Research on Key Nodes Identification based on Clustering Coefficient in Topological Networks

Jianxi Wang
School of Computer Science, Pingdingshan University, Pingdingshan, 467000, China

Abstract: Traditional identification methods of key nodes in network mainly study the unauthorized and undirected network. The identification effect of nodes in network with changing topological relations is unstable, and the identification method has low accuracy and poor robustness. Aiming at the above problems, the key nodes identification method based on clustering coefficient is studied. Based on the topology network model, the aggregation coefficients of nodes in the network are calculated. The aggregation coefficient is taken as the node characteristic attribute and the improved KNN algorithm is used to identify the key nodes. Compared with the traditional node recognition method, the experimental results show that the recognition accuracy of the proposed method can reach 97.15% on average. Moreover, the proposed method has good robustness in application and can be applied in practice.

Keywords: Agglomeration coefficient; Topological network; Key nodes; Node identification; Improved KNN

1. Introduction

Topological network means that devices and communication nodes in the network are connected to each other according to the established connection mode, so that the network has an obvious topology structure. With the development of related concepts, most systems in the real world can be represented by topological networks. It is found that topological networks all have the characteristics of non-homogeneous topological structures, which determines the different status of each node in the network: nodes with great influence on the network account for a small part of the network nodes, and these nodes are called the key nodes of the network [1]. The security of the network can be improved by identifying and protecting the key nodes of the topological network.

The traditional key node recognition method based on information entropy USES the principle of information entropy, but information entropy can only be applied to undirected network, and the recognition effect for network nodes with topology changes is not good. All the research objects based on degree centrality algorithm are unauthorized networks and are based on the premise that the integrity of the network will not be damaged. However, when identifying key nodes in the topology network, the importance of nodes in the topology network cannot be accurately assessed, so it has certain limitations [2, 3]. Agglomeration coefficient is one of the most significant features of network structure, and it can accurately represent the key degree of network nodes. Based on the above analysis, this paper will carry out research on the clustering coefficient based recognition method of key nodes in topological networks.

2. Research on Key Node Recognition Method based on Clustering Coefficient in Topology Network

2.1. Set up topology network model

In order to analyze the strength and direction of interaction between nodes in the network, a directed weighted topology network model is established. Node strength refers to the sum of weights of all edges connected with the node. Since the connection between nodes has directivity, node strength can be divided into incoming strength and outgoing strength according to the different direction of the connection.

The modeling process of directed weighted topology network model is as follows [4]:

Step 1: If the network in the initial state is a fully connected network composed of n_0 nodes, that is, all nodes in the network are interconnected, set the initial weight of each connection in the network to w_0;

Step 2: In each time step of model evolution, new nodes and directed edges are added to the original network according to the following principles:

A. A new node v was added to the original network with a fixed probability of q, and then m nodes were selected from the original network according to certain principles to be connected with it, so that the number of outgoing edges, that is, the outgoing degree of this node, was subject to binomial distribution of $B(n,p)$, and the number of incoming edges, that is, the incoming degree of this node was subject to binomial distribution of $B(n,1-p)$, among which p met $0 \le p \le 1$.

B. The number of nodes is fixed and only new connections are added to the original network. The probability of adding new edges is $1-q$ and the number $n \le n_0$.

Step 3: For the two network growth modes in step 2, the newly added nodes or edges will select the nodes in the original network to connect in a preferred way. When connecting nodes with an edge are selected, nodes with high degree of intensity in the network are more likely to be selected, and the probability of node selection is [5]:

$$\prod n \leftarrow i = \frac{S_{ii}}{\sum\limits_{j\in t(ii)} S_{ji}} \tag{1}$$

In formula (1), $t(ii)$ represents the set of all nodes that pointed to node i; $S_{ii} = \sum\limits_{j\in t(ii)} w_{ji}$ represents the input strength of node i, that is, the sum of all input edge weights of node i.

Similarly, when selecting a node with a chain out of an edge, nodes with high strength out of the network are more likely to be selected. The probability of node i being selected can be expressed as follows:

$$\prod n \leftarrow i = \frac{S_{io}}{\sum\limits_{j\in t(io)} S_{jo}} \tag{2}$$

In formula (2), $t(io)$ represents the set of all nodes that pointed to node i; $S_{io} = \sum\limits_{j\in t(io)} w_{ij}$ represents the outgoing strength of node i, namely the sum of all outgoing edge weights of node i.

Step 4: Dynamic evolution of network weights, set to w_0 for each new edge added to the network. Suppose that every time a new edge is generated in the network, the edge weight between the node and its neighboring nodes will change accordingly, and the intensity of access between the node and its neighbor nodes will also change. The topology network model is constructed by updating the network weight according to the extra traffic load brought by the newly added network node connection edge. The aggregation coefficients of nodes in the topology network model are calculated.

2.2. Network node aggregation coefficient calculation

In topological network model, nodes show clustering characteristics. According to the clustering characteristics of nodes, clustering coefficient is divided into point clustering coefficient and edge clustering coefficient. The point aggregation coefficient of the node is calculated as follows [6]:

$$C(u) = \frac{2E_u}{d_u(d_u-1)} \tag{3}$$

In formula (3), d_u represents the degree of node u; E_u represents the number of edges existing in a small network of node u and d_u neighboring nodes of the node. By calculating the aggregation coefficient C of each node in the network and calculating the average value of them, the average aggregation coefficient of the whole network can be obtained, which reflects the degree of clustering among the interacting nodes to a certain extent.

The concept of side aggregation coefficient is developed on the basis of point aggregation coefficient. The side aggregation coefficient is defined as the ratio between the actual number of triangles formed by side uv and the maximum number of possible triangles formed by side uv in the current network. The calculation formula is as follows [7]:

$$ECC(u,v) = \frac{t_{u,v}}{\min(d_u-1, d_v-1)} \tag{4}$$

In formula (4), $t_{u,v}$ represents the actual number of triangles formed by sides uv in the network; d_u represents the degree of node u; $\min(d_u-1, d_v-1)$ refers to the degree of node u and node u reduced by 1, which is smaller, that is, the maximum number of triangles formed by side uv in the network. The edge aggregation coefficient characterizes the degree of close connection between the front two endpoints and the surrounding nodes. The larger the edge aggregation coefficient is, the more likely the edge is to exist in the community structure of the network. After calculating the clustering coefficient of network nodes, the k-nearest neighbor principle is used to identify the key nodes.

2.3. Key nodes identification in topological network

The clustering coefficient of the nodes in the topology network is taken as the corresponding attribute of the nodes and the k-neighbor principle is used to identify the key nodes in the topology network. In this paper, Manhattan distance is used to calculate the similarity between nodes of topological network, and the distance of each feature is multiplied by its accuracy. The concept and calculation formula of Manhattan distance [8]:

Suppose two nodes A and B in the complex network, $A = (f_{a1}, f_{a2}, \mathbf{L}, f_{an})$ means node A has N-dimensional characteristics, similarly, $B = (f_{b1}, f_{b2}, \mathbf{L}, f_{bn})$ means node B has N-dimensional characteristics. Manhattan distance is used to calculate the distance between nodes A and B. The calculation formula is as follows [9]:

$$dis(A,B)=\sum_{i=1}^{n}\left|f_{ai}-f_{bi}\right| \tag{5}$$

The idea of Bootstrap resampling was introduced and the characteristics and samples were resampled respectively. Bootstrap sampling enables KNN algorithm to achieve the best integration effect. Let the eigenmatrix established in this paper be $FT=\left[ft_{ij}\right]_{M\times N}$, M represents the number of nodes, N represents the eigen-dimensions, and ft_{ij} represents the j th feature of the i th node. For the characteristic matrix FT in the training set, row sampling is carried out first, and then column sampling is carried out for the submodel's training sample. For the number of sub-models is N, repeat sampling from FT is needed N times. After the heavy sampling of Bootstrap, the sub-training samples obtained after sampling were respectively modeled, and N sub-models were built according to the N sub-training samples after sampling [10]. Finally, this paper summarized the results of Bootstrap-KNN for N times and got the final score after sorting. The score result is the key degree of topology network nodes, and the score is arranged in descending order to complete the identification of key nodes. At this point, the key node recognition method based on clustering coefficient is completed.

3. Simulation Experiment and Result Analysis

In the above part, the design of key node identification method based on clustering coefficient in topological network is completed. This section will apply this method in the computer simulation platform to carry out simulation test, and verify the feasibility and effectiveness of the key node identification method studied above.

3.1. Preparation of experimental data

This simulation experiment was carried out on four different topological networks, taking the four topological networks as experimental objects. These four networks are GAMA network, Usair97 network, Football network and yeast protein action network in DIP database. GAMA network is a signed graph describing the aggregation of the Gahuku-Gama system in the central highlands of eastern New Guinea, in which there are 668 network nodes and 306 key nodes. USair97 network is a network formed by 332 airports and routes between them, including 332 network nodes and 259 key nodes. Football Network is a network in which members of 22 football teams that participated in the World Championship in Paris in 1998 signed contracts with 35 countries, with 17,710 network nodes and 8,986 key nodes. Yeast protein interaction network is from the DIP database, which contains 5037 nodes and 22061 edge interactions. According to the DEG database, this network contains a total of 978 key proteins, 3322 non-key proteins, and the rest of the proteins are key unknown.

The number of key nodes and non-key nodes in the above experimental topological network was recorded, and the recorded value was taken as the basis for the calculation of experimental indexes in order to draw experimental conclusions.

3.2. Content of simulation experiment

This experimental study was completed in a computer simulation platform equipped with Windows 10 and configured with Intel E5-2609 V2 with a main frequency of 3.75ghz CPU. The key node identification method based on clustering coefficient is compared with the traditional key node identification method based on information entropy and the key node identification method based on degree centrality. The comparison index of this experiment is the sensitivity, accuracy and robustness of the recognition method when different key node recognition methods are applied to identify the experimental network.

The sensitivity of the recognition method is the ratio between the number of key nodes correctly identified and the total number of original key nodes. The accuracy is the ratio between the number of key nodes correctly identified in the experimental results and the total number of nodes in the experimental network. When comparing the robustness of the recognition methods, the robustness of the three key node recognition methods is judged by deliberately attacking the network, deleting the three key nodes identified by the three key nodes in turn, and measuring the running state of the network after being attacked. After deleting the nodes in the topological network according to the recognition results of each key node recognition method, the number of subgraph S after the failure of the network and the ratio of the size of the maximum connected subgraph G before and after the failure of the network are used to measure the running state of the topology network, so as to compare the robustness of the applied key node recognition method.

3.3. Simulation results and analysis

The accuracy and sensitivity results of the three key node recognition methods for the experimental network identification are shown in table 1, and the data in the table are processed and analyzed.

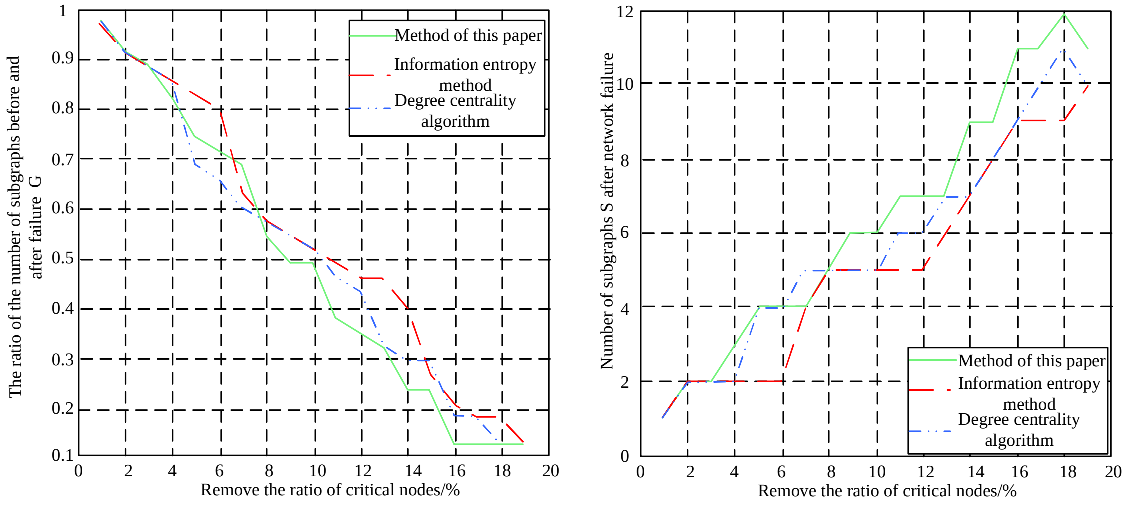
Table 1. Comparison of recognition accuracy and sensitivity of key node recognition methods

Key node identification method	GAMA net-	Usair97 network	Football network	Yeast protein action

		work			network
Method of this paper	Accuracy rate	96.2	98.3	96.6	97.5
	Sensitivity	94.2	94.3	94.1	94.3
The method based on information entropy	Accuracy rate	77.3	78.5	77.4	77.9
	Sensitivity	75.3	75.7	75.4	75.3
Recognition based on degree centrality algorithm	Accuracy rate	91.8	91.6	90.7	90.1
	Sensitivity	91.4	90.5	91.2	89.5

As can be seen from the above table, the accuracy and sensitivity of the key node recognition method studied in this paper are much higher than the two traditional methods in the recognition of the four experimental networks. The average recognition accuracy and sensitivity of the three key node recognition methods in this experiment were calculated. The average recognition accuracy of this method was 97.15%, and the average recognition sensitivity was 94.23%. The recognition method based on information entropy has an average recognition accuracy of 77.78% and an average recognition sensitivity of 75.43%. The average recognition accuracy and sensitivity of the algorithm based on degree centrality were 91.05% and 90.65% respectively. According to the above data, the key node identification method based on clustering coefficient is more accurate in identifying the key nodes.

In the yeast protein action network, the robustness of the three key node recognition methods was verified. The effect of removing the corresponding key nodes on the network operation state after the application of the three key node recognition methods. The relationship between the curves in the figure is analyzed and the final experimental conclusion is drawn as show in Figure 1.



(a)The effect of removing nodes on the metric G

(b)Eliminate the influence of nodes on index S

Figure 1. Shows the impact of removing nodes on network metrics

It can be seen from the change trend of G that the method proposed in this paper, the recognition method based on information entropy and the recognition method based on degree centrality algorithm can all cause the drop of G, but the method proposed in this paper can cause a larger drop of G. It can be seen from the change trend of S that the method proposed in this paper, the recognition method based on information entropy and the recognition method based on degree centrality algorithm can all cause the rise of S, but the method proposed in this paper can cause a more substantial rise of S. Based on the above experimental results, it can be concluded that deleting the key nodes identified by the proposed method in this paper in turn can cause a greater drop in G and a greater rise in S. Thus, it can be concluded that the efficiency of the proposed method in identifying key nodes is higher than that of the identification method based on information entropy and the identification method based on degree centrality algorithm.

To sum up, the clustering coefficient based recognition method in this paper has high accuracy and sensitivity and better application effect, so it can be applied to the recognition of key nodes in topological networks.

4. Conclusion

In this paper, the key node identification method based on clustering coefficient in topological network is studied. Through the comparison experiment with two commonly used key node identification methods, it is verified that the identification method studied in this paper is superior to the two methods compared in each index. At the same time, the method in this paper also has some shortcomings. Other algorithms need to be integrated to improve the recognition rate of the recognition method.

References

[1] Tian You. An algorithm of network node importance evaluation based on network structured index. Mobile Communications. 2018, 42(08), 62-66.

[2] Shao Hao, Wang Lunwen, Deng Jian. Important node identification method for dynamic networks based on H operation. Journal of Computer Applications. 2019, 39(9), 2669-2674.

[3] Jiang Yisen. Key Node Identification of navigation network based on structural centrality. Computer and Modernization. 2018(07), 108-113.

[4] Duan Xinyu, Jin Guoqing. Emergy evaluation of complex network center nodes under cloud computing. Computer Simulation. 2018, 35(11), 352-355.

[5] Chen Mili, Li Zijian, Deng Xigui, et al. Identification of key nodes in container maritime network based on information entropy and grey relational method. China Harbour Engineering. 2019, 39(09), 8-12.

[6] Zou Yanli, Yao Fei, Wang Yang, et al. Critical node identification for power systems based onnetwork structure and power tracing. Journal of Guangxi Normal University(Natural Science Edition). 2019, 37(01), 133-141.

[7] Deng Xiaoyi, Yang Yang, Jin Chun. Identifying influential nodes based on network topology. Operations Research and Management Science. 2019, 28(07), 91-99.

[8] Mo Zhenchun, Fu Lihua, Peng Yaohui, et al. Identification of ecological space network key nodes based on a comprehensive importance evaluation. Journal of Hunan University of Technology. 2018, 32(02), 64-69.

[9] Yang Gui, Zheng Wenping, Li Jinyu. Research on essential protein identification method based on multi attribute decision analysis. Journal of Shanxi University(Natural Science Edition). 2018, 41(01), 128-134.

[10] Long Jun, Wang Yulou, Yuan Xinpan, et al. A key node identification algorithm in P2P streaming networks. Computer Engineering and Science. 2019, 41(01), 56-64.