

Received March 5, 2020, accepted March 20, 2020, date of publication March 24, 2020, date of current version April 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2983053

A Novel Centrality of Influential Nodes Identification in Complex Networks

YUANZHI YANG¹, XING WANG¹, YOU CHEN¹, MIN HU², AND CHENGWEI RUAN³

¹Aeronautics Engineering College, Air Force Engineering University, Xi'an 710038, China

²School of New Energy and Materials, Southwest Petroleum University, Chengdu 610500, China

³95910 Army, Jiuquan 735018, China

Corresponding authors: You Chen (chenyousky@163.com) and Min Hu (hu_min_min@163.com)

This work was supported by the National Natural Science Foundation of China under Grant 61472443.

ABSTRACT Influential nodes identification in complex networks is vital for understanding and controlling the propagation process in complex networks. Some existing centrality measures ignore the impacts of neighbor node. It is well-known that degree is a famous centrality measure for influential nodes identification, and the contributions of neighbors also should be taken into consideration. Furthermore, topological connections among neighbors will affect nodes' spreading ability, that is, the denser the connections among neighbors, the greater the chance of infection. In this paper, we propose a novel centrality, called DCC, to identify influential nodes by comprehensively considering degree and clustering coefficient as well as neighbors. The weights of degree and clustering coefficient are calculated by entropy technology. To verify the feasibility and effectiveness of DCC, the comparisons between DCC and other centrality measures in four aspects are conducted based on four real networks. The experimental results demonstrate that DCC is more effective in identifying influential nodes.


INDEX TERMS Complex networks, influential nodes, degree centrality, neighbor node, clustering coefficient.

I. INTRODUCTION

In recent years, a wide range of real-world complex systems, such as social network [1], [2], power grids [3]–[5], computer system [6], [7] and traffic system [8], [9], can be described by complex networks. Although complex systems bring us great convenience, the related hazards will also occur, such as the high speed and large scale of WannaCry's spread, the outbreak of infectious diseases, and the North American blackout. These hazards usually start with a small number of nodes and can quickly spread to all the network [10]. Therefore, influential nodes identification is of great interest for the robustness and stability of networks. For example, the spread of information can be accelerated or prevented with the help of prominent individuals [11]–[13]. Critical nodes in infectious disease network are identified to control and diagnose disease [14]–[16]. Key nodes in power grids can be identified to prevent power outages [17], [18].

The research of influential nodes identification has received increasing attention, and many classical

centrality measures have been proposed. For example, degree centrality [19], betweenness centrality [20], closeness [21], eigenvector centrality [22] and K-shell decomposition [23] have been widely used for influential nodes identification. The existing centrality measures can be divided into three categories: local centrality measures, semi-local centrality measures and global centrality measures [24]. Local centrality measures, such as degree centrality and clustering coefficient, only make use of the nearest neighbors' information and ignore the global information, resulting in low accuracy, but they are suitable for large-scale networks due to low time complexity. On the contrary, global centrality measures, such as betweenness centrality, closeness centrality and eigenvector centrality, determine the spreading ability of nodes based on the information of the entire network and thus exhibit higher accuracy, but they are not suitable for large-scale networks. Local and global centrality measures have their advantages and limitations as we mentioned above. Semi-local centrality measures, which are taken as a trade-off between local and global centrality measures, have been developed in recent years, they have high accuracy and low time complexity with considering more information of

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott .

neighbors. Nowadays, some researchers have proposed to combine different kinds of measures. Lü *et al.* [25] considered global and local information to detect influential nodes, and proposed M-centrality combining K-core decomposition and degree variation at a local level. What is more, there are some methods combining different attributes to identify influential nodes, such as evidential method [26], the Vlsekriterijumska Optimizacija I Kompromisno Resenje (VIKOR) method [27], the Technology for Order Preference by Similarity to an Ideal Solution (TOPSIS) method [28], etc. Gao *et al.* [29] pointed out that multi-attribute methods can make a further enhanced identification. However, some methods consider the contributions of different attributes equally important, which is not scientific and unreasonable. In addition, in the field of search engine, there are also some well-known centrality measures. PageRank [30] was developed to rank the importance of web-pages via link structure. LeaderRank [31] was a simple variant of PageRank and introduced a so-called ground node connected with every other node by a bidirectional edge.

In this paper, a novel centrality (abbreviated as DCC) belonging to semi-local centrality measures is proposed to identify influential nodes in complex networks. As indicated in [32], local characteristics including degree and clustering coefficient play important roles in identifying influential nodes, but the method in [32] (denoted as NP) does not consider neighbor information comprehensively. In our method, we take degree and clustering coefficient into consideration as well as neighbor information. We consider not only the degree of the target node but also the degree of its neighbors. Besides, clustering coefficient reflects the density of connections between neighbors and the target node, and it is an important index to reflect the spreading ability of node. We consider two aspects of clustering coefficient: clustering coefficient of the target node and second-level neighbors. It is well known that a node's clustering coefficient has negative effect on the spreading ability (the greater the clustering coefficient, the less influential the node), while clustering coefficient of second-level neighbors has positive effects on the spreading ability (the greater the clustering coefficient of second-level neighbors, the more influential the node). In fact, a node with a high degree, high neighbors' degree, low clustering coefficient and dense second-level neighbors can be identified as a structural hole. To verify the feasibility and efficiency of DCC, Susceptible–Infected (SI) model [33] is adopted to simulate the spreading process on four real networks. Some classical centrality measures (degree centrality, closeness centrality, betweenness centrality, eigenvector centrality, K-shell, local centrality and NP) are used for comparison. The experimental results indicate the superiority of DCC.

The rest of the paper is organized as follows. We review the related work in section II. In section III, the proposed centrality measure is discussed. Experiments based on four real networks are conducted in section IV. And we draw a conclusion in Section V.

II. RELATED WORK

Centrality measures can be divided into three categories, and we review some classical centrality measures of each category in this part. Given an undirected and unweighted network $G(V, E, A)$ with $|V| = n$ nodes and $|E| = m$ edges. $A = \{a_{ij}\}$ is the adjacency matrix, a_{ij} has two values, which is 1 (node i and node j are connected) and 0 (node i and node j are disconnected).

Degree centrality is the most important and famous centrality belonging to local centrality measures. It only uses the information of the nearest neighbors and has low accuracy. Degree centrality [19] of node i is defined as

$$C_D(i) = \sum_{j=1}^n a_{ij} \quad (1)$$

Betweenness centrality [20] and closeness centrality [21] are the most essential centrality measures belonging to global centrality measures, they need the information of the entire network and are not suitable for large-scale networks. Betweenness reflects nodes' ability to control the information traveling along the shortest path in the network, which is denoted as

$$C_B(i) = \sum_{s,t \neq i} \frac{g_{st}(i)}{g_{st}} \quad (2)$$

where $g_{st}(i)$ represents the number of shortest paths between node s and node t passing through node i , and g_{st} is the number of all possible shortest paths.

Closeness centrality, which determines the spreading ability of nodes by the average spreading time of information, is defined as

$$C_C(i) = \frac{1}{\sum_{j \in V/i} d_{ij}} \quad (3)$$

where d_{ij} is the shortest distance between node i and node j .

The K-shell decomposition [23] proposed by Kitsak is also a global centrality measure. The outer nodes are stripped layer by layer, and the inner nodes have high influence. Specifically, this method begins with removing all the nodes with degree 1 from the graph, and the removing process continues until there is no node with degree 1, then the removed nodes are assigned with $ks = 1$. A similar process is followed for nodes with degree 2. Finally, all the nodes in the graph will be assigned a ks value. It can be regarded as a coarse-grained ranking method based on nodes' degree. It is easy to execute and can be applied to large-scale networks. However, it results in a poor performance in distinguishing nodes' centrality value.

The well-known semi-local centrality measure called local centrality (LC) [34] considers both the nearest and the next nearest neighbors. It is defined as

$$Q(j) = \sum_{w \in \Gamma_j} N(w) \quad (4)$$

$$C_L(i) = \sum_{j \in \Gamma_i} Q(j) \quad (5)$$

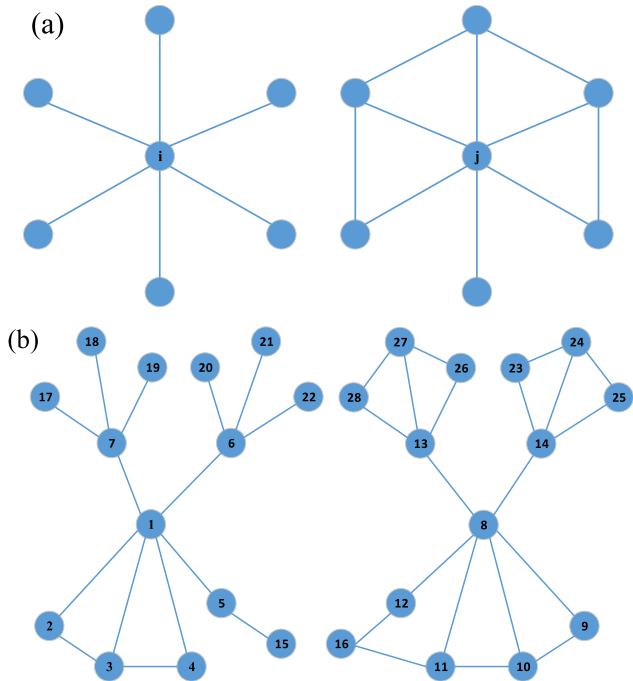


FIGURE 1. The example networks. (a) Node i and node j have the same degree ($C_D(i) = C_D(j) = 6$) but different clustering coefficient ($C_i = 0$, $C_j = 0.2667$). (b) Degree and clustering coefficient of node 1 and node 8 are the same ($C_D(1) = C_D(8) = 6$, $C_1 = C_8 = 0.1333$), but the sum of the second-level neighbors' clustering coefficient is different (2.6667 for node 1, 8 for node 8), the spreading ability of node 8 (3.86) is stronger than that of node 1 (3.80). The spreading ability of nodes is determined at the spreading probability (0.1 in (a) and 0.04 in (b)) by performing 1000 Monte Carlo simulations of the SI model per seed node.

where Γ_i represents the set of the nearest neighbors of node i and $N(w)$ denotes the total number of the nearest and the next nearest neighbors of node w . LC has its limitations and ignores the topological connections among the neighbors. To solve this issue, the local structural centrality (LSC) [35] considering clustering coefficient was proposed. The LSC is defined as

$$C_{LS}(i) = \sum_{j \in \Gamma_i} \left(\lambda N(j) + (1 - \lambda) \sum_{w \in \Gamma_j^2} C_w \right) \quad (6)$$

where C_w denotes clustering coefficient of node w , and $\lambda \in [0, 1]$ is the adjustment parameter. But the value of λ in LSC is artificially given, which will lead to different identification results. Clustering coefficient [36] is denoted as

$$C_i = \frac{2e_i}{C_D(i)(C_D(i) - 1)} \quad (7)$$

where e_i represents the number of connected edges between all adjacent nodes of node i .

III. THE PROPOSED CENTRALITY

The main purpose of our work is to propose a semi-local centrality measure, which considers not only degree of the target node and the neighbors, but also clustering coefficient of the target node and the second-level neighbors.

Degree and clustering coefficient are two essential centrality measures for influential nodes identification. At the same time, the effects of neighbors on the target node should also be considered. Here, we illustrate the problem according to some examples. Sometimes, nodes may have the same degree but different clustering coefficient. For example, in FIGURE 1(a), degree of node i and node j is 6, but their clustering coefficient is 0 and 0.2667, respectively. In Susceptible–Infected (SI) model, the spreading ability of node i and node j is 3.75 and 4.52, respectively. Clustering coefficient has a negative impact on node's spreading ability [37]. In addition, clustering coefficient of the second-level neighbors plays an important role in identifying influential nodes. Sometimes, nodes may have the same degree and clustering coefficient, but the sum of the second-level neighbors' clustering coefficient is different. For example, we compare the spreading ability of node 1 with node 8 in FIGURE 1(b), they have the same degree ($C_D(1) = C_D(8) = 6$) and the same clustering coefficient ($C_1 = C_8 = 0.1333$), nonetheless node 8 has stronger spreading ability than node 1 since node 8 has higher value of the sum of the second-level neighbors' clustering coefficient (2.6667 for node 1, 8 for node 8). The spreading ability of node 1 and node 8 estimated by the SI model is 3.80 and 3.86, respectively, and we can find that clustering coefficient of the second-level neighbors has a positive impact on nodes' spreading ability. It can be seen from the topological connections that node 8 is a key connector among the dense areas and expected to be a structural hole. We consider not only the degree of target node and neighbors, but also clustering coefficient of target node and the second-level neighbors, a novel centrality measure (DCC) is defined as

$$DCC(i) = \alpha I_D(i) + \beta I_C(i), \quad \alpha + \beta = 1 \quad (8)$$

where $I_D(i) = C_D(i) + \sum_{j \in \Gamma_i} C_D(j)$ represents the effect of degree and neighbors' degree of node i , $I_C(i) = e^{-C_i} \sum_{j \in \Gamma_i^2} C_j$ denotes the effect of clustering coefficient and second-level neighbors' clustering coefficient of node i , α and β represents the weight of $I_D(i)$ and $I_C(i)$, respectively.

Furthermore, we need to determine the values of α and β . Some multi-attribute decision-making methods assign equal weights ($\alpha = \beta$) to attributes, this will lead to the ranking results being affected by human factors, which is not scientific and unreasonable. There are many methods to compute the weights, such as Analytic Hierarchy Process (AHP) method, Delphi method, principle element analysis and entropy technology [38]–[40]. Here, we choose entropy technology to calculate the weights of $I_D(i)$ and $I_C(i)$ for its excellent performance [41]. The process of entropy technology is represented as follows.

First, we establish a decision matrix D according to the values of $I_D(i)$ and $I_C(i)$ of all nodes in the network.

$$D = d(i, j)_{2 \times n} = \begin{bmatrix} I_D(1) & I_D(2) & \cdots & I_D(n) \\ I_C(1) & I_C(2) & \cdots & I_C(n) \end{bmatrix} \quad (9)$$

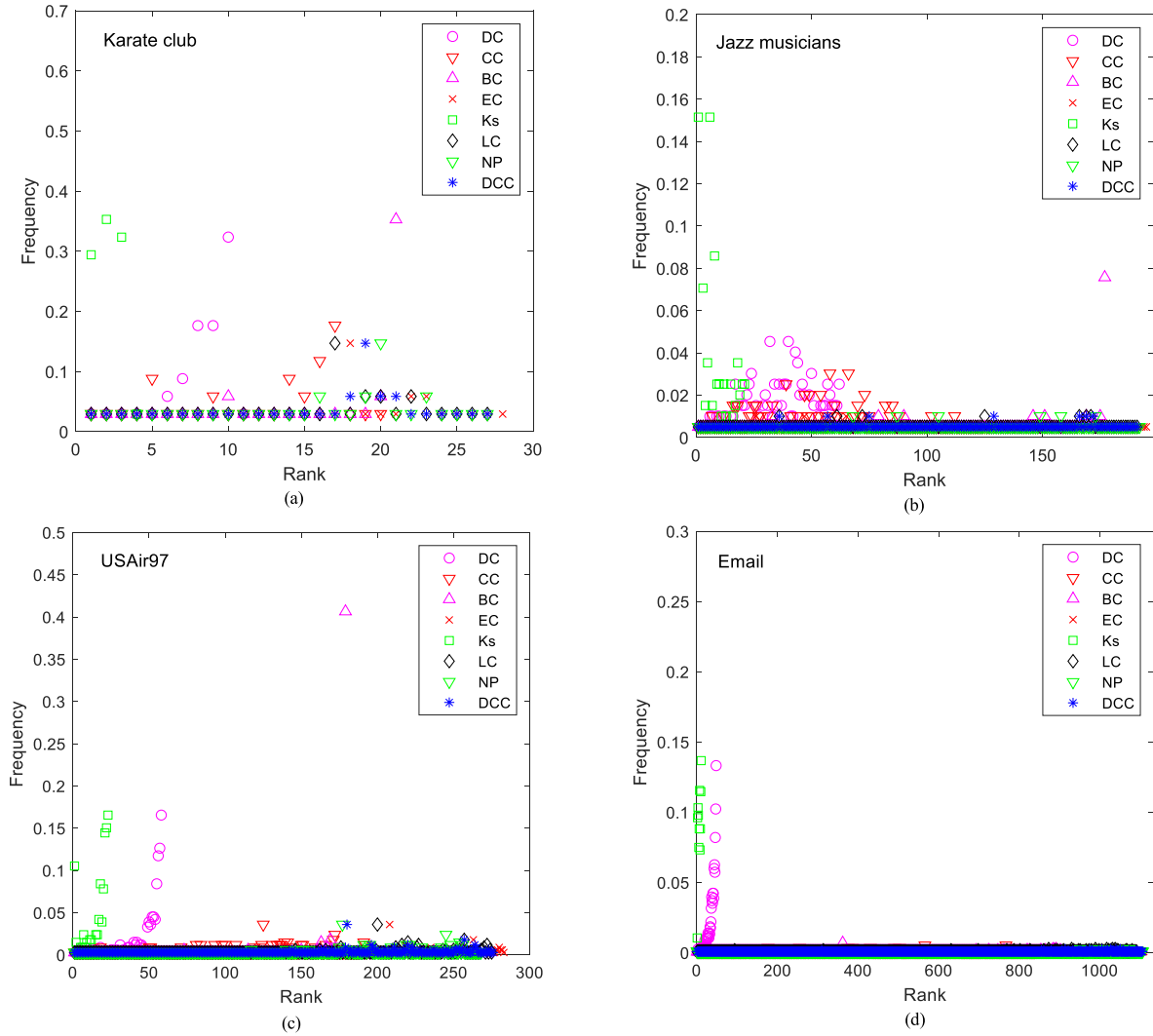


FIGURE 2. The frequency of nodes with the same ranking value in four datasets. (a) Karate club network. (b) Jazz musicians network. (c) USAir97 network. (d) Email network.

Next, we need to normalize the decision matrix D and build the standard matrix R .

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \end{bmatrix}, \quad r_{ij} = d_{ij} / \sqrt{\sum_{j=1}^n (d_{ij})^2} \quad (10)$$

Then, we calculate the entropy of the i th attribute as follows.

$$E_i = -\frac{1}{\ln n} \sum_{j=1}^n r_{ij} \ln r_{ij}, \quad i = 1, 2; \quad j = 1, 2, \dots, n \quad (11)$$

Finally, we can obtain the weight of the i th attribute.

$$w_i = \frac{1 - E_i}{2 - \sum_{i=1}^2 E_i}, \quad i = 1, 2 \quad (12)$$

To gain a better understanding of the calculation process of DCC, we take FIGURE 1(b) as an example to show how to

identify influential nodes. For node 1,

$$\begin{aligned} I_D(1) &= C_D(1) + \sum_{j \in \Gamma_1} C_D(j) = 6 \\ &\quad + (2 + 3 + 2 + 2 + 4 + 4) = 23, \\ I_C(1) &= e^{-C_1} \sum_{j \in \Gamma_1} C_j = e^{-C_1} \\ &\quad \times (C_2 + C_3 + C_4 + C_{15} + C_{17} + C_{18} \\ &\quad + C_{19} + C_{20} + C_{21} + C_{22}) = 2.3338. \end{aligned}$$

And we can obtain the values of $I_D(i)$ and $I_C(i)$ of all nodes in the same way. Then we can establish the decision matrix.

$$D = \begin{bmatrix} 23 & 11 & \cdots & 9 \\ 2.3338 & 0.6622 & \cdots & 0.7848 \end{bmatrix}.$$

Next, the standard matrix is built as follows.

$$R = \begin{bmatrix} 0.3739 & 0.1788 & \cdots & 0.1463 \\ 0.2035 & 0.0577 & \cdots & 0.0684 \end{bmatrix}.$$

TABLE 1. The basic statistics of four real networks. These statistics include the number of nodes (N), the number of edges (E), the average degree (< k >), clustering coefficient (C) and average shortest distance (< d >).

Network	N	E	<k>	C	<d>
Karate club	34	78	4.59	0.59	2.41
Jazz musicians	198	2742	27.70	0.63	2.21
USAir97	332	2126	12.81	0.75	2.74
Email	1133	5451	9.62	0.22	3.61

TABLE 2. The comparison of TRF using different centrality measures.

Network	DC (%)	CC (%)	BC (%)	EC (%)	Ks (%)	LC (%)	NP (%)	DCC (%)
Karate club	82.35	58.82	47.06	26.47	97.06	32.35	32.35	32.35
Jazz musicians	93.43	57.07	14.65	3.03	99.49	7.07	6.06	7.07
USAir97	92.17	63.25	49.70	23.49	99.40	26.20	27.71	26.20
Email	99.38	45.72	21.89	3.8	100	6.88	4.24	4.24

And we calculate the entropy of the first attribute ($I_D(i)$).

$$E_1 = -\frac{1}{\ln 28} \sum_{j=1}^{28} r_{1j} \ln r_{1j} = 2.4011.$$

Then, we can obtain the weight of the first attribute.

$$\alpha = \frac{1 - E_1}{2 - (E_1 + E_2)} = 0.6742.$$

Finally, the novel centrality value of node 1 is obtained.

$$DCC(1) = \alpha I_D(1) + \beta I_C(1) = 16.2677.$$

We can calculate the values of all the nodes in the same way and rank the spreading ability of nodes.

The proposed method is made up with four parts: degree, neighbors' degree, clustering coefficient, and second-level neighbors' clustering coefficient. The time complexities for calculating degree and neighbors' degree of all nodes are $O(n)$. The time complexity for calculating clustering coefficient of each node is $O(\langle k \rangle)$, where $\langle k \rangle$ is the average degree of a graph, and calculating second-level neighbors' clustering coefficient of each node has time complexity $O(\langle k \rangle^2)$. In addition, the calculation of entropy method has a complexity of $O(2n)$. Therefore, the time complexity for DCC is $O(n\langle k \rangle^2)$.

IV. EXPERIMENTAL ANALYSIS

A. DATASETS

We choose four real networks with varying sizes as the datasets to conduct experiments.

(1) Karate club [42]. Zachary karate club is a classical dataset in social networks, which reflects the social relationship among 34 members of the karate club in a university in the United States.

(2) Jazz musicians [43]. It is also a classical dataset in social networks. Each node represents a jazz musician and each edge denotes the cooperation between two musicians.

(3) USAir97. It is a North American transportation network consisting of 332 nodes and 2126 edges. Each node

TABLE 3. The top-5 nodes using different centrality measures in Karate club network.

Rank	DC	CC	BC	EC	Ks	LC	NP	DCC
1	34	1	1	34	34	1	34	1
2	1	3	34	1	33	3	1	34
3	33	34	33	3	1	34	33	3
4	3	32	3	33	2	33	3	33
5	2	9	32	2	3	9	32	32

TABLE 4. The top-5 nodes using different centrality measures in Jazz musicians network.

Rank	DC	CC	BC	EC	Ks	LC	NP	DCC
1	136	136	136	60	60	60	136	60
2	60	60	153	132	132	136	60	136
3	132	168	60	136	168	132	168	132
4	168	70	149	168	33	168	132	168
5	70	83	168	108	100	108	70	108

TABLE 5. The top-5 nodes using different centrality measures in USAir97 network.

Rank	DC	CC	BC	EC	Ks	LC	NP	DCC
1	118	118	118	118	118	118	118	118
2	261	261	8	261	261	261	261	261
3	255	67	261	255	67	255	255	255
4	152	255	201	182	201	182	182	182
5	182	201	47	152	230	152	152	152

TABLE 6. The top-5 nodes using different centrality measures in Email network.

Rank	DC	CC	BC	EC	Ks	LC	NP	DCC
1	105	333	333	105	299	105	578	105
2	333	23	105	16	389	42	135	42
3	16	105	23	196	434	333	23	16
4	23	42	578	204	552	16	233	333
5	42	41	76	42	571	23	105	23

and edge represent airport and route between airports, respectively. This dataset is available at <http://vlado.fmf.uni-lj.si/pub/networks/pajek/data/gphs.htm>.

(4) Email [44]. It is an email network of the University at Rovira i Virgili, each node represents a user and an edge indicates that two users have email exchange.

The basic statistics of four real datasets are presented in TABLE 1.

B. ANALYSIS AND RESULTS

In order to assess the performance of DCC, four experiments are conducted based on the four datasets. DC (Degree centrality), CC (Closeness centrality), BC (Betweenness centrality), EC (Eigencentrality), Ks (K-shell centrality), LC (Local centrality) and NP (the centrality proposed in [32]) are applied to the same datasets for comparison. In addition, SI model is adopted to simulate the spreading ability.

1) CAPABILITY OF DIFFERENT CENTRALITIES MEASURES TO DISTINGUISH NODES' SPREADING ABILITY

We rank the spreading ability of nodes according to their centrality values in each network using different centrality measures. When ranking the spreading ability of nodes in a

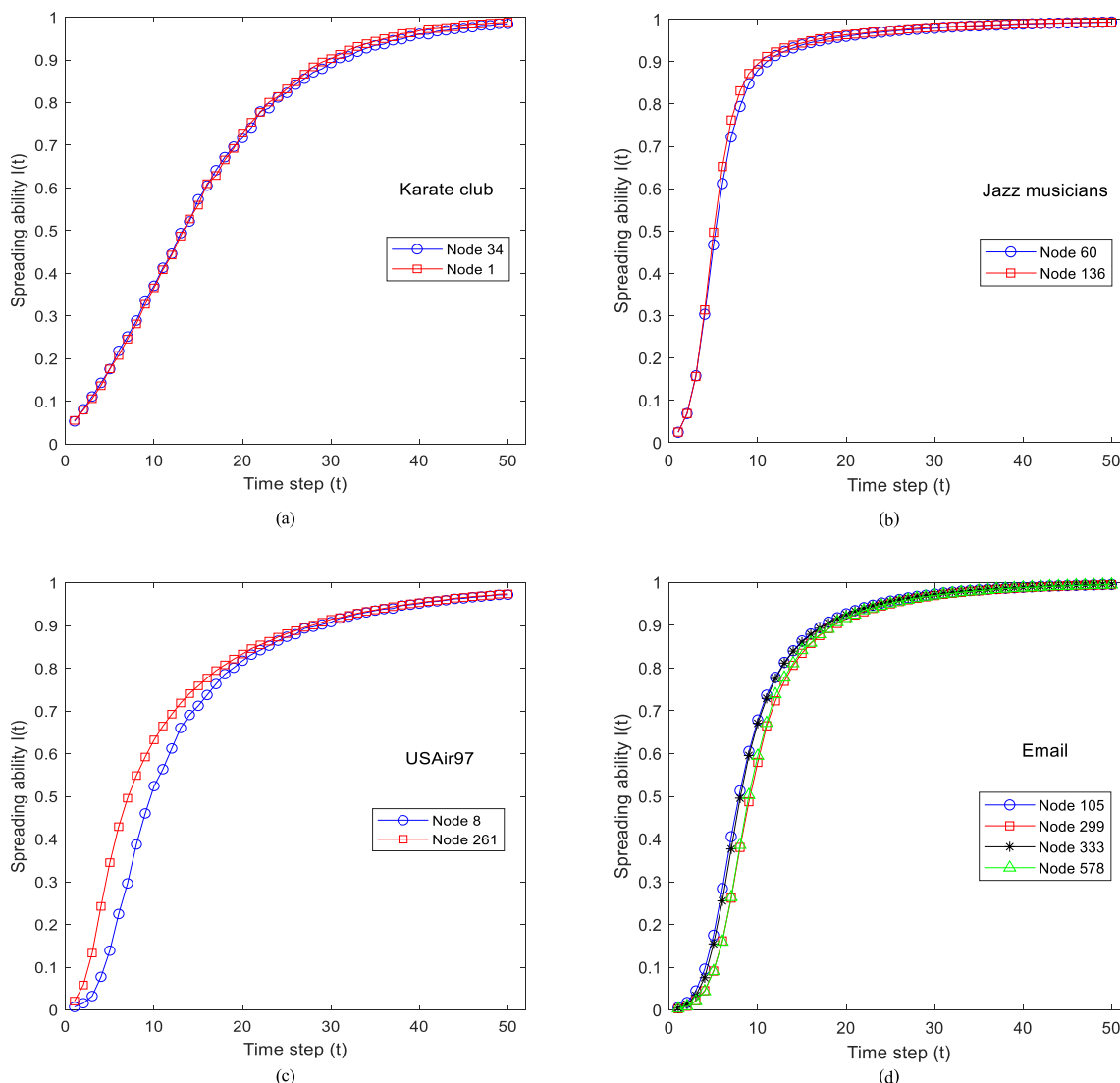


FIGURE 3. Comparison of spreading ability of top-1 nodes ranked by different centrality measures. The maximum time step is set as $t = 50$ and the simulation runs 1000 Monte Carlo. (a) Karate club network. (b) Jazz musicians network. (c) USAir97 network. (d) Email network.

network, it usually occurs that some nodes have the same centrality value and thus cannot be distinguished. For a centrality measure, high frequency of nodes with the same centrality value indicates that the centrality measure has poor performance. Therefore, the frequency of nodes with the same centrality value can be adopted as an indicator to evaluate the performance of a centrality measure. In this part, the frequency of nodes with the same centrality value using different centrality measures is compared in FIGURE 2. Obviously, in the four networks, the frequency of DC and Ks is higher than that of other centrality measures, which means that there are much more nodes with the same centrality value obtained by DC and Ks than other centrality measures, that is, DC and Ks have the worst capability of distinguishing nodes' spreading ability. Comparatively, almost in all cases, DCC, EC, LC and NP own the least nodes with the same centrality

value, therefore, these measures can better detect the differences between nodes and they have stronger capability of distinguishing nodes' spreading ability.

Moreover, we define a parameter to conduct a further comparison between the classical centrality measures and DCC.

$$TRF(\text{Total Repetition Frequency}) = \frac{n_s}{n} \quad (13)$$

where n_s denotes the total number of nodes with the same centrality value. The minimum value $TRF = 0$ represents that all nodes are assigned different centrality values, while the maximum value $TRF = 1$ indicates that all nodes have the same centrality value.

Obviously, a smaller TRF value indicates a better performance of a centrality measure. The TRF values using different centrality measures are shown in TABLE 2. In the four

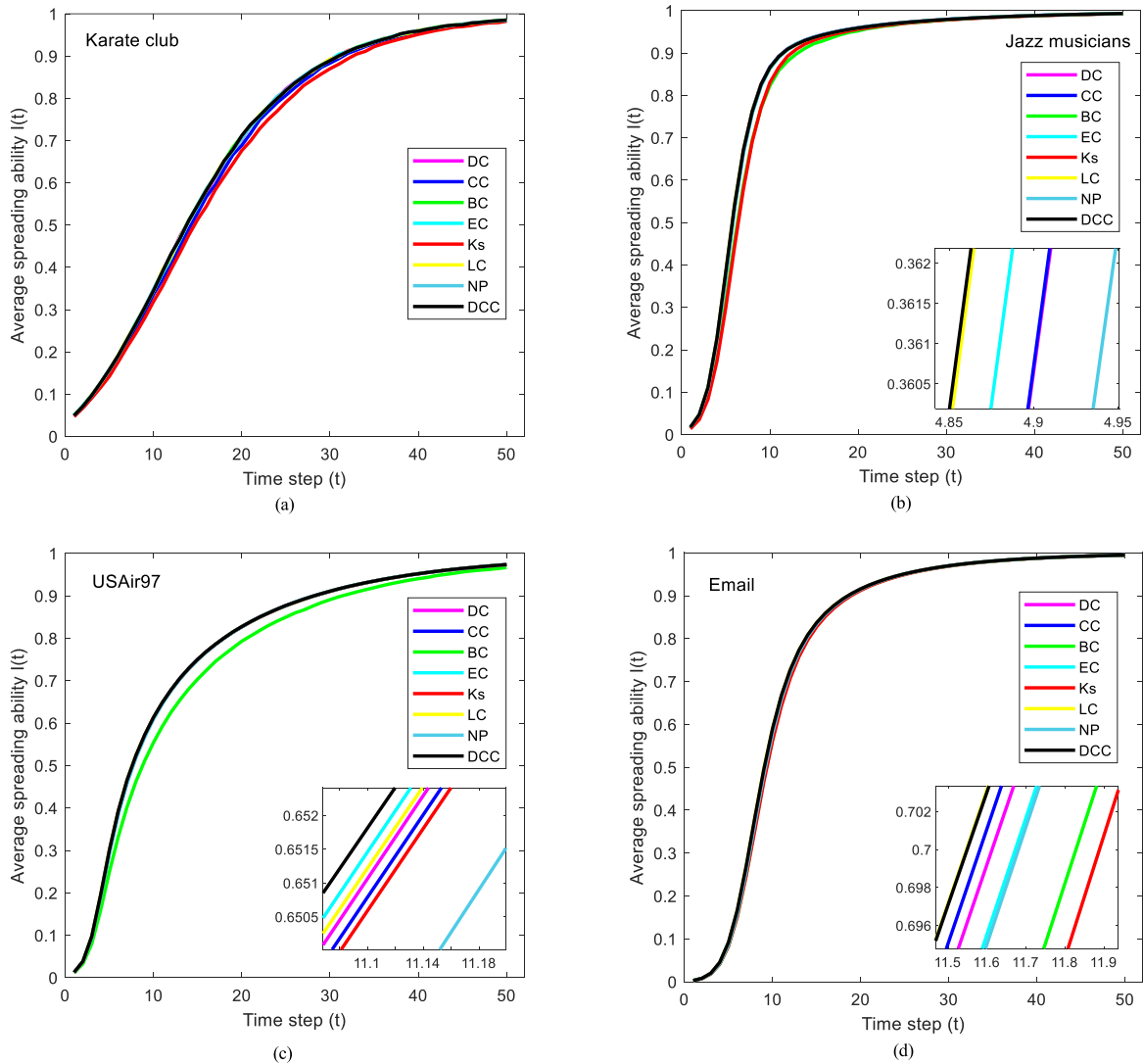


FIGURE 4. Comparison of average spreading ability of top- L nodes ranked by different centrality measures. The maximum time step is set as $t = 50$ and the simulation runs 1000 Monte Carlo. (a) Karate club network. (b) Jazz musicians network. (c) USAir97 network. (d) Email network.

networks, the TRF value of EC is the smallest, while the TRF values of DC and Ks are almost the greatest. DCC, LC and NP have similar performance because the TRF values of them are almost the same. Specifically, DCC owns smaller TRF value than LC and NP in Jazz musicians, which indicates that DCC distinguishes nodes' spreading ability more effectively.

2) CAPABILITY OF DIFFERENT CENTRALITIES MEASURES TO DISTINGUISH NODES' SPREADING ABILITY

We can obtain the ranking results using different centrality measures in four networks, and the top-5 nodes of each measure in four networks are presented in TABLE 3-6.

The spreading ability of the top-1 node simulated by the SI model is compared. Nodes in SI model has two states, one is susceptible state, and the other is infected state. Nodes in susceptible state will be infected by nodes in infected state with a certain probability and will never be recovered.

The top-1 nodes obtained by different centrality measures are chosen as source nodes, and the number of infected nodes will reach n_t after t ($t = 1, 2, \dots$) time step. Then the spreading ability of top-1 nodes can be denoted as $I(t) = n_t/n$. The maximum time step is set as $t = 50$ and the comparison results with 1000 Monte Carlo simulations are presented in FIGURE 3.

We define a symbol " $>$ " which denotes "more influential than". As shown in FIGURE 3, the number of infected nodes increases with time step and finally reaches a stable value. In Karate club network, we can find that node 1 $>$ node 34 slightly, which is consistent with DCC, CC, BC and LC, but contrary to DC, EC, Ks and NP. As for Jazz musicians network, the same situation exists that node 136 $>$ node 60 slightly, DC, CC, BC and NP identify node 136 as the most influential node, which perform slightly better than EC, Ks, LC and DCC. In USAir97 network, the top-1 nodes with

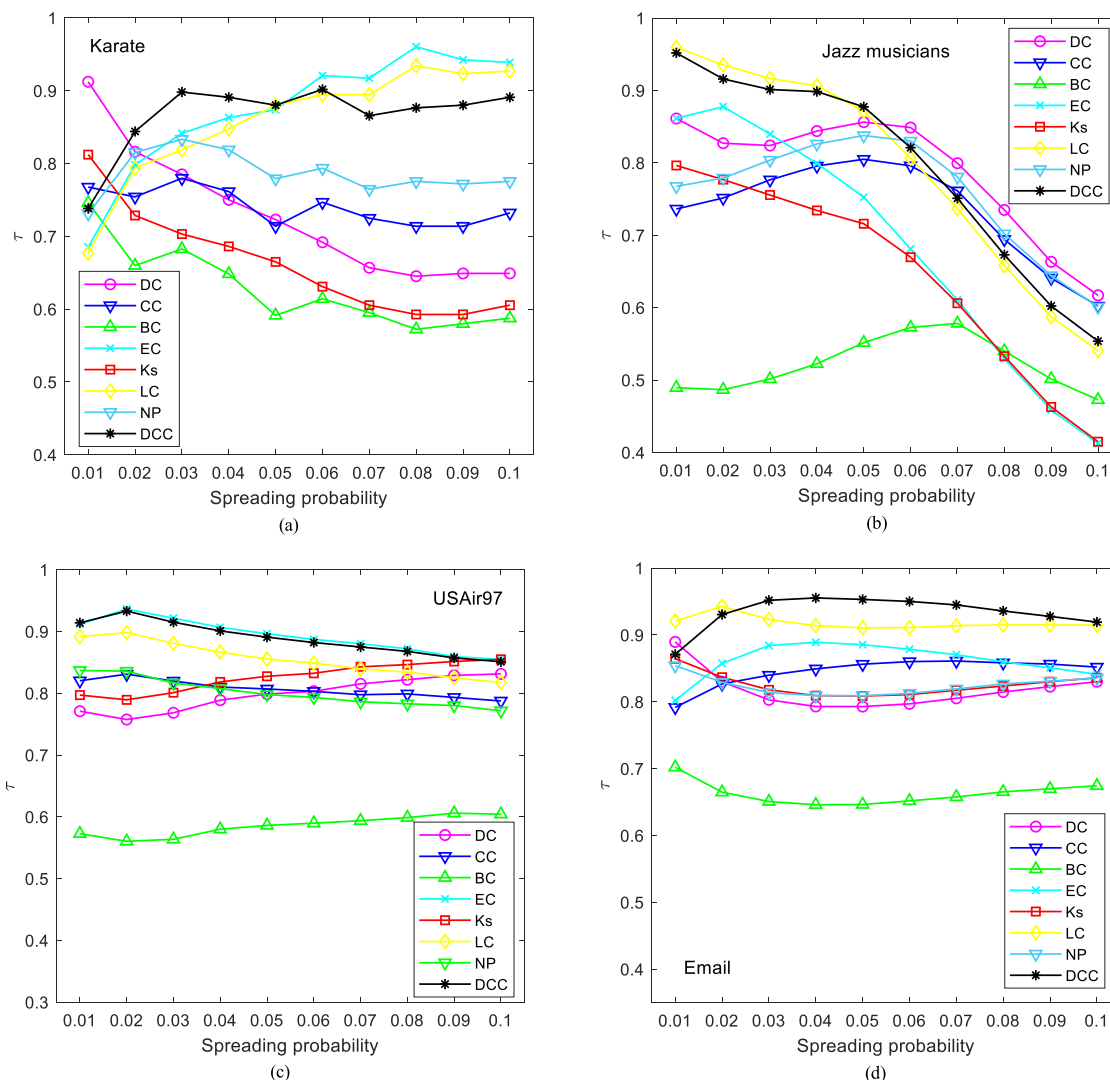


FIGURE 5. The Kendall's tau τ between different centrality measures and the ranking list ϵ generated by the SI model at $t = 10$. (a) Karate club network. (b) Jazz musicians network. (c) USAir97 network. (d) Email network.

eight centrality measures are all the same, therefore, we compare the second influential nodes. We can see that node $261 > \text{node } 8$, which is contrary to BC. In Email network, it is clear that node $105 > \text{node } 333 > \text{node } 578 > \text{node } 299$, which proves that DCC, DC, EC and LC perform better than others. Especially for NP, it performs the worst in identifying the most influential node. In a word, DCC is more accurate than the other centrality measures in identifying the most influential nodes.

3) COMPARISON OF AVERAGE SPREADING ABILITY OF TOP-L NODES

In the previous section, we approved the superiority of DCC in identifying the most influential node. In this part, we focus on the spreading ability of a group of nodes. A node i is chosen from the top- L list as a source node, and the number of infected nodes will reach n_{it} after t ($t = 1, 2, \dots$) time step. The average spreading ability of top- L nodes can be denoted as $I(t) = (\sum_{i=1}^L n_{it} / L) / n$. Here, we set $L = n \times 10\%$ to

pay attention to the centrality measures' ability to identify a group of influential nodes. In the same way, the maximum time step is set as $t = 50$ and the simulation runs 1000 Monte Carlo. We show the comparison results in FIGURE 4.

According to the FIGURE 4, in Karate club network, the average spreading abilities of DC, BC, EC, LC, NP and DCC are similar because the top-4 nodes of them are all the same, the average spreading ability of CC is worse than the mentioned six measures, while Ks has the worst performance. In Jazz musicians network, DCC and LC have similar performances, which are slightly better than that of other centrality measures, BC and Ks have the worst average spreading abilities. In USAir97 network, DCC performs slightly better than others, and BC has the worst average spreading ability. As for Email network, DCC shows similar performance with LC, while they slightly outperform all the other centrality measures. It is clear that DCC has a marginally better performance than the other centrality measures in all the four networks.

4) EVALUATION OF CORRELATION BETWEEN THE SIX CENTRALITY MEASURES AND THE ACTUAL SPREADING SITUATION

The Kendall's tau [45] is adopted as a correlation coefficient between the six centrality measures and the actual spreading situation. The ranking list ε at $t = 10$ simulated by the SI model is considered as the actual spreading situation, and ε is compared with the ranking lists using the six centrality measures. Suppose X and Y are the ranking lists of two centrality measures, a pair of distinct nodes i and j is concordant if $((X_i > X_j) \text{ and } (Y_i > Y_j))$ or $((X_i < X_j) \text{ and } (Y_i < Y_j))$. If $((X_i > X_j) \text{ and } (Y_i < Y_j))$ or $((X_i < X_j) \text{ and } (Y_i > Y_j))$, the pair is discordant. The Kendall's tau τ is defined as

$$\tau = \frac{n_c - n_d}{0.5 \times n \times (n - 1)} \quad (14)$$

where n_c and n_d are the number of concordant pairs and discordant pairs, respectively. Obviously, better performance of centrality measure has higher τ value.

We compare the τ value of the correlation between the eight centrality measures and ε with the increase of spreading probability. As shown in FIGURE 5, all the centrality measures are positively correlated with ε . In Karate club network, the τ value of DCC is greater than that of EC and LC when spreading probability is less than 0.05, as spreading probability increases, DCC shows a worse performance than EC and LC. In Jazz musicians network, DCC outperforms others when spreading probability is small, while DC shows better performance with the increase of spreading probability. In USAir97 network, DCC has similar correlation with EC, and they perform better than other centrality measures. As for Email network, DCC performs better than others almost across the entire range of spreading probability. Overall speaking, in these four datasets, BC has the weakest correlation with ε , DCC owning the strongest correlation with ε is closer to the actual spreading situation than others.

V. CONCLUSION

The identification of influential nodes in complex networks is an essential and open issue, which is of great significance to the robustness and stability of networks. There is increasing attention on this issue using new measures with local information. In this paper, a novel centrality, called DCC, considering the semi-local information is proposed for influential nodes identification. DCC comprehensively considers four aspects including degree, neighbors' degree, clustering coefficient and second-level neighbors' clustering coefficient. It is well-known that degree and neighbors' degree are positively correlated with spreading ability, that is, the higher the degree and neighbors' degree, the stronger the spreading ability. However, clustering coefficient is negatively correlated with spreading ability. Here, we focus on clustering coefficient of the second-level neighbors, the larger value of the sum of the second-level neighbors' clustering coefficient means that the second-level neighbors of the target node are in a denser part in the network. Therefore, a node has a high

degree, high neighbors' degree, low clustering coefficient and high second-level neighbors' clustering coefficient, it is identified as a structural hole. Furthermore, we divide these four aspects into two parts: degree effect and clustering coefficient effect. The entropy technology is adopted to assign weights to the two parts objectively, and we obtain comprehensive and reasonable ranking results. We conduct four experiments to verify the feasibility and efficiency of DCC based on four real datasets, and seven centrality measures are used for comparison. The experimental results demonstrate that DCC performs better. DCC is an effective complement to influential nodes identification with semi-local information.

ACKNOWLEDGMENT

The authors would like to thank the valuable comments and suggestions of the reviewers.

REFERENCES

- [1] S. Pei, L. Muchnik, J. S. Andrade, Jr., Z. Zheng, and H. A. Makse, "Searching for superspreaders of information in real-world social media," *Sci. Rep.*, vol. 4, Jul. 2015, Art. no. 5547.
- [2] A. N. Arularasan, A. Suresh, and K. Seerangan, "Identification and classification of best spreader in the domain of interest over the social networks," *Cluster Comput.*, vol. 22, no. S2, pp. 4035–4045, Mar. 2019.
- [3] W.-L. Fan, X.-M. Zhang, S.-W. Mei, and S.-W. Huang, "Vulnerable transmission line identification considering depth of K-shell decomposition in complex grids," *IET Gener., Transmiss. Distrib.*, vol. 12, no. 5, pp. 1137–1144, Mar. 2018.
- [4] P. H. J. Nardelli, N. Rubido, C. Wang, M. S. Baptista, C. Pomalaza-Raez, P. Cardieri, and M. Latva-Aho, "Models for the modern power grid," *Eur. Phys. J. Special Topics*, vol. 223, pp. 2423–2437, Oct. 2014.
- [5] Y. Yang, T. Nishikawa, and A. E. Motter, "Small vulnerable sets determine large network cascades in power grid," *Science*, vol. 358, no. 6365, Nov. 2017, Art. no. eaan3184.
- [6] Y. Yao, R. Zhang, F. Yang, J. Tang, Y. Yuan, and R. Hu, "Link prediction in complex networks based on the interactions among paths," *Phys. A, Stat. Mech. Appl.*, vol. 510, pp. 52–67, Nov. 2018.
- [7] H.-J. Li, H. Li, and C. Jia, "A novel dynamics combination model reveals the hidden information of community structure," *Int. J. Mod. Phys. C*, vol. 26, no. 4, Apr. 2015, Art. no. 1550043.
- [8] W.-B. Du, X.-L. Zhou, O. Lordan, Z. Wang, C. Zhao, and Y.-B. Zhu, "Analysis of the chinese airline network as multi-layer networks," *Transp. Res. E, Logistics Transp. Rev.*, vol. 89, pp. 108–116, May 2016.
- [9] S. Zhao, P. Zhao, and Y. Cui, "A network centrality measure framework for analyzing urban traffic flow: A case study of Wuhan, China," *Phys. A, Stat. Mech. Appl.*, vol. 478, pp. 143–157, Jul. 2017.
- [10] M. Zhao, T. Zhou, B.-H. Wang, and W.-X. Wang, "Enhanced synchronizability by structural perturbations," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 72, no. 5, Nov. 2005, Art. no. 057102.
- [11] A. Sheikahmadi, M. A. Nematbakhsh, and A. Shokrollahi, "Improving detection of influential nodes in complex networks," *Phys. A, Stat. Mech. Appl.*, vol. 436, pp. 833–845, Oct. 2015.
- [12] O. Hinz, C. Schulze, and C. Takac, "New product adoption in social networks: Why direction matters," *J. Bus. Res.*, vol. 67, no. 1, pp. 2836–2844, Jan. 2014.
- [13] M. Jalili and M. Perc, "Information cascades in complex networks," *J. Complex Netw.*, vol. 5, no. 5, pp. 665–693, Oct. 2017.
- [14] D. Helbing, D. Brockmann, T. Chadefaux, K. Donnay, U. Blanke, O. Woolley-Meza, M. Moussaid, A. Johansson, J. Krause, S. Schutte, and M. Perc, "Saving human lives: What complexity science and information systems can contribute," *J. Stat. Phys.*, vol. 158, no. 3, pp. 735–781, Feb. 2015.
- [15] Z. Wang, C. T. Bauch, S. Bhattacharyya, A. d'Onofrio, P. Manfredi, M. Perc, N. Perra, M. Salathé, and D. Zhao, "Statistical physics of vaccination," *Phys. Rep.*, vol. 664, pp. 1–113, Dec. 2016.

- [16] M. J. Alvarez, Y. Shen, F. M. Giorgi, A. Lachmann, B. B. Ding, B. H. Ye, and A. Califano, "Functional characterization of somatic mutations in cancer using network-based inference of protein activity," *Nature Genet.*, vol. 48, no. 8, pp. 838–847, Aug. 2016.
- [17] R. Albert, I. Albert, and G. L. Nakarado, "Structural vulnerability of the north American power grid," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 2, Feb. 2004, Art. no. 025103.
- [18] A. E. Motter, "Cascade control and defense in complex networks," *Phys. Rev. Lett.*, vol. 93, no. 9, Aug. 2004, Art. no. 098701.
- [19] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *J. Math. Sociol.*, vol. 2, no. 1, pp. 113–120, 1972.
- [20] M. E. J. Newman, "A measure of betweenness centrality based on random walks," *Social Netw.*, vol. 27, no. 1, pp. 39–54, Jan. 2005.
- [21] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Netw.*, vol. 1, pp. 215–239, 1978.
- [22] P. Bonacich and P. Lloyd, "Eigenvector-like measures of centrality for asymmetric relations," *Social Netw.*, vol. 23, no. 3, pp. 191–201, Jul. 2001.
- [23] M. Kistak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identifications of influential spreaders in complex network," *Natura Phys.*, vol. 6, pp. 888–893, Aug. 2010.
- [24] A. Ibnoulouafi, M. El Haziti, and H. Cherifi, "M-centrality: Identifying key nodes based on global position and local degree variation," *J. Stat. Mech., Theory Exp.*, vol. 2018, no. 7, Jul. 2018, Art. no. 073407.
- [25] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou, "Vital nodes identification in complex networks," *Phys. Rep.*, vol. 650, pp. 1–63, Sep. 2016.
- [26] H. Mo, C. Gao, and Y. Deng, "Evidential method to identify influential nodes in complex networks," *J. Syst. Eng. Electron.*, vol. 26, no. 2, pp. 381–387, Apr. 2015.
- [27] Y. Yang, L. Yu, X. Wang, Z. Zhou, Y. Chen, and T. Kou, "A novel method to evaluate node importance in complex networks," *Phys. A, Stat. Mech. Appl.*, vol. 526, Jul. 2019, Art. no. 121118.
- [28] Z. Liu, C. Jiang, J. Wang, and H. Yu, "The node importance in actual complex networks based on a multi-attribute ranking method," *Knowl.-Based Syst.*, vol. 84, pp. 56–66, Aug. 2015.
- [29] C. Gao, L. Zhong, X. Li, Z. Zhang, and N. Shi, "Combination methods for identifying influential nodes in networks," *Int. J. Mod. Phys. C*, vol. 26, no. 06, Jun. 2015, Art. no. 1550067.
- [30] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan, "Searching the Web," *ACM Trans. Internet Technol.*, vol. 1, no. 1, pp. 2–43, Aug. 2001.
- [31] L. Lü, Y.-C. Zhang, C. H. Yeung, and T. Zhou, "Leaders in social networks, the delicious case," *PLoS ONE*, vol. 6, no. 6, Jun. 2011, Art. no. e21202.
- [32] K. Berahmand, A. Bouyer, and N. Samadi, "A new centrality measure based on the negative and positive effects of clustering coefficient for identifying influential spreaders in complex networks," *Chaos, Solitons Fractals*, vol. 110, pp. 41–54, May 2018.
- [33] 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," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 74, no. 5, Nov. 2006, Art. no. 056109.
- [34] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," *Phys. A, Statist. Mech. Appl.*, vol. 391, pp. 1777–1787, Feb. 2012.
- [35] S. Gao, J. Ma, Z. Chen, G. Wang, and C. Xing, "Ranking the spreading ability of nodes in complex networks based on local structure," *Phys. A, Stat. Mech. Appl.*, vol. 403, pp. 130–147, Jun. 2014.
- [36] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, Jun. 1998.
- [37] X. Zhao, F. Liu, J. Wang, and T. Li, "Evaluating influential nodes in social networks by local centrality with a coefficient," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 2, p. 35, Jan. 2017.
- [38] B. Kim, J. Cho, S. Jeon, and B. Lee, "An AHP-based flexible relay node selection scheme for WBANS," *Wireless Pers. Commun.*, vol. 89, no. 2, pp. 501–520, Jul. 2016.
- [39] J. Jin, K. Xu, N. Xiong, Y. Liu, and G. Li, "Multi-index evaluation algorithm based on principal component analysis for node importance in complex networks," *IET Netw.*, vol. 1, no. 3, pp. 108–115, Sep. 2012.
- [40] Z. Wang, C. Du, J. Fan, and Y. Xing, "Ranking influential nodes in social networks based on node position and neighborhood," *Neurocomputing*, vol. 260, pp. 466–477, Oct. 2017.
- [41] D. He, J. Xu, and X. Chen, "Information-theoretic-entropy based weight aggregation method in multiple-attribute group decision-making," *Entropy*, vol. 18, no. 6, p. 171, Jun. 2016.
- [42] W. W. Zachary, "An information flow model for conflict and fission in small groups," *J. Anthropol. Res.*, vol. 33, no. 4, pp. 452–473, 1977.
- [43] P. M. Gleiser and L. Danon, "Community structure in jazz," *Adv. Complex Syst.*, vol. 6, no. 4, pp. 565–573, Dec. 2003.
- [44] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 68, no. 6, Dec. 2003, Art. no. 065103.
- [45] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, nos. 1–2, pp. 81–93, Jun. 1938.



YUANZHI YANG received the B.S. and M.S. degrees from Air Force Engineering University, Xi'an, China, in 2014 and 2017, respectively, where he is currently pursuing the Ph.D. degree with the Aeronautics Engineering College.

His current research interests include complex networks, failure propagation, and machine learning.



XING WANG is currently a Professor with Air Force Engineering University.

His research interests include electronic countermeasures, machine learning, and artificial intelligence.



YOU CHEN received the Ph.D. degree from Air Force Engineering University, Xi'an, China, in 2011.

He is currently a Lecturer with Air Force Engineering University. His research interests include pattern recognition, data mining, machine learning, and artificial intelligence.



MIN HU received the B.S. degree from the School of Petroleum and Chemical Technology, Liaoning Shihua University, China, in 2017. She is currently pursuing the master's degree with the School of New Energy and Materials, Southwest Petroleum University, China.

Her research interests include the theoretical calculation of materials, and corrosion and protection.



CHENGWEI RUAN received the B.S., M.S., and Ph.D. degrees from Air Force Engineering University, Xi'an, China, in 2009, 2013, and 2017, respectively.

He works with 95910 Army. His current research interests include complex networks, pattern recognition, and artificial intelligence.

• • •