AN IMPROVED LABEL PROPAGATION ALGORITHM WITH EXTENSIONAL LOCAL CLUSTERING COEFFICIENT FOR COMMUNITY DETECTION IN COMPLEX NETWORKS

Lei Fang[1,2], Qun Yang[1,2] and Xiangmao Chang[1,2]

[1]College of Computer Science and Technology
[2]Collaborative Innovation Center for Novel Software and Industrialization
Nanjing University of Aeronautics and Astronautics
No. 29, Jiangjun Avenue, Jiangning District, Nanjing 211106, P. R. China
Qun.Yang@nuaa.edu.cn

Received February 2016; accepted May 2016

Abstract. *The time complexity of label propagation algorithm (LPA) is lower than most of the existing algorithms for community detection. However, there are still some drawbacks in the algorithm. Therefore, we proposed the extensional local clustering coefficient to measure an edge's ability to transfer labels and used the measure to improve the label propagation algorithm. The novel algorithm was tested on several real-world complex networks. Experimental results show that the optimized algorithm can improve the robustness and effectiveness for community detection.*

Keywords: Community detection, Extensional local clustering coefficient, Label propagation algorithm, Complex networks

1. **Introduction.** Complex networks contain some statistical properties, such as "small-world" [1], scale-free properties [2] and community structure. Among them, the community structure gains more focus due to its significant effect and widespread use. Many approaches, such as user interaction based algorithm [3], clustering algorithm [4,5], information cascade-based model [6], and semantic network-based algorithm [7] have been proposed to tackle the community detection problem. Usually, a community is a set of nodes closely connected to each other but sparsely connected to those nodes which are not in the community [8,9]. The nodes in the different communities usually form different functional modules. For instance, communities in a citation network are sets of relevant papers on the same topic [10].

The time complexity of most existing algorithms for community discovery is higher than expectation. Some algorithms even require prior parameters, such as the total number of communities or the rough size of communities. Such factors can have a negative impact on the application of the algorithm. In view of these shortcomings, Raghavan et al. proposed the label propagation algorithm to detect communities [11]. The algorithm uses both the structural characteristics of the network and the propagation characteristics at the same time. In addition, it does not need to specify the number of communities in networks or design the objective function beforehand. More importantly, it is easy to implement and costs relatively less time, which makes it possible to be applied to large-scale networks [12].

Nonetheless, compared to the previous algorithms, LPA does not perform well in terms of both accuracy and stability, because of its randomness in the initial label allocation phase and the following step of updating labels. It exploits the propagation character of the network instead of many structural properties, which causes the label propagation process out of control and the randomness of the algorithm becomes obvious.

Many research works have paid attention to LPA. Algorithms, including multivariate graph-based method [13], balanced label propagation [14] and parallel SLPA [15], have been proposed. Xie and Szymanski proposed the neighborhood strength driven label propagation algorithm [16]. It modifies the label update rule of LPA but additional parameters, which are often difficult to determine, are introduced. Lin et al. proposed the label propagation algorithm with community kernel (CKLPA) [17]. The algorithm is more stable than the original LPA, but the community kernel is required to be detected first and the size of the community kernel is required to be selected manually. Therefore, based on the structural characteristics of the network, the extensional local clustering coefficient is presented in this paper. It can be used to control the propagation phase and help the labels cluster more easily. Besides, it can effectively reduce the randomness in the propagation of the labels. The calculation process is consistent with the definition of the community, which makes the results much closer to the definition of the community.

The rest of the paper is organized as follows. The original label propagation algorithm, the local clustering coefficient and the Jaccard similarity are introduced in Section 2. The label propagation algorithm with extensional local clustering coefficient (ELCLPA) is described in Section 3. And the experiments for community detection are described in Section 4. Finally, conclusions are given in Section 5.

2. Problem Statement and Preliminaries.

$G(V, E)$ is an unweighted and undirected graph which can represent the complex network. V is the set of all the nodes in graph G and E is the set of edges connected between the nodes in V. e_{ij} stands for the edge that connects the node v_i with the node v_j.

$L(v)$ is the label of the node v ($v \in V$) and $N(v)$ is the set of the node v's neighbors. The label of each node is only influenced by its neighbors. After several iterations, the close-connected nodes receive the same label. And the nodes with the same label are clustered in a community. The formula of $L(v)$ is shown below [12]:

$$L(v) = \arg\max_{l} \left|N^l(v)\right| \tag{1}$$

$N^l(v)$ is the set of all the nodes, which have the same label l, in the neighborhood of v. $L(v)$ is the label that most of the node v's neighbors obtain currently. Below are the key steps of the original label propagation algorithm.

- First, each node is labeled with a unique label (an integer). The label represents the community that it belongs to.
- Then, it updates all nodes' labels in iterations. For the node v, the label will be updated by Equation (1). It means that the new label is the largest one among the labels of all the neighbors of the node v. However, if there are several candidate labels, the node v will select one label randomly. Repeat this step until all nodes' labels do not change.
- Finally, the nodes with the same label are clustered into the same community.

Watts and Strogatz introduced the local clustering coefficient [1] which can quantify how close a node's neighbors are to be a clique (complete graph). The local clustering coefficient determines whether a graph is a small-world network.

The neighborhood of the node i is defined as below:

$$N_i = \{v_j | e_{ij} \in E\} \tag{2}$$

$|N_i|$ is the radix of the set N_i and its value is the number of the nodes in the neighborhood of the node i, too. e_{ij} and e_{ji} are considered identical in an undirected graph. Therefore, if the node v_i has $|N_i|$ neighbors, $\frac{|N_i| * (|N_i| - 1)}{2}$ edges could exist among its neighbors at most.

The local clustering coefficient C_i of the node v_i is given by the below formula [1].

$$C_i = \frac{2|\{e_{jk}|v_j, v_k \in N_i, e_{jk} \in E\}|}{|N_i| \times (|N_i| - 1)} \tag{3}$$

According to the above Equation (3), if there is a complete graph constituted by the neighbors of node v_i, the local clustering coefficient C_i of the node v_i would be 1. The subgraph, which consists of the node v_i and its neighbors, is also a complete graph. If a graph is close to the complete graph, the value of its Modularity Q would be close to 1. Thus, the local clustering coefficient represents the important structural information of a node. However, it is only a measurement for one node and cannot be combined with the label propagation. We need a measure function which can evaluate the edge's ability to transfer labels.

For an undirected graph, the number of common neighbors of two nodes can indicate the degree of closeness between them. The more the common neighbors of the two nodes are, the higher the possibility that they belong to the same community is. In LPA, if there are many nodes which are both connected with the node v_i and v_j and there is an edge e_{ij} between v_i and v_j, it can be regarded that the labels would transfer more easily between them. In most cases, the Jaccard similarity is used to calculate the similarity of two sets. Here we use the following Formula (4) to calculate the similarity between the two nodes v_i and v_j:

$$S_{jaccard} = \frac{|N_i \cap N_j|}{|N_i \cup N_j|} \tag{4}$$

where $N_i \cap N_j$ shows the common neighbors of v_i and v_j. And $N_i \cup N_j$ are the total neighbors of these two nodes.

The Jaccard similarity is only influenced by the number of the common neighbors. If the two nodes v_i and v_j have the same neighbors, their $S_{jaccard}$ would be one. However, it cannot indicate the topological information among these neighbors. Whether these neighbors are connected completely or there are not any edges between any two nodes of them, the $S_{jaccard}$ is only related to the number of them.

In view of this problem, this paper extends the similarity parameter with the local clustering coefficient to show the similarity and joint degree between two nodes at the same time.

3. **Label Propagation Algorithm with Extensional Local Clustering Coefficient.** In LPA, as all labels spread on edges, in this paper we focus on the coefficient of two nodes of an edge. The local clustering coefficient is a concept based on the single node. And it quantifies the closeness of its neighbors, which determines whether they

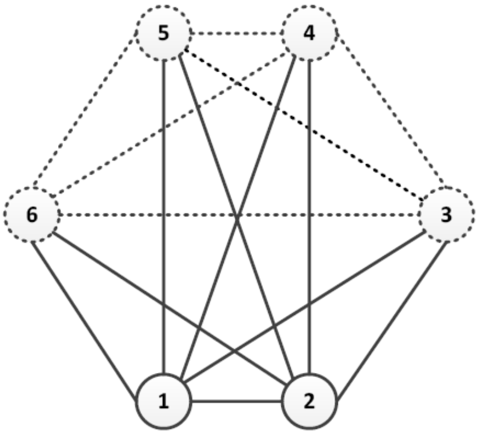

Figure 1. An example for extensional local clustering coefficient

are to be a clique (complete graph). We need a measure to quantify the closeness of the common neighbors which belong to the two nodes in the same edge. If the two nodes connect to all the nodes in a clique (complete graph) and these two nodes are connected, the edge is in the clique. For example, in Figure 1 above, the node 1 and the node 2 are connected by a solid line and the dashed part is a complete sub-graph in the whole graph. When the node 1 and the node 2 are both connected with all the nodes in the dashed subgraph, the whole graph becomes a complete graph which means that labels may transfer more easily between the two nodes of the edge e_{12}. Thus, we introduce the extensional local clustering coefficient to calculate the edge's ability to transfer labels.

The neighbors N_i of the node i are replaced by common neighbors $N_i \cap N_j$ which belong to the two nodes in the edge e_{ij}. Equation (3) is transformed into the following form:

$$C_{ij} = \begin{cases} |N_i \cap N_j| \times \dfrac{\frac{2|\{e_{mn}|v_m, v_n \in N_i \cap N_j, e_{mn} \in E\}|}{|N_i \cap N_j| \times (|N_i \cap N_j| - 1)} + 1}{2} + 1 & |N_i \cap N_j| > 1 \\ 2 & |N_i \cap N_j| = 1 \\ 1 & |N_i \cap N_j| = 0 \end{cases} \tag{5}$$

Equation (5) is a good measurement of the edge's ability to transfer labels. It can avoid some particular cases. For example, when two nodes of an edge do not share a neighbor, the label cannot transfer on the edge. The extensional local clustering coefficient is a measure of the similarity and the tightness of the two nodes in an edge. The more closely they are connected with each other, the larger the extensional local clustering coefficient of the edge is. Usually, the measure is smaller on the edges between communities than the edges within communities. Thus, each edge's ability to transfer labels in a complex network can be measured by the extensional local clustering coefficient. Then LPA is transformed to the new algorithm, ELCLPA. In ELCLPA, the formula of $L(v_i)$ is as below:

$$L(v_i) = \arg\max_{l} \sum_{e_{ij \in E}} C_{ij}^{l} \tag{6}$$

$L(v_i)$ is the label of node v_i ($v_i \in V$) and if node v_i is connected with node v_j, edge $e_{ij} \in E$. C_{ij} is the extensional local clustering coefficient of e_{ij}. And C_{ij}^{l} means that the label of the node v_j is l.

The main steps of ELCLPA are as follows.

- ELCLPA calculates the extensional local clustering coefficient of each edge e_{ij} first.
- And then, it also updates labels in a random order, but it uses $L(v_i)$ in Equation (6) instead of $L(v)$ in Equation (1). In the initial label allocation phase, each node v_i can obtain the label from its neighbor v_j, and the C_{ij} of the edge e_{ij} is the maximum one among all the edges connected with v_i. This is more reasonable than the random selection in LPA's initialization. And at the step of updating labels, each node selects the label whose sum of C_{ij} is the maximum instead of one of the labels that most of the node's neighbors obtain currently. If there are several candidates, ELCLPA randomly selects one from them.
- All the nodes with the same label are clustered into the same community at last.

4. **Experiments and Discussion.**

Modularity Q is a significant measurement of community's quality to distinguish between results of different community detection algorithms.

$$\mathrm{Q} = \sum_{i} \left(e_{ii} - a_i^2\right) \tag{7}$$

where e_{ii} is the fraction of the edges in community i, and a_i is the fraction of the edges that link to the nodes in community i. Q is between zero and one, and the more Q is close to one, the better the quality of communities is.

The normalized mutual information (NMI) is widely used in clustering problems to measure the similar degree of two clustering results.

$$NMI(A,B) = \frac{2I(A,B)}{H(A)+H(B)} \tag{8}$$

Here we use the NMI to evaluate the stability of different community detection algorithms. A and B stand for two community partitions in a network. If A is identical to B, $NMI(A,B) = 1$. If A is totally different from B, $NMI(A,B) = 0$.

There are six different real-world network datasets in Table 1. Their topological structures are different from each other. The "Clusters" shows the real number of communities in the networks. And the "–" means that the value is unavailable.

We test LPA, CKLPA and ELCLPA on the networks provided in Table 1 for 50 times. The number of communities that they discovered is shown in Table 2. And the number is the mode of all 50 results here. The table shows that the number of communities which is discovered by ELCLPA, is closer to the real number compared with the other two algorithms.

Average modularity Q and average NMI of these results is displayed in Table 3. From the table we can see that ELCLPA's average modularity Q and NMI are larger than the results of LPA and CKLPA in all networks. The average modularity Q of CKLPA is larger than LPA sometimes but its NMI is not larger than LPA in most cases. It must be affected by the parameters which are manually selected. So ELCLPA's results are better and more stable.

Table 1. The real-world networks and their information

Network	Nodes	Edges	Clusters
Karate [18]	34	78	2
Dolphins [19]	62	159	2
Books [20]	105	441	3
Football [9]	115	613	12
Blogs [21]	1490	16715	–
Netsci [22]	1589	2742	–

Table 2. The number of communities discovered in the networks

Network	Clusters	LPA	CKLPA	ELCLPA
Karate	2	3	3	2
Dolphins	2	6	3	2
Books	3	3	5	3
Football	12	8	7	12

Table 3. The average modularity Q and NMI

Network	Modularity Q			NMI		
	LPA	CKLPA	ELCLPA	LPA	CKLPA	ELCLPA
Karate	0.31	0.35	0.39	0.65	0.62	0.87
Dolphins	0.45	0.42	0.54	0.54	0.48	0.75
Books	0.47	0.52	0.56	0.52	0.55	0.57
Football	0.53	0.32	0.58	0.89	0.84	0.94
Blogs	0.40	0.35	0.45	0.35	0.31	0.42
Netsci	0.84	0.73	0.90	0.32	0.33	0.54

5. **Conclusions.** Label propagation algorithm is a linear time algorithm for community detection without using any prior parameters. However, its performance of accuracy is not good enough and its randomness can hardly be ignored. For this reason, ELCLPA is proposed for community detection. ELCLPA gives each edge a weight which measures the edge's ability to transfer labels, involving adequate local topological information. The experimental results show that the results of ELCLPA are more accurate than LPA and CKLPA. Besides, ELCLPA is more stable than the other two algorithms in different sizes of networks. Furthermore, the communities, which are detected by ELCLPA, are closer to real communities in complex networks. Here, we can draw a conclusion that ELCLPA optimizes the robustness and effectiveness of LPA. In the future research, we will improve ELCLPA for community detection in directed networks.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (No. 61202350).

REFERENCES

[1] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small world' networks, *Nature*, vol.393, no.6684, pp.440-442, 1998.

[2] A. L. Barabási and R. Albert, Emergence of scaling in random network, *Science*, vol.286, no.5439, pp.509-512, 1999.

[3] H. Dev, A user interaction based community detection algorithm for online social networks, *Proc. of the ACM SIGMOD International Conference on Management of Data*, pp.1607-1608, 2014.

[4] W. Q. Lin, X. N. Kong, P. S. Yu, Q. Y. Wu, Y. Jia and C. Li, Community detection in incomplete information networks, *Proc. of the 21st International Conference on World Wide Web*, pp.341-350, 2012.

[5] H. Jin, S. L. Wang and C. Y. Li, Community detection in complex networks by density-based clustering, *Physica A: Statistical Mechanics and Its Applications*, vol.392, no.19, pp.4606-4618, 2013.

[6] N. Barbieri, F. Bonchi and G. Manco, Cascade-based community detection, *Proc. of the 6th ACM International Conference on Web Search and Data Mining*, pp.33-42, 2013.

[7] Z. Y. Xia and Z. Bu, Community detection based on a semantic network, *Knowledge & Information Systems*, vol.26, no.1, pp.30-39, 2012.

[8] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, *Proc. of the National Academy of Sciences of the United States of America*, vol.99, pp.7821-7826, 2002.

[9] M. E. J. Newman, The structure and function of complex networks, *SIAM Review*, vol.45, no.2, pp.167-256, 2003.

[10] P. Chen and S. Redner, Community structure of the physical review citation network, *Journal of Informetrics*, vol.4, no.3, pp.278-290, 2009.

[11] U. N. Raghavan, R. Albert and S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, *Physical Review E*, vol.76, no.3, 2007.

[12] L. Šubelj and M. Bajec, Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction, *Physical Review E*, vol.83, no.3, pp.885-896, 2011.

[13] M. E. Stokes, M. M. Barmada, M. I. Kamboh and S. Visweswaran, The application of network label propagation to rank biomarkers in genome-wide Alzheimer's data, *BMC Genomics*, vol.15, no.4, pp.1-13, 2014.

[14] J. Ugander and L. Backstrom, Balanced label propagation for partitioning massive graphs, *Proc. of the 6th ACM International Conference on Web Search and Data Mining*, pp.507-516, 2013.

[15] K. Kuzmin, S. Y. Shah and B. K. Szymanski, Parallel overlapping community detection with SLPA, *Proc. of the IEEE International Conference on Social Computing*, pp.204-212, 2013.

[16] J. Xie and B. K. Szymanski, Community detection using a neighborhood strength driven label propagation algorithm, *IEEE Network Science Workshop*, West Point, New York, USA, pp.188-195, 2011.

[17] Z. Lin, X. Zheng, N. Xin and D. Chen, CK-LPA: Efficient community detection algorithm based on label propagation with community kernel, *Physica A: Statistical Mechanics and Its Applications*, vol.416, no.C, pp.386-399, 2014.

[18] W. W. Zachary, An information flow model for conflict and fission in small groups, *Journal of Anthropological Research*, vol.33, no.4, pp.452-473, 1977.

[19] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten and S. M. Dawson, The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations, *Behavioral Ecology and Sociobiology*, vol.54, no.4, pp.396-405, 2003.

[20] V. Krebs, *http://www.orgnet.com/*, 2008.

[21] L. A. Adamic and N. Glance, The political blogosphere and the 2004 U.S. election: Divided they blog, *LinkKDD: Proc. of the 3rd International Workshop on Link Discovery*, pp.36-43, 2005.

[22] M. Fiedler, Algebraic connectivity of graphs, *Czechoslovak Mathematical Journal*, vol.23, no.2, pp.298-305, 1973.