A Community Discovery Algorithm Using Increment of Modularity to Optimize the Label Propagation Process

Xinqi Xu(✉) and Xiaoyan Zheng

Tianjin University of Technology and Education, Tianjing 300222, China
xxq_0414@163.com

Abstract. Due to the strong randomness of label selection, the label propagation algorithm makes the community results unstable. Especially in asynchronous updates, the final results are quite different due to the different order of selecting listeners. This paper proposes a community discovery algorithm using increment of modularity to optimize the label propagation process. In the process of label propagation, the increment of modularity is introduced to ensure that the increment of modularity is positive in each update. At the same time, the selection of popular nodes in traditional label propagation algorithms is retained. Combining these two methods, each node selection improve the division of the community and reduce the possibility of poor results due to asynchronous updates. The algorithm is verified in real network, and the results show that the algorithm is feasible and effective.

Keywords: Community detection · Label propagation · Modularity · Optimization

1 Introduction

A complex network is an abstraction of a complex system. In a complex network, nodes represent individuals in a complex system, and edges between nodes represent connections between individuals according to specific rules. There are also some statistical characteristics commonly found in complex networks, such as the "scale-free characteristic" [1], which reflects the characteristics of network nodes obeying a power-law distribution, the "small-world" [2], which reflects the characteristics of short path length and high clustering coefficient of the network. As well as the "community structure" that reflects the tight connections between nodes in the same community and the sparse connections between nodes in different communities. It is ubiquitous in complex networks [3].

With the rapid development of computer and information technology, especially the emergence of social networks, people are divided into countless small groups for various reasons. Meanwhile, the study of community structure has become increasingly valuable. It can help people to extract useful association

Y. Cao and X. Shao (Eds.): MONAMI 2022, LNICST 474, pp. 245–252, 2023.
https://doi.org/10.1007/978-3-031-32443-7_17

information from known social networks in complex networks. Therefore, community discovery has more and more applications in scientific research, commercial promotion, public safety and other fields. More and more researchers are devoted to related fields.

The rest of the paper is arranged as follows: Section 2 discusses the preparatory work related to the algorithm; Sect. 3 describes how the increment of modularity helps labels to spread better; Sect. 4 verifies the feasibility of the algorithm in real networks; Finally, the full text is summarized. and discuss the next stage of work.

2 Related Work

After years of exploration, researchers have proposed many classic community discovery algorithms.

The Label Propagation Algorithm (LPA) was proposed by Zhu et al. [4] in 2002, which uses the relationship between samples to build a relational complete graph model. In 2007, Raghavan et al. [5] first proposed the application of LPA to community discovery, which is referred to as the RAK algorithm for short. The algorithm will give each node a unique label, select the most popular label to record in the iterative process, and finally the nodes with the same label will be combined into the same community.

Since most of the communities in the real network are not independent of each other, the study of overlapping communities becomes very important. Gregory [6] extended the RAK algorithm. The algorithm allows each node to retain multiple labels, so that the node can select multiple different communities to discover overlapping community structures. However, as the number of iterations increases, the performance decreases significantly, which is not suitable for today's huge dataset environment. Xie et al. [7] proposed the SLPA algorithm, which retains the charac-teristics of low complexity and high efficiency of the LPA algorithm. Label all nodes with different labels, scan all labels of the nodes, the record with the most occurrences is the candidate label, the label with the most occurrence of the candidate label is the node community name, and the final community division is obtained after multiple traversals.

Modularity, as an important indicator for evaluating community structure, has been applied in label propagation by many scholars. In order to avoid all nodes in LPA selecting the same community, Barber et al. [8] proposed a modularity-specialized label propagation algorithm (LPAm), which is a constraint-based LPA monitoring network community. A variable is introduced to maximize the modularity value of the community. The community discovery problem is transformed into a solution problem of objective function optimization. On the basis of the number of connected vertices with the same label, an objective function H is defined, and the LPA algorithm is used to find the local optimal value of the H function. Qiao et al. [9] proposed an overlapping community detection algorithm in complex network big data, based on the idea of modularity clustering, graph computing and optimal modularity. A balanced

binary tree is used to index the increment of modularity to alleviate the impact of multiple modularity computations on the performance of the algorithm. Leung et al. [10] also pointed out that the module maximization method is not a scale-free interval measure method, and it is not feasible to detect communities only by relying on it. Therefore, modularity is generally not used as the final objective function, but it can still help us improve the quality of the found community structure.

Through numerous algorithm studies, we found that the LPA algorithm has low complexity and can be well adapted to the monitoring of large-scale communities. It does not require optimization of a predefined objective function, nor does it require prior information about the number and size of communities, and there is no limit to the size of communities. However, due to the uncertainty of label selection, the final community division result is unstable. This paper chooses the community structure evaluation index - modularity, to help nodes select labels. It avoids the shortcomings of modularity as an objective function, and at the same time, it can help nodes choose labels that are beneficial to themselves as much as possible. The increment of modularity after each label propagation is kept positive, so that each node label selection can improve the overall community division and reduce the possibility of poor results due to updates.

3 Description of the Proposed Algorithm

This chapter introduces the specific calculation formula of the increment of modularity, describes the community discovery algorithm using increment of modularity to optimize the label propagation process (IMO-LPA) in detail, and finally verifies that the time complexity of algorithm is close to linear.

3.1 The Increment of Modularity

Modularity [11] is an important indicator for calculating community accuracy. The specific formula is as follows:

$$Q = \sum_{c=1}^{n_c} \left(\frac{l_c}{m} - \left(\frac{d_c}{2m} \right)^2 \right) \tag{1}$$

where, m is the total number of edges in the network, n_c is the number of communities, l_c is the total number of edges in community c, and d_c is the sum of the degrees of all nodes in community c.

If the node i in the community c_2 is allowed to enter the community c_1, that is, the increment of modularity generated by the node i selecting the label represented by the community c_1

$$Q_{before} = \frac{l_{c_1}}{m} - \left(\frac{d_{c_1}}{2m} \right)^2 + \frac{l_{c_2}}{m} - \left(\frac{d_{c_2}}{2m} \right)^2 \tag{2}$$

$$Q_{after} = \frac{l_{c_1} + k_{i\_c1}}{m} - (\frac{d_{c_1} + k_i}{2m})^2 + \frac{l_{c_2} + k_{i\_c2}}{m} - (\frac{d_{c_2} - k_i}{2m})^2 \quad (3)$$

Formula (3)–(2) can obtain formula (4):

$$\Delta Q = \frac{1}{m}[k_{i\_c_1} - k_{i\_c_2} - \frac{k_i(d_{c_1} - d_{c_2} + k_i)}{2m}] \quad (4)$$

where, $k_{i\_c_1}$ represents the number of connecting edges between node i and nodes in community c_1. $k_{i\_c_2}$ represents the number of connecting edges with the nodes in the community c_2. If there is only node i itself in the community c_2, then $k_{i\_c_2} = 0$.

3.2 IMO-LPA Algorithm

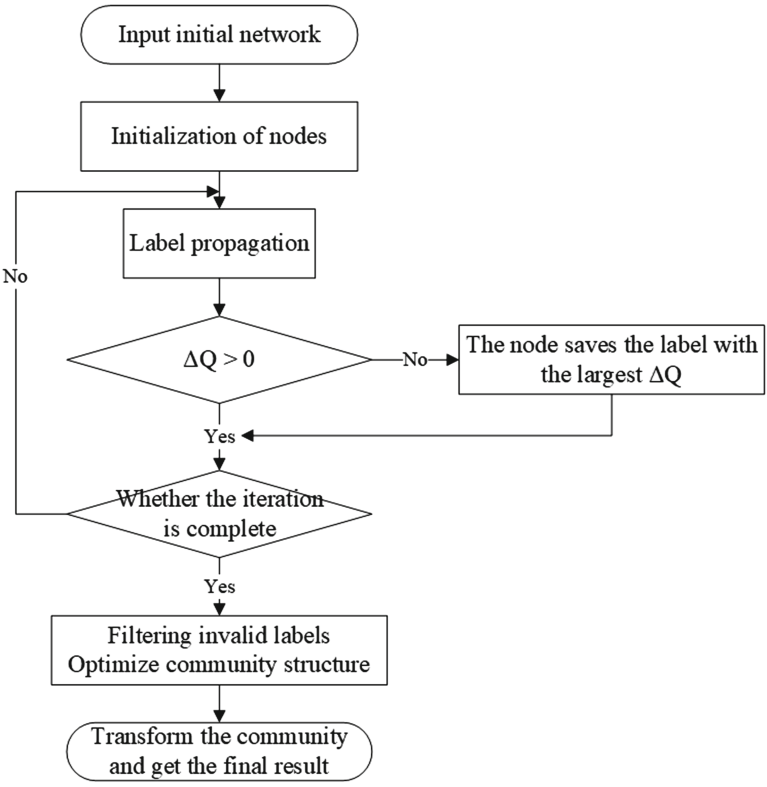

Fig. 1. The specific flow of the algorithm

Modularity, also known as modularity metric, characterizes the accuracy of the division of node communities in the network. The general basic idea of using

modularity for community division is: If the connection strength between nodes in a subgraph of a network is much greater than the connection strength between nodes in the subgraph under random division of the network, then the subgraph can be considered as a community of the network. The increment of modularity is the increase of the calculated modularity. If two nodes are in the same community and have a positive effect on the entire community structure, then the increment of modularity is positive. Similarly, the closer the connection between the two nodes is, the greater the increment of modularity will be. Due to the high randomness of label propagation, generally speaking, the higher the influence of the label, the faster the propagation, but there are a large number of random selections in the selection process, especially in the initial stage. At this time, the increment of modularity can be added to help nodes select more suitable nodes for message interaction. The specific process is as Fig 1:

3.3 Complexity Analysis

We perform a time complexity analysis of the IMO-LPA algorithm. The time complexity of initializing nodes and communities is $O(N)$. Each node will act as a listener, and the number of iterations is t_0, and the sounding nodes are the nodes of the listener, that is, the node degree $k_{listener}$ of the listener. For the calculation of the increment of modularity, at least $O(1)$, the most popular labels meet the judgment conditions; at most $O(k_{listener})$, then each label has the same weight. The final label optimization, if the final number of communities obtained is $N_{community}$, the time complexity is $O(NN_{community})$. The time complexity is $O((2+t_0k^2_{listener}+N_{community})N)$. In general, t_0 and $N_{community}$ tend to a fixed value, and $k_{listener}$ is much smaller than the number of nodes N, so the time complexity of the algorithm tends to be linear.

4 Experiments

4.1 Experiment Preparation

This paper also compares the IMO-LPA algorithm in four different real networks and one artificial synthetic network. The information and parameters of the experimental network are shown in Table 1.

Table 1. General information of the real network.

Network	Description	Nodes	Edges
Dolphins [12]	Lusseau's dolphins	62	159
Polbooks [13]	Amazon's american political book	105	441
Football [3]	American College football union	115	616
Powergrid [1]	The topology of the powergrid of the United States	4941	6594

The experiment will use two evaluation indicators to verify the community division results. For communities with accurate community division results, the Normalized Mutual Information (NMI) [14] can quantify the similarity between the generated algorithm partition community and the standard community, and measure the accuracy of the algorithm results. For communities without accurate division results, the overlapping modularity (Qov) [15,16] is used to help evaluate the merits of the division results. At the same time, Qov can also help us judge whether the impact of overlapping modularity on the algorithm is positive.

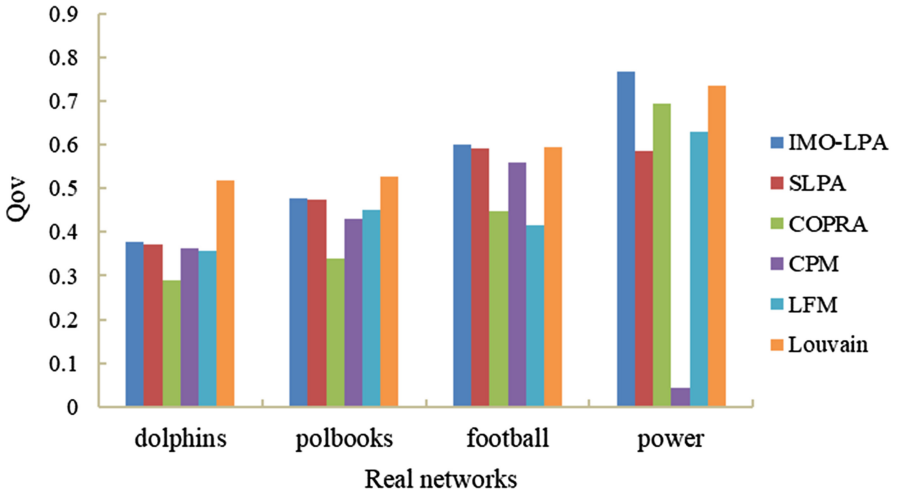

Fig. 2. Qov comparison experiment in real networks

4.2 Result Analysis

In this paper, a total of five classic algorithms and MIO-LPA algorithms are selected for comparison experiments, namely SLPA, COPRA, CPM [17], LFM [15] and Louvain [18]. As can be seen from Fig. 1, the IMO-LPA algorithm has the best modularity in the network football and power, and is slightly lower than the algorithm Louvain in the network dolphins and polbooks. Overall, the IMO-LOA algorithm outperforms most experimental algorithms in real networks Among the four real networks selected in the experiment, the network dolphins, polbooks and football all have accurate community division results. NMI is used for evaluation in these three networks. The evaluation results are shown in Fig. 2. It can be seen from the figure that the accuracy of the IMO-LPA algorithm is significantly higher than that of other algorithms. In general, the modularity of general real networks is close to 0.5, and the IMO-LPA algorithm is not the best in overlapping modularity, but has better accuracy. Therefore, it shows that the increment of modularity can indeed help the label to propagate better (Fig. 3).

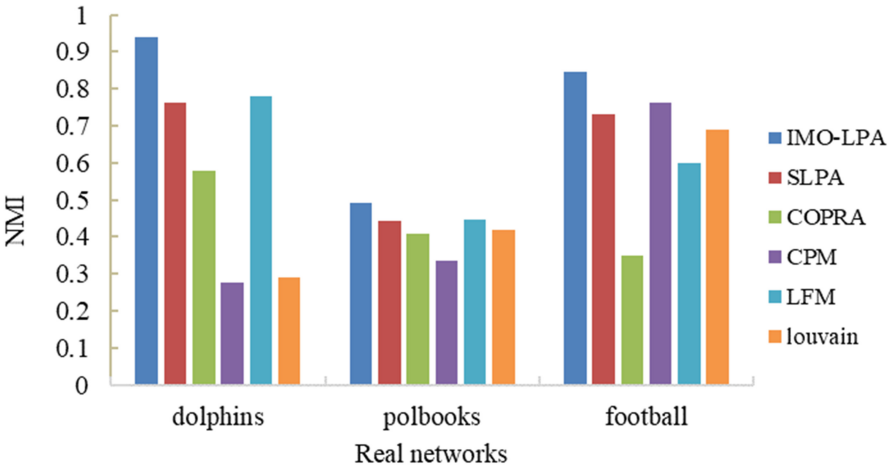

Fig. 3. NMI comparison experiment in real networks

5 Conclusion

Because the LPA algorithm has the characteristics of low complexity and high performance, it is more in line with the needs of today's huge data. However, because of the large amount of randomness in the process, the results of the algorithm are sometimes not as good as expected. In this paper, the increment of modularity is added to the label propagation process to help nodes select more reasonable labels and reduce the adverse effects of uncertainty. Experiments in real networks show that the algorithm increases the accuracy of the algorithm while retaining the advantages of the LPA algorithm.

Acknowledge. This work is supported by Tianjin Science and Technology Planning Project (Grant No. 64822KPXMRC00170), Science and Technology Think Tank Young Talent Program, China (Grant No. 64920220615ZZ07110153).

References

1. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world'networks. Nature **393**(6684), 440–442 (1998)
2. Barabási, A., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
3. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. **99**(12), 7821–7826 (2002)
4. Zhu, X.: Learning from labeled and unlabeled data with label propagation. Technical report (2002)
5. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E **76**(3), 036106 (2007)
6. Gregory, S.: Finding overlapping communities in networks by label propagation. New J. Phys. **12**(10), 103018 (2010)

7. Xie, J., Szymanski, B.K., Liu, X.: SLPA: uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. arXiv e-prints (2011)
8. Barber, M.J., Clark, J.W.: Detecting network communities by propagating labels under constraints. Phys. Rev. E **80**(2), 026129 (2009)
9. Qiao, S.J., Han, N., Zhang, K.F., Zou, L., Gutierrez, L.A.: Algorithm for detecting overlapping communities from complex network big data (2017)
10. Leung, I.X., Hui, P., Lio, P., Crowcroft, J.: Towards real-time community detection in large networks. Phys. Rev. E **79**(6), 066107 (2009)
11. Waltman, L., Van Eck, N.J.: A smart local moving algorithm for large-scale modularity-based community detection. Eur. Phys. J. B. Condens. Matter Phys. **86**, 1–14 (2013)
12. Lusseau, D., Newman, M.E.J.: Identifying the role that animals play in their social networks. Proc. R. Soc. London. Series B: Biol. Sc. **271**(suppl_6), S477–S481 (2004)
13. Newman, M.: Mark newman. http://www-personal.umich.edu/mejn/
14. Attea, B.A., Hariz, W.A., Abdulhalim, M.F.: Improving the performance of evolutionary multi-objective co-clustering models for community detection in complex social networks. Swarm Evol. Comput. **26**, 137–156 (2016)
15. Lancichinetti, A., Fortunato, S., Kertész, J.: Detecting the overlapping and hierarchical community structure in complex networks. New J. Phys. **11**(3), 033015 (2009)
16. Nicosia, V., Mangioni, G., Carchiolo, V., Malgeri, M.: Extending the definition of modularity to directed graphs with overlapping communities. J. Stat. Mech: Theory Exp. **2009**(03), P03024 (2009)
17. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**(7043), 814–818 (2005)
18. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. **2008**(10), P10008 (2008)