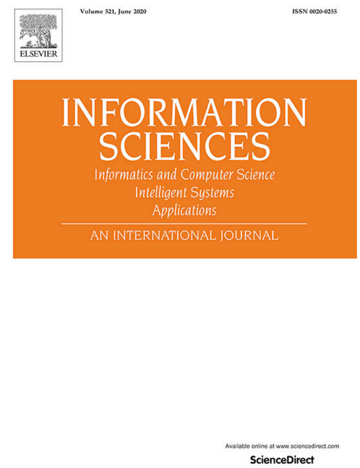
Journal Pre-proofs



Identifying influential nodes in complex networks: Effective distance gravity model

Qiuyan Shang, Yong Deng, Kang Hao Cheong

PII: S0020-0255(21)00091-8
DOI: https://doi.org/10.1016/j.ins.2021.01.053
Reference: INS 16182

To appear in: *Information Sciences*

Received Date: 11 October 2020
Accepted Date: 17 January 2021

Please cite this article as: Q. Shang, Y. Deng, K.H. Cheong, Identifying influential nodes in complex networks: Effective distance gravity model, *Information Sciences* (2021), doi: https://doi.org/10.1016/j.ins.2021.01.053

Identifying influential nodes in complex networks: Effective distance gravity model

Qiuyan Shang[a,b], Yong Deng[a,c,*], Kang Hao Cheong[d,e,*]

[a] Institute of Fundamental and Frontier Science, University of Electronic Science and Technology of China, Chengdu, 610054, China
[b] Yingcai Honors of school, University of Electronic Science and Technology of China, Chengdu, 610054, China
[c] School of Education, Shannxi Normal University, Xi'an, 710062, China
[d] Science, Mathematics and Technology Cluster, Singapore University of Technology and Design (SUTD), S487372, Singapore
[e] SUTD-Massachusetts Institute of Technology International Design Centre, Singapore

Abstract

The identification of important nodes in complex networks is an area of exciting growth due to its applications across various disciplines like disease control, data mining and network system control. Many measures have been proposed to date, but they are either based on the locality of nodes or the global nature of the network. These measures typically use the traditional Euclidean Distance, which only focuses on local static geographic distance between nodes but ignores the dynamic interaction between nodes in real-world networks. Both the static and dynamic information should be considered for the purpose of identifying influential nodes. In order to address this problem, we have proposed an original and novel gravity model with effective distance for identifying influential nodes based on information fusion and multi-level processing. Our method is able to comprehensively consider the global and local information of complex networks, and also utilizes the effective distance to incorporate static and dynamic information. Moreover, the proposed method can help us mine for hidden topological structure of real-world networks for more accurate results. The susceptible infected model, Kendall correlation coefficient and eight existing identification methods are utilized to carry out simulations on twelve different real networks.

Keywords: Complex networks, Influential nodes, Gravity model, Effective distance, Susceptible infected model

1. Introduction

In recent years, the study of complex networks has attracted immense attention [13, 14]. Many real-world problems [47, 44, 7] can be analyzed as part of network science [11, 37, 25], for example, Internet security, network control system [50] and social network [39]. Hence, the identification of influential nodes

*Corresponding author at: Institute of Fundamental and Frontier Science, University of Electronic Science and Technology of China, Chengdu, 610054, China. E-mail: dengentropy@uestc.edu.cn, prof.deng@hotmail.com. (Yong Deng)
Science, Mathematics and Technology Cluster, Singapore University of Technology and Design, S487372, Singapore. mail: kanghao_cheong@sutd.edu.sg. (Kang Hao Cheong)

in complex networks plays an important role [40] in both structural and functional aspects [24], which becomes an important area of research [5]. It can be applied across various fields to solve real-world problems, such as disease and rumor controlling [23, 20], network system [9], biology [28, 40], social system [12, 18, 17], influence maximizing [46, 32], time series prediction [33], community finding [49, 16] and information propagation [45, 19].

There are many existing methods to assess the influence of nodes [43, 35, 45]. One type of the methods focuses on the local information of nodes but ignores the global information such as degree centrality (DC) [5], K-shell decomposition (KS) [23], enginevector centrality (EC) [2] and PageRank (PC) [31, 29]. Another type suggests that the influence of nodes is mainly dependent on the connectivity in the network. Closeness centrality (CC) [26] and betweenness centrality (BC) [30] are some examples. Although these identification methods can often give reasonable results, they have many limitations as well. For instance, DC only considers the neighbor's influence but ignores global influence. PC has good performance on directed network but does not perform very well on undirected network. CC and BC are very sensitive to network structure and the complexity of them is high [24]. Furthermore, these methods only consider either the local information or global information of nodes, and not realistic enough.

Recently, a variety of novel identification approaches has been proposed to address the current gap in that some methods only consider local or global information [36, 21, 1]. For instance, Chen et al. [4] utilized the path diversity for identification. Ren et al. [34]proposed a model based on iterative resource allocation. Deng et al. detected vital nodes by inverse-square law [10]. Inspired by the law of gravity, an algorithm based on the gravity model, called gravity model (GM) was proposed by Li et al. [24]. From the perspective of fractal property, the fuzzy local dimension was proposed by Jiang et al. to detect influential nodes [41]. Further, Wen et al. identified vital nodes by local information dimension [42]. Due to the efficiency of data fusion of different sources by evidence theory, some evidential methods to rank influential nodes have been studied. There are also other methods for influential nodes identification, such as random-walking-based method [22], TOPSIS-based method [48], entropy model [15] and Quasi-Laplacian centrality [27].

Most existing approaches only focus on the static information of nodes but ignore the dynamic information between nodes in real-world networks, thereby leading to unsatisfactory results. In the real-world, most complex networks have hidden dynamic topological structure that contains dynamic information between nodes. For instance, the power transmission between nodes in power network, the information flow between users in social network, the exchange of chemical substances in protein network, the spread of virus in a disease transmission network. Behind the observable geometric structure, there are also hidden dynamic structures in the network to dominate the dynamic process, especially in real-world applications. This dynamic topological structure contains information to help us identify vital nodes in the real-world better, like the dynamic interaction information between nodes. Such dynamic information has been assessed to be as significant as explicit static information which cannot be ignored. Hence, how to effectively mine the hidden topology of the network and the dynamic interaction information between nodes has become a topic of immense interest. The effective distance proposed by Brockmann et al. [3] can aid us in solving this problem. Effective distance is a distance abstracted from the probability. It mainly pays attention to the interaction of nodes in the network, which is the main basis for judging. The crux of effective distance is to discover the most probable path between two nodes by calculating the probability through the adjacency matrix, which fully considers the dynamic information interaction between nodes in the real-world network.

The main contribution of this work is in our proposed effective distance gravity model. First, the effective distance gravity model comprehensively considers the local information of the node and global information of the network, thereby has advantages over the traditional centrality methods that only focus on either local or global information. Second, our proposed method can mine the hidden dynamic structure of the network and the dynamic interaction information between nodes, which dominates the actual operation of the network. Third, the combination of dynamic information and explicit static information can better detect the important nodes in complex networks. Fourth, the cumulative centrality score calculation method in our model reduces the identification error caused by the unstable structure.

Based on our proposed method, we have carried out a variety of experiments on twelve real-world networks by using the susceptible infected (SI) model [42], and compared it with eight existing well-known identification methods. Our experimental results indicate the robustness and reasonableness of our proposed method over existing methods. The paper is organized as follows. In Section 2, we describe the parameters used in this paper, an overview of several well-known node identification measures is also given. The concept of effective distance will also be introduced. In Section 3, a new identification of influential nodes measure: effective distance gravity model is then proposed. In Section 4, a variety of experiments and comparisons with other measures are then illustrated to show the feasibility and effectiveness of our proposed method. We conclude our study in Section 5.

2. Preliminaries

In an undirected graph $G = (V, E)$, where the V represents the set of nodes and E represents the set of links. The number of nodes in the graph is denoted as n, where $n = |V|$. The adjacency matrix of graph G is $A = \{a_{ij}\}$, where $a_{ij} = 1$ if there is an edge between node i and node j.

2.1. Centrality measures

Definition 2.1. *Degree centrality (DC) identifies the importance of a node by comparing degree of the node. $DC(i)$ of each node i can be obtained by,*

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

where k_i *is the degree of node* i*. The node with large degree is of high influence [5].*

Definition 2.2. *The Betweenness centrality (BC) [30] measures the importance of a node by the number of shortest paths through it, which is shown as,*

$$BC(i) = \sum_{j,k \neq i} \frac{N_{jk}(i)}{N_{jk}}, \tag{2}$$

where N_{jk} *represents the number of shortest paths from node* j *to node* k*, and* $N_{jk}(i)$ *is the number of* N_{jk} *through node* i*. The more the number of shortest paths through node* i*, the more important is node* i *in the network.*

Definition 2.3. *Closeness centrality (CC) [5] evaluates the influence of nodes by the reciprocal of the sum of shortest path between nodes, which is shown as,*

$$CC(i) = \frac{1}{\sum_j^N d_{ij}}, \tag{3}$$

where d_{ij} *denotes the shortest length of path between node* i *and node* j*. The higher the* $CC(i)$ *is, the more important the node* i *is.*

Definition 2.4. *Eigenvector centrality (EC) [2] determines the influence of a node not only by the number of neighbors, but also by the importance of them.* $EC(i)$ *can be calculated by the formula below:*

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

The largest eigenvalue of A *is be represented by* λ *and* x_j *is the value of* j*th entry of the eigenvector corresponding to* λ*.*

Definition 2.5. *PageRank (PC) [31] uses an iterative approach to obtain the influence of nodes and it is very effective in calculating the importance of nodes in a directed network.* $PC(i)$ *of node* i *can be obtained by,*

$$PC(i)^q = \sum_{j=1}^{n} (a_{ij} \frac{PC(j)^{q-1}}{k_j}). \tag{5}$$

The influence score of node i *in step* q *is denoted as* $PC(i)^q$*. The higher the PC score is (when the PC finally converges), the more important the node is.*

2.2. Gravity model

The Gravity model (GM) is defined by the law of gravity. The influence $C(i)$ of node i can be estimated by GM [24] as follows:

$$C(i) = \sum_{i \neq j} \frac{k_i \times k_j}{(d_{ij})^2}. \tag{6}$$

2.3. Effective distance

Effective distance (ED) [3] is a distance abstracted from the probability, which reveals hidden pattern geometry of complex networks. The effective distance from node m to node n which are directly connected with each other is defined as [3]:

$$D_{n|m} = 1 - log_2(P_{n|m}), \tag{7}$$

where $D_{n|m}$ is the value of effective distance from node m to the node n if they are directly connected. $P_{n|m}$ is the probability of node m to the node n. $P_{n|m}$ can be obtained by [3]:

$$P_{n|m} = \frac{a_{mn}}{k_m}, (m \neq n), \tag{8}$$

where k_m is the out-degree of node m in directed graph and the degree of node m in undirected graph. a_{mn} is the element in the adjacency matrix of graph G.

Figure 1 reveals the computation of the effective distance between directly connected nodes. Clearly, $P_{n|m} \neq P_{m|n}$ and $D_{n|m} \neq D_{m|n}$ can be seen in Figure 1. The effective distance from the node i to itself is 0. An important difference from Euclidean distance is that effective distance is asymmetric.

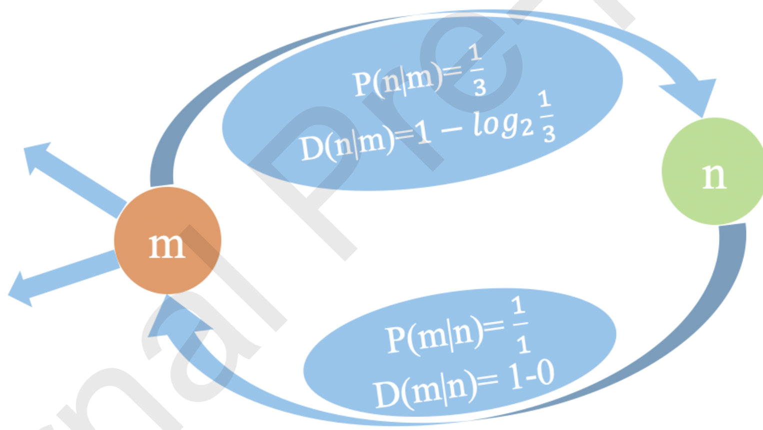


Figure 1: Effective distance between directly connected nodes.

For indirectly connected nodes, the effective distance between them can be obtained transitively. For example, the effective distance from node C to node F can be obtained by $D_{F|C} = D_{H|C} + D_{F|H}$-like form.

If there are multiple paths from node m to point n, we will use the shortest path between these two nodes. According to the idea of shortest path in the network, the shortest one of all effective distances from node m to n is selected as the final effective distance [3]:

$$D_{n|m} = \min\{D^1_{n|m}, D^2_{n|m}, D^3_{n|m}, ...\}, \tag{9}$$

where $D^i_{n|m}$ represents different effective distance from node m to node n.

It is worth noting that $D_{n|m}$ and $D_{m|n}$ are usually not equal. A simple example for the calculation of effective distance between indirectly connected nodes is presented in Figure 2. It can be found that there are three possible paths from node m to node n, among which $D^1_{n|m} = 2.5850 + 1 = 3.5850$, $D^2_{n|m} = 2.5850$, $D^3_{n|m} = 2.5850 + 1 + 2 = 5.5850$. According to the definition of effective distance shown in Equation (9), the final effective distance $D_{n|m}$ is the shortest path $D^2_{n|m}$, which is equal to 2.5850.

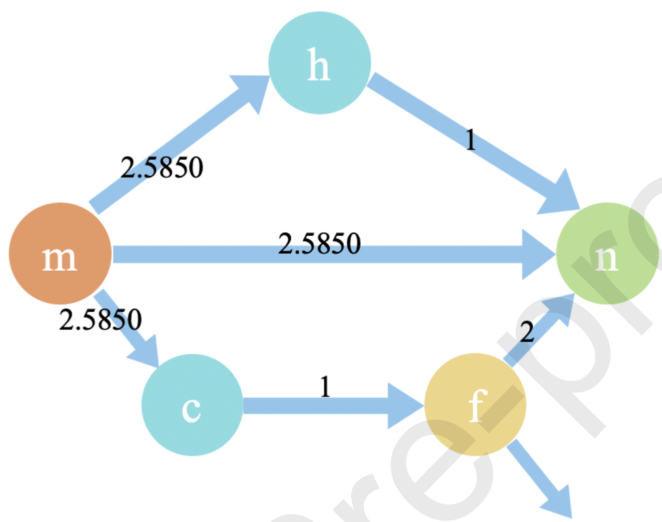


Figure 2: Effective distance between indirectly connected nodes.

3. Proposed method

3.1. Effective distance gravity model

In order to identity important nodes in real-world networks, it is not appropriate to only consider the local information of the node, or the global information of the network. Moreover, static and dynamic information need to be considered at the same time, since they are equally important. The lack of any type of information will lead to unsatisfactory results. Other than the geometric structure of the network that can be observed, there is often a hidden dynamic structure which dominates the dynamic process of the network. The dynamic structure contains effective information about the dynamic interaction between nodes, which can help us better identify important nodes in the real-world, since it may play a leading role. A large number of scientific experiments have proved that the phenomenon which can be easily observed is not necessarily the most essential. For instance, behind the falling of feathers and iron balls is the combined effect of gravity and air resistance, not like what can be easily observed, heavy things fall faster. Consequently, it is necessary to incorporate the explicit static information and the dynamic information contained in the hidden topology of network. Thus, an effective distance gravity model (EffG) is proposed. The framework of our proposed method is shown in Figure 3, and further details of the calculation will be given below.

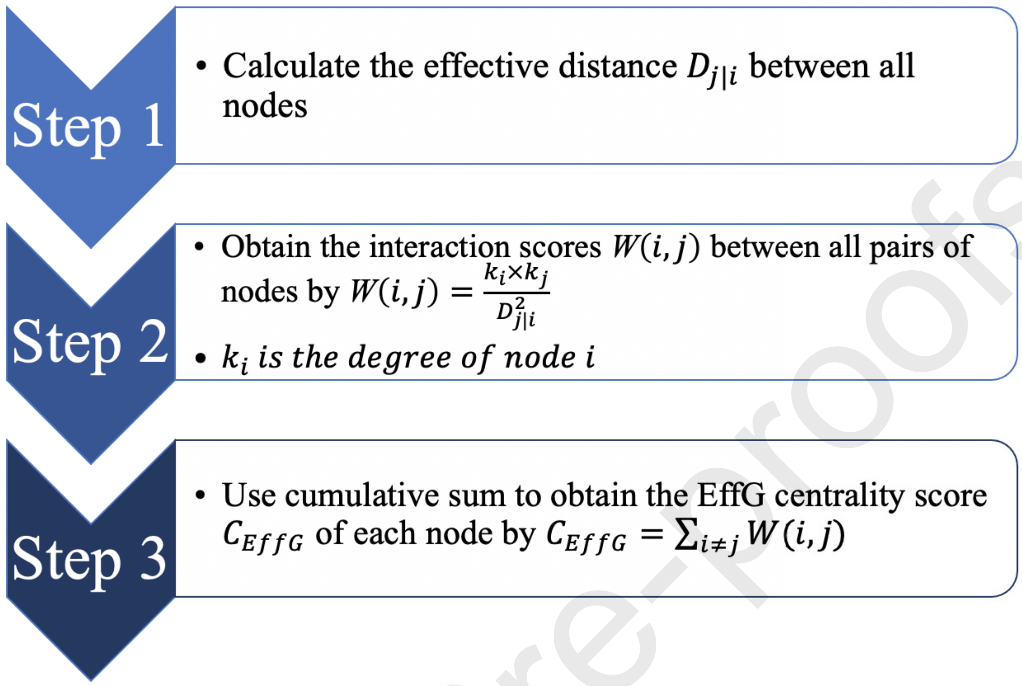


Figure 3: The implementation steps of our proposed EffG method.

Step 1: Calculate the effective distance between all nodes.

Euclidean distance only focuses on the static topological distance of the node, thus cannot effectively mine the hidden dynamic information. For instance, the Euclidean distance from A to B and that from B to A is the same because Euclidean distance is a non-directional distance measurement. However, this may not be the case in real-world propagation process. In the dynamic propagation process, the distance is often directional. It may be easier to propagate from A to B than from B to A. This difficulty can be measured by a directional distance. In order to overcome this problem, the effective distance is proposed by Dirk Brockmann and Dirk Helbing [3]. Their work states that if the probability is used to construct a new distance metric to replace the conventional geographic distance, the complex space-time patterns can be reduced to simple and homogeneous wave propagation patterns. Their experimental results indicate that when the parameters of the disease and the network structure are given, the flux information obtained from the effective distance can reliably predict the arrival time of disease. In their experiments, the predicted time of arrival of SARS disease in 2003 and the H1N1 pandemic in 2009 in each region is a good proof of this result [3]. Hence, it is reasonable to use the effective distance to effectively mine the hidden dynamic topology of network and the dynamic information between nodes. The effective distance between nodes can be calculated by the Equations (7) - (9).

Step 2: Calculate the interaction scores between all pairs of nodes.

The interaction scores between nodes can be obtained by the effective distance between nodes and the gravity model, which considers both global information of the network and the local information of the nodes. Moreover, the introduction of effective distance renders our method suitable for mining the hidden dynamic topology of network and the dynamic information between nodes. Consequently, the explicit static information and hidden dynamic information of the network can now be comprehensively considered. Based on the gravity formula, the specific interaction score between nodes is defined as follows.

$$W_{interaction}(i,j) = \frac{k_i \times k_j}{D_{j|i}^2}, \tag{10}$$

where k_i and k_j are the degree of node i and node j respectively. $D_{j|i}$ is the effective distance from node i to node j. $W_{interaction}(i,j)$ denotes the interaction score between node i and node j.

Step 3: Use cumulative sum to get the EffG centrality score of each node.

After obtaining the interaction scores between nodes, the EffG score of each node is obtained by the cumulative sum of its interaction scores $W_{interaction}$ with other nodes in the network. Hence, the EffG centrality scores can be defined as follows.

$$\begin{aligned} C_{EffG}(i) &= \sum_{j=1, j\neq i}^{N} W_{interaction}(i,j) \\ &= \sum_{j=1, j\neq i}^{N} \frac{k_i \times k_j}{D_{j|i}^2}, \end{aligned} \tag{11}$$

where N is the number of all nodes in the network and $C_{EffG}(i)$ represents the EffG centrality scores of node i.

3.2. A numerical example

In order to better explain our proposed identification method EffG, a numerical example is given below as a way of illustration of how the EffG works. We take node 2 as an example to calculate the EffG scores of it. Figure 4(a) is the graph of a network, and adjacency matrix is represented in Figure 4(b).

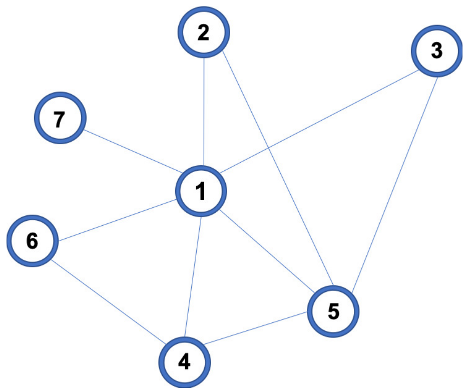
The degree of each node is shown in Table 1.

Table 1: The degree of each node in Figure 4.

Node	node1	node2	node3	node4	node5	node6	node7
degree	6	2	2	3	4	2	1

Step 1: Calculate the effective distance between all nodes.

First, we calculate the effective distance between nodes by Equations (7) - (9). Under normal circumstances, the $P_{i|j} \neq P_{j|i}$ and $D_{i|j} \neq D_{j|i}$. The effective distance between node 2 and node 7 can be



(a) The graph of a simple network.

$$\begin{Bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{Bmatrix}$$

(b) The adjacency matrix of network.

Figure 4: A simple network with seven nodes.

calculated as follows:

$$\begin{aligned} D_{2|7} &= \min\{D_{1|7} + D_{2|1}, D_{1|7} + D_{5|1} + D_{2|5}, ...\} \\ &= \min\{1 - \log_2(P_{1|7}) + 1 - log_2(P_{2|1}), 1 - \log_2(P_{1|7}) + 1 - log_2(P_{5|1}) + 1 - \log_2(P_{2|5}), ...)\} \\ &= 2 - \log_2(P_{1|7}) - log_2(P_{2|1}) \\ &= 4.5850, \qquad (12) \\ D_{7|2} &= \min\{D_{1|2} + D_{7|1}, D_{5|2} + D_{1|5} + D_{7|1}, ...\} \\ &= \min\{1 - \log_2(P_{1|2}) + 1 - log_2(P_{7|1}), 1 - \log_2(P_{5|2}) + 1 - log_2(P_{1|5}) + 1 - \log_2(P_{7|1}), ...) \\ &= 5.5850. \end{aligned}$$

As observed, $D_{2|7}$ and $D_{7|2}$ are different as we have discussed above. Using the same procedures, the effective distance from node 2 to the other nodes can be also calculated. The result is shown in Table 2.

Table 2: Effective Distance from node 2 to the other nodes in Figure 4.

	$D_{1\|2}$	$D_{2\|2}$	$D_{3\|2}$	$D_{4\|2}$	$D_{5\|2}$	$D_{6\|2}$	$D_{7\|2}$
ED	2.0000	0	5.0000	5.0000	2.0000	5.5850	5.5850

Step 2: Calculate the interaction scores between all pairs of nodes.

Then, the interaction scores $W_{interaction(i,j)}$ between node 2 and the other nodes can be calculated by Equation (10) as follows:

$$W_{interaction}(2,1) = \frac{k_2 \times k_1}{D^2_{1|2}} = \frac{2 \times 6}{2.0000^2} = 3.0000,$$
$$W_{interaction}(2,3) = \frac{k_2 \times k_3}{D^2_{3|2}} = \frac{2 \times 2}{5.0000^2} = 0.1600,$$
$$W_{interaction}(2,4) = \frac{k_2 \times k_4}{D^2_{4|2}} = \frac{2 \times 3}{5.0000^2} = 0.2400,$$
$$W_{interaction}(2,5) = \frac{k_2 \times k_5}{D^2_{5|2}} = \frac{2 \times 4}{2.0000^2} = 2.0000,$$
$$W_{interaction}(2,6) = \frac{k_2 \times k_6}{D^2_{6|2}} = \frac{2 \times 2}{5.5850^2} = 0.1282,$$
$$W_{interaction}(2,7) = \frac{k_2 \times k_7}{D^2_{7|2}} = \frac{2 \times 1}{5.5850^2} = 0.0641.$$

Step 3: Use cumulative sum to obtain the EffG centrality score of node 2.

The EffG scores of node 2 can be obtained by Equation (11) as follows:

$$C_{EffG}(2) = \sum_{j \neq 2} W_{interaction}(2,j) = 5.5923.$$

The EffG scores of the other nodes can be calculated by the same procedure, which are shown as follows:

$$\begin{aligned} C_{EffG}(1) &= \sum_{j \neq 1} \frac{k_1 \times k_j}{D^2_{j|1}} = 6.5360, \\ C_{EffG}(3) &= \sum_{j \neq 3} \frac{k_3 \times k_j}{D^2_{j|3}} = 5.5923, \\ C_{EffG}(4) &= \sum_{j \neq 4} \frac{k_4 \times k_j}{D^2_{j|4}} = 5.8511, \\ C_{EffG}(5) &= \sum_{j \neq 5} \frac{k_5 \times k_j}{D^2_{j|5}} = 6.1265, \\ C_{EffG}(6) &= \sum_{j \neq 6} \frac{k_6 \times k_j}{D^2_{j|6}} = 5.2011, \\ C_{EffG}(7) &= \sum_{j \neq 7} \frac{k_7 \times k_j}{D^2_{j|7}} = 2.1184. \end{aligned} \tag{13}$$

From the Figure 4(a). we can see that node 1 is located in the central position of the network, it has the strongest connection with the other nodes and covers the most number of shortest paths in the network. Without node 1, the network will be broken into multiple isolated parts. Thus, it is reasonable that to say that node 1 is the most influential node in this network. Moreover, node 7 can be seen as the least influential in the network and the EffG score that matches it is the lowest. The importance of node 2 and node 3 in this network is the same, they also have the same EffG score. This simple example shows that our proposed method EffG is practical and objective.

4. Application

To verify the feasibility and effectiveness of our proposed method, five experiments were performed on twelve real-world networks, in comparison with eight existing well-known methods including DC, CC, EC, PC, BC, Gravity, Quasi-Laplacian centrality (QL) [27] and iterative resource allocation-based method (Ira) [34].

4.1. Datasets

The experiment was conducted on twelve real-world networks, Jazz, NS, GrQc, Email, EEC, Facebook, PB, PDZBase, Haggle, Infectious, Physicians and USAir, including two communication networks (Email, EEC), one transportation network (USAir), two social networks (Facebook, PB), three cooperative networks (Jazz, NS, GrQc, Physicians), one protein-protein interaction network (PDZBase) and one wireless mobile device network (Haggle). Among them, Email is a network where users send emails and communicate with each other. EEC is a network where European researchers communicate via exchange of emails. Physicians is a network where physicians collaborate. USAir represents a US air transportation network. Infectious is a network related to infectious diseases. Facebook describes a social network derived from Facebook. PB is a blog network. Jazz describes a practical network of Jazz musician collaborations. NS is a network where scientists collaborate and work together. GrQc is a network published on preprint. PDZBase is a network that describes the interaction between protein. Haggle is a network of wireless mobile devices. Other relevant information about the networks are displayed in Table 3.

Table 3: **The basic topology information of the twelve real-world networks.** n and m are the number of nodes and edges of the network, $\langle k \rangle$ and $\langle d \rangle$ are the average degree and average distance of the network [8]. C and r are the network's clustering coefficient and assortative coefficient [38].

Networks	n	m	$\langle k \rangle$	$\langle d \rangle$	C	r
Jazz	198	2472	27.6970	2.2350	0.6334	0.0202
NS	379	914	4.4832	6.0419	0.7981	-0.0817
GrQc	4158	13422	6.4560	6.0494	0.6648	0.6392
EEC	986	16064	32.5842	2.5869	0.4505	-0.0257
Email	1133	5451	9.6222	3.7160	0.1101	0.0782
PB	1222	16714	27.3553	2.7375	0.3600	-0.2213
Facebook	4039	88234	43.6910	3.6925	0.6170	0.0636
USAir	332	2126	12.8072	2.7381	0.7494	-0.2079
Physicians	241	1098	9.1120	0.7964	0.2510	-0.0558
PDZBase	212	2672	24.6270	2.3300	0.3040	0.0959
Haggle	274	28244	206.16	2.4240	0.5660	-0.4743
Infectious	410	17298	84.3800	3.5700	0.4360	0.2258

4.2. Evaluating with susceptible infected model

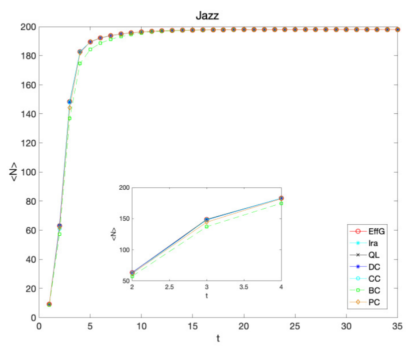
The susceptible infected (SI) model [42] can be used to estimate the node's capability of transmission in the network, which indirectly reflects the influence of the node. In the SI model, there are two components which deserve our attention: (1) susceptible state (2) infection state. In the process, the infected nodes infect the surrounding susceptible nodes with a certain probability. The parameters utilized in the SI model are t, $F(t)$, β and K. The experimental simulation time of the susceptible infected model is denoted by t. β represents the probability of nodes infection. K is the number of experiments. The average number of infected nodes at time t is denoted by $F(t)$. It can be easily understood that the more important the node, the greater its influence. Under the condition that the infection time t and the infection probability β are both the same, the more influential node will cause more surrounding nodes to be infected. Hence, $F(t)$ reflects the influence of the initial infected node. The node with higher $F(t)$ is of greater importance.

In order to estimate the capability of different measures in identifying the influential nodes, the SI model was applied on ten different real-world networks. In the experiments, the top-100 nodes ranked by different methods was selected firstly. After that, the top-100 nodes were used as the initial infection nodes in the SI model separately. Finally, the average number of infected nodes $F(t)$ was calculated for each method respectively. Furthermore, the SI model in experiments is given the same propagation probability β to control the variables and the β was set to be 0.2 in our experiment. In this experiment, the $F(t)$ is denoted as $\langle N \rangle$.

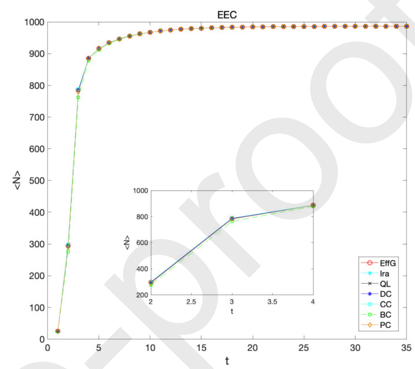
The experimental results are shown in Figure 5 and Figure 6. The node with more final infected nodes is of greater importance. Hence, the faster the curve rises and the higher the curve, the more influential the nodes in the initial infection set are. That is to say, the more effective the identification method is. In Figures 5 (a) – 5 (c), the difference among these methods is not obvious, which means they are basically consistent. The curve of BC is always the lowest while the curves corresponding to QL, Ira and DC are usually in the middle. In Figures 6 (a) – 6 (b), the curve of CC is always the highest while that of EffG is the second highest. In Figures 6 (c) – 6 (e), all methods performed significantly differently. The curves corresponding to EffG and CC are higher and rise faster than the others. This means that the top-100 nodes selected by CC and EffG are more influential. Moreover, as shown in Figure 6 (c), the curve of EffG in NS network is the highest and rises fastest. In summary, it can be seen that the curves corresponding to our proposed method EffG and CC are always at the highest or second highest position. In addition, the slope of the curve corresponding to them is also very large in all networks mentioned, which means that the initial node set selected by the two have a stronger infection ability. In other words, CC and our proposed method EffG can select influential nodes more accurately. However, in most of the networks mentioned above, the curves corresponding to PC and BC rise more slowly that the others.

4.3. Comparison ranking results

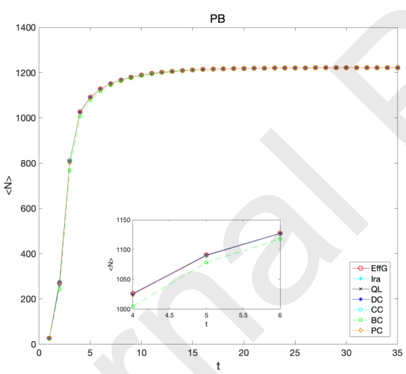
The top-10 vital nodes in the Jazz network ranked by different methods including our proposed method (EffG), DC, BC, CC, PC, EC, Gravity and SI model in which $\beta = 0.2$, $t = 20$, are listed in Table 4. As these methods consider different information and properties of the network, their ranking list may be different from other methods. In this experiment, the number of coincident nodes demonstrates the



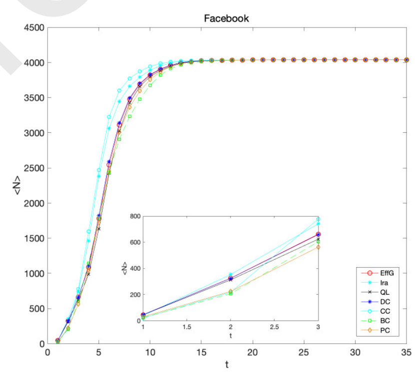
(a) Jazz



(b) EEC

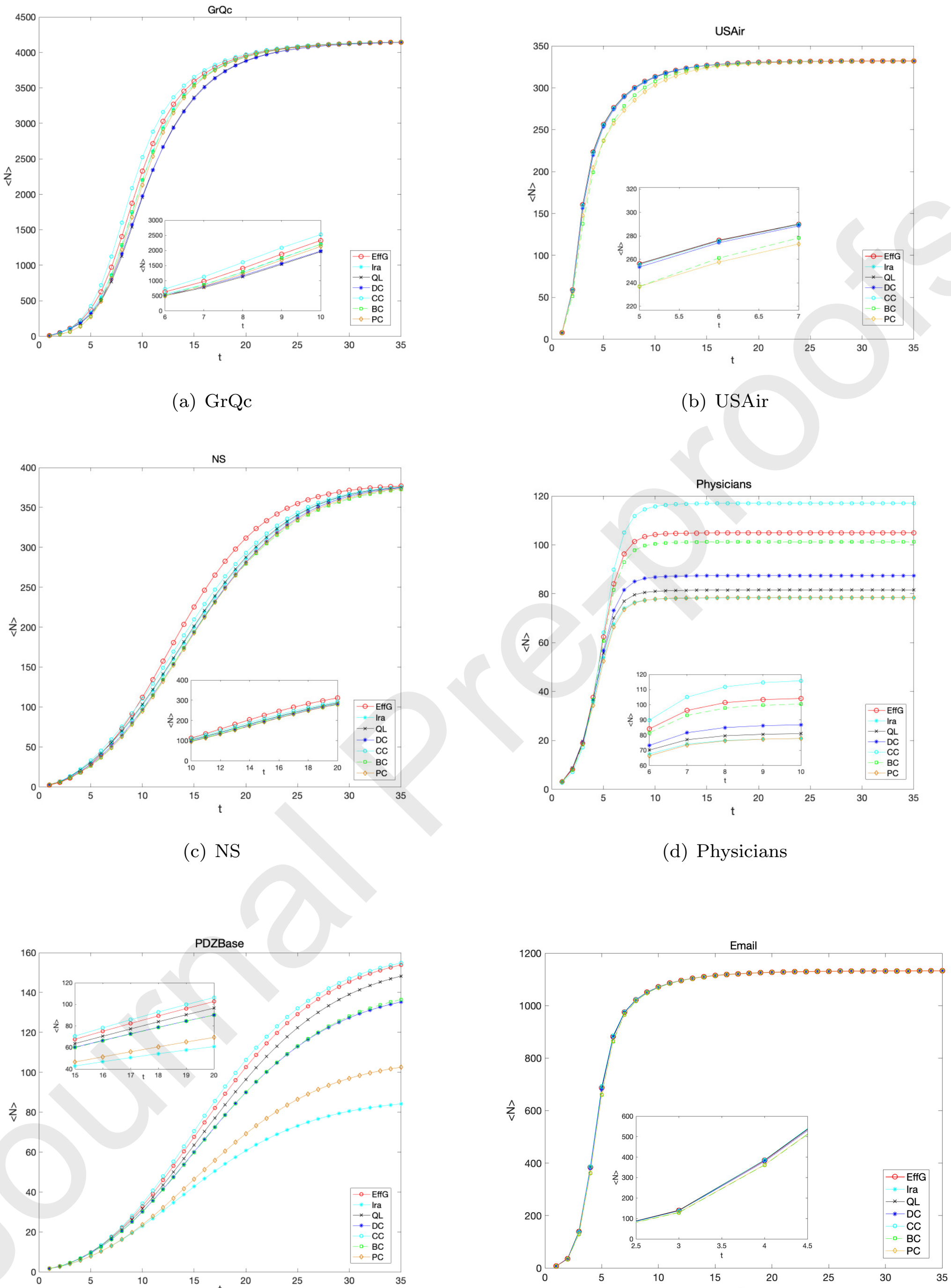


(c) PB



(d) Facebook

Figure 5: The infection ability of the top-100 nodes selected by different methods in different networks.



(a) GrQc

(b) USAir

(c) NS

(d) Physicians

(e) PDZBase

(f) Email

Figure 6: The infection ability of the top-100 nodes selected by different methods in different networks.

effectiveness of our method to a certain extent. This idea is similar to the degree of support in evidence theory. The greater the amount of evidence supporting A, the higher the credibility of A.

In Physicians network, the most similar lists to EffG are Gravity and BC, since the number of nodes consistent between them is up to 9. The number of nodes consistent between other methods and EffG is 7-8, which is also relatively high but lower than BC and Gravity. In the Jazz network, the most similar lists to EffG are Gravity and EC, since the top-10 nodes selected by them is the same. However, BC has the least consistent nodes with EffG, which is different from that in Physicians network. In the USAir network, it is interesting that the most consistent lists with EffG are still EC and Gravity while the most inconsistent list is still BC, as observed similarly in the Jazz network. The number of nodes consistent with EffG in CC and SI model decreases, but that of PC increases. Moreover, as can be observed from Table 4, all approaches successfully identify the node 118 as the most influential in USAir. Through observation of the results for the PB network, the most consistent list with EffG are DC and PC since the same top-10 nodes with EffG. CC, BC, SI model and Gravity all have 9 same nodes with EffG while EC only has 5. In summary, across the different networks, the proposed method has close performance with different methods in different networks. This phenomenon reflects that EffG has both global and local properties, as well as static and dynamic information, since it is consistent with different types of methods on different networks. This phenomenon also indicates that EffG is better at adapting to different networks. In addition, in 4 real-world networks shown in Table 4, the number of consistent nodes across EffG and the other methods are high. The high number of coincidences with other approaches confirms the justification for our proposed method.

4.4. Relation of proposed method with SI model

The Kendall correlation coefficient [6] is used to measure the correlation of two sequences. The absolute value of the Kendall coefficient is between 0 and 1. The larger the Kendall coefficient's absolute value, the stronger the correlation between the two sequences. If the Kendall coefficient between the two sequences is 0, it means the two sequences have no correlation. In this experiment, the Kendall coefficient is used to measure the correlation between sequences generated by different identification methods and the sequence generated by the SI model, thereby inferring the effectiveness of the identification method. The greater the absolute value of the Kendall coefficient, the more valid the identification method.

Given two sequences with S elements, $X = (x_1, x_2, x_3, \ldots, x_s)$ and $Y = (y_1, y_2, y_3, \ldots, y_s)$. Let (x_i, y_i) by a set of sequence pairs. For any pairs (x_i, y_i) and (x_j, y_j) that $i \neq j$, if both $x_i > x_j$ and $y_i > y_j$ or both $x_i < x_j$ and $y_i < y_j$, they are classified as concordant sequence pairs. If both $x_i > x_j$ and $y_i < y_j$ or both $x_i < x_j$ and $y_i > y_j$, they are classified as the discordant sequence pairs. The Kendall correlation coefficient of two sequences X and Y, is defined as follows.

$$tau = \frac{n_+ - n_-}{S \times (S-1)}, \tag{14}$$

Table 4: Top-10 nodes in four real-world networks selected by our proposed method (EffG), DC, BC, CC, PC, EC, Gravity and SI model.

Physicians							
EffG	DC	CC	BC	PC	EC	SI	Gravity
15	127	15	15	127	15	15	15
74	15	40	74	15	127	40	127
40	121	74	23	121	40	12	74
11	74	11	11	128	121	74	40
12	128	12	40	74	12	11	23
23	23	23	127	23	74	69	12
127	40	69	12	194	11	23	11
69	11	29	69	40	23	54	121
13	12	54	29	11	13	13	69
10	10	4	10	12	128	16	13
Jazz							
EffG	DC	CC	BC	PC	EC	SI	Gravity
8	8	8	8	8	100	8	8
100	100	100	155	100	4	100	100
4	4	131	100	131	8	131	4
131	131	194	186	4	131	194	131
194	194	69	131	186	80	111	194
80	80	4	136	136	129	53	80
69	129	32	127	194	5	69	129
129	69	53	60	69	194	4	69
162	162	111	28	28	69	59	162
53	77	162	69	175	53	67	53
USAir							
EffG	DC	CC	BC	PC	EC	SI	Gravity
118	118	118	118	118	118	118	118
261	261	261	8	261	261	261	261
255	255	67	261	182	255	201	255
182	152	255	201	152	182	166	230
152	182	201	47	255	152	182	152
166	230	182	182	230	230	255	182
230	166	47	255	166	112	47	166
67	67	166	152	201	67	67	67
112	112	248	313	67	166	248	147
147	201	112	13	8	147	112	112
PB							
EffG	DC	CC	BC	PC	EC	SI	Gravity
127	127	838	672	672	127	838	127
838	838	127	127	127	48	127	838
48	672	497	768	768	497	497	48
497	48	48	838	838	566	566	497
672	497	890	497	497	283	48	672
566	768	566	1178	48	147	1178	566
768	1006	768	48	1006	838	922	1006
1006	566	922	782	922	84	1006	922
922	922	1178	922	1178	384	890	768
1178	1178	672	566	566	498	768	890

where n_+ and n_- are the number of concordant sequence pairs and discordant sequence pairs respectively, S is the total number of sequence pairs. The value of Kendall correlation coefficient is denoted as tau.

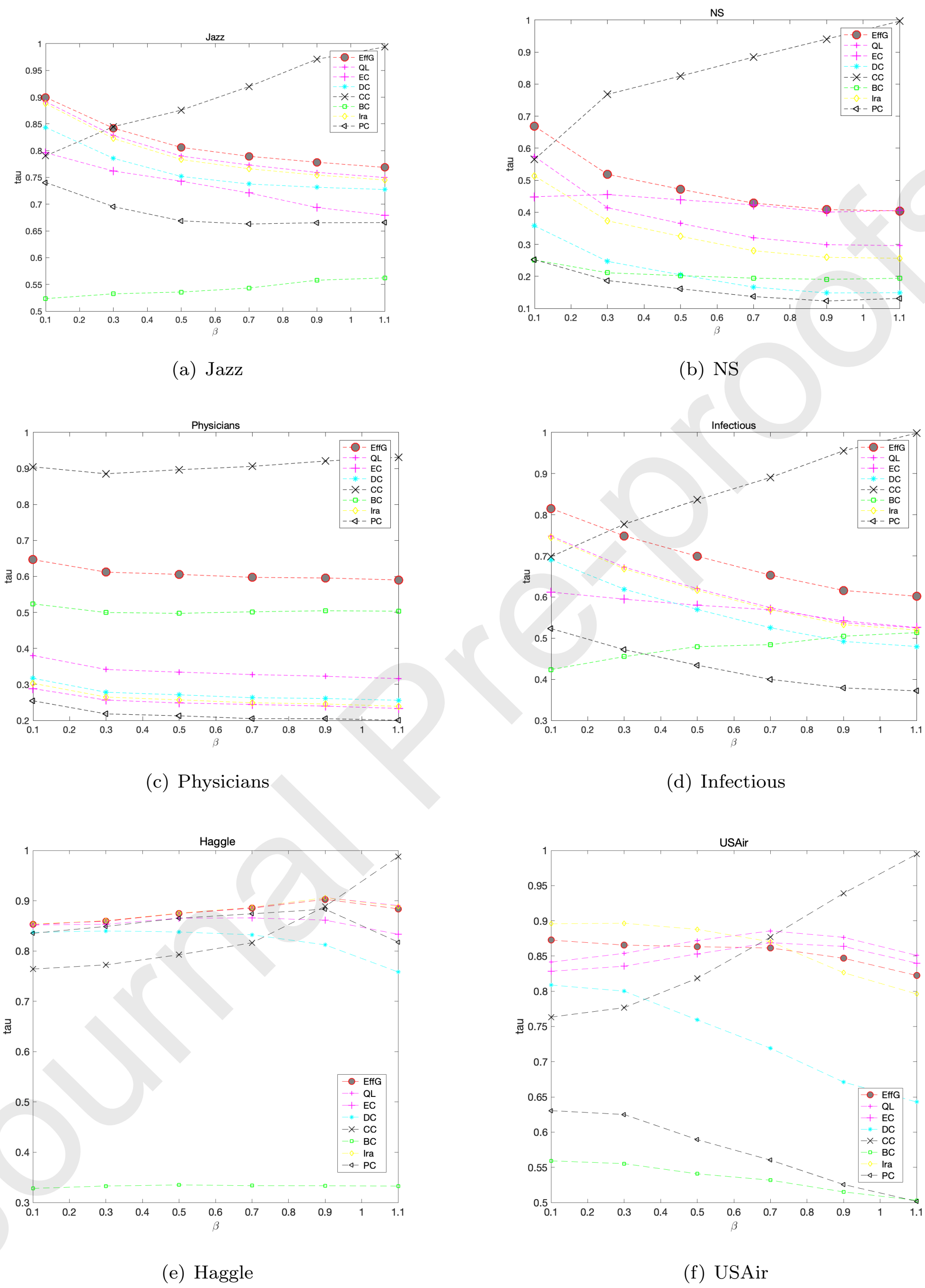
In this experiment, the evaluation of the effectiveness of the method is based on the correlation with the SI model. In all data sets, the different infection probability β was given to the SI model respectively to obtain a standard centralized sequence. Then Kendall correlation coefficient of SI model sequence and the other method's sequence was calculated. In the experiment, the infection probability β changed from 0.1 to 1.1 and the SI model independently ran 120 times with the infection time $t = 20$ to take the average across different networks. The experimental results are shown in Figure 7, where tau represents the value of Kendall correlation coefficient. Higher tau value indicates stronger positive correlation between centrality method and SI model.

In Figures 7 (a)-7 (d), the curve corresponding to EffG is always at the second highest position and that of CC is the highest. However, the curves of BC and PC is lower than other methods. In the Figure 7 (d), as β increases, all curves except the CC show a downward trend. As observed in Figure 7 (e), the curve of EffG is the highest when $0.1 < \beta < 0.9$ while that of BC is the lowest in Haggle network. Moreover, the Kendall correlation coefficients of other methods are very close. As shown in Figure 7 (f), the Kendall correlation coefficient is the smallest in USAir while that of CC is the largest when $0.7 < \beta < 1.1$. In summary, the correlation between CC and SI sequence is the strongest in most cases. The correlation between SI sequence and EffG is always at least the second strongest, even being the strongest in some cases. This means that EffG has a stronger correlation with the SI model than most existing methods, which demonstrates its effectiveness and superiority.

4.5. Compare the correlations between proposed method and SI model.

In this experiment, four real-world networks were used to evaluate the feasibility of our method, including PDZBase, NS, Email and USAir. First, the ranking of nodes on each network is derived by different methods, DC, BC, CC, EC and our proposed method (EffG). Then, each node will be used as the initial infected node in the SI model, and the final number of infected nodes will be calculated by $t = 15$. Finally, the correlation between the node ranking and the average number of nodes infected by them at time t will be established. The results are shown in Figure 8, where the Rank represents the node ranking.

The node with higher ranking is of stronger capability to infect other nodes, which means the node is more influential. This means that the higher the ranking of a node is, the greater the number of nodes which should eventually be infected by it. The lower the node's ranking is, the smaller the final number of infected nodes should be. Hence, the curve corresponding to a good identification method should basically continue to decline. As can be observed in Figure 8 (a), the curves of BC and DC have obvious fluctuation while the downward trend in them is not very obvious. However, the curve corresponding to EffG does not fluctuate much and maintains a continuous decline. As shown in Figure 8 (b), the



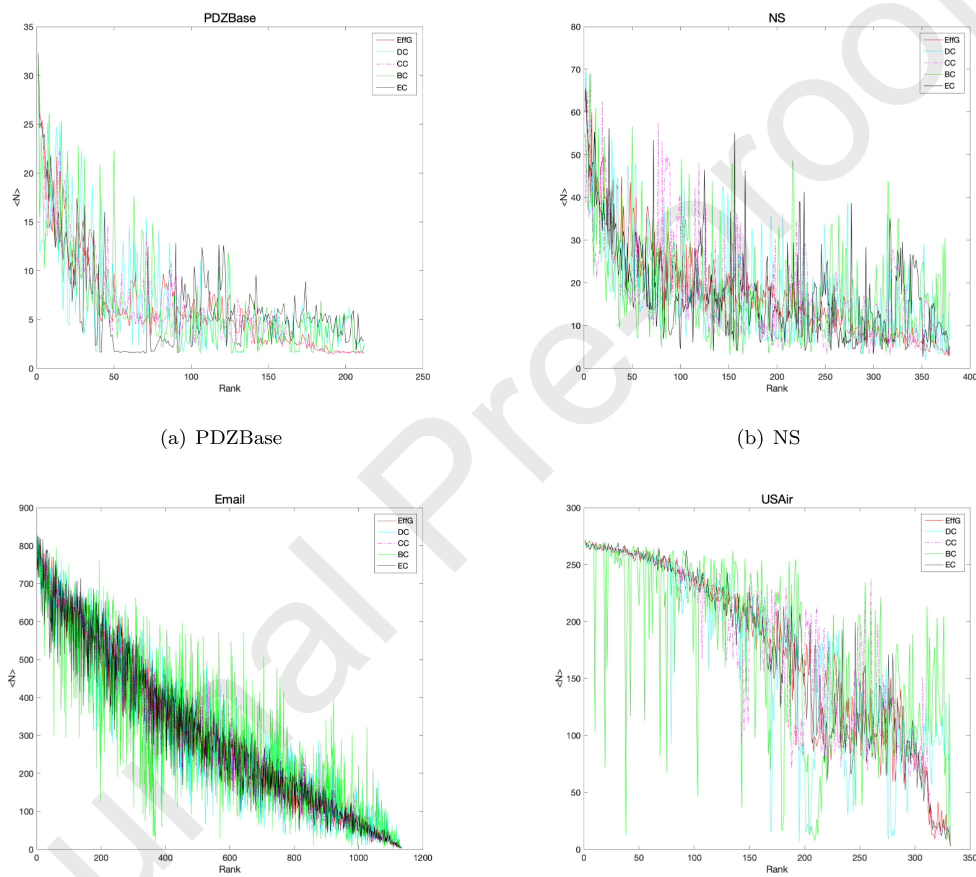
(a) Jazz (b) NS

(c) Physicians (d) Infectious

(e) Haggle (f) USAir

Figure 7: The Kendall correlation coefficient between SI model sequence and other method's sequence in different networks.

fluctuation of all methods except for EffG is very obvious while the downward trend is also not obvious in NS. In the Email network, the curves of all methods show a clear downward trend. As shown in Figure 8 (d), the curve corresponding to BC has almost no downward trend and fluctuates a lot. The other methods show a clear and continuous downward trend. In summary, the curve corresponding to our proposed method EffG is continuously declining, and has little fluctuation compared with other methods. Furthermore, it can be easily verified that the curve corresponding to BC fluctuates greatly during the decline and does not clearly show a downward trend. Therefore, it can be inferred that our proposed method EffG is valid and reasonable compared to other methods to some extent.



(a) PDZBase (b) NS

(c) Email (d) USAir

Figure 8: The correlation between different methods and SI model in different networks.

4.6. The relationship between different methods

In this experiment, the centrality scores obtained by different methods are compared to obtain the relationship between different methods. The coordinate axes in the figures represent different centrality scores. The color of each point represents the influence of node obtained by the SI model. In the SI model used in this experiment, the infection time $t = 20$, the experiment times $K = 120$ and the infection rate $\beta = 0.2$. The relationship between EffG and other methods is shown in Figures 9-14.

In Figures 9-14, EffG and other centrality methods show a strong positive correlation. Moreover, it can be seen from the color of the point that the obtained centrality score of the node is also in line with expectations. The more important the node identified, the stronger the infection ability. As shown in Figures 13-14, compared with EC in the NS network and Infectious network, our method can well avoid the situation where many nodes are tied due to the same centrality score, which is very beneficial for practical applications. In summary, the proposed EffG can effectively identify important nodes in complex networks. In addition, compared to some other centrality methods, it can identify the importance of nodes with close values obtained by some existing methods.

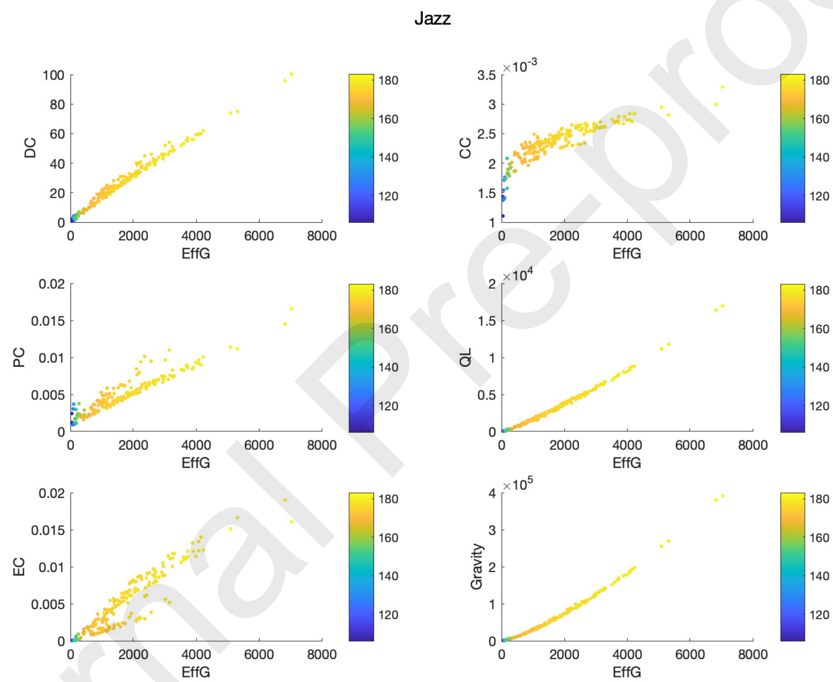


Figure 9: The correlation between different methods in Jazz.

5. Conclusion

An original and novel method for identifying influential nodes based on an effective distance gravity model is proposed. On top of single-dimensional factors, our proposed EffG model also comprehensively considers the local information of the node and global information of the network based on the idea of multi-source information fusion. An important contribution is that the EffG model uses the concept of effective distance to replace the traditional static Euclidean Distance. EffG is able to take full advantage

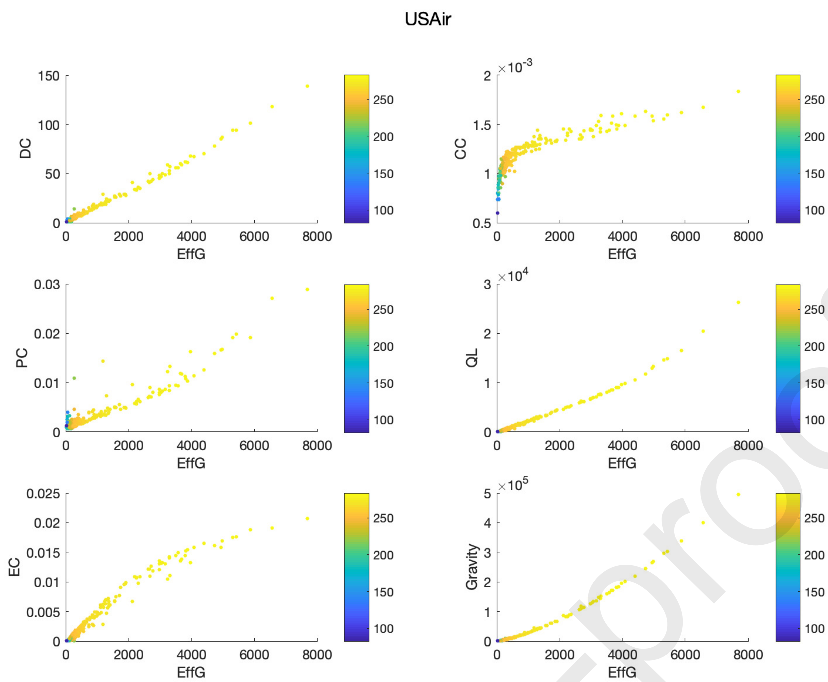


Figure 10: The correlation between different methods in USAir.

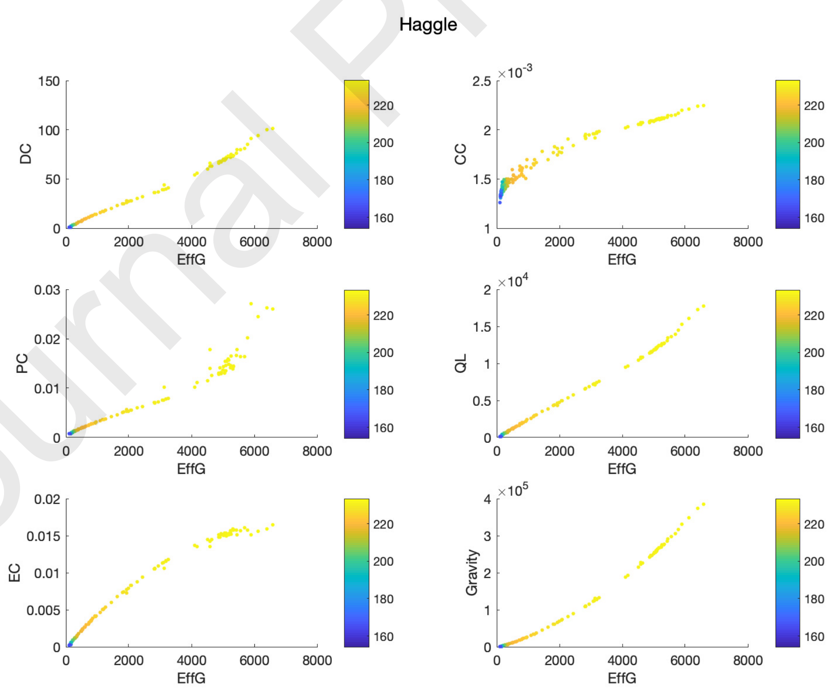


Figure 11: The correlation between different methods in Haggle.

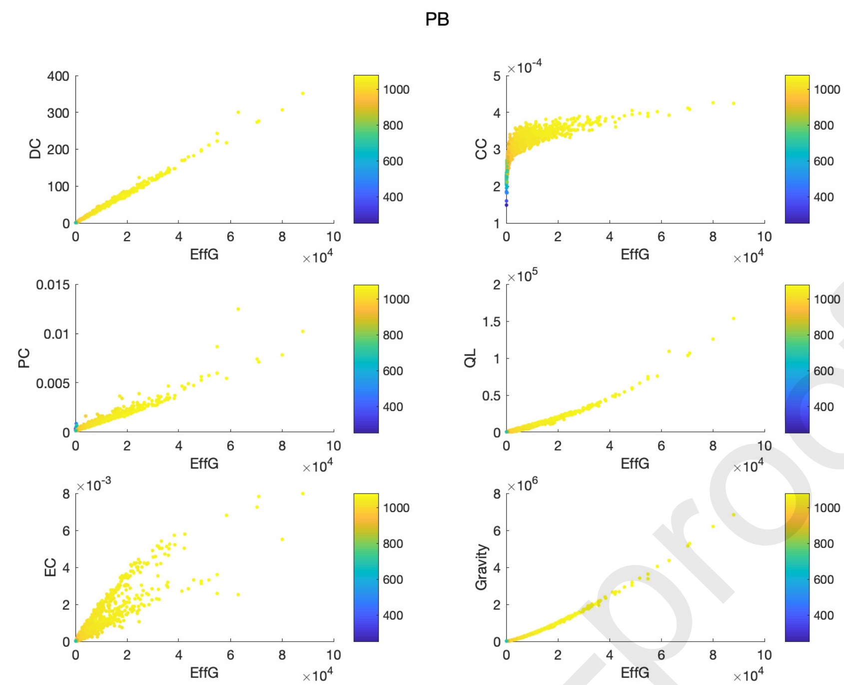


Figure 12: The correlation between different methods in PB.

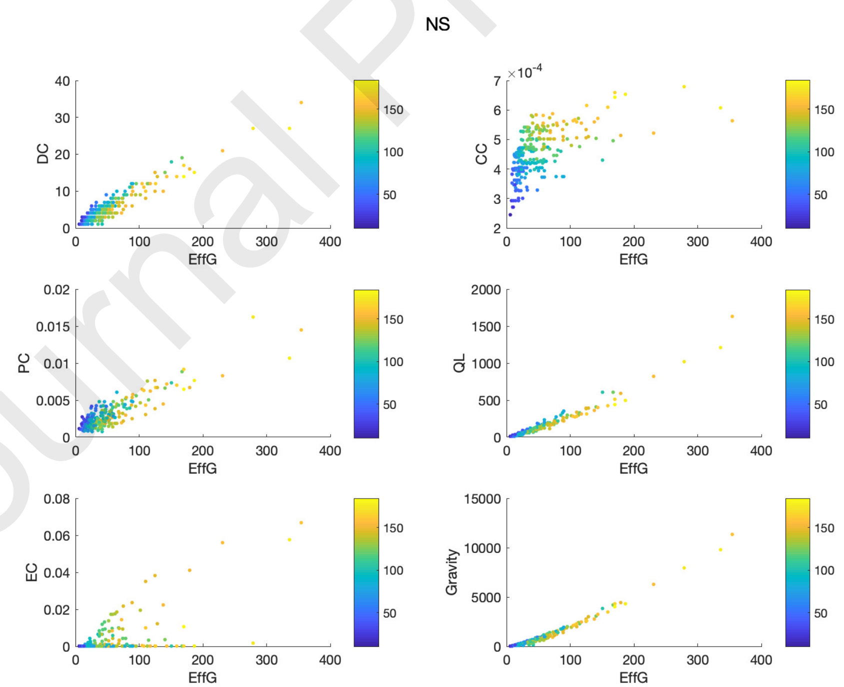


Figure 13: The correlation between different methods in NS.

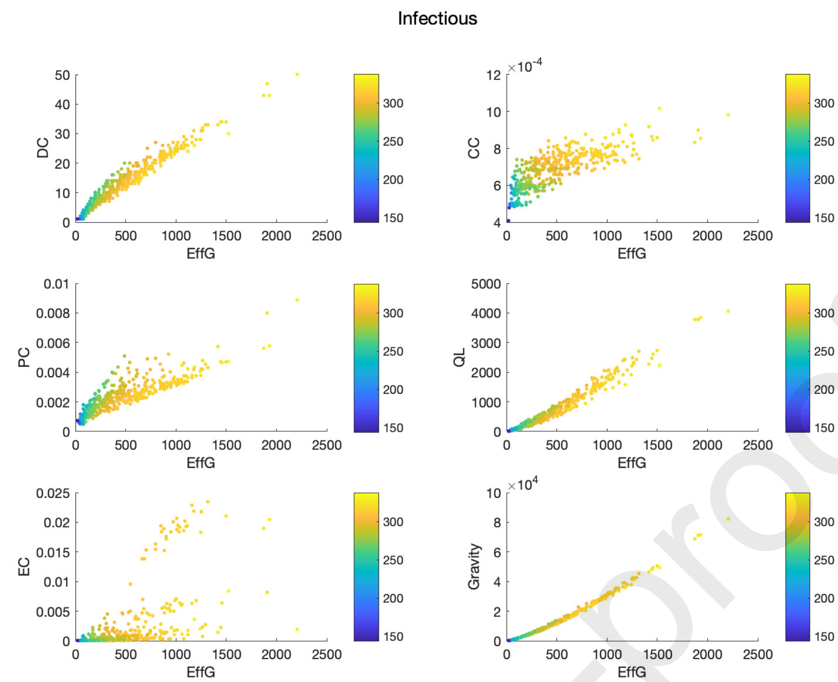


Figure 14: The correlation between different methods in Infectious.

of the dynamic information exchange between nodes in the network. In addition, EffG can help to unravel the hidden topology of the network that drives many dynamic information propagation processes. Importantly, the identification of influential nodes by EffG is aligned to real-world conditions. In order to verify the effectiveness and feasibility of this method, a variety of experiments has been conducted on twelve real-world networks and EffG has been compared with eight existing well-known methods. The experimental results indicate that our method performs well under dynamic information propagation and across several test-examples, thereby demonstrating its potential applications in network science, biological and social system, time series and information propagation. Although EffG has achieved great success in influential node identification, our future work is motivated by whether EffG can be improved to lower its time complexity. In this way, the method can be adapted for use with large-scale networks.

Conflict of interests

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

Data Availability Statements

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgment

The authors greatly appreciate the reviews' suggestions and the editor's encouragement. The work is partially supported by National Natural Science Foundation of China (Grant No. 61973332).

References

[1] Mohammed Alshahrani, Zhu Fuxi, Ahmed Sameh, Soufiana Mekouar, and Sheng Huang. Efficient algorithms based on centrality measures for identification of top-k influential users in social networks. *Information Sciences*, 2020.

[2] Phillip Bonacich and Paulette Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Social networks*, 23(3):191–201, 2001.

[3] Dirk Brockmann and Dirk Helbing. The hidden geometry of complex, network-driven contagion phenomena. *science*, 342(6164):1337–1342, 2013.

[4] Duan-Bing Chen, Rui Xiao, An Zeng, and Yi-Cheng Zhang. Path diversity improves the identification of influential spreaders. *EPL (Europhysics Letters)*, 104(6):68006, 2014.

[5] Duanbing Chen, Linyuan Lü, Ming-Sheng Shang, Yi-Cheng Zhang, and Tao Zhou. Identifying influential nodes in complex networks. *Physica A: Statistical mechanics and its applications*, 391(4):1777–1787, 2012.

[6] Regino Criado, Esther García, Francisco Pedroche, and Miguel Romance. A new method for comparing rankings through complex networks: Model and analysis of competitiveness of major european soccer leagues. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 23(4):043114, 2013.

[7] RC De Souza, Daniel R Figueiredo, AA de A Rocha, and Artur Ziviani. Efficient network seeding under variable node cost and limited budget for social networks. *Information Sciences*, 514:369–384, 2020.

[8] Zhaogang Ding, Xia Chen, Yucheng Dong, and Francisco Herrera. Consensus reaching in social network degroot model: The roles of the self-confidence and node degree. *Information Sciences*, 486:62–72, 2019.

[9] Hossam Faris, Al-Zoubi Ala'M, Ali Asghar Heidari, Ibrahim Aljarah, Majdi Mafarja, Mohammad A Hassonah, and Hamido Fujita. An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks. *Information Fusion*, 48:67–83, 2019.

[10] Liguo Fei, Qi Zhang, and Yong Deng. Identifying influential nodes in complex networks based on the inverse-square law. *Physica A: Statistical Mechanics and its Applications*, 512:1044–1059, 2018.

[11] Lazaros K Gallos, Hernán A Makse, and Mariano Sigman. A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks. *Proceedings of the National Academy of Sciences*, 109(8):2825–2830, 2012.

[12] Lazaros K Gallos, Diego Rybski, Fredrik Liljeros, Shlomo Havlin, and Hernán A Makse. How people interact in evolving online affiliation networks. *Physical Review X*, 2(3):031014, 2012.

[13] Ya-Chun Gao, Chuan-Ji Fu, Shi-Min Cai, Chun Yang, and H Eugene Stanley. Repulsive synchronization in complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(5):053130, 2019.

[14] Marko Gosak, Rene Markovič, Jurij Dolenšek, Marjan Slak Rupnik, Marko Marhl, Andraž Stožer, and Matjaž Perc. Network science of biological systems at different scales: a review. *Physics of life reviews*, 24:118–135, 2018.

[15] Chungu Guo, Liangwei Yang, Xiao Chen, Duanbing Chen, Hui Gao, and Jing Ma. Influential nodes identification in complex networks via information entropy. *Entropy*, 22(2):242, 2020.

[16] Kun He, Yingru Li, Sucheta Soundarajan, and John E Hopcroft. Hidden community detection in social networks. *Information Sciences*, 425:92–106, 2018.

[17] Qiang He, Xingwei Wang, Fubing Mao, Jianhui Lv, Yuliang Cai, Min Huang, and Qingzheng Xu. Caom: A community-based approach to tackle opinion maximization for social networks. *Information Sciences*, 513:252–269, 2020.

[18] Dirk Helbing, Dirk Brockmann, Thomas Chadefaux, Karsten Donnay, Ulf Blanke, Olivia Woolley-Meza, Mehdi Moussaid, Anders Johansson, Jens Krause, Sebastian Schutte, et al. Saving human lives: What complexity science and information systems can contribute. *Journal of statistical physics*, 158(3):735–781, 2015.

[19] Adil Imad Eddine Hosni, Kan Li, and Sadique Ahmad. Analysis of the impact of online social networks addiction on the propagation of rumors. *Physica A: Statistical Mechanics and its Applications*, 542:123456, 2020.

[20] Adil Imad Eddine Hosni, Kan Li, and Sadique Ahmad. Minimizing rumor influence in multiplex online social networks based on human individual and social behaviors. *Information Sciences*, 512:1458–1480, 2020.

[21] Dongeun Kim, Dongmin Hyeon, Jinoh Oh, Wook-Shin Han, and Hwanjo Yu. Influence maximization based on reachability sketches in dynamic graphs. *Information Sciences*, 394:217–231, 2017.

[22] Seungkeol Kim, Dongeun Kim, Jinoh Oh, Jeong-Hyon Hwang, Wook-Shin Han, Wei Chen, and Hwanjo Yu. Scalable and parallelizable influence maximization with random walk ranking and rank merge pruning. *Information Sciences*, 415:171–189, 2017.

[23] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature physics*, 6(11):888–893, 2010.

[24] Zhe Li, Tao Ren, Xiaoqi Ma, Simiao Liu, Yixin Zhang, and Tao Zhou. Identifying influential spreaders by gravity model. *Scientific reports*, 9(1):1–7, 2019.

[25] Lien-Fa Lin and Yung-Ming Li. An efficient approach to identify social disseminators for timely information diffusion. *Information sciences*, 544:78–96.

[26] Linyuan Lü, Duanbing Chen, Xiao-Long Ren, Qian-Ming Zhang, Yi-Cheng Zhang, and Tao Zhou. Vital nodes identification in complex networks. *Physics Reports*, 650:1–63, 2016.

[27] Yue Ma, Zhulou Cao, and Xingqin Qi. Quasi-Laplacian centrality: A new vertex centrality measurement based on Quasi-Laplacian energy of networks. *Physica A: Statistical Mechanics and its Applications*, 527:121130, 2019.

[28] Soumen Majhi, Bidesh K Bera, Dibakar Ghosh, and Matjaž Perc. Chimera states in neuronal networks: A review. *Physics of life reviews*, 28:100–121, 2019.

[29] Naoki Masuda, Mason A Porter, and Renaud Lambiotte. Random walks and diffusion on networks. *Physics reports*, 716:1–58, 2017.

[30] Mark EJ Newman. A measure of betweenness centrality based on random walks. *Social networks*, 27(1):39–54, 2005.

[31] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[32] Sen Pei, Xian Teng, Jeffrey Shaman, Flaviano Morone, and Hernán A Makse. Efficient collective influence maximization in cascading processes with first-order transitions. *Scientific reports*, 7:45240, 2017.

[33] Sonja Pravilovic, Massimo Bilancia, Annalisa Appice, and Donato Malerba. Using multiple time series analysis for geosensor data forecasting. *Information Sciences*, 380:31–52, 2017.

[34] Zhuo-Ming Ren, An Zeng, Duan-Bing Chen, Hao Liao, and Jian-Guo Liu. Iterative resource allocation for ranking spreaders in complex networks. *EPL (Europhysics Letters)*, 106(4):48005, 2014.

[35] Kazumi Saito, Masahiro Kimura, Kouzou Ohara, and Hiroshi Motoda. Super mediator–a new centrality measure of node importance for information diffusion over social network. *Information Sciences*, 329:985–1000, 2016.

[36] Roosevelt Sardinha, Aline Paes, and Gerson Zaverucha. Revising the structure of bayesian network classifiers in the presence of missing data. *Information Sciences*, 439:108–124, 2018.

[37] Amir Sheikhahmadi, Mohammad Ali Nematbakhsh, and Ahmad Zareie. Identification of influential users by neighbors in online social networks. *Physica A: Statistical Mechanics and its Applications*, 486:517–534, 2017.

[38] Peng Gang Sun. Controllability and modularity of complex networks. *Information Sciences*, 325:20–32, 2015.

[39] Peng Gang Sun and Xiaoke Ma. Dominating communities for hierarchical control of complex networks. *Information Sciences*, 414:247–259, 2017.

[40] Peng Gang Sun, Yi Ning Quan, Qi Guang Miao, and Juan Chi. Identifying influential genes in protein–protein interaction networks. *Information Sciences*, 454:229–241, 2018.

[41] T. Wen and W. Jiang. Identifying influential nodes based on fuzzy local dimension in complex networks. *Chaos, Solitons & Fractals*, 119:332–342, 2019.

[42] Tao Wen and Yong Deng. Identification of influencers in complex networks by local information dimensionality. *Information Sciences*, 512:549–562, 2020.

[43] Tao Wu, Xingping Xian, Linfeng Zhong, Xi Xiong, and H Eugene Stanley. Power iteration ranking via hybrid diffusion for vital nodes identification. *Physica A: Statistical Mechanics and its Applications*, 506:802–815, 2018.

[44] Ying Xie, Xinmei Wang, Dan Jiang, and Rongbin Xu. High-performance community detection in social networks using a deep transitive autoencoder. *Information Sciences*, 493:75–90, 2019.

[45] Wenzheng Xu, Tong Li, Weifa Liang, Jeffrey Xu Yu, Ning Yang, and Shaobing Gao. Identifying structural hole spanners to maximally block information propagation. *Information Sciences*, 505:100–126, 2019.

[46] Ahmad Zareie and Amir Sheikhahmadi. A hierarchical approach for influential node ranking in complex social networks. *Expert Systems with Applications*, 93:200–211, 2018.

[47] Ahmad Zareie, Amir Sheikhahmadi, and Mahdi Jalili. Identification of influential users in social networks based on users' interest. *Information Sciences*, 493:217–231, 2019.

[48] Ahmad Zareie, Amir Sheikhahmadi, and Keyhan Khamforoosh. Influence maximization in social networks based on topsis. *Expert Systems with Applications*, 108:96–107, 2018.

[49] Zibin Zheng, Fanghua Ye, Rong-Hua Li, Guohui Ling, and Tan Jin. Finding weighted k-truss communities in large networks. *Information Sciences*, 417:344–360, 2017.

[50] Tian Zhu, Bai Wang, Bin Wu, and Chuanxi Zhu. Maximizing the spread of influence ranking in social networks. *Information Sciences*, 278:535–544, 2014.