ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

Detecting Overlapping Communities in Dynamic Networks Using Label Propagation

T. Sreeshanmugapriya[1], Dr.G.T.Prabavathi[2]

Research Scholar, Department of Computer Science, Gobi Arts & Science College, Gobichettipalayam, TamilNadu, India[1]

Associate Professor, Department of Computer Science, Gobi Arts & Science College, Gobichettipalayam, TamilNadu, India[2]

ABSTRACT: Social network analysis aims to analyze the relationship between the members of social group based on similar interests. A community is commonly considered as a social unit that shares something in common. Discovering the communities is an interesting area for studying the large-scale structure of networks. Many algorithms have been proposed for community detection and benchmarks have been created to evaluate their performance. There exist various disjoint and overlapping community detection algorithms in both static and dynamic networks. Label propagation algorithms (LPA) propagate labels throughout the network and can identify both overlapping and non overlapping communities in static and dynamic networks. In this paper, a multi-label propagation technique has been proposed which uses labels at different time stamps to identify overlapping communities in dynamic networks. The algorithm has been tested with synthetic networks and the results show that propagating multiple labels provides better results.

KEYWORDS: Social Networks, Dynamic networks, overlapping communities, label propagation

I. INTRODUCTION

A complex network is a graph-based representation of the interactions amongst entities that take place in the real world. Examples include social networks such as acquaintance networks, collaboration networks, technological networks such as the Internet and the World Wide Web, and biological networks such as neural networks, and metabolic networks [2]. Real networks are not random and they usually exhibit in homogeneity, indicating the coexistence of order and organization. Network scientists call this organization as the community structure of networks. Though there is a lack of consensus in the definition of communities, most popular and well-accepted definition suggests that: communities are the subsets of vertices within which vertex-vertex connections are dense, but between which connections are less dense [1]. Analysis of such communities is essential to understand the structural and the functional organizations of the network.

Overlap is one of the characteristics of complex networks, in which a node may belong to more than one group. In particular, communities in social networks overlap with each other since many active users can possibly participate in multiple groups at the same time. In a dynamic network a node can be removed from the network or a node can newly join a network. The addition or deletion of nodes and links at different time stamps form a dynamic network. Since each and every minute new nodes and links are formed and deleted in online social networks, detecting communities in dynamic network is an emerging area of interest.

Many algorithms require prior information about the size of communities which is often not predictable beforehand in many complex networks. Particularly, in social networks where the size of the network is huge and heterogeneous in nature, it is impossible to predict about the number of communities initially. Moreover, it is computationally expensive to determine the community size. Hence, algorithms that are able to detect communities from complex networks without having any knowledge about the number of communities or the size of communities

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

are essential. Label Propagation Algorithms [7] are such type of algorithms that detects community structure that uses network structure alone as its guide and requires neither a pre-defined objective function nor prior information about the communities.

In this paper, a review on algorithms that detects overlapping community in dynamic network using label propagation algorithm has been made in Section 2. Section 3 highlights the importance of proposed multi label propagation algorithm in dynamic networks and Section 4 discusses the results of the proposed algorithm in a small synthetic dynamic network. In Section 5 conclusion and suggestions are given.

II. LITERATURE REVIEW

There is a need for algorithms that can incrementally update and monitor communities whose evolution generates huge real-time data streams, such as the internet or on-line social networks. Xie et al. modified LabelRank [10] for dynamic networks called as LabelRankT [9] which is an online distributed algorithm for detection of communities in large-scale dynamic networks through stabilized label propagation. Results of tests on real-world networks demonstrate that LabelRankT has much lower computational costs than other algorithms. It also improves the quality of the detected communities compared to dynamic detection methods and matches the quality achieved by static detection approaches. Unlike most of other algorithms which apply only to binary networks, LabelRankT works on weighted and directed networks, which provides a flexible and promising solution for real-world applications.

Multi-label propagation algorithm [MLPA] [6] detects overlapping communities using more than one label during the community detection process. The focus of this algorithm is based on the decision of choosing multiple labels for propagation and storing multiple labels received from the propagation process. There are three important phases in this approach sending labels to other nodes, listening labels from other nodes for deciding which label to choose from the set of labels and receiving labels from other nodes and updating the nodes. All the three phases relies on the propagation process in the network. First label propagation constraints were formulated. The nodes are categorized as speaker nodes and listener nodes and the constraints are imposed on the two categories. The algorithm works on each node in the network as both speaker and listener. When a specific node starts gathering the labels from the adjacent nodes, it acts as an information consumer. Whereas, when it sends labels to the adjacent nodes it acts as an information provider. Until the listener asks for the labels from speaker, no labels are propagated from node to node. In this work, each node broad casts multiple labels to the neighbour nodes and at the same time receive multiple labels from its neighbours. The algorithm works only for the static networks and the time complexity is ***O(Tm).***

Nathan et al. [3] improved SCAN (Structural Clustering Algorithm for Networks) [8] by allowing it to form communities without the user-defined threshold for updating a dynamic network of timestamps. In DSCAN (Dynamic Scan**)**, SCAN is performed on the first timestamp, for the next consecutive timestamps, the difference in edges between the two timestamps are obtained. When there are edge changes in the network, it is updated. During updation of edges a node can become a core, or a node can no longer be a core node. An existing cluster id or a new cluster id is propagated through all structurally connected nodes to form a new community. Through propagation, a node can either become a hub or an outlier. DSCAN performs better then SCAN by removing the need for parameter tuning.

Nathan et al. modified DSCAN to detect overlapping communities using an algorithm named DSCANO [4]. First, it detects communities using DSCAN, then; it checks every node to extract their maximum similarity of all incident edges using a threshold value. The overlapping threshold is user specified similarity value. Any node whose edge is incident to another community with a similarity greater than the node's maximum similarity minus the similarity decrease becomes part of the other community. It shows the overlap between the two communities.

The authors [4] proposed Genetic Algorithm for dynamic networks namely Genetic Algorithm Dynamic (GAD). In GAD the genes repressent all individual nodes and their corresponding community. To find a gene that satisfies a good community structure of a network, the gene must maximize a given fitness function, such as modularity. A user-defined number of iteration is performed over the population which mutates each gene into a cross between two or more genes that produce the highest values due to the fitness function. At the end of the iteration, the gene with the highest fitness function value is used as the final community structure. GAD uses the previously found

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

"best" gene of the previous timestamp as a gene in the population of the next. This results in the next timestamp starting with a gene in the population with a relatively good fitness.

A brief literature survey on overlapping community detection has been stated in [5]. Next chapter deals with the multi-label propagation algorithm for dynamic networks.

III. PROPOSED METHOD

The proposed algorithm uses speaker, listener node concept same as that of MLPA [6] with the extension for dynamic networks. The modified LPA can be viewed as an extension of MLPA with an extra conditional update rule by which nodes that make edge changes from one timestamp to another timestamp alone accept the new distribution.

Updates of the labels between two time stamps is made only when (i) an existing node adds/deletes link to/from another node (ii) a new node is joining the network and (iii) an existing node is leaving a network. Communities split, merge and dissolve and those changes are stored in the modified label propagation algorithm using probability values. Initially the probability calculation is done for the Graph at Time T_0. When, edges are formed and dissolved, the label probabilities are calculated using the same preprocessing steps, inflation operator, cutoff operator and conditional update operator which is used in MLPA. The steps for the dynamic MLPA is given below:

1. Input: Graph G at different timestamps G[0,1,2,3,….N]
2. For n=1:N do
   - *(i)* Compare the edge changes(addition/removal) between *G(n-1) and G(n)*
   - (ii) If Node i changes edges between *G(n-1)* and *G(n)*, reinitialize the label distribution as MLPA
   - (iii) If there is no edge changes in node between *G(n-1)* and *G(n)*, keep the label probabilities as it is
3. Update only changed nodes labels and assign it to the communities C1,C2,C3…Cn respectively

The modified community detection algorithms quantify the strength of association between all pairs of nodes and communities. In this algorithm, a soft membership vector, or belonging factor is calculated for each node.

IV. RESULTS AND DISCUSSIONS

The modified label propagation algorithm was tested using synthetic sample networks and real-world social networks. The results of the small synthetic network are given below.

A sample graph containing 12 nodes and 15 edges is taken initially at Timestamp T_0 and modified for thirty time stamps (different intervals). Figure 1 and Table 1 shows the example network graph *G(0)* with number of nodes *n* and label distribution matrix of the network *G(0)*. Colors represent three communities indicated by Red, Blue, Green and nodes (4, 5) (yellow) indicates overlapping nodes.

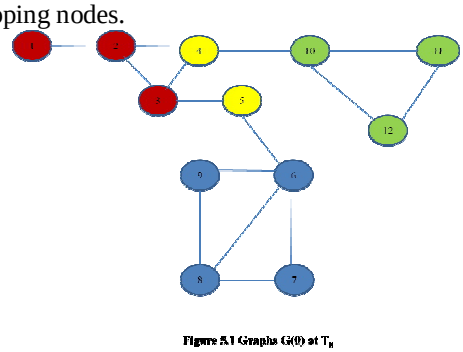

Figure 5.1 Graphs G(0) at T_0

Figure 1 Label Distribution probabilities at time T_0

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

	1	2	3	4	5	6	7	8	9	10	11	12
1	0	1	0	0	0	0	0	0	0	0	0	0
2	.33	0	0.33	0.33	0	0	0	0	0	0	0	0
3	0	.33	0	0.33	0.33	0	0	0	0	0	0	0
4	0	.33	0.33	0	0	0	0	0	0	0.33	0	0
5	0	0	0.5	0	0	0.5	0	0	0	0	0	0
6	0	0	0	0	0.33	0	0.33	0	0.33	0	0	0
7	0	0	0	0	0	0.5	0	0.5	0	0	0	0
8	0	0	0	0	0	0.33	0.33	0	0.33	0	0	0
9	0	0	0	0	0	0.5	0	0.5	0	0	0	0
10	0	0	0	0.33	0	0	0	0	0	0	0.33	0.33
11	0	0	0	0	0	0	0	0	0	0.5	0	0.5
12	0	0	0	0	0	0	0	0	0	0.5	0.5	0

Table 1 Initial Label Distribution probabilities at time T_0

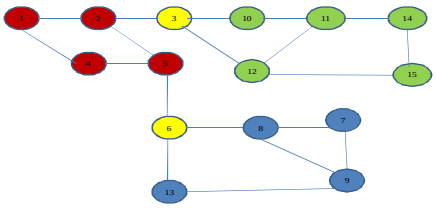

Figure 1 Label Distribution probabilities at time T_1

Figure 2 and shows the example network G(1) with node n=15 at Timestamp T_1. Colors represent communities at next time stamp. At time stamp T_1, edge from 4 to 10; 4 to3; 6 to 9; 5 to 6 is removed and new edge from 4 to 1; 6 to 8; 6 to 13; 9 to 13; 12 to 15; 11 to 14; 3 to10 is formed. At T_1, same three communities exists but the overlapping nodes changes from 4 to 3 and 5 to 6. This formation of change is identified by the modified label propagation algorithm using edge change calculations and the probability values.

Similarly the synthetically generated graph changes were done for 30 different timestamps. After few iterations, the probabilities that have smaller value begin to drop and the probabilities that have higher value becomes more strong if the node continues to exists as overlapping node between two or more communities. A threshold value of 0.1 to 0.5 is applied after each iteration to drop the labels that are week. After the post-processing phase, if there exists two labels with same probabilities, then that node forms the overlap between two communities. The communities were formed at the end of the seventh iteration and proved to stabilize after that.

Though there are numerous small and large real-world networks, the networks listed in Table 2 are widely used by researchers for checking the community detection algorithms due to its known community structure. The commonly used overlapping quality metric has been used to quantify the quality of the output produced by the developed algorithm.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

Networks	Nodes	Edges
AS-Internet Routers Graph	6477	7488
arXivHEP-TH	27770	352285
Enron	36692	367662

Table 2 Real World Social Networks

For the above specified networks, 90% of the nodes were correctly classified within 7 to 15 iterations. Modularity values for the above mentioned graphs were 0.62,0.65, 0.64 which shows the algorithms provides better results.

V. CONCLUSION

The study of complex networks is an active area of research and various researches has revealed number of properties about the behavior and topology of naturally occurring network. One of the interesting and important properties of complex network is community structure. Since, many real world networks commonly depict community structures; much research effort has been devoted to develop algorithms that efficiently highlight the hidden community structure of a network. A study and analysis on various community detection approaches in complex network for detecting overlapping communities in social network are studied and a method has been proposed to identify overlaps in dynamic networks using multiple label propagation. The developed algorithm has been tested with synthetic networks and real-world networks and evaluated with quality measures. The results show that the developed algorithm is capable of producing consistently good overlap between communities in the tested small networks.

In this research work, the modified label propagation algorithm has been tested with small synthetic network and few real-world social networks. To understand the modularity values properly, the algorithm should be tested with very large synthetic networks such as LFR benchmarks. The time complexity of the algorithm can be still reduced by using hub nodes.

REFERENCES

1. Girvan M., Newman M. E. J., “Community structure in social and biological networks”, Proc. National Academy of Sciences of the United States of America, Vol. 99, no. 12, pp. 7821-7826, 2002.
2. Li XJ., Zhang P., and Di ZR., and Fan Y., “Community structure of complex networks”, Complex systems and complexity science, Vol. 5, no. 3, pp. 19-42, 2008.
3. Nathan Aston., Jacob Hertzler., and Wei Hu., “Overlapping Community Detection in Dynamic Networks (DSCAN)”, Vol. 7, pp. 872- 882, 2014.
4. Nathan Aston., Wei Hu., “Community Detection in Dynamic Social Networks”, Vol. 6, pp. 124-136, 2014.
5. Prabavathi GT., SreeShanmugapriya T., “A Review on Community Detection in Dynamic Social Networks”, International Journal of Innovative Research in Computer Science and Communication Engineering, Vol. 4, issue. 7, pp. 13390-13393, 2016.
6. Prabavathi GT., Thiagarasu V., “Design and development of overlapping community detection algorithm using Multi-Label Propagation”, International Journal of Advance Research in Computer Science and Management Studies, pp. 19-59, 2014.
7. Raghavan U.N., Albert R., and Kumara S., “Near linear time algorithm to detect community structures in large-scale networks”, Physics Review E, Vol. 76(3), pp. 036106, 2007.
8. Xiaowei Xu., Nurcan Yuruk., and Zhidan Feng., and Thomas A. J. Schweiger., “SCAN: A Structural Clustering Algorithm for Networks”, In Proceeding of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining ACM, pp. 824-833, 2007.
9. Xie J., Chen M., and Szymanski BK., “LabelRankT: Incremental community detection in dynamic networks via label propagation”, arXiv: 1305.2006 v2 [cs.SI], 2013.
10. Xie J., Szymanski B.K., “LabelRank: A stabilized label propagation algorithm for community detection in networks”, IEEE Network Science Workshop, pp. 138–143, 2013.