A Novel Label Propagation Algorithm based on Core Node

Shanshan Chen [a], Tao Wang [b] and Jinfang Wang [c]

School of Mathematics and Physics, North China Electric Power University, Baoding 071003, China

[a]1462981577@qq.com, [b]wtwxx@126.com,[c]897311324@qq.com

Keywords: CN-LPA algorithm, core node, label initialization, community detection.

Abstract: Detecting community structure can give a significant insight into structural and functional properties in complex networks. In this paper, we propose a novel label propagation algorithm based on core node. In the label initialization process, the node importance index is calculated to get the core node and a label is assign to the node through the core node. In the label propagation process, when the maximum number label of neighbor nodes is not unique, the importance of the neighbor nodes is sorted and the most important node is selected to update the current node. The proposed algorithm is test on both real network and synthetic network, and is compared with classical algorithms in community detection. The experimental results confirm the feasibility and effectiveness of the proposed algorithm.

1. Introduction

There are many kinds of community detection methods. Among them, the LPA algorithm is well known because of its linear complexity [1]. However, the traditional LPA has some shortcomings. In the node initialization label process, the LPA algorithm sets a unique label for each node, and treats each node equally. This may cause small communities to be swallowed up. In addition, in the label propagation process, when the maximum number label of neighbor nodes is not unique, random selection of label for update may lead to a countercurrent phenomenon. In order to solve the above problem, we propose a CN-LPA algorithm, which not only resets the initialization label of the node, but also defines the node importance index. In the label initialization process, calculate the importance of each node and use the average of all node importance as a criterion. Then, if the importance value of the node is greater than the average value, the node is the core node and set a unique label for the node through the core node. Moreover, in the label propagation process, when the maximum number label of neighbor nodes is not unique, the importance of the neighbor nodes is sorted and the most important node is selected to update the current node.

2. Definition

Given an unweighted and undirected network $G=(V,E)$, E represents a collection of edges, V represents a set of nodes, n represents the number of nodes, $m=|E|$ represents the number of edges. The adjacent matrix $\boldsymbol{A}$ is symmetric, and element $a_{i,j}$ indicates whether the edge between node i and node j is truly existed. If edge e_{ij} exists, then $a_{i,j}=1$, otherwise $a_{i,j}=0$.

2.1 Node Importance Index

We define the node importance index as:

$$NI(i)=s_{i,i}+\sum_{j\in N(i)} d_j\, sim(i,j) s_{j,j} \tag{1}$$

Where $N(i)$ represents the neighbor set of node i, s_{ii} represents the signal propagation amount of the node [2], $sim(i,j)$ represents the similarity between adjacent nodes [3] and d_j represents the degree of the node. The greater the $NI(i)$ value of a node, the greater the importance of the node.

2.2 Asynchronous Update

The method of label update can be divided into synchronous update method and asynchronous update method. Synchronous updates may cause label oscillating in a binary network or an approximate binary network. However, asynchronous updates can well resolve this problem [4]. The asynchronous update formula is as follows:

$$c_j(t) = f\left(c_{j_1}(t-1), \cdots, c_{j_m}(t-1), c_{j_{m+1}}(t), \cdots, c_{j_k}(t)\right), j_i \in N_j \tag{2}$$

where the f function denotes the most frequently occurring label among neighbor nodes of node j. $c_j(t)$indicates the label of node j at the t-th iteration, and $c_{j_m}(t-1)$represents the label of the m-th neighbor node of node j at the (t-1)-th iteration. m is the label number of non-updated neighbor nodes of node j in this iteration.

3. Proposed Methods

Algorithm steps:

Input: undirected and unweighted network $G = (V, E)$, the maximum number of iterations t_{max}

Output: divided communities C.

Step 1: Calculate NI for each node according to equations (1).

Step 2: Calculate the average of the node importance values, using the average as the standard. If the node's importance value is greater than the average, the node is the core node and a unique label is set for the core node.

Step 3: Set $i = 1$, determine if node i has a label. If the node has a label or does not have a core node directly connected to the node, skip to continue to the next step. Otherwise, find the core node directly connected to the node, and select the label of the most important core node to update current node.

Step 4: $i = i + 1$

Step 5: Repeat steps (3)-(4) until $i = n$. If each node in the network has a label, skip step (6) and directly perform step (7). Otherwise, perform the next step.

Step 6: Continue with steps (3)-(5) until each node in the network has a label.

Step 7: Set the iterations number $t = 1$.

Step 8: Update the label for each node in the network by equations (2). When the labels with the maximum number of neighboring nodes are unique, the label is updated with its own label. Otherwise, the label of neighbor node with the most importance is updated.

Step 9: $t = t + 1$,

Step 10: Repeat steps (8)-(9) until $t = max$ or the label of each node no longer changes Then the nodes with the same label are divided into the same community, and the algorithm ends.

4. Experimental Results

In order to test the performance of CN-LPA algorithm on real networks and synthetic network, some classical algorithms are contrasted in the experiment, including LPA algorithm [1], GN algorithm [5], FN algorithm [6] and BLDLP algorithm [7]. The performance of algorithm is compared by analyzing normalized mutual information (NMI) [8] and modularity.

4.1 Real Network

The CN-LPA algorithm was tested in three real networks, including the college football network and the dolphin network.

Fig. 1 shows the communities number in two real networks under different experimental times. From Fig. 1, the following analysis can be derived. The community detection result obtained in the CN-LPA are more stable. On the contrary, the community detection result obtained in the LPA algorithm are more volatile. The experimental results show that the stability of CN-LPA algorithm is better.

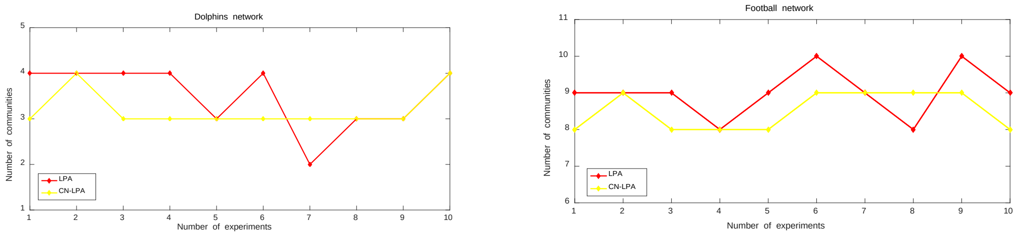

Fig. 1 Communities number under different experiment times

4.2 Experiments on Artificial Network

LFR benchmark network is often used to simulate the network structure with large community scale, distribution of nodes in various communities unevenly and unfixed degree. The LFR network is closer to the actual situation of real-world network. In this paper, LFR benchmark network are set as follow: $N = 1000$, $k_{max} = 50$, $\langle k \rangle = 20$, $c_{min} = 15$, $c_{max} = 50$, $\tau_1 = 2$, $\tau_2 = 1$, and mixing parameter μ varies from 0.1 to 0.8.

The performance of five algorithms on LFR benchmark network is compared by NMI curve, as shown in Fig. 2. Form Fig. 2 observation, the following analysis can be derived. When $\mu \in [0.1, 0.4]$, the NMI of most methods stay at about 1. However, the NMI value of BLDLPA begin to decrase partly. When $\mu \in [0.4, 0.8]$, the NMI value of CN-LPA is higher than others. When μ is 0.6, the NMI value of CN-LPA is higher than others at least by 0.1. Therefore, The NMI of CN-LPA are better than other contrast algorithms in the range [0.4, 0.8].

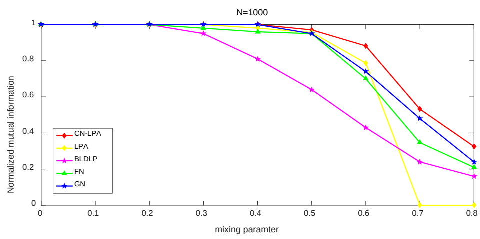

Fig. 2 Comparison of five methods on LFR benchmark.

5. Conclusion

In this paper, a label novel propagation algorithm based on core node is proposed. First, we propose a node importance formula and calculate the core node. Secondly, in the process of label initialization, set a unique label for the node through the core node. Finally, in the process of label propagation, when the neighbor node with the most occurrence of the label is not unique, the most important node is selected to update the current node. The simulation results show that the detection accuracy of CN-LPA is superior to contrasted methods on synthetic and real-world complex networks.

References

[1] U.N. Raghavan, R. Albert, S. Kumara, Near linear-time algorithm to detect community structures in large-scale networks, Phys. Rev. E. Vol. 76 (2007) No. 3, p.036106.

[2] S. Pahwa, M. Youssef, P. Schumm, Optimal intentional islanding to enhance the robustness of power grid networks, Physica A. Vol. 392 (2013), p.029.

[3] Z. Deng, H. Qiao, M. Gao, et al. Complex network community detection method by improved density peaks model, Physica A. Vol.526 (2019), p.212070.

[4] X. Zhang, J. Ren, C. Song, et al. Label propagation algorithm for community detection based on node importance and label influence, Phys.Lett.A. Vol.381 (2017) No.33, p.2691-2698.

[5] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. Vol.99 (2002) No.12, p.7821–7826.

[6] R. Guimera, M. Sales-Pardo, L.A. Amaral, Modularity from fluctuations in random graphs and complex networks, Phys. Rev. E. Vol.70 (2004) No.2, p.188.

[7] E. Jokar, M. Mosleh, Community detection in social networks based on improved label propagation algorithm and balanced link density, Phys. Lett.A. Vol.383 (2019) No.8, p.718-727.

[8] M.E. Newman, M. Girvan, Finding and evaluating community structure in networks. Phys. Rev. E. Vol. 69 (2004), p.026113.