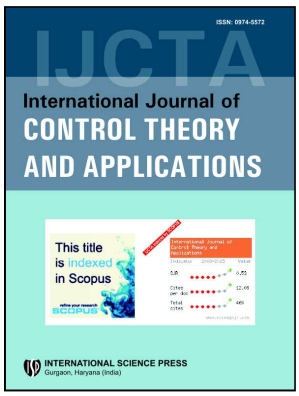

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 19 • 2017

Degree centrality based multi label propagation algorithm for social community detection

S. Rao Chintalapudi[1] and M.H.M. Krishna Prasad[1]

[1] *Department of CSE University College of Engineering Kakinada (A) JNTUK, Kakinada, India,*
Emails: srao.chintalapudi@gmail.com, krishnaprasad.mhm@jntucek.ac.in

Abstract: Detection of communities in social networks is useful for various real time applications such as recommendation systems and target marketing. Community detection using Label Propagation has a limitation that, sometimes it may produce a single community due to random selection of seed node. To overcome this limitation authors adopted the concept of degree centrality in selection of seed node. The proposed approach Degree Centrality based Multi Label propagation Algorithm (DCMLPA) is experimented on real-world networks and synthetic networks. And the performance of the algorithm is measured in terms of Modularity, ONMI, Omega index and F-Score. The result shows that DCMLPA produce the good number of quality communities over its baseline algorithms. Moreover, DCMLPA can detect both disjoint and overlapping communities in the network by varying post processing threshold.

Keywords: Link Analysis, Social Network Analysis, Graph Clustering, Community Detection, Degree Centrality, Multi-Label Propagation.

1. INTRODUCTION

A social network is a collection of users linked with different types of relationships. Analyzing such social networks [1, 2] plays a vital role in various areas such as telephone networks, co-authorship networks, biological networks, transportation networks, World Wide Web, citation networks and food webs. These networks consists latent information in the form of community structures. Community can be defined as a densely connected group of vertices with sparser connections to the other groups. A sample network and its two communities are depicted in Figure 1. Here, nodes with same color are belonged to one community. Identifying such structures in social networks gains much attention from industry and academia since it has many real time applications such as friend recommendation, movie recommendation, opinion mining, trend prediction, and target marketing. Detecting communities can be helpful to understand the properties of nodes from the network topology alone.

There has been several community detection algorithms are proposed in the last few years [1, 2, 3]. These algorithms are broadly falling into two categories: structure based and attribute based. Structure based community detection algorithms detects communities based on link analysis techniques whereas attribute based community detection algorithms uses node attributes and edge attributes for identifying communities. Here, this paper

focused on detecting communities based on topology alone that means clustering can be done without any node attributes and edge attributes into consideration. Communities in the networks are of two types namely, disjoint communities and overlapping communities. In disjoint communities, each node is exactly member of only one community where as nodes in the overlapping communities has more than one community membership. There are several disjoint community detection algorithms available such as Modularity Optimization [4], Label Propagation [5], Infomap [6], Edge Betweenness [7] and so on.

Overlapping communities are more realistic because a node can be participated in more than one community. For example, in the real world scenario a facebook user can be member of both family and friend communities at the same time. Based on this fact, many overlapping algorithms have been proposed over the years they include Clique Percolation [8], COPRA [9], SLPA [10], BIGCALM [11] and so on.

The size of the network datasets becomes larger in the recent years, so scalability and the time complexity are important issues. In future, algorithm should handle thousands of nodes in a reasonable time. So, there is a need to develop an algorithm for detecting both disjoint and overlapping communities with very low time complexity.

The structure of the remaining paper is as follows. Section II discusses label propagation based algorithms and their limitations and the concept of degree centrality. The proposed algorithm Degree Centrality based Multi Label Propagation Algorithm (DCMLPA) is discussed in section III. The experiments conducted and the datasets used are described in section IV. Conclusions and the possible future directions are discussed in the section V.

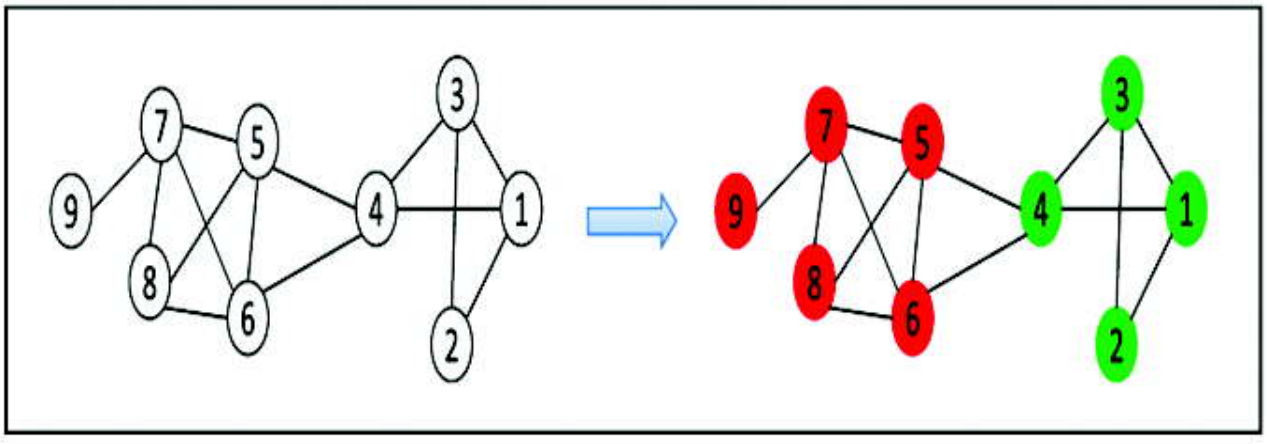

Figure 1: Sample network and its two communities (nodes with same color denotes one community)

2. RELATED WORK

Label propagation [5] is the one of the fast community detection algorithm that identifies disjoint communities in the network. In this approach each node of the network is associated with a label usually node id. The label of each node is updated by most frequent label in its neighborhood. If a neighbor has same number of different labels, then any one label can be selected at random. After a few iterations all the members that belong to same community have same label. Here, nodes are selected randomly to propagate labels and ties are broken randomly. Because of this random nature, algorithm produces different results at different runs and some of them are of poor quality. The advantages with this algorithm are simple mechanism, no need of input parameters and linear time complexity even for sparse networks. The limitations of this approach are harder to estimate iterations required for classifying all the nodes, non deterministic nature and it is used only for detecting disjoint communities.

Community Overlap Propagation Algorithm (COPRA) [6] is an extended version of label propagation that allows a node belongs to up to *v* communities and this can be provided as input parameter. Each node is associated with belonging coefficient for representing strength of community membership. Each propagation step updates node labels in conjunction of its neighbors and normalizes the sum of all belonging coefficients. It is also a non deterministic algorithm, so there is a scope to improve determinacy.

Speaker Listener Propagation Algorithm (SLPA) [7] works based on general speaker listener interaction rules. In this approach, each node is associated with to store labels received and the probability of a label in node's memory is treated as community membership strength. The advantage with this approach is that no need of prior knowledge about number of communities but one needs to provide number of iterations as a parameter of the algorithm. It is also non deterministic in nature because listeners are selected at random and ties in assigning community label are broken randomly.

Selection of seed node is a crucial step in the above algorithms. Hence, to select seed node in the algorithm this paper adopts the concept of degree centrality [12]. For an undirected network, it can be defined as the ratio of the degree of the node and maximum possible degree of nodes. The maximum possible degree is one less than the number of nodes in the network. The degree centrality of a node *DC(x)* can be computed using Eq. (1), where *x* is node and *n* is total number of nodes.

$$DC(x) = \frac{\deg(x)}{n-1} \tag{1}$$

The computation of the degree centrality is clearly depicted in Figure. 2 and Figure. 3. Figure. 2 is a sample network with seven nodes and Figure. 3 represent degree centrality of each node as node label. Consider Figure. 2, the degree of node 2 is 3 and maximum possible degree in the network with 7 nodes is 7 – 1 = 6, then the degree centrality of the node 2 is 3 / 6 (i.e. 0.50). Usually, nodes with higher degree centrality value acts as hubs, so labels can be easily propagated to the maximum extent in less number of iterations. Hence, authors incorporated degree centrality measure in the proposed approach.

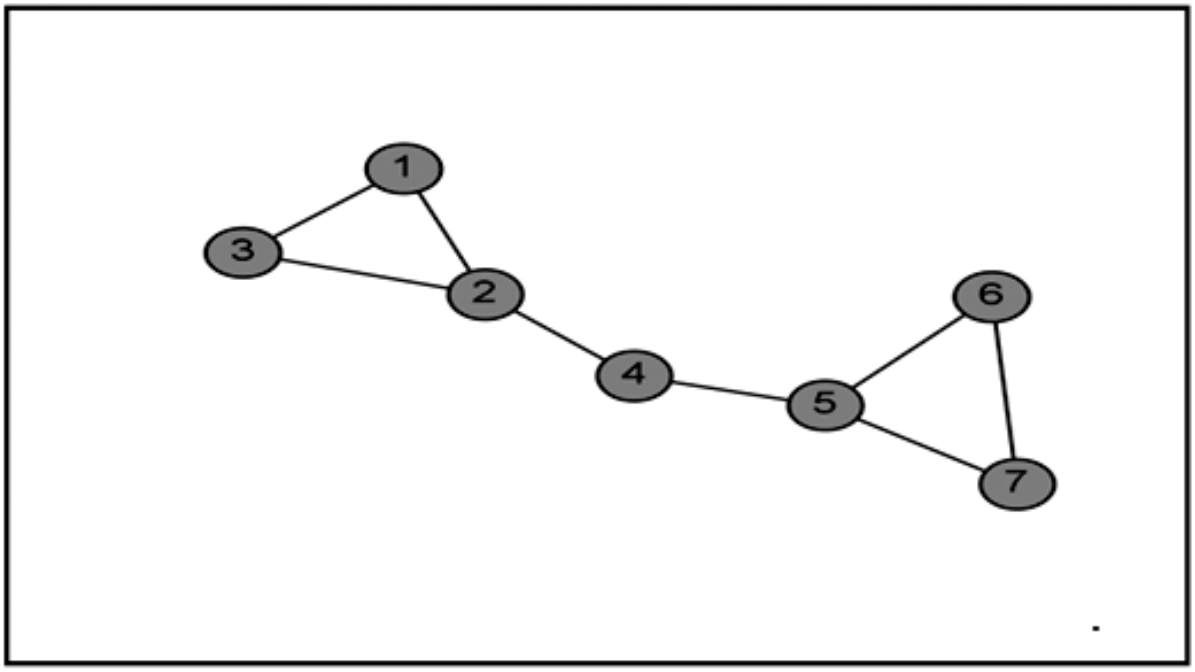

Figure 2: Sample Network with 7 nodes.

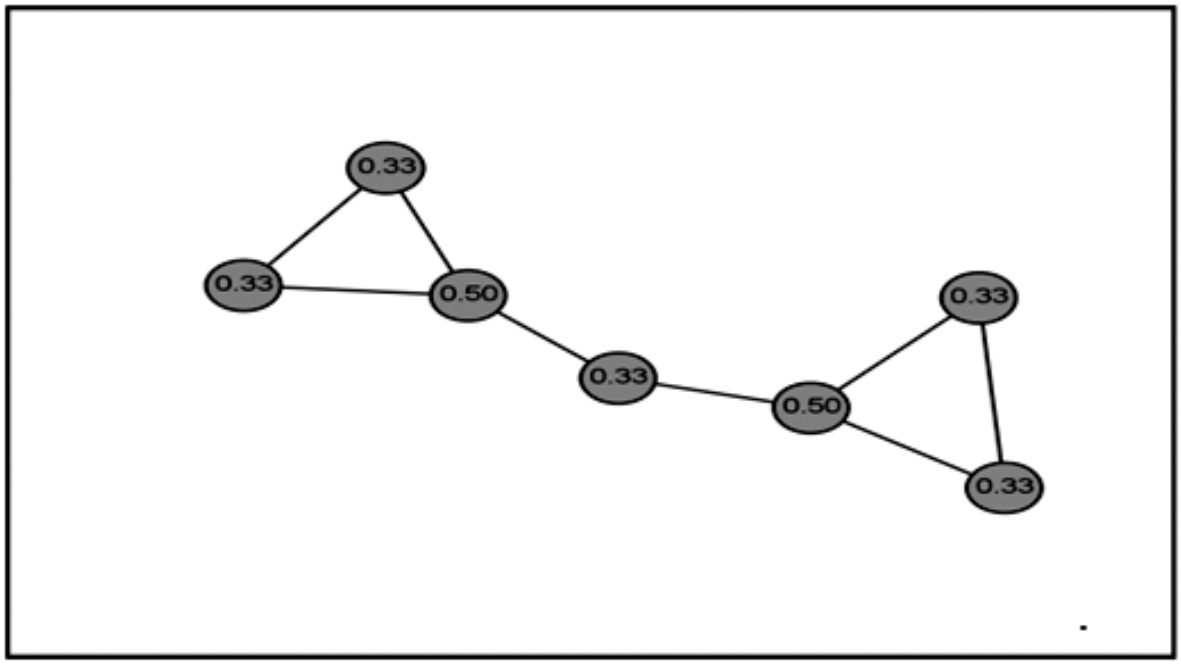

Figure 3: Sample Network with corresponding degree centrality value as node label.

3. DEGREE CENTRALITY BASED MULTI LABEL PROPAGATION ALGORITHM (DCMLPA)

DCMLPA is extended version of Label propagation algorithm that detect both disjoint and overlapped communities. In this algorithm, each node maintains multiple labels in its memory and current decisions can be done based on past labels in its memory. And also Degree Centrality measure is incorporated in the selection of seed node instead of random selection. The major steps involved are listed in the following algorithm.

Algorithm: DCMLPA

Input: Network G= (V, E)

V: node set, E: edge set, N: maximum number of iterations

T: Post-Processing threshold (< 0.5 for overlapping communities)

Output: Communities (Community id, <node id list>)

1 : Read input network from a file.

2 : Find neighbors of each node and Compute degree centrality of each node.

3 : Sort nodes according to descending order of degree centrality.

4 : Select top node as a seed node and point to next node for next iteration.

5 : Receive label from each neighbor of seed node and find most popular label for the node.

6 : Add popular label to its memory and increment number of communities.

7 : Repeat step 4 to 6 until all nodes are considered.

8 : Post process the result based on threshold T.

9 : return disjoint communities if T e" 0.5

else return overlapping communities.

In the above algorithm, each node is considered as one community and produce desired communities at the end based on post processing threshold T. The advantage with this algorithm is that no need to give number of communities as input. Each time step 6 of this algorithm increases the size of the node memory by one for each node. Even though, this algorithm is non-deterministic due to randomness in ties, produce a stable output for N = 20.

Post processing of the result can be done in step 8 based on threshold value T. The range of the threshold is [0, 1] and the experiments revealed that the algorithm produce disjoint communities when $T \geq 0.5$. Initially, the memory of each node will be converted into the probability distribution of labels. The labels with lower probability distribution values are deleted i.e. [0, 0.5]. After that, community can be formed using the nodes with the same label. The nodes that have more than one label are overlapping nodes. The ties happened in forming communities are broken randomly, so it will produce different communities in different runs for the same dataset. Lower values of threshold may result more number of communities with overlapping nodes.

4. EXPERIMENTAL RESULTS

4.1. Datasets

The proposed approach is tested on seven synthetic networks and three real-world networks. The synthetic network generator proposed by Lancichinetti-Fortunato-Radicchi (LFR) [13] produces benchmark networks with the properties of real world networks such as heterogeneous community size and degree distribution. The parameters of the network generator are: number of nodes (N=1000), average degree ($\bar{k} = 10$), maximum degree (maxk=30), mixing parameter ($\mu = 0.1 - 0.7$), exponents of the power law distribution of vertex degrees (t_1=2, t_2=1), minimum size of the community (minc=10), maximum size of the community (maxc=50), number of memberships of the overlapping nodes (O_m=2), total number of overlapping nodes in the network (O_n=100). This network generator can also allow overlapping communities using the parameters O_m and O_n. Seven benchmark networks are generated by varying the value of μ from 0.1 to 0.7 with the above fixed parameters.

Three real large datasets have been used to test the performance and scalability of the DCMLPA namely, Amazon, DBLP and YouTube. These datasets are acquired from Stanford Large Network Dataset Collection http://snap.stanford.edu/data/. Amazon dataset represents co-purchasing network that consists of co-purchasing relationship between products. Here, nodes represent products and edge represents connection between co-purchased products and each product category is considered as ground truth community. DBLP is a co-authorship network in which each node denotes author and the edge denotes co-authorship relationship that is those two authors have a publication together. Here, publication venue is considered as ground truth community. Youtube is a social network where node represents users and edge represents friendship between users and users can create channels where other users can join. These user defined channels/groups are considered as ground truth communities. Table 1 shows the summary of real datasets used in the experiments.

Table 1
Summary of real-world network datasets

Dataset	*Nodes*	*Edges*	*Ground Truth*
Amazon	334863	925872	75149
DBLP	317080	1049866	13477
YouTube	1134890	2987624	8385

4.2. Result Analysis

The experiments done in this paper are performed on a system with 2.10 GHz CPU, 32 GB RAM. The framework for the experiments is developed in java and the result analysis is done on R for computing modularity, Overlapping Normalized Information (ONMI), Omega index and F-Score.

In these experiments, DCMLPA algorithm is implemented and compared the results with two overlapping community detection algorithms (i.e, COPRA, SLPA). Initially, detected communities generated from these three algorithms are compared with ground truth communities. The results are shown in Table 2 and these experiments are conducted on three real-world networks namely Amazon, DBLP, YouTube. These results shows that proposed algorithm can detect more number of communities over COPRA and SLPA. After that, the proposed algorithm is evaluated using the popular measures Modularity, ONMI, Omega index and F-score. These measures are calculated for three real and seven synthetic datasets and are shown in Table 3. Here, the naming convention for synthetic networks for example, LFR_0.1 represents these networks are generated from LFR benchmark network generator and 0.1 represents mixing parameter (ì) used. Among all the networks in Table.3, first three networks shows lower values for all measures than the remaining seven networks because real-world networks have less modular structures than the synthetic networks. Proposed algorithm detects good modular structures from DBLP and the quality of communities is very good for YouTube network. The performance of the proposed approach in terms of modularity is shown in Figure.4 and it shows that DCMLPA can detect good modular structure in the real-world network than the baseline algorithms. After analyzing the measures of synthetic networks it is observed that all the measures are degraded with increase of mixing parameter (ì) and the reason behind this is that with the increment of mixing parameter (ì) the network lost its modular structure or community property. The behavior of the proposed algorithm and its baseline algorithms with respect to ONMI is shown in Figure.5 and it shows proposed algorithm will produce good quality communities over baseline algorithms.

Table 2
Number of communities detected using COPRA, SLPA and DCMLPA

Dataset	*COPRA*	*SLPA*	*DCMLPA*
Amazon	69146	70128	72585
DBLP	10181	10367	11256
YouTube	7849	7993	8023

Table 3
Evaluation Metrics for DCMLPA (* represents real-world networks)

Dataset	*Modularity*	*ONMI*	*Omega Index*	*F-Score*
Amazon*	0.363	0.746	0.694	0.765
DBLP*	**0.418**	0.667	0.632	0.653
Youtube*	0.392	**0.786**	**0.794**	**0.783**
LFR_0.1	**0.753**	**0.824**	**0.893**	**0.863**
LFR_0.2	0.647	0.739	0.852	0.834
LFR_0.3	0.582	0.717	0.806	0.813
LFR_0.4	0.554	0.603	0.783	0.757
LFR_0.5	0.528	0.564	0.694	0.721
LFR_0.6	0.483	0.458	0.613	0.658
LFR_0.7	0.412	0.409	0.534	0.613

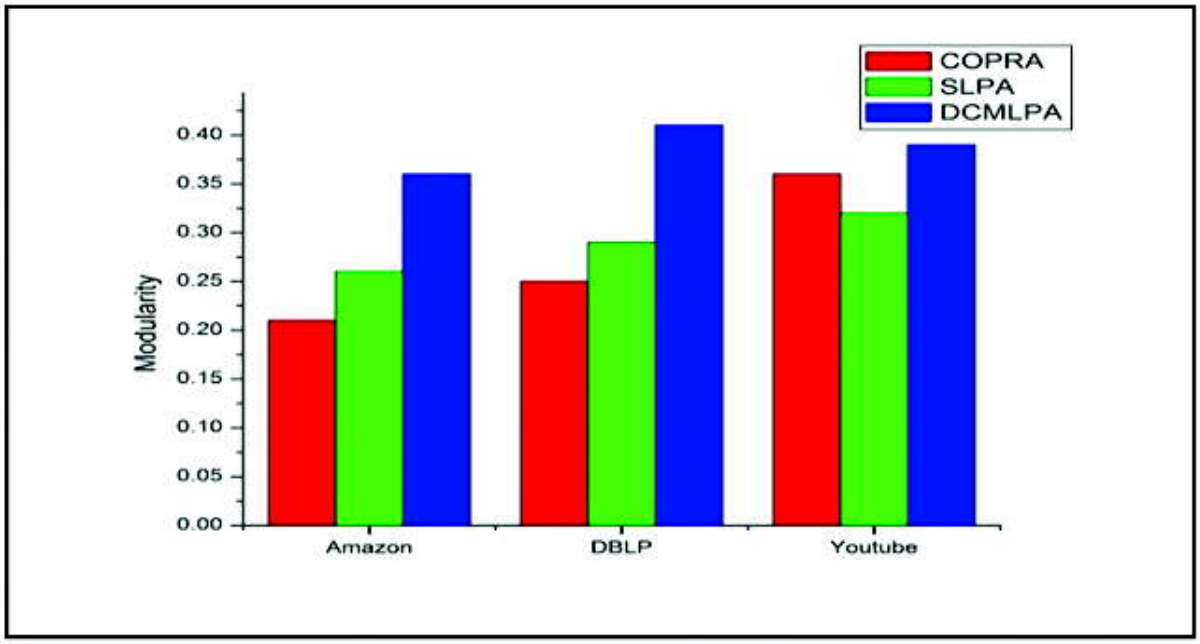

Figure 4: Comparison of DCMLPA with COPRA and SLPA with respect to modularity over three real networks namely Amazon, DBLP and YouTube.

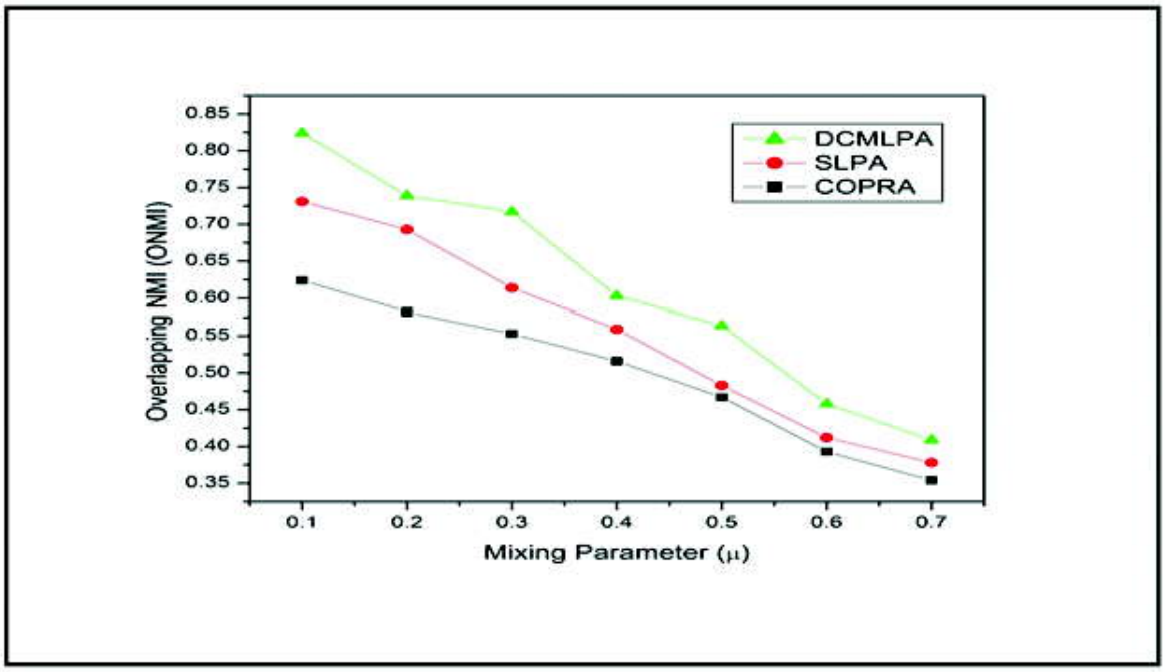

Figure 5: Performance of DCMLPA, SLPA and COPRA on LFR Benchmark networks with respect to ONMI.

5. CONCLUSION AND FUTURE SCOPE

In this paper, label propagation based community detection algorithm is proposed called Degree Centrality based Multi Label Propagation Algorithm (DCMLPA). The algorithm adopts multiple labels for detecting overlapping communities and the concept of degree centrality to reduce the number of iterations. The experiments revealed that DCMLPA produce both disjoint and overlapping communities based on post processing threshold

and a stable output in 20 iterations. The scalability of the algorithm is tested using large real network datasets such as Amazon, DBLP, Youtube and detects good number of communities over its baseline algorithms. And also DCMLPA is evaluated using Modularity, ONMI, Omega Index and F-Score on both real and benchmark networks. The results demonstrated that DCMLPA produce quality communities. This algorithm will work for undirected networks only; hence there is a scope for research to extend it for directed networks and bipartite networks in future.

REFERENCES

[1] S. Harenberg, G. Bello, L. Gjeltema, S. Ranshous, H. Jitendra, R. Seay, K. Padmanbhan and N. Samatova, "Community detection in large-scale networks: a survey and empirical evaluation," WIREs Comp. Stat. vol. 6, pp. 426-439, 2014.

[2] J. Xie, S. Kelley and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," ACM Comput. Surv. vol. 45, pp. 1-35, 2013.

[3] Ch. S Rao and M. H. M. Krishna Prasad, "A survey on community detection algorithms in large scale real world networks," 2nd International Conference on Computing for Sustainable Global Development (INDIACom), pp.1323-1327, IEEE, New Delhi, 2015.

[4] M. E. J. Newman," Fast algorithm for detecting community structure in networks", Phys. Rev. E. vol.69,pp. 066133,2004.

[5] U. N. Raghavan, R. Albert and S. Kumara, "Near linear time algorithm to detect community structures in large scale networks," Phys. Rev. E. vol.76, pp.036106, 2007.

[6] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," Proc. Natl. Acad. Sci. vol.105,pp.1118-1123, 2008.

[7] M. Girvan, M. and M. E. J. Newman, " Community structure in social and biological networks," Proceedings of the National Academy of Sciences of the United States of America, Vol. 99, No. 12, pp. 7821–7826, 2002.

[8] G. Palla, I.Derenyi, I. Farkas and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," Nature. vol.435,pp.814-818,2005.

[9] S. Gregory, "Finding overlapping communities in networks by label propagation," New J. Phys. vol.12,pp.103018, 2010.

[10] J. Xie, B. K. Szymanski, X. Liu, "SLPA: Uncovering overlapping communities in social networks via a speaker listener interaction dynamic process," Eleventh International Conference on Data Mining Workshops (ICDMW), pp. 344-349, IEEE, Vancouver, 2011.

[11] J. Yang and J. Leskovec, " Overlapping community detection at scale: A nonnegative matrix factorization approach," Sixth ACM International Conference on Web search and Data Mining, pp. 587-596,ACM, Rome, Italy, 2013.

[12] Charu C, Aggarwal. Data Mining: The Textbook, 1st ed., Springer International Publishing, Switzerland, 2015.

[13] A. Lancichinetti, S. Fortunato and F. Radicchi, "Benchmark graphs for testing community detection algorithms," Phys. Rev. E., vol.78, pp.046110, 2008.