Identifying influential nodes in complex networks: Effective distance gravity model

Qiuyan Shang^{a,b}, Yong Deng^{a,*}, Kang Hao Cheong^{c,d,1,*}

^aInstitute of Fundamental and Frontier Science, University of Electronic Science and Technology of China, Chengdu, 610054, China

^bYingcai Honors of school, University of Electronic Science and Technology of China, Chengdu, 610054, China

^cScience and Math Cluster, Singapore University of Technology and Design (SUTD), S487372, Singapore

 dSUTD -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 controlling, community finding, data mining, network system controlling, just to name a few. Many measures have thus been proposed to date, and these measures are either based on the locality of nodes or the global nature of the network. These measures typically use distance based on the concept of traditional Euclidean Distance, which only focus on the local static geographic distance between nodes but ignore the interaction between nodes in real networks. However, a variety of factors should be considered for the purpose of identifying influential nodes, such as degree, edge, direction and weight. Some methods based on evidence theory have also been proposed. In this paper, 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 the complex network, and also utilize the effective distance to replace the Euclidean Distance. This allows us to fully consider the complex

^{*}Corresponding author at: Institute of Fundamental and Frontier Science, University of Electronic Science and Technology of China, Chengdu, 610054, China. E-mail: dengen-tropy@uestc.edu.cn, prof.deng@hotmail.com. (Yong Deng)

Email address: kanghao_cheong@sutd.edu.sg (Kang Hao Cheong)

topological structure of the real-world network, as well as the dynamic interaction information between nodes. In order to validate the effectiveness of our proposed method, we have utilized the susceptible infected (SI) model to carry out a variety of simulations on eight different real-world networks using six existing well-known methods. The experimental results indicate the reasonableness and effectiveness of our proposed method.

Keywords: Complex networks, Influential nodes, Gravity model, Effective distance, SI model

1. Introduction

In recent years, the study of complex network[1, 2] has attracted immense attention[3]. Many real-world problems[4, 5] can be analyzed as part of network science[6, 7] for further research[8], such as Internet security, network control system[9, 10] and social network. Hence, the identification of influential nodes in complex networks play an important role[11] in both structural and functional aspects[12, 13], and is an important area of research[14]. The identification of influential nodes can be applied across various fields[15] such as disease[16], network system[17], biology[18], social system[19, 20, 4], time series[21], information propagation[22] and Parrondo's paradox[23, 24, 25, 26, 27]. Besides, identifying the vital nodes[28] can allow us to discover and address real-world problems[29, 30] such as transportation hubs identifying, influence maximizing, rumor controlling[31], disease controlling[32], advertising and community finding[33, 34].

Many methods have been proposed to assess the influence of nodes[35, 36]. These methods can be classified under two broad categories: locality of nodes and the global nature of the network. One view is that the influence of nodes often depends on its neighbors, such as degree centrality (DC)[14], K-shell decomposition method (KS)[32], semi-local centrality[14] and PageRank[37]. For DC, the influence of nodes is determined by the number of neighbors; a node with many neighbors is of high influence. The KS suggests that the influence of nodes is related to their topological properties in the local area. The more central the node is in the local structure, the more influential the node is. PageRank, the algorithm for random walking[38] states that the influence of nodes is not only dependent on the number of neighbors, but also related to the quality of neighbors[12]. It works by simulating the process of browsing the web. In general, it has good performance on directed network, but does not perform very well on undirected network.

Another view is that the influence of nodes mainly depends on paths in the network. For example, closeness centrality (CC)[39] and betweenness centrality (BC)[12] are representatives of such algorithm. CC suggests that the shorter the average distance between a node and other nodes, that is, the closer the node is located to the center, the more influential the node is. However, BC claims that the influence of a node is mainly determined by the number of shortest paths through it. Although BC and CC algorithms can often give better results than other algorithms, they are very sensitive to network structure[40] and the complexity of these algorithms is high with many limitations as well. The above algorithms are either neighborhood-based local method, or the path-based global method.

Recently, inspired by the law of gravity, Li et al. proposed an algorithm based on the gravity model, called gravity model (GM)[12]. Liu et al. further proposed a more generalized weighted gravity model, called generalized mechanics model (GMM)[41]. The two models take into account both the neighborhoodbased local information and the path-based global information. They are also applicable on both directed and undirected networks, and have proven to be effective and feasible. However, the distance in the algorithm based on the gravity model mainly utilizes the traditional Euclidean Distance[42], focusing only on the local static geographical location between nodes, while ignoring the interaction between nodes in the actual network. In order to address this critical gap, we propose an original and novel method called the effective distance gravity model. On the basis of GM, the original Euclidean Distance[43] is now replaced by the effective distance proposed by Brockmann et al. Effective distance is an abstract concept of distance derived from the idea of probability. It mainly pays attention to the interaction of nodes in the network and uses it as the main basis for judging. The core of effective distance is to discover the most probable path between two points by calculating the probability through the adjacency matrix. The effective distance fully considers the potential dynamic information interaction between nodes in the actual network. Therefore, our proposed effective distance gravity model takes into account not only the network local and global information, but also the potential dynamic interaction between nodes, such that a node with more neighbors and shorter effective distance between the other nodes is more influential. Based on our proposed method, we have carried out a variety of experiments on eight real networks using the susceptible infected (SI) model[44], and compared it with six 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 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 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[45]. And 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. The node with large degree is of high influence[46].

DC(i) of each node *i* can be obtained by the following formula.

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

Where k_i is the degree of node i/47.

Definition 2.2. The definition of Betweenness centrality(BC)[41] is as follows. BC measures the importance of a node by the number of shortest paths through it. The more the number of shortest paths through node *i*, the more important node *i* is in the network.

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

Definition 2.3. Closeness centrality(CC)[39] evaluates the influence of nodes by the reciprocal of the sum of shortest path between nodes. The higher the CC(i)is, the more important the node *i* is.

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

where d_{ij} denotes the length of shortest path between node *i* and node *j*.

Definition 2.4. Eigenvector centrality(EC)[41] is a complex method, which claims the influence of a node is determined not only by the number of neighbors, but also by the importance of the them. EC(i), the centrality scores of node *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 jth entry of the eigenvector corresponding to λ .

Definition 2.5. PageRank(PC) uses an iterative approach to obtain the influence of nodes, and it is very effective to calculate the importance of nodes in the directed network. PC(i) of node i[37] can be obtained by following formula.

$$PC(i)^{q} = \sum_{j=1}^{n} \left(a_{ij} \frac{PC(j)^{q-1}}{k_{j}}\right)$$
(5)

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

2.2. Gravity model (GM)

The GM is defined by the gravity formula. The influence of a node can be estimated by GM[12] as follows.

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

C(i) represents the centrality score of node *i*, the degree of the node *j* is denoted as k_j . The shortest path between node *i* and node *j* can be represented as d_{ij} . In particular, the distance here uses Euclidean Distance.

2.3. Effective distance (ED)

Effective distance [43] is an abstract distance based on probability, which mainly focuses on the potential information interaction between nodes in the real complex networks. The definition of ED is as follows.

$$D_{mn} = \min\{1 - \log_2(P_{mn}^*)\}\tag{7}$$

Where D_{mn} is the value of effective distance from node m to node n and P_{mn} is the probability from node m to node n, which can also be obtained by the product of multiple probabilities in the graph. For example P_{mn}^* can be calculated by $P_{mn}^* = P_{ml} \times P_{ls} \times ... \times P_{kn}$, which is similar to Markov Chain. The P_{mn} is calculated as follows.

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

Where k_m is the degree of node m, a_{mn} is the element in the adjacency matrix of graph G.

3. Proposed method

3.1. Effective distance gravity model (EffG)

In reality, many problems can be analyzed as part of network science for further research. The structure and function of these actual networks are often more complicated than we think. Thus, for the identification of influential nodes, only use neighborhood-based local properties of the network but ignore the global connectivity are unadvisable. Similarly, it is not feasible to only consider the path-based globality of the network but ignore the local properties of the node. Individually speaking, the two properties should be considered together in order to achieve a good effect. Therefore, we consider a comprehensive consideration of the degree of nodes and the path of the network.

At the same time, considering the complex structure and the evolution of the complex network, the structure of network is likely to be unstable. Therefore, there may be some errors for the node identification. However, multiple cumulative summing can effectively solve this problem.

In addition, since the distance of most of the measures is conventional Euclidean Distance, which is the static geographic distance between two points. However, considering the multiple complexity fusion of complex networks, we cannot simply think that the true structure of the network is the exhibited geometry structure currently. The potential interaction of information and energy between nodes may lead to changes in the network structure. Therefore, we believe that complex networks are likely to have a potential geometric structure, which often drives many dynamics propagation processes, such as the spread of epidemics and rumors. In this situation, it is obviously inadequate to use a simple static geometric metric such as Euclidean distance. Dirk Brockmann and Dirk Helbing claimed that if probability is used to construct a new distance metric to replace conventional geographic distance, then the complex space-time patterns can be reduced to surprisingly simple, homogeneous wave propagation patterns. Their experimental results showed that the effective distance can reliably predict the arrival time of the disease. Thus, we believe that the effective distance fully takes into account a potential topology of complex network due to the dynamic information interaction between nodes[43], which has certain significance for the identification of influential node in the complex network. Consequently, we consider replacing the conventional Euclidean Distance with the effective distance proposed by Brockmann to further optimize the algorithm.

In summary, we proposed an effective distance gravity model, which not only comprehensively considers the local and global network structural indicators, but also reduces the identification error caused by the unstable structure of the complex network through cumulative summation. At the same time, by replacing Euclidean Distance with effective distance, the dynamic interaction between nodes and the potential complex topology of the network are fully considered. Therefore, the influence of the node can be estimated as

$$C_{EffG}(i) = \sum_{i \neq j} \frac{k_i \times k_j}{D_{ij}^2} \tag{9}$$

Where k_i and k_j are the degree of node *i* and node *j*, D_{ij} is the effective distance from node *i* to node *j*. And $C_{EffG}(i)$ represents the centrality scores of node *i*. The whole steps and calculation process of proposed method are shown in Fig.1.

3.2. An example

In order to better explain our proposed identification method EffG, here a simple example is given to help understand how EffG works in the network. We take *node2* as an example to calculate the EffG scores of it. The Fig.2(a) is the



Figure 1: **The flow chart of our proposed method.** The first step is to build a network. The second step is to calculate the degree of each node. The third and fourth steps are combined to obtain the effective distance between the nodes. Finally, the fifth step calculates the centrality scores of each node.

graph of a network, which adjacency matrix is the Fig.2(b).



Figure 2: A simple network with seven nodes

The degree of each node is presented in the Table I.

Table 1: The degree of each node in Fig.2.

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

First, we calculate the effective distance between nodes by the Function (7).

Table 2: Effective Distance (ED) between nodes in Fig.2.

D_{ij}	D_{21}	D_{22}	D_{23}	D_{24}	D_{25}	D_{26}	D_{27}
ED	2.0000	$+\infty$	4.0000	4.0000	2.0000	4.5850	4.5850

In particularly, Under normal circumstances , the $P_{ij} \neq P_{ji}$ and $D_{ij} \neq D_{ji}$. Besides, because P_{ii} , the probability from node *i* to itself is often zero, the distance from node *i* to itself which denoted as D_{ii} is also infinite. The effective distance between *node*2 and *node*7 is calculated as:

$$D_{27} = \min \{1 - \log_2(P_{27}^*)\}$$

$$= 1 - \log_2(\max\{P_{27}^*\})$$

$$= 1 - \log_2(\max\{P_{27}, P_{21} \times P_{17}, P_{25} \times P_{51} \times P_{57}, ...\})$$

$$= 1 - \log_2(P_{21} \times P_{17})$$

$$= 1 - \log_2(\frac{1}{2} \times \frac{1}{6})$$

$$= 4.5850$$

$$D_{72} = \min \{1 - \log_2(P_{72}^*)\}$$

$$= 1 - \log_2(\max\{P_{72}^*\})$$

$$= 1 - \log_2(P_{71} \times P_{12})$$

$$= 1 - \log_2(\frac{1}{1} \times \frac{1}{6})$$

$$= 3.5850$$

$$(10)$$

Using the same method of D_{27} and D_{72} , the effective distance between other nodes and *node2* can be calculated. The result is in the Table II.

Then, the EffG scores of node 2 can be calculated as follows:

$$C_{EffG}(2) = \sum_{j \neq 2} \frac{k_2 \times k_j}{D_{2j}^2} = 5.9104$$

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

$$C_{EffG}(1) = \sum_{j \neq 1} \frac{k_1 \times k_j}{D_{1j}^2} = 6.5358$$

$$C_{EffG}(3) = \sum_{j \neq 3} \frac{k_3 \times k_j}{D_{3j}^2} = 5.9104$$

$$C_{EffG}(4) = \sum_{j \neq 4} \frac{k_4 \times k_j}{D_{4j}^2} = 6.0704$$

$$C_{EffG}(5) = \sum_{j \neq 5} \frac{k_5 \times k_j}{D_{5j}^2} = 6.2865$$

$$C_{EffG}(6) = \sum_{j \neq 6} \frac{k_6 \times k_j}{D_{6j}^2} = 5.5981$$

$$C_{EffG}(7) = \sum_{j \neq 7} \frac{k_7 \times k_j}{D_{7j}^2} = 1.0115$$

From the Fig2.(a). we can see that the *node*1 is the central of the network which has the most strong connection with others and covers the most shortest paths in the network. Without *node*1, the network will be broken into multiple isolated parts. Thus, it is reasonable that the *node*1 is the most influential node in this network. Besides, the *node*7 can be seen that is the least influential in the network, and the EffG score that matches it is the lowest. And the importance of *node*2 and *node*3 in this network is basically 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, six experiments were performed on eight actual networks, in comparison with six existing well-known methods.

4.1. Datasets

The experiment was conducted on eight real-world networks, Jazz, NS[48], GrQc, Email, EEC, Facebook[41], PB[49] and USAir[12], including two communication networks (Email, EEC), one transportation network (USAir), two social networks (Facebook, PB) and three cooperative networks (Jazz, NS, GrQc). Among them, Email is a network where users send emails and communicate with each other. EEC is a network where European researchers do exchange of emails. USAir represents a US air transportation network. 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. Other relevant information about the network is displayed on Table III.

Table 3: The basic topology information of the eight actual 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. C and r are the network's clustering coefficient[50] and assortative coefficient[51].

Networks	n	m	< k >	< d >	С	r
Jazz	198	2472	27.6970	2.2350	0.6334	0.0202
\mathbf{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

4.2. Centrality scores of nodes

In this experiment, our proposed method (EffG)was used to calculate centrality scores in six real-world networks we provided. Five existing well-known methods (DC, CC, BC, PC, Gravity) were used in the same networks for comparison. The experimental results are shown in Figs.3-8. The importance of the node is reflected by the color of the node in the heat map. The darker the color of node is, the more influential the node is.

As can be seen, the distribution of relative importance of nodes is basically consistent, although the centrality scores calculated by CC and EffG are higher and the value of BC is lower. Besides, it can be easily found in figs.4-8. that the centrality scores calculated by BC, DC and PC are difficult to distinguish in the figures because the scores are all quite low. However, the centrality scores calculated by CC, Gravity model and EffG are well distinguished, especially CC and EffG. Moreover, the distribution of relative importance of nodes in CC and EffG is more similar. Consequently, our proposed method, EffG, can be found to be convenient for us to distinguish the relative importance of nodes with high accuracy.



Figure 3: This figure compares the centrality scores of different measures in Jazz. The distribution of Gravity model, CC and our proposed method is similar.



Figure 4: This figure compares the centrality scores of different measures in NS. DC, Gravity, BC and PC are basically the same. Our method and CC are similar in the distribution of cital node.



Figure 5: This figure compares the centrality scores of different measures in EEC. The value of proposed method and CC is higher, and BC is difficult to distinguish the different nodes.



Figure 6: This figure compares the centrality scores of different measures in Email. DC and PC is almost the same, and the distribution of CC and proposed method is similar although the value of our method is higher.



Figure 7: This figure compares the centrality scores of different measures in PB. The value of CC is highest, the value of BC is lowest. The distribution of our method is little different from the others except for CC.



Figure 8: This figure compares the centrality scores of different measures in USAir. DC and PC is basically the same, and CC can clearly be seen the difference in global influence of nodes.

4.3. Evaluating with susceptible infected (SI) model

The susceptible infected (SI) model[44] 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 compartments 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 N. The experimental simulation time of the susceptible infected model is denoted by t. β represents the probability of nodes infection. N is the number of experiments. The average number of infected nodes at time t, denoted by F(t). It can be easily understood that the more important the node has the greater the influence. And under the condition that the infection time t and the infection probability B 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 vital nodes, the SI model was applied on eight 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. In particular, 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.

The experimental results are shown in Figs.9-16. The node with more final infected nodes is of greater importance. Hence, the faster the curve rises and the higher the curve is, the more influential the nodes in the initial infection set are. That is to say, the more effective the identification method is.

It can be seen that the curves corresponding to our proposed method 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 has a stronger infection



Figure 9: The figure compares the infection ability of the top-100 nodes selected by different methods in Jazz. All methods except for BC are basically the same and our method is the highest and the curve rises faster than others.



Figure 10: The figure compares the infection ability of the top-100 nodes selected by different methods in EEC. All methods performance is similar while BC curve is sightly lower than other curves.



Figure 11: The figure compares the infection ability of the top-100 nodes selected by different methods in Email. The performance of our method is better than others slightly.



Figure 12: The figure compares the infection ability of the top-100 nodes selected by different methods in GrQc. All methods performed different significantly. The top-100 nodes of CC and EC are the most influential, our method is the second.



Figure 13: The figure compares the infection ability of the top-100 nodes selected by different methods in NS. Our method's curve rises fastest but PC's curve rises slowly.



Figure 14: The figure compares the infection ability of the top-100 nodes selected by different methods in PB. The difference among these methods are not obvious, which means they are basically consistent.



Figure 15: The figure compares the infection ability of the top-100 nodes selected by different methods in Facebook. The trends of our method and DC are consistent. The CC curve rises fastest and BC rises slowly.



Figure 16: The figure compares the infection ability of the top-100 nodes selected by different methods in USAir. Our method, EC, CC and DC performed similarly. The BC and PC rise more slowly than other methods.

ability. That is to say, CC and our proposed method EffG can select influential nodes more accurately.

4.4. Comparison ranking results

The top-20 vital nodes in the Jazz network ranked by different methods, our proposed method (EffG), DC, BC, CC, PC, EC, Gravity and SI model in which the β =0.2,t=20,N=50, are listed in Table IV. The number of overlapping nodes in the set of the top-20 nodes sorted by our proposed method and the top-20 nodes set sorted by other methods are shown in Table V. The number of coincident nodes demonstrates the effectiveness of our method to a certain extent. As can be seen, in the Jazz network, the number of nodes that are consistent with the top-20 nodes obtained by our method and the top-20 nodes obtained by other methods are high. The high number of coincidences with other measures confirms the justifiability of our method and the unity with other methods.

4.5. Relation of proposed method with other centrality methods

The Kendall coefficient, Kendall Tau[52], 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, 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 is, the more valid the identification method is.

Given two sequences with N elements, $X = (x_1, x_2, x_3, \ldots, x_n)$ and $Y = (y_1, y_2, y_3, \ldots, y_n)$. 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. While if both $x_i > x_j$

our method	DC	$\mathbf{C}\mathbf{C}$	BC	\mathbf{PC}	EC	SI	Gravity
8	8	8	8	8	100	8	28
100	100	100	155	100	4	100	186
4	4	131	100	131	8	4	136
131	131	194	186	4	131	194	175
194	194	69	131	186	80	53	98
80	80	4	136	136	129	131	158
69	69	53	60	69	194	111	113
162	162	111	28	28	69	133	33
53	77	162	69	175	53	162	23
5	5	129	175	155	32	67	9
77	53	59	194	162	84	5	87
59	32	67	9	129	85	59	86
32	59	80	32	80	130	80	4
67	186	186	111	59	162	129	77
133	67	133	4	77	77	32	131
111	133	77	59	53	133	115	38
84	28	5	79	32	115	77	178
85	111	55	113	5	89	151	72
9	9	79	151	113	59	28	16

Table 4: Top-20 ranking of influential nodes in Jazz network by our proposed method (EffG), DC, BC, CC, PC, EC, Gravity and SI model.

Table 5: The number of same nodes between other methods and EffG.

DC	CC	BC	PC	EC	SI	Gravity
18	17	10	14	17	17	4

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's Tau of two sequences X and Y, is defined as follows.

$$tau = \frac{n_{+} - n_{-}}{N \times (N - 1)} \tag{10}$$

Where n_+ and n_- are the number of concordant sequence pairs and discordant sequence pairs respectively, N is the total number of sequence pairs.

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'sTau* of SI model sequence and other method's sequence was calculated. In the experiment, the infection probability β changed from 0.2 to 1.6, and the SI model was independently work 50 times with the infection time t = 5 to take the average on different networks. The experimental results are shown in Figs.17-22, where *tau* represents the value of *Kendall'sTau*. Higher tau value indicates stronger positive correlation between centrality method and SI model.

As can be seen that CC has the strongest correlation with SI and CC is gradually close to SI as β increases, which means that it has a strong positive correlation with the SI model. Our proposed method performs well in general although the correlation is a little weak on some networks. For instance, the *tau* of our method is the second highest when $\beta > 0.8$ in Figs.20-21.

4.6. 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 Jazz, 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. 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 = 20. Finally, the correlation between the node ranking and the final number of nodes infected by them, denoted as $\langle N \rangle$, will be established. The results



Figure 17: The figure compares the tau of SI model sequence and other method's sequence in Jazz. CC has the strongest correlation with SI and CC is gradually close to SI as β increases.



Figure 18: The figure compares the tau of SI model sequence and other method's sequence in NS. The tau of BC is the highest while the tau of our method is the lowest when $\beta > 0.24$.



Figure 19: The figure compares the tau of SI model sequence and other method's sequence in GrQc. The tau of our method is in the middle basically, and as β increases, all methods have large fluctuations in tau value.



Figure 20: The figure compares the tau of SI model sequence and other method's sequence in PB. The tau of our method is highest when $\beta > 0.8$ while when $\beta < 0.6$ it is lowest. And tau of DC is lowest when $\beta > 0.9$.



Figure 21: The figure compares the tau of SI model sequence and other method's sequence in Facebook. The tau of our method, PC and EC are the lowest when $\beta = 0.6$. As β changes, CC and tau are gradually close to the SI and the tau of our method is the second highest when $\beta > 0.9$.



Figure 22: The figure compares the tau of SI model sequence and other method's sequence in USAir. The tau of our method and BC are lower.

are shown in Figs.23-26.

The node with higher ranking is of stronger capability to infect other nodes, which means the node is more influential. That is to say, the higher the ranking of a node is, the greater the number of nodes eventually infected by it should be. The lower the node 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 seen in Figs.23-26, the curve corresponding to our method is continuously declining, and has little fluctuation compared with other methods. However, it can be easily found 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.



Figure 23: The figure describes the correlation between different methods and SI model in Jazz. As the ranking of nodes changes, the N value of BC fluctuates greatly. And our method continues to decline and is most stable

5. Conclusion

In this paper, an original and novel method for identifying the influence node is proposed: an effective distance gravity model. Instead of considering



Figure 24: The figure describes the correlation between different methods and SI model in NS. The curves of CC, BC, EC and DC fluctuates greatly. While our method continues to decline and is stable.



Figure 25: The figure describes the correlation between different methods and SI model in Email. Overall trend of these curves is declining. And the curve of BC fluctuates the most greatly. While our method and CC are more stable.



Figure 26: The figure describes the correlation between different methods and SI model in USAir. Overall trend of these curves is declining. The curve of BC fluctuates greatly. While our method and EC are more stable than the others

single-dimensional factors, our proposed EffG 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 uses the effective distance to replace traditional static Euclidean Distance. EffG is able to take full advantage of the dynamic information exchange between nodes in the real-world network. In addition, EffG can help us unravel the 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 were conducted on eight real-world networks and compared with six existing well-known methods. The experimental results indicated 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.

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 work is partially supported by National Natural Science Foundation of China (Grant Nos. 61573290, 61503237).

References

- Y. Xie, X. Wang, D. Jiang, R. Xu, High-performance community detection in social networks using a deep transitive autoencoder, Information Sciences 493 (2019) 75–90.
- [2] R. De Souza, D. R. Figueiredo, A. d. A. Rocha, A. Ziviani, Efficient network seeding under variable node cost and limited budget for social networks, Information Sciences 514 (2020) 369–384.
- [3] Y.-C. Gao, C.-J. Fu, S.-M. Cai, C. Yang, H. Eugene Stanley, Repulsive synchronization in complex networks, Chaos: An Interdisciplinary Journal of Nonlinear Science 29 (5) (2019) 053130.
- [4] A. Zareie, A. Sheikhahmadi, M. Jalili, Identification of influential users in social networks based on users' interest, Information Sciences 493 (2019) 217–231.
- [5] B. Wei, F. Xiao, Y. Shi, Synchronization in kuramoto oscillator networks with sampled-data updating law, IEEE Transactions on Cybernetics (2019) 1–9doi:10.1109/TCYB.2019.2940987.

- [6] M. Gosak, R. Markovič, J. Dolenšek, M. S. Rupnik, M. Marhl, A. Stožer, M. Perc, Network science of biological systems at different scales: a review, Physics of life reviews 24 (2018) 118–135.
- [7] L. K. Gallos, H. A. Makse, M. 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) (2012) 2825–2830.
- [8] Z. Wang, M. Jusup, L. Shi, J. H. Lee, Y. Iwasa, S. Boccaletti, Exploiting a cognitive bias promotes cooperation in social dilemma experiments, Nature Communications 9 (2018) 7. doi:10.1038/s41467-018-05259-5.
 URL <GotoISI>://WOS:000439970700010
- [9] T. Zhu, B. Wang, B. Wu, C. Zhu, Maximizing the spread of influence ranking in social networks, Information Sciences 278 (2014) 535–544.
- [10] B. Wei, F. Xiao, Y. Shi, Fully distributed synchronization of dynamic networked systems with adaptive nonlinear couplings, IEEE Transactions on Cybernetics (2019) 1–9doi:10.1109/TCYB.2019.2944971.
- [11] P. G. Sun, Y. N. Quan, Q. G. Miao, J. Chi, Identifying influential genes in protein–protein interaction networks, Information Sciences 454 (2018) 229–241.
- [12] Z. Li, T. Ren, X. Ma, S. Liu, Y. Zhang, T. Zhou, Identifying influential spreaders by gravity model, Scientific reports 9 (1) (2019) 1–7.
- [13] T. Wen, W. Jiang, An information dimension of weighted complex networks, Physica a-Statistical Mechanics and Its Applications 501 (2018) 388-399. doi:10.1016/j.physa.2018.02.067.
- [14] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, T. Zhou, Identifying influential nodes in complex networks, Physica a: Statistical mechanics and its applications 391 (4) (2012) 1777–1787.

- [15] M. Li, Q. Zhang, Y. Deng, Evidential identification of influential nodes in network of networks, Chaos, Solitons & Fractals 117 (2018) 283–296.
- [16] Z. Wang, C. T. Bauch, S. Bhattacharyya, A. d'Onofrio, P. Manfredi, M. Perc, N. Perra, M. Salathe, D. W. Zhao, Statistical physics of vaccination, Physics Reports-Review Section of Physics Letters 664 (2016) 1–113. doi:10.1016/j.physrep.2016.10.006. URL <GotoISI>://WOS:000390637400001
- [17] H. Faris, A.-Z. Ala'M, A. A. Heidari, I. Aljarah, M. Mafarja, M. A. Hassonah, H. Fujita, An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks, Information Fusion 48 (2019) 67–83.
- [18] S. Majhi, B. K. Bera, D. Ghosh, M. Perc, Chimera states in neuronal networks: A review, Physics of life reviews 28 (2019) 100–121.
- [19] L. K. Gallos, D. Rybski, F. Liljeros, S. Havlin, H. A. Makse, How people interact in evolving online affiliation networks, Physical Review X 2 (3) (2012) 031014.
- [20] A. Zareie, A. Sheikhahmadi, K. Khamforoosh, Influence maximization in social networks based on topsis, Expert Systems with Applications 108 (2018) 96–107.
- [21] S. Pravilovic, M. Bilancia, A. Appice, D. Malerba, Using multiple time series analysis for geosensor data forecasting, Information Sciences 380 (2017) 31–52.
- [22] W. Xu, T. Li, W. Liang, J. X. Yu, N. Yang, S. Gao, Identifying structural hole spanners to maximally block information propagation, Information Sciences 505 (2019) 100–126.
- [23] Z. X. Tan, K. H. Cheong, Nomadic-colonial life strategies enable paradoxical survival and growth despite habitat destruction, Elife 6 (2017) e21673.

- [24] Z.-X. Tan, J. M. Koh, E. V. Koonin, K. H. Cheong, Predator dormancy is a stable adaptive strategy due to parrondo's paradox, Advanced Science (2019) 1901559.
- [25] K. H. Cheong, Z. X. Tan, Y. H. Ling, A time-based switching scheme for nomadic-colonial alternation under noisy conditions, Communications in Nonlinear Science and Numerical Simulation 60 (2018) 107–114.
- [26] K. H. Cheong, Z. X. Tan, N.-g. Xie, M. C. Jones, A paradoxical evolutionary mechanism in stochastically switching environments, Scientific reports 6 (2016) 34889.
- [27] K. H. Cheong, J. M. Koh, M. C. Jones, Multicellular survival as a consequence of parrondo's paradox, Proceedings of the National Academy of Sciences 115 (23) (2018) E5258–E5259.
- [28] A. Zareie, A. Sheikhahmadi, A hierarchical approach for influential node ranking in complex social networks, Expert Systems with Applications 93 (2018) 200-211. doi:10.1016/j.eswa.2017.10.018. URL <GotoISI>://WOS:000416498300016
- [29] D. Helbing, D. Brockmann, T. Chadefaux, K. Donnay, U. Blanke, O. Woolley-Meza, M. Moussaid, A. Johansson, J. Krause, S. Schutte, et al., Saving human lives: What complexity science and information systems can contribute, Journal of statistical physics 158 (3) (2015) 735–781.
- [30] S. Pei, X. Teng, J. Shaman, F. Morone, H. A. Makse, Efficient collective influence maximization in cascading processes with first-order transitions, Scientific reports 7 (2017) 45240.
- [31] A. I. E. Hosni, K. Li, S. Ahmad, Minimizing rumor influence in multiplex online social networks based on human individual and social behaviors, Information Sciences 512 (2020) 1458–1480.

- [32] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley,
 H. A. Makse, Identification of influential spreaders in complex networks, Nature physics 6 (11) (2010) 888–893.
- [33] Z. Zheng, F. Ye, R.-H. Li, G. Ling, T. Jin, Finding weighted k-truss communities in large networks, Information Sciences 417 (2017) 344–360.
- [34] K. He, Y. Li, S. Soundarajan, J. E. Hopcroft, Hidden community detection in social networks, Information Sciences 425 (2018) 92–106.
- [35] T. Wu, X. Xian, L. Zhong, X. Xiong, H. E. Stanley, Power iteration ranking via hybrid diffusion for vital nodes identification, Physica A: Statistical Mechanics and its Applications 506 (2018) 802–815.
- [36] K. Saito, M. Kimura, K. Ohara, H. Motoda, Super mediator-a new centrality measure of node importance for information diffusion over social network, Information Sciences 329 (2016) 985–1000.
- [37] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, T. Zhou, Vital nodes identification in complex networks, Physics Reports 650 (2016) 1–63.
- [38] N. Masuda, M. A. Porter, R. Lambiotte, Random walks and diffusion on networks, Physics reports 716 (2017) 1–58.
- [39] Y. Wang, S. Wang, Y. Deng, A modified efficiency centrality to identify influential nodes in weighted networks, Pramana 92 (4) (2019) 68.
- [40] T. Wen, D. Pelusi, Y. Deng, Vital spreaders identification in complex networks with multi-local dimension, Knowledge-Based Systems 195 (2020) 105717. doi:10.1016/j.knosys.2020.105717.
- [41] L. Fan, Z. Wang, Y. Deng, GMM: A Generalized Mechanics Model for Identifying the Importance of Nodes in Complex Networks, Knowledge-Based Systems (2020) 10.1016/j.knosys.2019.105464.
- [42] P.-E. Danielsson, Euclidean distance mapping, Computer Graphics and image processing 14 (3) (1980) 227–248.

- [43] D. Brockmann, D. Helbing, The hidden geometry of complex, networkdriven contagion phenomena, science 342 (6164) (2013) 1337–1342.
- [44] X. Meng, Y. Yang, S. Zhao, Adaptive evolution of virulence-related traits in a susceptible-infected model with treatment, in: Abstract and Applied Analysis, Vol. 2014, Hindawi, 2014.
- [45] N.-T. Le, B. Vo, L. B. Nguyen, H. Fujita, B. Le, Mining weighted subgraphs in a single large graph, Information Sciences 514 (2020) 149–165.
- [46] Y. Li, W. Cai, Y. Li, X. Du, Key node ranking in complex networks: A novel entropy and mutual information-based approach, Entropy 22 (1) (2020) 52.
- [47] Z. Ding, X. Chen, Y. Dong, F. Herrera, Consensus reaching in social network degroot model: The roles of the self-confidence and node degree, Information Sciences 486 (2019) 62–72.
- [48] T. Wen, W. Jiang, Identifying influential nodes based on fuzzy local dimension in complex networks, Chaos, Solitons & Fractals 119 (2019) 332–342. doi:10.1016/j.chaos.2019.01.011.
- [49] T. Wen, Y. Deng, The vulnerability of communities in complex networks: An entropy approach, Reliability Engineering & System Safety 196 (2020) 106782. doi:10.1016/j.ress.2019.106782.
- [50] T. Wen, M. X. Song, W. Jiang, Evaluating topological vulnerability based on fuzzy fractal dimension, International Journal of Fuzzy Systems 20 (6) (2018) 1956–1967. doi:10.1007/s40815-018-0457-8.
- [51] T. Wen, W. Jiang, Measuring the complexity of complex network by tsallis entropy, Physica a-Statistical Mechanics and Its Applications (2019). doi: Accept.
- [52] R. Criado, E. García, F. Pedroche, M. Romance, A new method for comparing rankings through complex networks: Model and analysis of com-

petitiveness of major european soccer leagues, Chaos: An Interdisciplinary Journal of Nonlinear Science 23 (4) (2013) 043114.