Analysis of Overlapping Community Based on Complex Network

Xiaoyuan Xu

Department of Information and Technology, Nanjing, China

Keywords: network community; spectral clustering; edge; detection;

Abstract: Community detection is the basis of studying complex network structure. Based on the analysis of existing overlapping community detection algorithms, an edge-based overlapping community detection algorithm SAEC is proposed. The algorithm regards the community as a set of edges. By defining the similarity of edges, the probability transition matrix is obtained. The number of communities is automatically determined by spectral clustering method. Finally, overlapping communities are divided by K-means algorithm. The validity of the algorithm is verified by the test of random generated network and real network.

1. Introduction

With the deepening of network research in the world today, researchers have found that many real networks not only have community structure, but also have overlapping and interrelated characteristics among communities. That is, there are some special nodes in the network, which are closely connected with many communities and belong to many communities. For example, a particular individual in a social network may belong to different communities such as schools, hobbies, families, etc. If the nodes are strictly divided into a community, it can not reflect the actual relationship between the nodes and the community in the real world. Therefore, the discovery of overlapping communities is more in line with the law of community organization in the real world.

In the complex network structure, the edge reflects the interaction, cooperation and interaction among nodes in the network, which can only belong to a community. Therefore, the use of edge characteristics to achieve community partition can make the partition results more truly reflect the role and function of nodes in complex networks. Based on this, from the perspective of edge, this paper proposes a kind of overlapping community discovery algorithm SAEC (spectral analysis based on edge clustering), which represents the edge in the network as a point in the vector space, and divides the edge with similar random track distribution into the same community based on the spectral analysis method of probability transfer matrix. In order to realize the division of border communities.

2. Relevant Work

In order to accurately and effectively analyze the community structure in complex networks, many community structure partitioning algorithms have been proposed by researchers. Early community partitioning algorithms divide the network into several separate communities, ignoring the overlap between communities [2-6]. In order to discover overlapping community structure, Palla et al. [7] proposed an overlapping community discovery algorithm CPM, and applied it to protein network, scientist co-authorship network and communication network, and got good results. Baumes et al. [8] based on the idea of local optimization, proposed a local optimization function to evaluate the community structure according to the edge density. According to certain strategies, all communities in the network were optimized locally, and the final optimization results were taken as the result of community division. Because a node may be added to the community in different optimization processes, overlapping communities can be found.

Based on the GN algorithm, the algorithm CONGA [9] splits the nodes with the highest node median into multiple copies, and then uses the traditional GN algorithm for community discovery, and finally merges the splitted nodes. Because the splitted nodes are divided into different

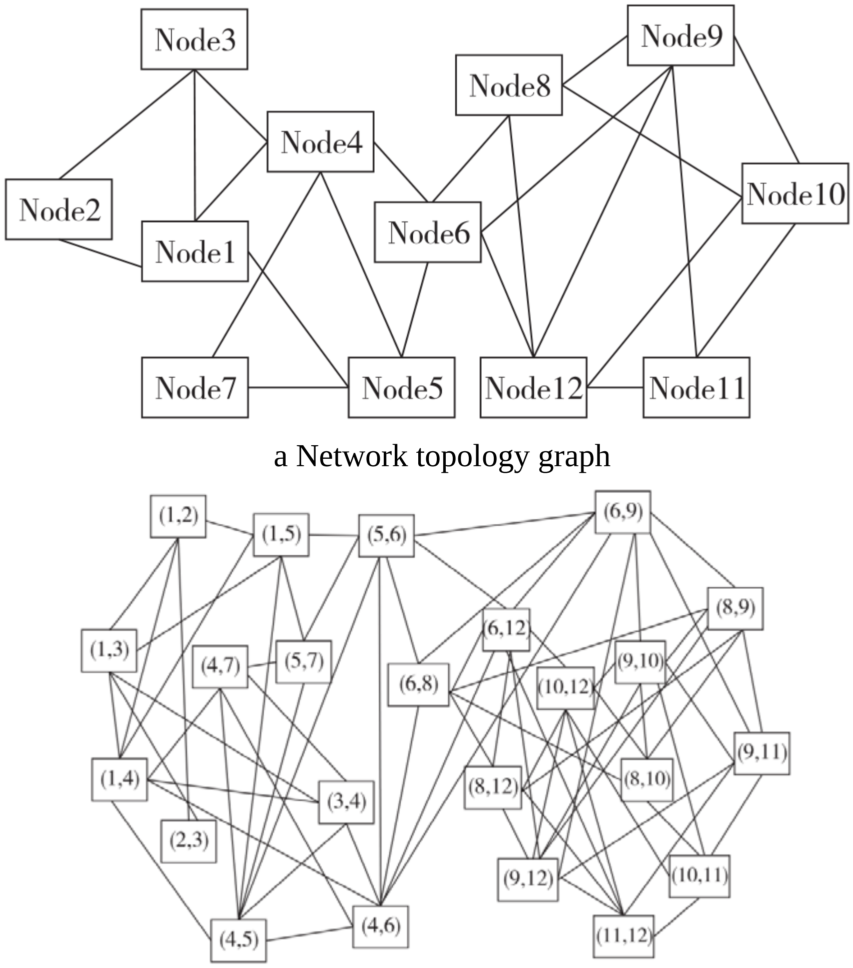
DOI: 10.25236/aisct.2019.089

communities, the discovery of overlapping communities is realized. Nicosia V et al. [10] extended the definition of modularity to directional networks and proposed the concept of node membership coefficient, which also made the judgement of overlapping communities more feasible. However, this method is still arbitrary in determining the contribution weight of each community. Literature [11] Based on the idea of local optimization of fitness function, a LFK algorithm is proposed to simultaneously discover overlapping communities and hierarchical structures in networks. The algorithm assumes that the community is essentially a local structure, including nodes belonging to the module itself, plus at most their neighbors. By defining the fitness function of the nodes, the fitness of all neighbor nodes to the community is calculated, and the nodes with the greatest fitness to the community are selected to join the community in order to achieve community division. The algorithm allows nodes to be assigned to multiple communities, so overlapping communities can be found.

3. SAEC Algorithm

3.1 Line Graph

Based on edges, this article uses the relation between edge and edge to partition communities. To illustrate it, this article uses line graph[14] to describe the edge relation in the network. Given an undirected graph $G = \langle V, E \rangle$, the corresponding line graph $L(G)$regards the edges in graph G as a set of vertexes. If any two edges share a same node in graph G, there will be an edge existing at this node in line graph $L(G)$. As shown in Figure 1, Graph 1a is a randomly generated network topology with 12 nodes and 23 edges and Graph 1b is the corresponding line graph.



a Network topology graph

b Line graph

Figure 1 Network topology graph and line graph

As is shown in Figure 1a, the nodes in the areas on the left side and the right side of node 6 are connected tightly respectively, but there is no any connection between these two areas. Also, as is shown in Figure 1b, the structural features of these bilateral communities made up of edges are very obvious.

3.2 Calculate Similarity

The spectral clustering algorithm based on transition probability matrix regards that if a walker departs from a node in any community in a network, it will keep walking between nodes in this community for a long time before leaving this community. When this walker enters another community and selects the next starting node, the probability of selecting nodes inside the community is far bigger than outside the community. Therefore, if two nodes start random walk with same number of steps and the walk trajectories are very consistent, it is highly possible that the edge connecting these two nodes is inside the community; otherwise, if the walk trajectories are very different, it is highly possible that the edge connecting these two nodes is the edge between communities. This kind of walk trajectory can be described by a network random walk transition probability matrix P^M with M steps. In this matrix, each row represents a random walk trajectory which takes a node in network as staring point. A large number of experimental results show that if node i and node j locate in same community, they will have very high degree of similarity and the values of P_i^M and P_j^M are close.

To get the similarity of the edges, this article defines the similarity degree between two nodes as formula (1):

$$S\left(e_{ij}, e_{kd}\right) = \frac{\left|\left(n_i \cup n_j\right) \cap \left(n_k \cup n_d\right)\right|}{\left|n_i \cup n_j \cup n_k \cup n_d\right|}, S \in [0,1] \tag{1}$$

n_i is the set of neighbor nods of node and $n_i = \left|\bigcup_{e_{ij} \in E} j\right|$ E is the set of all edges in a network topology graph.

Through analysis, it is found that there are mainly three cases regarding the similarity between edge and edge. The first one is that if two edges share a same node, it is highly possible that they are in the same community. The second one is that if two edges do not share a same node but they have many mutual “friends’, it is possible that they are in the same community. The final one is that that if two edges do not share a same node and they have less or do not have mutual “friends’, it is possible that they are not in the same community.

In order to partitio the community more efficiently, this article also takes the step length between two nodes into account when the similarity between edge and edge are considered. If the minimum path length between two nodes in a line graph is bigger than step M, the degree of similarity is 0. Because the step length of these two nodes has exceeded the scope measured by step length. When the length of possible path between two nodes in a line graph is defined as $TL_{mn} = \{tl_1, tl_2, tl_3, \cdots\}$, the values of elements in similarity matrix ***W*** of nodes are as below.

$$W_{nn} = \begin{cases} S(m,n), \text{ if } 0 < \min\left(TL_{mn}\right) \le M \\ 0, \text{ if } \min\left(TL_{mn}\right) = 0 \text{ or } \min\left(TL_{mn}\right) > M \end{cases} \tag{2}$$

In transition probability matrix $P = D^{-1}W$, D is diagonal matrix and its definition is as below.

$$D = \left\{ d_{ii} = \sum_{j=1}^{n} w_{ij}, d_{ij} = 0 \right\} \tag{3}$$

Based on the definition above, Figure 2 shows the corresponding transition probability matrix ***W***, diagonal matrix ***D***and transition probability matrix ***P*** (Only the top 12 columns are shown due to limited page layout).

3.3 Determine the Number of Communities

The key problem is how to automatically determine the number of communities and how to make the result close to the real community structure. Ideally, based on the heuristic rule of spectral clustering method, if a data set contains k cluster, there will be a large difference between the first k largeeigenvalues and the following $k+1$ small eigenvalues in the corresponding characteristic matrix, which is called as characteristic interval. Therefore, after getting the transition probability matrix ***P***, eigenvectors$V=\left(v_2,\cdots v_n\right)$ and eigenvalues$\lambda=\left(\lambda_1,\lambda_2\cdots,\lambda_n\right)$ can be calculated. Based on that heuristic rule, after sorting eigenvaluesλ_n from largest to smallest, the k value that meets $\max\left(\lambda_k-\lambda_{k+1}\right)$ can be found and it is the number of the communities.

For the line graph in Figure 1b, the eigenvectors, eigenvalues and matrix ***H*** are shown in Figure 3. Through calculation, the value of maximum characteristic interval$\max\left(\lambda_k-\lambda_{k+1}\right)$is 0.613, k=2. From the result, it is found that this network is divided into two communities after structuring matrix ***H*** with first two column of eigenvectors and calling K-means algorithm. One community contains nodes{1、2、3、4、5、6、7}and the other one contains nodes{6、8、9、10、11、12}. Obviously, node 6 is the overlapping node of these two communities.

4. Experiment and Analysis

To verify the effectiveness of the algorithm, this article uses the network constructed in Reference[4] and the protein function network provided by CFinder, a free software which is used to resolve network overlapping community problem based on CPM algorithm and contains lots of real data, as the testing data and compares the results with classical CPM algorithm.,

In terms of the network in Reference [4], maximum characteristic interval is $\max\left(\lambda_k-\lambda_{k+1}\right)=0.3736$, k=4, based on the transition probability matrix calculated with the algorithm in this article. Therefore, there are four communities in this network and the partition result is consistent with the one of CPM algorithm as shown in Figure 4.

The overlapping nodes found by CPR algorithm are {3,10,11,12} and the algorithm in this article finds 4 more overlapping nodes, which are {4,17,22}. Because this algorithm uses edge to partition communities, resulting that edge (5,22) is allocated to a community and the nodes on the two ends are also allocated to the community. Algorithm allocates edge (5,22) to the community represented by square node, so node 22 becomes an overlapping node without doubt. As shown in Figure 4, node 4, 7 and 22 are the join node between two communities and the communication between these two communities must be realized by information transmission among those nodes.

5. Conclusion

This algorithm focuses on the relationship between edges in the network, and uses the similarity between edges to achieve community partition. Because a node may have multiple edges associated with it, and as different edges are divided into different communities, nodes are also divided into different communities accordingly, so that overlapping nodes belonging to multiple communities can be found. The experimental results show that the algorithm can effectively find overlapping communities. Because the algorithm performs edge clustering based on the block matrix partitioned by spectral analysis, the number of communities generated does not depend on external conditions, and is completely determined by the distribution of data objects, reflecting a smaller external dependence. In the future research work, we will consider the hierarchy of the network and select a larger network to verify the algorithm.

Acknowledgement

Natural science foundation for universities of Jiangsu Province, 19KJB520026

References

[1] Xie Jie-rui,Kelley S,Boleslaw K,et al.Overlapping community detection in networks:The state of the art and comparative study[J].ACM Computing Surveys,2012,45(4):43.

[2] Newman M E J.Communities, modules and large-scale structure in networks[J].Nature Physics,2012(8):25-31.

[3] Newman M E J.Fast algorithm for detecting community structure in networks[J].Physical Review E,2004,69(6):066133.

[4] Clauset A,Newman M E J,Moore C.Finding community structure in very large networks[J].Physical Review E,2004,70(6):066111.