

Article

Identifying Influential Nodes in Complex Networks Based on Node Itself and Neighbor Layer Information

Jingcheng Zhu and Lunwen Wang *

College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China; zhujingcheng20@nudt.edu.cn

* Correspondence: wanglunwenmust@nudt.edu.cn

Abstract: Identifying influential nodes in complex networks is of great significance for clearly understanding network structure and maintaining network stability. Researchers have proposed many classical methods to evaluate the propagation impact of nodes, but there is still some room for improvement in the identification accuracy. Degree centrality is widely used because of its simplicity and convenience, but it has certain limitations. We divide the nodes into neighbor layers according to the distance between the surrounding nodes and the measured node. Considering that the node's neighbor layer information directly affects the identification result, we propose a new node influence identification method by combining degree centrality information about itself and neighbor layer nodes. This method first superimposes the degree centrality of the node itself with neighbor layer nodes to quantify the effect of neighbor nodes, and then takes the nearest neighborhood several times to characterize node influence. In order to evaluate the efficiency of the proposed method, the susceptible–infected–recovered (SIR) model was used to simulate the propagation process of nodes on multiple real networks. These networks are unweighted and undirected networks, and the adjacency matrix of these networks is symmetric. Comparing the calculation results of each method with the results obtained by SIR model, the experimental results show that the proposed method is more effective in determining the node influence than seven other identification methods.

Keywords: complex network; node influence; neighbor layer information; SIR model



Citation: Zhu, J.; Wang, L. Identifying Influential Nodes in Complex Networks Based on Node Itself and Neighbor Layer Information. *Symmetry* **2021**, *13*, 1570. <https://doi.org/10.3390/sym13091570>

Academic Editor: Jan Awrejcewicz

Received: 9 August 2021

Accepted: 16 August 2021

Published: 26 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The research direction of node influence in complex networks has attracted wide attention in recent years. The main research focus is to rank the influence of nodes to evaluate their importance. In fact, many systems in real life can be considered as complex network systems [1], such as animal networks [2], traffic networks [3], social networks [4], and economic networks. For example, identifying critical nodes in social networks and monitoring them can prevent the mass spread of infectious diseases. Identifying effective drug targets in cellular network systems allows for faster drug action in the body [5]. Numerous studies [6–13] have shown the significant theoretical and practical importance of identifying influential nodes that have a role in network structure and function. In the network, there are often a small number of nodes that support the entire network architecture, and the accurate identification of these nodes helps to maintain the stability of the network and ensure the normal operation of network functions.

Currently, researchers have proposed many classical methods for identifying influential nodes, including degree centrality [6], betweenness centrality [7], closeness centrality [8], eigenvector centrality [9], bridging centrality [10], LeaderRank [11], k -shell decomposition [12], and H-index [13]. Among these methods, the most widely used are degree centrality and k -shell decomposition. Degree centrality is simpler and more intuitive because it can be expressed directly in terms of the number of connected nodes, and it can be easily applied to large-scale networks. However, degree centrality and k -shell decomposition also have certain limitations. Degree centrality considers the most local information

of nodes, while k -shell decomposition divides many nodes into the same layer; both easily cause the node sorting to be too coarse-grained. Researchers have made many efforts to improve the problem of coarse-graining. Joonhyun et al. [14] proposed an improved coreness method that introduces information on neighborhoods to define node influence. Ahmed et al. [15] proposed a method to identify the influence of nodes according to the circular area density formula by taking the nodes' own degree centrality and the shortest path distance between two nodes as mass and radius, respectively. Li et al. [16] used the gravity model to propose a method that uses neighborhood information and path information to measure the influence of nodes in the spreading process. Li et al. [17] considered the role of the clustering coefficient, and they combined the clustering coefficient and the sum of the degree centrality of the nearest neighbor node to quantify node influence. Sheng et al. [18] combined the global and local structural characteristics of the network, global information reflecting the proximity to other nodes in the network, and local information as the contribution value of the nearest neighbor node to the measured node. Yang et al. [19] first proposed an improved k -shell decomposition method based on the k -shell value and the number of iterations of node removal in the k -shell decomposition, and then combined the improved method with degree centrality and the shortest path length to characterize the node influence. Zareie et al. [20] proposed an improved clustering ranking approach, which takes the common hierarchical structure of nodes and their neighborhood set into account. Yan et al. [21] propose, a new method that considers the local topological characteristics of nodes, center position of nodes, and effect of neighbor nodes.

The above methods are from different perspectives to solve the problem of insufficient differentiation in the process of node identification, but there is still room for improvement in recognition accuracy. In order to improve the accuracy of identification, we propose a new method of identifying node influence by combining information on the node itself and neighbor layers, which first superimposes the degree centrality of the node itself and neighborhood nodes within a certain range, and then considers the contribution of multiple fetches of nearest neighbor information to the accuracy of the results. To verify the effectiveness of the proposed method, this paper uses the SIR model [22,23] to perform 1000 independent simulations on multiple real networks to obtain the information dissemination ability of nodes. The proposed method in this paper is compared with degree centrality, betweenness centrality, closeness centrality, area density centrality [15], GM method [16], CLD method [17], and GLI method [19] in terms of discrimination and accuracy. The experimental results show that the proposed method in this paper can distinguish the influence among nodes and improve the recognition accuracy.

This section clarified the background of the study and the current research progress in the field. The rest of the paper is framed as follows: Section 2 explains the definition of comparison methods and the detailed idea of the proposed method based on information on the node itself and neighbor layers. The datasets used in this paper and the evaluation criteria for the experiments are given in Section 3. In Section 4, we present verification experiments in terms of the discrimination and accuracy comparing the proposed method with others. The final summary of this paper is given in Section 5.

2. Materials and Methods

An unweighted and undirected network is represented by $G = (V, E)$, where $V = \{v_1, v_2, \dots, v_n\}$ represents the set of n nodes, and $E = \{e_1, e_2, \dots, e_m\}$ represents the set of m edges. We can use adjacency matrix $A = (a_{ij})_{n \times n}$ to represent the structural characteristics of a complex network, where the adjacency matrix is symmetric. The element a_{ij} in the matrix can represent the edge information between any two nodes. If $a_{ij} = 1$, it means that there is a connection relationship between two nodes; otherwise, there is no connection relationship. Hereafter, k_i represents the degree centrality of node i .

2.1. Node Influence Based on Node and Neighbor Layer Information

Degree centrality represents the most local information of nodes in the network. It only reflects the importance of nodes according to the nodes themselves and ignores the influence of the surrounding nodes. Numerous studies have shown that the neighbor layer information of a node has an indispensable effect on node influence. During the spreading process, the node will certainly spread information to neighbor nodes through their connection in sequence until it cannot be transmitted. It is destined that the surrounding environment of the node will directly affect the spreading ability of the node. In order to explore the role of neighbor layer information on node spreading, this paper proposes a node influence identification method based on the node itself and neighbor layer information, called the NINL method. The main idea of this method is to consider the local information within a certain range of nodes. First, a radius is defined according to the average path length of the network, and the main consideration is the influence of the surrounding environment caused by the set of nodes within this radius. Then, the information of the nodes themselves is combined with the information of nodes within this range, and the initial local information is defined as

$$NINL_0(i) = k_i + \sum_j k_j, j \in \Gamma_i^{\lceil L \rceil}, \quad (1)$$

where Γ is used to represent nodes within a certain distance from node i , L refers to the average path length of the network, and $\lceil x \rceil$ refers to the smallest integer not smaller than x . To reflect the influence of nearest neighbor information more comprehensively, Equation (1) is extended to define the influence of nodes by using recursion, and the recursive formula is shown below.

$$\begin{aligned} NINL_1(i) &= \sum_{j \in \Gamma_i} NINL_0(j) \\ NINL_2(i) &= \sum_{j \in \Gamma_i} NINL_1(j), \\ &\dots\dots\dots \\ NINL_p(i) &= \sum_{j \in \Gamma_i} NINL_{p-1}(j) \end{aligned} \quad (2)$$

where p refers to the number of iterations, and Γ_i represents the nearest neighborhood nodes of node i . We can speculate the p -value according to some networks. In this paper, we chose $p = 3$, for reasons discussed in the next section. We use an example network below to illustrate the calculation process in detail.

Taking the example network of 13 nodes in Figure 1, the average path length of the network is calculated as $L = \text{sum}(d_{ij}) / (n \times (n-1)) = 358 / (13 \times (13-1)) \approx 2.2949$. Thus, we can get $\lceil L \rceil = 3$. The following takes node 1 as an example to give the calculation process of $NINL_0(1)$. First, we need to calculate degree centrality of each node, as shown below.

$$k_1 = 1, k_2 = 1, k_3 = 4, k_4 = 6, k_5 = 4, k_6 = 3, k_7 = 1, k_8 = 4, k_9 = 5, k_{10} = 2, k_{11} = 2, k_{12} = 4, k_{13} = 1.$$

Then, according to $\text{ceil}(L) = 3$, we can get that the first-level neighbor node set of node 1 is $\{3\}$, the second-level neighbor node set is $\{4, 5, 6\}$, and the third-level neighbor node set is $\{2, 7, 8, 9\}$. To sum up, the set of neighbor nodes within the three layers is $\{2, 3, 4, 5, 6, 7, 8, 9\}$; thus, $NINL_0(1) = k_1 + (k_2 + k_3 + k_4 + k_5 + k_6 + k_7 + k_8 + k_9) = 1 + (1 + 4 + 6 + 4 + 3 + 1 + 4 + 5) = 1 + 28 = 29$. Table 1 shows the calculation results of the node influence of $NINL_0 - NINL_3$ of the example network.

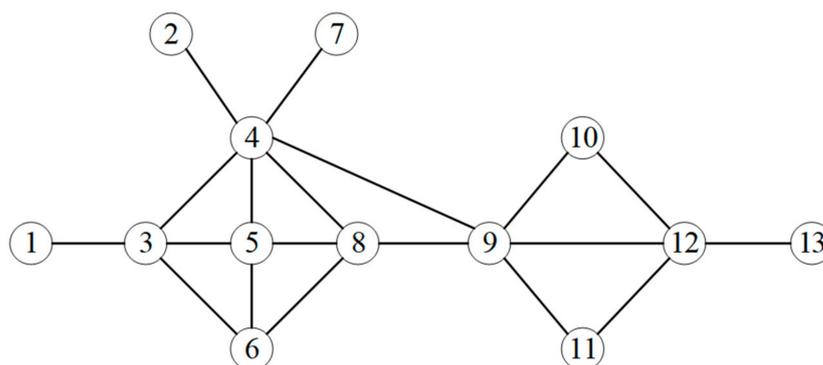


Figure 1. Example network.

Table 1. Node influence calculation results of the example network.

Node	1	2	3	4	5	6	7	8	9	10	11	12	13
$NINL_0$	29	37	37	38	37	37	37	38	38	37	37	37	24
$NINL_1$	37	38	141	224	150	112	38	150	187	75	75	136	37
$NINL_2$	141	224	523	704	627	441	224	673	660	323	323	374	136
$NINL_3$	523	704	1913	2931	2341	1823	704	2432	2397	1034	1034	1442	374

From the table, we can see that $NINL_0$ defines nodes 2, 3, 5, 6, 7, 10, 11, and 12 as the same value, and nodes 4, 8, and 9 as the same value. After getting the nearest neighbor node information for the first time, $NINL_1$ defines nodes 1 and 13 as the same value, nodes 2 and 7 as the same value, nodes 5 and 8 as the same value, and nodes 10 and 11 as the same value. $NINL_2$ also distinguishes the influence of nodes 1 and 13. The influence of nodes 5 and 8 was also identified. As for nodes 2 and 7, they could not be distinguished because they had the same information about themselves and their neighbors as nodes 10 and 11. At the same time, we can observe that nodes 2 and 7 were locally symmetrical, just like nodes 10 and 11. The above phenomenon shows that the proposed $NINL$ method can effectively distinguish the influence of the nodes in the sample network.

The specific computational flow of Algorithm 1 is shown below.

Algorithm 1. The $NINL_p$ Method

Input: the network $G = (V, E)$
 Output: node influence of $NINL_p$ centrality
 1: for $i = 1$ to $|V|$
 2: for $j = 1$ to $|V|$
 3: calculate the shortest path length between node i and node j
 4: end for
 5: end for
 6: calculate average path length L
 7: for $i = 1$ to $|V|$
 8: calculate the Degree centrality of node i
 9: end for
 10: for $i = 1$ to $|V|$
 11: find the neighbor nodes with a radius of $ceil(L)$ from the node i
 12: calculate $NINL_0$ of node i according to Equation (1)
 13: end for
 14: for $i = 1$ to $|V|$
 15: find the nearest neighbor nodes of node i
 16: end for
 17: set the value of p
 18: Recursively calculate $NINL_p$ centrality according to Equation (2)
 19: return $NINL_p$ centrality

2.2. Benchmark Methods

2.2.1. Degree Centrality

Degree centrality [6] in complex network reflects the node's connection information, which can be expressed by the number of nearest neighbor nodes. A greater value of the degree denotes a greater influence of the node. The degree centrality of node i can be defined as

$$k_i = \sum_{j=1}^n a_{ij}. \quad (3)$$

In addition, degree centrality can be normalized to

$$DC(i) = \frac{k_i}{n-1}. \quad (4)$$

2.2.2. Betweenness Centrality

Betweenness centrality [7] refers to the ratio of the number of shortest paths through a node to the number of shortest paths for all pairs of nodes in the network. The betweenness centrality of node i can be expressed as

$$BC(i) = \sum_{s \neq t \neq i \in V} \frac{g_{st}^i}{g_{st}}, \quad (5)$$

where g_{st}^i represents the number of shortest paths through node i , and g_{st} refers to the number of shortest paths for all node pairs in the node set.

2.2.3. Closeness Centrality

Closeness centrality [8] refers to the reciprocal of the sum of the shortest path lengths between a node and other nodes. This method takes the global information of the network into account. The formula is as follows:

$$CC(i) = \frac{n-1}{\sum_{j \neq i \in V} d_{ij}}, \quad (6)$$

where d_{ij} refers to the shortest path length between node i and node j .

2.2.4. Density Centrality

The principle of density centrality [15] is the circular area density formula. The degree centrality of the node and the path length between two nodes are used as the mass and radius in the density formula, which highlights the influence of the number of neighbor nodes. The density centrality of node i can be defined as

$$DNC(i) = \sum_{j \in \Gamma_i^r} \frac{k_i}{\pi d_{ij}^2}, \quad (7)$$

where Γ_i^r refers to the set of nodes whose path length is less than or equal to r ; $r = 3$ was applied in the original paper.

2.2.5. Gravity Model

The gravity model [16] is derived from the law of gravity. It considers the relationship between node neighbor layer information and path length. The degree centrality is regarded as quality, and the shortest path length is taken as the distance. The influence of node i is expressed as

$$GM(i) = \sum_{j \neq i} \frac{k_i k_j}{d_{ij}^2}. \quad (8)$$

2.2.6. Clustered Local-Degree (CLD) Method

The clustered local-degree method [17] first obtains the sum of degree centrality of the nearest neighbor nodes, and then links the obtained results with the clustering coefficient of the nodes to propose a method for identifying the influence of nodes, as expressed below.

$$CLD(i) = (1 + C_i) \sum_{j \in \Gamma_i} k_j, \quad (9)$$

where C_i represents clustering coefficient of node i , and Γ_i represents the nearest neighborhood.

2.2.7. GLI Method

The GLI method [19] contains both global location information and local structure information. First, considering that the k -shell decomposition method divides many nodes into the same layer, an improved k -shell decomposition method (Iks) was defined as follows:

$$Iks(i) = ks(i) + nit(i), \quad (10)$$

where $ks(i)$ refers to the level of node i after k -shell decomposition, and $nit(i)$ refers to the number of iterations when the node is deleted during the iteration. Then, researchers proposed a new node influence identification method named the GLI method by linking the obtained improved method with degree centrality and path information, expressed as follows:

$$GLI(i) = \exp\left(\frac{Iks(i) + k_i}{\sum_{i=1}^n (Iks(i) + k_i)}\right) \times \sum_{j \in \Gamma_i} \frac{Iks(j) + k_j}{d_{ij}}. \quad (11)$$

In this case, the neighbors within three hops of the node are considered, i.e., $r = 3$.

3. Experimental Data and Evaluation Criteria

3.1. Datasets

In order to verify the effectiveness of the method proposed in this paper, several real networks with different structures were selected for experiments. The real networks used in this article were the Contiguous network [24], Dolphin network [25], Polbooks network [26], Word network [27], Jazz network [28], Slavko Facebook network [26], USAir network [29], Netscience network [27], Infectious network [30], and Email network [31]. The datasets can also be found at <https://github.com/Ismileo/Datasets>, accessed on 19 July 2021 [32]. Table 2 shows the basic topological properties of the networks. In the table, n and m denote the number of nodes in the network and the number of connected edges, respectively. k_{\max} refers to the maximum degree, $\langle k \rangle$ refers to the average degree of all network nodes, D is the network diameter, L is the average path length, C refers to the average clustering coefficient, and r refers to the assortative coefficient of the network [33].

Table 2. Basic topological characteristics of 10 real networks.

Networks	n	m	k_{\max}	$\langle k \rangle$	D	L	C	r
Contiguous	49	107	8	4.367	11	4.163	0.497	0.2334
Dolphins	62	159	12	5.129	8	3.357	0.259	−0.0436
Polbooks	105	441	25	8.4	7	3.079	0.488	−0.1279
Word	112	425	49	7.589	5	2.536	0.173	−0.1293
Jazz	198	2742	100	27.697	6	2.235	0.618	0.0202
Slavko	324	2218	58	13.691	7	3.054	0.466	0.2473
USAir	332	2126	139	12.807	6	2.738	0.625	−0.2079
Netscience	379	914	34	4.823	17	6.042	0.741	−0.0817
Infectious	410	2765	50	13.488	9	3.631	0.456	0.2258
Email	1133	5451	71	9.622	8	3.606	0.220	0.0782

3.2. Spreading Model and Evaluation Criteria

3.2.1. SIR Model

The SIR model can be used to describe the process of information dissemination, which is widely used in the field of infectious diseases. In this paper, the SIR model was used to simulate the node infection process to obtain the influence of nodes. Nodes in the SIR spreading model can exist in the following three states: (i) susceptible (S), (ii) infected (I), and (iii) recovered (R). The susceptible state means that the node is not currently infected but has the possibility of being infected, the infected state means that it has been infected and can infect other nodes with a certain probability, and the recovered state means that it cannot infect other nodes and cannot be infected by other nodes. The specific process is to first set a node to be in an infected state and other nodes to be in a susceptible state, and then the infected node can infect the susceptible nodes with a certain spread probability β , while the infected nodes will also be converted to the recovery state with a recovery probability $\lambda = 1$, until the whole process is in a stable state. In this paper, the spreading probability β was set to take a value around the spreading threshold [34], and the total number of nodes infected by a node during spreading was considered as the influence of a node. In this article, we carried out 1000 independent simulations, and the average value was used to represent the node influence according to SIR model. The proposed method and other methods were compared on the basis of the ranking results obtained from the SIR spreading model. A closer ranking result to the SIR model denotes a higher accuracy.

3.2.2. CCDF Method

In the process of obtaining node influence, there will be multiple nodes with the same value. The complementary cumulative distribution function (CCDF) can show the probability distribution of the sorting results, and the effect of each method can be observed through the change trend. The specific formula is as follows:

$$CCDF(r) = 1 - \frac{\sum_{i=1}^r n_i}{n}, \quad (12)$$

where n_i refers to the numerical value in the i -th place in a ranking list, and n refers the number of all nodes. When the number of different values in a ranking list is closer to n , it means that the method can distinguish the influence of each node more effectively, and the decrease rate of CCDF will be smaller.

3.2.3. Kendall Correlation Coefficient

The Kendall correlation coefficient [35] is usually used to evaluate the correlation of two sorting results, with a value range of $[-1,1]$; when the value is 1, it means that the two groups have the same sorting result, whereas a value of -1 means that two sets of numbers are completely negatively correlated, and a value of 0 means the sorting results are independent of each other. Suppose there exist two sequences X and Y with n elements, whereby the sequence $XY_i = (x_i, y_i)$ is formed by taking the elements at the same position in the sequences X and Y . For any two elements XY_i and XY_j in the newly composed sequence, there are three cases: (i) if $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$, it is said that the node pairs are concordant; (ii) if $x_i > x_j$ and $y_i < y_j$ or $x_i < x_j$ and $y_i > y_j$, the node pairs are discordant; (iii) if the above conditions are not met, the node pairs are neither concordant nor discordant. The Kendall correlation coefficient can be expressed by the following formula:

$$\tau(X, Y) = \frac{2(C - D)}{n(n - 1)}, \quad (13)$$

where C refers to the number of concordant pairs, and D is the number of discordant pairs. It is worth noting that, in addition to the Kendall correlation coefficient, there are the Pearson correlation coefficient, Spearman's rank correlation coefficient, and Gamma correlation coefficient [36].

3.2.4. Jaccard Similarity Coefficient

Jaccard similarity [37] can be used to evaluate the similarity of the two methods in the sorting results, expressed as the ratio of the number of intersection nodes to the number of union nodes. The specific formula is

$$J_r(X, Y) = \frac{|X(r) \cap Y(r)|}{|X(r) \cup Y(r)|}, \quad (14)$$

where $X(r)$ and $Y(r)$ refer to the first r elements in the two lists X and Y , respectively. A closer J_r result to 1 denotes that the two lists are closer, which can be used to evaluate the accuracy of a certain ratio of node identification.

4. Experiment and Analysis

In this section, the CCDF, Kendall correlation coefficient, influence consistency, and Jaccard similarity coefficient were used to evaluate the proposed method. The experimental process and results are presented below.

4.1. Discrimination Experiment

The easy accessibility of degree centrality has led to it being widely used. Degree centrality can be expressed only by the number of nearest neighbor nodes. This is its advantage and its disadvantage. The disadvantage is that it is too simple to represent node information using degree centrality, which will cause multiple nodes to be defined with the same influence. The information of the nodes is also influenced by the environmental elements, which leads to the majority of nodes having different influence results. Therefore, the first step of an effective method should be able to distinguish the importance of each node. Only by distinguishing the influence of each node can an effective ranking be carried out. To verify the discernibility of each method on the node, this paper used the CCDF method to determine the discrimination effect of the obtained sorting results, as shown in Figure 2.

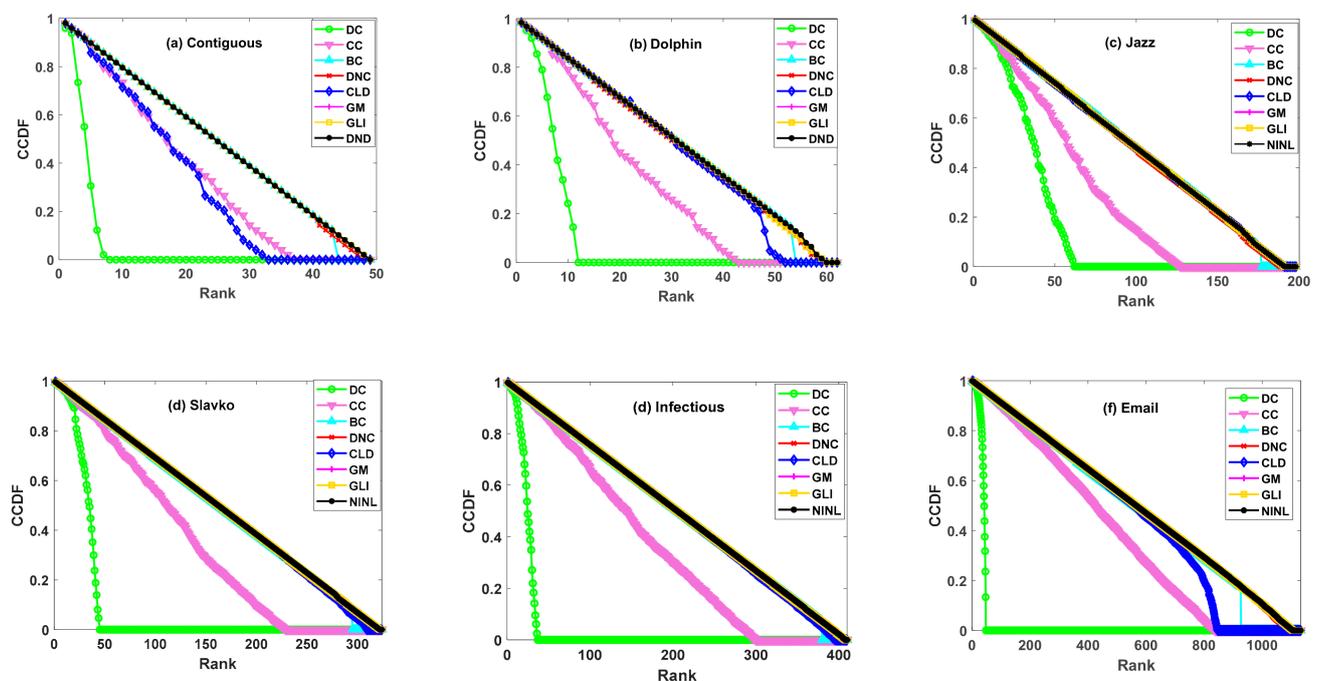


Figure 2. CCDF of the ranking results obtained using each method. (a–f): CCDF on the network of Contiguous, Dolphin, Jazz, Slavko, Infectious, Email.

Figure 2 shows the change trend of CCDF obtained using each method on the six networks of Contiguous, Dolphin, Jazz, Slavko, Infectious, and Email. The downward trend of CCDF at a certain point is determined by the frequency of the corresponding ranking point. If the number of corresponding ranking nodes is higher, the change trend is more drastic. The closer the CCDF is to a straight line, the better the effect is of distinguishing node influence. From the results in Figure 2, we can see that DC and CC methods declined faster among the six networks compared to other methods. DC represents the most local information of nodes; hence, it is very easy to define multiple nodes as the same value, which also highlights the biggest shortcoming of the DC method. Furthermore, CC considers the global shortest path length, and the result of superposition is likely to have the same value. The CLD method has a certain limitation because it only considers the node's own clustering coefficient and the degree centrality of the nearest neighbor nodes, which results in it not being able to distinguish the influence of nodes well. It can also be observed that the BC method can make an effective distinction at an early stage, but there is a sudden decrease at a later stage. The BC method considers the pivotal role of nodes in the network, but there are also several nodes with insignificant pivotal roles in the network. Such nodes will also vary in importance when they are affected by the environment; however, the BC method cannot distinguish this phenomenon. Overall, GM, GLI, and the method proposed in this paper best discriminated the influence of network nodes, basically demonstrating a straight line. It can be concluded that the method proposed in this paper can successfully distinguish the spreading influence of each node.

4.2. Accuracy Experiment

4.2.1. Selection of p -Value

In order to ensure the rationality of the value of p , this paper used the Kendall coefficient to carry out simulation research on 10 networks, and the simulation results are shown in Figure 3. We set the value of p as the x -axis and Kendall coefficient as the y -axis. To observe the changes more clearly, a partially enlarged subfigure is presented. According to the experimental results, the Kendall coefficient value increased with the increase in p at the beginning, but stopped increasing after reaching a threshold. In the partial enlarged image, we can see that the Kendall values of the networks except for Jazz and Infectious began to decrease after a p -value of 3, whereas the Jazz and Infectious networks began to decrease after a p -value of 4. Therefore, this paper selected $p = 3$ for subsequent experiments.

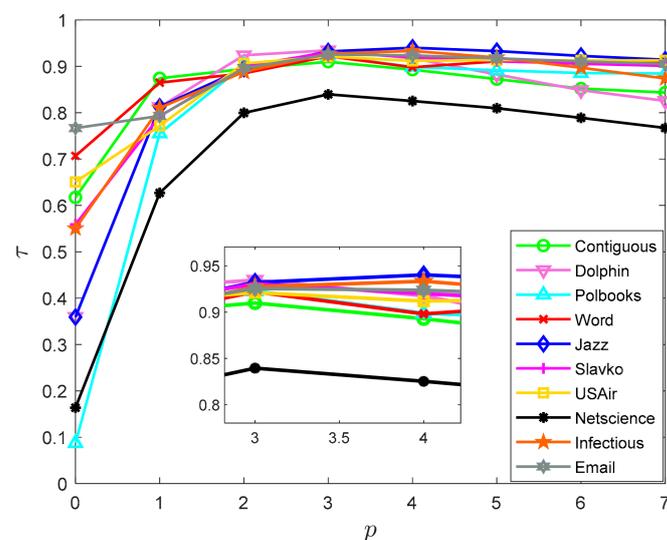


Figure 3. Kendall values of different p -values on 10 networks.

Table 3 shows the Kendall values obtained by comparing all methods and the SIR model using the 10 networks. In this paper, the spreading probability β set in the SIR model was the value near the spreading threshold β_{th} , and 1000 independent SIR model simulation experiments were conducted to represent the influence of a node with the average value. According to the experimental results in Table 3, the BC method had the worst effect in identifying nodes. Comparing the Kendall coefficients of the CLD and DC methods, it can be found that superimposing the information of the nearest neighbor nodes could improve the accuracy of node identification. Comparing the GM, GLI, and DC methods, it can be found that the combination of the node itself and the neighbor layer information could more effectively identify influential nodes. From this table, we can see that the NINL method proposed in this paper had a more obvious recognition effect and a more accurate effect compared with other methods, thus verifying that accumulating the nearest neighbor information several times after combining node and neighbor layer information can more accurately reflect the node influence and achieve effective identification.

Table 3. Kendall coefficient between ranking list obtained by SIR model and by other methods using 10 networks.

Networks	β_{th}	β	τ_{DC}	τ_{CC}	τ_{BC}	τ_{DNC}	τ_{CLD}	τ_{GM}	τ_{GLI}	τ_{NINL}
Contiguous	0.2027	0.20	0.7126	0.7253	0.5587	0.8155	0.8435	0.8690	0.8469	0.9099
Dolphin	0.1470	0.15	0.7721	0.6187	0.5389	0.8355	0.7916	0.8731	0.8355	0.9344
Polbooks	0.0838	0.09	0.7518	0.3679	0.3505	0.7679	0.8139	0.8198	0.5141	0.9229
Word	0.0726	0.08	0.8311	0.8549	0.6523	0.8822	0.8388	0.9086	0.8784	0.9218
Jazz	0.0259	0.03	0.8069	0.7080	0.4569	0.8175	0.8655	0.8505	0.8937	0.9322
Slavko	0.0466	0.05	0.7719	0.7128	0.3625	0.8234	0.8538	0.8411	0.7938	0.9305
USAir	0.0225	0.03	0.7251	0.8043	0.5081	0.8157	0.8854	0.8243	0.8522	0.9211
Netscience	0.1247	0.13	0.5955	0.3292	0.3048	0.7724	0.7980	0.7788	0.6950	0.8395
Infectious	0.0534	0.06	0.7281	0.6095	0.3707	0.7877	0.8105	0.8186	0.6984	0.9273
Email	0.0535	0.06	0.7615	0.8138	0.6203	0.8330	0.8622	0.8226	0.8345	0.9255

4.2.2. Influence Consistency Experiment

The influence consistency experiment was used to show the correlation between each method and the node influence and the correlation between the proposed method and each method. A higher correlation denotes a higher accuracy of node recognition. Figure 4 shows the experimental results of six methods on the Word, Jazz, and USAir networks. The x -axis is the node influence calculated by each method, and the y -axis is the node influence calculated by 1000 independent SIR spreading models. The experimental plots show that all six methods had positive correlations with node influence. Among them, the DNC and GM methods had an obvious upward convex trend, the GLI method had an obvious downward convex trend, and the DC, CLD, and NINL methods had a straight-line trend. The NINL method proposed in this paper showed better linearity in the three networks, and the scattered points were basically maintained around the straight line, suggesting a relatively strong correlation with the influence of nodes. In summary, the method proposed in this paper is more advantageous for discovering influential nodes.

To explore the intrinsic connection between the methods, Figure 5 shows the nodal influence correlation diagram between the proposed method and the other five methods in this paper. Each node in the figure represents the value obtained by a node in the network according to different methods. The x -axis refers to the NINL method values, and the y -axis refers to five method values of DC, DNC, CLD, GM, and GLI. As shown in Figure 5, the NINL method showed a positive correlation with the other five methods, and in the Netscience network, a more divergent scatter plot was obtained. In the Email network, the NINL method and CLD method seemingly presented a straight line. There was no such correlation when using the other methods. By observing the scatter plots of the NINL method and the DC method in the three networks, we can find that the value of DC method was concentrated in a very small interval, highlighting that the weak distinguishing ability of degree centrality for nodes was the main reason for the divergence of the results.

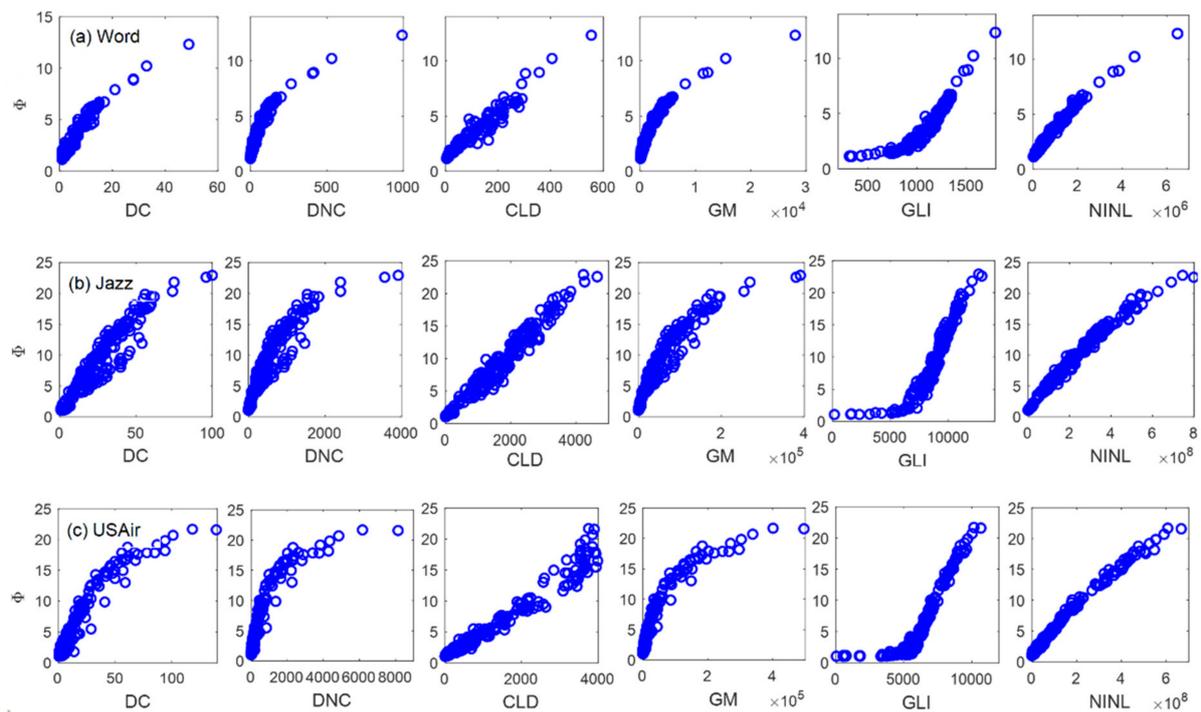


Figure 4. The correlation between six methods and the influence of nodes by SIR model on three networks. (a–c): The correlation on the network of Word, Jazz, USAir.

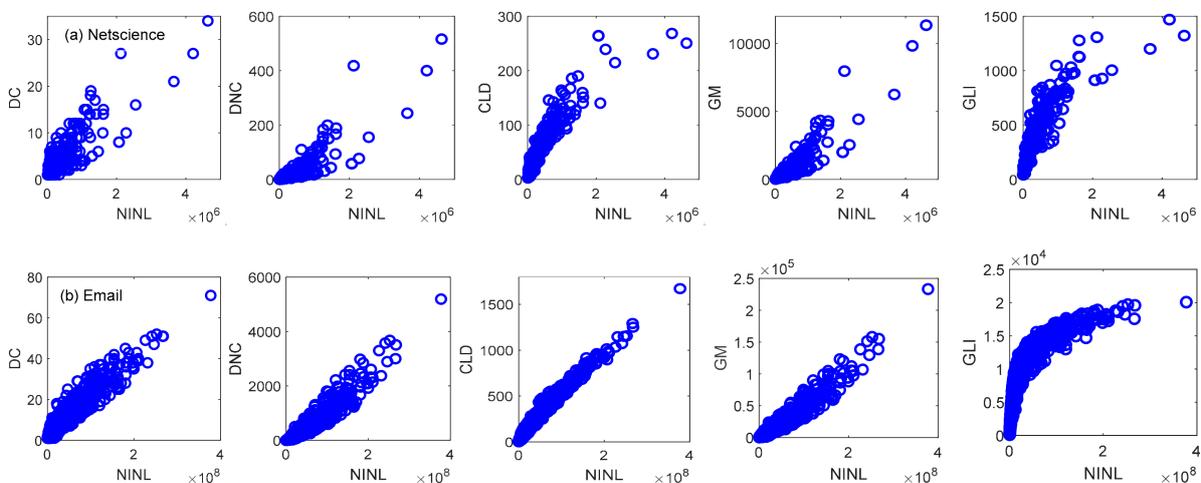


Figure 5. The correlation between several methods and the proposed method on two networks. (a,b): The correlation on the network of Netscience, Email.

4.2.3. Recognition Effect of Each Method under a Certain Range of Propagation Probability

In order to more comprehensively evaluate the influence of the spreading probability on the ranking accuracy, this paper used the SIR model to obtain the node influence within a certain range of spreading probability, and the effectiveness of the proposed method was verified by the Kendall correlation coefficient. The results are shown in Figure 6. It can be seen from the figure that the BC method had the weakest ability to accurately identify influential nodes on the six networks. When the spreading probability was small, it can be seen that the DC, DNC, and GM methods had higher Kendall correlation coefficients, which was due to the fact that the infected node could only infect a portion of nodes close to it. At this time, the spreading process was limited to a certain local area, and the degree of a node in the local area had a great impact on the spreading influence. As the spreading

probability was near the spreading threshold, the NINL method fully considered the relationship between the nearest neighbor and neighbor layer information within a certain range. Compared with other methods, NINL had a higher correlation with node influence, reflecting a better recognition effect. When the spreading probability was increased, we can see that the correlation curve had a significant decline in the USAir network. This is because the clustering coefficient of the network was very high, which led to the connection between nodes becoming closer. Under these circumstances, the information between nodes could be easily transmitted through the closely connected nodes layer by layer. As a result, the recognition accuracy of the NINL method approached that of the other methods. In general, the proposed NINL method could more accurately evaluate the influence of nodes than other methods around the propagation threshold.

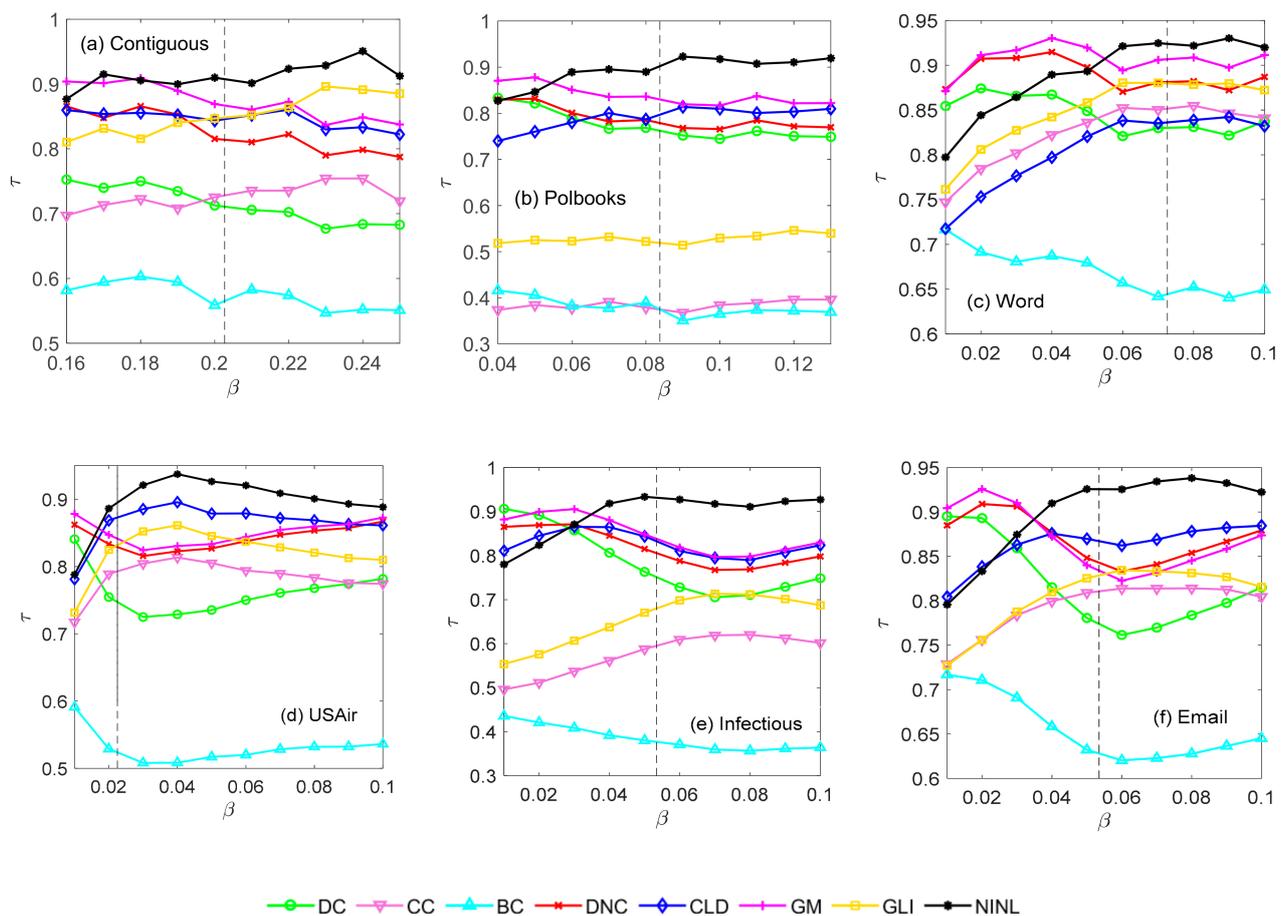


Figure 6. Accuracy comparison of different methods under different probabilities. (a–f): The Kendall correlation coefficient on the network of Contiguous, Polbooks, Word, USAir, Infectious, Email.

4.2.4. Recognition Effect of Each Method under a Certain Percentage of Ranking Results

There are a very small number of nodes in the network responsible for the normal operation of the entire network. It is of great significance to identify the most influential nodes. Table 4 shows the top 10 nodes obtained using each method in the Word and USAir networks. In this table, the Φ value given in the last column of the table is the ranking result calculated by the SIR model under the condition of spreading probability β . By observing the results of each method and the SIR model, it can be found that, in the Word network, the top 10 nodes identified by the DC, DNC, GM, GLI, and NINL methods had nine identical nodes in terms of Φ , whereas eight of the CC and CLD methods were identical. The BC method had only seven identical nodes. In the USAir network, the GM, GLI, and NINL methods identified nine identical nodes in terms of Φ , whereas DC

and DNC only identified eight identical nodes, CC identified six identical nodes, and BC and CLD only identified five identical nodes. According to the above analysis, the NINL method could accurately identify the top 10 nodes.

Table 4. Top 10 nodes measured using various methods in Word and USAir networks.

Word Network										USAir Network									
Rank	DC	CC	BC	DNC	CLD	GM	GLI	NINL	Φ	Rank	DC	CC	BC	DNC	CLD	GM	GLI	NINL	Φ
1	18	18	18	18	18	18	18	18	18	1	118	118	118	118	109	118	118	118	261
2	3	3	3	3	3	3	3	3	3	2	261	261	8	261	131	261	261	261	118
3	44	52	44	52	52	52	52	52	52	3	255	67	261	255	112	255	255	255	255
4	52	44	52	44	44	44	44	44	44	4	152	255	201	182	299	182	182	182	182
5	105	28	10	105	51	105	105	105	105	5	182	201	47	152	118	152	152	152	230
6	10	105	80	10	105	10	25	51	10	6	230	182	182	230	255	230	230	230	176
7	25	10	105	28	22	25	51	10	25	7	166	47	255	166	176	166	67	112	152
8	28	27	28	25	55	51	28	26	51	8	67	166	152	67	147	67	166	166	147
9	51	25	2	51	25	28	26	25	28	9	112	248	313	112	261	112	112	67	67
10	2	26	29	26	32	26	10	55	55	10	201	112	13	201	301	147	147	147	166

In order to explore the recognition accuracy of the top nodes, the Jaccard similarity coefficient was used as the evaluation standard to carry out related experiments. Figure 7 presents the results of Jaccard similarity experiments on six types of networks. The x-axis represents the range of ranking results considered, and the y-axis represents the Jaccard similarity coefficient. A larger Jaccard coefficient denotes a higher similarity and, thus, a more effective recognition result. It can be seen from Figure 7 that, as the range of the ranking results increased, the Jaccard similarity coefficients obtained by each method became more and more stable. At the same time, the NINL method proposed in this paper clearly showed a superior similarity curve to the other seven methods in the Slavko, Netscience, Infectious, and Email networks. Thus, it can be considered that the NINL method was highly correlated with the influence of nodes and, thus, could more accurately identify the first r nodes.

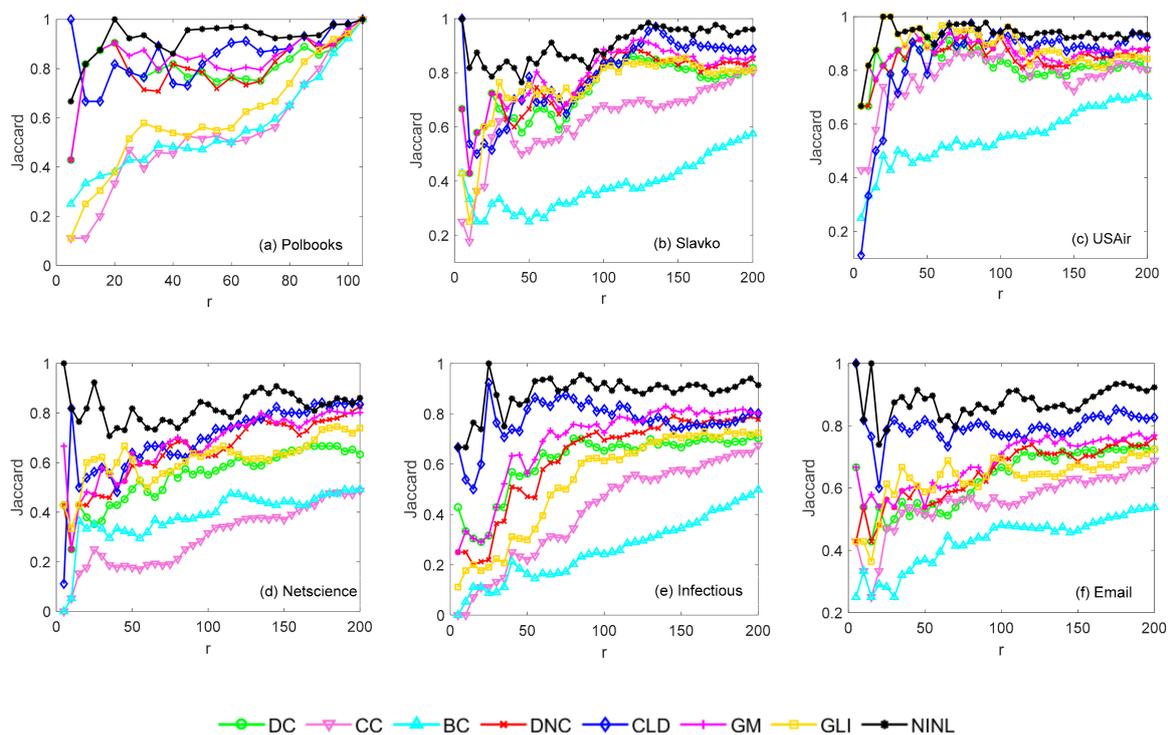


Figure 7. Comparison of Jaccard similarity of different methods. (a–f): The Jaccard similarity on the network of Polbooks, Slavko, USAir, Netscience, Infectious, Email.

5. Conclusions

In this paper, we studied the problem of identifying influential nodes in the network. Identifying influential nodes accurately in the network can provide a clearer understanding of the overall network function implementation process and the information dissemination between nodes. This paper mainly uses the symmetric adjacency matrix of the unweighted and undirected network to obtain various information. The paper first defines a node's own information and neighbor layer nodes within a certain range as the initial node influence, and then takes the nearest neighbor information multiple times as the final influence of nodes. In different networks, the number of times to fetch the nearest neighbor information is different. In order to ensure the rationality of the p -value, a verification experiment of p selection was done. At the same time, this paper used the CCDF curve to perform a discrimination experiment on multiple real networks, while the Kendall coefficient and Jaccard similarity coefficient were used to carry out recognition and accuracy experiments. The method proposed in this paper effectively avoided the phenomenon of most nodes having the same value. Furthermore, it had higher accuracy in identifying the influence of nodes near the propagation threshold. The experimental results show that the proposed method was more effective than other methods in identifying influential nodes, which is significant for understanding the node information dissemination process.

Author Contributions: Conceptualization, J.Z. and L.W.; methodology, J.Z.; writing—original draft preparation, J.Z.; writing—review and editing, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets that support the findings of this study are openly available at <https://github.com/lsmileo/Datasets>, accessed on 19 July 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pawan, K.; Ravins, D. Formalising and detecting community structures in real world complex networks. *J. Syst. Sci. Complex.* **2021**, *34*, 180–205.
2. Robitaille, A.L.; Webber, Q.M.R.; Turner, J.W.; Wal, E.V. The problem and promise of scale in multilayer animal social networks. *Curr. Zool.* **2021**, *67*, 113–123. [[CrossRef](#)] [[PubMed](#)]
3. Li, Q.; Jing, R.Z. Characterization of delay propagation in the air traffic network. *J. Air Transp. Manag.* **2021**, *94*, 102075. [[CrossRef](#)]
4. Liu, Q.; Chen, Y.; Zhang, G.Q.; Wang, G.Y. A novel functional network based on three-way decision for link prediction in signed social networks. *Cognit. Comput.* **2021**, 1–13. [[CrossRef](#)]
5. Nguyen, T.; Le, H.; Quinn, T.P.; Nguyen, T.; Venkatesh, S. GraphDTA: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics* **2021**, *37*, 1140–1147. [[CrossRef](#)] [[PubMed](#)]
6. Freeman, L.C. Centrality in social networks conceptual clarification. *Soc. Netw.* **1978**, *1*, 215–239. [[CrossRef](#)]
7. Freeman, L.C. A set of measures of centrality based on betweenness. *Sociometry* **1977**, *40*, 35–41. [[CrossRef](#)]
8. Sabidussi, G. The centrality index of a graph. *Psychometrika* **1966**, *31*, 581–603. [[CrossRef](#)]
9. Borgatti, S.P. Centrality and network flow. *Soc. Netw.* **2005**, *27*, 55–71. [[CrossRef](#)]
10. Hwang, W.; Cho, Y.; Zhang, A.; Cho, Y.R.; Hwang, W. Bridging Centrality: Identifying Bridging Nodes in Scale-free Networks. In Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06), Philadelphia, PA, USA, 20–23 August 2006.
11. Lü, L.Y.; Zhang, Y.C.; Yeung, C.H.; Zhou, T. Leaders in social networks, the Delicious case. *PLoS ONE* **2011**, *6*, e21202. [[CrossRef](#)]
12. Kitsak, M.; Gallos, L.K.; Havlin, S.; Liljeros, F.; Muchnik, L.; Stanley, H.E.; Makse, H.A. Identification of influential spreaders in complex networks. *Nat. Phys.* **2010**, *6*, 888–893. [[CrossRef](#)]
13. Lü, L.Y.; Zhou, T.; Zhang, Q.M.; Stanley, H.E. The H-index of a network node and its relation to degree and coreness. *Nat. Commun.* **2016**, *7*, 440–442. [[CrossRef](#)]
14. Bae, J.; Kim, S. Identifying and ranking influential spreaders in complex networks by neighborhood coreness. *Phys. A* **2014**, *395*, 549–559. [[CrossRef](#)]
15. Ahmed, I.; Mohamed, E.H. Density centrality: Identifying influential nodes based on area density formula. *Chaos Solitons Fractals* **2018**, *114*, 69–80.

16. Li, Z.; Ren, T.; Ma, X.Q.; Zhou, T. Identifying influential spreaders by gravity model. *Sci. Rep.* **2019**, *9*, 355–359. [[CrossRef](#)] [[PubMed](#)]
17. Li, M.T.; Zhang, R.S.; Hu, R.J.; Yang, F.; Yao, Y.; Yuan, Y.B. Identifying and ranking influential spreaders in complex networks by combining a local-degree sum and the clustering coefficient. *Int. J. Mod. Phys. B* **2018**, *32*, 1850118. [[CrossRef](#)]
18. Sheng, J.F.; Dai, J.Y.; Wang, B.; Duan, G.H.; Long, J.; Zhang, J.K.; Guan, K.R.; Hu, S.; Chen, L.; Guan, W.H. Identifying influential nodes in complex networks based on global and local structure. *Phys. A* **2020**, *541*, 123262. [[CrossRef](#)]
19. Yang, Y.Z.; Hu, M.; Huang, T.Y. Influential nodes identification in complex networks based on global and local information. *Chin. Phys. B* **2020**, *29*, 664–670. [[CrossRef](#)]
20. Zareie, A.; Sheikahmadi, A.; Jalili, M.; Fasaeei, M.S.K. Finding influential nodes in social networks based on neighborhood correlation coefficient. *Knowl. Based Syst.* **2020**, *194*, 105580. [[CrossRef](#)]
21. Yan, X.L.; Cui, Y.P.; Ni, S.J. Identifying influential spreaders in complex networks based on entropy weight method and gravity law. *Chin. Phys. B* **2020**, *29*, 664–672. [[CrossRef](#)]
22. Babu, M.; Marimuthu, S.; Joy, M.; Nadaraj, M.; Jeyaseelan, L. Forecasting COVID-19 epidemic in India and high incidence states using SIR and logistic growth models. *Clin. Epidemiol. Glob. Health* **2021**, *9*, 26–33.
23. Kermack, W.O.; McKendrick, A.G. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. A* **1927**, *115*, 700–721.
24. Contiguous USA Network Dataset—KONECT. 2017. Available online: <http://konect.cc/networks/contiguous-usa> (accessed on 5 July 2021).
25. Lusseau, D.; Schneider, K.; Boisseau, O.J.; Haase, P.; Slooten, E.; Dawson, S.M. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behav. Ecol. Sociobiol.* **2003**, *54*, 396–405. [[CrossRef](#)]
26. Blagus, N.; Šubelj, L.; Bajec, M. Self-similar scaling of density in complex real-world networks. *Phys. A* **2012**, *391*, 2794–2802. [[CrossRef](#)]
27. Newman, M.E.J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **2006**, *74*, 036104. [[CrossRef](#)] [[PubMed](#)]
28. Gleiser, P.M.; Danon, L. Community structure in jazz. *Adv. Complex Syst.* **2003**, *6*, 565–573. [[CrossRef](#)]
29. Batagel, V.; Mrvar, A. Pajek-program for large Network analysis. *Connections* **1998**, *21*, 47–57.
30. Isella, L.; Stehlé, J.; Barrat, A.; Cattuto, C.; Pinton, J.F.; Broeck, W.V.D. What’s in a crowd? Analysis of face-to-face behavioral networks. *J. Theor. Biol.* **2011**, *271*, 166–180. [[CrossRef](#)]
31. Guimerà, R.; Danon, L.; Díaz, G.A.; Giralt, F.; Arenas, L. Self-similar community structure in a network of human interactions. *Phys. Rev. E* **2003**, *68*, 065103. [[CrossRef](#)]
32. Datasets. Available online: <https://github.com/ismileo/Datasets> (accessed on 19 July 2021).
33. Newman, M.E.J. Assortative mixing in networks. *Phys. Rev. Lett.* **2002**, *89*, 208701. [[CrossRef](#)]
34. Castellano, C.; Pastor-Satorras, R. Thresholds for epidemic spreading in networks. *Phys. Rev. Lett.* **2010**, *105*, 218701. [[CrossRef](#)]
35. Knight, W.R. A computer method for calculating kendall’s tau with ungrouped data. *J. Am. Stat. Assoc.* **1966**, *61*, 436–439. [[CrossRef](#)]
36. Lorentz, J.; Bolboacă, S.-D. Pearson versus Spearman, Kendall’s Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds. *Leonardo J. Sci.* **2006**, *5*, 179–200.
37. Amirkhani, D.; Bastanfard, A. An objective method to evaluate exemplar-based inpainted images quality using Jaccard index. *Multimed. Tools. Appl.* **2021**, *80*, 26199–26212. [[CrossRef](#)]