Chaos, Solitons and Fractals 186 (2024) 115227

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

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

Influential nodes identification based on hierarchical structure

Longyun Wang [a], Jianhong Mou [a], Bitao Dai [a], Suoyi Tan [a], Mengsi Cai [a], Huan Chen [a], Zhen Jin [b], Guiquan Sun [b,c], Xin Lu [a,*]

[a] College of Systems Engineering, National University of Defense Technology, Changsha, 410073, Hunan, China
[b] Complex Systems Research Center, Shanxi University, Taiyuan, 030006, Shanxi, China
[c] Department of Mathematics, North University of China, Taiyuan, 030051, Shanxi, China

ARTICLE INFO

Keywords:
Complex networks
Influential spreaders
Hierarchical structure
Ranking method

ABSTRACT

Identifying influential nodes is an important research topic in complex network analysis, with significant implications for understanding and controlling propagation processes. While extant methods for assessing node influence rely heavily on network topology, often overlooking the dynamic interactions and propagation patterns within networks. In this paper, we propose the Hierarchical Structure Influence (HSI) method. The HSI method evaluates the potential outbreak size of nodes by modeling their infection sequences and paths according to a network's hierarchical structure, and integrating propagation probabilities to estimate these outbreak sizes accurately. It accounts for infections occurring across different node layers, intra-layer, and heterogeneous infection routes of varying lengths. To validate its effectiveness, HSI is compared with seven state-of-the-art methods across nine real-world networks. Experimental results reveal that HSI outperforms other methods in terms of ranking accuracy, top-k nodes, and distinguishing ability. Furthermore, HSI exhibits high consistency in evaluating node outbreak sizes when compared to SIR simulations. Our method offers valuable insights that can be leveraged for network management and the development of intervention strategies.

1. Introduction

Complex networks are widely used to depict relationships between various entities, such as social networks [1,2], biological networks [3,4], transportation networks [5,6], and power networks [7,8]. In these networks, a small number of special nodes, namely influential nodes, significantly influence the structure and function of the network. Evaluating and ranking nodes' influence is a vital aspect of research in spreading dynamics. This approach could locate influential nodes of complex networks, enhancing the understanding and controlling propagation processes within the networks, including viral marketing [9,10], epidemic prevention [11–13], rumor control [14], and information dissemination [15].

In recent years, extensive research has been conducted from various perspectives to identify influential nodes in complex networks. A considerable number of methods are based on topological structures, such as degree centrality (DC) [16], neighbors' degrees centrality (ND) [17], betweenness centrality (BC) [18], closeness centrality (CC) [19], and K-shell centrality [20]. Degree centrality focuses only on the number of connecting nodes without considering the global structure. Betweenness centrality and closeness centrality are global measures, but their high computational complexity limits their application in large-scale networks. K-shell centrality is also a global index that enables the quick ranking of nodes in large-scale networks. However, its main drawback is that it may assign the same K-shell value to many nodes, leading to monotonicity issues in ranking. To address this problem, several efforts have been made to finely distinguish the spreading capability of influential nodes, such as extended neighborhood coreness (CNC+) [21], K-shell iteration factor (KSIF) [22], and mixed degree decomposition (MDD) [23]. Other approaches, like local structural centrality (LSC) [24] and entropy-based ranking measure (ERM) [25], extend the evaluation of a node's influence beyond its immediate neighbors. Furthermore, Li et al. [26] proposed a gravity model (GM) [26] that considers neighborhood information and path information to identify influential nodes. Mou et al. [27] proposed a spindle vector based on hierarchical structure to capture the relative order of nodes in diffusion propagation, effectively approximating the spatiotemporal evolution of diffusion dynamics on networks. Yang et al. [28] considered the positional attributes of nodes and proposed an improved gravity model based on the K-shell algorithm (KSGC). Ullah et al. [29] approached the identification of influential nodes from the perspective of local and global topological features and designed a new algorithm (LGC) using measurements such as degree centrality and shortest path.

* Corresponding author.
E-mail address: xin.lu.lab@outlook.com (X. Lu).

https://doi.org/10.1016/j.chaos.2024.115227
Received 13 May 2024; Received in revised form 19 June 2024; Accepted 28 June 2024
Available online 6 July 2024

While these algorithms have yielded promising results in identifying influential nodes, most algorithms only consider the topological characteristics of nodes, which inadequately describe their involvement in information and virus propagation dynamics. Recent research has shown that the propagation influence of nodes is also affected by the dynamics of the spreading process [30,31]. The ranking of node influence will change with the variation of dynamic parameters, while centrality algorithms based on topological structure maintain the same ranking under different dynamic parameters [32]. Šikić et al. [30] argued that in the susceptible–infected–recovered (SIR) model [33], the influence of nodes largely depends on the spreading rate and recovering rate. By incorporating the dynamic propagation characteristics of nodes in a network, the performance and accuracy of identifying influential nodes have been improved to some extent [34]. For example, Liu et al. [30] combine spreading rate, recovery rate, and limited-time steps to find influential nodes based on the susceptible–infected–recovered (SIR) model. Chen et al. [35] introduced the concept of path diversity and considered that the influence of nodes is related to the number of information dissemination paths and path diversity. Xu et al. [36] proposed an algorithm based on the local propagation probability model, which measures the influence of nodes through the aggregated scores of neighbor nodes within a three-level neighborhood. Lin et al. [32] combine Markov chains with propagation processes and propose a dynamic Markov process (DMP) method to estimate the outbreak size of the initial spreader. Ai et al. [37] argue that the propagation ability of source nodes is measured by the probability of uninfected nodes being directly or indirectly infected by the source, and propose a centrality method based on node spreading probability (SPC). However, this method only considers the distance between nodes and source nodes, as well as the propagation probability, without considering the diversity of propagation paths.

To overcome the above limitations, this paper proposes the Hierarchical Structure Influence (HSI) method, aimed at evaluating the outbreak size and identifying influential nodes in the network. We consider the propagation process as a hierarchical infection, where low-order neighbors of source nodes infect higher-order neighbors layer by layer. The propagation depth is influenced by the infection of different-order neighbors, while the propagation width is affected by the infection of neighboring nodes within the same layer. And the propagation probability determines the final infection scale. Based on the hierarchical structure, we consider the comprehensive impact of nodes being infected by multiple paths with different lengths and calculate the probability of nodes at different layers being infected by source nodes separately. Then, a node's influence is determined by summing up the infected probabilities of all nodes in the network.

The rest of the paper is structured as follows: Section 2 introduces the hierarchical structure and its relationship with propagation, and describes the HSI method. Section 3 gives the datasets, comparative algorithms, and evaluation metrics used in the experiments. Section 4 presents the experimental results and analysis. Finally, the conclusion and some future recommendations of our study are shown in Section 5.

2. Methodology

2.1. Hierarchical structure and network spreading

The hierarchical structure of a root node delineates its immediate neighborhood and higher-order neighborhoods extending up to the maximum distance. Given an undirected and unweighted network $G(V, E)$, where V represents the nodes in the network and E represents the edges. For any source node u in network G, traversing all nodes in the network using breadth-first search (BFS) [38] generates a hierarchical structure centered on node u. The hierarchical structure's first layer includes node u's first-order neighbors, with subsequent layers containing higher-order neighbors. For a node u in graph G, the hierarchical structure is defined as

$$L(u) = \{l_u^1, l_u^2, l_u^3, \ldots, l_u^K\}, \tag{1}$$

where l_u^i denotes the ith order neighbors of node u, $i \in [1, 2, 3, \ldots, K]$. Here, $K < D$ represents the depth of the hierarchical structure, which is constrained to be less than the maximum permissible number of orders, i.e., the network diameter D.

Table 1
The hierarchical structure of nodes in the toy network.

Source node	Layer 1	Layer 2	Layer 3	Layer 4	Node counts
0	1, 2, 4	3, 5, 6	7		3, 3, 1
1	0, 2, 3	4	5, 6	7	3, 1, 2, 1
2	0, 1, 3, 4	5, 6	7		4, 2, 1
3	1, 2, 4	0, 5, 6	7		3, 3, 1
4	0, 2, 3, 5, 6	1, 7			5, 2
5	4, 6	0, 2, 3, 7	1		2, 4, 1
6	4, 5, 7	0, 2, 3	1		3, 3, 1
7	6	4, 5	0, 2, 3	1	1, 2, 3, 1

Consider a toy network depicted in Fig. 1. The hierarchical structure of node 2 reveals layer-wise node counts as $\{4, 2, 1\}$. Similarly, the hierarchical structures for all nodes are detailed in Table 1. Node 4, occupying a central position within the network, features two layers, with the first layer containing the highest number of nodes. Conversely, Node 7, located at the network's periphery, spans four layers, with the node count initially increasing and then decreasing across these layers. The contrasting hierarchical structures of central and peripheral nodes suggest that centrally located nodes are characterized by a greater number of nodes within fewer, smaller layers. This hierarchical distribution effectively encapsulates the positional information of nodes within the network, thereby serving as a critical attribute for measuring node importance.

The spread of infection from a single node is intricately associated with its hierarchical structure. Using the SIR model for illustration, consider an extreme case where both the infection and recovery rates are set to 1. Under such conditions, the infection propagation mimics a BFS process. At the time step t, the tth order neighbors become fully infected, the $(t-1)$-th order neighbors shift to the recovered state, and nodes beyond the tth order remain susceptible. The propagation unfolds layer by layer, akin to the ripple effect of water waves, a dynamic effectively represented by the hierarchical structure.

When the infection rate falls below 1, the spread does not strictly follow the hierarchical layers. However, the general direction of transmission still progresses from lower order layers to higher order layers. This pattern emerges because an infection in higher order neighbors implies that the lower order neighbors must have served as a conduit for the spread. The source node cannot directly infect nodes in the higher-order layers that are not directly connected. In the shortest path from the source node to any infected node, each layer traversed must contain at least one infected node. Consequently, the hierarchical structure effectively illustrates the sequence and pathways through which the infection spreads, providing a robust framework for modeling the propagation dynamics.

2.2. Description of the proposed HSI method

Building on the hierarchical structure, we model the infection propagation as proceeding the transmission from lower order layers to higher order layers. To appropriately simplify the intricate propagation process for estimating node influence, we consider two spreading scenarios for the first infection time of each node: direct infection from nodes in the preceding layer and indirect infection via nodes within the same layer. For simplicity, this assumption neglects the "backward" infection from higher order neighbors to lower order neighbors. This

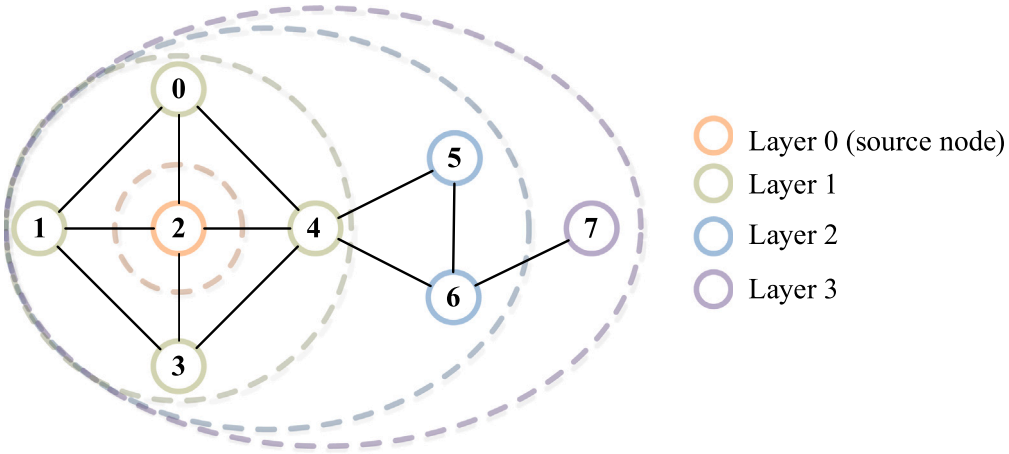

Fig. 1. The hierarchical structure of node 2 in the toy network. The dashed circles divide the nodes of each layer. The first layer consists of {0, 1, 3, 4}, the second layer consists of {5, 6}, and the third layer consists of {7}.

premise allows us to calculate the infection probabilities for nodes across each layer.

(1) Calculate the propagation influence between different layers.

In the hierarchical structure, the source node u can infect its first order neighbors, each second order neighbor node will infect its third order neighbor nodes, and so forth. Here, we only consider the scenario where adjacent layers are infected by lower-order neighbors. For a node v in layer i, it may have multiple neighboring nodes in layer $(i-1)$, all of which may infect node v. The infection of node v by multiple nodes is a nonlinear coupling process, and its probability is given by

$$P_v^1 = 1 - \prod_{q \in \Gamma(v) \cap l_u^{i-1}} \left(1 - P_q \cdot P_{qv}\right), \tag{2}$$

where $\Gamma(v)$ denotes the set of neighbor nodes of node v, $l_u^{(i-1)}$ denotes the $(i-1)$-th order neighbors of node u, P_q denotes the probability that node q is in an infected state, P_{qv} denotes the probability that node q infects its neighbor node v.

If node v is in the first layer, then its only neighboring node in the lower-order layer is node u, and the corresponding infection probability is $P_v^1 = 1 - (1 - P_u P_{uv}) = P_{uv}$.

(2) Calculate the propagation influence in same layer.

In addition to lower-order neighbors infecting higher-order neighbors, there is also infection among nodes within the same layer. The more neighbors a node has in the same layer, the higher the likelihood of it being infected. If the node v in layer i is uninfected state and is infected by multiple nodes in the same layer, the corresponding infection probability is

$$P_v^2 = 1 - \prod_{q \in \Gamma(v) \cap l_u^{i}} \left(1 - P_q^1 \cdot P_{qv}\right), \tag{3}$$

where P_q^1 is the probability of node q being infected by nodes in layer $(i-1)$, i.e., the propagation influence between different layers.

(3) Calculate the multi-step mixed propagation influence.

The node v may be infected by lower-order neighboring nodes as well as by nodes within the same layer. By considering these two scenarios together, the total probability of node v being in an infected state can be obtained as

$$P_v = P_v^1 + \left(1 - P_v^1\right) P_v^2, \tag{4}$$

where P_v^1 is the probability of infection from the lower-order layer, and P_v^2 is the probability of infection from the same layer.

In reality, infections by nodes within the same layer can be viewed as indirect infections by lower-order neighbors through two steps. This means that Eq. (4) calculates the probability of node v infected by lower-order neighbors via multiple paths of one or two steps. The infection path originating from the source node u may traverse 1 or 2 nodes per layer, resulting in 2^m possible combinations for m layers. The shortest path length is m steps, while the longest is $2m$ steps. Therefore, for a node in layer m, it is as if the source node u infects it through multiple paths spanning from m to $2m$ steps. Consider the toy network shown in Fig. 1, where node 7 is in layer 3. Theoretically, the propagation path from the source node 2 to node 7 should have $2^3 = 8$ possible combinations. However, since node 7 has no neighboring nodes in the same layer, there is no indirect infection in layer 3. This results in 2^2 actual path combinations, yielding 6 distinct propagation paths (one path of length 3, three paths of length 4, and two paths of length 5), as shown in Fig. 2.

(4) Calculate the Hierarchical Structure Influence.

Based on Eqs. (2) to (4), the probability of each node being in an infected state can be calculated layer by layer. By summing the probabilities of all nodes being in an infected state, the Hierarchical Structure Influence (HSI) of the source node u can be determined:

$$HSI(u) = \sum_{v \in V} P_v \tag{5}$$

2.3. Computational complexity of HSI

The Hierarchical Structure Influence (HSI) method involves constructing a hierarchical structure using a Breadth-First Search (BFS) process, which has a time complexity of $O(N+M)$, where $N = |V|$ corresponds to the number of nodes, and $M = |E|$ denotes the number of edges in the network. When calculating the propagation influence between different layers P_v^1, we need to compute it for each node, where each node has an average of d neighbors. This results in a complexity of $O(N \times d)$ for traversing each node and its neighbors. Similarly, for calculating the propagation influence in same layer P_v^2, we also traverse each node and its neighbors, resulting in a time complexity of $O(N \times d)$. When updating P_v for each node, the time complexity is $O(N)$. Finally, computing the HSI of the source node requires summing the values for all nodes, which also has a time complexity of $O(N)$. Combining these analyses, the total time complexity of the HSI method is $O(N+M) + 2 \times O(N \times d) + 2 \times O(N)$. Since for most graphs, the total number of edges M is approximately equal to the product of the number of nodes N and the average degree d ($M \approx N \times d$), the overall time complexity simplifies to $O(N+M)$.

Notably, the computational complexity of HSI is significantly lower than that of prevalent global centrality metrics, such as betweenness centrality and closeness centrality, which manifest complexities of $O(MN^3)$ and $O(MN^2)$, respectively.

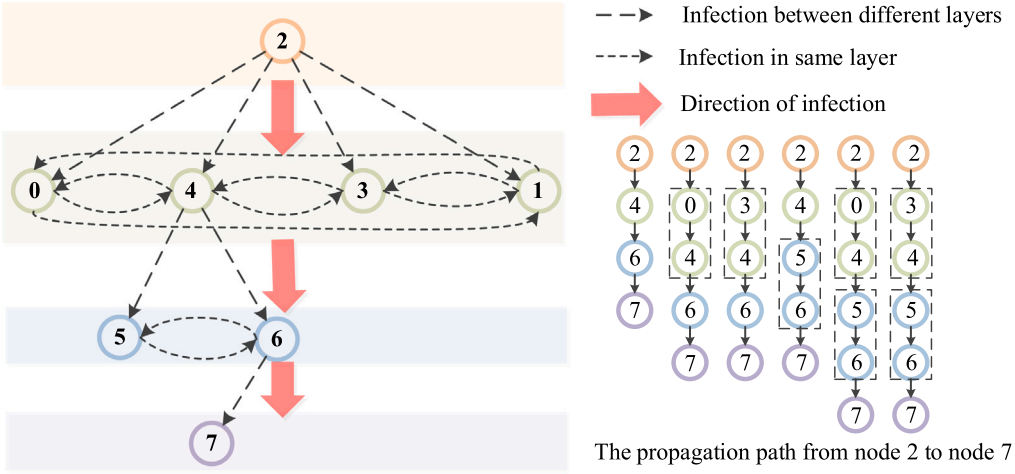

Fig. 2. The multi-step mixed infection paths of toy network.

Table 2
Statistical attributes of the real-world networks.$\langle k\rangle$, $K_{\max}$, $\langle Q\rangle$, $\langle P\rangle$, $\langle C\rangle$ and $\langle L\rangle$ denote the average degree, maximum degree, modularity, density, clustering coefficient and average shortest path length of the network, respectively. β_c represents the epidemic threshold and is defined as $\beta_c = \frac{\langle k\rangle}{\langle k^2\rangle - \langle k\rangle}$, where $\langle k^2\rangle$ denotes the second-order degree of the network.

Network	$\lvert V\rvert$	$\lvert E\rvert$	$\langle k\rangle$	$K_{\max}$	$\langle Q\rangle$	$\langle P\rangle$	$\langle C\rangle$	$\langle L\rangle$	β_c
NS	1,461	2,742	3.8	34	0.959	0.003	0.694	–	0.168
GrQC	5,242	14,484	5.5	81	0.863	0.001	0.530	–	0.063
Figeys	2,239	6,432	5.7	314	0.463	0.003	0.040	–	0.018
Facebook	4,039	8,8234	43.7	1045	0.835	0.011	0.606	3.69	0.009
Hamster	2,426	16,631	13.7	273	0.563	0.006	0.538	–	0.024
LastFM	7,624	27,806	7.3	216	0.8130	0.001	0.219	5.23	0.041
PowerGrid	4,941	6,594	2.7	19	0.935	0.001	0.080	18.99	0.348
Vidal	3,023	6,149	4.1	129	0.644	0.001	0.064	–	0.069
Sex	1,6730	3,9044	4.7	305	0.487	0.001	0.007	4.20	0.032

Note:An empty $\langle L\rangle$ indicates that the network is a non-connected graph.

3. Experiments

3.1. Data description

To rigorously assess the efficacy of the proposed algorithm, this study encompasses 9 real-world networks spanning domains such as social, scientific collaboration, proteomics, and power distribution systems. The networks are briefly introduced as follows: (1) NS [39]: A collaboration network among scientists researching network theory and experimentation. (2) GrQC [40]: A research collaboration network among authors in General Relativity and Quantum Cosmology. (3) Figeys [41]: A human protein interaction network derived from the first large-scale study on protein–protein interactions in human cells, utilizing mass spectrometry. (4) Facebook [42]: An ego-network on Facebook, featuring users and their friends. (5) Hamster [43]: A network from hamsterster.com, capturing friendships and family relationships among its users. (6) LastFM [44]: A social network of LastFM users in Asia. (7) PowerGrid [45]: A network detailing the electricity distribution system in the western United States. (8) Vidal [46]: A network mapping binary protein–protein interactions in the human proteome. (9) Sex [47]: A bipartite web community centered on sexual activity. All networks are undirected and unweighted, with their statistical properties summarized in Table 2.

3.2. Algorithms for comparison

To illustrate the performance of our proposed method (HSI), we used ND [17], GM [26], ERM [25], KSIF [22], KSGC [28], LGC [29], SPC [37] and DMP [32] as the comparison algorithms. They are briefly described as follows.

Neighbors' degrees centrality (ND) [17]: An approach considers the degree of a vertex's neighbors.

Gravity model(GM) [26]: A gravity model considering neighborhood information and path information. GM deems the more influential node with more immediate nodes and shorter average distances.

Entropy-based ranking measure (ERM) [25]: A method to measure the spreading capability of nodes based on node entropy centrality, as well as the entropy centrality of their neighbors and second-order neighbors.

K-shell iteration factor (KSIF) [22]: KSIF utilizes iterative information based on k-shell decomposition to differentiate the influence of nodes with the same k-shell index.

Improved Gravity Centrality based on the K-Shell algorithm (KSGC) [28]: An improved gravity model algorithm considers the position information of nodes to measure the interactions between them.

Local-and-Global-Centrality (LGC) [29]: A method for identifying influential nodes based on degree centrality and shortest distance, considering both local and global topological characteristics.

A centrality method based on node spreading probability (SPC) [37]: SPC measures the infectivity of source nodes based on the probability of uninfected nodes being directly or indirectly infected by the source nodes.

Dynamic Markov process method (DMP) [32]: A dynamic Markov process (DMP) method by integrating the Markov chain and the spreading process to evaluate the outbreak size of the initial spreader.

3.3. The spreading model

We employ the SIR model to delineate the influence of each node [48]. The SIR model categorizes nodes into three states: susceptible, infected, and recovered. At each step, nodes in the susceptible state may transition to the infected state. Infected nodes attempt to infect susceptible neighbors with a probability β. Subsequently, infected nodes may recover and transition to the recovered state with a probability γ, after which they are considered immune to reinfection. The process concludes either upon reaching the maximum time step or when no infected nodes remain in the network. To estimate the influence of an individual node, we iteratively initialize each node as infectious while the remaining nodes are susceptible, and the epidemic propagation size is adopted as the node's spreading influence.

In this experiment, the β/β_c ratio is considered to range from 0.2 to 1.8, with nodes recovering in the subsequent time step post-infection, i.e., $\gamma = 1$. Given the stochastic nature of simulation outcomes, the node's spreading influence is determined by averaging 1000 simulation iterations.

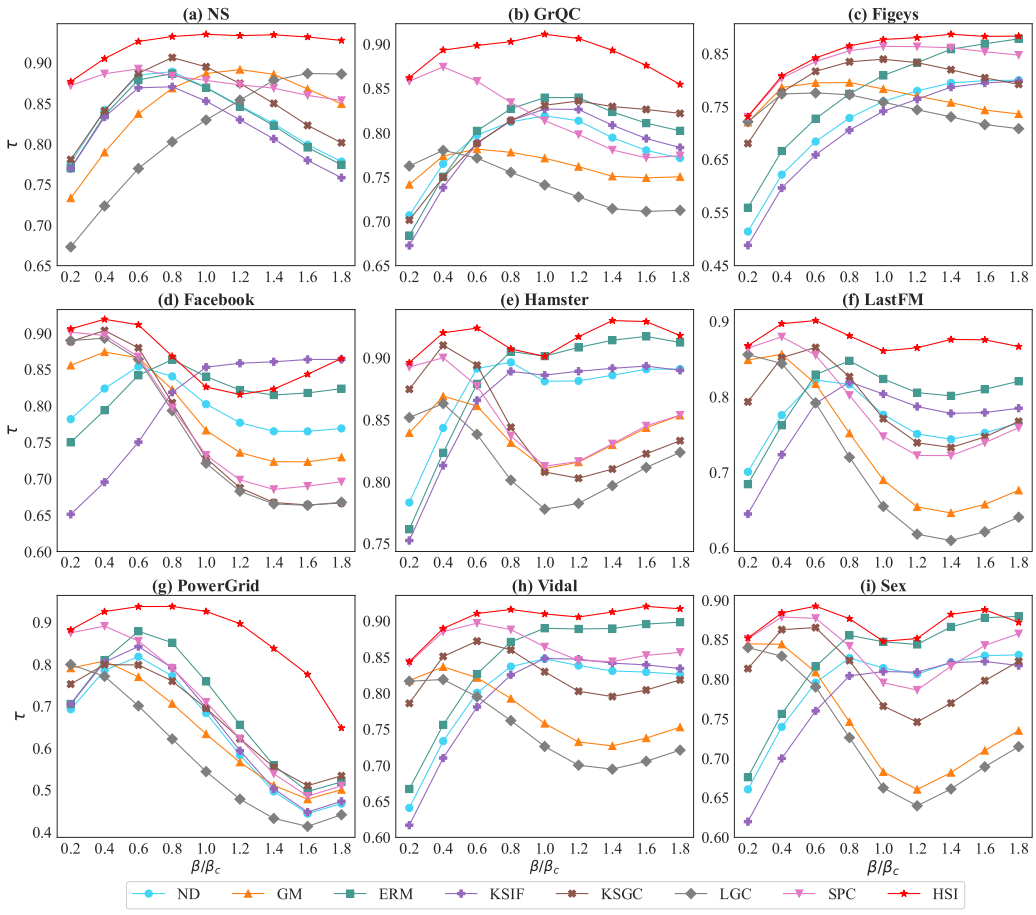

Fig. 3. The Kendall's τ coefficient between different ranking methods and SIR with different infection probabilities.

3.4. Evaluation metrics

3.4.1. Kendall correlation coefficient

The Kendall's τ coefficient [49] quantifies the congruence between two ranking sequences of identical length, serving as a prevalent metric for assessing the precision of influential spreader identification algorithms. Assuming there are two ranked sequences A and B, each containing n elements. (A_i, B_i) denotes the ith element-pair of A and B. For any element-pair (A_i, B_i) and (A_j, B_j), if $(A_i - A_j)(B_i - B_j) > 0$, they are regarded as a concordant pair. On the contrary, they are regarded as a discordant pair. Then the Kendall's τ coefficient is defined as

$$\tau = \frac{2\left(n_c - n_d\right)}{n(n-1)}, \tag{6}$$

where n_c and n_d denote the number of concordant pairs and discordant pairs, and n denotes the total number of elements of the ranked sequence. The range of Kendall's τ lies between [0,1], and a larger τ corresponds to better performance.

3.4.2. Ranking monotonicity

Ranking monotonicity [21] is employed to assess the discriminatory capacity of various methods regarding node influence, calculated as follows,

$$M(R) = \left(1 - \frac{\sum_{r \in R} n_r \left(n_r - 1\right)}{n(n-1)}\right)^2 \tag{7}$$

where R denotes the ranking list of network nodes, n denotes the number of nodes, and n_r denotes the number of nodes with the same rank r. The range of $M(R)$ lies between $[0, 1]$, with values closer to 1 indicating better discrimination. A value of 0 means that all nodes have been assigned the same rank, and the ranking list R is invalid.

To illustrate the distribution of ranking values, we also employ the complementary cumulative distribution function (CCDF) to compare the discriminative power of different methods.

$$\mathrm{CCDF}(Z) = \mathrm{Prob}(Z > z) = 1 - \mathrm{CDF}(Z), \tag{8}$$

where $\mathrm{CDF}(Z)$ represents the cumulative distribution function, indicating the probability that a node's rank is equal to or less than a specific value z, i.e., $\mathrm{CDF}(Z) = \mathrm{Prob}(Z < z)$. In general, if the CCDF of a ranking method rapidly drops to zero in a small ranking range, then the method suffers from a serious monotonicity problem.

3.4.3. Jaccard similarity coefficient

The Jaccard similarity coefficient is employed to evaluate the similarity of the top-k most influential nodes in two ranking lists. Given two ranked sequences A and B, Jaccard similarity is defined as

$$\text{Jaccard } @k(A, B) = \frac{\left|A_k \cap B_k\right|}{\left|A_k \cup B_k\right|}, \tag{9}$$

where A_k and B_k represent the sets containing the top-k nodes from lists A and B, respectively. The Jaccard similarity ranges from [0,1], with values closer to 1 indicating a higher degree of overlap between two lists.

4. Results

4.1. Nodal spreading influence ranking

To validate the enhanced accuracy of the HSI method, a comparative analysis of ranking accuracy on 9 real networks was conducted

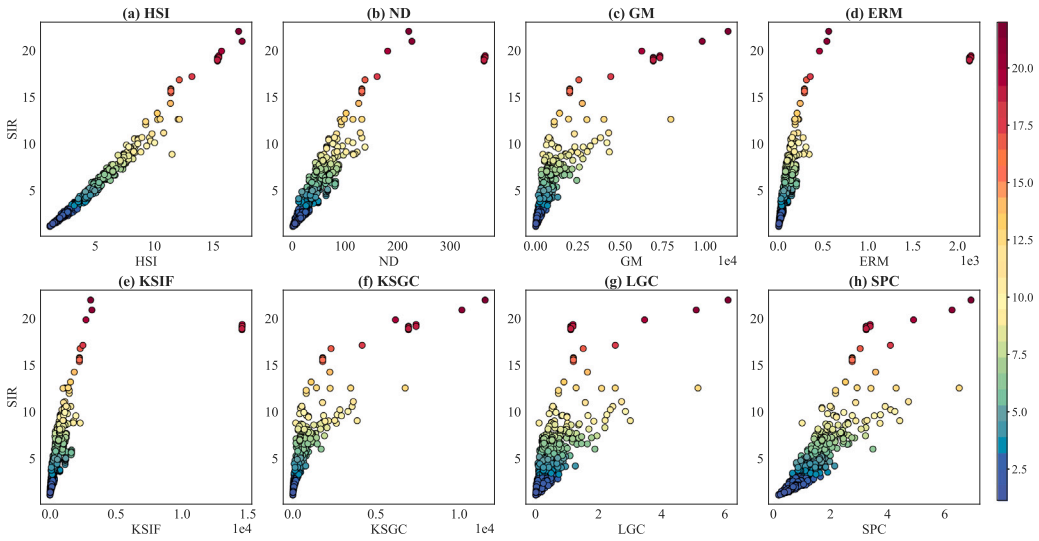

Fig. 4. The correlation between the results simulated by SIR and the results measured by different ranking methods on NS. The spreading rate is set as $\beta = \beta_c$.

Table 3
The mean Kendall's τ coefficient of different ranking methods under varying infection probabilities.

Network	HSI	ND	GM	ERM	KSIF	KSGC	LGC	SPC	Improvement (%)
NS	**0.9225**	0.8346	0.8456	0.8310	0.8188	0.8510	0.8115	0.8742	5.5
GrQC	**0.8888**	0.7845	0.7621	0.7978	0.7835	0.7998	0.7418	0.8182	8.6
Figeys	**0.8504**	0.7201	0.7651	0.7747	0.7037	0.7996	0.7445	0.8347	1.9
Facebook	**0.8644**	0.7978	0.7886	0.8188	0.8018	0.7656	0.7602	0.7740	5.6
Hamster	**0.9160**	0.8717	0.8394	0.8804	0.8634	0.8444	0.8163	0.8518	4.0
LastFM	**0.8764**	0.7669	0.7330	0.7981	0.7678	0.7881	0.7062	0.7877	9.8
PowerGrid	**0.8628**	0.6381	0.6405	0.6927	0.6508	0.6696	0.5780	0.6975	23.7
Vidal	**0.9030**	0.7983	0.7751	0.8428	0.7938	0.8245	0.7490	0.8640	4.5
Sex	**0.8717**	0.7917	0.7460	0.8243	0.7737	0.8075	0.7281	0.8385	4.0
$\mu(\tau)$	**0.8840**	0.7782	0.7662	0.8067	0.7730	0.7944	0.7373	0.8156	8.4

Note: Improvement=$\frac{\tau_H - \tau_\theta}{\tau_\theta}$, where τ_H denotes the Kendall's τ coefficient of the HSI method, and τ_θ represents the maximum Kendall's τ coefficient in the baseline method. $\mu(\tau)$ represents the mean Kendall's τ coefficient of a method on all networks.

using Kendall's τ coefficient. As depicted in Fig. 3, HSI consistently outperforms baseline methods in most cases with the infection rate β/β_c varies from 0.2 to 1.8. Specifically, HSI has the highest Kendall's τ coefficient for all infect probabilities on NS, GrQC, Figeys, Hamster, LastFM, PowerGrid and Vidal networks. And HSI achieves the best performance on Facebook when $\beta/\beta_c \leq 0.8$ and Sex when $\beta/\beta_c \leq 1.6$. The τ value of HSI on all network mostly stays above 0.85, indicating a high concordance between HSI's rankings and the ones generated by the simulation result. On Facebook, when β/β_c exceeds 0.8, the performance of HSI is weaker than the baseline method. This is because the nodes in the Facebook network are densely connected, and "backward" infections have a more pronounced impact at higher infection probabilities. The baseline method shows significant differences across different networks, while HSI exhibits relative stability, demonstrating better performance under varying infection probabilities.

Table 3 presents the mean Kendall's τ coefficient of different ranking methods under varying infection probabilities. It can be found that HSI exhibits the best performance, showing improvements from 1.9% to 23.7% over the baseline methods, with an average improvement of 8.4%. This indicates that HSI can identify and rank influential nodes more accurately. The improvement of PowerGrid is the largest, the highest mean Kendall's τ coefficient of the baseline method is 0.6975, and the HSI can reach 0.8682. This is due to the sparse nature of the PowerGrid network, where HSI is more effective in identifying the nodes' influence compared to other methods.

Taking the NS network as an example, we compared the correlation between SIR-simulated results and the indices measured by various ranking methods, as shown in Fig. 4. One can see that there exists a strong correlation between the outcomes of the HSI method and the real node influential capability simulated by SIR. The data points align almost linearly, demonstrating a favorable monotonicity. In contrast, the distribution of data points for other methods is more scattered. And it is worth noting that a group of nodes with significantly higher ND values also exhibit much higher ERM and KSIF values compared to other nodes (as shown in Fig. 4(b), (d), (e)), although their actual influence is not the greatest. The HSI method accurately evaluates the influence of this group of nodes. We also check the performance of the HSI method and other methods by Pearson correlation coefficient r, as shown in Table 4. It can be observed that HSI has the highest r, all exceeding 0.95, with an average improvement of 8.5%, indicating a strong linear correlation between HSI and SIR.

4.2. Evaluating the spreading influence of node

Beyond ranking nodes' spreading capabilities, the HSI method skillfully evaluates the spreading influence of node, i.e., the outbreak size. In contrast to other methods that only compare the relative magnitude of spreading capabilities, HSI can provide the outbreak size of a node. There is a strong linear relationship between the HSI value of nodes and the results of SIR simulations, as shown in Figs. 5-(a), (e), and (i). As the infection probability increases, the outbreak size of node also expands. And HSI can represent how the outbreak size of node varies with changes in infection probability.

Table 4
The Pearson correlation coefficient r between different methods and SIR.

Network	HSI	ND	GM	ERM	KSIF	KSGC	LGC	SPC	Improvement (%)
NS	**0.9868**	0.8207	0.8047	0.6712	0.6652	0.7736	0.7248	0.8851	11.5
GrQC	**0.9619**	0.9161	0.9086	0.9088	0.8993	0.8671	0.8681	0.8578	5.0
Figeys	**0.9932**	0.6812	0.9084	0.9421	0.6406	0.8258	0.9128	0.9761	1.7
Facebook	**0.9662**	0.9510	0.7813	0.9538	0.9428	0.6671	0.7712	0.8220	1.3
Hamster	**0.9895**	0.9721	0.8856	0.9686	0.9656	0.7351	0.9171	0.9573	1.8
LastFM	**0.9730**	0.9366	0.8210	0.9379	0.9261	0.7324	0.8321	0.8811	3.7
PowerGrid	**0.9578**	0.7872	0.7368	0.7376	0.7490	0.7403	0.6754	0.7983	20.0
Vidal	**0.9899**	0.9250	0.8476	0.9593	0.9058	0.7677	0.8647	0.9691	2.1
Sex	**0.9849**	0.9294	0.8671	0.9622	0.9071	0.7757	0.8869	0.9698	1.6
$\mu(r)$	**0.9781**	0.8799	0.8401	0.8935	0.8446	0.7650	0.8281	0.9019	8.5

Note: Improvement $= \frac{r_H - r_\theta}{r_\theta}$, where r_H denotes the Pearson correlation coefficient of the HSI method, and r_θ represents the maximum Pearson correlation coefficient in the baseline method. $\mu(r)$ represents the mean Pearson correlation coefficient r of a method on all networks.

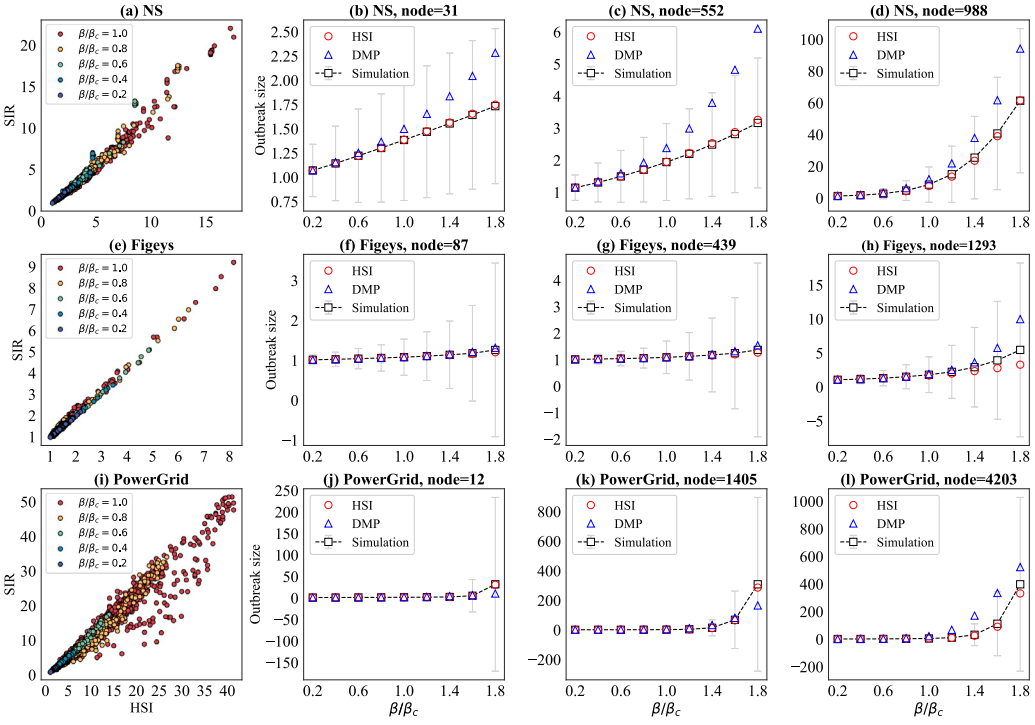

Fig. 5. The performance of the HSI method for evaluating the outbreak of different initial nodes on NS, Figeys and Powergrid. The symmetric bars indicate the fluctuations around the average value computed on 10^5 simulations of the SIR.

We selected 3 nodes with distinct infective capacities from each of the three networks: NS, Figeys, and Powergrid, totaling 9 nodes. The outbreak size obtained from HSI, DMP and SIR simulations under various infection rates β for these nodes is illustrated in Fig. 5. The results indicate that HSI could evaluate the outbreak size more accurate than DMP. One could also observe that as the infection probability increases, the outbreak size of node does not increase linearly. And HSI has the ability to capture this trend. Since HSI method does not rely on extensive simulations, the predictions are free from randomness interference and are more time-efficient.

We further analyzed the correlations for the PowerGrid, Figeys, and NS networks with β/β_c ranging from 0.2 to 1.0, and calculated the errors and determination coefficients between HSI and SIR simulation results, as shown in Fig. 6. The results demonstrate a very strong correlation between HSI and SIR. For the same network, as the infection probability increases, both root mean square error (RMSE) and mean absolute error (MAE) continuously increase. However, even when $\beta/\beta_c = 1.0$, the errors are still relatively small, with a determination coefficient R^2 greater than 0.94. In the NS network, when β/β_c ranges from 0.4 to 0.8, HSI tends to underestimate the infection capability of a few high-infection-capability nodes, but the estimates for the majority of nodes remain accurate.

4.3. Top influential spreaders identification

In most scenarios, it is more important to identify the most influential nodes instead of focusing on the correctness of all nodes' ranks. Therefore, we evaluated the performance of different methods in identifying the top-k% influential spreaders using the Jaccard similarity coefficient. As shown in Fig. 7, the HSI method achieved the best performance on all networks. Its superiority was especially evident in sparse networks with lower average degrees, such as PowerGrid and NS. Other methods showed reasonable performance in networks with higher average degrees like Facebook and Hamster, yet they remained inferior to HSI. This suggests that their effectiveness in identifying influential nodes is significantly affected by the network's density.

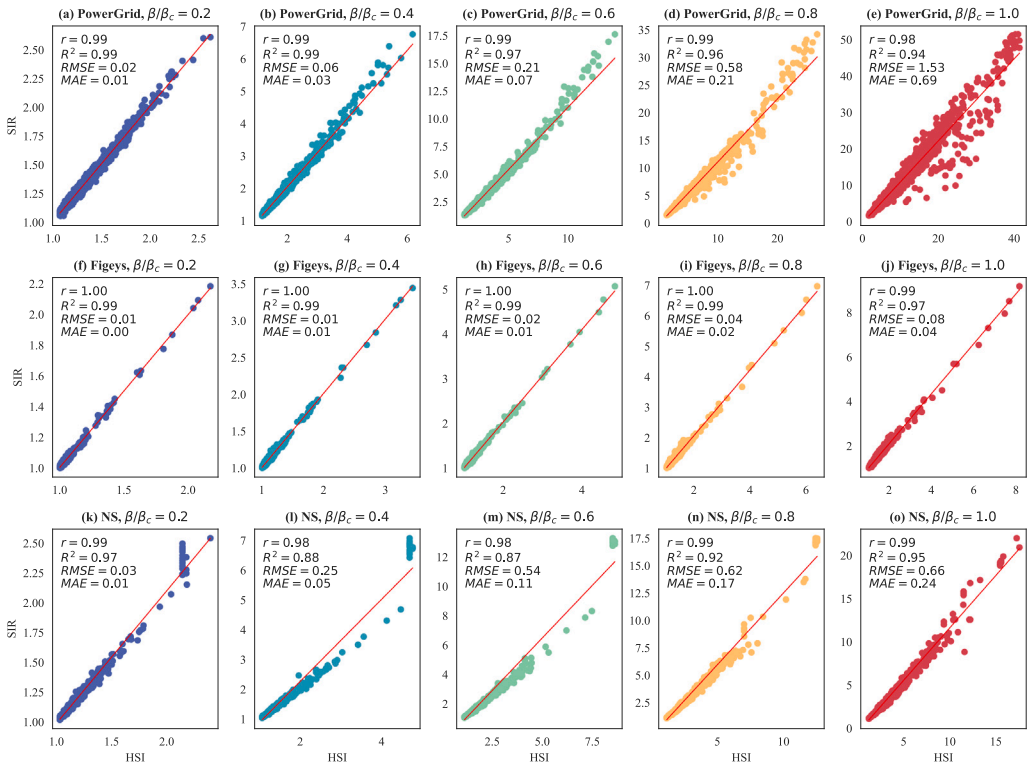

Fig. 6. The correlations between HSI and SIR simulation results on PowerGrid, Figeys, and NS networks. The red solid lines represent the linear fitting results. The Pearson correlation coefficient (r), coefficient of determination (R^2), root mean squared error ($RMSE$), and mean absolute error (MAE) between the HSI and SIR simulation results are indicated in the upper left corner.

Table 5
The monotonicity results of different ranking methods.

Network	ND	GM	ERM	KSIF	KSGC	LGC	SPC	HSI
NS	0.8966	0.9166	0.9095	0.9163	0.9171	0.9172	0.9094	**0.9176**
GrQC	0.9649	0.9871	0.9821	0.9854	0.9872	0.9872	0.9858	**0.9873**
Figeys	0.9887	0.9938	0.9939	0.9934	0.9929	**0.9946**	0.9938	0.9940
Facebook	0.9995	**0.9999**	**0.9999**	**0.9999**	**0.9999**	**0.9999**	0.9972	**0.9999**
Hamster	0.9823	0.9857	0.9849	0.9856	0.9853	0.9857	0.9854	**0.9860**
LastFM	0.9887	**0.9999**	**0.9999**	0.9984	0.9998	**0.9999**	0.9998	**0.9999**
PowerGrid	0.8866	**0.9999**	**0.9999**	0.9805	**0.9999**	**0.9999**	0.9945	**0.9999**
Vidal	0.9713	**0.9923**	0.9799	0.9882	**0.9923**	**0.9923**	0.9922	**0.9923**
Sex	0.9938	**0.9999**	**0.9999**	0.9995	**0.9999**	**0.9999**	**0.9999**	**0.9999**
$\mu(M)$	0.9636	0.9861	0.9833	0.9830	0.9860	**0.9863**	0.9842	**0.9863**

Note: $\mu(M)$ represents the mean monotonicity of a method on all networks.

The experimental results confirm that the HSI method can effectively identify high-influence nodes in networks, and it exhibits more stable performance.

4.4. Monotonicity analysis

Good monotonicity implies that a ranking method can differentiate the influence of all nodes and assign them distinct values as much as possible. We compared the monotonicity results of different methods across various networks, as shown in Table 5. HSI consistently exhibits the best monotonicity on 8 of 9 networks in the experiments, indicating that our method can finely distinguish the influence of nodes. To visually demonstrate and compare the ranking value distributions of different methods, we select the PowerGrid network with the smallest average degree and the Facebook network with the largest average degree in the experiments, as representatives. As shown in Fig. 8, it can be observed that HSI performs the best on both networks, with the CCDF curve declining most slowly. Additionally, the rate of decline in the CCDF for ND and KSIF shows notable differences between the PowerGrid and Facebook networks. This indicates that the network density has a significant impact on the monotonicity of the methods. However, HSI is almost unaffected by network density. The above experiments demonstrate that HSI exhibits excellent monotonicity, effectively distinguishing the influence of nodes.

5. Conclusion and discussion

In this study, we introduce the Hierarchical Structure Influence (HSI) method, an innovative approach leveraging the hierarchical structure of nodes within networks to effectively identify influential nodes. Diverging from methods that depend exclusively on network topology, HSI factors in the nodes' positions, propagation paths, and propagation probabilities, providing a more nuanced and accurate evaluation of nodal influence. Extensive experiments on nine real-world networks demonstrate that HSI outperforms other state-of-the-art methods. Compared to the best baseline method, HSI achieved an average improvement of 8.4% in Kendall's τ coefficient and an average improvement of 8.5% in Pearson correlation coefficient. HSI not only ranks and evaluates nodal influence with high accuracy but also performs well in identifying top-k% influential nodes and maintaining ranking monotonicity. In addition, it is expected that by incorporating the omitted terms for "backward" infection, further performance improvements can be obtained.

The practical implications of our method are manifold. By providing a reliable tool for identifying key spreaders, the HSI method can help optimize resource allocation for information dissemination campaigns or epidemic containment efforts. Moreover, its efficacy across various types of networks makes it a versatile tool for a wide range of applications, from social media analytics to public health planning. Future research can extend this framework to dynamic networks and integrate it with other analytical frameworks, aiming to boost its predictive

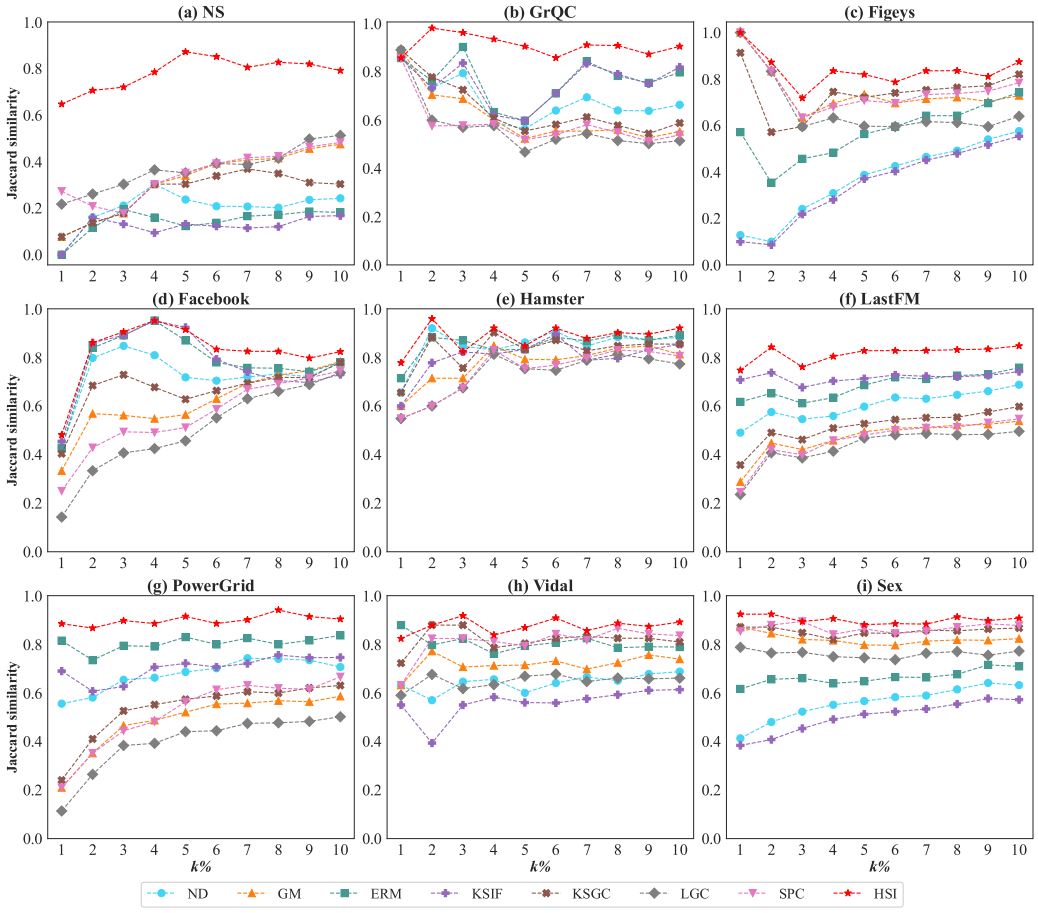

Fig. 7. The Jaccard similarity coefficients on the top-k% influential spreaders.

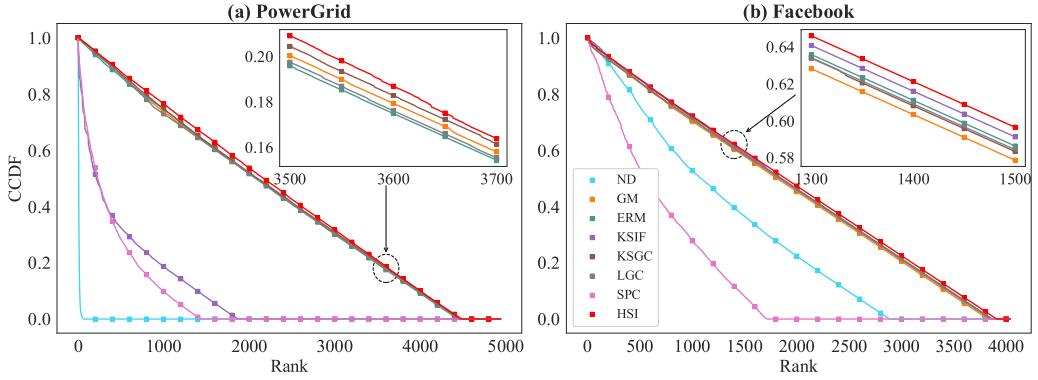

Fig. 8. The complementary cumulative distribution of the node influential ranks.

capabilities and applicability across a broader range of networked systems.

CRediT authorship contribution statement

Longyun Wang: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Conceptualization. **Jianhong Mou:** Writing – review & editing, Methodology, Formal analysis. **Bitao Dai:** Writing – review & editing. **Suoyi Tan:** Writing – review & editing. **Mengsi Cai:** Writing – review & editing. **Huan Chen:** Writing – review & editing, Data curation. **Zhen Jin:** Writing – review & editing. **Guiquan Sun:** Writing – review & editing. **Xin Lu:** Writing – review & editing, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 72025405, 72088101, 72001211, 72301285),

the National Social Science Foundation of China (22ZDA102), and the Natural Science Foundation of Hunan Province (2023JJ40685).

References

[1] Arularasan A, Suresh A, Seerangan K. Identification and classification of best spreader in the domain of interest over the social networks. Cluster Comput 2019;22(Suppl 2):4035–45.

[2] Aral S, Walker D. Identifying influential and susceptible members of social networks. Science 2012;337(6092):337–41.

[3] Cannistraci CV, Alanis-Lobato G, Ravasi T. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. Sci Rep 2013;3(1):1613.

[4] Ahmed H, Howton T, Sun Y, Weinberger N, Belkhadir Y, Mukhtar MS. Network biology discovers pathogen contact points in host protein-protein interactomes. Nature Commun 2018;9(1):2312.

[5] Waniek M, Raman G, AlShebli B, Peng JC-H, Rahwan T. Traffic networks are vulnerable to disinformation attacks. Sci Rep 2021;11(1):5329.

[6] Sugishita K, Masuda N. Recurrence in the evolution of air transport networks. Sci Rep 2021;11(1):5514.

[7] Li M, Lü L, Deng Y, Hu M-B, Wang H, Medo M, et al. History-dependent percolation on multiplex networks. Natl Sci Rev 2020;7(8):1296–305.

[8] Boers N, Goswami B, Rheinwalt A, Bookhagen B, Hoskins B, Kurths J. Complex networks reveal global pattern of extreme-rainfall teleconnections. Nature 2019;566(7744):373–7.

[9] Anastasia B, Marcos G, Colin F C. The golden age of social science. Proc Natl Acad Sci USA 2021;118(5):e2002923118. http://dx.doi.org/10.1073/pnas.2002923118.

[10] Huang H, Shen H, Meng Z, Chang H, He H. Community-based influence maximization for viral marketing. Appl Intell 2019;49(6):2137–50. http://dx.doi.org/10.1007/s10489-018-1387-8.

[11] Yao S, Fan N, Hu J. Modeling the spread of infectious diseases through influence maximization. Optim Lett 2022;16(5):1563–86. http://dx.doi.org/10.1007/s11590-022-01853-1.

[12] Chen B-L, Jiang W-X, Yu Y-T, Zhou L, Tessone CJ. Graph embedding based ant colony optimization for negative influence propagation suppression under cost constraints. Swarm Evol Comput 2022;72:101102.

[13] Schadt EE. Molecular networks as sensors and drivers of common human diseases. Nature 2009;461(7261):218–23.

[14] Ni Q, Guo J, Huang C, Wu W. Community-based rumor blocking maximization in social networks: Algorithms and analysis. Theor Comput Sci 2020;840:257–69. http://dx.doi.org/10.1016/j.tcs.2020.08.030.

[15] Ou Y, Guo Q, Xing J-L, Liu J-G. Identification of spreading influence nodes via multi-level structural attributes based on the graph convolutional network. Expert Syst Appl 2022;203:117515. http://dx.doi.org/10.1016/j.eswa.2022.117515.

[16] Bonacich P. Factoring and weighting approaches to status scores and clique identification. J Math Sociol 1972;2:113–20.

[17] Namtirtha A, Dutta A, Dutta B, Sundararajan A, Simmhan YL. Best influential spreaders identification using network global structural properties. Sci Rep 2021;11.

[18] Freeman LC. A set of measures of centrality based on betweenness. Sociometry 1977;35–41.

[19] Albert R, Barabási A-L. Statistical mechanics of complex networks. Rev Modern Phys 2002;74(1):47.

[20] Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, et al. Identification of influential spreaders in complex networks. Nature Phys 2010;6(11):888–93.

[21] Bae J, Kim S. Identifying and ranking influential spreaders in complex networks by neighborhood coreness. Phys A 2014;395:549–59.

[22] Wang Z, Zhao Y, Xi J, Du C. Fast ranking influential nodes in complex networks using a k-shell iteration factor. Physica A 2016;461:171–81.

[23] Zeng A, Zhang C-J. Ranking spreaders by decomposing complex networks. Phys Lett A 2013;377(14):1031–5.

[24] Gao S, Ma J, Chen Z, Wang G, Xing C. Ranking the spreading ability of nodes in complex networks based on local structure. Phys A 2014;403:130–47.

[25] Zareie A, Sheikhahmadi A, Fatemi A. Influential nodes ranking in complex networks: An entropy-based approach. Chaos Solitons Fractals 2017;104:485–94.

[26] Li Z, Ren T, Ma X, Liu S, Zhang Y, Zhou T. Identifying influential spreaders by gravity model. Sci Rep 2019;9(1):8387.

[27] Mou J, Dai B, Tan S, Holme P, Lehmann S, Liljeros F, et al. The spindle approximation of network epidemiological modeling. New J Phys 2024;26(4):043027.

[28] Yang X, Xiao F. An improved gravity model to identify influential nodes in complex networks based on k-shell method. Knowl-Based Syst 2021;227:107198.

[29] Ullah A, Wang B, Sheng J, Long J, Khan N, Sun Z. Identifying vital nodes from local and global perspectives in complex networks. Expert Syst Appl 2021;186:115778.

[30] Šikić M, Lančić A, Antulov-Fantulin N, Štefančić H. Epidemic centrality—is there an underestimated epidemic impact of network peripheral nodes? Eur Phys J B 2013;86:1–13.

[31] Liu J-G, Lin J-H, Guo Q, Zhou T. Locating influential nodes via dynamics-sensitive centrality. Sci Rep 2016;6(1):21380.

[32] Lin J, Chen B-L, Yang Z, Liu J-G, Tessone CJ. Rank the spreading influence of nodes using dynamic Markov process. New J Phys 2023;25(2):023014.

[33] Hethcote HW. The mathematics of infectious diseases. SIAM Review 2000;42(4):599–653.

[34] Radicchi F, Castellano C. Leveraging percolation theory to single out influential spreaders in networks. Phys Rev E 2016;93(6):062314.

[35] Chen D-B, Xiao R, Zeng A, Zhang Y-C. Path diversity improves the identification of influential spreaders. Europhys Lett 2014;104(6):68006.

[36] Xu G, Meng L. A novel algorithm for identifying influential nodes in complex networks based on local propagation probability model. Chaos Solitons Fractals 2023;168:113155.

[37] Ai J, He T, Su Z, Shang L. Identifying influential nodes in complex networks based on spreading probability. Chaos Solitons Fractals 2022;164:112627.

[38] Moore EF. The shortest path through a maze. In: Proc. of the international symposium on the theory of switching. Harvard University Press; p. 285–92.

[39] Newman ME. Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 2006;74(3):036104.

[40] Leskovec J, Kleinberg JM, Faloutsos C. Graph evolution: Densification and shrinking diameters. ACM Trans Knowl Discov Data 2006;1:2.

[41] Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, et al. Large-scale mapping of human protein–protein interactions by mass spectrometry. Mol Syst Biol 2007;3:89.

[42] Leskovec J, Mcauley J. Learning to discover social circles in ego networks. Adv Neural Inf Process Syst 2012;25.

[43] Kunegis J. KONECT: the koblenz network collection. In: Proceedings of the 22nd international conference on World Wide Web. 2013.

[44] Rozemberczki B, Sarkar R. Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models. In: Proceedings of the 29th ACM international conference on information & knowledge management. 2020.

[45] Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature 1998;393:440–2.

[46] Rual J-F, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, et al. Towards a proteome-scale map of the human protein–protein interaction network. Nature 2005;437:1173–8.

[47] Rocha LEC, Liljeros F, Holme P. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. PLoS Comput Biol 2010;7.

[48] Newman ME. Spread of epidemic disease on networks. Phys Rev E Stat Nonlinear Soft Matter Phys 2002;66(1):016128.

[49] Kendall MG. A new measure of rank correlation. Biometrika 1938;30:81–93.