BRING PageRank TO THE INFRASTRUCTURE NETWORK

Xiufeng Meng

Shanxi Institute of Energy, Jinzhong 030600, China

Zequan Li[†], Zhao Yang

School of Management, North China Institute of Science and Technology
Beijing 101601, China

(Received January 15, 2018; accepted April 23, 2018)

The PageRank algorithm was introduced to rank the important nodes on an actual air traffic control network. In order to verify its ability to identify the important nodes in the network, PageRank algorithm was compared with the other three node ranking methods and simulation verification was conducted via the disaster spreading dynamics model. The results showed that as compared with the other three ranking methods, PageRank algorithm had a stronger ability to identify important nodes, so it could be used to rank the important nodes on infrastructure networks.

DOI:10.5506/APhysPolB.49.1497

1. Introduction

It is well-known that almost all of the complex systems such as biology, society, nervous system, computer network, traffic and transportation [1–6] can be represented by network in which the node represents each entity of the system, and the edge between nodes represents the relationship between the entities. Similarly, infrastructures including electricity, communications, water supply, gas supply, aviation and roads [7–11] can also be represented as complex networks and then their characteristics are studied via network topology and dynamics feature.

Many scholars [10, 12] studied the statistical characteristics and dynamics process of these infrastructure networks. The results showed that the brokenness of network under disturbance was varied to some extent due to

[†] Corresponding author: lzquancumtb@126.com

the different position or location of nodes in the real network. Therefore, it was of great significance in the research on network survivability to rank the node importance in the infrastructure network and to identify the key nodes in the network. This work is called optimal percolation in the complex network, and currently it becomes one of the important research directions in network science [13, 14].

At present, the ranking methods for important nodes in the network mainly include degree centrality [15], k-shell decomposition method [16], information index [17], betweenness centrality [18] and their weighting methods. It can be seen that there are many assessment methods for important nodes in the network and their focuses are different. In order to facilitate the selection of suitable methods for practical issues, Lü *et al.* [19, 20] systematically analyzed over 30 representative mining methods for important nodes in the complex network field, and these methods were classified into four categories, namely, ranking methods based on node neighbor, on path, on feature vector, and on removal and contraction.

As compared with other ranking methods, the method based on feature vector not only takes into account the neighbor numbers of node, but also considers the impact of its quality on the node importance, thus it has attracted wide attention both in theory and in commerce in recent years. In particular, PageRank, the core algorithm of Google search engine, has been widely used in the field of web page ranking, moreover, many scholars have also introduced it to other aspects, such as identifying leaders in social networks [21], scientific papers citation analysis [22], assessment of node importance in the water network [10].

Shanxi Dashu Network project was selected as the network background, the literature [10] studied the shortcomings of four single-index ranking methods (degree centrality, closeness centrality, betweenness centrality and k-shell decomposition) for node importance in water network, and the multi-attribute decision-making method based on TOPSIS was proposed. The direction and the level difference of water network were taken into account, and PageRank algorithm was used to assess the importance of nodes with directed weight in water network. Perhaps, in the literature, it is rare to find that PageRank algorithm is used to rank the important nodes in infrastructure networks like water network, and the findings in literature [10] are not verified by relevant models. In order to solve this issue, four ranking methods for important node were selected for node ranking comparison in this paper, and the disaster spreading model was used for simulation verification.

This paper was arranged as follows: Section 2 presents an actual infrastructure network, and briefly introduces the research ideas and methods of this paper; Section 3 introduces a universal disaster spreading dynamics model; Section 4 mainly analyzes the simulation results; Section 5 proposes some important research conclusions.

2. Network and methods

The infrastructure network used in this paper was Air Traffic Control (US Air Traffic Control Network, hereinafter referred to as ATC) which was taken from the Federal Aviation Administration National Flight Data Center (hereinafter referred to as FAA-NFDC) in the United States. In this network, a node represented an airport or service center, and edge represented the preferred flight routes recommended by the NFDC. The ATC network was a directed and scale-free network with 1,226 nodes and 2,615 edges. The maximum degree of nodes in the network was 37 and the power law constant was 3.7.

As mentioned above, literature [19] analyzed the common methods used by academia and industry to rank important nodes in network, and four basic types were summarized. Its idea was referred in this paper, and one method was selected from four types respectively, namely, degree centrality (ranking method based on node neighbor), betweenness centrality (ranking method based on path), residual closeness centrality (ranking method based on node removal and contraction) and PageRank method (ranking method based on feature vector). The brief introduction for these four methods was as follows.

2.1. Degree centrality

The importance of a node is considered equivalent to the ability of such node to establish direct contact with its surrounding nodes, which is defined as the number of edges of a node, denoted as

$$\mathrm{DC}(i) = \frac{k_i}{n-1} \tag{1}$$

in which $k_i = \sum_j a_{ij}$, a_{ij} is the element in line i and column j of adjacent matrix A of network, n is the number of nodes in the network, $n-1$ is the possible maximum of edges of the node.

2.2. Betweenness centrality

The importance of a node is considered to be described through the size of the information or energy loaded in the node, that is, the more the shortest paths passing through the node, the more important it is, given as

$$\mathrm{BC}(i) = \frac{2\sum_{j<k} g_{jk}(i)}{(n-1)(n-2)g_{jk}} \tag{2}$$

in which g_{jk} is the number of shortest paths between node j and node k, $g_{jk}(i)$ is the number of shortest paths passing node i between node j and node k, $(n-1)(n-2)/2$ is the maximum possible node betweenness.

2.3. Residual closeness centrality

If the deletion of a node increases the vulnerability of the network, such node is considered more important. This is used to measure the impact of node removal on the network, denoted as

$$\mathrm{RCC}(i) = \sum_{j} \sum_{k \neq j} \frac{1}{2^{d_{jk}(-i)}} \tag{3}$$

in which $d_{jk}(-i)$ is is the shortest distance between node j and node k after node i is deleted.

2.4. PageRank algorithm

It was originally mainly used for web page ranking. The importance of a page is considered to depend on the quantity and quality of other pages directing to it. If a page is directed to by many high-quality pages, the quality of such page is also high, given as

$$\mathrm{PR}_i(t) = (1-c)\sum_{j=1}^{n} a_{ji} \frac{\mathrm{PR}_j(t-1)}{k_j^{\mathrm{out}}} + \frac{c}{n} \tag{4}$$

in which k_j^{out} is the out-degree of node j, c is the probability of random skip, which is generally taken at 0.15. As the most classic algorithm for ranking of nodes in directed network, PageRank and its improved algorithms have been widely used in the fields of journal ranking, social online user ranking and scientist influence ranking and so on.

In addition, the research ideas of this paper were described as follows: first, all the nodes in the ATC network were ranked by the four ranking methods mentioned above. Then, the attack stimulation was conducted via disaster spreading dynamics model on the nodes of the first five (Top-5),

the first ten (Top-10), and the first twenty (Top-20) selected respectively, and the simulation time was 20 steps. Finally, the total number of broken nodes in a network at a certain time step after the phase was compared. It can be seen from the simulation results that, when $t = 10$, it had reached equilibrium.

3. Model

Traditionally, the epidemic models (namely, SIS model and SIR model, such as virus spreading on communication network [23], successive failure on power network [24]) were used to assess the mining algorithm of various node importance. However, as for an infrastructure network like a power system, an epidemic model could not effectively describe the spreading of disasters on the network. It is well-known that the spreading of information and viruses on the network is very different. The spreading of the virus requires physical contact, and literature [25] can be referred for a detailed discussion on this issue. Perhaps, since Buzna *et al.* [6] proposed the disaster spreading dynamics model, many scholars turned to this model for disaster spreading research on infrastructure networks. Therefore, this dynamic model was also used as the assessment and verification model for ranking algorithm in this paper.

For infrastructure networks, we mainly focus on its vulnerability, namely, the spreading speed or scope of broken status or disaster after the breaking of a (some) node(s) on the network. Based on this, Buzna *et al.* established a universal disaster spreading dynamical model to simulate the spreading process of disasters on the network.

A given directed network $G=(N,S)$ contains nodes $i \in N := \{1,2,\ldots,n\}$ and edges $(i,j) \in N \times N$, which represent the mutual relation between the system node and various nodes; x_i represents the attribute value of the node, when $x_i = 0$, it means the node is in a steady state, when x_i deviates from zero, it means the node is broken. Therefore, dynamics model for the node against the time evolution can be expressed as

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = -\frac{x_i}{\tau} + \Theta(x_i)\left(\sum_{j\neq i}\frac{M_{ij}x_j\,(t-t_{ij})}{f(o_i)}e^{-\frac{\beta t_{ij}}{\tau}}\right) + \xi_i(t)\,, \tag{5}$$

$$\Theta(x_i) = \frac{1-\exp(-\alpha x_i)}{1+\exp(-\alpha(x_i-\theta_i(t)))}\,, \tag{6}$$

$$f(o_i) = \frac{ao_i}{1+bo_i}\,. \tag{7}$$

This dynamics equation includes three parts: the first item on the right-hand side of equation (5) represents the self-healing function of the node;

the second item represents the disaster spreading mechanism of the node; and the third item represents the internal random noise of the node. $1/\tau$ is the self-healing speed of nodes, M_{ij} is the influence degree of node i on node j, t_{ij} is the influence delay time between node i and node j, and β is the damping effect during spreading. Equation (6) is the Sigmoid function, α is a fixed value, θ_i is the threshold of node i. Equation (7) is the out-degree function of node i, reflecting the influence degree of node i on other nodes, o_i is the out-degree value, a and b are fixed values.

4. Results

According to the above ideas, the ranking results of four ranking methods were given, see Table I for details. It can be seen that the ranking overlapping ratio was high on the degree centrality and betweenness centrality for the first 20 important nodes, but residual closeness centrality and PageRank algorithm had very little overlap with the former two, especially for the PageRank algorithm, the value of the first five important nodes were completely different from those of the other three methods.

TABLE I

Node ranking results by four methods (Top-20).

Ranking	Degree centrality	Betweenness centrality	Residual closeness centrality	PageRank
1	67	67	68	312
2	51	211	52	61
3	43	311	213	105
4	109	51	689	19
5	112	134	522	842
6	134	212	291	187
7	45	522	617	578
8	603	121	358	86
9	211	147	220	52
10	212	688	690	311
11	147	220	136	38
12	26	109	521	26
13	5	357	308	68
14	522	603	362	201
15	357	118	250	306
16	305	290	454	109
17	66	43	148	44
18	220	616	1149	157
19	52	305	763	89
20	688	112	81	266

In order to get a better understanding on node features under PageRank algorithm, the schematic diagram for the neighbor numbers and the quality of three nodes were given here, as shown in figure 1. It can be seen that, in addition to more edge numbers, the top ranked nodes were prominent in the neighbor quality, in other words, their neighbors had more neighbors. In principle, the PageRank algorithm is a method based on feature vector, which not only takes into account the neighbor numbers, but also consideres the impact of the neighbor quality on the node importance. It was exactly fitted with the characteristics shown in figure 1.

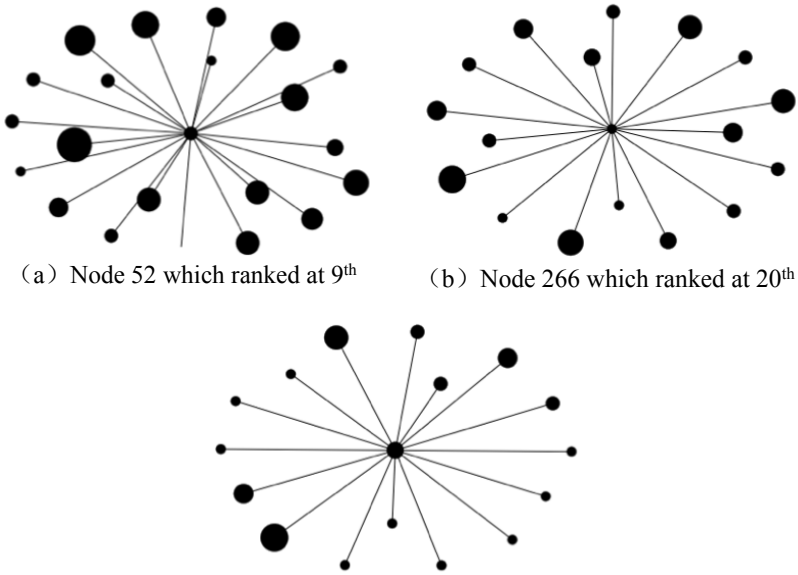
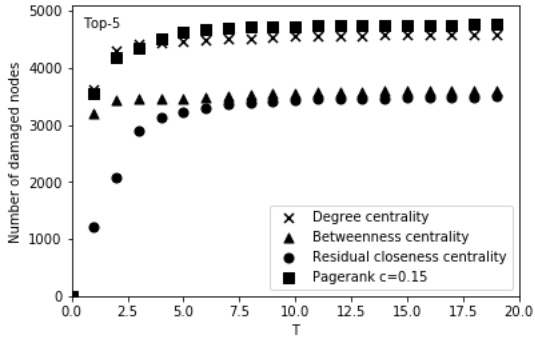


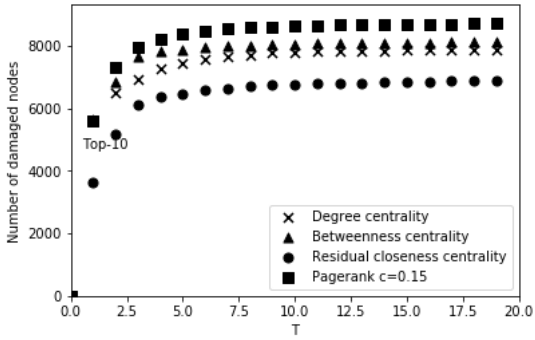
Fig. 1. Quality of neighbor of nodes with PageRank algorithm. (The size of circles represents the number of edges.)

The ranking results for important nodes by four algorithms were shown in Table I. The analog stimulation was conducted via the above disaster spreading dynamics model to assess and verify the ranking results of the algorithm, and the specific results were shown in figure 2. In Fig. 2 (a), the analog stimulation results of the PageRank algorithm and the degree centrality method were basically similar. As compared with other methods, especially the degree centrality ranking method, the advantage of PageRank algorithm was not obvious. However, as the number of ranking nodes increased, as shown in (b) and (c) of figure 2, the cumulative number of broken nodes in the network was gradually increased after the analog simulation with the ranking nodes obtained by PageRank. For example, PageRank far

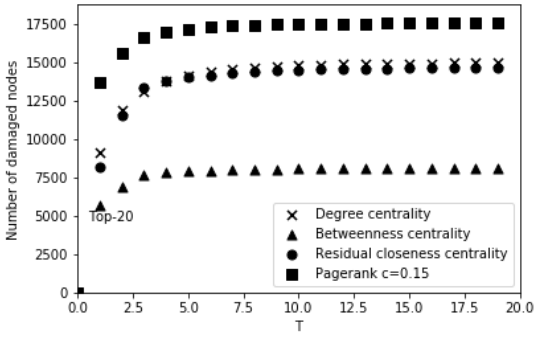
surpassed betweenness centrality in Fig. 2 (c). Therefore, it was indicated that the ranking nodes solved by the PageRank algorithm had a greater impact on the network.



(a) Top-5 important nodes



(b) Top-10 important nodes



(c) Top-20 important nodes

Fig. 2. The cumulative number of damaged nodes in three statistical ways.

5. Conclusion

In this paper, the ranking results for important nodes by four ranking algorithms (degree centrality, betweenness centrality, residual closeness centrality and PageRank) were compared in an actual air traffic infrastructure network (U.S. Air Traffic Control), and the stimulation verification was conducted via disaster spread dynamics model. The results showed that the PageRank algorithm can be used to rank the important nodes on the infrastructure network. As compared with other ranking methods, the verification results showed that the number of its final broken nodes in the network was always the highest, indicating that the PageRank algorithm had a better ability to identify important nodes than other methods.

REFERENCES

[1] G. Bianconi, A.L. Barabási, *Phys. Rev. Lett.* **86**, 5632 (2001).
[2] J.M. Beggs, D. Plenz, *J. Neurosci.* **23**, 11167 (2003).
[3] R. Lambiotte *et al.*, *Physica A* **387**, 5317 (2008) [arXiv:0802.2178 [physics.soc-ph]].
[4] L. Valentini, D. Perugini, G. Poli, *Physica A* **377**, 323 (2007).
[5] R. Albert, H. Jeong, A.L. Barabási, *Nature* **401**, 130 (1999).
[6] L. Buzna, K. Peters, D. Helbing, *Physica A* **363**, 132 (2006).
[7] P. Ormerod, A.P. Roach, *Physica A* **339**, 645 (2004).
[8] K. Sun, Z.X. Han, Y.J. Cao, *Power Sys. Technol.* **29**, 1 (2005).
[9] J. Guo, D.L. Wang, *Tele. Ele. Pow. Sys.* **30**, 6 (2009).
[10] Z.H. Liu, H. Yu, F.T. Yang, *Sci. Sin. Technol.* **44**, 1280 (2014).
[11] W.J. Liu, B.H. Wang, T. Zhou, *Pro. Nat. Sci.-Mater* **18**, 601 (2008).
[12] Z.Q. Li, R.X. Zhang, Z. Yang, *Acta Phys. Sinica* **61**, 238902 (2012).
[13] A.E. Mother, Y.C. Lai, *Phys. Rev. E* **66**, 065102 (2002).
[14] Y. Tan, J. Wu, H. Deng, *Sys. Eng. Theor.* **11**, 79 (2006).
[15] L.C. Freeman, *Soc. Networks* **1**, 215 (1979).
[16] M. Kitsak, L.K. Gallos, *Nature. Phys.* **6**, 888 (2010).
[17] M. Altmann, *Soc. Networks* **15**, 1 (1993).
[18] E. Estrada, V.J.A. Rodriguez, *Phys. Rev. E* **71**, 056103 (2005).
[19] X.L. Ren, L.Y. Lü, *Chin. Sci. Bull.* **59**, 1175 (2014).
[20] L.Y. Lü *et al.*, *Phys. Rep.* **650**, 1 (2016) [arXiv:1607.01134 [physics.soc-ph]].
[21] L.Y. Lü, Y.C. Zhang, C.H. Yeung, T. Zhou, *PLOS ONE* **6**, e21202 (2011).
[22] N. Ma, J.C. Guan, *Inf. Pro. Man.* **44**, 800 (2008).
[23] X.L. Peng, X.J. Xu, X. Fu, *Phys. Rev. E* **87**, 022813 (2013).
[24] C.D. Brummitt, R.M. D'Souza, E. Leicht, *Proc. Nat. Acad. Sci.* **109**, E680 (2012).
[25] L.Y. Lü, D.B. Chen, T. Zhou, *New J. Phys.* **13**, 123005 (2011).