Improved Community Detection Algorithm Based on Distance Centrality in Social Network

Juhi Prajapati[1], Dr.S.M.Shah[2]

[1] Gujarat Technology University, Department of Computer Engineering, GEC Modasa, Gujarat, India
[2]Professor, Department of Computer Engineering, GEC Modasa, Gujarat, India

ABSTRACT

In today's era, online social networks have become quite prominent, and a lot of research is getting done for making these networks useful and more enjoyable, by suggesting useful communities and similar interest communities to users. Also due to freely available web space and interactions, there are multi communities per user, making it really hard for highly iterative algorithms to perform fast and give meaningful suggestions. So, using community detection algorithm provides some important information from social network. Social network is one of the most vital complicated networks. Many existing community detection algorithms are developed. Distance Centrality based Community Detection algorithm required the more recalculations of center node for each community. In this paper, we propose improved Distance Centrality based Community Detection algorithm which reduce the recalculations of center node for each community. First we calculate the distance between nodes and get the pre-community then choose the center node by calculating the closeness centrality of each node using their distance data. Then check the similarity between pre-community nodes and non pre-community nodes and assign every node to the foremost similar community. We will get communities from the social network. We demonstrate that the proposed improved Distance Centrality based Community Detection (iDCCD) algorithm terminated on a decent community number and additionally has comparable detection accuracy with different existing approaches.

Keyword: - *Social Network, Community Detection, Distance Centrality, Similarity.*

1. INTRODUCTION

We can extract the necessary information from massive resources of information using data mining technique. It provides extracting relevant information from the massive volume using suitable algorithms. Social Network Analysis is the way to study the social phenomena, particularly social setting. There are many social networks, like Facebook, YouTube, Email, Google+, Twitter, etc. Social network is represented as graph. A social network includes group of nodes and nodes are connected by the edges. In the social network nodes are shown as objects, like peoples, cities, organizations, etc. The edges show the relationships between nodes like friendship, common interest, interaction (like comment, post, share, tag, message etc.), religion, education so on. Example of social network is circle of friends, the nodes represent people, and the edge between the nodes represent the friendship between people.

Community is a very important structure in social networks. Communities are also called "clusters or modules, groups of nodes which probably share common properties and/or play similar roles within the graph [7]". In general, Communities of network are groups of nodes within which nodes are much more connected with each other than to the other nodes of the network. Detection of community is a famous topic in social network analysis and also in knowledge mining. The aim of a "community detection algorithm is to divide the vertices of a network into some number of groups, while maximizing the number of edges inside these groups and minimizing the number of edges established between vertices in different groups, these groups are the communities of the network [8]". Fig 1 shows the three communities in the graph.

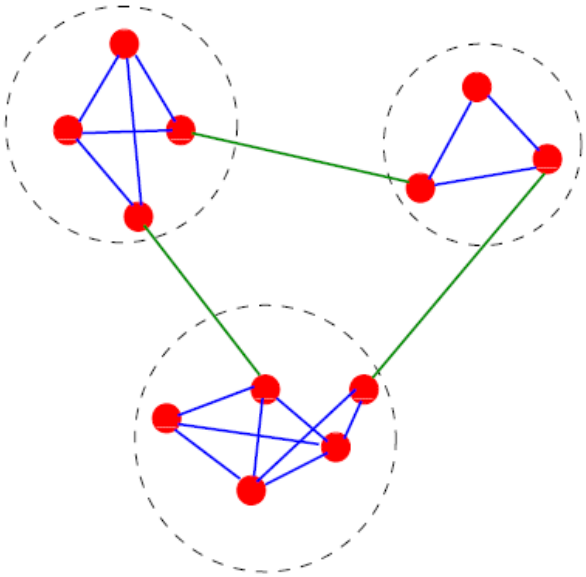

Fig 1: Three communities in the graph [7]

Various applications of community detection are social communities (e.g. families, friendship, villages, states etc.), we obtain communities from protein-protein interaction networks (e.g. group of proteins which have the same function in cell) [15,16, 17], we get the communities from the World Wide Web (e.g. group of the pages which is in the same or related topics) [14,13], analyses marketing strategy of organization.

2. RELATED WORK

In this section, we present the different kind of algorithm which exists for detecting community from social network.

In 2013, Longju Wu, Tian Bai, Zhe Wang, Limei Wang, Yu Hu, Jinchao Ji proposed a DCCD algorithm [1], which is based on distance centrality. Here, we do not need to predefine the number of community for detecting the community. The drawback of this algorithm is that the more recalculations of center nodes of communities are required. Hierarchical agglomerative algorithm [2] is used for community detection in social networks and which is based on enhanced similarity. Drawback is that it has more computational complexity. Expectation Maximization (EM) algorithm [3] is used to detect communities from social network and needed to predefine the number of communities. Authors proposed an adaptive approximation algorithm for community detection in social network [4] that is performed based on vertex relation and their modularity and makes this method more precise based on dynamically calculating modularity. It checks the status of nodes adaptively and gets the communities from network. And dynamically calculate the modularity. Drawback is that it cannot contain the many overlapping nodes in communities. In 2014, Chang Su, Yukun Wang proposed a community detection algorithm based on seed nodes [5] which is not required to predefine the number of community. Author proposes an algorithm for community detection based on hierarchical clustering (CDHC Algorithm) [6] which proposes the concept of extensive modularity, removing some weakness of modularity. The extensive modularity can good calculate the effectiveness of community detection algorithm. It is independent of the number of communities that we detect.

3. BASIC TERMINOLOGY

3.1 Centrality

Centrality is the degree for a node to be the middle of the network [1]. There are different types of centrality of the node available. Here, we use the closeness centrality. It relies on the distance between any nodes.

The **Closeness Centrality** of a node is defined by "the inverse of the average length of the shortest paths to/from all the other nodes in the graph". If the node has the higher closeness centrality then the smaller average distance of all

or any nodes of the network. So closeness centrality specifies the importance of the nodes in a network by the higher closeness centrality of node [9]. First find the closeness centrality, we should calculate the average distance of a node to all others nodes within the network.

$$D_{avg}(V_i) = \frac{1}{n-1} \sum_{j \neq i}^{n} d(V_i, V_j),$$

where, n means the number of node in the network, and $d(V_i, V_j)$ is distance between node V_i and V_j .

So the Closeness Centrality of node V_i is the inverse of $D_{avg}(Vi)$,

$$C_c(V_i) = [D_{avg}(V_i)]^{-1}$$

3.2 Similarity

"Similarity refers to the similar degree between two nodes"., The higher possibility of two nodes to come into the same community, if two nodes have the larger similarity [1].

Jaccard similarity could be similarity index supported local information. And therefore the local information is common neighbors of two nodes. Nodes between higher similarities then there is more common neighbours between nodes [1]. Defined the Jaccard similarity as below:

$$S_{Jaccard} = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|},$$

Where, $\Gamma(i)$ is the neighbors of node V_i , and $\Gamma(j)$ is the neighbors of node V_j .

4. PROPOSED METHODOLOGY

Given a static social network G (V, E), where V=set of nodes and E=set of edges. Our algorithm aims to find communities in social network.
In a proposed algorithm, we have used closeness centrality and jaccard similarity because both closeness centrality and jaccard similarity are very well known and precise measures in the field of data mining.
Input: Static undirected graph G (V, E)
Output: Set of communities

Explanation:
In this algorithm, we find a closeness centrality for each node in the graph. After getting closeness centrality, center node of pre-community will be assigned based on closeness centrality in descending order. In the next step this algorithm will find a pre-communities using center nodes. After getting all the pre-communities, number of nodes in the pre-community will be checked. If it is less than threshold value (Thr), where threshold value (In karate club dataset Thr=3), algorithm will remove this pre-community and make nodes of these communities unassigned to any community.
Now we will have some nodes without being in any community. So to include those nodes we will use an approach which is based on jaccard similarity measure. It described as follows:

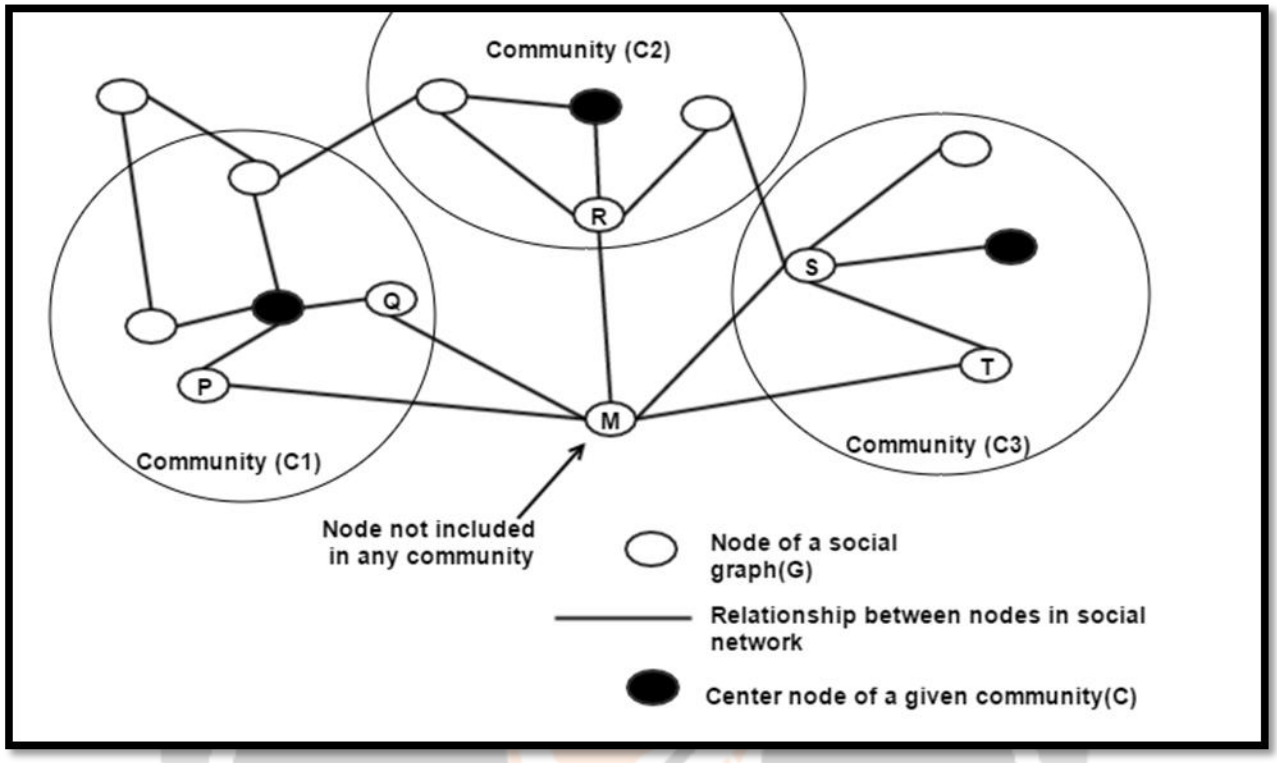

Fig 2: Proposed approach using similarity measure

As shown in fig 2, M is a node which is not included in any communities. To include this node M into any one community, we will use Jaccard similarity. Jaccard similarity is explained in detail in section 3.2. M have five neighbors labeled P, Q, R, S, T. As described in figure:1, node P and Q are part of community (C1) , node R is a part of community(C2) and node S and T are part of community (C3).

To include M into any one of community C1, C2, C3,…., C_i ,

Similarity of a node M and community(C(i)),

$S\_C(i)= J(M,N_1) + J(M,N_2) + J(M,N_3) + \ldots + J(M,N_j)$

Where N1 ,N2, N3, …., Nj are the neighbours of node M and also part of community C(i) and $J(M,N_j)$ is jaccard similarity between node M and node N_j.

As per above equation, we will get similarity score of each community with node M. Now if similarity score is higher, node M will move to the community.

For example,

As shown in figure,

S_C (1) = J (M, P) + J (M, Q)

S_C (2) = J (M, R)

S_C (3) = J (M, S) + J (M, T)

Here suppose S_C (1) is greater than similarity scores of all other communities then M will assigned to community C1.

Now using above approach all nodes will be assigned to each community. And if, some still remains those will be considered as a noise.

Algorithm:

Steps of proposed algorithm are shown below:

Step-1: Distance between all nodes are calculated and less than threshold then come into pre-community.

Step-2: Assign a center node (V_i) to the node with highest closeness centrality in given pre-community.

Step-3: For all the nodes of pre-community (C_i),

Visited (V_n) =1, where $\forall$ $V_n \in$ Vertex-Set (C_i)

Step-4: Now this pre-community (C_i) is becomes a community (C_j) with a centre node (V_i).

Step-5: For each community,

If size_of_pre-communities (C_i)<=3

Remove pre-community (C_i) and make all the nodes of C_i unassigned to community.

Step-6: For each unvisited and unassigned nodes,

Using Jaccard similarity with their neighbors, assign that node to any community (C_i) using a proposed approach explained above.

Step-7: Result of step-6 will form a final expanded community (C_j).

5. EXPERIMENT

5.1. Evaluation Measures

Here, we use the modularity and purity for checking the accuracy of algorithm.

1. Modularity

Modularity measures the good partition or divided the different types of nodes from the network [11].Modularity is proposed by Newman. It is used for measure the quality of community in community detection algorithm [11]. The modularity is defined as actual interactions of nodes and differences between expected numbers of connections for a randomized network (with same nodes where edges created randomly but with respect the node degrees). Community strength is shown as below:

$$\sum_{i\in C, j\in C} A_{ij} - d_i d_j / 2m .$$

Where,

m = edges

A_{ij} = adjacency matrix in row i and column j

d_i = the degree of node i

d_j = the degree of node j

So for the network G, it has divided into k communities, and its modularity can be defined as [1]:

$$Q = \frac{1}{2m}\sum_{l=1}^{k}\sum_{i\in C, j\in C} A_{ij} - d_i d_j / 2m \cdot$$

2. Purity

Purity is "the number of correctly assigned nodes divided by the total number of nodes in V."Purity is a usual measure to compare with the ground truth [10]. Let V be the set of all nodes in network that V= {V_1 ,V_2 , ..., V_n } . A partitioning *Ri* is a subset of V, and $V = \bigcup_1^k R_i$, where $R_i \bigcap R_j = \varnothing (i \neq j)$. Moreover, the partitioning G is the standard partition, and showed as V= {G1 ,G2 , ..., Gn }[1]. assume R and G are two different partitioning of the same data. So purity is computed as:

$$purity(R,G) = \frac{1}{n} \times \sum_j \max_i \left| R_j \cap G_i \right| \cdot$$

5.2 Experimental results

1) Karate club

Karate club is a classic data set used to validate the social network analysis community detection algorithm [12]. The dataset is obtained by sociologists Zachary in the early 1970s, with two years watched the social relations between members of a university in karate club in the US. It includes 34 nodes and 78 edges.

The results of proposed algorithm pre-communities on karate club are shown in Fig 3

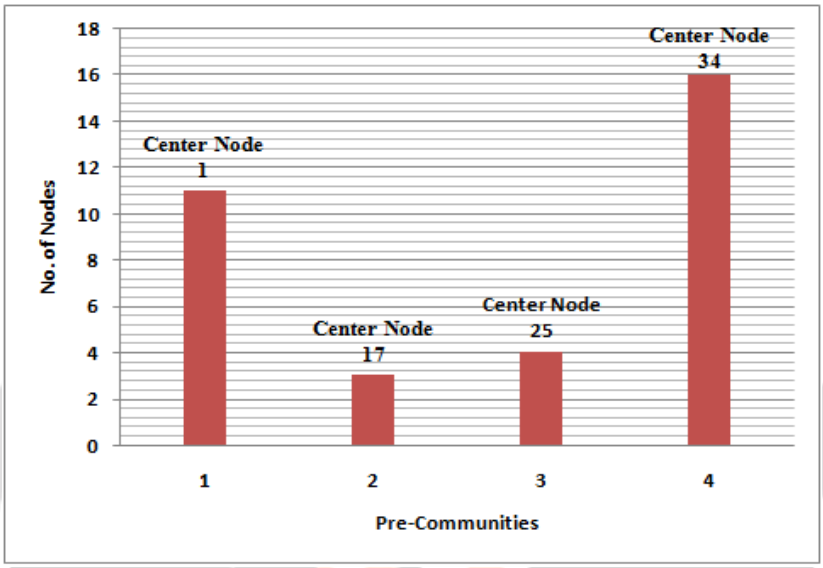

Fig 3: karate club results of pre-communities in graph

The results of proposed algorithm final communities on karate club are shown in Fig 4

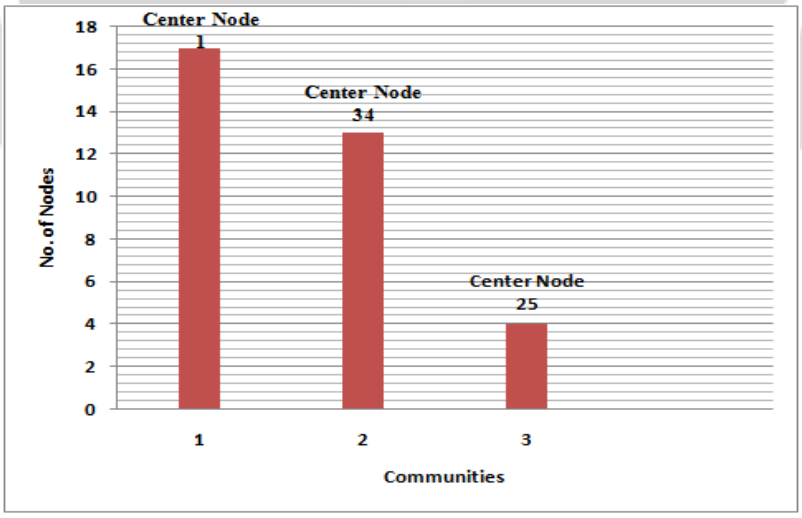

Fig 4: karate club results of communities in graph

Algorithm	Dataset	Number of communities
DCCD	Karate club	2
Improved DCCD		3

Table 1: Number of communities

Comparison on karate club results with existing algorithms

Algorithm	**Modularity(Q)**	**Purity**
Improved DCCD	0.376	0.97
DCCD[1]	0.371	1.00
CFinder[1]	0.182	0.65
SCAN[1]	0.312	0.76
TopLeader [1,18]	0.371	1.00
TopLeader [1,19]	0.374	1.00
TopLeader[1,9]	0.361	1.00
Fast Modularity [1]	0.380	0.97

Table 2: Results of Modularity and Purity on karate club

As shown in Figure 2, the network has been divided into four pre-communities. The center nodes of communities are node 1, node 17, node 25 and node 34.And we will get the four pre-communities with center nodes. Then if pre-community with less than 3 nodes than delete this pre-community with center node and we will check the similarity between pre-community nodes and non pre-community nodes. We will get the final three communities of network with center nodes are 1, 25 and 34 which is shown in fig: 3.

Modularity value of improved DCCD algorithm is good than DCCD algorithm and purity value of improved DCCD algorithm is near to ground truth value. Benefit of improved DCCD algorithm is that we do not require recalculations of center node for each community whereas in DCCD requires more number of recalculations of center node for each community. It is compared with different methods. Top Leader gives the good results. But it needs few predefined data. And the results of modularity value do not give good result if number of cluster value is set to 4. Although other algorithms do not would like additional predefined information, the results show even worse. But in our approach, it gives better result on modularity value with less prior information.

6. CONCLUSIONS

Here, we proposed an improved community detection algorithm based on distance centrality in social network. In the algorithm, we uses distance centrality and jaccard similarity for discover community. Using distance centrality we can get the center node of each pre-community which is based on the distance between nodes. After, we get pre-community with center node. Then check the similarity between pre-community nodes and the non pre-community nodes. Finally we get the set of community with center node. Here, benefit of this proposed algorithm is that we do not require to recalculations of center node for each community whereas in DCCD algorithm required more number of recalculations of center node for each community. For calculate the accuracy of proposed methodology, we did experiment on Zachary's karate club data set. Then we compared the results with DCCD algorithm and some other detection of community algorithms. The comparisons of result show that our proposed methodology obtains a greater modularity value and purity value is near to ground truth value. And we get communities of network with less prior information. In future, we will do experiment on large dataset using this proposed methodology.

7. REFERENCES

1) Longju Wu, Tian Bai, Zhe Wang, Limei Wang, Yu Hu, Jinchao Ji "A new community detection algorithm based on distance centrality"10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), IEEE, 2013

2) Bing Kong1, Lei Li1, Lihua Zhou1, Chongming Bao2. " A Hierarchical Agglomerative algorithm of Community Detecting in social network based on Enhanced Similarity " 2014 Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics, IEEE, 2014

3) Ahmed Ibrahem Hafez"', Abaul ella Hassanien2,., Aly A. Fahm/ and M.F.Talba3. "Community Detection in Social Networks by using Bayesian network and Expectation Maximization technique" 10th International Conference on Hybrid Intelligent System(HIS), IEEE, 2013

4) Kamal Sutaria, Dipesh Joshi, Dr.C.K.Bhensdadiya, Kruti Khalpada. "An Adaptive Approximation Algorithm for Community Detection in Social Network" 2015 International Conference on Computational Intelligence & Communication Technology, IEEE, 2015

5) Chang Su , Yukun Wang "A New Method for Community Detection Using Seed Nodes" International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM

6) Chuantao Yin,1,2 Shuaibing Zhu,1 Hui Chen,1 Bingxue Zhang,3 and Bertrand David3 "A Method for Community Detection of Complex Networks Based on Hierarchical Clustering" International Journal of Distributed Sensor Networks, 2015

7) Santo Fortunato, "Community detection in graphs," Physics Reports, vol. 486, no. 3-5, pp. 75-174,2010.

8) Michele Cosciaa,b,c, Fosca Giannottib, Dino Pedreschia,b "A Classification for Community Discovery Methods in Complex Networks " 2012

9) T. Opsahl, F. Agneessens and J. Skvoretz, "Node centrality in weighted networks: Generalizing degree and shortest paths," Social Networks, vol.32, pp. 245-251, 2010.

10) C.D. Manning, P. Raghavan and H. Schutze, "Introduction to Information Retrieval," Cambridge: Cambridge University Press, 1 edition, 2008.

11) M.E.J. Newman, M. Girvan. "Finding and evaluating community structure in networks," Physical review E, vol. 69, pp. 1-15, 2004.

12) W.W. Zachary, "An information flow model for conflict and fission in small groups," Journal of Anthropological Research, vol. 33, pp. 452-473,1977.

13) Flake, G. W., S. Lawrence, C. Lee Giles, and F. M. Coetzee, 2002, IEEE Computer 35, 66.

14) Dourisboure, Y., F. Geraci, and M. Pellegrini, 2007, inWWW'07: Proceedings of the 16th international conference on the World Wide Web (ACM, New York, NY, USA), pp. 461-470.

15) Chen, J., and B. Yuan, 2006, Bioinformatics 22(18), 2283.

16) Spirin, V., and L. A. Mirny, 2003, Proc. Natl. Acad. Sci. USA100(21), 12123.

17) Rives, A. W., and T. Galitski, 2003, Proc. Natl. Acad. Sci.USA 100(3), 1128.

18) Z. Xie and X.F. Wang, "An Overview of Algorithms for Analyzing Community Structure in Complex Networks," Complex Systems and Complexity Science, vol. 2, pp. 1-12, 2005.

19) R.R. Khorasgani, J. Chen, Osmar and R. Zaïane, "Top Leaders Community Detection Approach in Information Networks," in Proceedings of the 2010 International Conference on Knowledge Discovery and Data Mining (KDD'10), Washington DC: USA, pp. 1-9, 2010.