A novel approach for community detection in Social Networks

Sreedhar Bhukya, G Shiva Krishna and B Shobini
[1]Student,[2]Assistant Professor,[3]Assistant Professor
[1]Department of Computer Science and Engineering,
Swathi Institute of Technology & Science, Hyderabad, India
Email: sreedharbhukya@gmail.com, shivaguhjju@gmail.com, shobini.b@gmail.com

***Abstract:*Social Network Analysis (SNA) is computing and mapping of relationships and movements between people, groups, organizations, URLs, computers and other connected information/knowledge bodies. The people and groups are represented as nodes whereas the show movements or relationships between the nodes are represented as links. Community discovery is one of the challenging problems in any networks, we have proposed a community discovery algorithm which is producing better result than existing algorithm. Proposed and Existing algorithms are based on edge betweenness [1] score. We show that novel approach algorithm performs quite well on bench mark data sets and a few real-world networks. We have also considered the core nodes of the graph and identifies the importance for each node.**

***IndexTerms* - Edge betweenness, Betweenness centrality, GN algorithm, Novel approach**

I. INTRODUCTION

II. Social networks are modelled as directed or undirected graphs G = (V, E), V and E the set of nodes and edges respectively. Where V represents user of a group or node of any graph and E represents relationship between two vertices of a social network graph. Social networks are forming through real world or it may form a group of a community networks are said to exhibit strong community structure. Communities are groups of nodes in a network where the nodes within a group are more densely connected than between the groups. Partitioning the graph network into communities such that the graph is modularized optimally is an NP-Hard problem.

III. In networks, fundamental tools like Community detection algorithms enable us to identify organizational principles. While identifying communities, the structure of network, the features and the node characteristics can be the probable resources of information one can make use of. Detecting communities can be considered as an issue due to clustering of group of nodes into communities. The nodes in a community share common properties, and hence several relationships may exist among themselves, and hence a node can be a member of several communities. Efficient community detection is the important challenging problem that exists in social networks.

IV. The core nodes are central and essential in a community and they are known as the special nodes. The core nodes are used to build the community or graph and if they are removed, the community will shatter. Numerous edges are shared by the nodes which are located in the central position along with the other group partners. Hence the core nodes have agglomeration and high degree

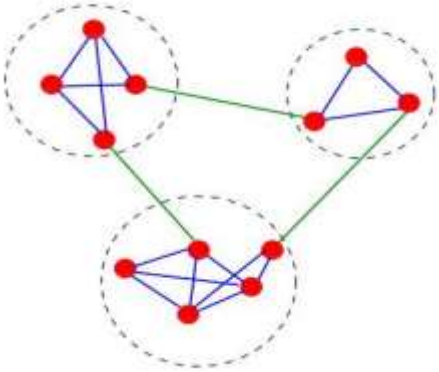

Figure 1: A snapshot of three communities within the circle, having three bridge edges between communities and each node has density with three communities.

The major objective is to find the communities existing in the social networks using density-based approach. We proposed a novel approachalgorithm for community detection with the help of density as well as some fixed thresholds between nodes are becomes a particular community, it is calculating with dynamically. Consider a graph has different threshold will appear in every node example by taking any community graphs like Author's collaboration network, Twitter network and Facebook network and each network has different density of each user or node. We observed that in Facebook network all the user they do not have equal number of friends in each group. Our proposed algorithms also find the communities by identifying core nodes and their corresponding neighbors nodes following density-based approach.

Many algorithms are proposed for detect the various communities with different algorithms. One of the challenging is detect the exact community without any disturbing of any nodes is important, but one papers talks about the node density and core nodes, community structure and others. As we can see some of the problem with community detection is quite well-studied [1-5]. One of the classical algorithms for community discovery is that of Girvan and Newman (GN) who propose a simple way to identify communities in a graph by detecting edges that connect vertices of different communities, removing which will make the clusters get disconnected from each other. Since ground truth is not known in most cases [6], we consider the results given by GN algorithm as the true communities.

Methods to break a network into set of connected components called regions:

Divisive Method: Repeatedly identify and remove edges connecting density connection regions.

Agglomerative Method: Repeatedly identify and merge nodes that likely belongs in the same iteration.

GN algorithm uses a centrality measure called edge betweenness (EB) score that measures the fraction of all shortest paths passing through the given edge [7-13]. By removing links with high betweenness score, the network is progressively split into disconnected components, until the network is decomposed into the desirable number of communities. Implementation of Brandes [14-17] for the GN algorithm is considered an efficient algorithm that gives exact communities without disturbing internal edges within the community. Other algorithms are based on [18-23] density approach.

There are several approaches to identify the community structure where each method relies on one of the distinguishing features of the network. Community discovering algorithms can be broadly classified

into two categories namely accommodative (Agglomarative) algorithms where similar nodes are added up to form a community starting from a null graph. Second is a divisive algorithm where edges joining dissimilar nodes are removed iteratively.

Agglomerative methods involve in considering the number of node independent paths or number of edge independent paths as metric. This leads to dendogram structure [24]. This method is good in discovering strongly linked core of communities [24]. Another simpler traditional method is the graph partitioning algorithm.

This is the philosophy of divisive algorithms and GN is the most representative method of divisive methods. It is based on the edge betweenness [25] that measures the fraction of all shortest paths passing on a given link. By removing links with high betweenness, we can progressively split the whole network into disconnected components, until the network is decomposed in communities consisting of one single node. But this method is computationally inefficient. So far, many algorithms proposed for detecting communities for complex networks, biological, computer-generated and social networks, propose a local community structure which is call local modularity, which works on a fast-agglomerative algorithm that maximizes the local modularity in a greedy fashion. [27-30] Proposed community structure based on modularity, initially it divides two communities and further it divides more communities.

a new local algorithm based on node strength is proposed to detect the overlapping community structures. The main strategies are to find an initial community from a node with maximal node strength and to expand the partial community from the initial one by adding nodes that are tight with the community [30-36] proposed through the community discovery based on strongly connected components on mutual accessibility. The idea is simple: the values of the eigenvector components are close for vertices in the same community, so one can use them as coordinates to represent vertices as points in a metric space. So, if one uses M eigenvectors, one can embed the vertices in an M-dimensional space. Communities appear as groups of points well separated from each other. Here we have considered and mainly focused on a community detection algorithm based on edge betweenness and we proposed a enhanced (Girvan and Newman, 2002) algorithm for Community detection based on edge betweenness score, more efficient algorithm for detecting communities are edge betweenness, this gives exact communities without disturbing internal edges within the community and more focus on removing overlap links only. So far, community detection algorithms are proposed as fellows on betweenness.

1. First time discuss about point centrality and betweenness centrality calculation by [2] to calculate betweenness for nodes in a graph or network.
2. 2. Firstly [1] proposed algorithm for calculating the edge betweenness, introduce more efficient algorithms based on a new accumulation technique that integrates well with traversal algorithms solving the single-source shortest-paths problem.
3. Detecting communities an efficient algorithm of (Girvan and Newman) calculating edge betweenness score for all edges and removing with highest betweenness edge until formation of desired community, it occupies high time complexity for running the algorithm. Algorithm is computationally expensive because it requires the repeated evaluation, for each edge in the system, but this algorithm works for a good community detection for sparser and complex graphs. The algorithm runs $O(n3)$ on sparse graphs.

4. For community detection [16] proposed a self-contained version of the GN algorithm, this algorithm is also calculating the edge betweenness score for edges, it is also computationally expensive, to overcome to this problem they consider the edge-clustering coefficient, defined with node-clustering coefficient as number of triangles which given a edge belongs. But this algorithm is not suitable for sparser graphs for detecting communities.

2. BETWEENNESS CENTRALITY MEASURE:

The edge betweenness centrality is defined as the number of the shortest paths that go through an edge in a graph or network (Girvan and Newman 2002). Each edge in the network can be associated with an edge betweenness centrality value. An edge with a high edge betweenness centrality score represents a bridge-like connector between two parts of a network, and the removal of which may affect the communication between many pairs of nodes through the shortest paths between them. Figure 1 illustrates an example of fourteen nodes in a network, and the red edge between two red nodes has a high edge betweenness centrality value of 49. The removal of this edge will result in a partition of the network into two densely connected subnetworks.

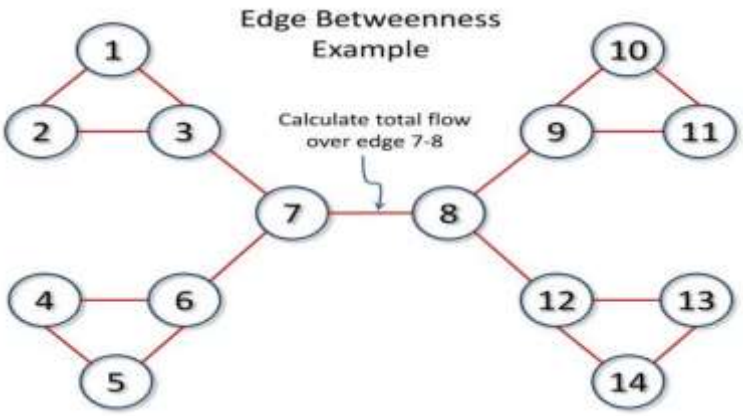

Figure 2 illustrates an example of fourteen nodes in a network

In another way the betweenness centrality measure plays a key role for community discovery. If one removes the bridge edges with highest betweenness score then the graph is divided into more than one community. Let σu,v(i, j) denote the number of shortest paths from u to v containing the edge (i, j) and σu,v denotes the number of shortest paths from u to v in the undirected graph G and n is a number of nodes in a given graph G. Then edge betweenness $C_B(i, j)$ is defined as

$$C_B(i,j) - \sum_{u,v \in E, u<v} \sigma_{u,v}(i,j)$$

Normalized betweenness centrality measure:

$$C'_B(i,j) = \frac{2C_B(i,j)}{n(n-1)}$$

3. EXISTING SYSTEM:

Several techniques are proposed based on Edge betweenness score which is called as divisive method, partition clustering is already exist for detecting communities in social networks.

GN proposed on Edge betweenness score for detect the community.

1. Step1: Calculate the Edge betweenness for all edges
2. Step2: Remove edge with highest betweenness score
3. Step3: Recalculate the edge betweenness score for all edges
4. Step4: If found the communities then exit
5. Step5: repeat steps 1 to 4.

This algorithm is works well for smaller graph with average time complexity where as if graph becomes larger than the performance of the algorithm is less due to high time complexity.

4.PROPOSED NOVEL APPROACH:

Our proposed novel approach algorithm is also works based on edge betweenness score and it shows the community structure.

Steps:

1. Calculate the Edge betweenness and remove highest betweenness edge [same as GN algorithm].
2. Additionally, remove other top edges with highest edge betweenness score [Threshold].
3. Recalculate the edge betweenness score for all remaining edges.
4. If desire the number of communities then exit
5. Repeat step 1 to 4.

The proposed approach is works well comparative the existing system and our proposed algorithm is removing more than one edge at a time where as earlier algorithm in always remove one edge at a time in each iteration. Our approach has less time complexity based on fallowing's.

- The proposed is based on edge betweenness score and density-based approach is used in our proposed framework to detect the communities in a social network.
- The edge betweenness score is calculated as well as density is calculated with a node, it has number connections establishes with others in social.

- Our proposed approach is also works based on edge betweenness score but it removes more edges to detect community.
- We refer several techniques based on divisive for partition clustering and latent space clustering etc., already exist for detecting communities in social networks.
- A node has to be selected from a group of nodes in the graph and identifying the core nodes.

The density-based approach is used in our proposed framework to detect the communities in a social network. The density is calculated with a node has number connections establishes with others in social graph and density is divided by a node has number of connections are divided by number of nodes in a connected graph. In our proposed system, based on input resolution parameter (η), the values of a minimum cluster size (μ) and a global neighborhood threshold (ε) are automatically computed for each node dynamically. Our proposed architecture for community detection follows density-based approach

6. RESULTS:

Datasets Zachary karate club: This is a well-known bench mark network in which each node represents a member of the club, and edges represent interactions. The interactions among the members are supposed to reveal a political division among the members has different densities with size of 28 users.

	Zachary	
It.	Edges	Score
1	32-1	71.39
2	3-1	66.89
3	9-1	77.31
4	34-14	82.03
5	34-20	123.23
6	33-3	100.20
7	31-2	143.62
8	9-3	95.08
9	28-3	122.40
10	29-3	171.16
11	10-3	288.00

	Zachary	
Itere.	Edges	Score
1	32 - 1	71.39
1	34 - 14	38.05
1	34 - 20	33.31
2	9 - 1	91.28
2	3 - 28	45.00
2	31 - 2	32.49
3	33 - 3	180.30
3	29 - 3	49.06
3	9 - 3	121.00
4	34 - 10	289.00

Table 1: shown the Existing GN algorithm runs 11 iterations to detect the community where as our novel proposed community discovery algorithm takes the 4 iterations to detect the same community.

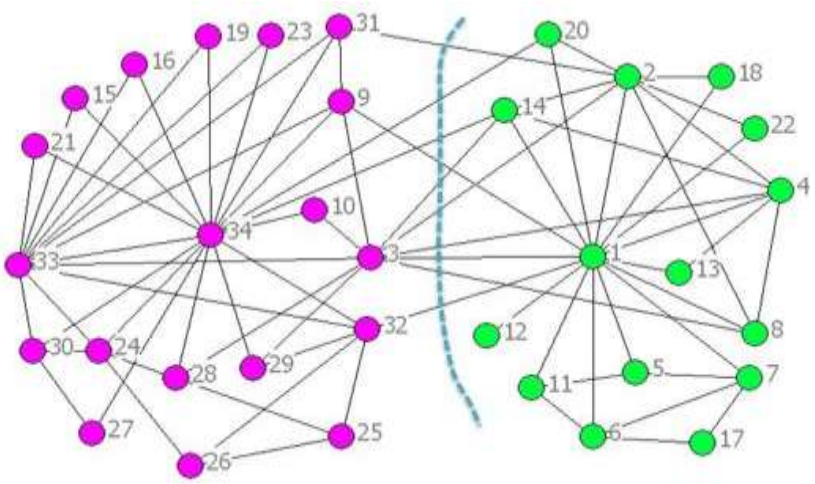

Fig 3. the above figure shows two communities, the pink-coloured node no. 33 and 34 are acting as core nodes and another Green coloured community has node 1 and 2 as considered as core nodes. This graph from Zachary Karate club of 34 members.

TABLE 1. shows core nodes and its neighbor nodes from the dataset. It also shows that total node is 19, total core node is 7 and the remaining node is not core node because it doesn't have the neighbors (nodes). So, the core node represents the community.

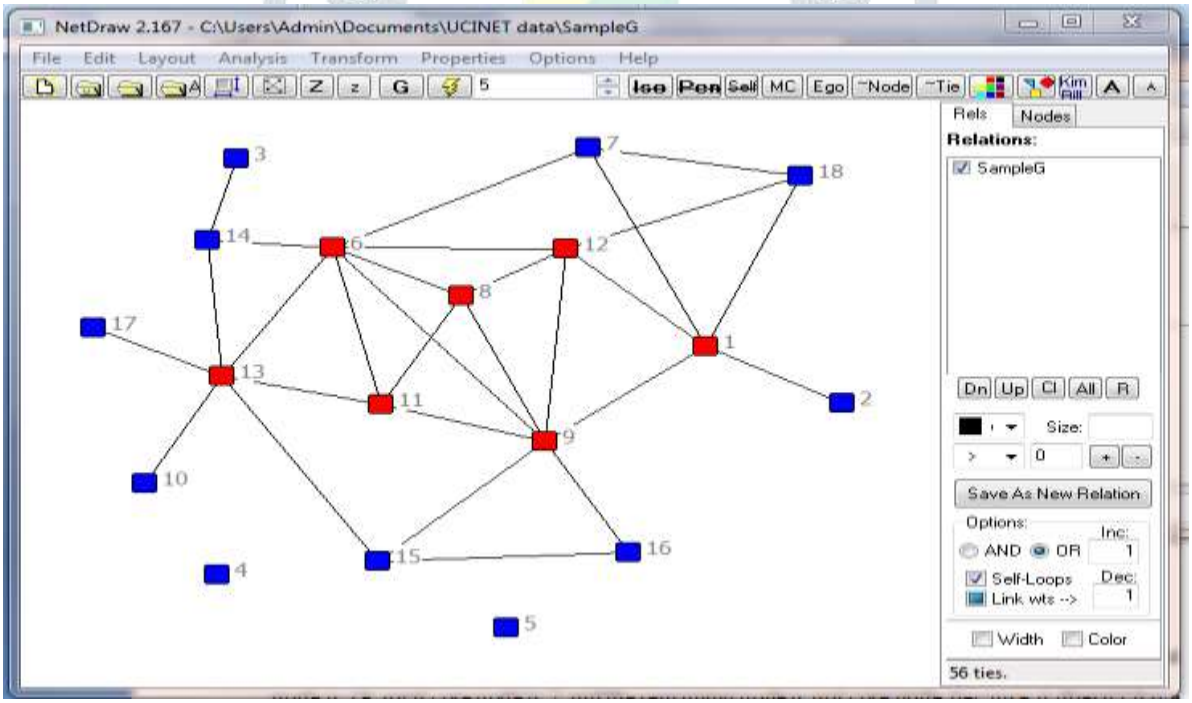

Figure 4. Shows the core nodes from the above graph

Core Node	Neighbour Nodes
1	2,7,9,12,19
6	7, 8, 9, 11, 12, 13, 14
8	6,9,11,12
9	16,8,11,12,15,16,
11	6, 8, 9, 13
12	1, 19, 6, 8, 9
13	17, 6, 10, 11, 14, 15

Table 2: Shows each core nodes with their neighbours.

The importance of the core nodes is that which has highest neighbors they are the leading the community and we have been observed in the real time networks.

6.CONCLUSION

The betweenness centrality is a one of the main important to detect the exact community or cluster of graphs and our proposed algorithm is saving the half of the time complexity as well as density-based approach is also used in our proposed framework to identify the communities of a social network. The main aim of our proposed approach is fallowing the existing community detection methods as well as it follows density-based approach. The assumption of values for two input parameters such as a threshold for global neighborhood and minimum size of a cluster. At the beginning of the process the assumption of neighborhood parameter is not required in our proposed framework. Based on input resolution parameter the neighborhood threshold and a local version of minimum cluster size is calculated automatically for each node locally. We presented experimental results in detecting communities by following our proposed framework. By applying our proposed algorithm and framework, it works efficiently to identified the communities exist in various real time datasets.

REFERENCES

1. Sreedhar Bhukya, A novel model for social networks," *2011 Baltic Congress on Future Internet and Communications*, pp. 21-24, 2011. doi: 10.1109/BCFIC-RIGA.2011.5733213.
2. Newman, M. E. J., Community structure in social and biological networks, Proc Natl Acad Sci, USA, volume.101, pp.5200-5205, 2004.

3. Newman, M. E. J., The structure of scientific collaboration networks, Proc Natl Acad Sci, USA, volume.98, pp.404-409, 1998.
4. G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," Nature, vol. 435, no. 7043, pp. 814–818, 2005.
5. T. Falkowski, Community Analysis in Dynamic Social Networks.Gottingen, Germany: Sierke, 2009.
6. P. Kumar, L. Wang, J. Chauhan, and K. Zhang, "Discovery and visualization of hierarchical overlapping communities from bibliography information," in Proc. IEEE 8th Int Conf. Depended,Autonom. Secure Comput., 2009, pp. 664–669.
7. Zhang, Tiantian, and Bin Wu. "A Method for Local Community Detection by Finding Core Nodes", 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2012.
8. R. Cazabet, F. Amblard, and C. Hanachi, "Detection of overlapping communities in dynamical social networks," in Soc. Comput.(SocialCom), 2010 IEEE Second Int. Conf., Aug. 2010, pp. 309–314.
9. McDaid and N. Hurley, "Detecting highly overlapping communities with model-based overlapping seed expansion," in Proc. Int. Conf. Adv. Soc. Netw. Anal. Min., 2010, pp. 112–119.
10. Stanford Large Network Dataset Collection. Available [Online]: http://snap.stanford.edu/data/
11. Wikipedia Adminship Election Data. Available [Online]: http://snap.stanford.edu/data/wiki-Elec.html
12. J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in Proc. 19th Int. Conf. World Wide Web, 2010, pp. 641–650.
13. S. Gregory, "Finding overlapping communities in networks by label propagation," New J. Phys., vol. 12, no. 10, p. 103018, 2010.
14. D. Greene, D. Doyle, and P. Cunningham, "Tracking the evolution of communities in dynamic social networks," in Proc. Int. Conf. Adv. Soc. Netw. Anal. Min. 2010, pp. 176–183.
15. J. Huang, H. Sun, J. Han, and B. Feng, "Density-based shrinkage for revealing hierarchical and overlapping community structure in networks," Physica A: Statist. Mech. Appl., vol. 390, no. 11, pp. 2160– 2171, 2011.
16. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, "Finding statistically significant communities in networks," PLoSONE, vol. 6, no. 4, pp. 1–18, 2011.
17. J. Xie and B. K. Szymanski, "Towards linear time overlapping community detection in social networks," in Proc. 16th Pacific-Asia Conf. Adv. Knowl. Discovery Data Min., 2012, vol. 2, pp. 25–36.
18. Jaewon Yang, Julian McAuley and Jure Leskovec, "Community Detection in Networks with Node Attributes", in IEEE 13th International Conference on Data Mining, 2013.
19. http://www.orgnet.com/sna.html
20. https://en.wikipedia.org/wiki/Social_network_analysis.
21. Newman, M. E. J., The Structure and Function of Complex Networks, SIAM Review, volume.45, pp.167-256, 2003.
22. Clauset, Aaron,Finding local community structure in networks, American Physical Society, volume.72, pp.026-132, 2005.
23. J.Yang, J.Leskovec, Defining and evaluating network communities based on ground-truth, Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics,pp.3:8-3:1, 2012.
24. J.Yang, J.Leskovec, De_ning and evaluating network communities based on ground-truth,Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics,pp.3:8-3:1, 2012.
25. Du, Nan, Wang, Bai and Wu, Bin, Community detection in complex networks,J. Comput. Sci. Technol., volume.23, pp.672{683, 2008.

26. Ulrik Brandes, A Faster Algorithm for Betweenness Centrality,Journal of Mathematical Sociology, volume.25, pp.163-177, 2001.

27. M.E.J. Newman, arXiv:1307.7729v1, July, 2013.

28. S. M. Van Dongen, Graph Clustering by Flow Simulation, PhD thesis, University of Utrecht, May 2002, 2002.

29. Steve Gregory, An Algorithm to Find Overlapping Community Structure in Networks, PKDD, pp.91-102, 2007.

30. Lee, M-J Lee, J, Park, J Y, Choi, R.H, Chung, C-W , QUBE: A quick algorithm for updating betweenness centrality, Proc.of 21st international conference on World Wide Web WWW '12, pp.351-360,2012.

31. V.U_mtsev, S.Bhowmick, Application of Group Testing in Identifying High Betweenness Centrality Vertices in Complex Networks, Eleventh Workshop on Mining and Learning with Graphs, ACM Press, 2013.

32. Davis,D, Lichtenwalter,R.N, Chawla, N.V., Supervised methods for multirelational link prediction, Social Network Analysis and Mining, volume.3, pp.127-141, 2013.

33. Watts, D.J., Strogatz, S.H., Collective dynamics of small-world networks, Nature, volume.393, pp.440-442, 1998.

34. Sreedhar Bhukya,A Model for Mobile Social Network Growth, International Journal of Computer Applications 180(27):16-19. 2018.

35. Sreedhar Bhukya, Community discovery in a growing model of social networks," *2010 IEEE International Workshop on: Business Applications of Social Network Analysis (BASNA)*, 2010, pp. 1-5, doi: 10.1109/BASNA.2010.5730299.

36. Sreedhar Bhukya, "Discover Academic Experts in Novel Social Network Model", *Advances in Social Networks Analysis and Mining (ASONAM) 2011 International Conference on*, pp. 696-700, 2011.