A sorting method of node based on Eigenvector and Closeness centrality

Jingcheng Zhu¹, Lunwen Wang¹* and Tao Wu¹
1 College of Electronic Engineering, National University of Defense Technology, Hefei, Anhui, 230000, China
*Corresponding author’s e-mail: wanglunwenmust@nudt.edu.cn

Abstract. Accurately identifying influential nodes in a complex network is of great significance to information dissemination. At present, researchers have put forward methods such as Degree centrality, Betweenness centrality, Closeness centrality and Eigenvector centrality, but these methods have certain limitations. There are many factors that affect the results of the node sorting, such as the number of neighbor nodes, node location information, etc. If multiple factors are combined, the proposed method can show more characteristics. First, closeness centrality of neighboring nodes is accumulated to reflect the influence of location information, and then integrated with Eigenvector centrality, the ECCN centrality is proposed to identify the network Influential nodes, this method integrates the path length and the number and quality of neighbor nodes. The SIR propagation model is used to simulate the influence of nodes on multiple real networks, and all centralities are compared with the results of the SIR model. Experimental analysis shows that the proposed method can more accurately identify influential nodes than other methods.

1. Introduction

In real life, we are surrounded by various complex networks, such as social networks [1], transportation networks [2], computer networks [3], and food chain networks [4]. Identifying influential nodes in the network is one of the important contents in complex network research. Many studies have shown that there are a small number of nodes that spread more widely during the propagation process. Therefore, accurate identification of the influential nodes in the network plays a vital role in information transmission and disease spreading.

At present, researchers have conducted a lot of research on identifying influential nodes, and proposed many classic centrality methods. Commonly used methods are degree centrality [5], betweenness centrality [6], and closeness centrality [7], eigenvector centrality [8], K-shell centrality [9] and clustering coefficient [10], etc. Degree centrality can be expressed by the number of directly connected neighbors of a node, which is simple and intuitive. But it has certain limitations because it only considers the local attributes of the node and does not consider the location information. Betweenness centrality and closeness centrality can be represented by the shortest path of the node, and its complexity is relatively large. Eigenvectors centrality considers both the number of neighbors of the node and the influence of quality on the importance of the node. K-shell centrality is to sort the nodes by gradually stripping off the nodes, but the node importance ranking obtained by this index is too coarse-grained, and it is impossible to distinguish the importance of the nodes at the same level [11]. In order to solve the problem of too coarse-grained, Bae et al. [12] proposed an improved coreness index, which introduced the coreness information of neighbors of the node to rank the influence of nodes, and
quantified the nodes more accurately. Berahmand et al. [13] proposed a new node influence recognition method that combines the node's own degree centrality, the clustering coefficient and the clustering coefficient of the node's secondary neighbors. In reference [14], a new strategy for searching influential nodes based on local topology and global location information is proposed from the basic factors affecting propagation. Experiments show that this method has a good balance between accuracy and network survivability.

Most of the above studies are based on the node's own information and location information to consider the influence of the node. This paper introduces the shortest path between the node and other nodes as an influencing factor. Firstly, the closeness centrality of the neighbor nodes is accumulated, and then the eigenvector centrality is combined with it. Finally, we propose the ECCN method which considers both the number and quality of neighbor nodes and the shortest path between nodes. In different networks, SIR model [15] is used for 1000 independent experiments to obtain the average value of the node's information dissemination ability. The proposed index is compared with the degree centrality [5], betweenness centrality [6], closeness centrality [7], eigenvector centrality [8] and the DCSC index [13]. The experimental results show that the ECCN index proposed in this paper is more accurate in evaluating the influence of nodes.

2. Method

Compared with other nodes, nodes in the center of the network need shorter time to reach all nodes. The closeness centrality takes the shortest path length between the node and other nodes into account and avoids the influence of extreme values by averaging. For an unweighted undirected network with $n$ nodes, the average distance from node $v_i$ to other nodes can be expressed as

$$d_i = \frac{1}{n-1} \sum_{j \in V \backslash \{i\}} d_{ij}$$

(1)

The $d_{ij}$ is the shortest path length of node $v_i$ and node $v_j$. The smaller $d_i$ is, the more central it is. To make the value positively correlated with importance, the reciprocal of closeness centrality $d_i$ is defined, so the closeness centrality of node $v_i$ is

$$CC(v_i) = \frac{1}{d_i} = \frac{n-1}{\sum_{j \in V \backslash \{i\}} d_{ij}}$$

(2)

Closeness centrality analyzes the importance of nodes from the perspective of the shortest path, but the evaluation of the importance of nodes from the node itself is not comprehensive enough. It is easy to ignore the structural information and cannot fully reflect the influence of nodes. Node propagation is based on the link between nodes as a medium to propagate along the connection path, so it is obviously unreasonable to ignore the connection information between nodes. In order to reflect the impact of node environmental information on node propagation, the method of accumulating neighbor's closeness centrality is used to correct node influence, which is expressed as

$$CCN(v_i) = \sum_{j \in A_i} CC(v_j)$$

(3)

In the sample network of Figure 1, the infection probability $\beta=0.325$ is set, and the SIR propagation model is used for 1000 independent simulation experiments. The average value of independent simulation results is used as the information propagation ability of the network with 11 nodes. The values of the top three nodes obtained by using closeness centrality are $CC(v_7)=0.5556$, $CC(v_4)=0.5263$, $CC(v_5)=0.5$. By accumulating neighbor closeness centrality $CCN$ method, the top three nodes are $CCN(v_4)=2.1545$, $CCN(v_5)=1.8113$, $CCN(v_7)=1.5025$. From the example network of Figure 1, we can see that the top three nodes are respectively $\Phi(v_4)=3.416$, $\Phi(v_5)=3.188$, $\Phi(v_7)=3.086$. In terms of the results, closeness centrality believes that node $v_7$ is more important than node $v_4$ and $v_5$, while the influence result obtained by SIR model is $v_4 > v_5 > v_7$, which is consistent with the conclusion obtained by CCN index. From this recognition result, we can see that CCN index is closer to centrality, which can better reflect the influence of nodes.
Considering the influence of the number and quality of neighbor nodes on node influence, this paper introduces the eigenvector centrality which can describe the long-term influence of nodes [16], combines the eigenvector centrality EC with the CCN which accumulates neighbor information, finally proposes a new ECCN method. In order to avoid the problem that the results are biased due to the large difference between the values of different indexes, the square and root numbers are normalized to make different indexes in a dimension, and the specific formula can be expressed as

\[ ECCN(v_i) = \frac{EC(v_i)}{\sqrt{\sum_{j=1}^{n}(EC(v_j))^2}} + \frac{CCN(v_i)}{\sqrt{\sum_{j=1}^{n}(CCN(v_j))^2}} \]  

(4)

It can be seen from equation (4) that this method combines the number and quality of node’s neighbors and the shortest path information of nodes, and the normalization processing makes the information fusion more reasonable.

3. Datasets and experimental analysis

3.1. Datasets

To evaluate the effectiveness of the node influence ranking method proposed in this paper, nine real networks with different structures are selected. The nine real networks are Karate Club Network [17], Dolphins Network [17], Jazz Network [18], Word Network [19], USAir Network [17], Netscience Network [19], Elegans network [20], Infectious network [21], Email network [21], the topological structure characteristics of the datasets are shown in Table 1. In the table, \( n \) and \( m \) respectively represent the number of nodes in the network and the number of edges in the network, \( \beta_{th} = <k>/<k^2> \) represents the propagation threshold of the network [22], \( <k> \) represents the average degree of nodes in the network, \( <k^2> \) represents the second-order average degree of network nodes, \( D \) is the network diameter, \( L \) is the average path length of the network [23].

| Networks   | \( n \) | \( m \) | \( \beta_{th} \) | \( <k> \) | \( D \) | \( L \) |
|------------|---------|---------|-----------------|--------|------|------|
| Karate     | 34      | 78      | 0.129           | 4.588  | 5    | 2.408|
| Dolphins   | 62      | 159     | 0.147           | 5.129  | 8    | 3.357|
| Word       | 112     | 425     | 0.073           | 7.589  | 5    | 2.536|
| Jazz       | 198     | 2742    | 0.026           | 27.697 | 6    | 2.235|
| USAir      | 332     | 2126    | 0.023           | 12.807 | 6    | 2.738|
| Netscience | 379     | 914     | 0.125           | 4.823  | 17   | 6.042|
| Elegans    | 453     | 2025    | 0.025           | 8.94   | 7    | 2.664|
| Infectious | 410     | 2765    | 0.053           | 13.488 | 9    | 3.631|
| Email      | 1133    | 5451    | 0.054           | 9.622  | 8    | 3.606|
3.2. Experiments and analysis

The experiments in this paper focus on the two aspects of the Kendall correlation coefficient [24] and the frequency of different values of nodes under each method. Equation (4) is used to calculate the node influence corresponding to the ECCN method proposed in this paper. And we also obtain other centrality results. The propagation threshold $\beta_{th}$ of the SIR propagation model is set as propagation rate $\beta$, and the calculation results of the propagation threshold retain a three decimal. The recovery rate is set to 1, which means that the infected node will be converted to the recovery state in the next step.

We use the Kendall coefficient to do comparative experiments on nine real networks, as shown in Table 2. From the experimental results in Table 2, the effect of betweenness centrality is the worst. The reason is that betweenness centrality emphasizes the pivotal role in the process of node path transmission, and ignores many nodes with weak pivotal effects. It can be seen from the table that the CCN method of accumulating neighbor closeness centrality is better than closeness centrality on the Kendall correlation coefficient, which also verifies that neighbor information has a certain effect on the influence of nodes. It can be seen in Table 2 that the ECCN method that integrates CCN and eigenvector centrality has the greatest correlation with the node influence, which demonstrates the accuracy of the method proposed in this paper.

| Networks      | $\beta_{th}$ | $\tau_{DC}$ | $\tau_{CC}$ | $\tau_{BC}$ | $\tau_{EC}$ | $\tau_{DCSC}$ | $\tau_{CCN}$ | $\tau_{ECCN}$ |
|---------------|--------------|-------------|-------------|-------------|-------------|---------------|--------------|--------------|
| Karate        | 0.129        | 0.6774      | 0.7130      | 0.5882      | 0.8717      | 0.7807        | 0.7843       | 0.9091       |
| Dolphin       | 0.147        | 0.7869      | 0.6293      | 0.5674      | 0.6695      | 0.7916        | 0.8731       | 0.8842       |
| Word          | 0.073        | 0.8343      | 0.8544      | 0.6573      | 0.9003      | 0.7444        | 0.8847       | 0.9369       |
| Jazz          | 0.026        | 0.8175      | 0.7082      | 0.4715      | 0.8807      | 0.8004        | 0.8419       | 0.9363       |
| USAir         | 0.023        | 0.7365      | 0.7986      | 0.5206      | 0.8915      | 0.8149        | 0.8027       | 0.9021       |
| Netscience    | 0.125        | 0.6071      | 0.3309      | 0.3115      | 0.3411      | 0.7554        | 0.6931       | 0.7803       |
| Infectious    | 0.053        | 0.7488      | 0.5939      | 0.3785      | 0.6544      | 0.6508        | 0.8097       | 0.8671       |
| Elegans       | 0.025        | 0.6557      | 0.6575      | 0.4914      | 0.8294      | 0.6754        | 0.7632       | 0.8627       |
| Email         | 0.054        | 0.7673      | 0.8134      | 0.6257      | 0.8795      | 0.7912        | 0.8116       | 0.8943       |

In the process of using various methods to get the sorting results of network nodes, there will be the same value of multiple nodes. Taking the great practical significance of distinguishing the influence of the top-ranking nodes into account, this paper conducts experiments on Karate, Word, Elegans and Email networks with the frequency of occurrence of the same values. As shown in Figure 2, the abscissa is the ranking result of ranking the influence of nodes calculated by centrality from large to small, and the ordinate shows the proportion of nodes with the same value in the corresponding ranking. It can be seen from Figure 2 that the DC and CC methods have poor effect on distinguishing the importance of nodes. In particular, DC only considers the edge information, so that many ranking nodes have the same value, and the distinguishing effect is poor. BC in the front ranking part of the discrimination is better. The main problem is that there are a large number of unobvious nodes that cannot be sorted. It can be seen in Figure 2 that the repetition rate of CCN centrality in the same ranking is basically lower than that of closeness centrality, especially in the top part of the ranking. It can be concluded that the CCN method with accumulated neighbor information is superior to the CC centrality in discrimination. In summary, it can be seen that the distinguishing effect of feature vector centrality EC and the proposed ECCN method is the best from the ranking frequency diagram of the four networks. The figure shows that the ECCN method will appear the same value only in the nodes with very low ranking, which has little influence on the ranking results. Therefore, it has the ability to effectively distinguish the influence of nodes.
Figure 2. Frequency of same ranking nodes in four networks.

In order to more intuitively show the influence of the change of propagation rate $\beta$ on the index recognition effect, a certain range of values near the propagation threshold are taken as the propagation rate for experiments, as shown in Figure 3. From the experimental results, the Kendall correlation coefficient of the betweenness centrality is the lowest, and the Kendall correlation coefficient of the CCN method is basically better than that of the near centrality index under each propagation rate, which reflects its advantages in the ranking of node influence. When the probability of spread is small, the infected nodes are often limited to a local area. At this time, the number of infected nodes is easily affected by the degree of nodes, and the ECCN index is not completely superior to other methods. When the propagation probability is near the propagation threshold, the ECCN method has a higher correlation with the result obtained by the SIR model than other methods. Therefore, we can get that it has certain advantages for node evaluation. From an overall point of view, the CCN index is superior to the closeness centrality in the evaluation of influence, and the ECCN index can better evaluate the influential nodes than other methods.

4. Conclusion

The identification of node influence in complex networks is a problem widely concerned by many researchers. The combination research of each method is helpful to show more attributes of nodes in the network, and is more helpful to clarify the structure and function of the network. Accurate identification of influential nodes in the network has practical significance for controlling rumor propagation and disease prevention. This paper considers the influence of the number and quality of neighbor nodes and the influence of the path length, and uses the closeness centrality of accumulated node neighbors to form the CCN method. The ECCN method is proposed by combining the normalized eigenvector centrality with the normalized CCN method. In this paper, the discrimination experiment is carried out on the real networks, and the results are also compared with those obtained by the SIR model. The Kendall correlation coefficients with different propagation rates are studied and analyzed. The results show that
the proposed method in this paper can more accurately identify influential nodes. The next research will focus on the recognition effect in the directed network and how to better combine the methods.

Figure 3. Kendall correlation coefficient between node influence under different propagation rate $\beta$ and each method.

References

[1] Arularasan, A.N., Suresh, A., Seerangan, K. (2019) Identification and classification of best spreader in the domain of interest over the social networks. Cluster Comput, 22: 4035–4045.

[2] Zhao, S.M., Zhao, P.X., Cui, Y.F. (2017) A network centrality measure framework for analyzing urban traffic flow: A case study of Wuhan, China. Physica A, 478: 143–157.

[3] Yao, Y., Zhang, R., Yang, F., Tang, J., Yuan, Y., Hu, R. (2018) Link prediction in complex networks based on the interactions among paths. Physica A, 510: 52–67.

[4] Li, Z.C., Ji, L.X., Liu, S.X., Pu, N.M. (2020) Link prediction method based on topological stability in directed network. Application Research of Computers, 37: 3744-3748.

[5] Freeman, L.C. (1978) Centrality in social networks conceptual clarification. Soc Networks, 1: 215-239.

[6] Freeman, L.C. (1977) A Set of Measures of Centrality Based on Betweenness. Sociometry, 40: 35-41.

[7] Sabidussi, G. (1966) The centrality index of a graph. Psychometrika, 31: 581-603.

[8] Borgatti, S.P. (2005) Centrality and network flow. Social Networks, 27: 55-71.

[9] Kitsak, M., Gallos, L. K., Havlin, S., et al. (2010) Identification of influential spreaders in complex networks. Nat. Phys., 6: 888–893.

[10] Ren, Z.M., Shao, F., Liu, J.G., et al. (2013) Node importance measurement based on the degree and clustering coefficient information. Acta Physica Sinica, 62: 522-526

[11] Wang, K.L., Wu C.X., Ai J., Su, Z. (2019) Complex network centrality method based on multi-order K-shell vector. Acta Physica Sinica, 68: 235-245.
[12] Bae, J., Kim, S. (2014) Identifying and ranking influential spreaders in complex networks by neighborhood coreness. Physica A, 395: 549–559.

[13] Berahmand, K., Bouyer, A., Samadi, N. (2018) A new centrality measure based on the negative and positive effects of clustering coefficient for identifying influential spreaders in complex networks. Chaos Solitons Fractals, 110: 41–54.

[14] Chen, G., Zhou, S., Liu, J., Li, M., Zhou, Q. (2020) Influential node detection of social networks based on network invulnerability. Phys. Lett. A, 384: 126879.

[15] Malavika, B., Marimuthu, S., Joy, M., et al. (2021) Forecasting COVID-19 epidemic in India and high incidence states using SIR and logistic growth models. Clin. Epidemiology Glob. Health, 9: 26–33.

[16] Lü, L., Chen, D., Ren, X. L., Zhang, Q.M., Zhang, Y.C., Zhou, T. (2016). Vital nodes identification in complex networks. Phys. Rep., 650: 1–63.

[17] Batagelj, V., Mrvar, A. (2006): Pajek datasets. http://vlado.fmf.uni-lj.si/pub/networks/data

[18] Gleiser, P. M., Danon, L. (2003) Community structure in jazz. Adv. Complex Syst., 6: 565-573.

[19] Newman, M.E.J. (2006) Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E, 74: 036104.

[20] Duch, J., Arenas, A. (2005) Community detection in complex networks using extremal optimization. Phys. Rev. E, 72: 027104

[21] Rossi, R.A., Ahmed, N.K. (2015). The network data repository with interactive graph analytics and visualization. In AAAI’15 Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin Texas. pp. 4292–4293.

[22] Castellano, C., Pastor-Satorras, R. (2010) Thresholds for epidemic spreading in networks. Phys. Rev. Lett., 105: 218701.

[23] Zhang, Q., Xue, Y., Wang, D., Niu, M. (2019) Asymptotic formula on average path length in a hierarchical scale-free network with fractal structure. Chaos Solitons Fractals, 122: 196–201.

[24] Knight, W.R. (1966) A Computer Method for Calculating Kendall’s Tau with Ungrouped Data. J. Am. Stat. Assoc., 61: 436–439.