



OPEN

Identifying influential nodes in complex networks using a gravity model based on the H-index method

Siqi Zhu , Jie Zhan  & Xing Li

Identifying influential spreaders in complex networks is a widely discussed topic in the field of network science. Numerous methods have been proposed to rank key nodes in the network, and while gravity-based models often perform well, most existing gravity-based methods either rely on node degree, k-shell values, or a combination of both to differentiate node importance without considering the overall impact of neighboring nodes. Relying solely on a node's individual characteristics to identify influential spreaders has proven to be insufficient. To address this issue, we propose a new gravity centrality method called HVGC, based on the H-index. Our approach considers the impact of neighboring nodes, path information between nodes, and the positional information of nodes within the network. Additionally, it is better able to identify nodes with smaller k-shell values that act as bridges between different parts of the network, making it a more reasonable measure compared to previous gravity centrality methods. We conducted several experiments on 10 real networks and observed that our method outperformed previously proposed methods in evaluating the importance of nodes in complex networks.

Complex networks are a pervasive presence in various domains of both human society and the natural world. In each system, individuals and their relationships can be represented as networks consisting of nodes and edges^{1,2}. Recently, the identification of significant nodes in complex networks has gained significant attention from researchers, providing a new perspective for understanding the objective world and facilitating a better comprehension of the spread of diseases^{3–5}, power grid protection⁶, information dissemination^{7–9}, protein discovery¹⁰, and immunization strategies^{11,12}, among other fields^{13–15}.

To date, numerous centrality methods have been proposed to detect key nodes in complex networks. Centrality measurement methods can be primarily categorised into three types: local indices, global indices, and hybrid indices¹⁶. Local-index-based centrality methods include classical measures such as degree centrality¹⁷ (DC) and H-index¹⁸. Local-index-based methods have low computational complexity and are suitable for large-scale networks as they only consider the local neighbourhood information of nodes. However, their ability to identify influential nodes that are not central but have high impact is limited. To address this limitation, many researchers have proposed improvements, such as extended H-index centrality¹⁹ (EHC) and local clustering H-index centrality²⁰ (LCH) methods. Global-index-based centrality methods assess individuals' influence by considering the global structural information of the network, such as closeness centrality²¹ (CC) and betweenness centrality²² (BC). The main drawbacks of these measurement methods are their high computational complexity and inapplicability to large-scale networks²³. Among them, the K-shell decomposition method²⁴ (KS), as a global approach, determines the influence of nodes by differentiating their core levels and operates at a faster speed. However, the main limitation of k-shell is that it assigns the same k-shell value to many nodes, resulting in low differentiation in node influence ranking. Many efforts have been made to address this issue, such as extended neighbourhood coreness²⁵ (CNC+), classifying neighbourhood²⁶ (CN), k-shell iteration factor²⁷ (KSIF), and Mixed Degree Decomposition²⁸ (MDD). The primary limitations of these global methods are their typically high computational costs as they consider the entire topological structure of the network. Hybrid-index-based centrality methods, such as local and global influence²⁹ (LGI), local and global centrality³⁰ (LGC) and global and local information³¹ (GLI) integrate both local and global information about nodes, aiming to strike a balance between algorithm accuracy and computational complexity.

Physical and Electronic Sciences College, Hunan University of Science and Technology of China, Xiangtan 411100, People's Republic of China. ✉email: zhushiqi@mail.hnust.edu.cn; jiezhawl@163.com

The gravity model not only considers the attributes held by two nodes but also takes into account the shortest path information between nodes, which represents their mutual interactions and provides a basis for integrating local and global information. Inspired by this formula, Ma et al.³² proposed two models (G and G+) based on the gravity formula. These models adopt the k-shell value of a node as its mass and use the shortest path distance between two nodes as the distance. Building upon this, Wang et al.³³ improved the model by considering the degree values of neighbouring nodes, resulting in the improved gravity centrality (IGC). Li et al.³⁴ introduced the gravity model (GM), which employs the degree of nodes as their mass, and developed the local gravity model (LGM), which only considers node pairs within a truncated radius. Furthermore, Li et al.³⁵ combined the local clustering coefficient and degree value as the mass of nodes, proposing the generalized gravity centrality (GGC). In addition, Yang et al.³⁶ introduced a gravity centrality (KSGC) based on the K-shell value of nodes, considering the variations in interactions when nodes are located in different shell layers. Li et al.³⁷ combined the k-shell value and k-shell iteration factor as the mass of nodes, presenting the DK-based gravity model (DKGM) to enhance the model's performance. Subsequently, they considered multiple features of nodes and proposed the multi-characteristics gravity model³⁸ (MCGM). Liu et al.³⁹ introduced the spreading entropy gravity Model (SEGM), incorporating the spreading information entropy of nodes into consideration.

From the above, we can observe that many of the gravity models mentioned are either based on node degree, related to the k-shell value, or a combination of both. However, It is not enough to evaluate the importance of a node solely on the basis of its single attributes; it is also necessary to consider the location of the node and the overall influence of neighbouring nodes on it. For instance, some nodes may have a relatively small k-shell index but possess significant influence since they act as bridges connecting different communities within the network. Similarly, there are nodes with lower degree or k-shell values compared to others but are closer to the most important nodes in the network, surrounded by highly influential nodes, as a result, their importance will also be enhanced. To address this issue, we propose the H-index-based gravity centrality method (HVG), which not only considers the path information of nodes but also incorporates the overall influence of neighbouring nodes, structural hole position information of nodes, and the differential gravitational impact of nodes positioned at different locations. Experimental results demonstrate that our proposed method exhibits significant competitiveness compared to other advanced gravity models, Particularly in networks with evident community structures, it exhibits outstanding accuracy, unlike other algorithms that are prone to identifying false core nodes.

Preliminaries

Centrality measures

In the context of an undirected and unweighted simple network $G = \langle V, E \rangle$, V and E respectively represent the sets of nodes and links. The cardinality of V and E can be expressed as $|V| = N$ and $|E| = M$, indicating the presence of N nodes and M links within the network. The network's connectivity structure is typically captured by its adjacency matrix $A = (a_{ij})_{N \times N}$, where $a_{ij} = 1$ if node i and node j are linked, and 0 otherwise.

Degree centrality¹⁷ of node i is defined as

$$DC(i) = k(i), \quad (1)$$

where $k(i) = \sum_{j=1}^N a_{ij}$.

The maximum integer fulfilling that there are at least $H(i)$ neighbors of node i whose degrees are all at least $H(i)$, represented by $H(i)$, is known as the H-index¹⁸ of the node i .

The k-shell decomposition method²⁴(KS), operates through an iterative process of decomposing the network into distinct shells. Initially, KS removes nodes with a degree of 1 from the network, resulting in a decrease in the degree values of the remaining nodes. This process is repeated by removing nodes with residual degrees less than or equal to 1 until all remaining nodes have residual degrees greater than 1. The nodes removed in the first step constitute the 1-shell, and their k-shell values are assigned as 1. This process is then iteratively applied to obtain the 2-shell, 3-shell, and so on. The decomposition process continues until all nodes in the network have been accounted for.

Gravity centrality³² (G) of node i is defined as

$$G(i) = \sum_{j \in \psi_i} \frac{k_s(i)k_s(j)}{d^2(i,j)}, \quad (2)$$

where $k_s(i)$ is the k-shell value of node i , $d(i,j)$ is the shortest path distance from node i to node j , and ψ_i is the set of nodes whose distance from node i does not exceed 3.

Extended gravity centrality³² (G+) of node i is described as

$$G_+(i) = \sum_{j \in \Lambda_i} G(j), \quad (3)$$

Λ_i is the nearest neighborhood of node i .

The improved gravity centrality³³ (IGC) of node i is measured by

$$IGC(i) = \sum_{d(i,j) \leq R, j \neq i} \frac{k_s(i)k(j)}{d^2(i,j)}, \quad (4)$$

where R is the truncation radius, and the optimal truncation radius R^* can be estimated by

$$R^* \approx \frac{1}{2} \langle d \rangle, \quad (5)$$

where $\langle d \rangle$ is the average distance of the network.

Extended improved gravity centrality³³ (IGG+) of node i is described as

$$IGC_+(i) = \sum_{j \in \Lambda_i} IGC(j), \quad (6)$$

Λ_i is the nearest neighborhood of node i .

The local gravity model³⁴ (LGM) of node i is determined by

$$LGM(i) = \sum_{d(i,j) \leq R, j \neq i} \frac{k(i)k(j)}{d^2(i,j)}, \quad (7)$$

The generalized gravity centrality³⁵ (GGC) of node i is defined as

$$GGC(i) = \sum_{d(i,j) \leq R, j \neq i} \frac{S_p(i)S_p(j)}{d^2(i,j)}, \quad (8)$$

$$S_p(i) = e^{-\alpha C_i} \times k(i) \quad (9)$$

$$C_i = \frac{2n_i}{k(i)(k(i) - 1)} \quad (10)$$

where C_i is the local clustering coefficient of node i , n_i denotes the number of edges between neighbors of node i , and $\alpha = 2$.

The k-shell based on gravity centrality³⁶ (KSGC) is defined as

$$KSGC(i) = \sum_{d(i,j) \leq R, j \neq i} F(i,j), \quad (11)$$

$$F(i,j) = c_{ij} \frac{k(i)k(j)}{d^2(i,j)}, \quad (12)$$

$$c_{ij} = e^{\frac{k_s(i) - k_s(j)}{k_{s\max} - k_{s\min}}} \quad (13)$$

where c_{ij} is the coefficient of attraction exerted by node i on node j , $k_s(i)$ and $k_s(j)$ denote the k-shell values of node i and node j , respectively. $k_{s\max}$ and $k_{s\min}$ refer to the largest and smallest k-shell values present in the network. $d(i,j)$ is the shortest path distance from node i to node j .

The DK-based gravity model³⁷ (DKGM) is measured by

$$DKGM(i) = \sum_{d(i,j) \leq R, j \neq i} \frac{DK(i)DK(j)}{d^2(i,j)}, \quad (14)$$

$$DK(i) = k(i) + k_s^*(i), \quad (15)$$

$$k_s^*(i) = k_s(i) + \frac{p(i)}{q(k) + 1}, \quad (16)$$

assume that the value of the k-shell of node i is $k_s(i)$. For the process of the k-degree iteration, the total iteration number is $q(k)$, and node i is removed in the $p(i)$ iteration of the k-degree process. $k_s^*(i)$ is called the improved k-shell index of node i .

The multi-characteristics gravity model³⁸ (MCGM) is measured by

$$MCGM(i) = \sum_{d(i,j) \leq R, j \neq i} \frac{\left(\frac{k(i)}{k_{\max}} + \frac{\alpha k_s(i)}{k_{s\max}} + \frac{x(i)}{x_{\max}} \right) \left(\frac{k(j)}{k_{\max}} + \frac{\alpha k_s(j)}{k_{s\max}} + \frac{x(j)}{x_{\max}} \right)}{d^2(i,j)}, \quad (17)$$

$$\alpha = \frac{\max\left\{ \frac{k_{\text{mid}}}{k_{\max}}, \frac{x_{\text{mid}}}{x_{\max}} \right\}}{\frac{k_{\text{smid}}}{k_{s\max}}}, \quad (18)$$

where k_{mid} , k_{smid} and x_{mid} denote the median of degree value, k-shell value and eigenvector centrality value, respectively. k_{max} , k_{smax} and x_{max} denote the maximum values of degree value, k-shell value, and eigenvector centrality value.

The entropy-based gravity model³⁹ (SEGM) is defined as

$$SEGM(i) = \sum_{d(i,j) \leq R, j \neq i} \frac{SE(i)SE(j)}{d^2(i,j)}, \quad (19)$$

$$SE(i) = e^{E(i)}k(i), \quad (20)$$

$$E(i) = - \sum_{j \in \Gamma(i)} I(j) \ln I(j), \quad (21)$$

$$I(i) = \frac{k(i)}{\sum_{j \in \Gamma(i)} k(j)}, \quad (22)$$

where $E(i)$ is the information entropy of node i , $\Gamma(i)$ represents the set of neighboring nodes of node i , and $I(i)$ is the importance of node i .

The SIR model used in this paper

To evaluate the ranking of impact generated by the algorithm and the simulation, we employed the widely used SIR model⁴⁰. In the beginning, a single node in the network, referred to as the "source node," is in the infected state (I), while the remaining nodes are in the susceptible state (S). An infected node has the potential to infect its susceptible neighbors with a probability of β , and the probability of each infected node entering the recovery (R) state is λ , after which it ceases to participate in the dynamics. This propagation process continues until no infected nodes remain in the network. The impact of any given node i can be estimated by

$$F(i) = N_r/N \quad (23)$$

the number of nodes that recover after the diffusion process has stabilized is represented by N_r . For the sake of simplicity, λ has been set to 1. Subsequently, the corresponding epidemic threshold⁴¹ can be computed by

$$\beta_c \approx \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} \quad (24)$$

where $\langle k \rangle$ and $\langle k^2 \rangle$ are the degree distribution's average degree and second-order moments.

Measures

Kendall's tau coefficient

Kendall's tau coefficient⁴² is a measure of correlation between two sequences, with a larger value indicating a greater similarity between the sequences. The definition of Kendall's tau coefficient is as follows: given two sequences X and Y of the same length, where the i th values are represented by x_i and y_i , respectively. Let each pair of elements x_i and y_i form a set, denoted by (x_i, y_i) . If $x_i > x_j$ and $y_i > y_j$, or $x_i < x_j$ and $y_i < y_j$, the pairs (x_i, y_i) and (x_j, y_j) are considered concordant. They are considered discordant if $x_i > x_j$ and $y_i < y_j$, or $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ and $y_i = y_j$, the pair is neither concordant nor discordant. Therefore, the Kendall's tau coefficient τ is defined as

$$\tau = \frac{2(n_+ - n_-)}{N(N - 1)} \quad (25)$$

where n_+ is the number of concordant pairs, and n_- is the number of discordant pairs.

Jaccard similarity coefficient

In some applications, concentrating on the top-rank nodes rather than all nodes may be appropriate. In contrast to the Kendall correlation coefficient, the Jaccard similarity coefficient is utilized to assess the similarity between the top-k nodes in two ranking lists^{25,43}. The Jaccard similarity is calculated by dividing the number of common nodes by the number of unique nodes in the two lists, and its expression is

$$Jaccard(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (26)$$

where X and Y represent the top-k nodes with the highest influence as determined by two different methods. In the context of our experiments, X represents the top-k nodes identified by HvGC and other baseline methods, while Y represents the top-k nodes obtained through the SIR simulation. We use the Jaccard similarity coefficient to measure the similarity between these two sets of top-k nodes. The Jaccard similarity coefficient ranges from 0 to 1, where a higher value indicates a greater degree of similarity between the two ranking results. A Jaccard similarity coefficient of 0 indicates completely distinct results, while a value of 1 indicates that the two sets of top-k nodes are identical.

The monotonicity index

The monotonicity²⁵ M is used to quantitatively measure the resolution of different indices in ranking list X , and can be calculated by

$$M(X) = \left[1 - \frac{\sum_{c \in V} N_c(N_c - 1)}{N(N - 1)} \right]^2 \quad (27)$$

where N is the size of network, and N_c is the number of nodes with the same index value c .

Results

Algorithms

Previous research has utilized the gravity model approach to analyze node importance in complex networks. Degree and k-shell values are commonly used metrics to consider the number of neighbors a node has and its position within the network, respectively. However, these metrics alone do not capture the overall influence of a node's neighbors. While the H-index considers the importance of a node's neighbors, it may overlook certain information from neighboring nodes, failing to account for the collective impact of all neighbors. We take the toy network shown in Fig. 1 to illustrate the problem for H-index, where the node spreading capacity derived from 1000 independent runs of the SIR model has been numerically labeled in Fig. 1. Obviously, $H(1) = H(2) = H(3) = H(4) = H(10) = 1, H(5) = H(7) = H(8) = 3, H(6) = H(9) = 2$, where $H(i)$ represents the H-index of node i . The H-index always assigns the same value to different nodes, which leads to a lack of excellence in the ability to differentiate the influence of nodes.

The same issue exists in DC¹⁷ and KS²⁴. Additionally, from Fig. 1, it can be observed that Node 3 has a higher propagation capability compared to Node 9, but Node 3 has a lower H-index than Node 9. This indicates that the H-index overlooks some information from the neighbors of a node. From this, we take out all neighboring nodes in the set of neighbors of node i with degree values greater than or equal to $H(i)$ and add up the degree values of these nodes to measure the overall influence of the neighboring nodes on node i . The value obtained is denoted as $HV(i)$, and the expression is

$$HV(i) = \sum_{j \in \Lambda_i} [k(j) | k(j) \geq H(i)], \quad (28)$$

where Λ_i is the nearest neighborhood of node i , $H(i)$ represents the H-index of node i .

By incorporating the overall influence of node neighbors into the definition, it enhances the discriminative power of node identification compared to the H-index. However, it is still insufficient to accurately distinguish cluster-like nodes, due to their close connections, these nodes can more easily achieve greater HV values, but, their actual influence may not be greater than that of nodes with lower HV values. As shown in Fig. 1. $HV(6) = 8, HV(9) = 7, HV(3) = 4$, and the actual propagation capacity from high to low is nodes 3, 9, and 6, a similar problem with the k-shell approach was noted by Liu et al.⁴⁴ In other words, removing node 3 from the network would result in nodes 1, 2, and 4 losing their interactions with the core nodes, while removing node 6 has a minimal impact on information transmission in the network. This finding demonstrates the higher importance of nodes that serve as bridges between different clusters compared to those within individual clusters.

Based on this, we considered the structural hole position of nodes to enhance the algorithm's ability to identify nodes within community networks. This allows us to identify those bridge nodes that may not have high HV values but play a crucial role in facilitating information flow across different parts of the network. The network constraint coefficient measures the level of constraints imposed on nodes forming a structural hole (SH) in a network⁴⁵, and it can be calculated as follows:

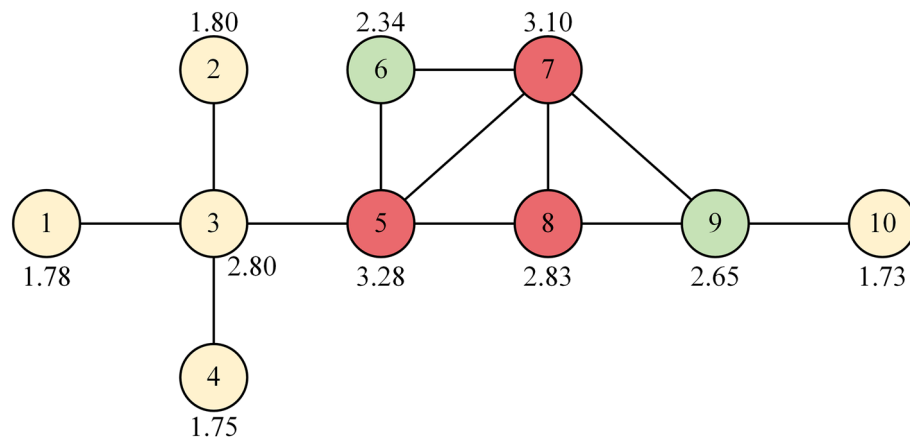


Figure 1. A toy network. The red node is ranked first in terms of H-index, while green and yellow represent second and third, respectively.

$$c(i) = \sum_{j \in \Gamma(i)} \left(p_{ij} + \sum_{w \in \Gamma(i) \cap \Gamma(j)} p_{iw} \cdot p_{wj} \right)^2, \tag{29}$$

$$p_{ij} = \frac{z_{ij}}{\sum_{w \in \Gamma(i)} z_{iw}}, \tag{30}$$

where $\Gamma(i)$ represents the set of neighboring nodes of node i , and $w \in \Gamma(i) \cap \Gamma(j)$ indicates the nodes that are common neighbors of both node i and node j . p_{ij} represents the proportion of energy invested by node i to maintain its relationship with node j . where $z_{ij} = 1$ ($i \neq j$) if there is a link between nodes i and j , otherwise $z_{ij} = 0$. Based on the above discussions, the gravity centrality based on the H-index (HVGC) measure proposed in this paper is defined as follows:

$$HVGC(i) = \sum_{d(i,j) \leq R, j \neq i} e^{-c(i)} \frac{Hv(i)Hv(j)}{d^2(i,j)}, \tag{31}$$

where $c(i)$ represents the structural hole constraint coefficient in Eq. (29). A smaller value of $c(i)$ indicates that the node occupies more structural holes and has a stronger ability to bridge different parts of the network. Finally, the metrics, including HVGC, H-index, HV, DC, and KS, were computed for each node in the toy network and compared with the node’s spreading capability (SC). The results are presented in Table 1, revealing that HVGC achieves a nearly identical ranking to SC, indicating excellent performance. The algorithmic description of the HVGC is provided in Algorithm 1.

In addition, Fig. 2 depicts a network with a clear community structure, where the four nodes with the strongest propagation capabilities are marked in green. The propagation capabilities of these nodes were determined

Node	SC	DC	KS	H-index	HV	HVGC
5	3.28	4	2	3	11	248.02
7	3.10	4	2	3	10	208.09
8	2.83	3	2	3	11	164.73
3	2.80	4	1	1	7	125.38
9	2.65	3	2	2	7	103.72
6	2.34	2	2	2	8	76.91
2	1.80	1	1	1	4	10.30
1	1.78	1	1	1	4	10.30
4	1.75	1	1	1	4	10.30
10	1.73	1	1	1	3	7.72

Table 1. The ranking results of SIR, DC, KS, H-index, HV, and HVGC on the toy network.

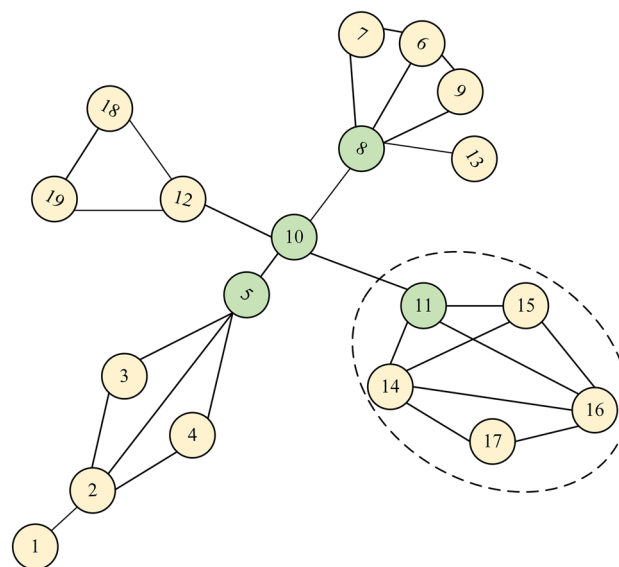


Figure 2. A sample network with an obvious community structure, where the four nodes with the strongest propagation capabilities are marked in green.

Rank	SIR	G+	IGC+	LGM	GGC	KSGC	DKGM	MCGM	SEGM	HVGC
1	10	11	11	10	10	11	11	11	10	10
2	11	14	14	8	8	8	14	16	11	11
3	8	16	16	11	5	10	16	14	14	8
4	5	15	10	14	11	14	10	10	16	5
5	16	10	15	16	2	16	8	15	8	16

Table 2. Comparison of the rankings of the top-5 nodes identified by different methods and the rankings based on the SIR propagation ability in the sample network.

through 1000 independent experiments using the SIR model. We compared HVGC with other gravity model-based methods in identifying the top 5 nodes in this network, and the results are presented in Table 2.

Algorithm 1: The gravity centrality based on the H-index

Input: graph: $G = \langle V, E \rangle$, truncation radius: R , number of nodes: N

Output: $Rank[v, HVGC(v)]$

for $i \leftarrow 1$ to N **do**

 Compute $k(v_i)$ of node v_i ;

 Compute $H(v_i)$ of node v_i ;

 Compute $c(v_i)$ of node v_i using Eq. 29;

end

for $i \leftarrow 1$ to N **do**

 Compute $HV(v_i)$ of node v_i using Eq. 28;

end

for $i \leftarrow 1$ to N **do**

 Find all neighbors of node v_i within the truncation radius R ;

 Compute $HVGC(v_i)$ of node v_i using Eq. 31;

end

return $Rank[v, HVGC(v)]$;

Data description

This paper evaluates the efficacy of HVGC by analyzing ten real networks from six distinct domains, including a transportation network (USAir⁴⁶), an infrastructure network (Power⁴⁷), a communication network (Email⁴⁸), a technology network (Router⁴⁹), two collaborative networks (Jazz⁵⁰ and NS⁵¹), and four social networks (Facebook⁵², PB⁵³, WV⁵⁴, and Sex⁵⁵). Table 3 presents the fundamental topological properties of these networks. N represents the number of nodes in the network, and M represents the number of links. The average degree of nodes is denoted as $\langle k \rangle$, and the average distance between pairs of nodes is denoted as $\langle d \rangle$. The clustering coefficient⁴⁷ of the network is denoted by C , while r represents the assortative coefficient⁵⁶. The degree heterogeneity⁵⁷ of the network is denoted by H . Additionally, β_c represents the epidemic threshold⁵⁸ of the SIR model⁴⁰ used to simulate the diffusion process.

Empirical results

Based on the aforementioned real network, we conducted simulations and compared the influence rankings of various algorithms utilizing the SIR model. In order to ensure the credibility of our findings and the standard ranking of nodes' influence, we conducted 1000 independent experiments for each given network and transmission probability β , with any one node being chosen as the seed node once during each run. The processor and runtime environment used for the calculations are i7-12700H and Python 3. The development platform used for this paper is Anaconda 3, and the code was executed in Jupyter Notebook. Kendall's tau (τ) was utilized to evaluate the accuracy of the algorithms, with a higher value indicating a greater correlation between the observed sequences and an improved algorithm performance. Table 4 provides a comparison of the accuracy of the proposed algorithm (HVGC) and ten benchmark algorithms, which include degree centrality¹⁷ (DC), k-shell decomposition method²⁴ (KS), the extended version of gravity centrality³² (G+), extended version of improved gravity centrality³³ (IGC+), local gravity model³⁴ (LGM), generalized gravity centrality³⁵ (GGC), the improved gravitational centrality based on k-shell values³⁶ (KSGC), the DK-based gravity model³⁷ (DKGM), multi-characteristics gravity model³⁸ (MCGM), and entropy-based gravity model³⁹ (SEGM). Additionally, Fig. 3 displays the accuracy of the different algorithms for varying values of β , within the range of $0.5\beta_c$ to $1.5\beta_c$.

Networks	N	M	$\langle k \rangle$	$\langle d \rangle$	C	r	H	β_c
Jazz	198	2742	27.6970	2.2350	0.6334	0.0202	1.3951	0.0266
NS	379	914	4.8232	6.0419	0.7981	0.0817	1.6630	0.1424
Email	1133	5451	9.6222	3.6060	0.2540	0.0782	1.9421	0.0565
PB	1222	16,714	27.3552	2.7375	0.3600	0.2213	2.9707	0.0125
Facebook	4039	88,234	43.6910	3.6925	0.6170	0.0636	2.4392	0.0095
WV	7066	100,736	28.5129	3.2475	0.2090	0.0833	5.0992	0.0069
Sex	15,810	38,540	4.8754	5.7846	0.0000	0.1145	5.8276	0.0365
USAir	332	2126	12.8072	2.7381	0.7494	0.2079	3.4639	0.0231
Power	4941	6594	2.6691	18.9892	0.1065	0.0035	1.4504	0.3483
Router	5022	6258	2.4922	6.4488	0.0329	0.1384	5.5031	0.0786

Table 3. The topological features of ten real networks.

Networks	DC	KS	G+	IGC+	LGM	GGC	KSGC	DKGM	MCGM	SEGM	HVGC
Jazz	0.8218	0.7536	0.8994	0.9077	0.8725	0.8301	0.8592	0.8843	0.9142	0.9084	<i>0.9184</i>
NS	0.6235	0.5327	0.8168	0.8316	0.8604	0.6604	0.8217	0.8653	0.8671	0.8938	0.8600
Email	0.7747	0.7730	0.9026	0.9038	0.8581	0.8189	0.8387	0.8579	0.9046	0.8997	<i>0.9088</i>
PB	0.8520	0.8563	0.9060	0.9050	0.8975	0.8529	0.8924	0.8990	0.8993	0.9016	<i>0.9063</i>
Facebook	0.7127	0.7369	0.8410	0.8400	0.7926	0.7428	0.7608	0.7994	0.8479	0.8387	<i>0.8494</i>
WV	0.7568	0.7604	0.8203	0.8206	0.8189	0.8031	0.8145	0.8180	0.8263	0.8232	<i>0.8268</i>
Sex	0.4919	0.5133	0.7738	0.7783	0.7782	0.7782	0.7514	0.7890	<i>0.8359</i>	0.8136	0.7980
USAir	0.7488	0.7657	0.8874	0.8901	0.8914	0.8055	0.8505	0.8918	0.8984	0.8971	<i>0.9001</i>
Power	0.5542	0.3786	0.8652	0.8705	0.7953	0.6467	0.7831	0.7938	0.8126	0.8665	0.8590
Router	0.3357	0.1919	0.7479	0.7342	0.7739	0.7694	0.7587	0.7808	<i>0.8035</i>	0.7894	0.7895

Table 4. The algorithms' accuracies for $\beta = \beta_c$, measured by the Kendall's Tau (τ). The top-ranked value in each row of the table is marked in italics, the second in bold.

According to Table 4, the methods that utilise the gravitational formula (G+, IGC+, LGM, GGC, KSGC, DKGM, MCGM, SEGM, and HVGC) exhibit significant advantages over classical methods (DC and KS). These advantages are especially prominent in the Power, Router, NS, and Sex networks. Furthermore, it is noteworthy that among all gravity-based algorithms tested on the ten networks, HVGC exhibited the best overall performance. Its Kendall coefficient ranked first in six out of ten networks, with a remarkable 70% proportion being in the top two ranks. Specifically, HVGC ranked first in the Jazz, email, Facebook, PB, WV, and USAir networks and second in the Router network. Additionally, as shown in Fig. 3, when $\beta = \beta_c$, although HVGC did not perform best in the NS, Power, and Sex networks, as β increases, its performance becomes very close to or even surpasses the previous best-performing algorithm. Taking into account HVGC's superior performance in community-type networks discussed earlier, it demonstrates a stronger overall performance, affirming the robustness of our findings. Furthermore, Fig. 4 displays the optimal truncation radius of HVGC in the ten real networks, revealing that the majority of networks concentrate their optimal truncation radius at $R = 1$. This indicates that HVGC achieves remarkably high accuracy by considering only the influence of the first-order neighbouring nodes of a node, while most other gravity model methods require considering information from second- or third-order neighbouring nodes. In other words, HVGC achieves a high level of accuracy while incurring lower time costs.

Discussion

This paper introduces a novel method called HVGC for identifying influential nodes in a network. While the original gravity model considered both neighbourhood and path information, this new method enhances the existing gravity centrality approaches by taking into account the overall influence of a node's neighbourhood, considering the structural hole position of nodes, and incorporating the differences in interactions between nodes. This method addresses the limitations of existing gravity centrality methods and strengthens the ability to identify important nodes in networks with clear community structures. Therefore, this approach demonstrates a high level of comprehensive performance. We conducted an analysis of the SIR dynamic propagation process in 10 real networks to compare the performance of HVGC with previous state-of-the-art methods. The results, as shown in Table 4, indicate the strong competitiveness of our method.

In certain scenarios, it is necessary to identify the top-k influential nodes for controlling information propagation. Therefore, in addition to evaluating the different ranking methods for individual nodes, we also assessed their performance in identifying the top-k influential spreaders. In other words, we compared the ranked lists of node influence obtained from the ranking methods with the ranked lists of node influence obtained from the SIR simulation, both sorted in descending order. Subsequently, we analysed the similarity between the two lists

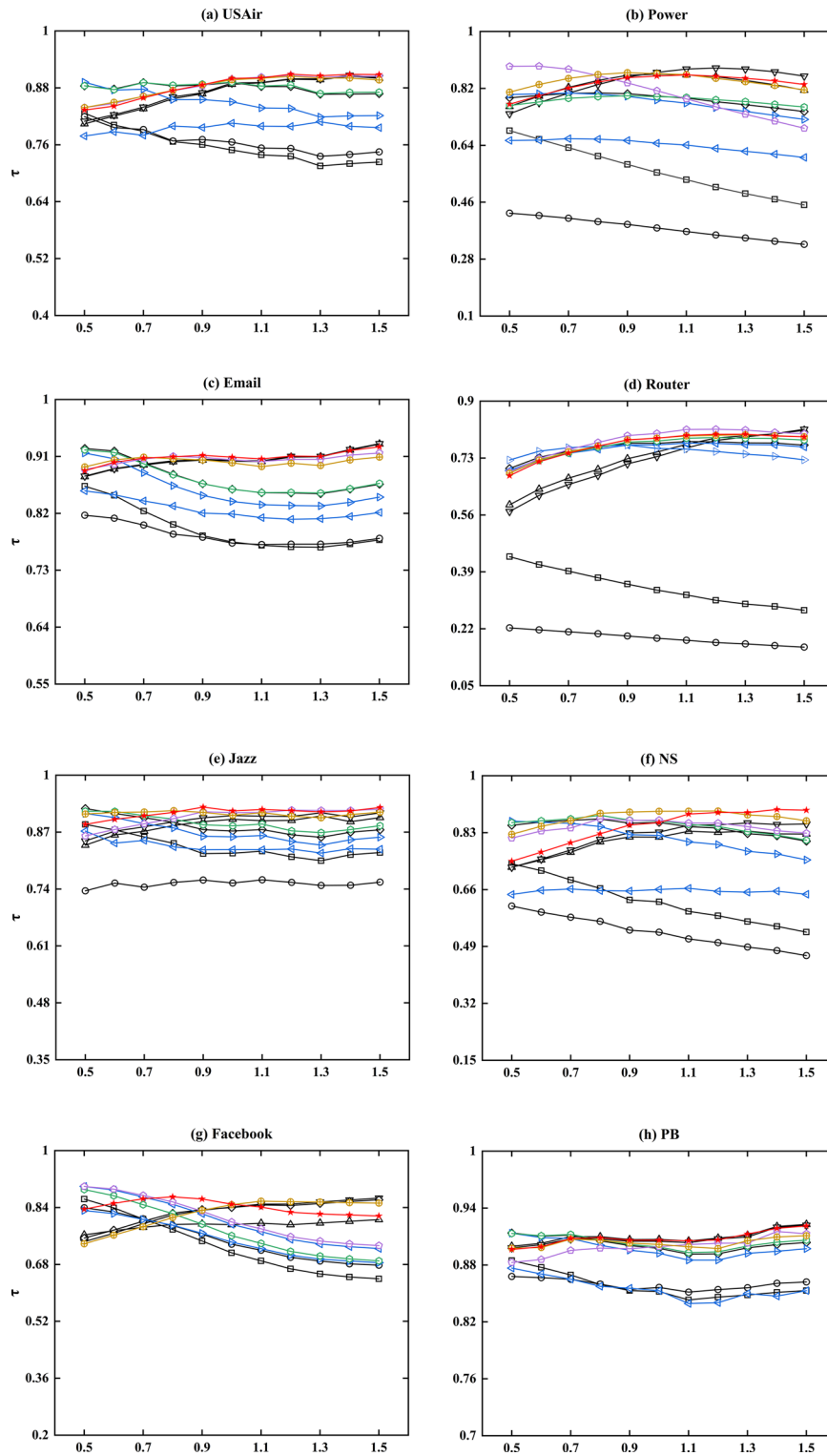


Figure 3. Kendall's Tau was utilized to measure the accuracy of the algorithms at various β values. The different colour symbols represent different methods, and the red symbol represents HVGC algorithms.

by considering the top- k nodes. Figure 5 illustrates the results of the Jaccard coefficient for identifying the top- k influential spreaders, ranging from 5 to 100 with a step size of 5. The X-axis shows the number of top influential spreaders, and the Y-axis shows the Jaccard similarity coefficients.

We can observe that, except for the Sex, Power, PB, and Router networks, HVGC exhibits the best and most stable overall performance in identifying the top- k influential spreaders in other networks. Specifically, across all networks, as the number of selected top- k nodes increases, HVGC consistently maintains a high-level or

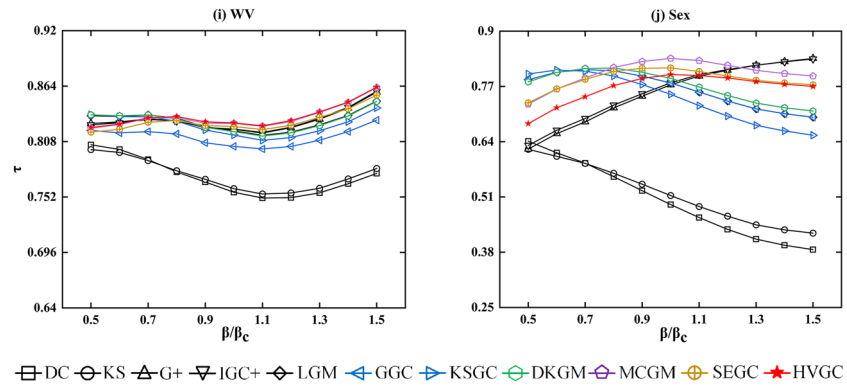


Figure 3. (continued)

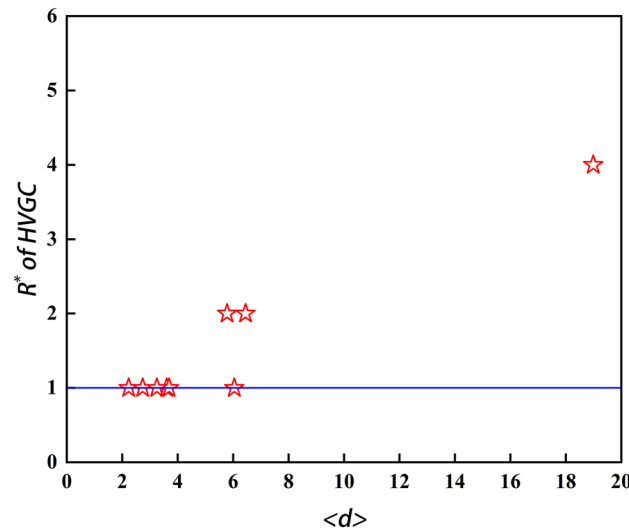


Figure 4. The optimal truncation radius R^* of HVGC at $\beta = \beta_c$ is presented in the graph. Each pentagram in the graph corresponds to a network, with a total of ten networks represented. The blue line corresponds to $R^* = 1$. Specifically, for HVGC, the value of R^* is 1 in Email, Facebook, Jazz, PB, USAir, NS and WV networks, 2 in Router and Sex networks, and 4 in Power network. The majority of the networks have an optimal truncation radius of 1, with the next most common radius being 2. This outcome aligns with the characteristics of domain centrality, which typically considers first-order and second-order neighbor nodes. HVGC represents a significant advancement over the H-index in domain centrality to obtain centrality, which is consistent with this characteristic. However, this does not impede its competitiveness relative to other algorithms, as it manages to achieve both simplicity and accuracy.

steadily increasing Jaccard coefficient, while other methods display varying degrees of fluctuations. Furthermore, we provide detailed plots for the top-25 nodes, revealing that HVGC consistently ranks among the top three in identifying the top-25 influential spreaders and, in some cases, even secures the first position, except for the Sex network. Therefore, we can conclude that HVGC not only accurately ranks the influence of all nodes in the network but also successfully identifies the top-k nodes with the highest impact.

After applying monotonicity²⁵, we assessed the resolution of various algorithms. Table 5 illustrates that HVGC and MCGM demonstrate similar performance in terms of monotonicity. However, HVGC excels in the majority of networks by solely considering the first-order neighbour information of nodes, whereas MCGM, even with the inclusion of second-order neighbour information, does not necessarily outperform HVGC and incurs higher computational complexity. Furthermore, HVGC demonstrates significantly better performance in identifying important nodes in networks with community structure compared to MCGM. Therefore, overall, HVGC surpasses other gravity model algorithms. Based on the results presented in Table 5, HVGC consistently ranks either at the top or very close to the best-performing algorithm in terms of monotonicity.

Based on the above discussion, it is evident that centrality based on the gravitational model is more accurate than classical centrality. However, many of these models tend to identify false core nodes in the network and do not take into account the influence of neighbouring nodes. In our proposed HVGC (H-index-based Gravity Centrality), we address this limitation by comprehensively considering the overall impact of a node’s neighbours and its position within the network’s structural holes. This approach effectively overcomes the drawbacks of gravity-based methods and demonstrates superior performance compared to other algorithms.

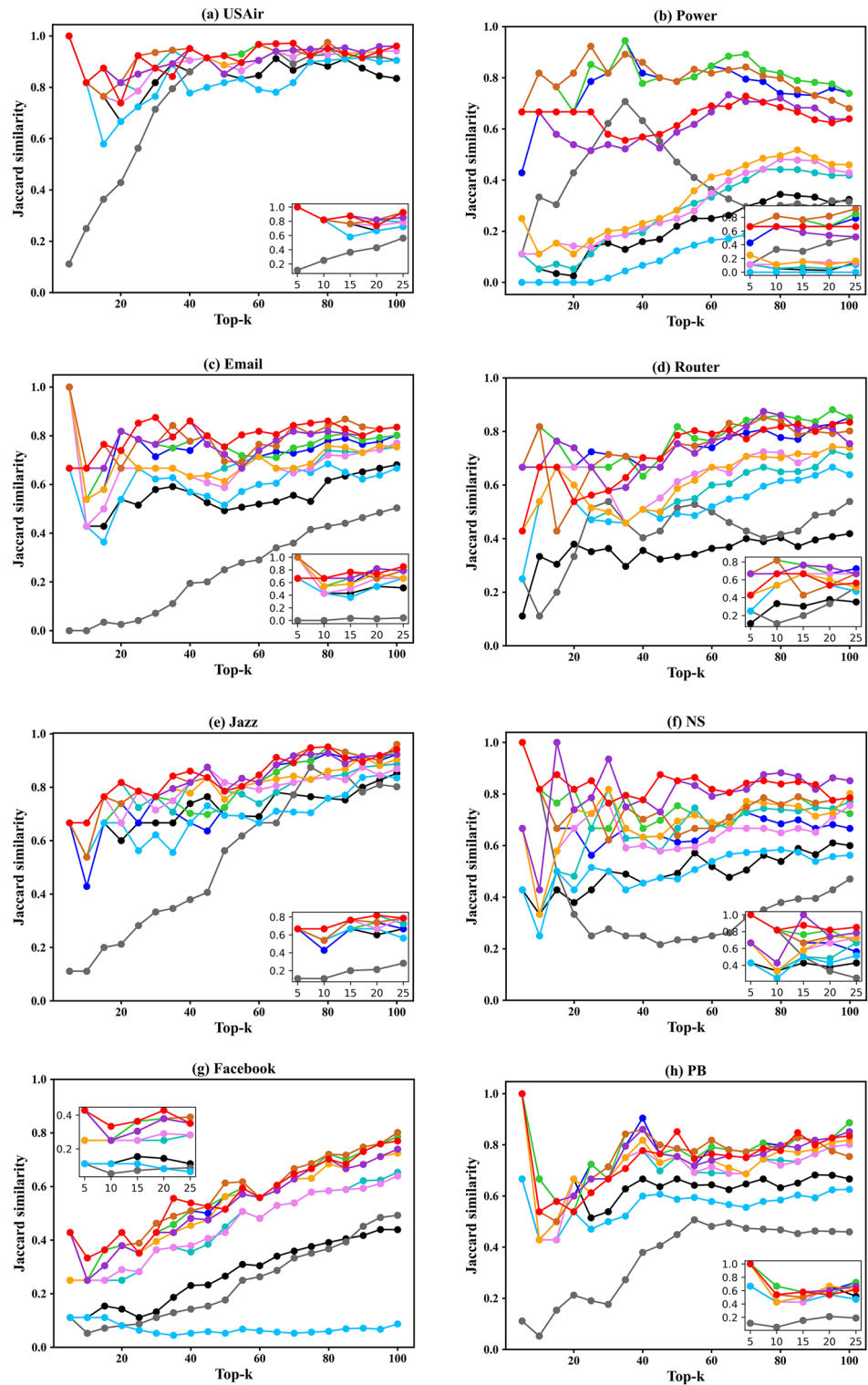


Figure 5. The Jaccard similarity coefficients on the top-*k* influential spreaders.

Despite the excellent performance exhibited by HVGC, it shares a common limitation with other gravity-based methods, namely the need to determine the optimal truncation radius *R*. However, this disadvantage is mitigated by the fact that most real networks exhibit small-world characteristics^{47,59}, and the optimal truncation radius is approximately linearly related to the average distance³⁴. Furthermore, since HVGC is derived from the domain centrality method, even considering only the first-order neighbor nodes in the ten real networks studied can lead to very high performance and accurate results.

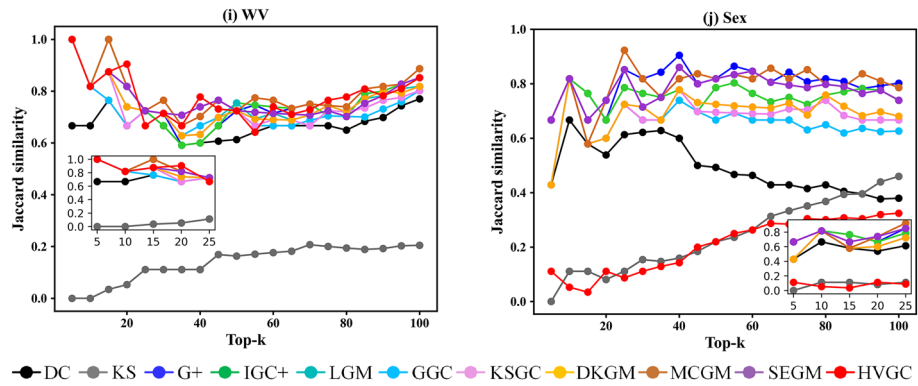


Figure 5. (continued)

Networks	DC	KS	G+	IGC+	LGM	GGC	KSGC	DKGM	MCGM	SEGM	HVGC
Jazz	0.9659	0.7944	0.9993	0.9994	0.9990	0.9992	0.9992	0.9993	0.9994	0.9994	0.9994
NS	0.7642	0.642	0.9951	0.9951	0.9950	0.9953	0.995	0.9951	0.9955	0.9954	0.9955
Email	0.8873	0.8088	0.9999	0.9999	0.9976	0.9997	0.9995	0.9995	0.9999	0.9999	0.9999
PB	0.9328	0.9063	0.9992	0.9992	0.9991	0.9992	0.9992	0.9992	0.9993	0.9993	0.9993
Facebook	0.9739	0.9419	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9999	0.9998	0.9999
WV	0.776	0.7673	0.9996	0.9996	0.9985	0.9996	0.9995	0.9993	0.9996	0.9996	0.9996
Sex	0.6001	0.5287	0.9996	0.9996	0.9989	0.9989	0.9996	0.9997	0.9997	0.9997	0.9997
USAir	0.8585	0.8114	0.9951	0.9951	0.9933	0.995	0.9942	0.9941	0.9951	0.9951	0.9951
Power	0.5926	0.246	0.9996	0.9997	0.9998	0.9998	0.9998	0.9998	0.9999	0.9998	0.9999
Router	0.2886	0.0691	0.9964	0.9964	0.9964	0.9965	0.9965	0.9965	0.9965	0.9966	0.9965

Table 5. Monotonicity of the various algorithms is observed, with the best algorithm for each network highlighted in bold.

In conclusion, while HVGC demonstrates better overall performance compared to other gravity-based methods and introduces improvements to existing gravity models, there are still areas that require further refinement. For example, the current approach does not consider the influence of weight factors associated with different indicators. Instead, it directly operates on the indicator values of the nodes. The weights of HV and the structural hole constraint coefficient $c(i)$ in the computation process may affect the accuracy of the algorithm. In networks with clear community structures, a higher weight for $c(i)$ may lead to better performance, while in other types of networks, a lower weight may yield better results. Therefore, future work may involve incorporating adjustable parameters to balance the weights of different indicators, which is a direction for further exploration. Additionally, these algorithms have not been evaluated in weighted networks, where the impact of the path from node i to node j may differ from that of the path from node j to node i , and the link heterogeneity⁶⁰ in a weighted network may result in varying node impact. Lastly, future research may involve incorporating adjustable parameters to modify the interplay of gravitational forces among nodes and balance the weights of different metrics in order to improve the performance of the algorithm.

Data availability

All relevant data are available at <https://github.com/MLIF/Network-Data>.

Received: 14 March 2023; Accepted: 26 September 2023

Published online: 29 September 2023

References

1. Strogatz, S. H. Exploring complex networks. *Nature* **410**, 268–276 (2001).
2. Newman, M. E. J. The structure and function of complex networks. *SIAM Rev.* **45**, 167–256 (2003).
3. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
4. Zhu, P., Zhi, Q., Guo, Y. & Wang, Z. Analysis of epidemic spreading process in adaptive networks. *IEEE Trans. Circuits Syst. II Express Briefs* **66**, 1252–1256 (2019).
5. Yao, S., Fan, N. & Hu, J. Modeling the spread of infectious diseases through influence maximization. *Optim. Lett.* **16**, 1563–1586 (2022).
6. Albert, R., Albert, I. & Nakarado, G. L. Structural vulnerability of the North American power grid. *Phys. Rev. E* **69**, 025103 (2004).
7. Hosni, A. I. E., Li, K. & Ahmad, S. Minimizing rumor influence in multiplex online social networks based on human individual and social behaviors. *Inf. Sci.* **512**, 1458–1480 (2020).
8. Ahmed, W., Vidal-Alaball, J., Downing, J. & Seguí, F. L. COVID-19 and the 5G conspiracy theory: Social network analysis of Twitter data. *J. Med. Internet Res.* **22**, e19458 (2020).

9. Xu, W. *et al.* Identifying structural hole spanners to maximally block information propagation. *Inf. Sci.* **505**, 100–126 (2019).
10. Csermely, P., Korcsmáros, T., Kiss, H. J. M., London, G. & Nussinov, R. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacol. Ther.* **138**, 333–408 (2013).
11. Liu, Y., Wang, X. & Kurths, J. Framework of evolutionary algorithm for investigation of influential nodes in complex networks. *IEEE Trans. Evol. Comput.* **23**, 1049–1063 (2019).
12. Morone, F. & Makse, H. A. Influence maximization in complex networks through optimal percolation. *Nature* **524**, 65–68 (2015).
13. Sui, L. *et al.* The fractal description model of rock fracture networks characterization. *Chaos Solitons Fractals* **129**, 71–76 (2019).
14. Huang, Y., Dong, H., Zhang, W. & Lu, J. Stability analysis of nonlinear oscillator networks based on the mechanism of cascading failures. *Chaos Solitons Fractals* **128**, 5–15 (2019).
15. Zhao, J. & Deng, Y. Complex network modeling of evidence theory. *IEEE Trans. Fuzzy Syst.* **29**, 3470–3480 (2021).
16. Namtirtha, A., Dutta, A. & Dutta, B. Weighted kshell degree neighborhood: A new method for identifying the influential spreaders from a variety of complex network connectivity structures. *Expert Syst. Appl.* **139**, 112859 (2020).
17. Bonacich, P. Factoring and weighting approaches to status scores and clique identification. *J. Math. Sociol.* **2**, 113–120 (1972).
18. Lü, L., Zhou, T., Zhang, Q.-M. & Stanley, H. E. The H-index of a network node and its relation to degree and coreness. *Nat. Commun.* **7**, 10168 (2016).
19. Zareie, A. & Sheikahmadi, A. EHC: Extended H-index Centrality measure for identification of users' spreading influence in complex networks. *Phys. Stat. Mech. Appl.* **514**, 141–155 (2019).
20. Xu, G.-Q., Meng, L., Tu, D.-Q. & Yang, P.-L. LCH: A local clustering H-index centrality measure for identifying and ranking influential nodes in complex networks. *Chin. Phys. B* **30**, 088901 (2021).
21. Freeman, L. C. Centrality in social networks conceptual clarification in Hawaii Nets conferences. *Cent. Soc. Netw. Concept. Clarification Hawaii Nets Conf.* **1**, 215–239 (1979).
22. Freeman, L. C. A set of measures of centrality based on betweenness. *Sociometry* **40**, 35–41 (1977).
23. Lü, C. *et al.* Identifying Influential Nodes in Complex Networks.pdf (2012).
24. Kitsak, M. *et al.* Identification of influential spreaders in complex networks. *Nat. Phys.* **6**, 888–893 (2010).
25. Bae, J. & Kim, S. Identifying and ranking influential spreaders in complex networks by neighborhood coreness. *Phys. Stat. Mech. Appl.* **395**, 549–559 (2014).
26. Li, C., Wang, L., Sun, S. & Xia, C. Identification of influential spreaders based on classified neighbors in real-world complex networks. *Appl. Math. Comput.* **320**, 512–523 (2018).
27. Wang, Z., Zhao, Y., Xi, J. & Du, C. Fast ranking influential nodes in complex networks using a k-shell iteration factor. *Phys. Stat. Mech. Appl.* **461**, 171–181 (2016).
28. Zeng, A. & Zhang, C.-J. Ranking spreaders by decomposing complex networks. *Phys. Lett. A* **377**, 1031–1035 (2013).
29. Qiu, L., Zhang, J. & Tian, X. Ranking influential nodes in complex networks based on local and global structures. *Appl. Intell.* **51**, 4394–4407 (2021).
30. Ullah, A. *et al.* Identifying vital nodes from local and global perspectives in complex networks. *Expert Syst. Appl.* **186**, 115778 (2021).
31. Yang, Y.-Z., Hu, M. & Huang, T.-Y. Influential nodes identification in complex networks based on global and local information. *Chin. Phys. B* **29**, 088903 (2020).
32. Ma, L., Ma, C., Zhang, H.-F. & Wang, B.-H. Identifying influential spreaders in complex networks based on gravity formula. *Phys. Stat. Mech. Appl.* **451**, 205–212 (2016).
33. Wang, J., Li, C. & Xia, C. Improved centrality indicators to characterize the nodal spreading capability in complex networks. *Appl. Math. Comput.* **334**, 388–400 (2018).
34. Li, Z. *et al.* Identifying influential spreaders by gravity model. *Sci. Rep.* **9**, 8387 (2019).
35. Li, H., Shang, Q. & Deng, Y. A generalized gravity model for influential spreaders identification in complex networks. *Chaos Solitons Fractals* **143**, 110456 (2021).
36. Yang, X. & Xiao, F. An improved gravity model to identify influential nodes in complex networks based on k-shell method. *Knowl. Based Syst.* **227**, 107198 (2021).
37. Li, Z. & Huang, X. Identifying influential spreaders in complex networks by an improved gravity model. *Sci. Rep.* **11**, 22194 (2021).
38. Li, Z. & Huang, X. Identifying influential spreaders by gravity model considering multi-characteristics of nodes. *Sci. Rep.* **12**, 9879 (2022).
39. Liu, Y., Cheng, Z., Li, X. & Wang, Z. An entropy-based gravity model for influential spreaders identification in complex networks. *Complexity* **2023**, e6985650 (2023).
40. Hethcote, H. W. The mathematics of infectious diseases. *SIAM Rev.* **42**, 599–653 (2000).
41. Castellano, C. *et al.* Thresholds for epidemic spreading in networks. *Phys. Rev. Lett.* **105**, 218701. <https://doi.org/10.1103/PhysRevLett.105.218701> (2010).
42. Kendall, M. G. A new measure of rank correlation. *Biometrika* **30**, 81–93 (1938).
43. Zareie, A., Sheikahmadi, A., Jalili, M. & Fasaei, M. S. K. Finding influential nodes in social networks based on neighborhood correlation coefficient. *Knowl. Based Syst.* **194**, 105580 (2020).
44. Liu, Y., Tang, M., Zhou, T. & Do, Y. Core-like groups result in invalidation of identifying super-spreader by k-shell decomposition. *Sci. Rep.* **5**, 9602 (2015).
45. Lu, M. Node importance evaluation based on neighborhood structure hole and improved TOPSIS. *Comput. Netw.* **178**, 107336 (2020).
46. Pajek Datasets. <http://vlado.fmf.uni-lj.si/pub/networks/data/>.
47. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
48. Guimerà, R., Danon, L., Diaz-Guilera, A., Giralt, F. & Arenas, A. Self-similar community structure in a network of human interactions. *Phys. Rev. E* **68**, 065103 (2003).
49. Spring, N., Mahajan, R., Wetherall, D. & Anderson, T. Measuring ISP topologies with Rocketfuel. *IEEEACM Trans. Netw.* **12**, 2–16 (2004).
50. Gleiser, P. M. & Danon, L. Community structure in jazz. *Adv. Complex Syst.* **06**, 565–573 (2003).
51. Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **74**, 036104 (2006).
52. Leskovec, J. & McAuley, J. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems*. Vol. 25 (Curran Associates, Inc., 2012).
53. Adamic, L. A. & Glance, N. The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*. 36–43 (Association for Computing Machinery, 2005). <https://doi.org/10.1145/1134277>.
54. Leskovec, J., Huttenlocher, D. & Kleinberg, J. Predicting positive and negative links in online social networks. In *Proceedings of the 19th International Conference on World Wide Web*. 641–650 (Association for Computing Machinery, 2010). <https://doi.org/10.1145/1772690.1772756>.
55. Rocha, L. E. C., Liljeros, F. & Holme, P. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLOS Comput. Biol.* **7**, e1001109 (2011).
56. Newman, M. E. J. Assortative mixing in networks. *Phys. Rev. Lett.* **89**, 208701 (2002).

57. Hu, H.-B. & Wang, X.-F. Unified index to quantifying heterogeneity of complex networks. *Phys. Stat. Mech. Appl.* **387**, 3769–3780 (2008).
58. Castellano, C. & Pastor-Satorras, R. Thresholds for epidemic spreading in networks. *Phys. Rev. Lett.* **105**, 218701 (2010).
59. Amaral, L. A. N., Scala, A., Barthélémy, M. & Stanley, H. E. Classes of small-world networks. *Proc. Natl. Acad. Sci.* **97**, 11149–11152 (2000).
60. Bellingeri, M., Bevacqua, D., Scotognella, F. & Cassi, D. The heterogeneity in link weights may decrease the robustness of real-world complex weighted networks. *Sci. Rep.* **9**, 10692 (2019).

Author contributions

S.Z. devised the research project. S.Z. performed the research. S.Z. and J.Z. analyzed the data. S.Z. and X.L. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.Z. or J.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023