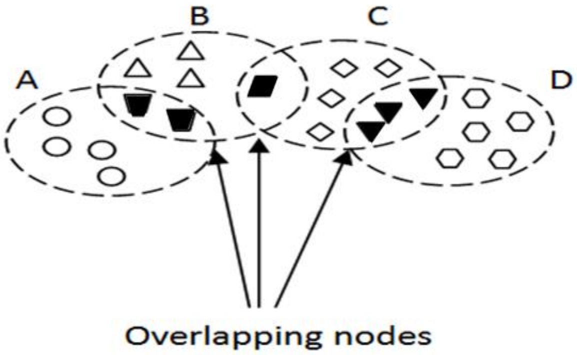

Fig.4. Overlapping Communities

Challenges encountered in the discovery of overlapping communities are:-

i. Identifying the nodes that are common in two or more communities [15].
ii. Identifying the degree of association of node to a particular community [16].

In overlapping community detection, a cover is formed comprising of different clusters C= {c_1, c_2,...., c_k} [15,16]. A node can be member of more than one cluster. In addition to this a belonging factor can be used to judge the associativity of a node with a particular cluster.

So overlapping community detection algorithms can be partitioned into two types crisp and fuzzy [16]. In crisp algorithms belonging factor is not considered, a node can either belong to a community or not. On the other hand, fuzzy algorithms give importance to belonging factor to judge the association of a particular node to a particular cluster.

3.4 Validation of Communities an Issue

Taking motivation from the work of Kleinsberg-Lin [3] and Girvan Newman [11], many community detection algorithms are developed. Individually most of the algorithms perform well, but comparing the performance of these algorithms is difficult. So the last big issue in this area is the issue of validation [17] of communities formed, and performance measure of algorithms used.

Challenges Involved

i. Are the discovered community reasonably correct? [17]
ii. Which algorithm works better under what condition? [2, 17]

Since a variety of different approaches are available for community detection, so a comparison is required to know the success and failure of each algorithm applied. Traditionally artificially generated networks or benchmark datasets such as Zachary's Karate Club, Lusseau's network of bottlenose dolphins [3] etc. are used as ground truth communities to test the algorithms. After this metrics such as Modularity, Rand Index, Normalized Mutual Information [3] etc. are applied to judge the performance.

But in real Social media network these correct partition and data generating process are unknown. So there is no ground truth to judge the performance. Moreover algorithms which work well for standard dataset, do not necessarily give good performance for real world dataset. For this reason, it is also an open area of research.

4. CONCLUSION

Community detection is an important area of research in the fields of computer science, biology, social science etc. Though research on community detection is being done since a long period, still it has various issues and challenges that are unsolved. The challenge start from the very beginning as there is more than one way to identify community. Also Communities overlap and they change over time. There is lack of ground truth to judge the performance of generated community. The research is still going on in this area but as we move deeper new challenges arise. So it is having vast scope for future research. In this paper we have tried to describe the major issues and challenges in CD that serve as a reference for both the scientist and researchers working in this area.

REFERENCES

1. Boccaletti, Stefano, et al. "Complex networks: Structure and dynamics." *Physics reports* 424.4-5 (2006): 175-308.
2. Zafarani, Reza, Mohammad Ali Abbasi, and Huan Liu. *Social media mining: an introduction*. Cambridge University Press, 2014
3. Fortunato, Santo. "Community detection in graphs." *Physics reports* 486.3-5 (2010): 75-174.
4. Gulbahce, Natali, and Sune Lehmann. "The art of community detection." *BioEssays* 30.10 (2008): 934-938.
5. Sahebi, Shaghayegh, and William W. Cohen. "Community-based recommendations: a solution to the cold start problem." *Proceedings of WOODSTOCK'97* (1997).
6. Sayyadi, Hassan, Matthew Hurst, and Alexey Maykov. "Event detection and tracking in social streams." *Icwsm*. 2009.
7. Campan, Alina, Yasmeen Alufaisan, and Traian Marius Truta. "Community Detection in Anonymized Social Networks." *EDBT/ICDT Workshops*. 2014.
8. Rosvall, Martin, et al. "Different approaches to community detection." *arXiv preprint arXiv:1712.06468* (2017).
9. Papadopoulos, Symeon, et al. "Community detection in social media." *Data Mining and Knowledge Discovery* 24.3 (2012): 515-554.
10. Malliaros, Fragkiskos D., and Michalis Vazirgiannis. "Clustering and community detection in directed networks: A survey." *Physics Reports* 533.4 (2013): 95-142.
11. Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." *Proceedings of the national academy of sciences* 99.12 (2002): 7821-7826.
12. Lancichinetti, Andrea, Santo Fortunato, and Filippo Radicchi. "Benchmark graphs for testing community detection algorithms." *Physical review E* 78.4 (2008): 046110.
13. Karrer, Brian, and Mark EJ Newman. "Stochastic blockmodels and community structure in networks." *Physical review E* 83.1 (2011): 016107.
14. Cazabet, Remy, and Frédéric Amblard. "Dynamic community detection." *Encyclopedia of Social Network Analysis and Mining*. Springer New York, 2014. 404-414.

15. Palla, Gergely, et al. "Uncovering the overlapping community structure of complex networks in nature and society." *nature* 435.7043 (2005): 814.
16. Xie, Jierui, Stephen Kelley, and Boleslaw K. Szymanski. "Overlapping community detection in networks: The state-of-the-art and comparative study." *Acm computing surveys (csur)* 45.4 (2013): 43.
17. Lancichinetti, Andrea, Santo Fortunato, and Filippo Radicchi. "Benchmark graphs for testing community detection algorithms." *Physical review E* 78.4 (2008): 046110.
18. Lin, Shen, and Brian W. Kernighan. "An effective heuristic algorithm for the traveling-salesman problem." *Operations research* 21.2 (1973): 498-516.
19. Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." *Proceedings of the national academy of sciences* 99.12 (2002): 7821-7826.
20. Newman, Mark EJ. "Fast algorithm for detecting community structure in networks." *Physical review E* 69.6 (2004): 066133.
21. Clauset, Aaron, Cristopher Moore, and Mark EJ Newman. "Hierarchical structure and the prediction of missing links in networks." *Nature* 453.7191 (2008): 98.
22. Zhou, Ding, et al. "Probabilistic models for discovering e-communities." *Proceedings of the 15th international conference on World Wide Web*. ACM, 2006.
23. Palla, Gergely, et al. "Uncovering the overlapping community structure of complex networks in nature and society." *Nature* 435.7043 (2005): 814.
24. Raghavan, Usha Nandini, Réka Albert, and Soundar Kumara. "Near linear time algorithm to detect community structures in large-scale networks." *Physical review E* 76.3 (2007): 036106.