WILEY | Hindawi

Research Article

Influential Spreader Identification in Complex Networks Based on Network Connectivity and Efficiency

Rong Qiang[1] **and Jianshe Yang**[2]

^{1}*School of Computer and Information Science, Southwest University, Chongqing 400715, China*
^{2}*Basic Medical School, Gansu Medical College, Pingliang 744000, China*

Correspondence should be addressed to Jianshe Yang; yangjs@impcas.ac.cn

Influential spreader identification is a vital research area in complex network theory, which has important influence on application and popularization. Each of the existing methods has its own advantages and disadvantages, and there are still various methods proposed to solve this issue. In this paper, we come up with a new centrality of influential spreader identification based on network connectivity and efficiency (CEC). The consequences of spreader deletion can be generally divided into two parts, one is that the connectivity of network topology is destroyed, and the other is that network's performance is degraded, which makes the network unable to meet the functional requirement. Therefore, the relative changes of connectivity and efficiency of network before and after removing spreaders are used to present the influence of spreaders. We adopt susceptible-infected (SI) model, a well-known infectious disease model, to verify the effectiveness of CEC through the spreading ability simulation of spreaders in actual networks. And the simulation results demonstrate the superiority of CEC.

1. Introduction

At present, complex networks are closely associated with our real life, for example, networks [1, 2], traffic systems [3, 4], power grids [5, 6], and ecological networks [7, 8]. Influential spreader identification remains an open and vital research issue that has attracted increasing attention, which helps to understand the structure of networks and control the propagation process. Some hazards caused by load propagation and cascading effect, for example, North American blackout and WannaCry's spread, often begin with a small portion of spreaders but spread rapidly to the entire network [9, 10]; this small portion of spreaders has a great impact on network. Therefore, accurate quantification and identification of influential spreaders is very important. For instance, we can effectively suppress the spread of the virus and prevent its large-scale outbreak by vaccinating key individuals in infectious disease network [11]. In power grids, we can effectively prevent the cascading failure with taking prior precau-

tions for circuits in vital areas [12]. In social network, such as MicroBlog and Twitter, we can control the dissemination of information to guide speech [13].

A variety of approaches have been proposed over the past few decades. Most of these approaches measure the influence of spreaders from the structural information of network.

There are a lot of methods proposed to search these key spreaders [14]. Degree centrality (DC) [15], one of the simplest and earliest methods, only counts the number of the directly connected spreaders and results in low complexity. Closeness centrality (CC) [16] measures spreader's capability to affect others through the network, while it will fail when applied to disconnected networks. Betweenness centrality (BC) [17] measures spreader's influence with the perspective from shortest path. Except these classical measures, some new methods have been proposed such as H-index centrality [18] and evidence theory [19]. Roberts et al. [20] suggested a centrality which considered the fourth-level

neighbors as a trade-off measure. However, these centralities ignore the connections among spreaders, and then, the ClusterRank [21] was proposed by taking the effect of clustering coefficient into consideration. What is more, Kitsak et al. [22] measured the influence of spreaders from location perspective and put forward a new method named K-shell decomposition (Ks). The spreaders were moved layer by layer based on continuously updated DC value. The biggest problem of Ks is the poor distinguish capacity of centrality value, i.e., poor monotonicity. Then, some approaches were put forward to solve this issue. Zeng and Zhang [23] came up with a MDD approach by considering the degree of initial spreaders and removed spreaders. Bae and Kim [24] summed the Ks value of neighbors to measure the importance of spreaders. In addition, there are also some approaches based on iteration such as PageRank [25], LeaderRank [26], and Hits [27]. Different centralities reflect the influence of spreaders from limited parts; some researchers have proposed multiattribute ranking approaches which combines several centralities to comprehensively rank the influence of spreaders. Liu et al. [28] proposed an improved Ks and used TOPSIS to fuse DC, CC, and BC and improved K-shell decomposition. Yang et al. [29] combined DC, CC, and BC with VIKOR method and adopted entropy weighting method to reasonably obtain the weights of attributes. Wen and Deng proposed a local information dimensionality (LD) to rank key spreaders [30]. Wang et al. focused on the contribution of spreaders to network efficiency and proposed EffC method to identify influential spreaders [31].

In this paper, we consider the importance of spreaders from global information perspective, and then, a novel centrality called connectivity and efficiency centrality (CEC) is put forward. The consequences of network spreaders removal can be generally divided into two aspects [32, 33], one is that the connectivity of network topology is destroyed, and the other is that the performance of the network is degraded, which makes the network unable to meet the service requirement. Therefore, we consider the relative changes of connectivity and efficiency of network before and after removing spreaders, and the combination of them is taken as an indicator to determine the influence of spreaders. Note that the removal of spreaders will also delete the links connected to them at the same time. To assess the effectiveness of CEC, we adopt susceptible-infected (SI) model [34] to measure spreading ability of spreaders in actual databases, and we compare the performance between CEC and others to verify the superiority of CEC.

2. Centralities

Given a network $G(V, E)$, where V and E, respectively, represent the set of spreaders and the set of edges, they meet $m = |E|$ and $n = |V|$. $A = (a_{ij})_{n \times n}$ indicates the adjacent matrix; if spreader i and spreader j are connected by edge $e_{ij} \in E$, $a_{ij} = 1$; otherwise, $a_{ij} = 0$.

Degree centrality [15], one of the simplest and earliest local centrality, only counts the number of the directly con-

nected spreaders and results in low complexity.

$$DC(i) = \sum_j^n a_{ij}. \quad (1)$$

Degree centrality indicates spreaders' ability to communicate directly with others.

Closeness centrality [16] considers the influence of spreaders based on the distance between them. It measures spreader's capability to affect others through the network.

$$CC(i) = \frac{n-1}{\sum_j^n d_{ij}}, \quad (2)$$

wherein d_{ij} represents the Euclidean distance between spreader i and spreader j. CC uses average transmission time of information to determine the influence of a spreader.

Betweenness centrality [17] measures spreader's influence with the perspective from shortest path. BC considers a spreader influential if it expressed as a "bridge."

$$BC(i) = \frac{1}{(n-1)(n-2)} \left(\sum_{s,t \neq i} \frac{g_{st}(i)}{g_{st}} \right), \quad (3)$$

wherein g_{st} represents the number of the shortest paths between spreader s and spreader t and $g_{st}(i)$ indicates the number of shortest paths passing through spreader i. BC can reflect the degree of independence between spreaders.

K-shell decomposition [22] measures the influence of spreaders from location perspective, which has important milestone significance. The spreaders were moved layer by layer based on continuously updated DC value.

3. The Proposed Centrality

We consider the influence of spreaders from global information perspective. The influence of spreaders can be measured by the relative changes of some global characteristic parameters of network before and after removing corresponding spreaders. The consequences of network spreader deletion can be generally divided into two parts, one is that the connectivity of network topology is destroyed, and the other is that the network efficiency is degraded, which makes the network unable to meet the service requirement. Both the two aspects should be taken into consideration to give comprehensive identification results.

Definition 1. The network connectivity represents the average influence of network to maintain connectivity, which is indicated as the mean value of the ratio of number of connected spreader pairs to the total number of spreader pairs in network.

$$CON(G) = \frac{1}{n(n-1)} \sum_{i \in G} \sum_{j > i} l_{ij}, \quad (4)$$

wherein l_{ij} represents the connection parameter from

TABLE 1: The top 10 spreaders using different centralities: karate club.

| Karate club | | | | | | |
Rank	DC	CC	BC	Ks	CEC	I(t)
1	34	1	1	1	1	1
2	1	3	34	34	34	34
3	33	34	33	33	3	3
4	3	32	3	31	33	33
5	2	9	32	3	32	2
6	4	14	9	14	2	9
7	32	33	2	2	14	14
8	9	20	14	4	9	32
9	14	2	20	8	4	4
10	24	4	6	9	20	20

TABLE 2: The top 10 spreaders using different centralities: Jazz musicians.

| Jazz musicians | | | | | | |
Rank	DC	CC	BC	Ks	CEC	I(t)
1	136	136	136	60	136	136
2	60	60	153	168	149	149
3	132	168	60	108	60	96
4	168	70	149	122	5	70
5	70	83	168	33	153	60
6	99	132	167	58	185	153
7	108	122	189	66	189	167
8	83	194	115	100	168	5
9	158	174	96	132	34	83
10	7	158	83	179	83	168

spreader i to spreader j; if they have a connected path, including directly connected path and indirectly connected path, then $l_{ij} = 1$; otherwise, $l_{ij} = 0$.

Definition 2. The residual network is denoted as G_k after removing spreader k from G, and the relative changes of network connectivity can be defined as

$$\text{RE\_CON}(k) = \frac{|\text{CON}(G_k) - \text{CON}(G)|}{\text{CON}(G)}. \tag{5}$$

Definition 3. The network efficiency refers to the effectiveness of information transmission on the network. It is denoted as

$$\text{EFF}(G) = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{d_{ij}}, \tag{6}$$

wherein d_{ij} refers to the shortest distance between spreader i and spreader j. Note that if spreader i and spreader j have no connected path, $d_{ij} = +\infty$ and $1/d_{ij} = 0$.

Definition 4. The residual network is denoted as G_k after removing spreader k from G, and the relative changes of network efficiency can be written as

$$\text{RE\_EFF}(k) = \frac{|\text{EFF}(G_k) - \text{EFF}(G)|}{\text{EFF}(G)}. \tag{7}$$

Definition 5. The proposed connectivity and efficiency centrality (CEC) can be defined as

$$\text{CEC}(k) = (1 + \text{RE\_CON}(k)) \times \text{RE\_EFF}(k). \tag{8}$$

The greater the value of $\text{CEC}(k)$, the more influential the spreader k.

TABLE 3: The top 10 spreaders using different centralities: USAir97.

| USAir97 | | | | | | |
Rank	DC	CC	BC	Ks	CEC	I(t)
1	118	118	118	67	8	118
2	261	261	8	94	261	67
3	255	67	261	109	118	261
4	152	255	201	112	13	8
5	182	201	47	118	201	182
6	230	182	182	131	152	313
7	166	47	255	146	182	201
8	67	166	152	147	313	152
9	112	248	313	150	67	248
10	201	112	13	152	258	258

TABLE 4: The top 10 spreaders using different centralities: email.

| Email | | | | | | |
Rank	DC	CC	BC	Ks	CEC	I(t)
1	105	333	333	299	333	105
2	333	23	105	389	58	333
3	16	105	23	434	355	135
4	23	42	578	552	578	42
5	42	41	76	571	105	3
6	41	76	233	726	21	52
7	196	233	135	756	270	21
8	233	52	41	788	376	355
9	21	135	355	885	42	233
10	76	378	42	886	233	270

4. Simulation and Analysis

4.1. Datasets. We choose four actual networks to conduct experiments and simulations, which cover multiple fields and network scales. (i) Karate club [35]: it is a widely used dataset describing the relationship between karate club
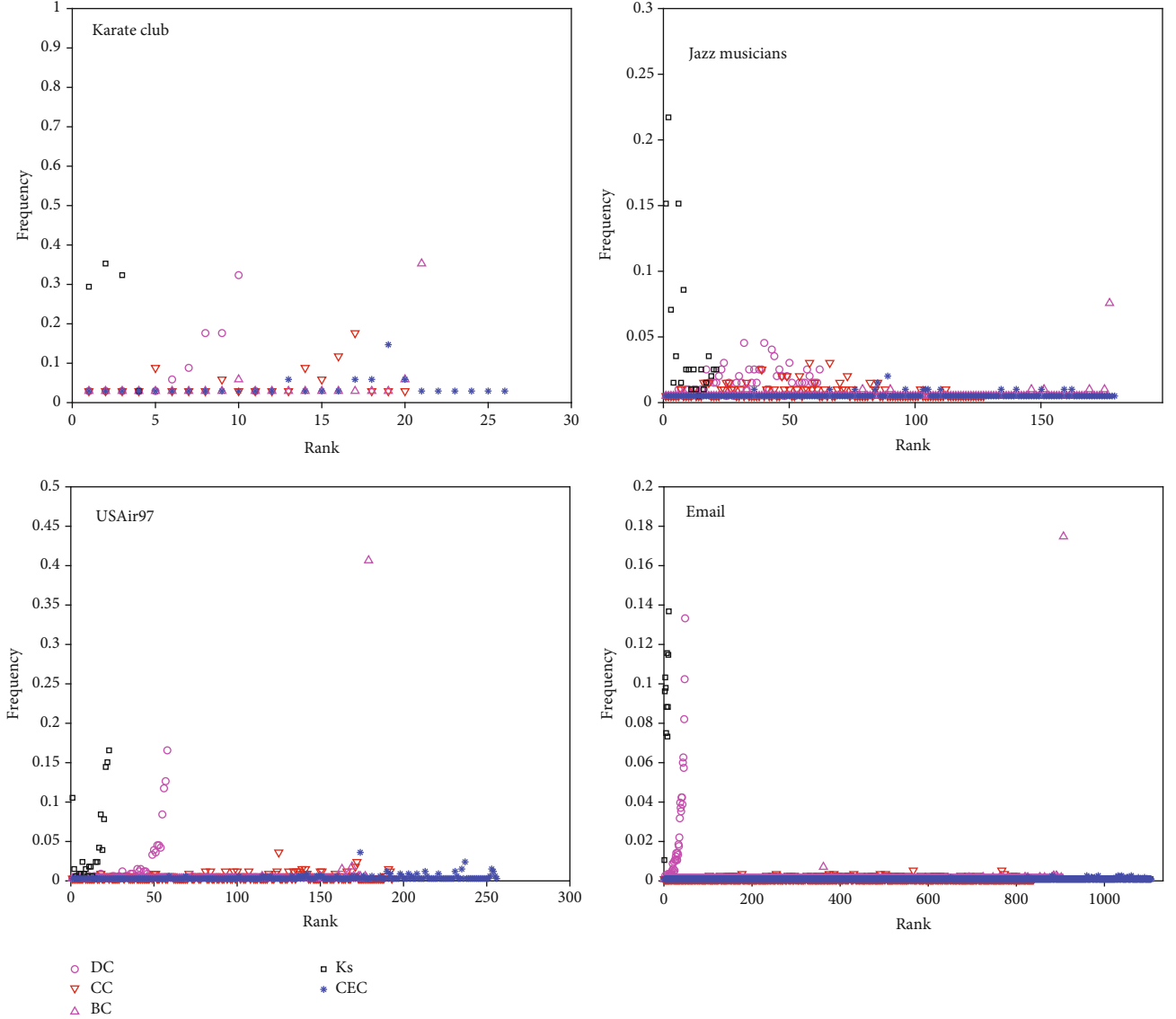
FIGURE 1: Example of a figure caption.

members. (ii) Jazz musicians [36]: it is a social dataset describing the cooperative relationship between jazz musicians. (iii) USAir97 [37]: it is a transportation dataset representing the airline relationship of American airports in 1997. (iv) Email: it describes the email exchange in a university.

4.2. Experiment and Analysis

4.2.1. Experiment 1: Comparison of Top 10 Spreaders Ranked by Different Centralities. The influence of each spreader in network is calculated using CEC and classical centralities. The actual spreading ability I(t) (t = 25) calculated by SI model is used as benchmark; the definition of I(t) will be introduced later. We pay attention to the top 10 spreaders sorted by several centralities. As shown in Table 1, in karate club network, the identification results of CEC and CC are the best due to their 10 same spreaders as I(t), and DC and BC have 9 same spreaders as I(t), while Ks owns 8 same spreaders. In Jazz musician network, shown in Table 2, there

are 5 same spreaders with I(t) in top 10 lists using DC and CC, while it is only 2 using Ks. CEC owned 7 same spreaders as I(t) performs slightly worse than BC. Besides, the top 2 spreaders of CEC are the same with I(t). In USAir97 network (Table 3), DC and CC both own 6 same spreaders; the number of same spreaders of CEC, BC, and Ks is 9, 7, and 3, separately. In email network, depicted in Table 4, CEC, CC, and BC have 6 same spreaders as I(t), which is lightly greater than DC, while there is no any same spreader between Ks and I(t). In a word, CEC has the most similar performance with actual ranking results; that is, CEC can identify spreaders more accurately.

4.2.2. Experiment 2: Comparison of Capability of Different Centralities to Distinguish Spreaders' Spreading Ability. When ranking the influence of spreaders, we find that some spreaders have the same centrality value and it is impossible to distinguish them. This phenomenon will reduce the accuracy of centrality. We consider the frequency of spreaders
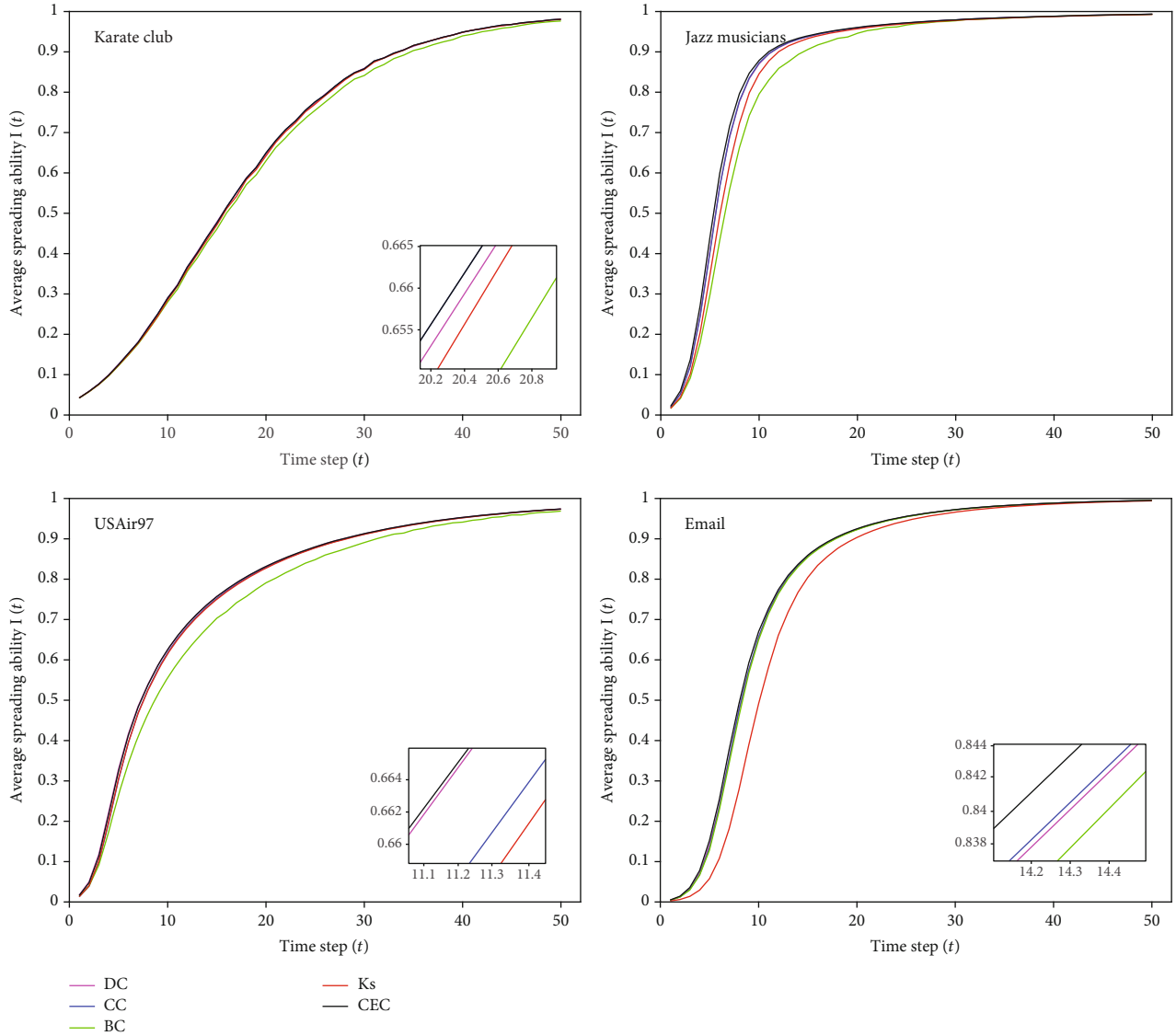
FIGURE 2: The average spreading ability of top 10 spreaders sorted with different centralities.

with same rank as an index to assess the distinguishing capability. The lower the frequency, the better the method. The experimental results of different centralities are shown in Figure 1. In the four networks, CEC has the lowest frequency; that is to say, CEC performs best in distinguishing spreaders' spreading ability. However, the frequency of DC and Ks is greater than other methods. The experimental result indicates the superiority of our method in distinguishing spreading ability.

4.2.3. Experiment 3: Comparison of the Average Spreading Ability of Top 10 Spreaders. We conduct transmission simulation with SI model [34] to examine the spreading ability of spreaders. We take spreader i as the source spreader and the spread process will start from the source spreader. The total number of infected spreaders will reach n_{it} after $t(t = 1, 2, \cdots)$ time step. Then, the spreading ability, denoted as $I_i(t) = n_{it}/n$, is expressed as ratio of infected spreaders to network size. And the average spreading ability of top 10 spreaders is represented as $I(t) = (\sum_{i=1}^{10} I_t(\text{t}))/10$. We set $t_{max} = 50$; the simulation results are presented in Figure 2.

From Figure 2, the average spreading ability of top 10 spreaders increases with t, and eventually almost the entire network is infected. In karate club network, we can see that the black curve and the blue curve overlap; that is to say, the spreading ability of CEC is the same with CC, because the top 10 spreaders of them are the same. It is clear that the spreading ability of CEC is superior to that of DC, BC, and Ks. In Jazz musician network, we can find that there are more infected spreaders of CEC than others, which demonstrates that the spreading ability of CEC is better than that of other methods. In USAir97 network, CEC is marginally better than CC, DC, and Ks, and BC is the poorest because the average number of infected spreaders of BC is much less than that of others. In email network, the number of infected spreaders at each step of CEC is marginally greater than DC, CC, and BC.
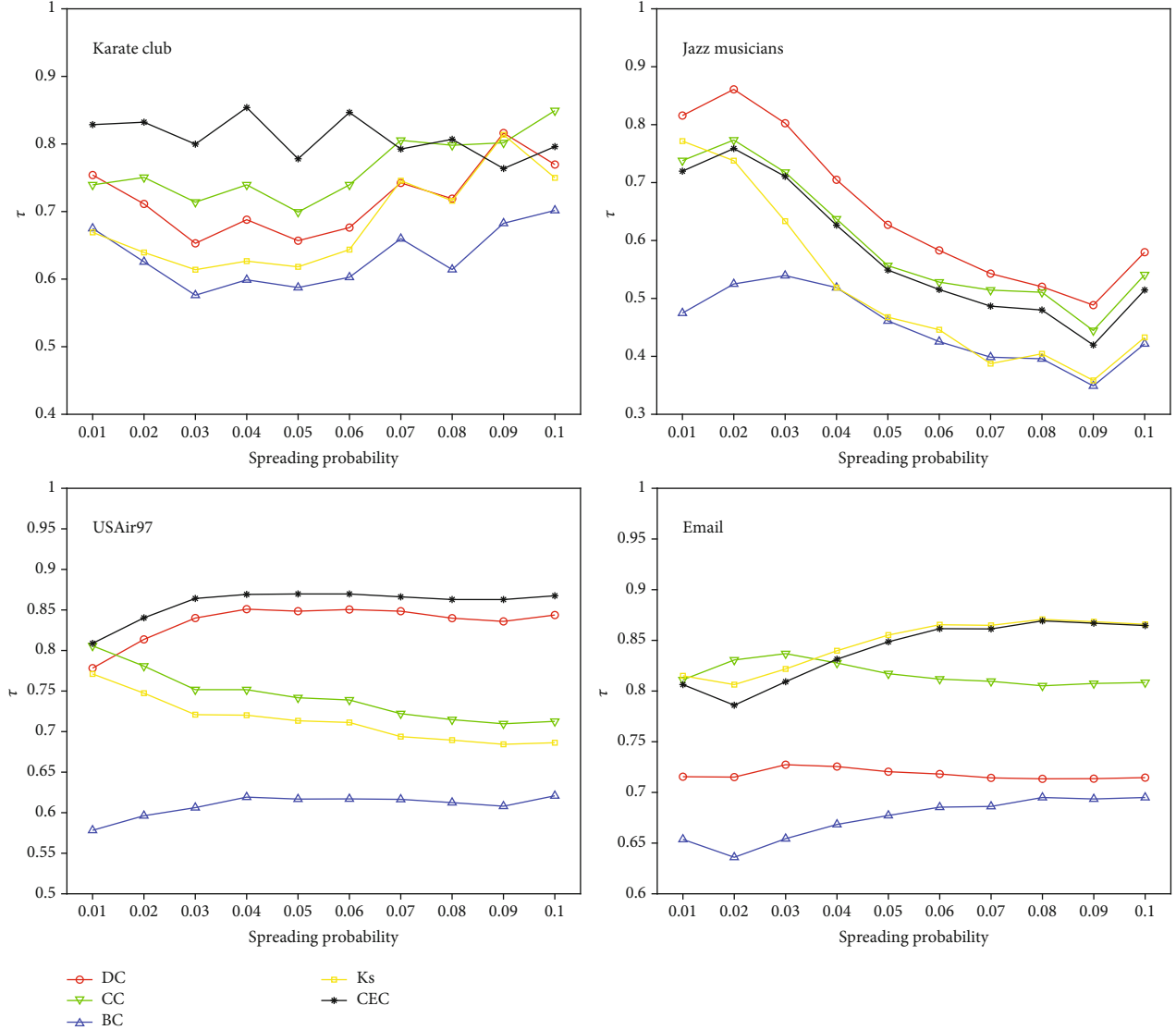
FIGURE 3: Comparison of the correlation between centralities and the actual ranking list.

4.2.4. Experiment 4: Comparison of the Correlation between Centralities and the Actual Ranking Result. We choose Kendall's tau coefficient (τ) [36] to be a linear correlation coefficient between the five methods and the actual ranking result. The value of τ ranges between [0, 1]; the larger the value of τ is, the more similar two sequences is. Give two sequences $X = \{x_1, x_2, \cdots, x_s\}$ and $Y = \{y_1, y_2, \cdots, y_s\}$. (x_i, y_i) is regarded as a positive sequence pair when $x_i > x_j$ and $y_i > y_j$, or $x_i < x_j$ and $y_i < y_j$, or else it will be considered as a negative sequence pair. Then, Kendall's tau can be denoted as $\tau = (n_+ - n_-/n(n-1))$, where n_+ and n_- indicate the number of positive sequence pairs and negative sequence pairs, respectively, and $n = n_+ + n_-$.

We consider the ranking list at $t = 10$ obtained by SI model as the actual ranking result Ia; then, we calculate the correlation between Ia and centralities. As shown in Figure 3, CEC outweighs other centralities before spreading probability 0.07 in karate club network, and it is lower than

CC after spreading probability 0.08. In Jazz musician network, DC has the greatest τ value, while it has very poor performance in email network, and the τ value of CEC is similar with CC. In USAir97 network, CEC outweighs other centralities across the spreading probability. In email network, the τ value of CEC is lower than CC and Ks before spreading probability 0.04, and it is similar with Ks after spreading probability 0.04. Overall speaking, CEC has the best correlation with actual ranking result in the four networks.

5. Conclusion

Identifying influential spreaders is essential for network invulnerability. In this paper, we pay attention to the approach of identifying influential spreaders based on global information, and the connectivity and efficiency centrality (CEC) are put forward to achieve this goal. Removing spreaders and the corresponding links will lead to two

consequences: the destruction of network connectivity and the decline of network efficiency. Therefore, we consider both the two aspects to provide a novel centrality in identifying influential spreaders. The relative changes of network connectivity and efficiency before and after removing spreaders are taken as indicators to measure the influence of spreaders; we combine the relative changes of network connectivity and efficiency to give comprehensive identifying results. The greater the relative changes, the more influential the spreader. We conduct several experiments based on actual datasets, and the results show that CEC performs better than other methods.

Data Availability

All data are available in the manuscript references, which can be accessed at Pubmed, google scholar and other web resources.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] J. Kim and M. Hastak, "Social network analysis: characteristics of online social networks after a disaster," *International Journal of Information Management*, vol. 38, no. 1, pp. 86–96, 2018.

[2] K. Q. Zhang, H. F. Du, and M. W. Feldman, "Maximizing influence in a social network: improved results using a genetic algorithm," *Physica A: Statistical Mechanics and its Applications*, vol. 478, pp. 20–30, 2017.

[3] R. Ding, N. Ujang, H. Hamid et al., "Detecting the urban traffic network structure dynamics through the growth and analysis of multi-layer networks," *Physica A: Statistical Mechanics and its Applications*, vol. 503, pp. 800–817, 2018.

[4] M. Y. Zhang, B. Y. Liang, S. Wang, M. Perc, W. B. Du, and X. B. Cao, "Analysis of flight conflicts in the Chinese air route network," *Chaos Solitons & Fractals*, vol. 112, pp. 97–102, 2018.

[5] Y. Yang, T. Nishikawa, and A. E. Motter, "Small vulnerable sets determine large network cascades in power grids," *Science*, vol. 358, no. 6365, p. eaan3184, 2017.

[6] B. Schäfer, D. Witthaut, M. Timme, and V. Latora, "Dynamically induced cascading failures in power grids," *Nature Communications*, vol. 9, no. 1, p. 1975, 2018.

[7] S. Pilosof, M. A. Porter, M. Pascual, and S. Kéfi, "The multilayer nature of ecological networks," *Nature Ecology & Evolution*, vol. 1, no. 4, 2017.

[8] G. Strona and K. D. Lafferty, "Environmental change makes robust ecological networks fragile," *Nature Communications*, vol. 7, no. 1, 2016.

[9] M. Zhao, T. Zhou, B. H. Wang, and W. X. Wang, "Enhanced synchronizability by structural perturbations," *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 72, no. 5, article 057102, 2005.

[10] Y. Z. Yang, L. Yu, Z. L. Zhou, Y. Chen, and T. Kou, "Node importance ranking in complex networks based on multicriteria decision making," *Mathematical Problems in Engineering*, vol. 2019, Article ID 9728742, 12 pages, 2019.

[11] M. J. Alvarez, Y. Shen, F. M. Giorgi et al., "Functional characterization of somatic mutations in cancer using network-based inference of protein activity," *Nature Genetics*, vol. 48, no. 8, pp. 838–847, 2016.

[12] R. Albert, I. Albert, and G. L. Nakarado, "Structural vulnerability of the North American power grid," *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 69, no. 2, article 025103, 2004.

[13] A. Sheikhahmadi, M. A. Nematbakhsh, and A. Shokrollahi, "Improving detection of influential nodes in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 436, pp. 833–845, 2015.

[14] L. Y. Lü, D. B. Chen, X. L. Ren, Q. M. Zhang, Y. C. Zhang, and T. Zhou, "Vital nodes identification in complex networks," *Physics Reports*, vol. 650, pp. 1–63, 2016.

[15] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *Journal of Mathematical Sociology*, vol. 2, no. 1, pp. 113–120, 1972.

[16] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.

[17] M. E. J. Newman, "A measure of betweenness centrality based on random walks," *Social Networks*, vol. 27, no. 1, pp. 39–54, 2005.

[18] A. Zareie and A. Sheikhahmadi, "EHC: extended H-index centrality measure for identification of users' spreading influence in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 514, pp. 141–155, 2019.

[19] P. Hu, W. L. Fan, and S. W. Mei, "Identifying node importance based on evidence theory in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 429, pp. 169–176, 2015.

[20] P. Roberts, D. Gaffney, J. Lee-Thorp, and G. Summerhayes, "Identifying influential nodes in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 4, pp. 1777–1787, 2012.

[21] D. B. Chen, H. Gao, L. Y. Lü, and T. Zhou, "Identifying influential nodes in large-scale directed networks: the role of clustering," *PLoS One*, vol. 8, no. 10, article e77455, 2013.

[22] M. Kitsak, L. K. Gallos, S. Havlin et al., "Identification of influential spreaders in complex networks," *Nature Physics*, vol. 6, no. 11, pp. 888–893, 2010.

[23] A. Zeng and C. J. Zhang, "Ranking spreaders by decomposing complex networks," *Physics Letters A*, vol. 377, no. 14, pp. 1031–1035, 2013.

[24] J. Bae and S. Kim, "Identifying and ranking influential spreaders in complex networks by neighborhood coreness," *Physica A: Statistical Mechanics and its Applications*, vol. 395, pp. 549–559, 2014.

[25] A. Arasu, J. H. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan, "Searching the web," *ACM Transactions on Internet Technology (TOIT)*, vol. 1, no. 1, pp. 2–43, 2001.

[26] L. Lü, Y. C. Zhang, C. H. Yeung, and T. Zhou, "Leaders in social networks, the delicious case," *PLoS One*, vol. 6, no. 6, 2011.

[27] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, vol. 98, pp. 668–677, 1998.

[28] Z. H. Liu, C. Jiang, J. Y. Wang, and H. Yu, "The node importance in actual complex networks based on a multi-attribute ranking method," *Knowledge-Based Systems*, vol. 84, pp. 56–66, 2015.

[29] Y. Z. Yang, L. Yu, X. Wang, Z. L. Zhou, Y. Chen, and T. Kou, "A novel method to evaluate node importance in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 526, 2019.

[30] T. Wen and Y. Deng, "Identification of influencers in complex networks by local information dimensionality," *Information Sciences*, vol. 512, pp. 549–562, 2020.

[31] S. S. Wang, Y. X. Du, and Y. Deng, "A new measure of identifying influential nodes: efficiency centrality," *Communications in Nonlinear Science and Numerical Simulation*, vol. 47, pp. 151–163, 2017.

[32] T. Zhou, J. G. Liu, W. J. Bai, G. Chen, and B. H. Wang, "Behaviors of susceptible-infected epidemics on scale-free networks with identical infectivity," *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 74, no. 5, article 056109, 2006.

[33] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.

[34] M. G. Pablo and L. E. Danon, "Community structure in jazz," *Advances in Complex Systems*, vol. 6, no. 4, pp. 565–573, 2003.

[35] "North American Transportation Atlas Data (NORTAD)," https://www.bts.gov/archive/publications/north_american_transportation_atlas_data/index.

[36] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 68, no. 6, 2003.

[37] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 1938.