Information Sciences 578 (2021) 725–749

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

The identification of crucial spreaders in complex networks by effective gravity model

Shuyu Li [a,b], Fuyuan Xiao [b,*]

[a] School of Computer and Information Science, Southwest University, Chongqing 400715, China
[b] School of Big Data and Software Engineering, Chongqing University, Chongqing 401331, China

ARTICLE INFO

Article history:
Received 12 May 2021
Received in revised form 16 July 2021
Accepted 8 August 2021
Available online 12 August 2021

Keywords:
Influence radius
Value information
Effective gravity model
Identification of key nodes
Complex networks

ABSTRACT

A complex network is a network that has the characteristics of small world, clustering, and power-law distribution. The discovery of crucial spreaders, as one of the significant research directions in complex networks, is mainly used to identify nodes that play a key role in the structure and function of the network. The gravity model is a special method in identifying influencers. However, it involves an open issue that is how to determine the range of interaction. In addition, the mass is merely represented by degree of nodes in traditional methods, which is also a thought-provoking question. For the sake of solving the above two problems, this paper presents an effective gravity model which is based on the precise radius and value information. The rough truncation radius is accurately calculated. And the value information, which represents the dissemination ability of the node, is modified to mass. In short, the node's influence range and value score are calculated according to the attributes of each node and the interaction of neighbor's nodes in the network. Compared with other similar methods and state-of-the-art measures, the rationality and superiority of our approach are demonstrated through six experiments on eleven real-world networks.

1. Introduction

The complex network, which has some or all of the properties involving small world, scale-free, attractors, self-organization and self-similarity, is composed of a huge number of nodes and intricate relationships between nodes. Due to the intersectionality and complexity of research, the complex network is widely used in different fields such as seismic networks [1] in the natural world, protein interaction networks [2] in the biological field, food chain networks [3], gene regulation networks [4] and metabolic networks [5], infectious disease transmission networks [6] in the medical field, the power networks [7], transportation networks [8] and the World Wide Web [9] in the engineering field, social networks [10], communication networks [11] and citation networks [12] in human society and so on. In order to master and use the characteristics of the network, the research on complex networks is continuously developed, including community detection, link prediction [13], identification of crucial spreaders and so on. In particular, compared with those less significant nodes, highly influential nodes with special properties are critical to the structure and function of the network.

* Corresponding author.
E-mail addresses: xiaofuyuan@cqu.edu.cn, doctorxiaofy@hotmail.com (F. Xiao).

https://doi.org/10.1016/j.ins.2021.08.026
0020-0255/

Therefore, these nodes in complex networks are gradually being valued by researchers. In the meanwhile, some methods for evaluating influential nodes have also been proposed.

In general, measures for identifying significant spreaders are roughly divided into different categories, which include neighborhood centrality [14], path centrality, eigenvector centrality, and fractal dimension. Degree centrality [15] (DC for short), H-index [16] and k-shell decomposition [17] are characteristic ways based on neighborhood centrality. Two emblematic path-based methods are closeness centrality [18] (CC for short) and betweenness centrality [19] (BC for short). Eigenvector centrality [20] (EC for short) and PageRank centrality [21] (PC for short) are representative of the approaches based on eigenvector centrality. The typical measures based on fractal dimension are the latest local information dimension [22] (LID for short) based on information entropy and fuzzy local dimension [23] (FLD for short) based on fuzzy set [24,25]. Nevertheless, there are certain limitations in the above metrics. DC only considers the local information of a node. Although the k-shell algorithm divides the nodes in the network into different hierarchies, the disadvantage of this method is that the nodes in the same hierarchy cannot be ranked. BC and CC focus on the global structure, which leads to the high computational complexity of the model. Since EC can merely be used in symmetric networks, it has greater limitations. Although the LID and the FLD distinguish the contribution degree of nodes which are at different distances from central nodes, sometimes the accuracy of measures are not high. Lately, some advanced and technical methods are adopt in this field. For example, Tang et al. [26] proposed a discrete shuffled frog-leaping approach which combines deterministic and random search strategies for influence maximization. As a population-based optimization method, gray wolf algorithm was presented by Zareie et al. [27] to study the issue of maximum impact. Pelusi et al. [28] proposed a gravitational search algorithm based on hyperbolic functions to find the optimal balance between exploration and exploitation. An influential marketer user detection algorithm was presented by Zareie et al. [29] to identify the most influential users in social networks which consider the user's interest on the messages.

Recently, a new method which uses gravity model [30] to discover vital nodes in complex networks has been presented by Li et al. The gravity centrality model (GC for short) is inspired by the law of universal gravitation. The degree of a node is taken as the mass, and the average path between the nodes is taken as the distance. Liu et al. [31] made the further amelioration to add a certain weight for each node. In addition, a temporal gravity model which considers using multiple global information was proposed by Bi et al. [32]. The temporal evolution of the networks is captured in this model. The latest generalized gravity model [33] (GGC for short) which was proposed by Li et al. accurately uses the spreading ability of the node as the mass. Nonetheless, it is complicated to properly determine the parameter of this measure.

It is worth noting that there are still two common problems in the above mentioned gravity-based measures. On the one hand, the truncation radius of the node is set to half of the average distance of the network. That means for every nodes in the network, their influence ranges are the same, which is not a reasonable and precise consideration. On the other hand, some traditional methods simply regard the degree of the node as the mass, which would ignore the neighbor's information of a node. This paper puts forward a new approach called effective gravity model (EGM for short) which is based on precise radius and value information to address these two questions mentioned above. The highlight of this paper is that the vague range of influence can be exactly calculated for every node in the network. Because we need to get the farthest influence of a node in the network, the relationship between node and its farthest node is used to obtain it. What's more, another contribution of this paper is that the value information which is used as the mass of the node is innovatively defined. Specifically, the degree of a node and its neighbors' degree distribution information in its local network are taken into consideration. All in all, the influence radius and node quality are amended by this paper, which can competently recognize crucial nodes in complex networks, as demonstrated by six different experiments in eleven real-world networks.

The remaining chapters of this article are organized as follows. Section 2 briefly introduces the preliminary knowledge of complex networks, incorporating classic representative measures and state-of-the-art approaches. An effective gravity model which is based on precise radius and value information is presented in Section 3. In Section 4, the reliability of EGM compared with other similar approaches and its superiority compared with the most advanced methods are demonstrated in six different experiments on eleven real networks. Ultimately, conclusions and reflections on the EGM measure are discussed in Section 5.

2. Preliminaries

To begin with, some basic ways for identifying significant spreaders in complex networks are reviewed in this section. Then newly proposed methods which are based on gravity model are described. Finally, state-of-the-art approaches which are based on fractal dimensions are introduced in detail.

2.1. The classic measures for identifying influencers

2.1.1. Degree centrality

The significance of a node depends on the number of neighbors. Given a network, the normalized DC [15] is defined as follows,

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

where $k_i = \sum_j a_{ij}$ is expressed as the degree of node i, a_{ij} is regarded as the connection between node i and node j in the adjacency matrix, and n signifies the number of nodes in the network.

2.1.2. Closeness centrality

The sum of the distance between a node and all other nodes needs to be calculated in the closeness centrality. A node with a smaller sum means it is closer to all other nodes, which indicates that the importance of this node is higher. The normalized CC [18], which is essentially the reciprocal of the distance, is defined as follows,

$$CC(i) = \frac{n-1}{\sum_{i \neq j} d_{ij}} \tag{2}$$

where n signifies the number of nodes in the network, and d_{ij} is used to stand for the shortest path distance between node i and node j.

2.1.3. Betweenness centrality

In betweenness centrality, the significance of a node is measured by the number of shortest paths that passes through the node. A node which appears in the shortest path between nodes many times means it has a great influence. BC [19], which reflects the importance of the node as a bridge, is defined as follows,

$$BC(i) = \sum_{u \neq i \neq j} \frac{\sigma_{uj}(i)}{\sigma_{uj}} \tag{3}$$

where $\sigma_{uj}(i)$ is used to indicate the number of shortest paths between node u and node j passing through node i, and the number of shortest paths between node u and node j is represented by σ_{uj}.

2.1.4. Eigenvector centrality

Eigenvector centrality, a way used to measure the transmission influence and connectivity between nodes, takes into account the importance of neighbor nodes. EC is of the opinion that links from critical nodes are more valuable than links from trivial nodes. Given the adjacency matrix $A = (a_{ij})$ of the graph, the eigenvector centrality [20] is defined as follows,

$$EC(i) = x_i = \frac{1}{\lambda} \sum_{j=1}^{n} a_{ij} x_j \tag{4}$$

where $\frac{1}{\lambda}(\lambda \neq 0)$ is a proportional constant. After many iterations, when the steady state is reached, $x = [x_1, x_2, x_3, \ldots, x_n]^T$ is expressed in the following matrix form,

$$Ax = \lambda x \tag{5}$$

where x signifies the eigenvector which is corresponding to the eigenvalue λ of matrix A.

2.1.5. PageRank centrality

PageRank centrality [21], a technology that is calculated based on the mutual hyperlinks between web pages, is used to measure the importance of a particular webpage relative to other webpages in the search engine index. The significance of a page is determined by the importance of all pages that only refer to it. In the initial stage, web pages construct a web map through link relationships, where each page is assigned the same PageRank value. Then each page evenly distributes its current PageRank value to the out-links contained in this page, which causes each link to obtain the corresponding weight. Finally, each page sums the weights passed in by all incoming links to this page, which can get a new PageRank score. When each page gets the updated PageRank value, a round of PageRank calculation is completed. The above process is iterated continuously, and when it stabilizes, it is the final result. The PR value of one of the pages is expressed in the following form,

$$PR(i) = \alpha \sum_{j \in O_i} \frac{PR(i)}{L(j)} + \frac{(1-\alpha)}{N} \tag{6}$$

where α denotes the probability of reaching a certain page and continuing to browse backwards, O_i represents the collection of all webpages that have links to i webpages. The number of links out of the webpage j is indicated by $L(j)$, and N signifies the total number of webpages.

2.2. The measures based on gravity model

2.2.1. The gravity centrality model

The gravity centrality method based on the law of universal gravitation takes both local and global information into consideration. In GC, the degree of a node is equivalent to mass, and the shortest distance between two nodes is taken as the distance. The GC [30] of node i is defined as follows,

$$GC(i) = \sum_{i \neq j, d_{ij} \leqslant R} \frac{k_i \times k_j}{d_{ij}^2} \tag{7}$$

where k_i and k_j represent the degrees of node i and node j respectively, and d_{ij} signifies the shortest distance between node i and node j. A truncation radius $R = 0.5\langle d \rangle$ is introduced to solve the noise and time-consuming problem, where $\langle d \rangle$ means the average path length.

2.2.2. The weighted gravity centrality model

The weighted gravity centrality model amends the GC by taking the weight of nodes into consideration. The WGC [31] method that introduces the concept of eigenvectors is defined as follows,

$$WGC(i) = \sum_{i \neq j, d_{ij} \leqslant R} e_i \times \frac{k_i \times k_j}{d_{ij}^2} \tag{8}$$

where d_{ij} is expressed as the shortest distance between node i and node j. The degrees of node i and node j are denoted by k_i and k_j, and e_i stands for the i-th value of the normalized eigenvector of the maximum eigenvalue. The truncation radius is $R = 0.5\langle d \rangle$.

2.2.3. The generalized gravity centrality model

In the generalized gravity centrality model, the mass of a node is represented by spreading ability. The spreading ability of a node depends on the degree and the clustering coefficient. The GGC [33] of node i is defined as follows,

$$GGC(i) = \sum_{d_{ij} \leqslant R \& j} \frac{sp_i \times sp_j}{d_{ij}^2} \tag{9}$$

where $sp_i = e^{-\alpha C_i} \times k_i$ denotes the communication capability of the node i. $C_i = \frac{2n_i}{k_i(k_i-1)}$ is the clustering coefficient. $\alpha(\alpha \geqslant 0)$ is a free parameter that can be flexibly adjusted in practical applications. d_{ij} signifies the distance between node i and node j, and $R = 0.5\langle d \rangle$ stands for truncation radius.

2.3. Two state-of-the-art measures

2.3.1. The local information dimensionality

Shannon entropy [34] is used to measure the number of nodes in each box in the local information dimension. The LID [22], which deliberates about the local structural characteristics around the central node, is defined as follows,

$$LID(i) = -\frac{d}{dlnl}(-p_i(l)lnp_i(l)) \tag{10}$$

where d represents the sign of the derivative, and l means the size of the box. The term $p_i(l) = \frac{n_i(l)}{N}$ is the probability that the information is contained in the box with the central node i of a given box size l, $n_i(l)$ signifies the number of nodes in the box, and N denotes the total number of nodes in the network.

2.3.2. The fuzzy local dimension

Fuzzy local dimension uses fuzzy sets [35] to study the nodes that are different distances [36] from the central node. The contribution of the node is affected by the distance, which is measured on the FLD through the fuzzy set. The FLD [23] that focuses on the local properties of each node is defined as follows,

$$FLD(i) = \frac{d}{dlogr_t} logN_i(r_t, \varepsilon) \tag{11}$$

where r_t is the radius which starts from the center node i. The term $N_i(r_t, \varepsilon) = \frac{\sum_{j=1}^{N} M_{ij}(\varepsilon)}{N_{i\_r}}$ represents the number of fuzzy nodes whose shortest distance is less than the box size ε obtained from the fuzzy set. In this term, N_{i_r} signifies the number of nodes when the shortest distance between nodes i and j is less than the box size ε. And $M_{ij}(\varepsilon) = exp(-\frac{d_{ij}^2}{\varepsilon^2})$ means the membership function when the distance from node j to node i is less than the box size ε, where d_{ij} stands for the shortest distance between the central node i and node j.

3. Proposed approach

The currently proposed gravity based methods set the influence radius to half of the average path length, which makes the range of interaction between nodes unable to be accurately measured in the network. At the same time, another direction needs to be optimized is that mass is only replaced by the degree and the neighbor's information of a node has been

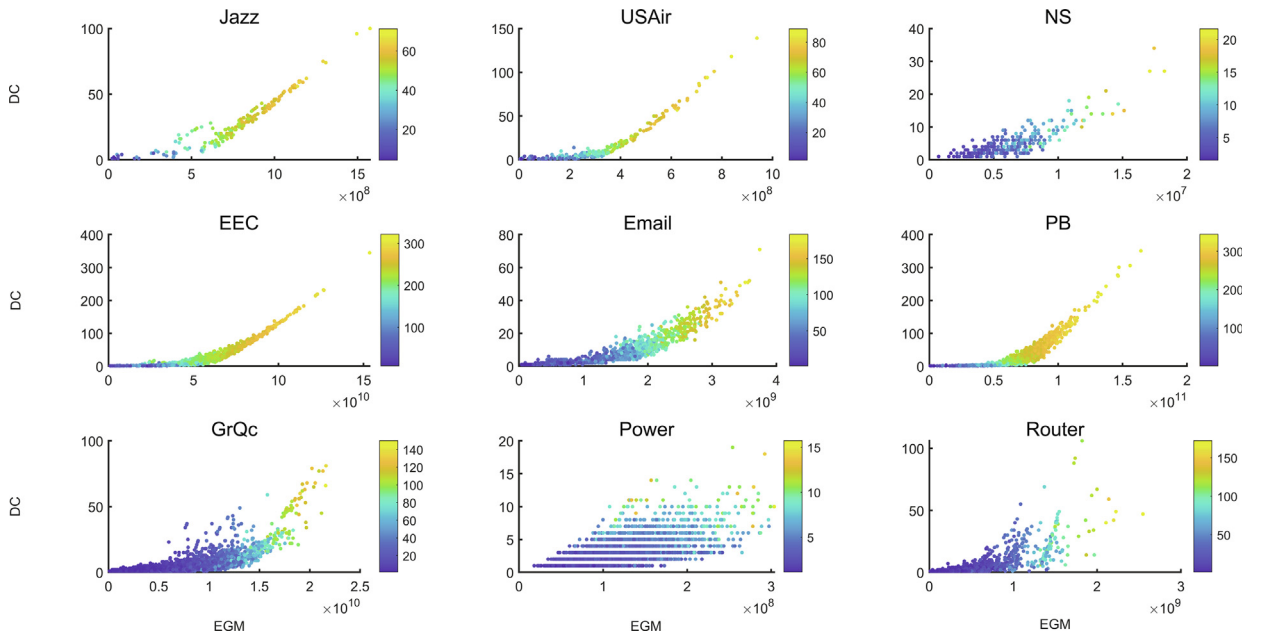


Fig. 1. The relationship between EGM and DC centrality measures on six small-scale networks which include Jazz, USAir, NS, EEC, Email, and PB, and three large-scale networks which include GrQc, Power, and Router. The X-axis denotes the centrality scores procured by the EGM, the Y-axis indicates the centrality scores procured by the DC measure, and each point corresponds to a node in the real-world network.

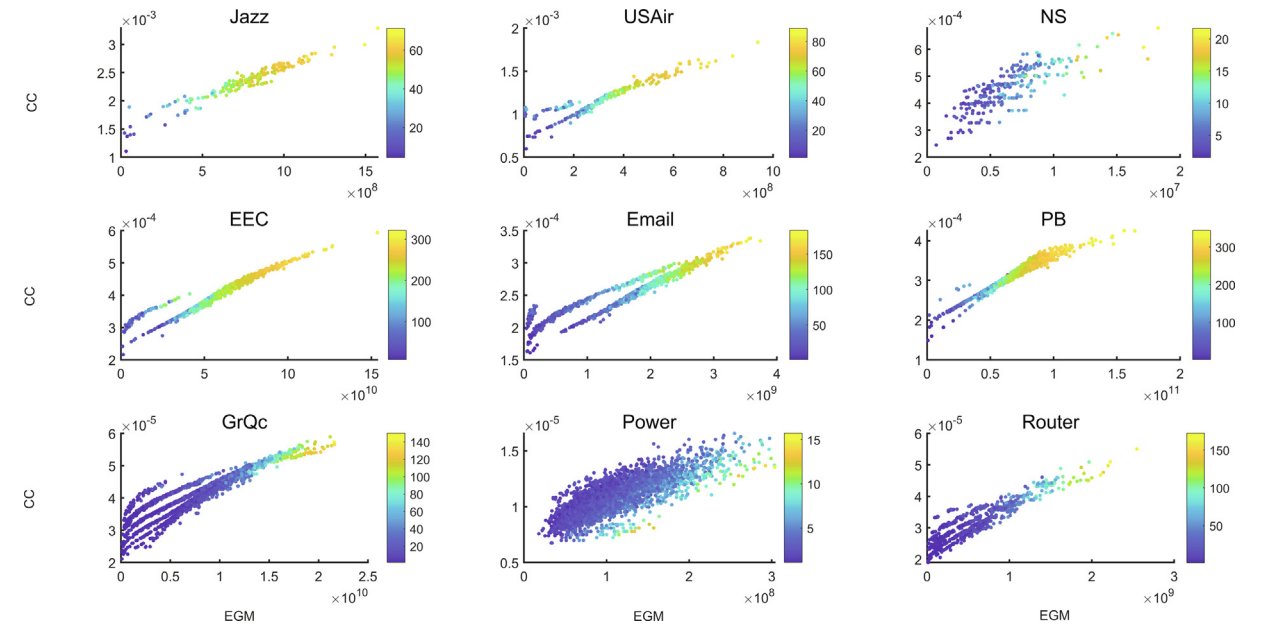


Fig. 2. The relationship between EGM and CC centrality measures on six small-scale networks which include Jazz, USAir, NS, EEC, Email, and PB, and three large-scale networks which include GrQc, Power, and Router.

ignored. Therefore, in order to amend approximate truncation radius, this paper defines the influence range based on the relationship between the node itself and its farthest node. What's more, for the sake of selecting the appropriate node mass, information entropy is used to evaluate the degree distribution in local network. Based on the above analysis, a approach called effective gravity model is presented, which is based on precise radius and value information to identify influencers in complex networks. EGM consists of the following five processes, whose flowchart is revealed in Fig. 9.

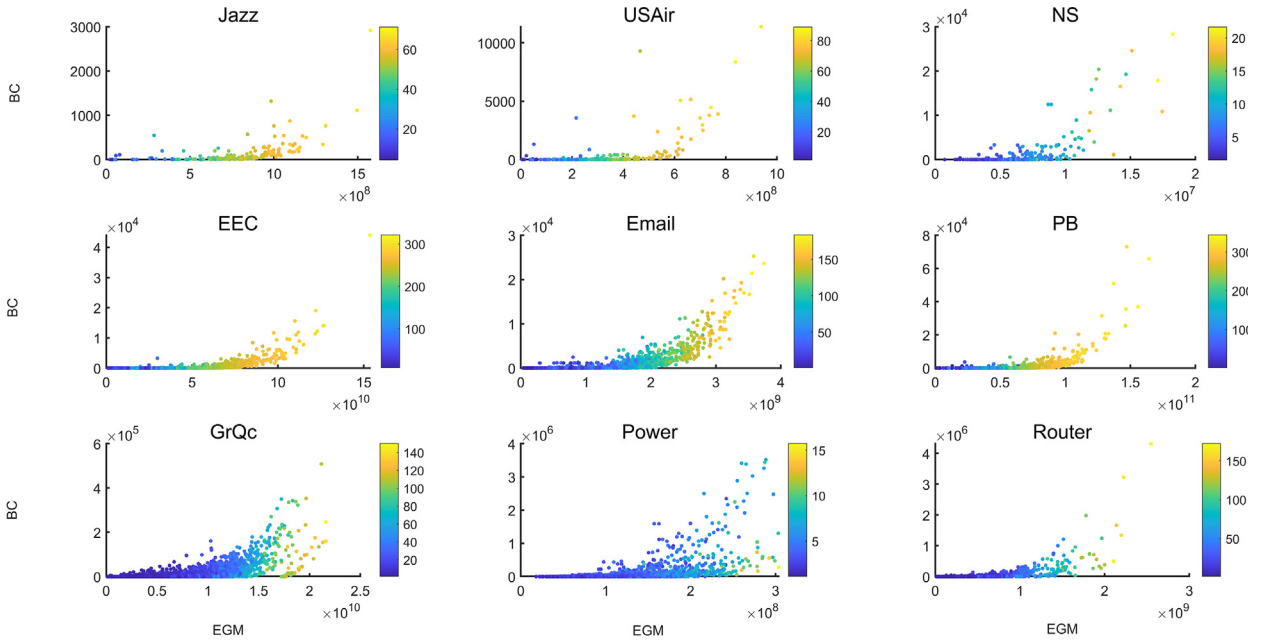


Fig. 3. The relationship between EGM and BC centrality measures on six small-scale networks which include Jazz, USAir, NS, EEC, Email, and PB, and three large-scale networks which include GrQc, Power, and Router.

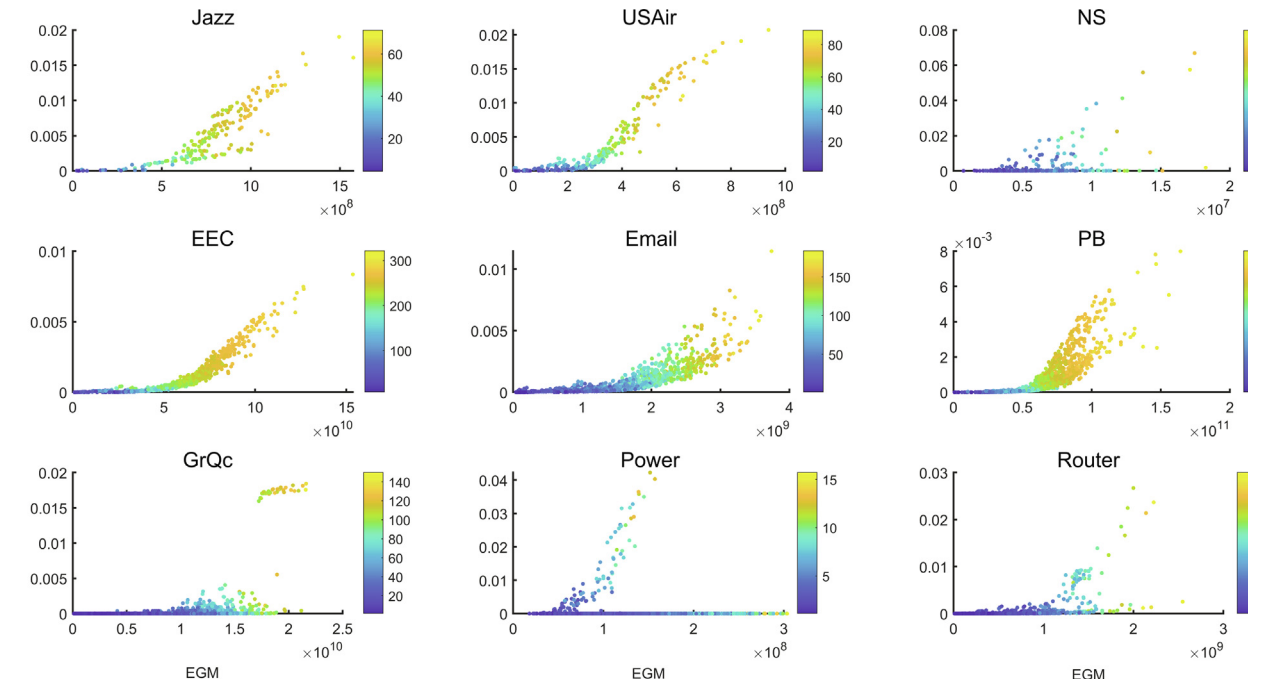


Fig. 4. The relationship between EGM and EC centrality measures on six small-scale networks which include Jazz, USAir, NS, EEC, Email, and PB, and three large-scale networks which include GrQc, Power, and Router.

(1) Build the network.
A given undirected network, which indicates a phenomenon in the real world, is used as input. The adjacency matrix of this network is used as output.

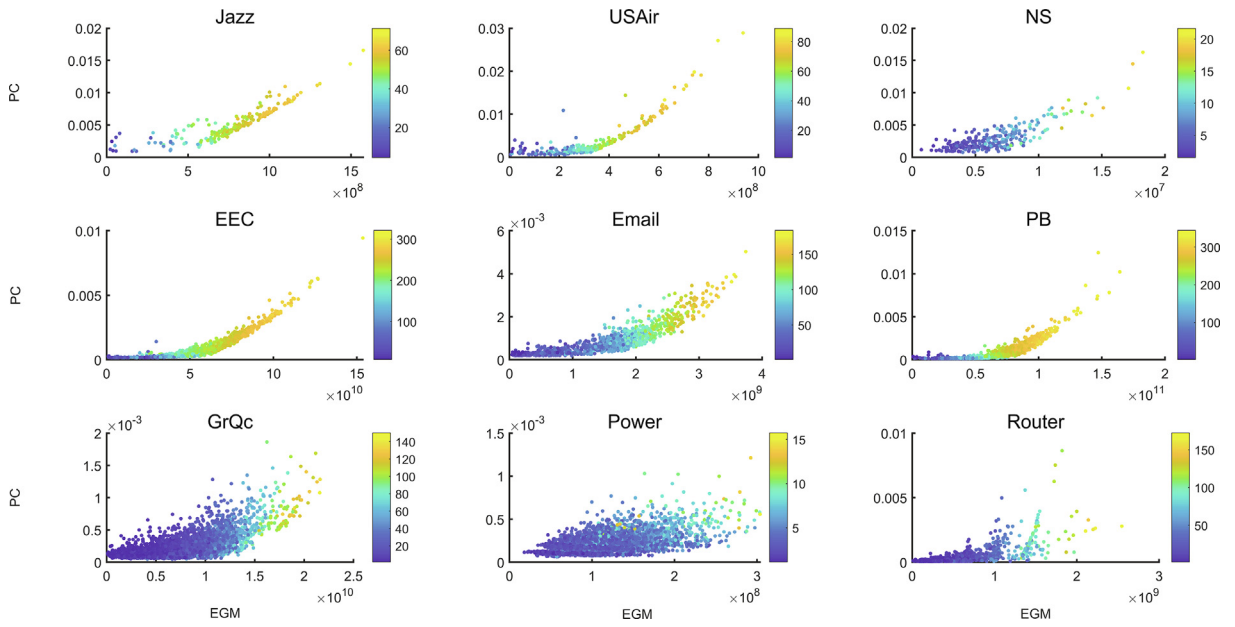


Fig. 5. The relationship between EGM and PC measures on six small-scale networks which include Jazz, USAir, NS, EEC, Email, and PB, and three large-scale networks which include GrQc, Power, and Router.

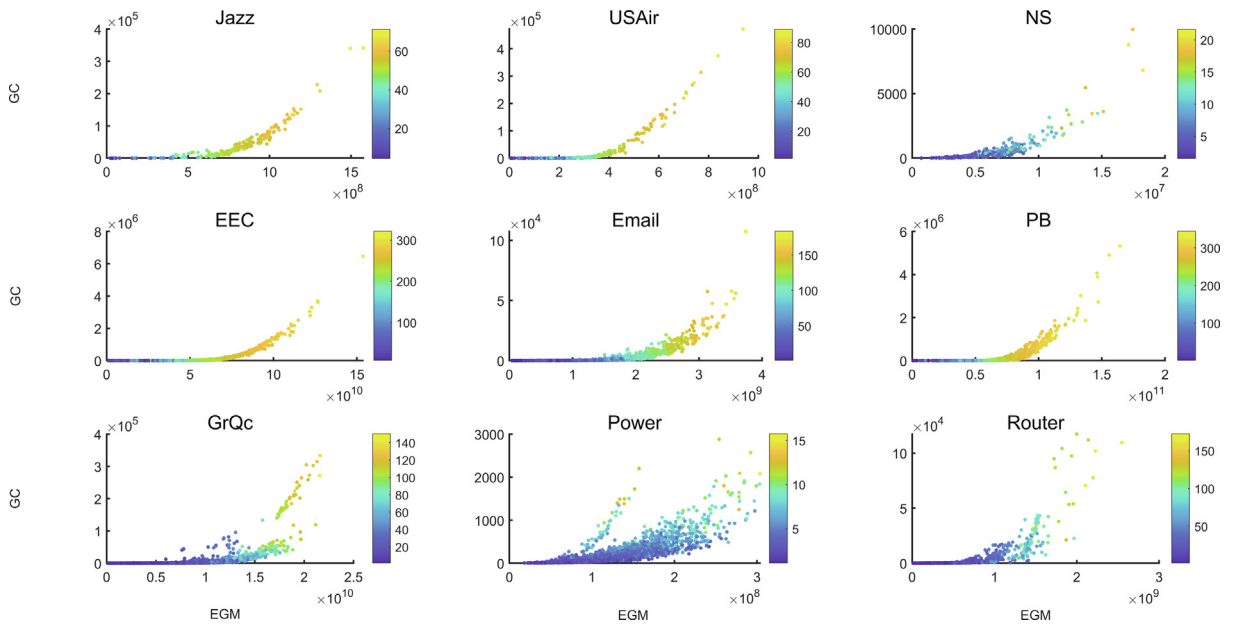


Fig. 6. The relationship between EGM and GC centrality measures on six small-scale networks which include Jazz, USAir, NS, EEC, Email, and PB, and three large-scale networks which include GrQc, Power, and Router.

(2) Obtain the degree distribution of nodes and distance matrix of the network.
The adjacency matrix obtained in the first step is used as input. And output the degree distribution matrix of entire network and the shortest path matrix which indicates the distance between the node pairs.
(3) Calculate the exact influence radius.
The degree distribution matrix and the shortest distance matrix are used as input. According to the way proposed in this paper, the exact influence radius of each node is output.

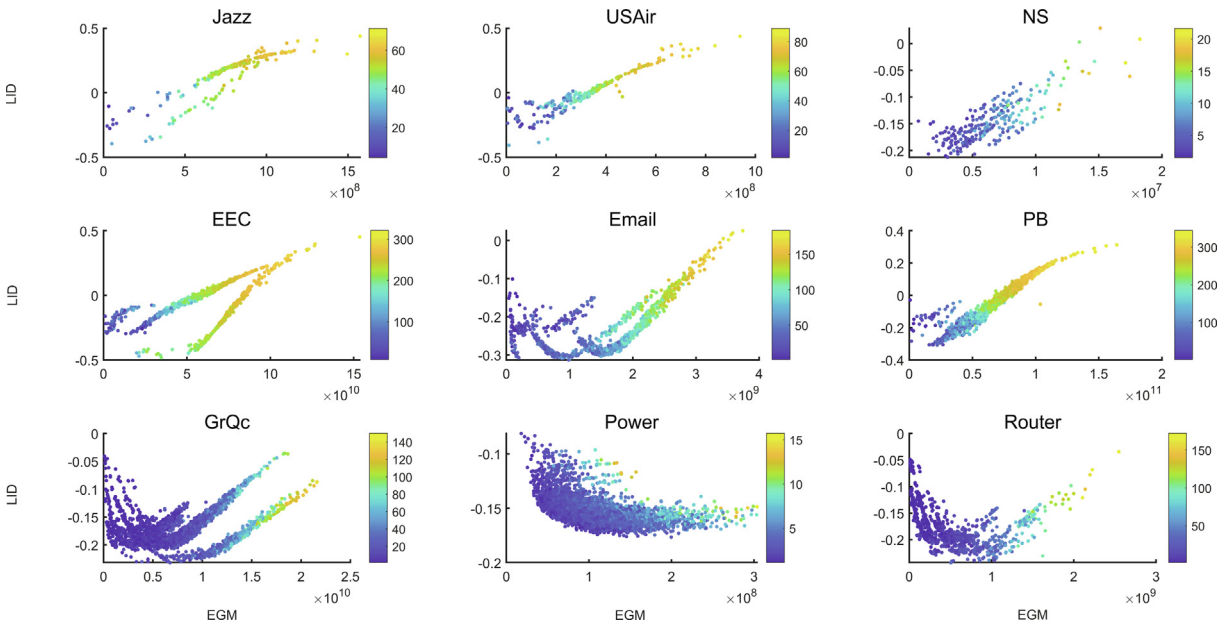


Fig. 7. The relationship between EGM and LID measures on six small-scale networks which include Jazz, USAir, NS, EEC, Email, and PB, and three large-scale networks which include GrQc, Power, and Router.

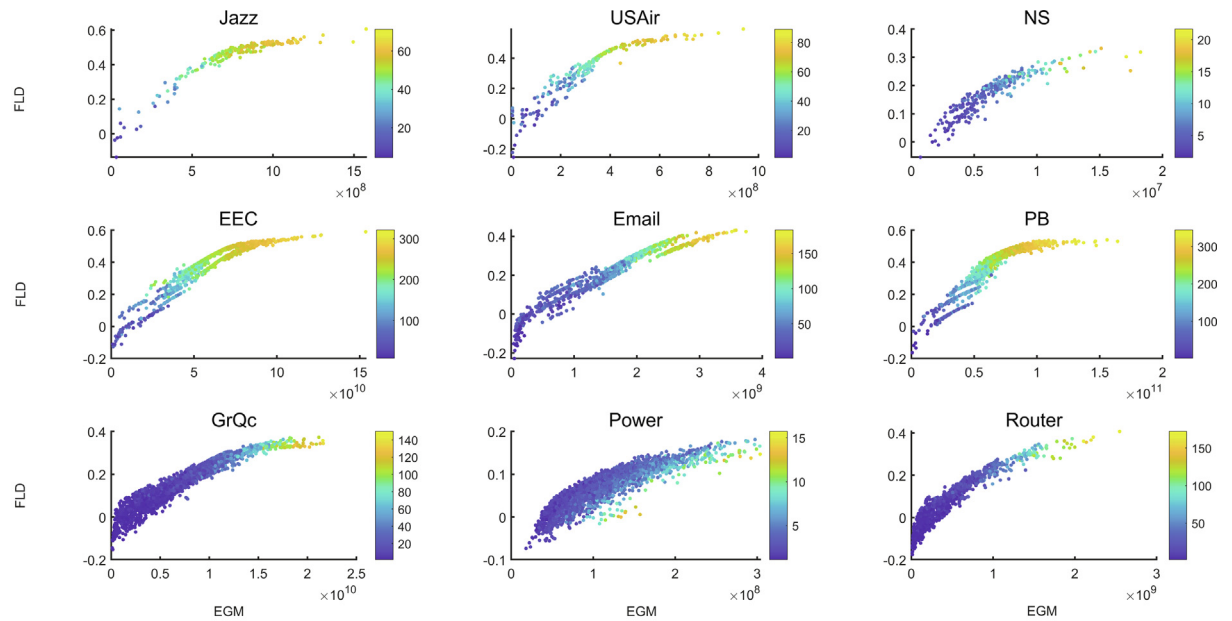


Fig. 8. The relationship between EGM and FLD measures on six small-scale networks which include Jazz, USAir, NS, EEC, Email, and PB, and three large-scale networks which include GrQc, Power, and Router.

(4) Calculate the value of the node.
Input the degree distribution matrix which is obtained in the second step. Output a value matrix which denotes the value of each node.
(5) Get the ranking of influence nodes.
Input the shortest distance which is obtained in the second step, the explicit influence radius and value matrix. Eventually, the influence ranking of all nodes is output in the network.

From Section 3.1 to Section 3.5, the content which covers each step of EGM is introduced in detail.

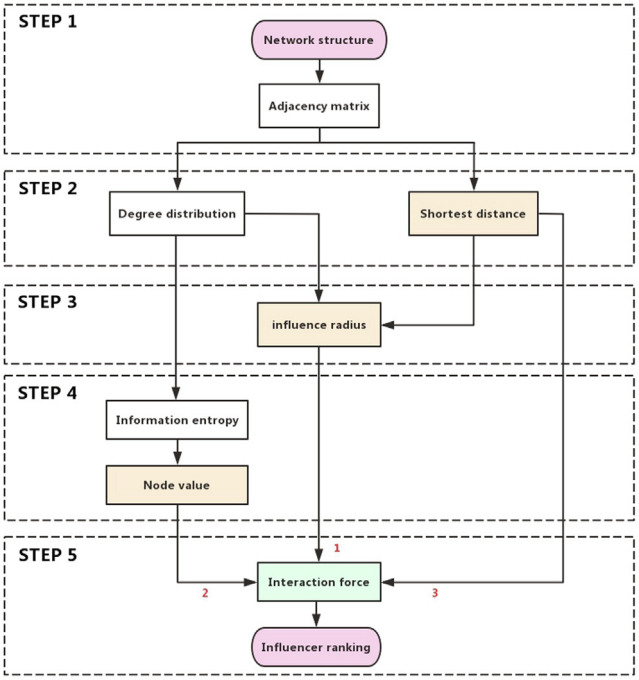


Fig. 9. The flow of the EGM algorithm. EGM has five steps, and the final result is composed of three factors.

3.1. Step 1: Build the network

A given undirected network which indicates the real-world system is represented by Graph $G(V,E)$. In graph G,V is denoted as node sets in the network, and E signifies edge sets. The graph G is stored as the adjacency matrix which is represented by A in this paper. The element a_{ij} in the matrix, which is located in the i-th row and the j-th column, indicates the connection relationship between the node i and the node j. $a_{ij} = 0$, which means that there is no edge between node i and node j. On the contrary, $a_{ij} = 1$ indicates that there is an edge.

3.2. Step 2: Obtain the degree distribution of nodes and distance matrix of the network

Based on the adjacency matrix of the network which is output in the first step, the degree distribution matrix is acquired. The degree of node i is defined as follows,

$$k_i = \sum_j a_{ij} \tag{12}$$

where k_i stands for the degree of node i. a_{ij} is got in the first step.

In addition, the shortest path between the source point and the end point is the path that contains the fewest number of edges. In this paper, if there is no path between two nodes, their shortest distance is assigned to infinity. And the distance from the node to itself is set to zero.

3.3. Step 3: Calculate the exact influence radius

In this part, we propose a new approach to calculate the influence range for every node. In fact, we believe that the influence radius of each node is different. Since we need to explore the maximum influence range of a node, the relationship between a node and its farthest point is considered. For node i, we assume that there exists a virtual node s on the path between node i and the node j furthest from i, where the influence from node i and j are equal. That's because when a certain node x is located between node i and node s, which is $x \in (i,s), x$ is only affected by node i. Similarly, when node x is located between node s and node j, which is $x \in (s,j), x$ is only affected by node j. Therefore, the node s is defined as the demarcation point between the influence of node i and node j, which can be obtained by the following steps.

Inspired by the formula of universal gravitation, the influence of a node on surrounding nodes is proportional to its own information and inversely proportional to the square of the distance. Therefore, given a network, the influence of node i on demarcation point s is defined as follows,

$$F_{is} = \frac{k_i \times k_s}{R_i^2} \tag{13}$$

where R_i is the radius of influence of node i, which means the distance between node i and demarcation point s. In the same way, the formula that is the influence of node j on demarcation point s is as follows,

$$F_{js} = \frac{k_j \times k_s}{(d_{ij} - R_i)^2} \tag{14}$$

where the distance between node i and node j is expressed as d_{ij}. For the demarcation point s, it receives equal influence from node i and node j, that is $F_{is} = F_{js}$. Then we can obtain the following equation,

$$\frac{k_i \times k_s}{R_i^2} = \frac{k_j \times k_s}{(d_{ij} - R_i)^2} \tag{15}$$

After solving the equation, the final influence radius formula of node i is procured, which is defined as follows,

$$R_i = \frac{d_{ij}}{1 + \sqrt{\frac{k_j}{k_i}}} \tag{16}$$

Because j is the furthest node from i, it can be rewrote as the following form,

$$R_i = \frac{d_{max(i)}}{1 + \sqrt{\frac{k_{max(i)}}{k_i}}} \tag{17}$$

In particular, if there are multiple nodes with the same furthest distance from i, these nodes constitute the furthest node domain $F_i = (k_1, k_2, \ldots, k_s)$. The average degree of nodes in the furthest node domain is expressed as $k_{mean(i)} = (k_1 + k_2 + \cdots + k_s) \times \frac{1}{s}$. $k_{max(i)}$ will be replaced by $k_{mean(i)}$.

3.4. Step 4: Calculate the value of the node

The value of the node, which evaluates the local information of the node, is presented in this part. In our opinion, the value of a node depends on the ability to influence the whole network. From a local view, a node with higher influence must be a node have higher degree, i.e. more neighbors. And from the perspective of influence propagation, if a node is more important, it has a greater probability of spreading outside in multiple potential directions. In other word, if a node has limited choices of spreading direction, the process is more likely to be obstructed. For a node with a certain local influence i.e. certain degree number in local network, the degree distribution of local network plays a significant role for node's importance. If the degree distribution of neighbors are more uniform, the information spreading uncertainty would increase. Based on the above analysis, information entropy can be used to measure the value of information from the perspective of information dissemination, which can well describe the uncertainty of social impact. Therefore, the value of a node, which takes into account both itself and neighborhood information, is defined in this paper as follows,

$$V_i = I_i \times k_i \tag{18}$$

where V_i is defined as the value of node i, I_i signifies the information entropy of node i, and k_i represents the degree of node i. The information entropy of each node contains two contents which are the characteristics of the node itself and the attributes of the neighbor nodes, which is more reasonable and comprehensive. The information entropy of node i is defined as follows,

$$I_i = -\sum_{j \in L_i} p_j log p_j \tag{19}$$

where p_j signifies the probability that the information is contained in the local area network of the node i, which is the ratio of the degree of the node j to the sum of the degrees in its local area network. The local area network of node i which is denoted by L_i, covers the node i, its neighbor's nodes and their degrees. An intuitive example of L_i is shown in Fig. 12. Consequently, p_j can be acquired in the following ways,

$$p_j = \frac{k_j}{\sum_{u \in L_i} k_u} \tag{20}$$

where k_j is expressed as the degree of node j. And $\sum_{u \in L_i} k_u$ indicates the sum of degrees which involves all nodes in the local area network of node i.

In summary, the value of node i can be written as the following form,

$$V_i = (-\sum_{j\in L_i} \frac{k_j}{\sum_{u\in L_i} k_u} log \frac{k_j}{\sum_{u\in L_i} k_u}) \times k_i \tag{21}$$

which consists of two parts of information belong to the node itself and neighbor nodes. The information of the node is reflected by the degree, and the neighbors' information is considered by the their degree distribution represented by information entropy.

3.5. Step 5: Get the ranking of influence nodes

Inspired by the law of universal gravitation, the EGM which is based on a modified precise radius and value information is defined as follows,

$$EGM(i) = \sum_{d_{ij}\leqslant R_i} \frac{V_i \times V_j}{d_{ij}^2} \tag{22}$$

where V_i and V_j stand for the value of node i and node j respectively. d_{ij} is represented as the shortest distance between i and j, and R_i is the influence radius of node i.

In conclusion, the vague influence radius, which was set to half of the average path length in the previous method, is replaced by the exact influence radius in this paper. At the same time, the value information of a node is regarded as mass, not just the degree of node. What's more, the local information and global information of the complex network are fully cogitated and combined in EGM. Specifically, for the consideration of local information, EGM measures the value of a node by making full use of the degree distribution information in its local network, and evaluates the spreading ability of the node through information entropy. For the combination of global information, EGM defines the range of the radius by considering the interaction between the node and its farthest node, and evaluates the importance by introducing the shortest distance between each node. In short, the modified exact influence range and the value information of a node that considers the surround degree distribution constitute the EGM proposed in this paper.

4. Experimental studies

In this section, eleven real-world networks are used to test the rationality of the proposed method. At the same time, the effectiveness of the proposed method is demonstrated by comparison with five classic measures. In addition, compared with three similar models and two state-of-the-art measures, the advancement of our approach is revealed. Six diverse experiments, which contain the top-ten nodes, node centrality score, Individuation, SI model, Kendall coefficient, and relationship analysis are used to fully prove the superiority of the proposed method for identifying crucial spreaders in complex networks.

4.1. The data sets

The data sets used in this paper are divided into six categories, including social networks, collaboration networks, communication networks, transportation networks, infrastructure networks and technology networks. In social networks, the Facebook network [37] consists of anonymized social circles in Facebook. Political blogs [38] is a network of direct hyperlinks between blogs about American politics. The voting data of the administrators in the Wikipedia community is described in the Wikipedia vote network [39]. In the collaborative network, the Jazz network [40] indicates the collaborative relationship between jazz musicians. Network science [41] shows the joint work process of scientists who study network science. Arxiv GR-QC [42] (General Relativity and Quantum Cosmology) cooperation network, which comes from electronic printing arXiv, covers scientific collaboration papers between authors which are published in the fields of general relativity and quantum

Table 1
Topological structure information of eleven real-world networks.

Type	Network	$\|N\|$	$\|E\|$	<d>	<k>	C
Small-scale	Jazz	198	2742	2.2350	27.6970	0.6175
	USAir	332	2126	2.7381	12.8072	0.6252
	NS	379	914	6.0419	4.8232	0.7412
	EEC	986	16064	2.5869	32.5842	0.4505
	Email	1133	10903	3.6060	9.6240	0.2202
	PB	1222	16714	2.7375	27.3552	0.3600
Large-scale	Facebook	4039	88234	3.6925	43.6910	0.6170
	GrQc	4158	13422	6.0494	6.4560	0.6648
	Power	4941	6594	18.9892	2.6691	0.1065
	Router	5022	6258	2.4922	6.4488	0.0329
	WV	7066	100736	3.2475	28.5129	0.2090

cosmology. In the communication network, the communication by emails is described in the Email network [43]. The EEC network [44] is a member of an E-mail exchange network which is related to European research institutions. The US Air lines network [45] in the transportation network involves air transportation in the United States. The Power network [46], in the infrastructure network, represents the topological structure of the state power grid in the western United States. The Router network [47] in the technical network is a symmetrical snapshot of the Internet structure.

Detailed information which contains eleven different types of networks is indicated in Table 1. $|N|$ and $|E|$ represent the set of nodes and edges respectively. $\langle d \rangle$ and $\langle k \rangle$ separately signify the average shortest distance and average degree of the network. The clustering coefficient, which denotes the degree of interconnection between adjacent nodes of a node, is represented by C. In addition, Jazz, USAir, NS, EEC, Email and PB are divided into small-scale networks. Facebook, GrQc, Power, Router and WV are divided into large-scale networks.

4.2. Similarity experiment: top-ten

The purpose of this experiment is to verify the similarity between the proposed approach EGM, classical ways which contain DC, CC, BC, EC, PC, similar approaches which include GC, WGC, GGC, and the most advanced measures including LID and FLD. The similarity between these approaches is measured by the number of the same top-ten nodes in their ranking lists. To be specific, diverse approaches evaluate the influencers in a complex network from different angles, and the top-ten nodes obtained are dissimilar. If there are more identical nodes among the top-ten nodes which are produced by the two approaches, it indicates that the two approaches have greater similarities. On the contrary, if the number of the same nodes is small, it means that the two ways do not have a strong correlation. The ranking list of the top-ten nodes generated by the real experiments carried out on eleven networks by the above approaches, which reflects the similarity between different measures, is revealed in Table 2. We use distinguishing colors to describe the same top-ten nodes which are acquired by different measures. For example, if there are identical nodes in the EGM ranking list and the ranking list which is generated by another approach, then these same nodes will be colored. On the contrary, the black indicates the dissimilar nodes which are acquired by EGM and another method.

As indicated in Table 2, the top-ten nodes acquired by DC, EC, GC, WGC and GGC are roughly the same as those obtained by EGM in Jazz. All methods have a higher matching degree with EGM, which also proves that the similarity between EGM and them is greater in USAir. In NS, the similitude between each measure is higher except for EC and WGC. For EEC, the top-ten nodes obtained by EGM are completely equivalent to those possessed by DC, which shows that the matching degree between them is relatively high. The results which are produced by CC and EGM are exactly the same in Email. The top-ten nodes of LID and EGM are equal in PB. In Facebook, EC and WGC have the lowest matching degree with EGM. The difference between LID and EGM is the largest in GrQc. For Power, the similarities between the various methods are not obvious. The results of CC, GC, GGC and FLD are approximately semblable to EGM in Router. It is worth noting that the results of DC are exactly the same as those of EGM in WV. In summary, the ranking of the top-ten nodes on each network for each approach is approximately the same as that of EGM except for Power, which indicates that our approach is similar to most existing measures and demonstrates its reliability.

4.3. Validity experiments

4.3.1. Centrality score

The purpose of this experiment is to calculate the standardized centrality score of nodes and the results are displayed by the way of heat map. Similar GC, WGC, GGC, the most advanced LID, FLD, and traditional BC, EC measures are used for comparison in this experiment. The relative importance distribution of each method is revealed in Figs. 10–13, where the color represents the significance of the node. If a node has a darker color, it means that the node has a high status.

As demonstrated in Figs. 10–13, the relative importance of node's distribution obtained by all measures are almost the same. The results on six small-scale networks are shown in Figs. 10 and 11. In Jazz, the centrality score acquired by FLD is too large. Both EGM and LID can distinguish nodes better. The performance of FLD and LID is better in USAir. For NS, EGM and LID have analogous results. And EGM, LID and FLD can distinguish nodes well. EGM, LID and FLD have strong distinguishing ability in EEC. The effectiveness of EGM is only inferior to FLD in Email. In PB, the score calculated by FLD is too large, and the performance of EGM and LID is better. The experimental results of the four large-scale networks are depicted in Fig. 13. For GrQc, the performance of EGM is only inferior to FLD. EGM, LID and FLD can distinguish the importance of nodes in Power and WV. In short, the accomplishment of EGM is better than similar methods like GC, WGC and GGC on all networks, which reveals that the improvement of our method for gravity model is effective for identifying influencers. At the same time, the distinguishing ability of EGM, LID and FLD surpasses other approaches.

4.3.2. Individuation

In order to compare the ability of diverse measures to distinguish nodes, the individuation experiment is used to count the frequencies of nodes with the same score which are obtained by each approach on ten networks. Individuation [48] is defined as the following formula,

Table 2
The top-ten nodes which are ranked by different measures in eleven real-world networks.

Jazz

	Classic					Similar			State-of-the-art		Proposed
Rank	DC	CC	BC	EC	PC	GC	WGC	GGC	LID	FLD	EGM
1	8	8	8	100	8	8	100	8	8	8	8
2	100	100	155	4	100	100	8	100	131	131	100
3	4	131	100	8	131	4	4	131	194	194	131
4	131	194	186	131	4	131	131	4	32	32	4
5	194	69	131	80	186	80	80	194	79	69	194
6	80	4	136	129	136	194	129	69	69	53	69
7	129	32	127	5	194	129	194	162	150	111	129
8	69	53	60	194	69	69	5	53	155	186	162
9	162	111	28	69	28	5	69	129	173	79	80
10	77	162	69	53	175	53	53	80	4	49	53

USAir

	Classic					Similar			State-of-the-art		Proposed
Rank	DC	CC	BC	EC	PC	GC	WGC	GGC	LID	FLD	EGM
1	118	118	118	118	118	118	118	118	118	118	118
2	261	261	8	261	261	261	261	261	67	261	261
3	255	67	261	255	182	255	255	255	261	67	255
4	152	255	201	182	152	182	182	182	201	255	182
5	182	201	47	152	255	152	152	152	47	201	152
6	230	182	182	230	230	230	230	166	255	182	166
7	166	47	255	112	166	166	166	230	166	166	230
8	67	166	152	67	201	67	67	67	248	47	67
9	112	248	313	166	67	112	112	201	182	248	112
10	201	112	13	147	8	147	147	112	112	112	201

NS

	Classic					Similar			State-of-the-art		Proposed
Rank	DC	CC	BC	EC	PC	GC	WGC	GGC	LID	FLD	EGM
1	4	26	26	4	26	4	4	4	51	51	26
2	5	95	51	5	4	5	5	26	26	95	4
3	26	51	169	16	5	26	16	5	52	231	5
4	16	231	95	15	95	16	15	16	67	26	51
5	67	100	67	45	67	15	45	51	95	52	95
6	70	52	5	46	16	51	1	95	5	5	231
7	95	5	231	47	32	95	46	231	169	169	16
8	15	44	100	176	51	231	47	67	16	100	52
9	32	234	44	177	8	67	176	52	23	170	169
10	51	297	66	250	70	70	177	169	231	76	67

EEC

	Classic					Similar			State-of-the-art		Proposed
Rank	DC	CC	BC	EC	PC	GC	WGC	GGC	LID	FLD	EGM
1	161	161	161	161	161	161	161	161	161	161	161
2	122	83	87	122	122	122	122	122	83	83	83
3	83	122	6	83	83	83	83	83	122	122	122
4	108	108	83	108	108	108	108	108	108	108	108
5	87	63	122	63	87	63	63	63	63	63	87
6	63	87	108	435	63	87	435	87	87	87	63
7	435	435	14	250	6	435	250	435	435	435	435
8	14	167	378	184	14	250	87	167	167	167	167
9	167	250	63	87	167	167	184	184	250	250	14
10	184	65	65	167	435	184	167	250	65	65	184

Email

	Classic					Similar			State-of-the-art		Proposed
Rank	DC	CC	BC	EC	PC	GC	WGC	GGC	LID	FLD	EGM
1	105	333	333	105	105	105	105	105	105	333	105
2	333	23	105	16	23	42	16	333	333	23	333
3	16	105	23	196	333	16	196	42	23	42	23
4	23	42	578	204	41	333	42	23	41	105	42
5	42	41	76	42	42	23	333	41	42	76	41
6	41	76	233	49	233	196	23	16	233	468	76
7	196	233	135	56	16	41	204	196	76	41	233
8	233	52	41	116	355	76	49	233	52	233	52
9	21	135	355	333	21	3	3	76	135	52	135
10	76	378	42	3	24	233	116	52	3	378	378

PB

	Classic					Similar			State-of-the-art		Proposed
Rank	DC	CC	BC	EC	PC	GC	WGC	GGC	LID	FLD	EGM
1	127	838	672	127	672	127	127	838	127	838	127
2	838	127	127	48	127	838	48	127	838	890	838
3	672	497	768	497	768	48	497	497	497	673	672
4	48	48	838	566	838	497	838	48	48	592	497
5	497	890	497	283	497	566	566	672	672	639	48
6	768	566	1178	147	48	672	283	566	768	692	768
7	1006	768	48	838	1006	1006	147	1006	566	497	566
8	566	922	782	84	922	922	498	922	890	566	1006
9	922	1178	922	384	1178	890	253	768	922	101	922
10	1178	672	566	498	566	830	384	890	1006	127	890

Facebook

	Classic					Similar			State-of-the-art		Proposed
Rank	DC	CC	BC	EC	PC	GC	WGC	GGC	LID	FLD	EGM
1	108	108	108	1913	3438	108	1913	108	108	108	108
2	1685	59	1685	2267	108	1913	2348	1913	1685	429	1685
3	1913	429	3438	2207	1685	1685	2544	1685	1913	564	1913
4	3438	564	1913	2234	1	2348	2267	3438	484	1466	1
5	1	1685	1086	2465	1913	2544	2234	2544	349	1719	3438
6	2544	172	1	2143	349	2267	1986	2348	415	1578	429
7	2348	349	699	2219	687	1986	2143	1889	1	607	564
8	1889	484	568	2079	3981	2234	2207	1801	429	367	484
9	1801	415	59	2124	415	1889	2411	1664	377	518	349
10	1664	377	429	1994	699	2143	2219	1353	476	527	1578

GrQc

	Classic					Similar			State-of-the-art		Proposed
Rank	DC	CC	BC	EC	PC	GC	WGC	GGC	LID	FLD	EGM
1	3348	2129	2129	3348	2174	3348	3348	3348	3261	2129	3348
2	3388	2205	1525	437	2129	1924	3388	3614	2139	3956	2754
3	1924	1525	2221	1924	2139	3388	1924	1924	3956	3261	1924
4	3614	2754	1232	3419	1525	3614	3614	3388	3534	1525	2129
5	1065	429	2139	1554	433	1065	1554	2129	2036	2036	3614
6	1554	3348	817	2277	3388	2754	1065	2754	1260	429	1065
7	3419	1947	2205	4012	1232	1554	3419	1065	1678	1947	437
8	2754	3956	433	1267	1012	437	2754	2139	461	1678	3388
9	437	1924	2174	2251	3614	3419	437	1554	433	2754	1554
10	3075	3614	2754	1995	3348	3075	3075	2978	2892	2139	1525

Power

	Classic					Similar			State-of-the-art		Proposed
Rank	DC	CC	BC	EC	PC	GC	WGC	GGC	LID	FLD	EGM
1	2554	1309	4165	4382	4459	2554	4346	4459	4823	2607	2555
2	4459	2595	2544	4346	3469	4459	4382	2555	4799	2529	1167
3	832	2606	1244	4337	832	4346	4333	2554	4788	4165	2609
4	3469	1132	4220	4333	2554	2576	4337	2576	4055	2544	2607
5	4346	2607	2529	4353	1225	2555	4353	1167	4	2613	2608
6	2383	1244	1268	4385	2383	2543	4396	1335	4932	2606	4459
7	2543	1477	1309	4403	2576	2618	4374	2609	4380	4121	4165
8	2576	2558	1245	4348	2440	2586	4403	1092	4351	4220	2544
9	2586	2529	427	4396	598	491	4385	2435	2304	1268	2529
10	3896	2533	2607	4374	3896	1167	4362	2618	2257	70	2618

Router

	Classic					Similar			State-of-the-art		Proposed
Rank	DC	CC	BC	EC	PC	GC	WGC	GGC	LID	FLD	EGM
1	3670	3668	3668	3369	3670	3369	3369	3668	3668	3668	3668
2	3639	3373	3373	3373	3639	3326	3338	3670	4	3373	3373
3	3338	3667	1480	3339	3338	3668	3339	3326	228	3667	3729
4	1453	3729	3326	3326	1453	3670	3326	3369	4875	3729	3326
5	3369	3666	3729	3342	2624	3373	3373	3338	4912	3666	3727
6	3339	623	3667	3325	3675	3339	3342	3339	243	3727	3369
7	3326	3326	4059	3352	3369	3338	3334	3639	1123	1480	3667
8	2624	3325	3597	3338	3672	3639	3354	3373	4683	3326	3734
9	3373	3727	1449	3334	3339	3729	3331	3729	3853	3325	3339
10	3675	3324	623	3327	3644	3727	3325	3727	3958	3687	3325

WV

	Classic					Similar			State-of-the-art		Proposed
Rank	DC	CC	BC	EC	PC	GC	WGC	GGC	LID	FLD	EGM
1	2355	2355	2355	2355	2355	2355	2355	2355	2355	2355	2355
2	703	703	9	703	9	1428	1428	703	4890	703	703
3	9	432	432	1428	703	703	703	1428	3644	432	1428
4	1428	1428	3644	1067	432	1067	1067	432	677	1067	432
5	432	1067	1428	2470	3644	432	432	1067	13	1428	1067
6	1067	1262	703	432	1428	9	2470	9	2203	66	9
7	2470	9	1067	3059	1067	2470	9	2470	4522	677	2470
8	1262	1055	13	9	2470	1262	3059	1262	2719	4890	1262
9	1055	2470	1262	1055	13	1055	1262	1055	3872	13	1055
10	4890	2282	2057	1262	1262	3059	1055	3059	294	294	4890

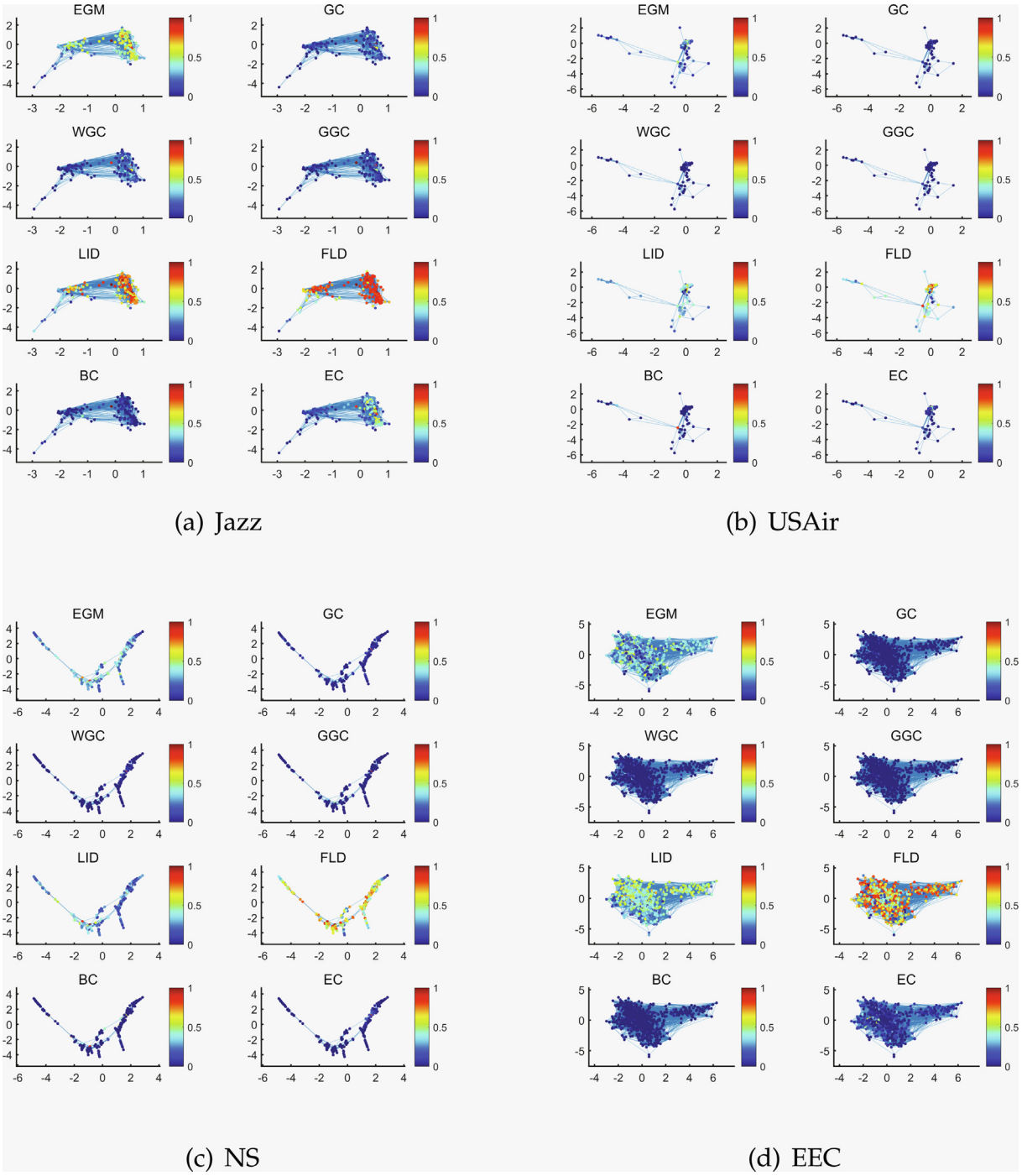


Fig. 10. (1) The centrality scores of nodes in six diverse small-scale networks which include Jazz, USAir, NS and EEC. The node with darker color indicates its greater influence.

$$Fre = \frac{|N_u|}{|N|} \tag{23}$$

where $|N_u|$ is expressed as the number of nodes with unique scores, and $|N|$ is the number of nodes in the entire network.

In this experiment, if an approach obtains more rankings, it works better. In addition, if a method has a higher individuation, it also proves that this method has a stronger distinguishing ability. The node frequencies obtained by diverse measures are indicated in Figs. 14 and 15 through the individuation experiment on ten networks. At the same time, the

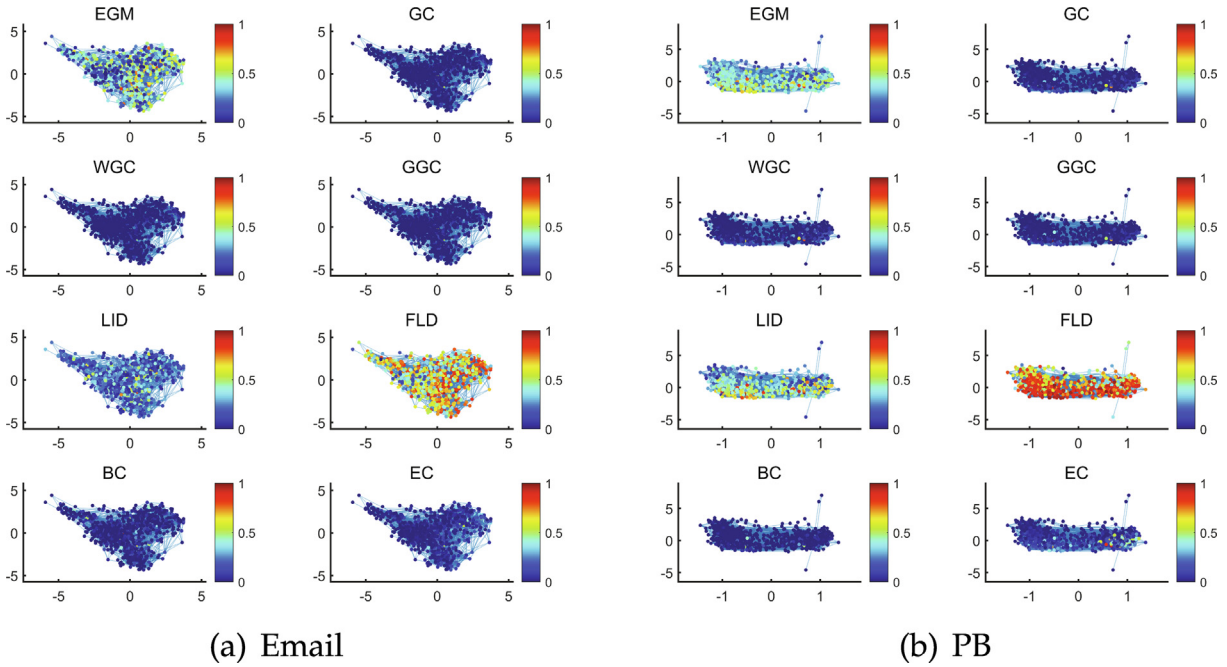


Fig. 11. (2) The centrality scores of nodes in six diverse small-scale networks which contain Email and PB.

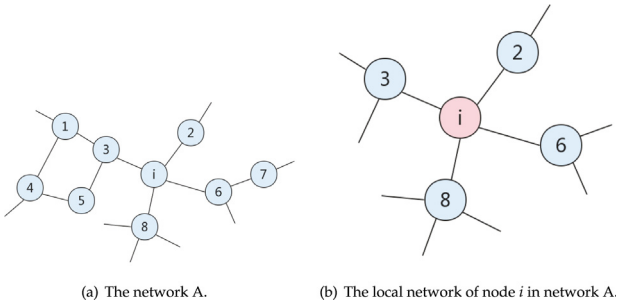


Fig. 12. A network and a local network of a node *i* in this network.

individuation of different approaches is recorded in Table 3 in detail, where the top-three individuation values are represented by red, yellow, and blue respectively.

We can see from Fig. 14 that WGC, GGC and EGM have fewer nodes in the same ranking for Jazz. For USAir, although EGM's performance is not the best, it beats DC, CC and BC. In NS and Email, the performance of EGM is only inferior to WGC. For EEC, EGM's distinguishing ability ranks third. In PB, the frequency of nodes in each ranking obtained by EGM is the lowest, and the ranking interval is the largest. As indicated in Fig. 15, EGM defeated all approaches except WGC in GrQc and Power. The lowest frequency in each ranking and the largest span are obtained by EGM in two large-scale networks which include Router and WV. What's more, it can be observed from Table 3 that the EGM's distinguishing ability ranks the top-three in all networks. Especially the lowest frequency and highest ranking are obtained by EGM in PB, Router and WV, which can distinguish the node's infection ability well. In general, through the analysis of Figs. 14 and 15, we can draw the conclusion that the performance of EGM is only inferior to WGC, but EGM is relatively better for large-scale networks.

4.3.3. SI model

For the sake of verifying the superiority of the proposed method in terms of transmission ability, the experiment on mathematical models of epidemic diseases is conducted on ten real-world networks. The SI model [49], one of the infectious disease models, is used in this experiment because the infection capacity of the selected node is proportional to the influence. The nodes in the complex network are divided into Class S and Class I in the SI model. Class S refers to susceptible nodes, which means that nodes of this type are not infected. Nevertheless, if the S-class node is in contact with the infected node, it is susceptible to infection. Class I refers to infectious nodes, which means that nodes of this type have been infected with

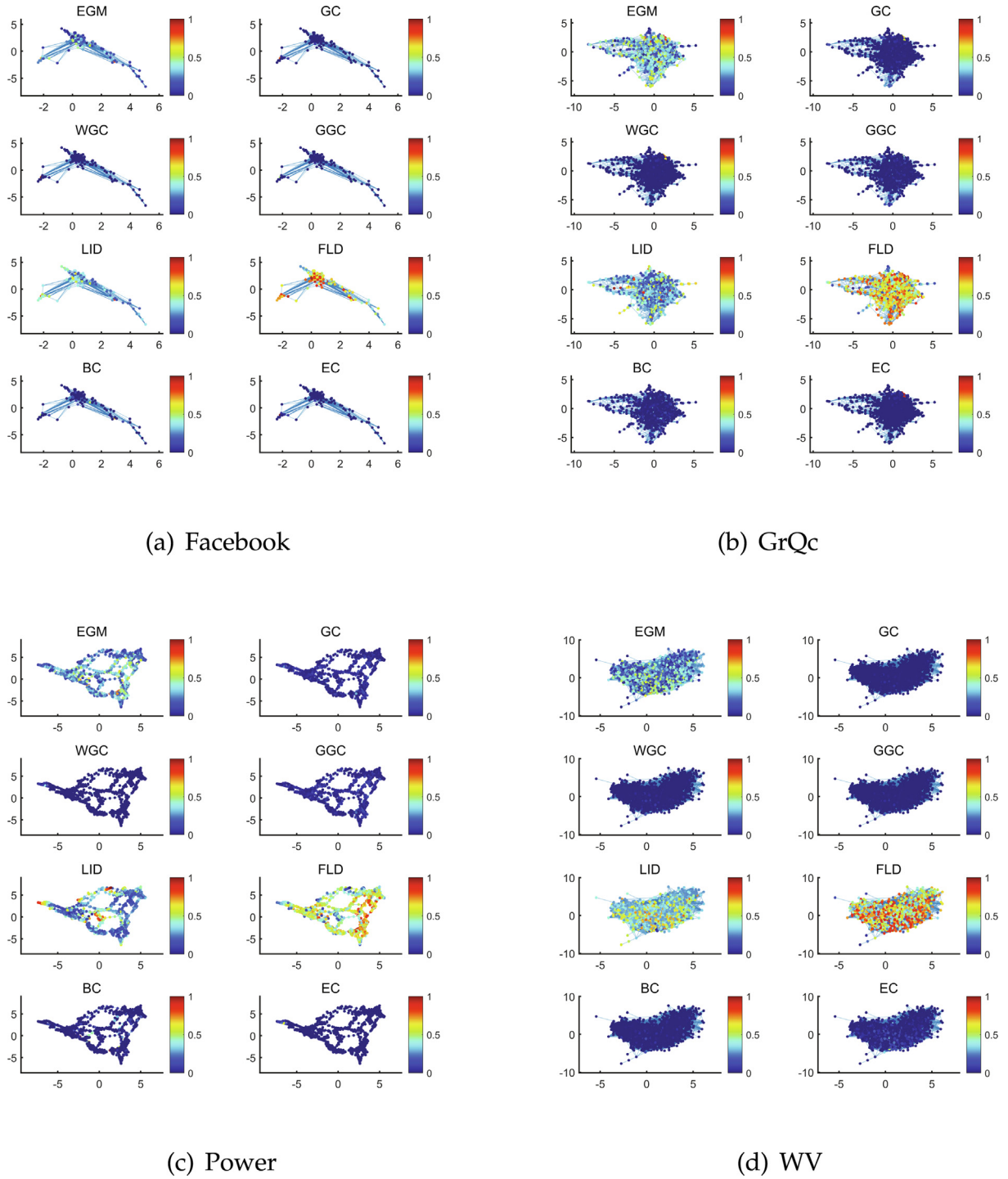
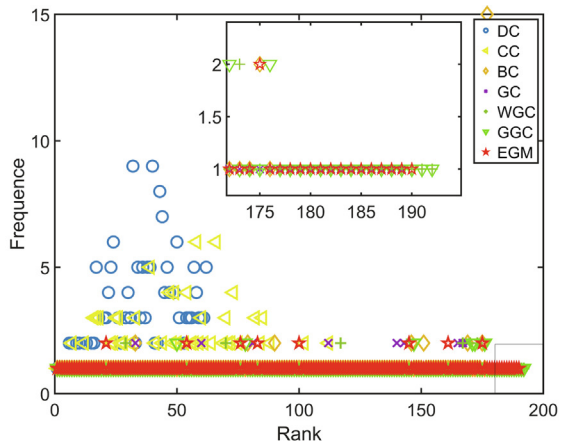


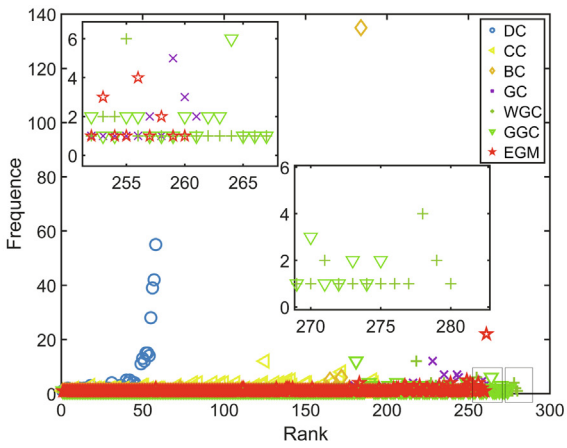
Fig. 13. The centrality scores of nodes in four diverse large-scale networks which contain Facebook, GrQc, Power and WV.

an infectious disease. And the I-class nodes have the infectious ability to change the type S into the type I, which infects their neighbor nodes with a certain rates of infection β. At a given time t, the differential equation of the SI model is as follows,

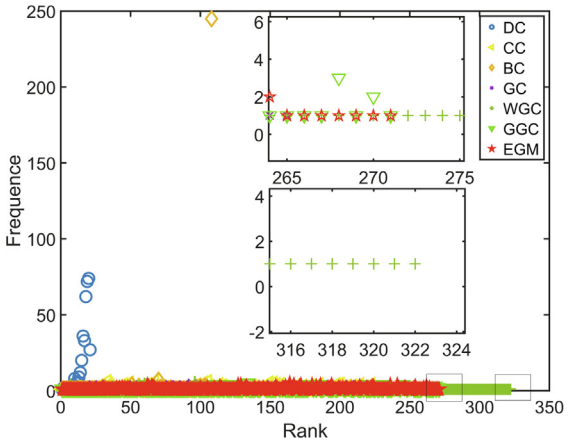
$$\frac{dS}{dt} = -\beta SI, \ \frac{dI}{dt} = \beta SI \tag{24}$$



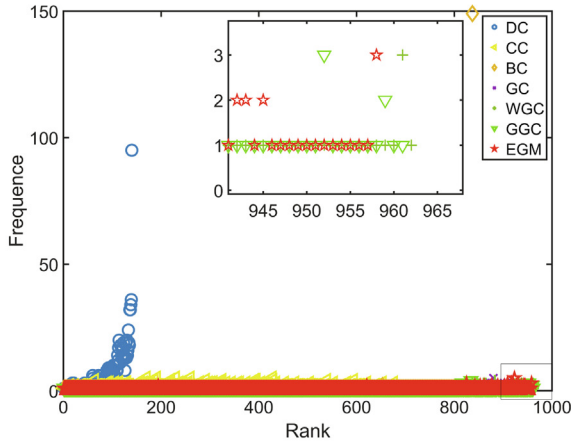
(a) Jazz



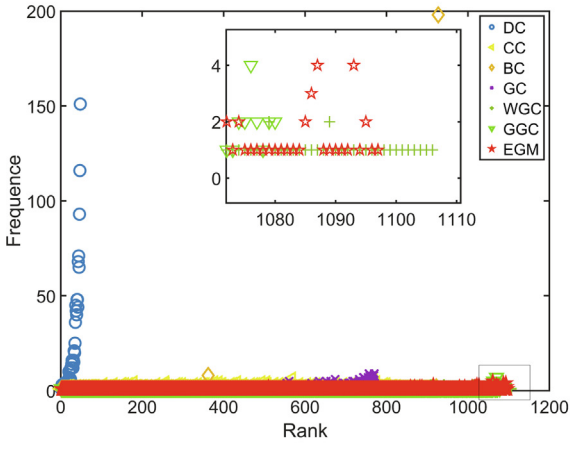
(b) USAir



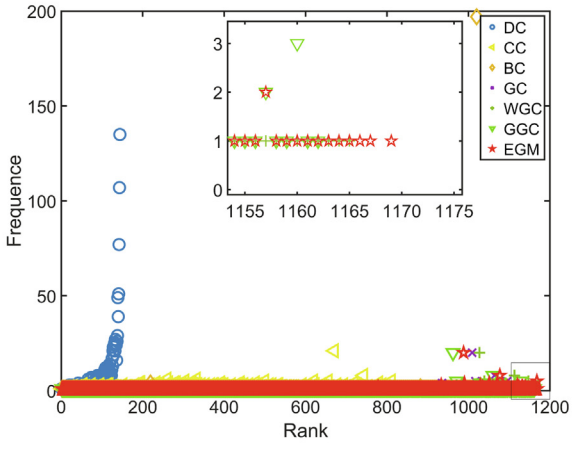
(c) NS



(d) EEC



(e) Email



(f) PB

Fig. 14. The frequencies of nodes with the same score in six diverse small-scale networks which include Jazz, USAir, NS, EEC, Email and PB. The X-axis indicates the ranking of nodes, the Y-axis signifies the frequence of nodes with the same rank.

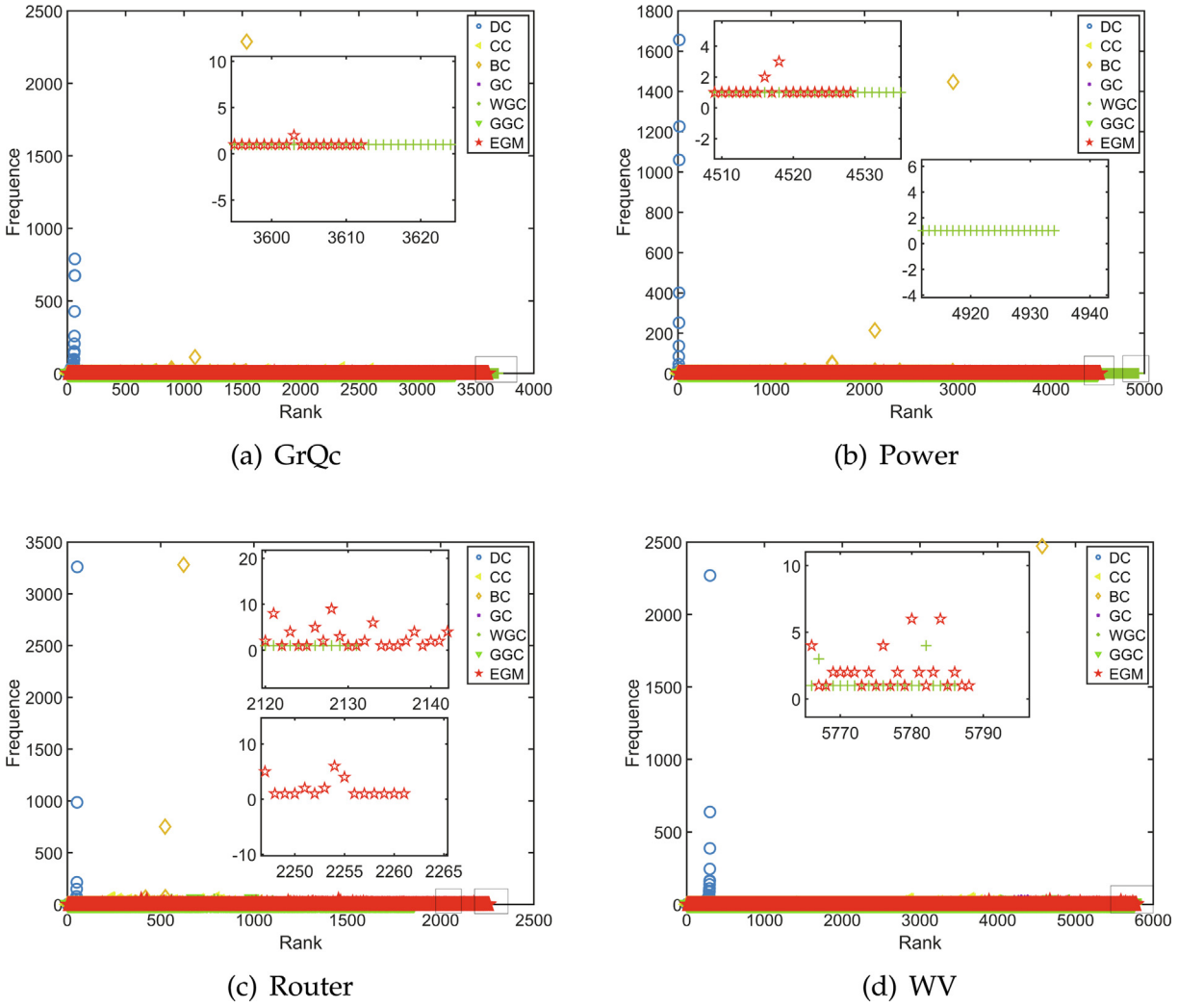


Fig. 15. The frequencies of nodes with the same score in four diverse large-scale networks which include GrQc, Power, Router and WV.

During the spread of the disease, the total number of nodes in the complex network $S(t) + I(t) = N$ remains unchanged. From this equation, it can be concluded that if the number of infected nodes increases faster, the source of infection is more important.

In this experiment, the top-ten nodes of each method in Section 4.2 are selected as the initially infectious nodes, and the remaining nodes are regarded as susceptible nodes. These infectious nodes will infect surrounding susceptible nodes with a certain probability β which is set to 0.1. The propagation time is represented by t. The number of infected nodes $F(t)$ is used as an index to determine the infection abilities of the initially selected nodes at a certain time t. Each experiment is independently performed one hundred times to get objective results, and the average results obtained are depicted in Figs. 16 and 17. In all measures, if the top-ten nodes which are initially selected by a method have greater influence, the infection power of these nodes will be stronger, and the final experiment will result in a higher number of infected nodes.

Fig. 16 describes the experimental results of the SI model in six small-scale networks. In Jazz, EGM achieves the optimal effect when $t \geqslant 40$. For USAir, the number of infected nodes procured by EGM exceeds other approaches at every moment., which demonstrates the superior capability of EGM. In NS, EGM has the strongest transmission capable when $t \geqslant 20$, which to a large extent demonstrates the advantage of EGM in almost the entire time period. For EEC, the maximum number of infected nodes is obtained by EGM when $25 \leqslant t \leqslant 40$ and $42 \leqslant t \leqslant 46$. In Email, EGM shows superior capability that beats other methods when $t \geqslant 2$. When $3 \leqslant t \leqslant 25$ and $t \geqslant 28$, EGM has the strongest transmission ability in PB. The propagation capability test in four large-scale networks is depicted in Fig. 17. Although the most infected nodes are not owned by EGM in Facebook, it is second only to CC. For GrQc, the transmission ability of EGM is better than other approaches when $t \geqslant 11$. In Power, the number of infected nodes possessed by EGM is significantly higher than other approaches when $t \geqslant 18$, which shows the strongest advantage of EGM. For WV, the best achievement is revealed by EGM when $t \geqslant 33$. It can be concluded

Table 3
Individuation experiment results of each measure on ten different networks.

Type	Network	DC	BC	CC	GC	WGC	GGC	EGM
Small-scale	Jazz	0.31313	0.89394	0.64141	0.95455	**0.96970**	**0.96970**	0.95960
	USAir	0.17470	0.55422	0.58133	0.78614	**0.84337**	0.82831	0.78614
	NS	0.05541	0.28496	0.60158	0.69921	**0.84960**	0.71504	0.71504
	EEC	0.14199	0.84888	0.68560	0.94828	**0.97566**	0.97465	0.97160
	Email	0.04237	0.81818	0.74051	0.68049	**0.97617**	0.95322	0.96823
	PB	0.11784	0.83470	0.67430	0.90671	0.95336	0.95090	**0.95663**
Large-scale	GrQc	0.01563	0.36989	0.66955	0.78523	**0.88793**	0.79966	0.86869
	Power	0.00324	0.59705	0.84639	0.91135	**0.99858**	0.91115	0.91641
	Router	0.01055	0.12405	0.29669	0.35743	0.42433	0.37017	**0.45022**
	WV	0.04246	0.64704	0.62780	0.62270	0.81885	0.81475	**0.81913**

that the strongest spread ability in experimental networks except Facebook is possessed by EGM within a certain period of time. What's more, the performance of EGM far exceeds analogous methods like GC and WGC, which indicates that EGM is the best measure based on gravity model so far.

4.3.4. Kendall coefficient

In order to test the reliability of the proposed method, the correlation between various centrality methods and the SI model which is the standard measure of infection capacity is evaluated by in the Kendall coefficient [50]. Kendall coefficient τ, a statistical value used to measure the correlation between two random sequences, is defined as the ratio of the subtraction between concordant pairs and discordant pairs to the total number of pairs. Kendall coefficient τ is expressed in the following form,

$$\tau = \frac{N_C - N_D}{\frac{1}{2}N(N-1)} \tag{25}$$

where N_C is used to denote the number of consistent sequence pairs, the number of discrepant sequence pairs is represented as N_D, and the length of sequence is denoted by N.

Specifically, suppose that the i-th value of two random sequences A and B is signified by a_i and b_i respectively, and any two corresponding values form two sequence pairs (a_i, b_i) and (a_j, b_j). If the conditions which is $a_i > a_j$ and $b_i > b_j$ or $a_i < a_j$ and $b_i < b_j$ are met, (a_i, b_i) and (a_j, b_j) are deemed to be consistent. On the contrary, if the conditions which is $a_i > a_j$ and $b_i < b_j$ or $a_i < a_j$ and $b_i > b_j$ are met, these two sequence pairs are considered to be inconsistent. In particular, if $a_i = a_j$ or $b_i = b_j$, this is deemed to be neither consistent nor discrepant. In addition, $\tau = 1$, which means their arrangement is the same. And $\tau = -1$, which shows that their arrangement is disparate.

In this experiment, the number of nodes infected with ten steps $(F(10))$ in the SI model is used to denote the infection ability of each node. For the sake of making the results more comprehensive, the propagation probability β of the SI model is changed to test dissimilar situations. Each experiment is independently performed one hundred times to get objective results. The average situations of the experimental results are shown in Figs. 18 and 19. If a method finally possesses a higher τ value, it indicates that this approach is more similar to the standard measure of the SI model, which also proves that this approach has better performance in terms of accuracy.

The experimental results on six small-scale networks are shown in Fig. 18. The τ value of EGM is the highest in Jazz when the infection rate $\beta = 0.18, 0.19$ and 0.20. In USAir, although EGM ranks fifth when the infection rate $\beta \leqslant 0.37$, it has a great amelioration for analogous algorithm like WGC. For NS, the achievement of EGM is best when $0.11 \leqslant \beta \leqslant 0.13$. The τ value of EGM is only lower than that of WGC and EC in EEC. In Email, the τ value second only to CC is owned by EGM when the infection rate $\beta = 0.45, 0.46, 0.47$ and 0.51. For PB, the τ value of EGM exceeds all approaches, which shows that EGM has the best performance. In Fig. 19 which describes the experimental results on four large-scale networks, the capability of EGM is much better than other measures in Facebook when the infection rate $0.03 \leqslant \beta \leqslant 0.08$. In GrQc, EGM has the best performance when $\beta = 0.11$. EGM defeats other approaches besides GC in Power. For Router, EGM ranks the fifth when $0.28 \leqslant \beta \leqslant 0.37$. In summary, in Jazz, NS, PB, Facebook and GrQc, the highest τ value is acquired by our approach at certain infection rates. In particular, larger amelioration to the gravity model is achieved by EGM in the networks like NS, PB, Facebook and GrQc.

4.4. Correlation experiment

The purpose of this experiment is to demonstrate the rationality of the proposed method by assessing the relationship between EGM and other methods. In addition, the SI model is introduced to evaluate the significance between nodes, where the propagation probability β is set to 0.1 and the time t is set to 10. Each experiment is tested on nine different networks to