


# An extended improved global structure model for influential node identification in complex networks

Jing-Cheng Zhu(朱敬成) and Lun-Wen Wang(王伦文)

**Citation:**Chin. Phys. B, 2022, 31 (**6**): 068904. DOI: 10.1088/1674-1056/ac380d
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

**What follows is a list of articles you may be interested in**

---

## Synchronization in multilayer networks through different coupling mechanisms

Xiang Ling(凌翔), Bo Hua(华博), Ning Guo(郭宁), Kong-Jin Zhu(朱孔金), Jia-Jia Chen(陈佳佳), Chao-Yun Wu(吴超云), and Qing-Yi Hao(郝庆一)
Chin. Phys. B, 2022, 31 (**4**): 048901. DOI: 10.1088/1674-1056/ac2b19

## Cascading failures of overload behaviors using a new coupled network model between edges

Yu-Wei Yan(严玉为), Yuan Jiang(蒋沅), Rong-Bin Yu(余荣斌), Song-Qing Yang(杨松青), and Cheng Hong(洪成)
Chin. Phys. B, 2022, 31 (**1**): 018901. DOI: 10.1088/1674-1056/ac1337

## Detection of influential nodes with multi-scale information

Jing-En Wang(王静恩), San-Yang Liu(刘三阳), Ahmed Aljmiai, and Yi-Guang Bai(白艺光)
Chin. Phys. B, 2021, 30 (**8**): 088902. DOI: 10.1088/1674-1056/abff2d

## LCH: A local clustering H-index centrality measure for identifying and ranking influential nodes in complex networks

Gui-Qiong Xu(徐桂琼), Lei Meng(孟蕾), Deng-Qin Tu(涂登琴), and Ping-Le Yang(杨平乐)
Chin. Phys. B, 2021, 30 (**8**): 088901. DOI: 10.1088/1674-1056/abea86

## Exploring individuals' effective preventive measures against epidemics through reinforcement learning

Ya-Peng Cui(崔亚鹏), Shun-Jiang Ni (倪顺江), and Shi-Fei Shen(申世飞)
Chin. Phys. B, 2021, 30 (**4**): 048901. DOI: 10.1088/1674-1056/abcfa5

---

# An extended improved global structure model for influential node identification in complex networks

Jing-Cheng Zhu(朱敬成) and Lun-Wen Wang(王伦文)†

*College of Electronic Engineering, National University of Defense Technology, Hefei* 230037, *China*

(Received 2 August 2021; revised manuscript received 29 October 2021; accepted manuscript online 10 November 2021)

Accurate identification of influential nodes facilitates the control of rumor propagation and interrupts the spread of computer viruses. Many classical approaches have been proposed by researchers regarding different aspects. To explore the impact of location information in depth, this paper proposes an improved global structure model to characterize the influence of nodes. The method considers both the node's self-information and the role of the location information of neighboring nodes. First, degree centrality of each node is calculated, and then degree value of each node is used to represent self-influence, and degree values of the neighbor layer nodes are divided by the power of the path length, which is path attenuation used to represent global influence. Finally, an extended improved global structure model that considers the nearest neighbor information after combining self-influence and global influence is proposed to identify influential nodes. In this paper, the propagation process of a real network is obtained by simulation with the SIR model, and the effectiveness of the proposed method is verified from two aspects of discrimination and accuracy. The experimental results show that the proposed method is more accurate in identifying influential nodes than other comparative methods with multiple networks.

**Keywords:** complex network, influential nodes, extended improved global structure model, SIR model

**PACS:** 89.75.Fb **DOI:** 10.1088/1674-1056/ac380d

## 1. Introduction

Complex networks have attracted the wide attention of researchers in recent years. One important research direction is to rank the nodes in the network, which is of great significance for clarifying network structure and maintaining network operation. In fact, various complex systems that accompany human daily life can be expressed as complex networks, such as scientific cooperation networks,[1] transportation networks,[2] social networks[3] and protein–protein interaction networks.[4] Research shows that a small number of nodes in the network play a supporting role in the whole network. Finding these nodes has become a research hotspot. Studying the structure information of these nodes can give a deeper understanding of the information propagation between nodes, which is of great significance to the study of node influence.

Many classical identification methods have been proposed for node influence, such as degree centrality,[5] betweenness centrality,[6] closeness centrality,[7] eigenvector centrality,[8] K-shell decomposition,[9] H-index,[10] etc. Among these methods, degree centrality is the simplest and most intuitive method, and it reflects the most local information of nodes, which can be expressed by the number of nearest neighbor nodes. However, the degree centrality method does not distinguish well among the influence of each node, and it is easy to define some nodes with the same value, which eventually leads to the low accuracy of the ranking results. To solve the problem of discrimination and accuracy, Zeng *et al.*[11] proposed a mixed degree decomposition method, which can be converted into degree centrality and K-shell decomposition method by adjusting parameters. Sheikhahmadi and Nematbakhsh[12] proposed a method of mixing core, degree and entropy information, considering the low discrimination between K-shell and degree centrality. Compared with K-shell and degree centrality, this method has a certain distinction, but some nodes have not been fully considered. Ibnoulouafi and Haziti[13] defined a new node influence identification method based on the density formula, which takes the degree centrality of the node itself as the mass and the path length from the node as the radius. This method improves the discrimination and accuracy, but does not fully consider the information of the surrounding nodes. Li *et al.*[14] proposed clustering local-degree (CLD) centrality by combining the clustering coefficient of nodes with the degree value of the nearest neighbor nodes. This method considers the information of the nearest neighbor nodes, but ignores the influence of other neighbor nodes. According to the gravity model, Li *et al.*[15] regarded the degree centrality of their own and neighbor layer nodes as the mass of mutually attractive celestial bodies in the gravity model, and regarded the path length between the nodes as the distance between celestial bodies. Yan *et al.*[16] proposed a node ranking method that integrated node information of itself, the neighbor layer and location information. This method first obtained a mixed multi-index method by using the entropy weight method, and then applied it to the gravitational law formula, and analyzed multiple combinations in

†Corresponding author. E-mail: wanglunwenmust@nudt.edu.cn

© 2022 Chinese Physical Society and IOP Publishing Ltd 

the datasets, indicating that the new method is more accurate than the classical method. Yang *et al.*[17] first proposed an improved K-shell decomposition method, and then combined it with degree centrality. At the same time, the influence of path length was introduced, and the performance of discrimination and accuracy was improved. Ullah *et al.*[18] proposed a global structure model to rank nodes. The model takes the K-shell value as its own influence, and combines the K-shell value of neighbor layer nodes with the path length as a global influence. Although the path length is introduced, the influence of path length is not fully considered.

The above methods solve the problems of the poor identification effect and low accuracy in the process of node identification from different perspectives, but do not fully consider the impact of the location information of nodes. In order to further explore the influence of location information between neighbor layer nodes and measured nodes, this paper proposes an extended improved global structure model to identify influential nodes in the network. This method first calculates the degree centrality of each node, and uses the degree information of the node itself to represent the influence of the node itself. Regarding the influence of the neighbor layer nodes, this paper introduces the power of the path length to show the attenuation effect in the information transmission process caused by the distance. Finally, the sum operation of the nearest neighbor node information is carried out to rank the influence of the node. To verify the effectiveness of the proposed method, 1000 independent simulation experiments are carried out on 11 real networks using the SIR propagation model[19] to obtain the propagation influence of nodes, and the results of each method are compared with the node influence obtained by the SIR model. The experiments show that the proposed EIGSM method in this paper performs better than DC,[5] BC,[6] K-shell,[9] MCDE,[12] and GSM.[18]

This section describes the research background and research status, the next sections describe the content as follows. Section 2 describes benchmark methods and the proposed EIGSM method. Section 3 describes the datasets used in this paper and the corresponding evaluation criteria. Section 4 shows the relevant experiments for discrimination and accuracy. Conclusions are given in Section 5.

## 2. Related work

An unweighted and undirected network can be represented by $G = (V, E)$, where $V$ denotes the set of nodes of the network and $E$ refers to the set of connected edge information. Meanwhile, we can represent the node and edge information of the whole network by the adjacency matrix $A = \{a_{ij}\}$, and use the size of the matrix to represent the number of nodes, and the matrix element $a_{ij}$ to represent the edge connection. If $a_{ij} = 1$, it means that the node $i$ and $j$ are directly connected, otherwise they are not directly connected.

### 2.1. Benchmark methods

#### 2.1.1. Degree centrality

Degree centrality[5] reflects the most local information of nodes, which is simple and intuitive, and can be directly expressed by the number of connected nodes. The larger the degree is, the greater the node influence is. The degree centrality formula of node $i$ can be expressed as

$$\mathrm{DC}(i) = \sum_{j=1}^{n} a_{ij}, \tag{1}$$

where $n$ means the total number of nodes in the network, which can be represented by the size of the adjacency matrix.

#### 2.1.2. Betweenness centrality

Betweenness centrality of a node in the network[6] refers to the ratio of the number of the shortest paths through the node to the number of the shortest paths of all nodes, and the formula can be expressed as

$$\mathrm{BC}(i) = \sum_{s \neq t \neq i \in V} \frac{g_{st}^{i}}{g_{st}}; \tag{2}$$

in the formula, $g_{st}^{i}$ denotes the number of shortest paths of any two nodes passing through node $i$ in node set $V$, and $g_{st}$ denotes the number of shortest paths of any pair of nodes.

#### 2.1.3. K-shell decomposition method

The K-shell decomposition method[9] separates the nodes into different layers, and the later the node is separated, the more important the node is. The specific steps of K-shell decomposition are as follows.

(i) Firstly, the degree of each node is obtained according to the adjacency matrix of the network.

(ii) Delete the node whose degree value is 1 and recalculate the degree value of the network. If there is a node whose degree value is 1, then continue to delete the node until the degree value is greater than 1 and divide all deleted nodes into the first layer.

(iii) Strip nodes of other degrees step by step according to (ii) until no node can be stripped.

#### 2.1.4. Mixed core, degree, and entropy

Mixed core, degree and entropy (MCDE)[12] takes the node's own core and degree values into account as well as the effects of the node's friends, as shown in the following equation:

$$\mathrm{MCDE}(i) = \alpha \mathrm{Core}(i) + \beta \mathrm{DC}(i) + \gamma \mathrm{Entropy}(i),$$

$$\mathrm{Entropy}(i) = -\sum_{j=0}^{\mathrm{Core}_{\max}} (p_j \times \log_2 p_j),$$

$$p_j = \frac{\text{Count}(i\text{'s friends in core } j)}{\text{DC}(i)}, \tag{3}$$

where $\text{Core}(i)$ and $\text{DC}(i)$ represent the kernel value and degree value of node $i$, respectively. Entropy information reflects the dispersion degree of friends of node $i$ in different kernel values.

#### 2.1.5. Global structure model

Global structure model (GSM)[18] considers the node's self-influence and global influence, and the influence of node $i$ can be expressed as

$$\text{GSM}(i) = \text{SI}(i) \times \text{GI}(i) = e^{\frac{\text{Ks}(i)}{n}} \times \sum_{i \neq j} \frac{\text{Ks}(j)}{d_{ij}}, \tag{4}$$

where $\text{Ks}(i)$ refers to the K-shell decomposition value of node $i$, and $d_{ij}$ refers to the path length between node $i$ and node $j$.

### 2.2. Proposed extended improved global structure model

In current research on node influence, researchers mostly focus on the K-shell decomposition method and the DC method. The main reason is that these two methods are very simple. Researchers can simply obtain the corresponding values of nodes in the network, which also allows these two methods to be widely used in large-scale networks and have a wider range of applications. However, both the K-shell decomposition method and the DC method are faced with a problem: they cannot well distinguish the importance of each node in a network. The K-shell decomposition method is characterized by a small number of levels, and it is easy to divide multiple nodes into the same level. The disadvantage of the DC method is that it only reflects the edge information of nodes, and there must be a large number of nodes with the same number of edges in a large-scale network. It is worth noting that nodes with the same value in the network also have different influences due to the influence of location information. Based on the GSM method, this paper makes improvements. First of all, considering that the DC method is better than the K-shell decomposition method in distinguishing the influence of nodes, this paper uses the DC method instead of the K-shell decomposition method, and uses the degree centrality information of nodes and the total number of network nodes to represent their own influence. At the same time, in addition to the node's own information, the location information of the node will have a certain impact on information dissemination. In this paper, the neighbor layer information is used to represent the influence of the location information on the measured node. A node with a distance of 1 is divided into neighbor layer 1, and a node with a distance of 2 is divided into neighbor layer 2. On account of information loss, the farther the distance from the measured node is, the weaker the ability to receive information is. This paper sets a path loss related to the shortest path length between nodes and the average degree. The formula is as follows:

$$\begin{aligned}
&\text{improved_SI}(i) = \exp\left(\frac{\text{DC}(i)}{n}\right),\\
&\text{improved_GI}(i) = \sum_{i \neq j} \frac{\text{DC}(j)}{d_{ij}^{\text{ceil}(\log_2(\text{aver_degree}))}},\\
&\text{IGSM}(i) = \text{improved_SI}(i) \times \text{improved_GI}(i)\\
&\qquad = \exp\left(\frac{\text{DC}(i)}{n}\right) \times \sum_{i \neq j} \frac{\text{DC}(j)}{d_{ij}^{\text{ceil}(\log_2(\text{aver_degree}))}}.
\end{aligned} \tag{5}$$

To consider the nearest neighbor information of nodes, an extended method called the EIGSM method is proposed according to formula (5), which defines the formula as

$$\text{EIGSM}(i) = \sum_{j \in \Gamma_i} \text{IGSM}(j), \tag{6}$$

where a refers to the nearest neighbor set of nodes $i$.

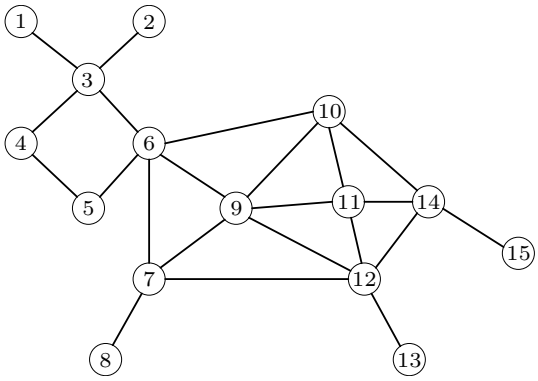



**Fig. 1.** Example network.

Figure 1 is an example network of 15 nodes and 22 edges, which shows that the degree values of each node are $\text{DC}(1) = \text{DC}(2) = \text{DC}(8) = \text{DC}(13) = \text{DC}(15) = 1$, $\text{DC}(4) = \text{DC}(5) = 2$, $\text{DC}(3) = \text{DC}(7) = \text{DC}(10) = \text{DC}(11) = \text{DC}(14) = 4$, $\text{DC}(6) = \text{DC}(9) = \text{DC}(12) = 5$. Then, we have $\text{average_degree} = (5 \times 1 + 2 \times 2 + 5 \times 4 + 3 \times 5)/15 = 2.9333$; we can get that $\text{ceil}(\log_2(\text{aver_degree})) = \text{ceil}(\log_2(2.9333)) = \text{ceil}(1.553) = 2$. Taking node 9 as an example, the first neighbor layer nodes $\text{layer1} = \{6,7,10,11,12\}$, the second neighbor layer nodes $\text{layer2} = \{3,5,8,13,14\}$, and the third neighbor layer nodes $\text{layer3} = \{1,2,4,15\}$. We can have $\text{improved_SI}(9) = \exp(5/15) = 1.3956$, $\text{improved_GI}(9) = \text{sum}(\text{DC}(\text{layer1})) + \text{sum}(\text{DC}(\text{layer2}))/2^2 + \text{sum}(\text{DC}(\text{layer3}))/3^2 = 22 + 12/2^2 + 5/3^2 = 25.5556$, and then IGSM can be calculated as $\text{IGSM}(9) = 1.3956 \times 25.5556 \approx 35.67$. Finally, we can get an extended IGSM(9) as $\text{EIGSM}(9) = \text{sum}(\text{IGSM}(\text{layer1})) \approx 149.19$. Table 1 shows DC, IGSM, and EIGSM values for each node, with IGSM and EIGSM values showing two decimal places.

**Table 1.** DC, IGSM, and EIGSM values for each node in the sample network.

| Node | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 1 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DC | 1 | 1 | 4 | 2 | 2 | 5 | 4 | 1 | 5 | 4 | 4 | 5 | 1 | 4 | 1 |
| IGSM | 9.22 | 9.22 | 18.84 | 11.60 | 15.03 | 33.11 | 27.82 | 10.87 | 35.67 | 29.59 | 28.43 | 30.24 | 11.77 | 24.52 | 10.17 |
| EIGSM | 18.84 | 18.84 | 63.14 | 33.87 | 44.70 | 126.95 | 109.89 | 27.82 | 149.19 | 121.72 | 120.01 | 128.20 | 30.24 | 98.44 | 24.52 |

In order to show the effect of distinguishing the influence of each node, Table 2 shows the ranking of each node. From the table, DC divides nodes into 4 levels, BC divides into 11 levels, K-shell decomposition divides into 3 levels, MCDE divides into 9 levels, GSM divides into 13 levels, IGSM and EIGSM divides into 14 levels. Node 1 and node 2 cannot be distinguished because the two nodes are locally symmetric and have the same information about themselves and their neighbors, so it can be considered that the influence of the nodes is the same. From the degree of distinction, DC can distinguish more levels than the K-shell decomposition method. The IGSM and EIGSM methods are better than the GSM method in node influence ranking. In terms of differentiation effects, DC can distinguish more layers than the K-shell decomposition method, and the IGSM and EIGSM methods distinguish better than the GSM method in terms of node influence ranking.

**Table 2.** Ranking results of different methods for each node in the example network.

| Rank | DC | BC | Ks | MCDE | GSM | IGSM | EIGSM |
|---|---|---|---|---|---|---|---|
| 1 | 6,9,12 | 6 | 6,7,9,10, 11,12,14 | 6 | 9 | 9 | 9 |
| 2 | 3,7,10,11,14 | 3 | 3,4,5 | 12 | 6 | 6 | 12 |
| 3 | 4,5 | 7 | others | 9 | 10 | 12 | 6 |
| 4 | others | 12 | | 7,14 | 12 | 10 | 10 |
| 5 | | 10 | | 3 | 7,11 | 11 | 11 |
| 6 | | 14 | | 10,11 | 14 | 7 | 7 |
| 7 | | 9 | | 5 | 3 | 14 | 14 |
| 8 | | 5 | | 4 | 5 | 3 | 3 |
| 9 | | 4 | | others | 13 | 5 | 5 |
| 10 | | 11 | | | 8 | 13 | 4 |
| 11 | | others | | | 4 | 4 | 13 |
| 12 | | | | | 15 | 8 | 8 |
| 13 | | | | | 1,2 | 15 | 15 |
| 14 | | | | | | 1,2 | 1,2 |

## 3. Datasets and evaluation criteria

### 3.1. Datasets

This paper selects several real network datasets with different structures to verify the effectiveness of the algorithm. The real networks are the Contiguous network,[20] Dolphin network,[21] Polbooks network,[22] Word network,[23] Jazz network,[24] Slavko network,[22] USAir network,[25] Netscience network,[23] Infectious network,[26] Elegans network,[27] and Email network.[28] Some basic parameters of the networks are shown in Table 3. The parameters shown in the table include the total number of nodes $n$, the total number of connected edges between nodes $m$, the propagation threshold $\beta_{\text{th}} = \langle k\rangle/\langle k^2\rangle$, $\langle k\rangle$ is the average degree, $\langle k^2\rangle$ is the second-order average degree, $\beta$ refers to the propagation probability, $k_{\max}$ refers the maximum degree, $D$ refers to the network diameter, $C$ is the average clustering coefficient and $r$ is the assortativity coefficient.[29]

**Table 3.** Basic parameters of real networks.

| Network | $n$ | $m$ | $\beta_{\text{th}}$ | $\beta$ | $k_{\max}$ | $\langle k\rangle$ | $D$ | $C$ | $r$ |
|---|---|---|---|---|---|---|---|---|---|
| Contiguous | 49 | 107 | 0.2027 | 0.21 | 8 | 4.367 | 11 | 0.497 | 0.2334 |
| Dolphin | 62 | 159 | 0.1470 | 0.15 | 12 | 5.129 | 8 | 0.259 | −0.0436 |
| Polbooks | 105 | 441 | 0.0838 | 0.09 | 25 | 8.4 | 7 | 0.488 | −0.1279 |
| Word | 112 | 425 | 0.0726 | 0.08 | 49 | 7.589 | 5 | 0.173 | −0.1293 |
| Jazz | 198 | 2742 | 0.0259 | 0.03 | 100 | 27.697 | 6 | 0.618 | 0.0202 |
| Slavko | 324 | 2218 | 0.0466 | 0.05 | 58 | 13.691 | 7 | 0.466 | 0.2473 |
| USAir | 332 | 2126 | 0.0225 | 0.03 | 139 | 12.807 | 6 | 0.625 | −0.2079 |
| Netscience | 379 | 914 | 0.1247 | 0.13 | 34 | 4.823 | 17 | 0.741 | −0.0817 |
| Infectious | 410 | 2765 | 0.0534 | 0.06 | 50 | 13.488 | 9 | 0.456 | 0.2258 |
| Elegans | 453 | 2025 | 0.0249 | 0.03 | 237 | 8.940 | 7 | 0.647 | −0.2258 |
| Email | 1133 | 5451 | 0.0535 | 0.06 | 71 | 9.622 | 8 | 0.220 | 0.0782 |

### 3.2. SIR model

The SIR model[19] is widely used to simulate the process of information dissemination and disease infection. In this paper, the SIR model is used to simulate the real influence of nodes. In the SIR propagation model, nodes can exist in three states: Susceptible (S) state, the nodes are in normal state and have the risk of being infected; infected (I) state, the nodes are infected with disease and will infect the surrounding nodes with a certain probability; recovery (R) state, the nodes recover from the infected state and will not be infected again. The specific step of simulating node influence is to set the initial state as an infected node; then the infected node infects the nearest neighbor node with a certain probability and the infected node recovers with the probability of $\lambda = 1$ until the whole process is stable. In this paper, the probability near the propagation threshold[30] is adopted for the experiment. The total number of infected nodes is used as the propagation influence of the nodes, and 1000 independent simulation experiments are carried out to obtain the average value. The results of the SIR model will be used for comparative analysis with various methods to test the excellence of the proposed method.

### 3.3. Comprehensive cumulative distribution function

Comprehensive cumulative distribution function (CCDF)[31] can fully describe the probability distribution of the values obtained by each method. The CCDF function

value corresponding to the coordinates ranked $r$ in the ranking list is the sum of probabilities greater than $r$, and the CCDF formula can be expressed as

$$\mathrm{CCDF}(r) = 1 - \frac{\sum_{i=1}^{r} n_i}{n}. \tag{7}$$

In the formula, $n$ is the total number of network nodes, and $\sum_{i=1}^{r} n_i$ refers to the number of numerical rankings less than or equal to $r$ in the ranking list. The distinction effect of the node influence can be observed by the decline of each ranking. If the ranking in the ranking list is closer to $n$, the distinction effect is better.

### 3.4. Kendall coefficient

The Kendell coefficient[32] is used to evaluate the correlation between the ranking results obtained by the SIR model and the ranking results obtained by specific methods. It is assumed that $(x_i, y_i)$ and $(x_j, y_j)$ are two elements in the two-dimensional array $(X, Y)$. If $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$, the pair of elements is said to be concordant. If $x_i > x_j$ and $y_i < y_j$ or $x_i < x_j$ and $y_i > y_j$, the pair of elements is said to be discordant. If $x_i = x_j$ or $y_i = y_j$, then these two elements are neither concordant pairs nor discordant pairs. The formula is defined as

$$\tau(X,Y) = \frac{2(C-D)}{n(n-1)}, \tag{8}$$

where $C$ and $D$ are the numbers of concordant pairs and discordant pairs, respectively.

### 3.5. Jaccard similarity coefficient

The Jaccard similarity coefficient[33] is used to compare the similarity between the ranking results obtained by each method and the SIR model. The Jaccard similarity $J_r$ can be expressed by the intersection and union of two arrays, and the specific formula is

$$J_r(X,Y) = \frac{|X(r) \cap Y(r)|}{|X(r) \cup Y(r)|}. \tag{9}$$

$X(r)$ and $Y(r)$ denote the first $r$ elements in the two lists $X$ and $Y$, respectively. The value range of $J_r$ is $[0, 1]$, and the closer to 1 this value is, the higher the similarity between the two lists.

## 4. Experiment and analysis

In this section, the distinction and accuracy of the identification method are experimentally and analytically studied. The distinction is used to avoid a large number of nodes being defined as the same value, and the accuracy is used to verify the rationality of the sorting results.

### 4.1. Discrimination experiments

In the network, there is a problem that some nodes are defined as the same value because the surrounding information of the nodes are fully considered. In this case, it is difficult to give the node influence ranking results of the same value nodes. In order to avoid this situation of the ranking results, it is necessary that the proposed method can distinguish the influence of each node as much as possible. Figure 2 shows the probability distribution of the ranking results of each method with six networks: Contiguous, Polbooks, USAir, Infectious, Elegans and Email. From the results of the figure, the probability distribution of the K-shell decomposition method and DC method is only concentrated in the front part, which shows that these two methods define many nodes as the same value. At the same time, we can also see that the DC method has better discrimination effect than the K-shell method. MCDE combines DC, K-shell and entropy information. It can be seen that the MCDE method can already distinguish the top-ranking nodes, but the effect of distinguishing the bottom-ranking nodes is not obvious. This is because the entropy information only considers the node's own degree value and the nearest neighbor's K-shell information, and there will be many similar structures for the nodes with small degree values, leading to the weakening of the distinguishing ability. BC aims to find nodes in the hub position of the network, so BC has a good distinction effect in the early stage. We will find that BC has a high point in the later period because there are also a large number of nodes not in the hub position in the network. These nodes will also show different influences due to the influence of the surrounding environment, which is not considered by BC. With regard to GSM, IGSM, and EIGSM, we can see that the influence of each node can be basically identified, and even if the same value node appears, it also remains at a very low point. In the Elegans network, it can be seen that the EIGSM method shows a better distinction effect than other methods.

Figure 3 shows the CCDF curves of each method. We can intuitively understand the discrimination of each method by observing the declining trend of the curves. When the curve has no inflection points as a straight line, it indicates that each node is defined as different values, and the faster the curve drops, the greater the number of nodes assigned to the same ranking. In Fig. 3, the DC, K-shell, and MCDE methods show an obvious downward trend. BC effectively distinguished nodes in the early stage, and there was a sudden drop point in the later stage. GSM, IGSM, and EIGSM fell more slowly than other methods. In the Contiguous and Polbooks networks, EIGSM completely distinguishes the influence of all nodes. In the Infectious and Email networks, the three approaches tend to decline approximately in a straight line. In the USAir and Elegans networks, the EIGSM method also obviously shows a slower decline rate than other methods. From the results, it can be seen that the EIGSM method can well distinguish the influence of each node.

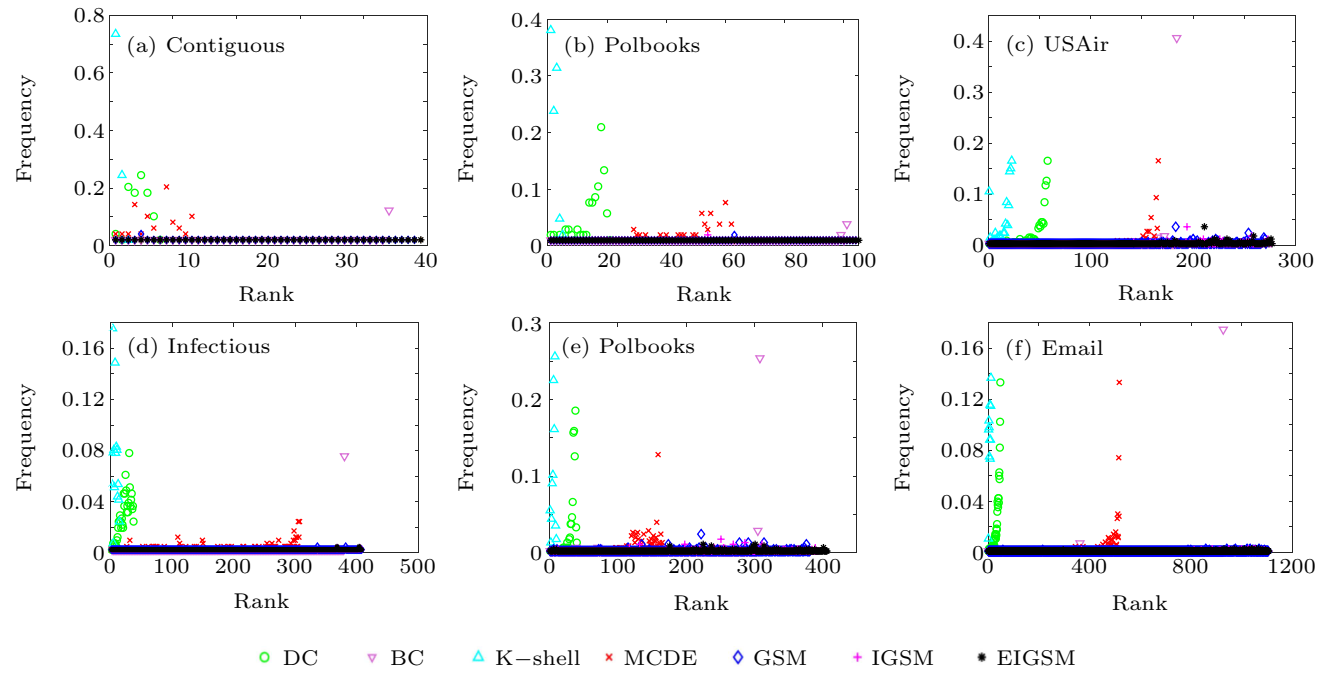



**Fig. 2.** Probability distribution in the ranking list obtained by each method.

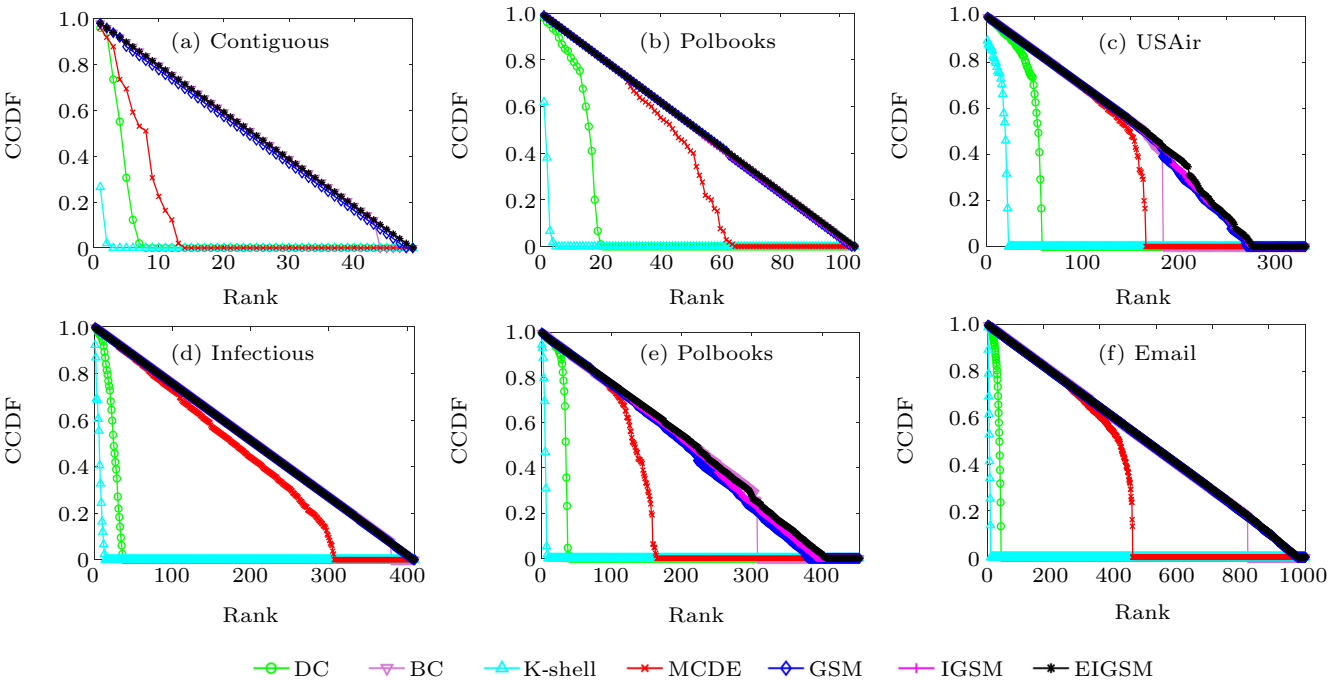



**Fig. 3.** CCDF diagram of ranking results of each method.

### 4.2. Accuracy experiment

The accuracy experiment mainly focuses on three aspects: the Kendell coefficient, the scatter plot of influence consistency and the Jaccard similarity coefficient.

#### 4.2.1. Selection of the power of the average path length

Figure 4 shows the influence of power selection in each network on the accuracy of node recognition. The $x$-axis is the value of the power, and the $y$-axis is the Kendall coefficient between the EIGSM and the SIR model. From the figure, we can see that different networks have different power values. At this time, we begin to think about how to choose the appropriate power value. Then, we think of the parameter of the average degree. The main reason for selecting the average degree is because this paper focuses on the degree of nodes, and the average degree is the average of all node degrees, so that we do not need to recalculate other things, which is simple and convenient. Figure 5 shows the relationship between the optimal

$k$ value of each network power in Fig. 4 and ceil$(\log_2\langle k\rangle)$. Because the coordinates are repeated, the size of the five-pointed star represents the number of repetitions. From Fig. 5, it can be obtained that the five-pointed stars are approximately distributed around the straight line $y = x$ by using the proportional function. So we finally determine that $k_{\text{optimal}} \approx \text{ceil}(\log_2\langle k\rangle)$.

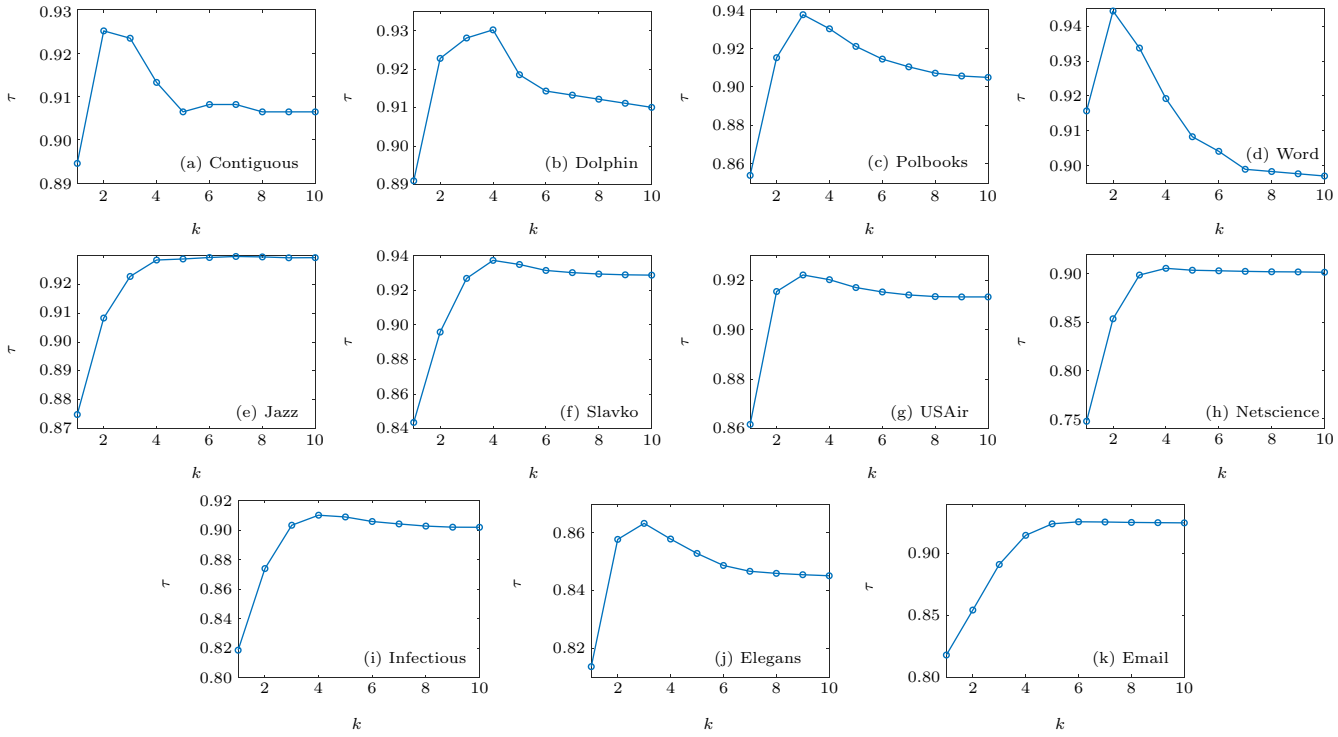



**Fig. 4.** Influence of power of the shortest path on each network.

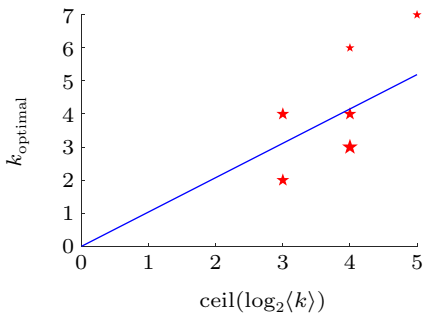



**Fig. 5.** The relationship between the optimal $k$ value and the average degree.

### 4.2.2. Influence consistency experiment

Figure 6 shows the scatterplot of the influence consistency between the node influence values obtained by each method on the Dolphin, Word and Slavko networks and the values obtained by SIR model simulation. Each scatter in the figure represents a node. The better the monotonicity of the curve trend shown in the scatter plot is, the higher the correlation is, and the closer the scatter is, the better the effect is. It can be seen from the figure that the scatter plot distribution of BC is very divergent. At the same time, it can also be observed that the K-shell decomposition method has a weak distinguishing ability, and the same K-shell value corresponds to many nodes. The DC, MCDE, GSM, IGSM, and EIGSM results are positively correlated with node influence. It can be seen that DC and MCDE are not very concentrated in the three networks, and GSM is also dispersed in the Dolphin and Slavko networks. Relatively speaking, the node number obtained by the proposed EIGSM method has the highest correlation with the node influence, and the nodes are relatively concentrated.

**Table 4.** Kendall correlation coefficients of different methods in the real network.

| Network | $\beta$ | $\tau_{\text{DC}}$ | $\tau_{\text{BC}}$ | $\tau_{\text{Ks}}$ | $\tau_{\text{MCDE}}$ | $\tau_{\text{GSM}}$ | $\tau_{\text{IGSM}}$ | $\tau_{\text{EIGSM}}$ |
|---|---|---|---|---|---|---|---|---|
| Contiguous | 0.21 | 0.7058 | 0.5825 | 0.3929 | 0.7100 | 0.8599 | 0.9048 | 0.9235 |
| Dolphin | 0.15 | 0.7721 | 0.5389 | 0.5717 | 0.7562 | 0.7705 | 0.9175 | 0.9281 |
| Polbooks | 0.09 | 0.7518 | 0.3505 | 0.6390 | 0.7647 | 0.6897 | 0.9114 | 0.9302 |
| Word | 0.08 | 0.8311 | 0.6523 | 0.7143 | 0.8197 | 0.9054 | 0.8887 | 0.9337 |
| Jazz | 0.03 | 0.8069 | 0.4569 | 0.7591 | 0.8397 | 0.8975 | 0.8917 | 0.9288 |
| Slavko | 0.05 | 0.7719 | 0.3625 | 0.7392 | 0.7980 | 0.8476 | 0.9112 | 0.9372 |
| USAir | 0.03 | 0.7251 | 0.5081 | 0.7476 | 0.7366 | 0.8622 | 0.8915 | 0.9202 |
| Netscience | 0.13 | 0.5955 | 0.3048 | 0.5219 | 0.6278 | 0.5688 | 0.8205 | 0.8985 |
| Infectious | 0.06 | 0.7281 | 0.3707 | 0.7534 | 0.7715 | 0.7888 | 0.8820 | 0.9101 |
| Elegans | 0.03 | 0.6391 | 0.4757 | 0.6586 | 0.6492 | 0.7208 | 0.7912 | 0.8579 |
| Email | 0.06 | 0.7615 | 0.6203 | 0.7689 | 0.7749 | 0.8580 | 0.9032 | 0.9147 |

Figure 7 shows the correlation scatter diagram of EIGSM and DC, MCDE, GSM, IGSM. The $x$-axis is the value calculated by EIGSM for nodes in the network, and the $y$-axis is the value of the other methods. It can be seen from the figure that the scatter plot is positively correlated. EIGSM and DC show a linear trend from the trend of the scatter plot curve, but due to the discrimination of DC, the scatter distribution is more

dispersed. From the density point of view, the scatter plots of EIGSM and IGSM are most concentrated, while the scatter plots of DC, MCDE, and GSM are more dispersed. Overall, EIGSM has the highest correlation with IGSM.

Table 4 gives the Kendall coefficient of various methods in the real network, and it can be seen that EIGSM has the best effect. This shows that providing neighbor layer nodes information attenuation and considering the nearest neighbor nodes information can improve the accuracy of identifying influential nodes. This shows that the proposed method has more advantages in identifying node influence than DC, BC, K-shell, MCDE, and GSM.

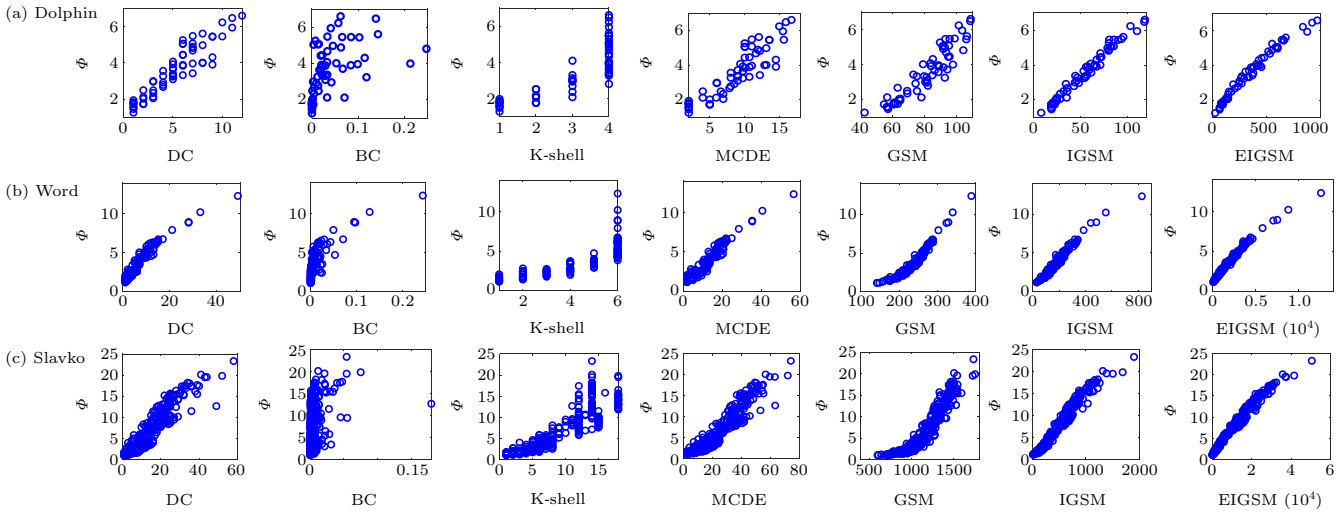



**Fig. 6.** Relationship between each method and node influence.

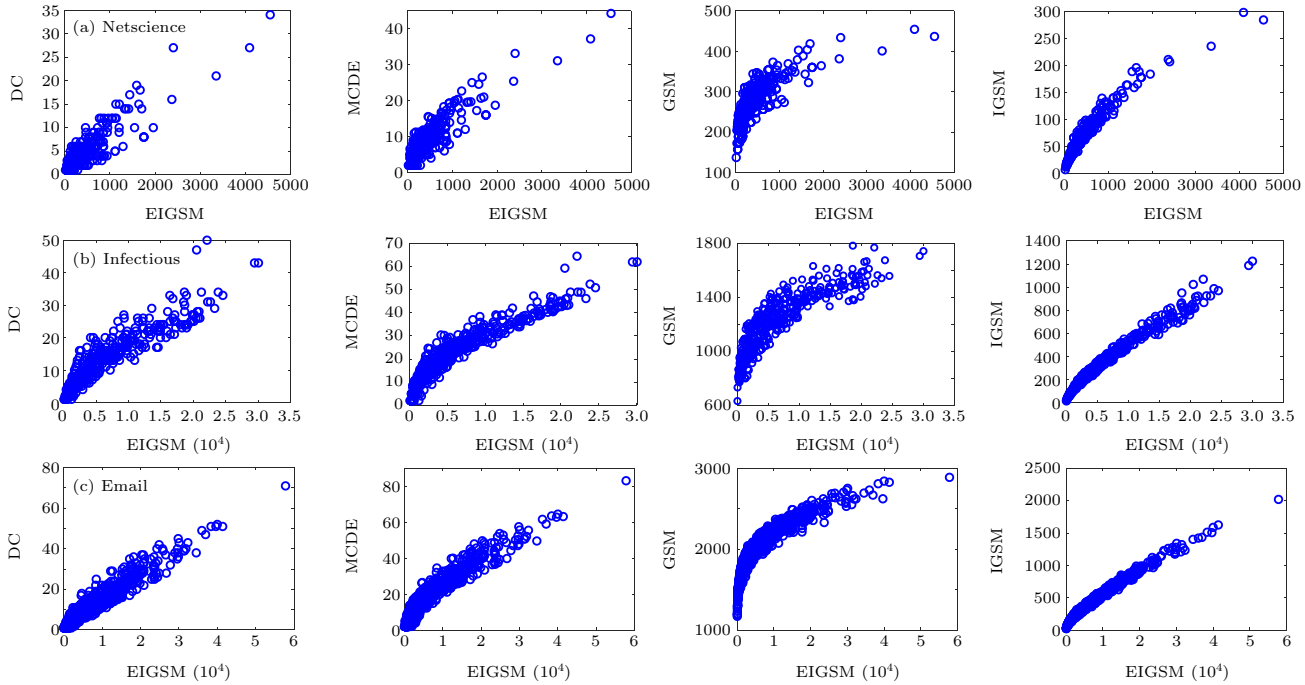



**Fig. 7.** Correlation between the proposed method and other methods.

### 4.2.3. Identification effect under certain propagation probability

In order to explore the influence of different propagation probabilities on the accuracy of the method proposed in this paper, the values around the propagation threshold $\beta_{\text{th}}$ are selected as the propagation probability for the experiments, and the Kendall coefficient was used as the evaluation criterion. The experimental results are shown in Fig. 8. From the figure, it can be seen that BC is always low in several networks, which shows that it is not good at evaluating the propagation influence of the nodes in the network compared to other methods. In the Word, Slavko, Infectious and Email networks, we can see that the Kendall coefficients of the DC and MCDE methods in the early stage are higher. This is because when the propagation probability is small, the information dissemination of the node is easily limited to a small part of the surrounding nodes. At this time, the

directly connected nodes are the easiest to receive information from, and the DC and MCDE methods have the closest relationship with the directly connected nodes. If the propagation probability is too small, it is easy to spread only over a local area, and if the propagation probability is too large, it will spread over a large area. From Fig. 8, it can be seen that EIGSM always maintains an advantage near the propagation threshold, and the identification accuracy is higher.

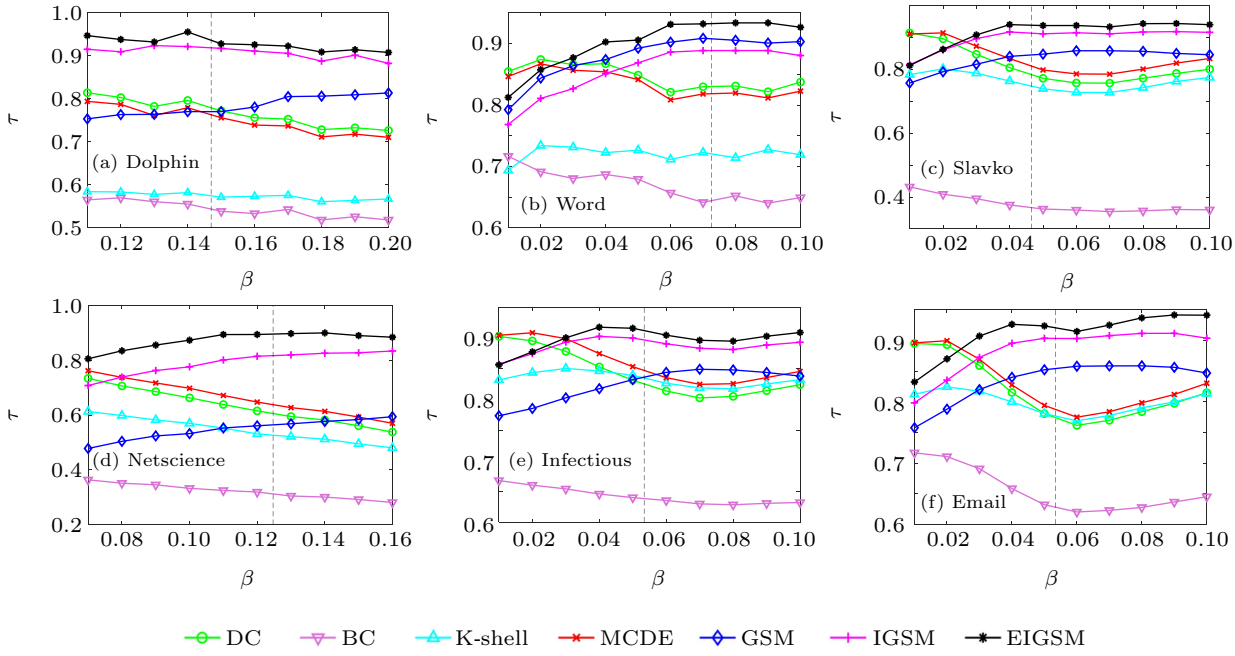



**Fig. 8.** Kendall coefficients between different propagation probabilities and the SIR model.

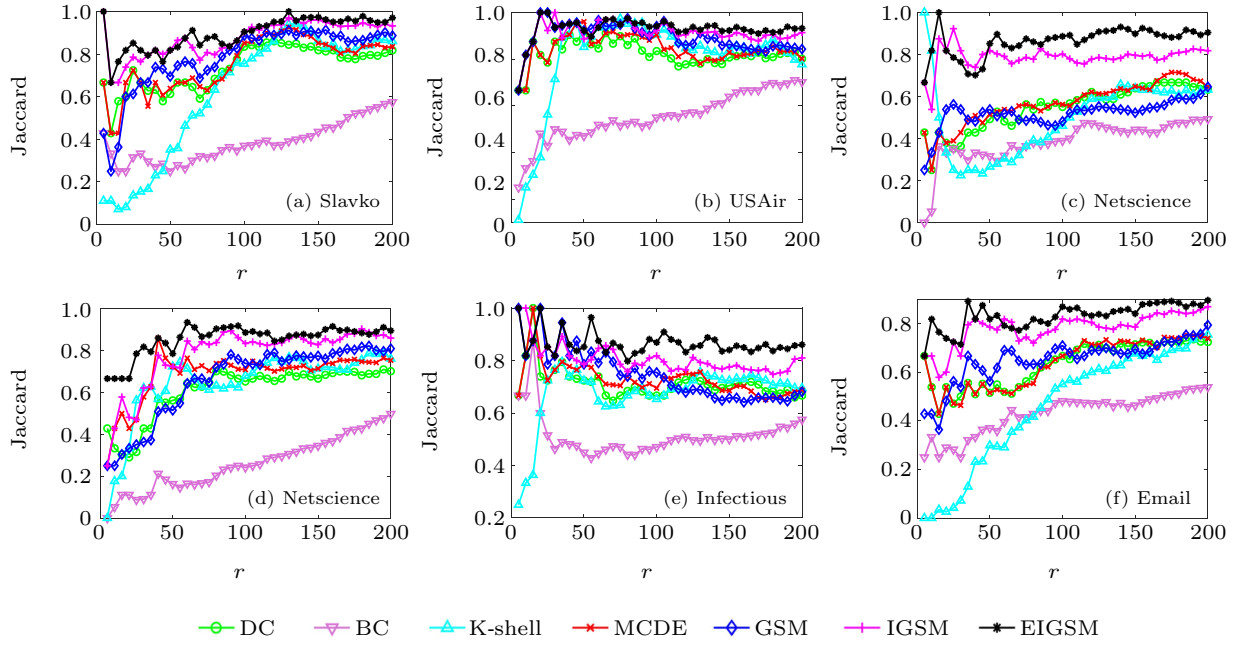



**Fig. 9.** The Jaccard similarity coefficients between each method and the SIR model.

### 4.2.4. Jaccard similarity coefficient

The most influential nodes form a small part, so we should not only judge the accuracy from the overall perspective but also consider the accuracy of the top nodes. Because the network scale is getting larger and larger, we selected the last six networks with a large number of nodes to do the following Jaccard similarity experiment, as shown in Fig. 9. Figure 9 studies the similarity between the simulation results of the SIR model of the forward node obtained by each method through the Jaccard similarity coefficient. The higher the similarity is, the more accurate the result is. The $x$-axis represents the number of front nodes, and the $y$-axis represents the Jaccard similarity coefficient. In Fig. 9, with an increase of the number of front nodes, the results tend to be stable. From

the results of Fig. 9, we see that BC is located below other methods in most parts, while IGSM and EIGSM are located above other methods. In Slavko, USAir, Infectious, Elegans, and Email networks, we can observe that the similarity coefficient of the K-shell decomposition method is very small at the beginning. With the increase of the abscissa, the similarity coefficient increases. This is because K-shell divides many nodes into the same layer, and the accuracy is limited by the discrimination. Moreover, the larger the network scale is, the closer the connection between the nodes and the surrounding nodes is. Therefore, the problem cannot be considered only from the node itself. In the Netscience network, K-shell has a high point at first because the top-ranking nodes are not only in the core position but also have a relatively large degree value. Then it will begin to decline with the discrimination problem. In the Infectious and Email networks, we can also see that EIGSM is significantly more similar than other methods. Overall, EIGSM can be considered to be more accurate in identifying node influence.

## 5. Conclusion

This paper studies the problem of identifying influential nodes in complex networks, which is crucial to the study of disease infection and information dissemination. This paper proposes an extended improved global structure model to identify influential nodes in the network. This method first uses the degree centrality information to represent its self-influence and then uses the neighbor layer information to represent the influence of the surrounding information. When considering the neighbor layer information, a path attenuation related to the path length and the average degree is introduced. Finally, the propagation influence of the node is characterized by accumulating the information of the nearest neighbor node. To evaluate the effectiveness of the proposed method, we use the SIR model to simulate the propagation process, and conduct experiments on two main aspects of discrimination and accuracy in multiple real networks. Firstly, we can see that the proposed method avoids the phenomenon that most nodes are defined as the same importance, and it effectively separates the importance of each node. Secondly, in terms of accuracy, it can be seen that the method proposed in this paper has been improved to a certain extent, especially near the curve; it is always better than other methods. About the computational complexity of the proposed method, because the IGSM method is similar to the GSM method, we think that the computational complexity of IGSM is the same as that of GSM, which is $O(n^2)$. In addition, EIGSM adds a nearest neighbor information overlay from 1 to $n$ on the basis of IGSM, so we believe the computational complexity of this method can be regarded as $O(n^2+n)$; when $n$ is infinite, the computational complexity of this method can also be regarded as $O(n^2)$. Overall, the experimental results show that the proposed method in this paper has more advantages than other methods in identifying influential nodes.

## Acknowledgment

Project supported by the National Natural Science Foundation of China (Grant No. 11975307).

## References

[1] Jung H and Phoa F K H 2021 *Entropy* **23** 502
[2] Mazzarisi P, Zaoli S, Lillo F, Delgado L and Gurtner G 2020 *J. Air Trans. Manag.* **85** 101801
[3] Reilly K H, Wang S M, Gould L H, Baquero M and Crossa A 2021 *J. Transp. Health.* **22** 101128
[4] Pan X Y, Hu L, Hu P W and You Z H 2021 *IEEEACM Trans. Comput. Biol. Bioinform.* **14** 3095947
[5] Freeman L C 1978 *Soc. Networks* **1** 215
[6] Freeman L C 1977 *Sociometry* **40** 35
[7] Sabidussi G 1966 *Psychometrika* **31** 581
[8] Solá L, Romance M, Criado R, Flores J, García D A A and Boccaletti S 2013 *Chaos* **23** 033131
[9] Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H E and Makse H A 2010 *Nat. Phys.* **6** 888
[10] Lü L Y, Zhou T, Zhang Q M and Stanley H E 2016 *Nat. Commun.* **7** 440
[11] Zeng A and Zhang C J 2013 *Phys. Lett. A* **377** 1031
[12] Sheikhahmadi A and Nematbakhsh M A 2017 *J. Inf. Sci.* **43** 412
[13] Ibnoulouafi A and Haziti M E 2018 *Chaos, Solitons Fractals* **114** 69
[14] Li M T, Zhang R S, Hu R J, Yang F, Yao Y B and Yuan Y N 2018 *Int. J. Mod. Phys. B* **32** 1850118
[15] Li Z, Ren T, Ma X Q, Liu S M, Zhang Y X and Zhou T 2019 *Sci. Rep.* **9** 355
[16] Yan X L, Cui Y P and Ni S J 2020 *Chin. Phys. B* **29** 048902
[17] Yang Y Z, Hu M and Huang T Y 2020 *Chin. Phys. B* **29** 088903
[18] Ullah A, Wang B, Sheng J F, Long J, Khan N and Sun Z J 2021 *Sci. Rep.* **11** 6173
[19] Ma Q and Ma J 2017 *Physica A* **465** 312
[20] Contiguous USA network dataset — KONECT, 2017
[21] Lusseau D, Schneider K, Boisseau O J, Haase P, Slooten E and Dawson S M 2003 *Behav. Ecol. Sociobiol.* **54** 396
[22] Blagus N, Šubelj L and Bajec M 2012 *Physica A* **391** 2794
[23] Newman M E J 2006 *Phys. Rev. E* **74** 036104
[24] Gleiser P M and Danon L 2003 *Adv. Complex. Syst.* **6** 565
[25] Batagelj V and Mrvar A 1998 *Connections* **21** 47
[26] Isella L, Stehlé J, Barrat A, Cattuto C, Pinton J F and Broeck W V 2011 *J. Theor. Biol.* **271** 166
[27] Duch J and Arenas A 2005 *Phys. Rev. E* **72** 027104
[28] Guimerà R, Danon L, Díaz-Guilera A, Giralt F and Arenas A 2003 *Phys. Rev. E* **68** 065103
[29] Newman M E J 2002 *Phys. Rev. Lett.* **89** 208701
[30] Castellano C and Pastor-Satorras R 2010 *Phys. Rev. Lett.* **105** 218701
[31] Helton J C 1996 *Reliab. Eng. Syst. Saf.* **54** 145
[32] Knight W R 1966 *J. Am. Stat. Assoc.* **61** 436
[33] Chen X, Zhou J, Liao Z, Liu S and Zhang Y 2020 *Entropy* **22** 848