



PAPER

OPEN ACCESS


RECEIVED
4 August 2023REVISED
31 October 2023ACCEPTED FOR PUBLICATION
21 November 2023PUBLISHED
1 December 2023

Original content from
this work may be used
under the terms of the
[Creative Commons
Attribution 4.0 licence](#).

Any further distribution
of this work must
maintain attribution to
the author(s) and the title
of the work, journal
citation and DOI.



Identifying key spreaders in complex networks based on local clustering coefficient and structural hole information

Hao Wang^{1,6}, Jian Wang^{2,6}, Qian Liu³ , Shuang-ping Yang¹, Jun-jie Wen⁴ and Na Zhao^{1,5,*}¹ Key Laboratory in Software Engineering of Yunnan Province, Yunnan University, Kunming, Yunnan 650091, People's Republic of China² Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650504, People's Republic of China³ School of Economics and Management, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong 518055, People's Republic of China⁴ Innovation & Digital Center, Yunnan Tin Industry Group (Holdings) Co, Kunming, Yunnan 650000, People's Republic of China⁵ The Key Laboratory for Crop Production and Smart Agriculture of Yunnan Province, Kunming, Yunnan 650201, People's Republic of China⁶ These authors contributed to the work equally and should be regarded as co-first authors.

* Author to whom any correspondence should be addressed.

E-mail: zhaonayx@126.com**Keywords:** complex network, clustering coefficient, structural hole, key spreader

Abstract

Identifying key spreaders in a network is one of the fundamental problems in the field of complex network research, and accurately identifying influential propagators in a network holds significant practical implications. In recent years, numerous effective methods have been proposed and widely applied. However, many of these methods still have certain limitations. For instance, some methods rely solely on the global position information of nodes to assess their propagation influence, disregarding local node information. Additionally, certain methods do not consider clustering coefficients, which are essential attributes of nodes. Inspired by the quality formula, this paper introduces a method called Structural Neighborhood Centrality (SNC) that takes into account the neighborhood information of nodes. SNC measures the propagation power of nodes based on first and second-order neighborhood degrees, local clustering coefficients, structural hole constraints, and other information, resulting in higher accuracy. A series of pertinent experiments conducted on 12 real-world datasets demonstrate that, in terms of accuracy, SNC outperforms methods like CycleRatio and KSGC. Additionally, SNC demonstrates heightened monotonicity, enabling it to distinguish subtle differences between nodes. Furthermore, when it comes to identifying the most influential Top-k nodes, SNC also displays superior capabilities compared to the aforementioned methods. Finally, we conduct a detailed analysis of SNC and discuss its advantages and limitations.

1. Introduction

The advancements in network science have propelled the analysis of complex networks into a prominent and burgeoning question within academia [1–6]. Within this realm of study, the identification of key nodes has persistently remained a crucial and foundational issue. Given the heterogeneity of network structures, it often transpires that a small number of nodes exert significant influence within a network, and certain nodes can even wield substantial impact on the structure and functionality of the network [7]. For instance, in networks modeling the spread of infectious diseases, the phenomenon of ‘super-spreaders’ is prevalent. These super-spreaders possess the capability to infect a majority of individuals in an exceptionally short span, thus augmenting the challenges of disease control and prevention. Rapidly pinpointing such ‘super-spreaders’ could potentially truncate the transmission pathways of the disease. Presently, various facets of life involve the issue of identifying key nodes, such as disease prevention and control [8], safeguarding vulnerable nodes in trade networks [9], and curtailing the propagation of rumors in online

social networks [10]. Consequently, the efficient and accurate localization of key nodes within networks constitutes a matter of significant practical importance.

Currently, a multitude of methods for identifying crucial nodes in complex networks have been proposed [7, 11], and researchers have introduced some novel approaches in recent years [12–14]. However, existing methods often neglect to consider the clustering coefficient of nodes, which is an essential attribute [15]. Additionally, the majority of methods are often designed based on two principles: that nodes with higher degrees exert greater influence, and nodes closer to the center possess higher influence. However, in reality, the contributions of degree and node position to influence are not as substantial. A typical example is the bridge node with low degree but high influence. If a node serves as a bridge between two communities in the network, its significance is self-evident. Similarly, if a node occupies a structural hole position at the network's periphery, it will also wield substantial influence.

Mass is a fundamental physical quantity in physics. Roughly speaking, it can be understood that objects with greater mass possess higher energy potential. In Newtonian mechanics, the magnitude of gravitational force is also directly proportional to mass. For instance, in systems like the solar system or other celestial arrangements, stars are typically the most massive entities. The computation of mass involves multiplying material density by volume. Inspired by this formula, we propose a novel method for identifying key nodes, which we term Structural Neighborhood Centrality (SNC). Its time complexity is related to the network's average degree, \bar{n} , and the number of edges, and is represented as $O(\bar{n} \cdot |E|)$. The principle behind SNC is straightforward. Initially, nodes are conceived as physical entities, with a node's local clustering coefficient regarded as its density, and its degree as its volume. The product of these attributes yields the node's mass. If a node's neighborhood includes numerous high-mass nodes, it suggests that the node itself possesses greater mass, enabling it to attract these high-mass nodes. Consequently, such a node may occupy a more significant position within the system. Beyond the aforementioned considerations, we also account for the structural hole position of nodes, as nodes with high mass are not inherently situated in vital locations.

In summary, unlike other methods, SNC incorporates three attributes: node degree, the less frequently considered local clustering coefficient, and structural hole information. Experimental results demonstrate that SNC exhibits enhanced accuracy and monotonicity, and it also facilitating a more effective identification of the most influential top-k nodes.

The structure of this paper is organized as follows: section 2 reviews some relevant literature; section 3 presents baseline methods and details the proposed SNC method; section 4 describes the experiments conducted to assess the performance of SNC and provides the details of these experiments; section 5 analyzes the experimental outcomes, evaluates the SNC method, and finally, the conclusion section summarizes the content of this paper.

2. Related works

In recent years, researchers have devised numerous methods for identifying key nodes in complex networks. Ma *et al* incorporated the universal law of gravitation from physics, where the k-shell (KS) value of a node was regarded as its mass, and the shortest distance between nodes was considered as the distance between them. Based on this, they proposed a key node identification method that simultaneously accounted for node positions and path information [16]. However, treating the KS value of a node as its mass seemed somewhat strained in Ma's approach. Therefore, Li *et al* made some improvements to Ma's method and introduced a novel key node identification approach known as LGM [17]. Li *et al* primarily made two modifications: first, they considered the node degree as node mass, and second, they expanded the calculation scope to the network's cutoff radius. Nevertheless, Li *et al* only treated the node degree as node mass, resulting in a relatively single-factor consideration, which inevitably weakened the performance of LGM.

The KS and its improved methods based on the KS approach are another effective approach for identifying key nodes in network analysis. The KS calculation is straightforward and accurately positions nodes within the network, determining their influence based on their network position [18]. However, a critical limitation of the KS method is its inability to differentiate the importance of same-layer nodes. Therefore, researchers have focused on enhancing the precision of the KS method by addressing this drawback. Zeng and Zhang made initial attempts in this direction by introducing the mixed degree decomposition (MDD) method, which considers the contribution of deleted nodes and their edges to the network [19]. MDD improves the accuracy of the KS method but increases computational costs, limiting its applicability to large-scale networks. Subsequently, Bae *et al* proposed the CNC+ method [20], which narrows the consideration to the core numbers of neighboring nodes, thus enhancing K-shell's precision. Following a similar approach, Li also proposed CN, which improved K-shell's accuracy by considering neighborhood information [21]. CN significantly enhances K-shell's precision and maintains the same time complexity. Additionally, other methods have been proposed, including the Layered KS introduced by Zareie

and Sheikahmadi [22], KS Iteration Factor suggested by Wang *et al* [23], and approaches by Magi *et al* aimed at refining K-shell's accuracy [24, 25].

In addition to refining the KS method, exploring important nodes based on the inherent information within the network structure is also a research direction worthy of thorough investigation. Koene first introduced the concept of degree centrality where nodes with higher degrees are deemed more important [26]. Subsequently, Chen *et al* proposed the concept of semi-local centrality, which considers both direct and indirect neighbors' impact on a node's importance [27]. In recent years, several novel methods have emerged. Tulu *et al* incorporated node degree information and community structure to assess node importance and proposed the CbM method [28]. Wang *et al* introduced the EFFC method, which evaluates node importance by measuring the change in information propagation efficiency upon node removal, thus considering information flow efficiency within the network [29]. Salavati *et al* combined community detection algorithms and closeness centrality (CC), introducing a new node importance ranking method that reduces computational complexity while rapidly identifying highly influential nodes in the network [30]. Fei *et al* presented an approach that combines node local information and shortest paths, determining node importance through the summation of the interactive forces between nodes [31]. Ullah *et al* considered both node-specific local information and the network's overall topology, proposing the LGC method for node importance identification, which significantly enhances the algorithm's recognition capability [32].

Another approach to studying important node identification methods is to integrate techniques from other disciplines. For instance, Wang *et al* combined information entropy of nodes with the KS method to rank node influence within networks [33]. Xu *et al* considered information entropy of nodes' different neighborhoods to distinguish the significance of various nodes [34]. Additionally, Fan *et al* pioneered the integration of reinforcement learning to discover the key node of complex networks with the Finder method, which offering a novel paradigm for complex network research [35]. Yu *et al* merged graph convolutional networks, transforming the problem of identifying important nodes in complex networks into a regression challenge, resulting in an innovative approach for uncovering significant network nodes [36]. Du *et al* incorporated statistical techniques, utilizing the TOPSIS model to fuse multiple node importance algorithms for assessing node influence [37]. Building upon this, Kuo *et al* simplified the TOPSIS approach and introduced new node importance ranking methods, yielding novel ranking criteria [38]. Moreover, Yang *et al* integrated aspects of possibility theory to devise a method for computing $(k, \eta) - cores$ in uncertain graphs. Although Yang's approach is not specifically aimed at identifying important nodes in complex networks, its underlying principles also offer valuable insights [39].

3. Proposed method

This section initiates with an introduction to some node centrality metrics, followed by a detailed exposition of the proposed Structural Neighborhood Centrality (SNC) method in this paper.

3.1. Baseline methods

3.3.1. CC

CC is a classic method for identifying important nodes in complex networks. The CC considers the average shortest path length from each node to all other nodes [40]. In essence, CC posits that nodes closer to other nodes are of higher importance. While CC exhibits high accuracy in computation, its complexity limits its application in networks with a large number of nodes. The formula for CC is defined as follows:

$$CC_i = \frac{|V| - 1}{\sum_{j \neq i} d_{ij}} \quad (1)$$

where d_{ij} represents the shortest path distance from node i to node j .

3.3.2. KS

The KS method, proposed by Kitsak *et al* [41], is a classic approach to rank the importance of nodes in complex networks based on their network position. The fundamental idea behind KS is to recursively remove nodes with degrees less than or equal to 'k' in the network. For example, KS starts by removing nodes with degrees of 1 or lower, and subsequently eliminates nodes with degrees of 1 or lower that are generated after the previous removal. This process continues until no nodes with degrees of 1 or lower remain in the network. Nodes that are removed are then assigned a 'ks' value of 1, while the remaining nodes in the network have degrees of at least 2. The procedure is repeated to determine nodes with 'ks' values of 2, and this process continues iteratively until all nodes in the network are assigned 'ks' values. The KS method offers a straightforward principle and computationally efficient process. However, a major limitation of the method is its inability to precisely differentiate the importance of nodes within the same shell.

3.3.3. *H-index (HI)*

In complex networks, the HI of a node is a method that utilizes the concept of HI to determine node importance by considering the degrees of neighboring nodes [42–44]. Its definition is as follows:

$$HI_i = H(d_{v_1}, d_{v_2}, d_{v_3}, \dots, d_{v_j}). \quad (2)$$

In this equation, v_1, v_2, \dots, v_j represent the neighbors of node i , and d_{v_1} represents the degrees of the neighboring nodes. H is a function that returns a value h , which indicates that in the set $\{d_{v_1}, d_{v_2}, d_{v_3}, \dots, d_{v_j}\}$, there are h values greater than or equal to h .

3.3.4. *Collective influence (CI)*

CI is also a method for assessing node importance in a network, introduced by Morone and Makse [45]. CI determines node influence based on the extent of disruption to the giant connected component in the network when nodes are removed. Its definition is as follows:

$$CI_i = (d_i - 1) \sum_{j \in \partial B(i, l)} (d_j - 1) \quad (3)$$

where $B(i, l)$ represents the set of nodes in the network surrounding node i and belonging to the ball of radius l , d_i denotes the degree of node i , and l is a pre-defined value typically set to 3 in large and medium-sized networks, and 2 in small networks.

3.3.5. *CycleRatio (CR)*

Cycle ratio was proposed by Fan *et al* [46]. The theoretical foundation of CR is the network's cycle structure, where the term 'cycle ratio' refers to the degree to which a node participates in the shortest cycles of other nodes. The term 'shortest cycle' refers to the smallest-length cycle containing the node. The definition of the cycle ratio is straightforward and given as follows:

$$\begin{cases} CR_i = 0, c_{ii} = 0 \\ CR_i = \sum_{j, c_{ij} > 0} \frac{c_{ij}}{c_{ij}}, c_{ii} > 0. \end{cases} \quad (4)$$

In equation (2), c_{ij} represents the number of shortest cycles in the network that pass through nodes i and j , while c_{ii} represents the number of shortest cycles in the network that pass through node i .

3.3.6. *KSGC*

KSGC was introduced by Yang and Xiao [47]. Unlike other methods, Yang *et al* regarded a node's position within the network as a crucial attribute of its importance. However, many algorithms for assessing node importance overlook this aspect. Thus, Yang *et al* devised an improved gravitational model based on the KS method. Its definition is as follows:

$$KSGC_i = \sum_{d_{ij} \leq R} c_{ij} \frac{k_i k_j}{d_{ij}^2} \quad (5)$$

where c_{ij} is a constraint coefficient, defined as $c_{ij} = e^{\frac{ks_i - ks_j}{ks_{\max} - ks_{\min}}}$, where ks_i represents the ks value of node i , d_{ij} represents the distance between nodes i and j , ks_{\max} is the maximum ks value in the network, and R denotes the network's truncation radius, which is half of the average shortest path length in the network.

3.2. Structural neighborhood centrality

3.2.1. *Clustering coefficient*

Clustering coefficient is a fundamental concept in graph theory, describing the degree to which nodes tend to cluster together. The clustering coefficient of a specific node in a graph is referred to as its local clustering coefficient, which quantifies the extent to which neighboring nodes form clusters or cliques [15]. The local clustering coefficient of a node is defined as the ratio of the number of actual edges between its neighbors to the number of potential edges that could exist among them, as outlined below:

$$C_i = \frac{2E}{k(k-1)}. \quad (6)$$

Here, k represents the number of nodes within the neighborhood, and E denotes the actual number of edges that exist between nodes within the neighborhood.

3.2.2. Structural hole

A structural hole refers to the gap formed between two nodes in a network where no redundant connections exist [48]. Nodes that occupy a structural hole position in a network have a competitive advantage, which can manifest in various ways. For example, individuals occupying a structural hole position can allow other individuals to get to know each other or compete with each other. To quantify the control that structural hole nodes exert over various relationships within the network, Burt introduced the concept of network constraint, defined as follows:

$$SH_i = \sum_j \left(\mu_{ij} + \sum_{q \neq i,j} \mu_{iq} \mu_{qj} \right)^2 \quad (7)$$

where node q is the common neighbor between nodes i and j . μ_{ij} denotes the proportion of the total energy invested by node i to maintain the relationship with node j , which is defined as follows

$$\mu_{ij} = \frac{e_{ij}}{\sum_{j \in N_i} e_{ij}}. \quad (8)$$

In the above equation, N_i is the set of neighboring nodes of node i . The value of e_{ij} is 1 when there are connected edges of i and j , and 0 when the opposite is true.

3.2.3. Structural neighborhood centrality (SNC)

Consider a simple graph $G(V, E)$, where V represents the set of nodes in the graph, and E represents the set of edges. We define N_{v1} as the first-order neighborhood set of a node in V and N_{v2} as the second-order neighborhood set of that node. The computation of SNC involves two main steps. The first step involves calculating the *mass* value and structural hole coefficient ω_i for each node. The second step involves calculating the *nc* value for each node, as detailed below:

Firstly, calculate the *mass* value and structural hole coefficient ω_i for each node, defined as follows:

$$mass_i = d_i \times e^{C_i} \quad (9)$$

and:

$$\omega_i = \frac{e^{-SH_i}}{2} \quad (10)$$

where d_i denotes the degree of node i and C_i denotes the local clustering coefficient of node i .

After this, we determine the node's *nc* value, which is calculated as follows

$$nc_i = \sum_{j \in N_{v1}, N_{v2}} mass_j \times mass_i \times \omega_j \quad (11)$$

where ω_j represents the structural hole coefficient of node j within the neighborhood of node i . It is noteworthy that, similar to the concept of CR, we posit that the node also exerts an influence on the node itself. So in our method, the node itself is also included in the calculation. Ultimately, by incorporating the structural hole coefficient onto the foundation of the *nc* value, we obtain the significance score, denoted as the *snc* value:

$$snc_i = nc_i \times \omega_i. \quad (12)$$

In equation (9), a higher local clustering coefficient of a node indicates a stronger tendency for nodes to cluster together, and the function $y = e^x$ amplifies the impact of the clustering coefficient on the node. And in equation (10), as node i occupies more structural holes, its structural hole constraint SH_i becomes smaller, leading to an increase in ω_i . Consequently, when a node i occupies more structural holes, its structural hole constraint is reduced, resulting in a higher structural hole coefficient.

Unlike most traditional node centrality measures, SNC takes into account both degree and local clustering coefficient, driven by two primary reasons. Firstly, neighbors of nodes with high clustering coefficients tend to cluster together. This not only enhances the robustness of local network structures but also promotes the efficiency of information propagation within the network. However, it is worth noting that nodes with high clustering coefficients may also restrict the range of information dissemination while speeding up its transmission. This happens because, once nodes cluster together to form a community structure, information might primarily circulate within these communities, rather than spreading beyond

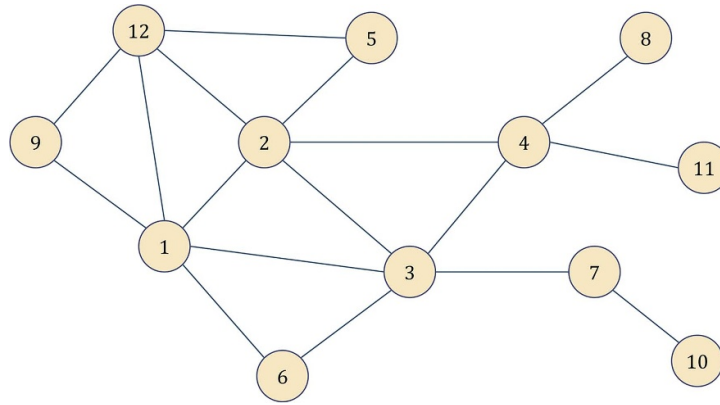


Figure 1. Example network. This example network has 12 nodes and 17 connected edges.

them. In such scenarios, if SNC only considers the clustering coefficient of nodes, it could lead to localized information propagation. By incorporating both degree and clustering coefficient, we can enhance network robustness and compensate for the limitations of clustering coefficient in measuring node influence. Secondly, the local clustering coefficient signifies the tendency of neighboring nodes to cluster together, and this might be influenced by higher-order neighboring nodes. Therefore, considering the clustering coefficients of first and second-order neighbors is akin to simultaneously taking into account the interaction information of nodes' third-order or even higher-order neighbors. This allows us to incorporate as much information as possible without increasing algorithm complexity significantly, thus compensating for some local centrality measures' limitations in terms of inadequate information consideration.

As mentioned earlier, in the SNC method, a node's degree is regarded as its volume, and its clustering coefficient is seen as density. The product of volume and density yields mass. Mass is a significant physical quantity in physics; typically, the greater an object's mass, the greater its energy, potentially rendering it an important element within a system. If a node is surrounded by numerous high-mass nodes, its mass may be higher, leading to a potentially more significant role within the system. Thus, we consider that a higher *snc* value for a node signifies greater importance. It is evident that the time complexity of the SNC method mainly revolves around two components. The first involves the calculation of the clustering coefficient and structural holes, while the second entails gathering information within the node's second-order neighborhood. Both the computation of structural holes and clustering coefficient necessitate considering a node's first-order neighbors and the edges connecting them. Consequently, the time complexity for this part is $O(\bar{n} \cdot |E|)$, where \bar{n} represents the network's average degree. On the other hand, the collection of information within the second-order neighborhood is only dependent on a node and its first-order neighbors' degrees, leading to a time complexity of $O(\bar{n}^2)$. Generally, the number of edges in a network far exceeds the average degree of nodes. As a result, the time complexity of SNC is $O(\bar{n} \cdot |E|)$.

Figure 1 illustrates an example network with 12 nodes and 17 edges. When computing the SNC value for each node, we should commence by determining the mass value for each node within the network. Subsequently, the first-order and second-order neighbor nodes of the current node are identified. Taking node 9 as an illustration, its first-order neighbors are nodes 1 and 12, while its second-order neighbors include nodes 1, 2, 3, 5, 6, 9, and 12. It is notable that node 1 functions as both a first-order neighbor and a second-order neighbor to node 9. So, when calculating the SNC value, we should not calculate node 1 only once. This is because node 1 can exert direct influence on node 9 and can also indirectly impact node 9 through its connection with node 12. Therefore, a comprehensive assessment of these influences is imperative when calculating the SNC value.

4. Experiments methods

This section focuses on the experiments and methods used to evaluate the performance of SNC. Section 4.1 describes the actual network dataset used in this study, while section 4.2 details the SIR infectious disease transmission model, which is a widely accepted approach for measuring the propagation ability of nodes. In addition, section 4.3 outlines the criteria used to evaluate the performance of all methodologies.

Table 1. Basic topological features of network data. The leftmost column is the name of the network, N and E denote the number of nodes and edges of the network, respectively; $\langle K \rangle$ and K_{\max} denote the average degree and maximum degree of the network, respectively; C and r denote the clustering coefficient and congruence coefficient of the network, respectively.

Network	N	E	$\langle K \rangle$	K_{\max}	C	r
Dolphins	62	159	5	12	0.2590	−0.0436
Lesmis	77	254	6	36	0.5731	−0.1652
Polbooks	105	441	8	25	0.4875	−0.1279
Adjnoun	112	425	7	49	0.1728	−0.1293
Jazz	198	2742	27	100	0.5203	0.0202
C_elegans	297	2148	15	134	0.3115	−0.1520
USAir97	332	2126	12	139	0.6252	−0.2079
NS_GC	379	914	4	34	0.7400	−0.0817
Email	1133	5451	9	71	0.2200	0.0782
Yeast	2375	11 693	9	118	0.3100	0.4539
Minnesota	2642	3303	2	5	0.0160	−0.1848
Kohonen	3772	12 718	5	740	0.2100	−0.1204

4.1. Datasets

In this paper, we investigate the impact of SNC on 12 actual networks from different domains. (1) Dolphins: a network of interaction relationships between 62 dolphins [49]. (2) Lesmis: a social network composed of characters from Victor Hugo's book *Les Misérables* [50]. (3) Polbooks: a network for buying books on American politics [50]. (4) Adjnoun: a network between adjectives and nouns commonly used in Charles Dickens's novel *David Copperfield* [51]. (5) Jazz: a collaborative network between jazz musicians [52]. (6) C_elegans: a neural network of elegant cryptobarc nematodes [15]. (7) USAir97: a network of American Airlines [50]. (8) NS_GC: a collaborative network of scientists [51]. (9) Email: the email communication network of the University of Rovira I Virgili, Spain [53]. (10) Yeast: a protein–protein interaction network in yeast [54]. (11) Minnesota: the Minnesota road network [50]. (12) Kohonen: a citation network from Pajek [50]. Table 1 provides additional information about these networks.

4.2. Susceptible-Infectious-Recovered (SIR) model

The SIR propagation model [55, 56] can faithfully simulate the spread of infectious diseases across networks, making it one of the standards for comparing node propagation ranking methods. In the standard SIR model, nodes can exist in three states: Susceptible (S), Infected (I), and Recovered (R). At the outset of the SIR model, all nodes are in the S state. In this time, a node is chosen as the initial spreading seed, transitioning its state to I. In each propagation round, an infected node (I) has a probability β of infecting susceptible nodes within its neighborhood, changing their states from S to I. Additionally, infected nodes transition to the R state with a probability λ ($\lambda = 1$), after which they remain unchanged. Propagation terminates when there are no nodes in the I state. At this point, the count of nodes in the R state is computed, representing the spreading capability of the initial infected nodes. To mitigate potential stochasticity, multiple simulation runs are often conducted, recording data from each run and subsequently averaging the results to determine the actual propagation capability of nodes.

In this study, each node in the network will be evaluated as a seed node to determine its propagation capability. The propagation capability test for each node will be performed 100 times and the resulting data will be averaged to reduce the impact of randomness. During the simulation process of SIR, it is necessary to set the magnitude of the propagation probability β . If β is too large, the nodes can be easily infected, causing most of the nodes in the network to become infected, which makes it difficult to differentiate nodes based on their propagation ability. On the other hand, if β is too small, the seed nodes will not be able to infect other nodes, resulting in most nodes remaining uninfected, which also makes it difficult to distinguish nodes based on their propagation. Therefore, in SIR simulation experiments, the value of β is usually set to a value slightly larger than the propagation threshold β_{th} , $\beta_{\text{th}} = \langle k \rangle / \langle k^2 \rangle$, where $\langle k^2 \rangle$ is the average second-order degree of the network and $\langle k \rangle$ is the average degree of the network.

4.3. Evaluation criteria

4.3.1. Kendall correlation coefficient

The Kendall tau correlation coefficient is used to evaluate the accuracy of the SNC method by measuring the correlation between the ranked list obtained by the SNC method and the ranked list obtained by the SIR simulation experiment. A higher Kendall tau correlation coefficient indicates a stronger similarity between the two lists, which implies a higher accuracy of the method.

The Kendall tau correlation coefficient is defined as follows [46]

$$(X, Y) = \frac{(n^+ - n^-)}{\sqrt{n^+ + n^- + t_x} \times \sqrt{n^+ + n^- + t_y}} \quad (13)$$

where X and Y represent the sorted lists of node influence obtained by two different methods. n^+ and n^- denote the number of homoscedastic and heteroscedastic pairs, respectively. Consider two sorted lists X and Y . Suppose $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_n)$. At this point, N double tuples $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ can be formed from the above two lists. For any pair of duals $(x_i, y_i), (x_j, y_j)$, if two duals are of the same rank, i.e., $x_i < x_j$ and $y_i < y_j$, or $x_i > x_j$ and $y_i > y_j$, then $(x_i, y_i), (x_j, y_j)$ is said to be a homogeneous pair, otherwise, it is a heterogeneous pair. In addition, if $x_i = x_j$ and $y_i \neq y_j$, the value of t_x is added by 1. Similarly, if $x_i \neq x_j$ and $y_i = y_j$, the value of t_y is added by 1. In particular, if $x_i = x_j$ and $y_i = y_j$, the element does not belong to any of the above cases and no change is made to any value.

4.3.2. Monotonicity

The primary purpose of assessing the monotonicity of ranking results is to determine whether a particular ranking method can effectively differentiate between the influence of individual nodes. Ranking methods with low monotonicity, such as KS or HI, may group nodes into the same rank, suggesting that the propagation abilities of various nodes are considered to be the same. However, there are often subtle differences between nodes, so a good ranking method should be capable of accurately measuring the propagation ability of each node, rather than roughly dividing them. The ordering monotonicity M is defined as follows [57]

$$M = \left[1 - \frac{\sum_{i \in L} s_i (s_i - 1)}{|V|(|V| - 1)} \right]^2 \quad (14)$$

where $|V|$ denotes the number of nodes in the network, L is the sorted list formed by the important node sorting method, and s_i denotes the number of nodes of rank i in the sorted list. M takes values in the range $[0, 1]$, and the closer its value is to 1, the stronger the monotonicity.

4.3.3. Jaccard similarity coefficient

The Jaccard similarity coefficient is commonly used to measure the similarity between two finite sets, with a higher Jaccard similarity coefficient indicating a greater similarity between the sets. In this study, we use this coefficient to evaluate the ability of SNC to identify Top-k nodes in a network, ranked by their propagation ability [20, 58, 59]. The Jaccard similarity coefficient is defined as follows

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (15)$$

where X and Y denote the Top-k node influence sorted lists obtained by two different methods. X typically corresponds to the list of Top-k nodes derived from SIR experiments, while Y represents the list of Top-k nodes obtained using another method. In this study, the closer the value of $J(X, Y)$ is to 1, the stronger the ability of method Y to find Top-k nodes.

5. Results and discussions

This section presents the results of experiments conducted to evaluate the performance of SNC. Section 5.1 describes the results of node importance ranking using each method on a manually designed miniature example network. Section 5.2 examines the accuracy of different methods on a real network using the Kendall tau correlation coefficient. Section 5.3 discusses the monotonicity of different methods, which refers to their ability to distinguish between the propagation abilities of different nodes. Section 5.4 examines the variation in the ability of different methods to identify the Top-k nodes with the highest propagation ability.

5.1. A simple experiment on a tiny artificial network

To assess the true spreading influence of nodes, we employed the classic SIR model for node evaluation. In the SIR experiments, each node underwent 100 propagation trials to reduce potential randomness. Figure 2 illustrates the top three nodes identified by each method in the example network from figure 1, while table 2 provides a more detailed ranking list generated by various sorting methods. From figure 2 and table 2, it can be observed that SNC, CI, CC, HI, and KSGC identify the same top three influential nodes as the SIR experiments, although with slight variations in rankings. CR identifies one different node in the top three in comparison to the SIR results, while the KS metric exhibits significant differences. Table 2 further reveals that

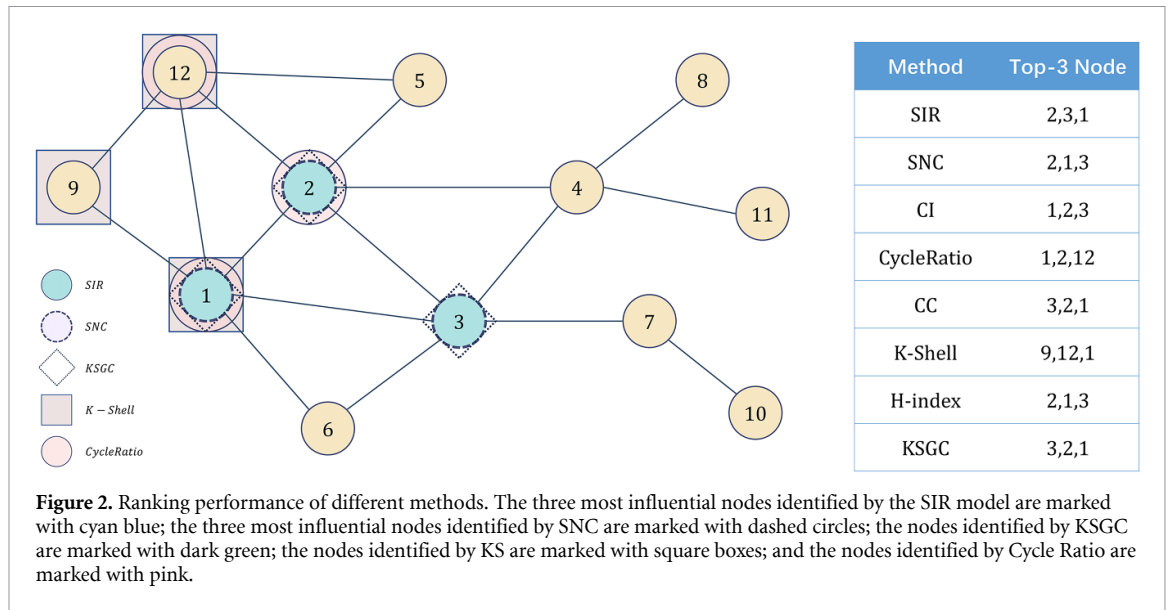


Table 2. Ranking list obtained by different ranking methods. The number below the method name indicates the node number, and the number below the rank indicates the ranking. SIR represents the ranking list of node influence given by the SIR model; SNC is the ranking list of node influence given by the method proposed in this study; CI represents Collective Influence; CR represents Cycle Ratio; KS stands for K-Shell; HI stands for H-index.

Rank	SIR	SNC	CI	CR	CC	KS	HI	KSGC
1	2	2	1,2	1,2	3	9,12,1,2,3,4,5,6,	2	3
2	3	1	3	12,3	2	7,8,10,11	1,3	2
3	1	3	4,12	4,5,6,9	1,4	—	4,5,6,9,12	1
4	12	12	6	7,8,10,11	12	—	7,8,10,11	4
5	4	4	5	—	6	—	—	12
6	6	6	9	—	7	—	—	6
7	5	5	7	—	5	—	—	9,5
8	9	9	8,10,11	—	9	—	—	7

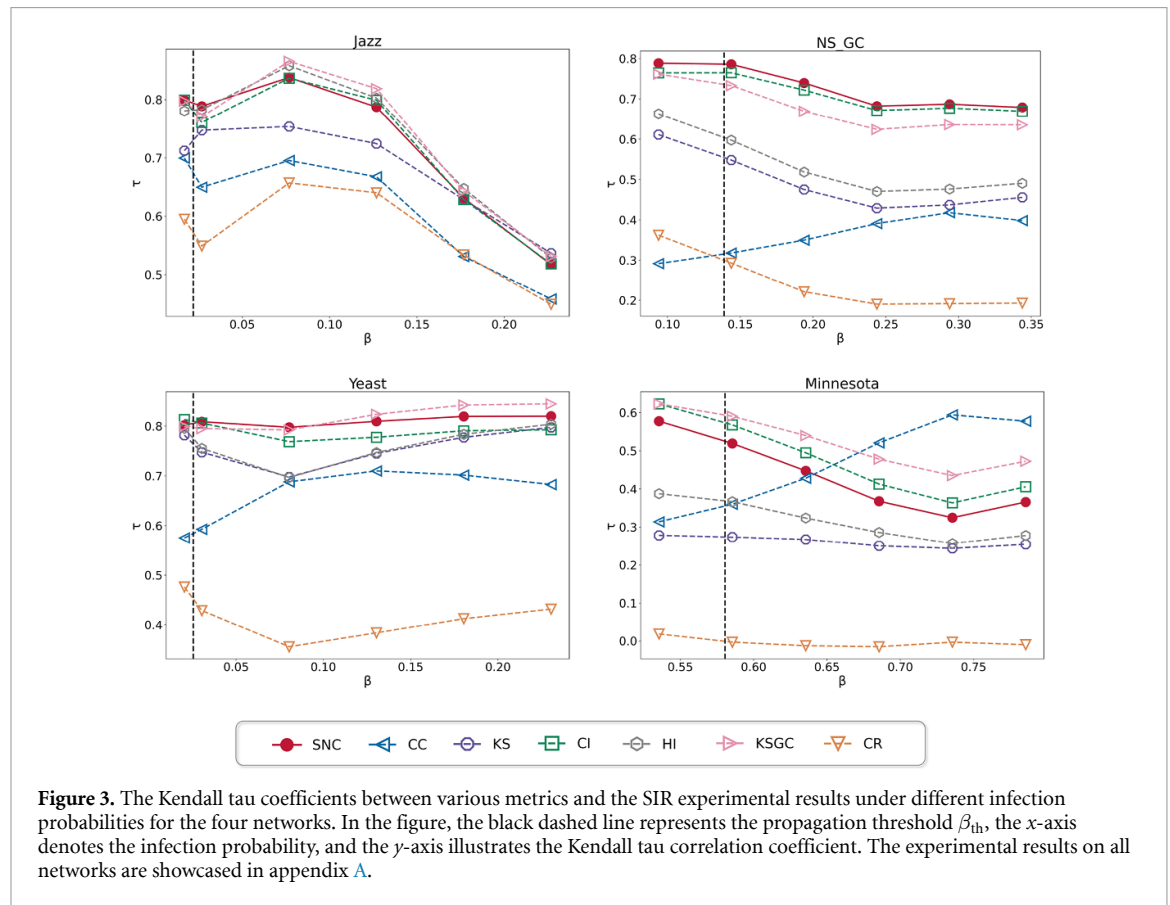
SNC precisely quantifies the spreading abilities of each node, ensuring that no two nodes share the same spreading capability and exhibiting the highest level of distinguishability. Its ranking list of node influence closely aligns with the results obtained from the SIR method. In contrast, KS, HI, and CR fail to differentiate between nodes' influences effectively, as these methods assign multiple nodes to the same rank, despite potentially substantial variations in their influence levels.

5.2. Accuracy of the SNC method

We calculated the Kendall tau correlation coefficient between the ranking lists obtained from the SNC method and the baseline method and the ranking lists obtained from the SIR experiments. The accuracy of the SNC method was evaluated based on this correlation coefficient. In this subsection, we set six different propagation rates for the SIR propagation experiment: $\beta - 0.05$ or $\beta - 0.01$, β , $\beta + 0.05$, $\beta + 0.10$, $\beta + 0.15$, and $\beta + 0.20$, where $\beta = 1.01 \times \beta_{th}$, and the propagation rate was set to $\beta - 0.01$ when $\beta - 0.05 \leq 0$. The experimental results are presented in table 3, appendix A and figure 3. It can be seen that the SNC method achieved the highest Kendall tau correlation coefficient and the best accuracy when the propagation rate is β . In the Lesmis, USAir97, and Polbooks networks, the SNC method maintained the highest accuracy at all five contagion probabilities. The KSGC method had higher accuracy in some networks but lower accuracy in others because it only considered node degree and did not take into account other local node information. In the Yeast and Dolphins networks, the CI method had the highest accuracy, but the difference between the SNC and CI methods was not as significant. In the Minnesota network, CI widens significant gap with SNC method. This is because the Minnesota network has a higher propagation rate, making nodes more easily infected, which weakened the performance of the SNC method. The CR method had the lowest accuracy in all datasets because the nodes identified by CR were different from those identified by SIR. CR identifies important nodes that are evenly distributed throughout the network, while other methods tend to identify important nodes that are clustered together, forming a distinct association structure.

Table 3. Results of Kendall tau coefficients between different ranking methods and SIR when the propagation rate is β . The value with the largest correlation is marked in bold.

Network	β	SNC	CC	KS	CI	HI	KSGC	CR
Dolphins	0.1739	0.8257	0.6054	0.7269	0.8164	0.8050	0.7896	0.5219
Lesmis	0.0914	0.8389	0.5484	0.7771	0.8087	0.8064	0.8124	0.6815
Polbooks	0.0924	0.7760	0.3735	0.7023	0.7189	0.7424	0.7569	0.4323
Adjnoun	0.0791	0.8533	0.7937	0.7823	0.8525	0.8323	0.8453	0.5791
Jazz	0.0268	0.7887	0.6496	0.7473	0.7606	0.7826	0.7707	0.5489
C_elegans	0.0403	0.7542	0.6458	0.6827	0.7197	0.7053	0.7445	0.4651
USAir97	0.0233	0.7700	0.7023	0.7528	0.7553	0.7486	0.7618	0.4039
NS_GC	0.1439	0.7858	0.3177	0.5477	0.7648	0.5972	0.7330	0.2928
Email	0.0271	0.8109	0.7644	0.7689	0.7972	0.7722	0.7775	0.5168
Yeast	0.0304	0.8087	0.5927	0.7475	0.8060	0.7552	0.7949	0.4288
Minnesota	0.5855	0.5576	0.3761	0.2986	0.5894	0.3964	0.6107	0.0976
Kohonen	0.0161	0.6307	0.5545	0.1522	0.6118	0.5507	0.6494	0.3373

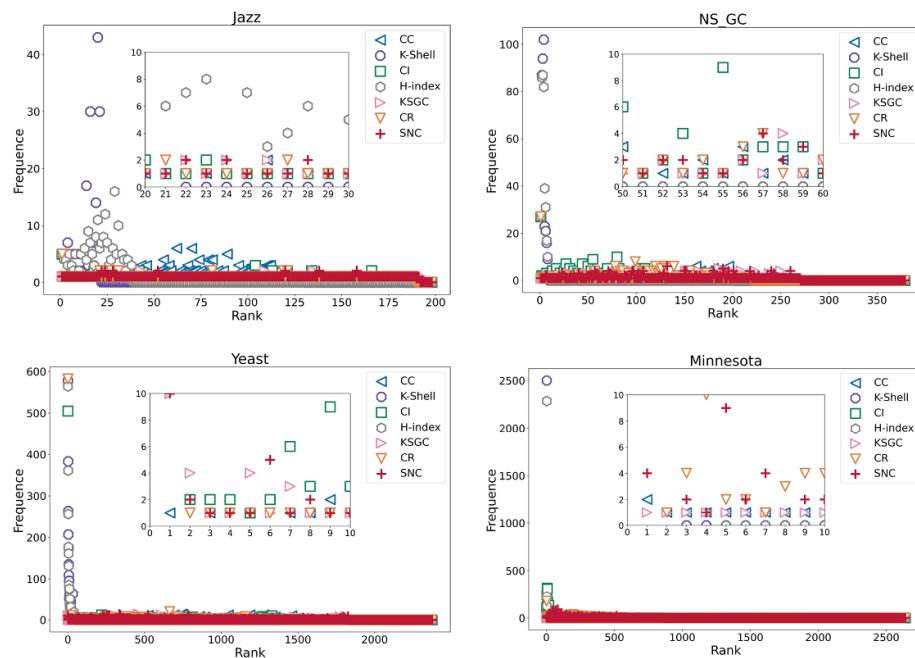


5.3. Monotonicity of the SNC method

Many contemporary node importance ranking methods face the challenge of accurately distinguishing the importance of each node. A good ranking method should not only identify the high-impact nodes but also uniquely determine the propagation ability of each node. The SNC method addresses this issue by considering not only the neighborhood information of nodes but also their clustering coefficients and SHs. Table 4 shows the monotonicity results of different methods, with KS and HI having the worst monotonicity, while SNC, KSGC, CI, and CR perform better, with SNC having the best monotonicity and the average monotonicity closest to 1. In figure 4 and appendix B, if the ranking result shows that a certain level appears more frequently, it means that there are more nodes of the same level, and the differential performance of this method is poor; if the ranking result can draw a straight line at the bottom, it means that nodes of the same level With fewer nodes, the differential performance of this method is better. It can also be seen from figure 4 and appendix B that the node ranking obtained by the SNC method forms a straight line at the bottom, indicating that it has only a limited number of nodes in each rank and can differentiate between nodes effectively. In fact, SNC, KSGC, CI, and CR methods all partially or fully consider the local

Table 4. Comparison of monotonicity results of different methods. The last row shows the average monotonicity results, and the highest monotonicity values in the table are marked in bolded black.

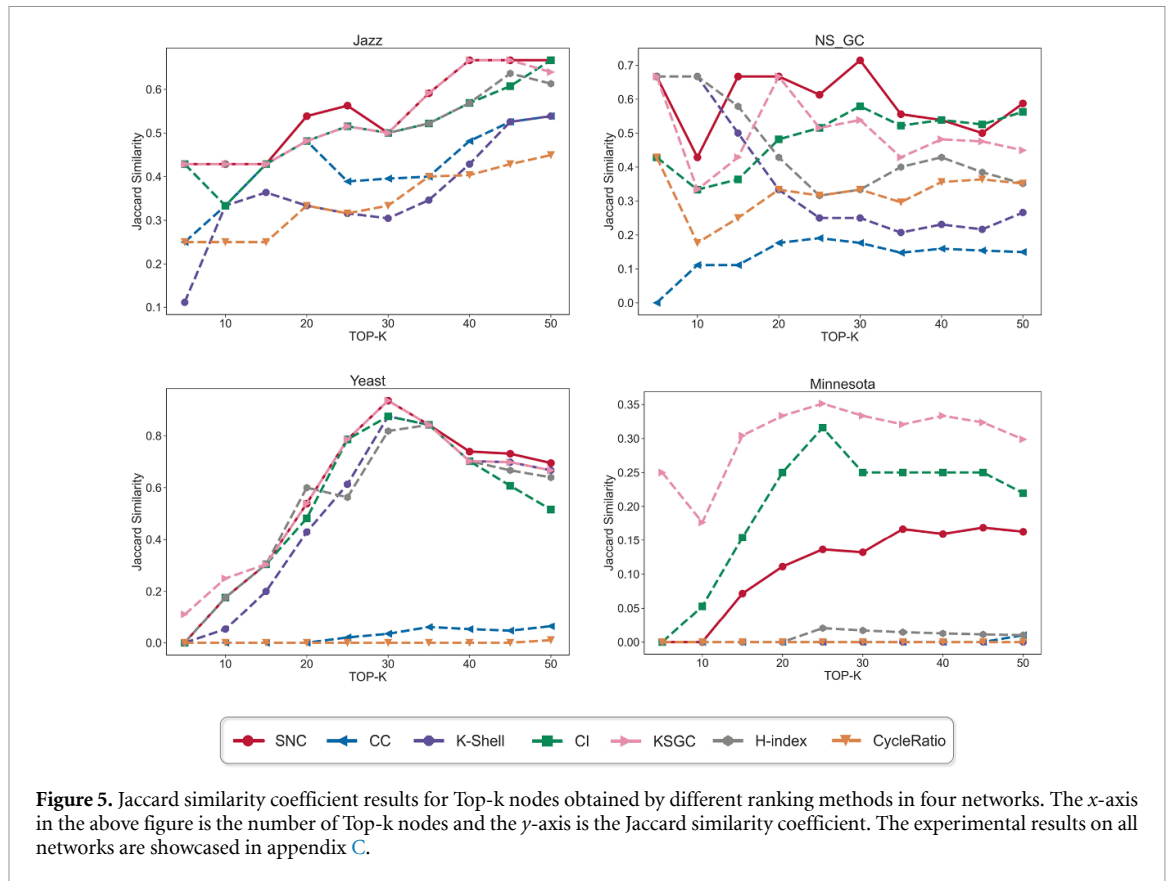
Network	SNC	CC	KS	CI	HI	KSGC	CR
Dolphins	0.9979	0.9737	0.3769	0.9613	0.6841	0.9852	0.9602
Lesmis	0.9587	0.9414	0.7441	0.8878	0.7625	0.9547	0.8769
Polbooks	0.9998	0.9847	0.4949	0.9992	0.7067	0.9985	0.9992
Adjnoun	0.9997	0.9837	0.5991	0.9846	0.8110	0.9980	0.9856
Jazz	0.9992	0.9878	0.7994	0.9979	0.9383	0.9993	0.9984
C_elegans	0.9976	0.9893	0.6094	0.9948	0.8745	0.9974	0.9952
USAir97	0.9951	0.9892	0.8114	0.9433	0.8355	0.9935	0.9451
NS_GC	0.9951	0.9927	0.6421	0.9807	0.6825	0.9950	0.9830
Email	0.9998	0.9988	0.8088	0.9649	0.8582	0.9982	0.9631
Yeast	0.9991	0.9988	0.7737	0.9111	0.7923	0.9988	0.8831
Minnesota	0.9856	0.9999	0.5810	0.8802	0.8599	0.9901	0.9841
Kohonen	0.9985	0.9981	0.7666	0.9417	0.9968	0.7335	0.9323
Average Value	0.9938	0.9865	0.6673	0.9539	0.8168	0.9701	0.9588

**Figure 4.** Rank distribution of ranked lists with different methods in four networks. The x-axis in the above figure is the rank of nodes, and the y-axis is the number of nodes with the same rank. The experimental results on all networks are showcased in appendix B.

information of nodes, which enhances their monotonicity. Since the local information of any two nodes is almost not the same, this contributes to the better performance of these methods.

5.4. Ability of the SNC method to identify the top-k influential spreaders

Apart from accurately identifying the importance of nodes within a network, correctly pinpointing the top-ranking influential nodes in the network also holds practical significance. In order to evaluate the ability of the SNC method in identifying the Top-k most influential nodes in a network, this paper selects the Top-k most influential nodes from the results lists obtained by different ranking methods and the results list obtained from SIR experiments. The Jaccard similarity coefficient is then employed to calculate the similarity between them. High similarity indicates that the method has successfully identified Top-k influential nodes that closely match those identified by SIR experiments, demonstrating its effectiveness in recognizing Top-k nodes. Figure 5 and appendix C present the correlation coefficient values between the Top-k nodes identified by different methods and those obtained from SIR experiments. The evaluation encompasses the ability of different methods to identify the Top-5 to Top-50 nodes in the network. As observed in figure 5, SNC, CI, KSGC, and HI exhibit the highest similarity coefficients. In some networks, SNC does not exhibit a significant difference from the other metrics. For example, in the NS_GC network, the ability of SNC to



identify the Top-10 nodes is weaker than that of HI, and its ability to identify the Top-45 nodes is weaker than CI. However, on the whole, the curve of the SNC method remains consistently at the highest position in almost all networks, with its similarity coefficients higher than other methods. Therefore, SNC not only accurately assesses the importance of each node in the network but also identifies the Top-k most influential nodes in the network.

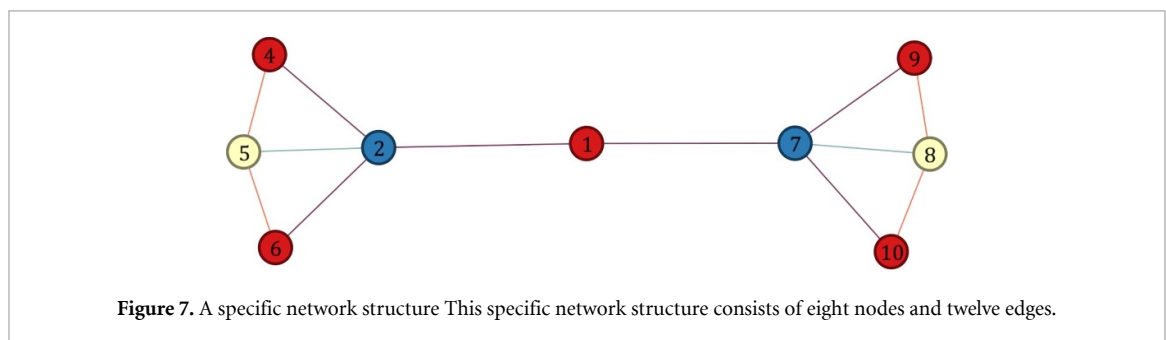
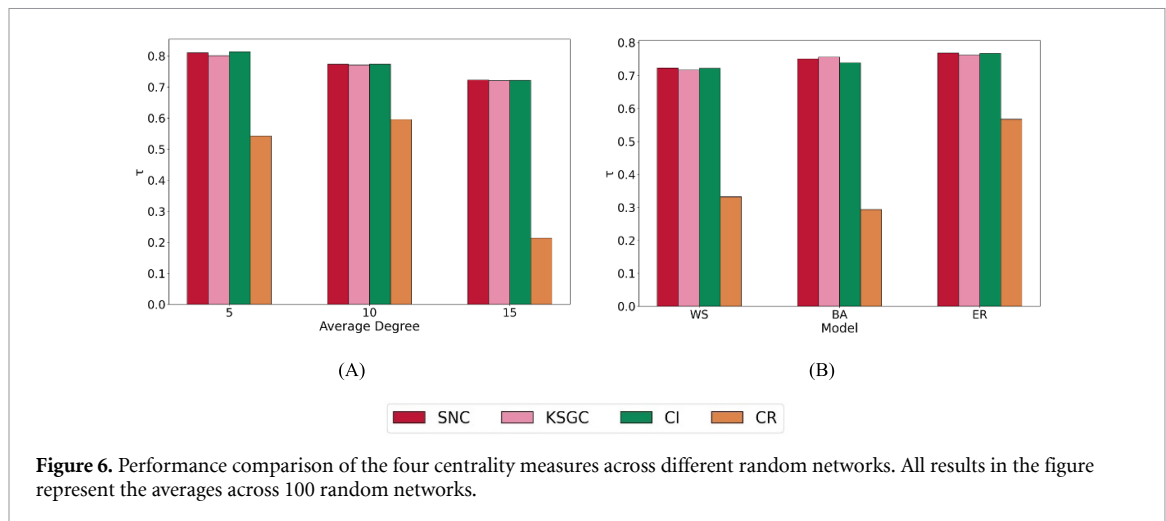
6. Further analysis of SNC

While we have tested various aspects of SNC's performance and discussed some of its advantages in the preceding sections, we still lack a comprehensive understanding of the underlying principles behind SNC. Therefore, it is necessary for us to delve further into the analysis of SNC. Sensitivity analysis can help identify the influence of network topology characteristics on SNC. Statistical analysis can determine the significance and stability of SNC scores, providing reliable grounds for the interpretability of the SNC metric. Additionally, correlation analysis can establish the relationships between SNC and other centrality measures, aiding in our comprehension of the strengths and limitations of SNC compared to traditional centrality metrics. Hence, this section primarily focuses on further analyzing SNC from these three perspectives.

6.1. Sensitivity analysis

The centrality scores of nodes are often influenced by the network's topological features, making the generality of a method across different networks an important property. In this section, we will discuss the performance of SNC under different link densities and degree distributions, comparing it with centrality methods like CI, CR, and KSGC.

To assess the impact of different link densities on SNC and other methods, we constructed three ER random networks, all with 1000 nodes but average degrees of 5, 10, and 15, as shown in figure 6(A). As the average degree increases, the network becomes denser, and the accuracy of these methods decreases. In sparse graphs, SNC's accuracy is slightly lower than that of CI, but it surpasses KSGC and CR. In other scenarios, SNC consistently exhibits the highest accuracy, but it does not show a significant gap when compared to KSGC and CI. CR is somewhat unique, as the influence of link density on it seems to lack a clear pattern. This is because CR is designed based on cycle structures within the network and is unrelated to the network's link density.



To test the influence of degree distributions on these centrality measures, we constructed three classic random graph models: ER graphs, Watts–Strogatz small-world networks (WS), and Barabasi–Albert scale-free networks (BA). These three models represent completely random, small-world effects, and scale-free degree distribution patterns, respectively. All of these networks consist of 1133 nodes and have edges close to 5451. Figure 6(B) illustrates that CI performs best in the BA network, while in other cases, SNC achieves the highest accuracy.

Indeed, due to SNC's consideration of both the node and its neighbors' local clustering coefficients, it tends to identify important nodes that are strongly clustered together. As a result, important nodes identified by SNC often aggregate and form a community structure. Consequently, when dealing with sparse graphs, SNC will rely primarily on degree to assess node importance, which can significantly diminish its performance.

Furthermore, specific network structures can also undermine SNC's performance. Consider a network structure like the one depicted in figure 7:

In figure 7, the node propagation capability sequence given by the SIR propagation model is 1, 7, 2, 5, 8, etc, while the node propagation capability sequence given by the SNC method is 7, 2, 5, 8, etc, without recognizing the importance of node 1, which serves as a bridge node. The reason for this phenomenon is that SNC considers not only its own structural hole information but also the structural hole information of neighboring nodes. Indeed, node 1 occupies the most structural hole positions with a structural hole coefficient of 0.30, the highest among all nodes. However, nodes 2 and 7 also have high structural hole coefficients of 0.29. Therefore, according to SNC's measurement criteria, nodes 2 and 7, which have more neighbors, are identified as more important bridge nodes. One feasible solution to address this issue is to regulate the contribution of node's local information and neighbor information to its importance by setting parameters. However, this approach will increase the time complexity of SNC and require extensive experiments to determine the parameter values.

6.2. Statistical analysis

In the problem of identifying important nodes in complex networks, various centrality metrics often have dimensionless absolute values. Moreover, due to the varying sizes and diverse structures of networks, we need

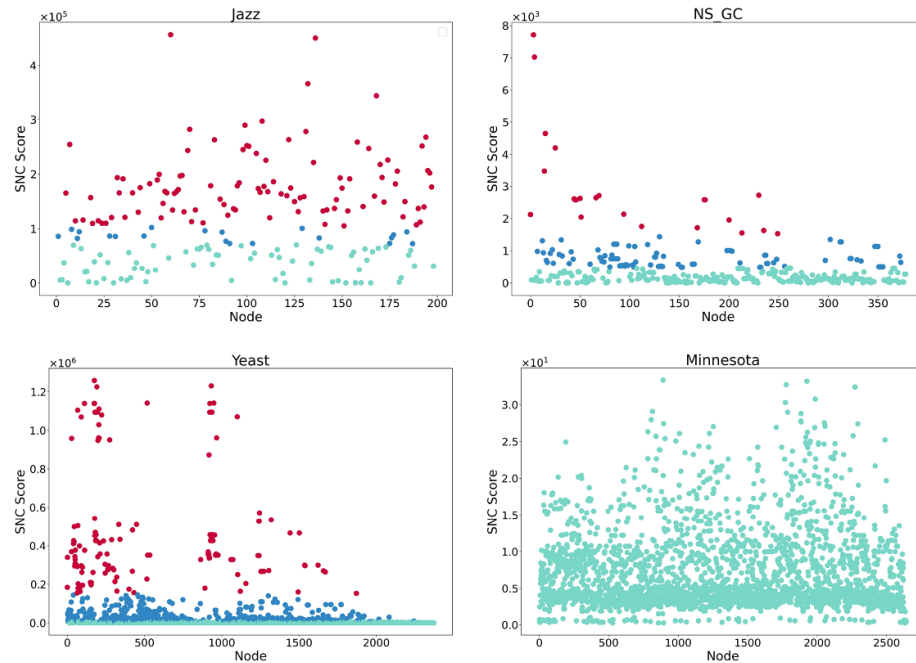


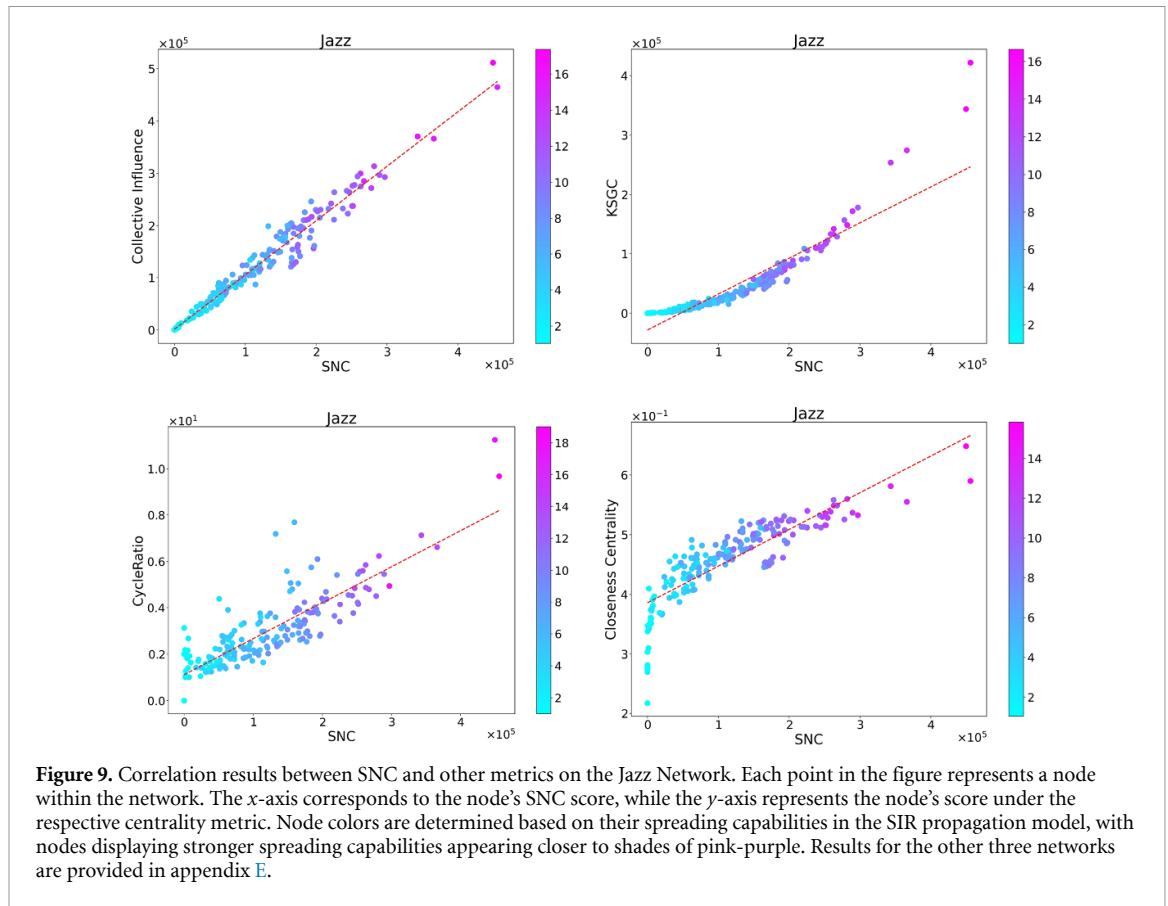
Figure 8. The distribution of SNC scores on four empirical networks. In these figures, each point represents a node within the network. Green nodes indicate that SNC scores fall within the confidence interval provided by the 0th-order null model. Blue nodes represent SNC scores within the confidence interval of the 1st-order null model but outside the confidence interval of the 0th-order null model. Red nodes signify SNC scores that exceed the confidence interval of the 1st-order null model. Results for all networks are provided in appendix D.

to use statistical comparisons of relative results to illustrate the empirical properties of nodes or networks. In this paper, we accomplish this task by constructing random networks that share certain properties with empirical networks. Such random networks are also referred to as null models in statistical methods [60, 61].

Generally, there are two types of methods for constructing null models of complex networks. One approach involves first calculating certain statistical properties of the network and then generating random networks based on these properties to create null models. While this method is straightforward, it is challenging to construct higher-order null models using it. The other approach involves random edge rewiring based on the original network, aiming to preserve the original network's properties while randomizing it. This method allows for the preservation or disruption of different network properties but is more complex to implement and computationally time-consuming, making it challenging for large-scale networks.

In this paper, we employ the method of constructing random networks to create two types of null models for complex networks. One null model has the same average degree as empirical networks, with all other properties randomized. We refer to this as the '0th-order null model.' The other null model maintains the same degree distribution as empirical networks, with all other properties randomized, and we call this the '1st-order null model.' In fact, higher-order null models could be constructed based on more constraints, such as having the same joint degree distribution or same joint degree distribution along with the average clustering coefficient as empirical networks [61]. However, as this paper's primary focus lies elsewhere and the generation of higher-order null models is more complex, we do not discuss them here, leaving them for future work.

To explore the potential impact of certain non-random characteristics in empirical networks on SNC, we constructed 500 instances of 0th-order null models and 500 instances of 1st-order null models. We calculated the distribution of node SNC scores in these null models and obtained confidence intervals for SNC scores. Figure 8 and appendix D depict the distribution of SNC scores on empirical networks. In these figures, each point represents a node in the network. Green nodes indicate SNC scores falling within the confidence interval provided by the 0th-order null model. Blue nodes represent SNC scores within the confidence interval of the 1st-order null model but outside the confidence interval of the 0th-order null model. Red nodes represent SNC scores exceeding the confidence interval of the 1st-order null model.



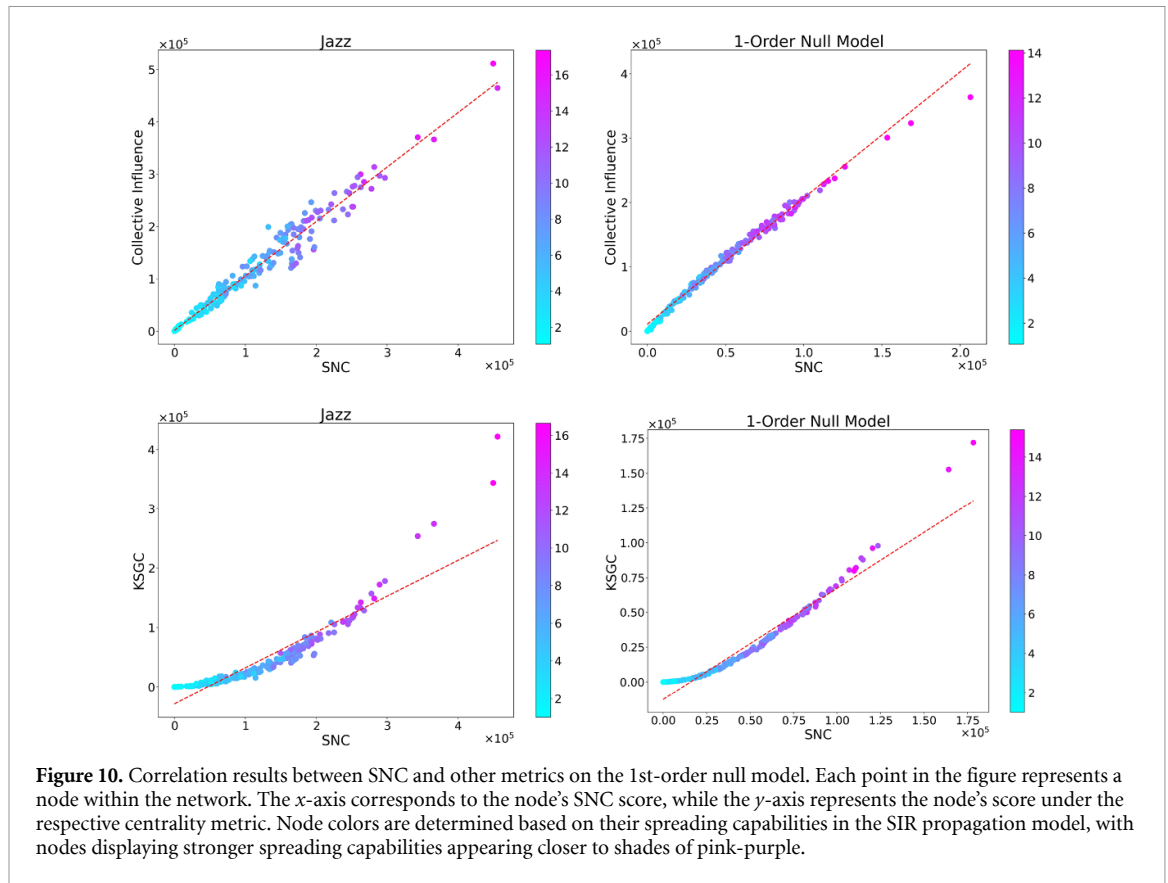
From the figures, it can be observed that in the Jazz, NS_GC, and Yeast networks, most of the nodes have SNC scores falling within the confidence interval of the 0th-order null model. There are also some nodes that exceed the confidence interval of the 0th-order null model but fall within the confidence interval of the 1st-order null model. Furthermore, many nodes surpass the confidence interval of the 1st-order null model. This suggests that SNC is not highly sensitive to the degree of nodes, and certain unique structures and properties existing in real networks may have a greater impact on it. Therefore, SNC can identify important nodes in empirical networks that possess special structures and properties not typically found in random networks. It is worth noting that in the Minnesota network, all nodes' SNC scores fall within the confidence interval of the 0th-order null model, which might be related to the sparsity of the Minnesota network.

In conclusion, determining which features in empirical network data differ from those in random networks and assessing the contributions of clustering coefficients and structural hole constraints to SNC rely on the analysis of higher-order null models. For example, constructing random networks with the same joint degree distribution and clustering coefficient as the empirical network would allow an analysis of the contribution of structural hole constraints to SNC. This is one of the tasks we are planning to undertake in our future work.

6.3. Correlation analysis

Correlation analysis can examine potential associations between different centrality metrics. If centrality metrics yield highly correlated rankings of node importance, it indicates that they provide similar information. Conversely, if they are uncorrelated, it suggests that they provide distinct information. In this study, we conducted correlation tests between SNC and four other centrality metrics: CI, KSGC, CR, and CC, on the Jazz, NS_GC, Yeast, and Minnesota networks, and the results are shown in figure 9 and appendix E.

In figure 9, each point represents a node in the network. The x-axis represents the node's SNC score, the y-axis represents the node's score under another centrality metric, and the color represents the node's spreading capability in the SIR model. Nodes with higher spreading capability are closer to pink-purple in color. From the figure, it can be seen that SNC has some correlation with CI and KSGC. Additionally, the correlation between SNC and KSGC is higher than that between SNC and CI. However, there is almost no



correlation between SNC and CR and CC. Furthermore, nodes with high SNC scores also tend to have high spreading capabilities, which indirectly supports the feasibility of SNC.

An interesting observation is the high exponential correlation between SNC and KSGC and CI. SNC assesses node importance using information about node degree, clustering coefficients, and structural hole constraints. CI relies on the degree information of high-order neighbors, while KSGC utilizes ks values and degree information. In essence, ks values are related to node degrees. Therefore, ks values, clustering coefficients, and structural hole constraints are not inherently correlated. In this context, SNC and KSGC should not exhibit such a high level of correlation. To investigate whether the correlation between SNC and KSGC is driven by degrees, we conducted tests using the 1st-order null model (which preserves the same degree distribution as the original network).

Figure 10 shows the correlation results between SNC and KSGC, as well as SNC and CI, on the 1st-order null model of the Jazz network. It can be observed that SNC and KSGC, as well as SNC and CI, exhibit stronger correlations in the 1st-order null model of the Jazz network. This suggests that in random networks, SNC and KSGC primarily rely on degree information to assess node importance, and the correlation between them is indeed driven by the shared use of degree information. In real networks, the correlation between SNC and these two metrics decreases. This may be due to SNC capturing certain features not present in random networks, such as nodes with high clustering coefficients or nodes occupying more structural hole positions in real networks. Consequently, SNC provides additional information that enhances its accuracy, which is challenging for KSGC to uncover.

7. Conclusion

Precisely locating highly influential nodes in a network is a problem of significant practical importance. The proposed SNC method in this paper draws inspiration from the mass formula in physics. It not only takes into account a node's degree and structural hole information but also incorporates the often overlooked clustering coefficient. Therefore, SNC not only identifies important nodes within a network but also provides nearly unique measurements of each node's importance. Experimental results from performance testing show that when the transmission rate is β , SNC achieved the highest Kendall Tau correlation coefficient in 10 out of 12 networks, and in the remaining two networks, SNC ranked close to the top with a minimal difference from the first place. Thus, overall, SNC exhibits better accuracy. Additionally, SNC also exhibits

higher monotonicity, with an average monotonicity value of 0.9938, surpassing other baseline methods. This indicates that SNC can accurately measure the importance of each node. In terms of identifying the Top-k most influential nodes within a network, SNC also performs admirably. While in some networks, the accuracy and ability of SNC to identify the Top-k nodes may be slightly weaker than certain methods, overall, SNC's performance remains superior.

Further analysis of SNC reveals that it consistently demonstrates top-tier performance in various networks and has the capability to capture latent structural features in real networks, such as clustering coefficients or structural hole constraints, providing a solid foundation for its high accuracy. Nevertheless, SNC still has some limitations. Firstly, SNC exhibits lower performance in sparse graphs. This is because SNC assesses node importance based on the local clustering coefficients of nodes and their neighbors. In sparse graphs or some low-clustering-coefficient random networks, SNC can only rely on node degrees to assess importance. Secondly, SNC still exhibits a high degree of correlation with KSGC. Compared to KSGC, SNC only provides some specific information that exists in real networks but is absent in random networks. Furthermore, the design philosophy of SNC does not consider the position of nodes within the network. Although node position is not necessarily entirely related to node influence, it remains one of the node's essential attributes. Additionally, SNC only considers a node's second-order neighbors, while higher-order neighbors of nodes may also have an impact, necessitating further exploration.

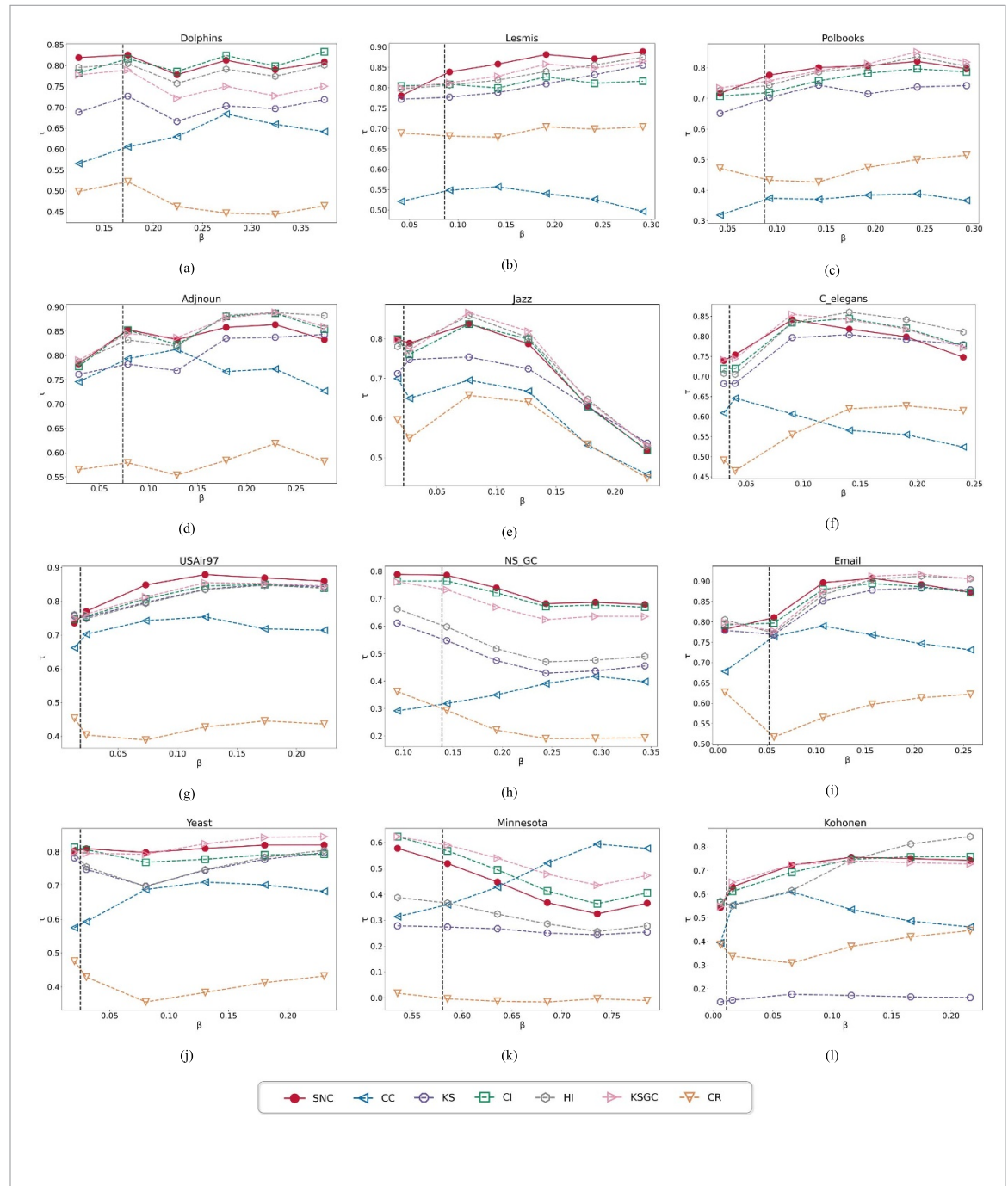
Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

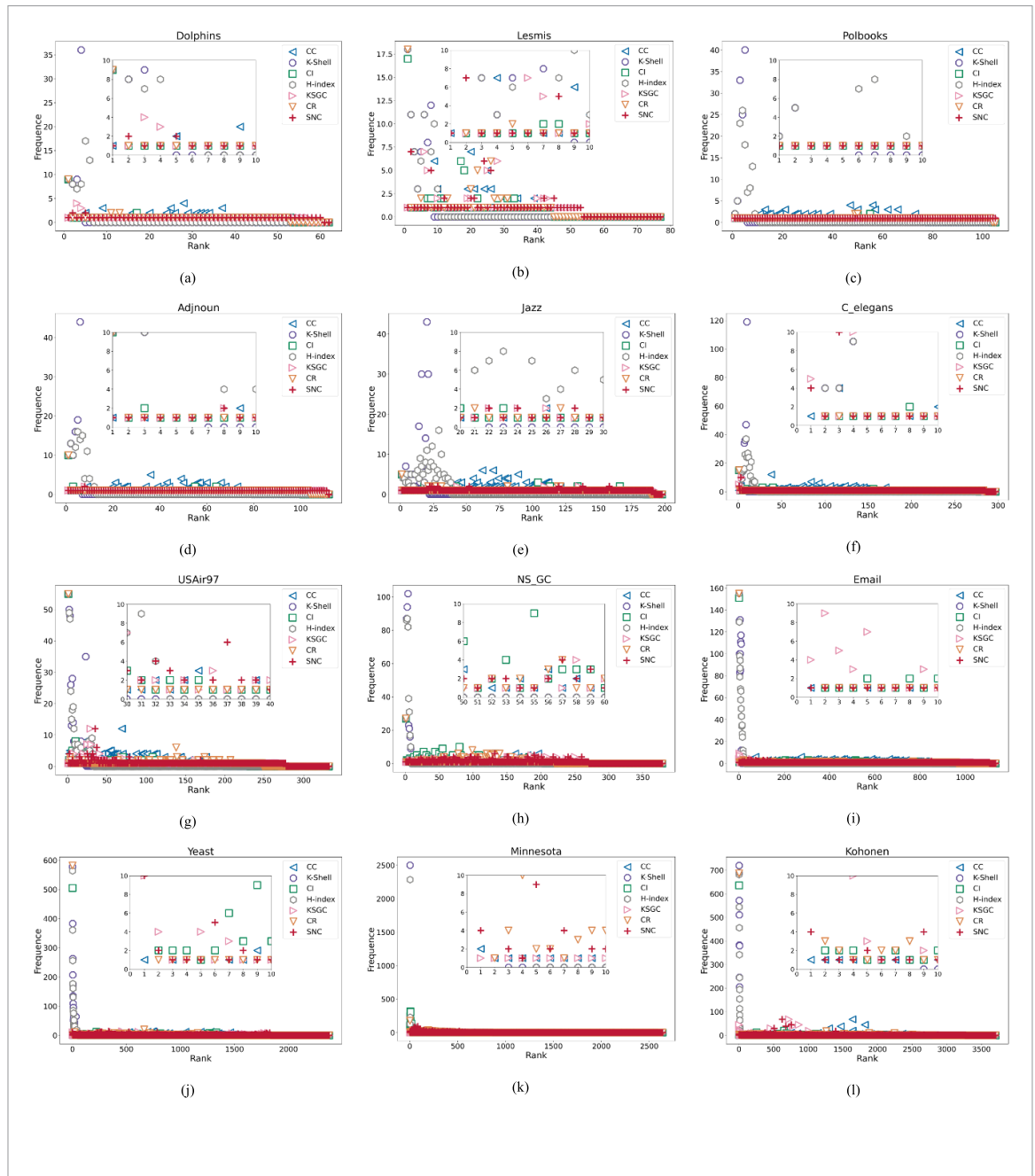
Acknowledgments

This work was funded by the Key Research and Development Program of Yunnan Province (202102AA100021); the National Natural Science Foundation of China (62066048) and (62366057); Demonstration project of comprehensive government management and large-scale industrial application of the major special project of CHEOS: 89-Y50G31-9001-22/23; the Science Foundation of Yunnan Province (202101AT070167) and supported by a grant from Key Laboratory for Crop Production and Smart Agriculture of Yunnan Province.

Appendix A. The accuracy performance of different node centrality metrics at various infection probabilities



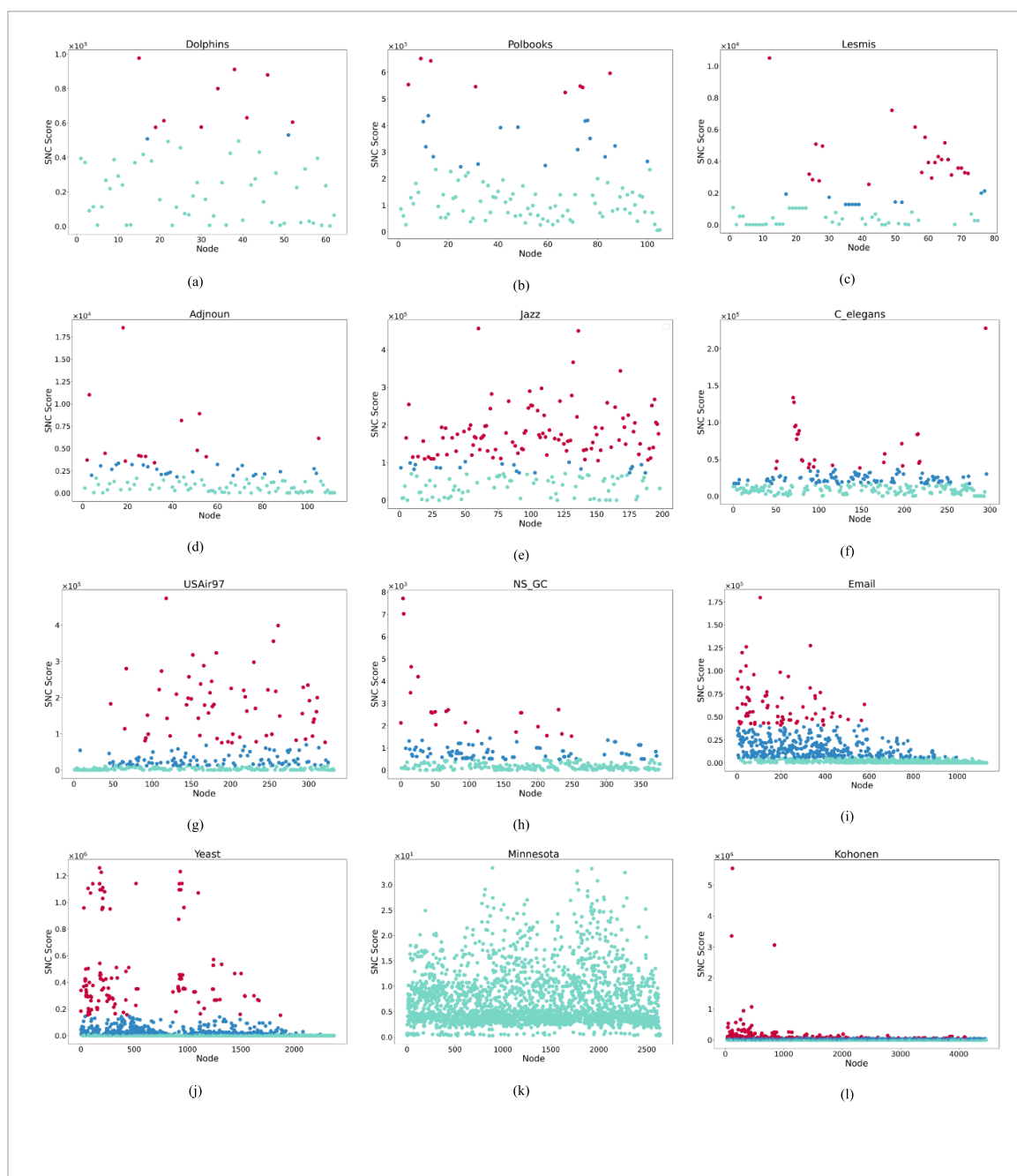
Appendix B. The rank distribution plots of different node centrality metrics across different networks



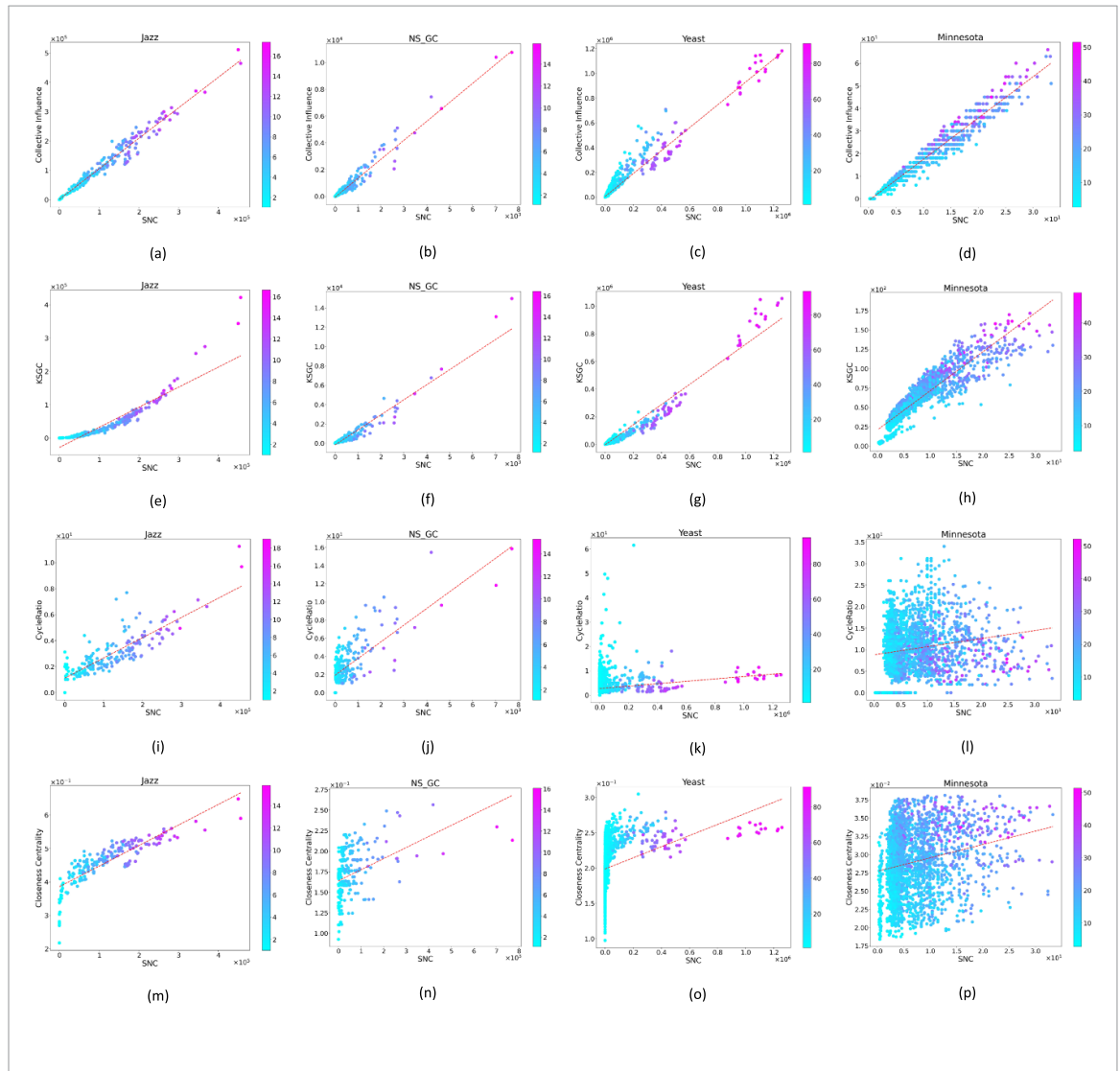
Appendix C. The Jaccard similarity coefficient results for different node centrality metrics



Appendix D. The distribution of SNC scores across different networks



Appendix E. The distribution of SNC scores across different networks



ORCID iD

Qian Liu  <https://orcid.org/0000-0002-3252-2845>

References

- [1] Boers N, Goswami B, Rheinwalt A, Bookhagen B, Hoskins B and Kurths J 2019 Complex networks reveal global pattern of extreme-rainfall teleconnections *Nature* **566** 373
- [2] Kamakshi S and Sriram V S S 2020 Modularity based mobility aware community detection algorithm for broadcast storm mitigation in VANETs *Ad Hoc Netw.* **104** 102161
- [3] Guanghui W, Yufei W, Jimei L and Kaidi L 2021 A multidimensional network link prediction algorithm and its application for predicting social relationships *J. Comput. Sci.* **53** 101358
- [4] Sui L, Yu J, Cang D, Miao W, Wang H, Zhang J, Yin S and Chang K 2019 The fractal description model of rock fracture networks characterization *Chaos Solitons Fractals* **129** 71–76
- [5] Zanin M, Sun X and Wandelt S 2018 Studying the topology of transportation systems through complex networks: handle with care *J. Adv. Trans.* **2018** 3156137
- [6] Grimm T W, Li C and Palti E 2019 Infinite distance networks in field space and charge orbits *J. High Energy Phys.* **JHEP03(2019)016**
- [7] Lu L, Chen D, Ren X-L, Zhang Q-M, Zhang Y-C and Zhou T 2016 Vital nodes identification in complex networks *Phys. Rep.* **650** 1–63
- [8] Yao S, Fan N and Hu J 2022 Modeling the spread of infectious diseases through influence maximization *Optim. Lett.* **16** 1563–86
- [9] Garas A, Argyrakis P, Rozenblat C, Tomassini M and Havlin S 2010 Worldwide spreading of economic crisis *New J. Phys.* **12** 113043
- [10] Ni Q, Guo J, Huang C and Wu W 2020 Community-based rumor blocking maximization in social networks: algorithms and analysis *Theor. Comput. Sci.* **840** 257–69

- [11] Liao H, Mariani M S, Medo M, Zhang Y-C and Zhou M-Y 2017 Ranking in evolving complex networks *Phys. Rep.* **689** 1–54
- [12] Alshahrani M, Fuxi Z, Sameh A, Mekouar S and Huang S 2020 Efficient algorithms based on centrality measures for identification of top-K influential users in social networks *Inf. Sci.* **527** 88–107
- [13] Wen T, Pelusi D and Deng Y 2020 Vital spreaders identification in complex networks with multi-local dimension *Knowl.-Based Syst.* **195** 105717
- [14] Zareie A, Sheikhamadi A and Jalili M 2019 Identification of influential users in social networks based on users' interest *Inf. Sci.* **493** 217–31
- [15] Watts D J and Strogatz S H 1998 Collective dynamics of 'small-world' networks *Nature* **393** 440–2
- [16] Ma L, Ma C, Zhang H-F and Wang B-H 2016 Identifying influential spreaders in complex networks based on gravity formula *Physica A* **451** 205–12
- [17] Li Z, Ren T, Ma X, Liu S, Zhang Y and Zhou T 2019 Identifying influential spreaders by gravity model *Sci. Rep.* **9** 8387
- [18] Batagelj V and Zaversnik M 2003 An $O(m)$ algorithm for cores decomposition of networks (arXiv:cs/0310049)
- [19] Zeng A and Zhang C-J 2013 Ranking spreaders by decomposing complex networks *Phys. Lett. A* **377** 1031–5
- [20] Bae J and Kim S 2014 Identifying and ranking influential spreaders in complex networks by neighborhood coreness *Physica A* **395** 549–59
- [21] Li C, Wang L, Sun S and Xia C 2018 Identification of influential spreaders based on classified neighbors in real-world complex networks *Appl. Math. Comput.* **320** 512–23
- [22] Zareie A and Sheikhamadi A 2018 A hierarchical approach for influential node ranking in complex social networks *Expert Syst. Appl.* **93** 200–11
- [23] Wang Z, Zhao Y, Xi J and Du C 2016 Fast ranking influential nodes in complex networks using a k-shell iteration factor *Physica A* **461** 171–81
- [24] Maji G, Namtirtha A, Dutta A and Malta M C 2020 Influential spreaders identification in complex networks with improved k-shell hybrid method *Expert Syst. Appl.* **144** 113092
- [25] Maji G 2020 Influential spreaders identification in complex networks with potential edge weight based k-shell degree neighborhood method *J. Comput. Sci.* **39** 101055
- [26] Koene J 1984 Applied network analysis—a methodological introduction *Eur. J. Oper. Res.* **17** 422–3
- [27] Chen D, Lu L, Shang M-S, Zhang Y-C and Zhou T 2012 Identifying influential nodes in complex networks *Physica A* **391** 1777–87
- [28] Tulu M M, Hou R and Younas T 2018 Identifying influential nodes based on community structure to speed up the dissemination of information in complex network *IEEE Access* **6** 7390–401
- [29] Wang S, Du Y and Deng Y 2017 A new measure of identifying influential nodes: efficiency centrality *Commun. Nonlinear Sci. Numer. Simul.* **47** 151–63
- [30] Salavati C, Abdollahpouri A and Manbari Z 2019 Ranking nodes in complex networks based on local structure and improving closeness centrality *Neurocomputing* **336** 36–45
- [31] Fei L, Zhang Q and Deng Y 2018 Identifying influential nodes in complex networks based on the inverse-square law *Physica A* **512** 1044–59
- [32] Ullah A, Wang B, Sheng J, Long J, Khan N and Sun Z 2021 Identifying vital nodes from local and global perspectives in complex networks *Expert Syst. Appl.* **186** 115778
- [33] Wang M, Li W, Guo Y, Peng X and Li Y 2020 Identifying influential spreaders in complex networks based on improved k-shell method *Physica A* **554** 124229
- [34] Xu X, Zhu C, Wang Q, Zhu X and Zhou Y 2020 Identifying vital nodes in complex networks by adjacency information entropy *Sci. Rep.* **10** 2691
- [35] Fan C, Zeng L, Sun Y and Liu Y-Y 2020 Finding key players in complex networks through deep reinforcement learning *Nat. Mach. Intell.* **2** 317–24
- [36] Yu E-Y, Wang Y-P, Fu Y, Chen D-B and Xie M 2020 Identifying critical nodes in complex networks via graph convolutional networks *Knowl.-Based Syst.* **198** 105893
- [37] Du Y, Gao C, Hu Y, Mahadevan S and Deng Y 2014 A new method of identifying influential nodes in complex networks based on TOPSIS *Physica A* **399** 57–69
- [38] Kuo T 2017 A modified TOPSIS with a different ranking index *Eur. J. Oper. Res.* **260** 152–60
- [39] Yang B, Wen D, Qin L, Zhang Y, Chang L and Li R-H 2019 Index-based optimal algorithm for computing K-cores in large uncertain graphs 2019 IEEE 35th Int. Conf. on Data Engineering (ICDE 2019) (IEEE) pp 64–75
- [40] Freeman L 1979 Centrality in social networks conceptual clarification *Soc. Netw.* **1** 215–39
- [41] Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H E and Makse H A 2010 Identification of influential spreaders in complex networks *Nat. Phys.* **6** 888–93
- [42] Hirsch J E 2005 An index to quantify an individual's scientific research output *Proc. Natl Acad. Sci. USA* **102** 16569–72
- [43] Lue L, Zhou T, Zhang Q-M and Stanley H E 2016 The H-index of a network node and its relation to degree and coreness *Nat. Commun.* **7** 10168
- [44] Wu X, Wei W, Tang L, Lu J and Lu J 2019 Coreness and h-index for weighted networks *IEEE Trans. Circuits Syst. I* **66** 3113–22
- [45] Morone F and Makse H A 2015 Influence maximization in complex networks through optimal percolation *Nature* **524** 65–8
- [46] Fan T, Lu L, Shi D and Zhou T 2021 Characterizing cycle structure in complex networks *Commun. Phys.* **4** 272
- [47] Yang X and Xiao F 2021 An improved gravity model to identify influential nodes in complex networks based on k-shell method *Knowl.-Based Syst.* **227** 107198
- [48] Burt R S 2004 Structural holes and good ideas *Am. J. Soc.* **110** 349–99
- [49] Lusseau D, Schneider K, Boisseau O J, Haase P, Slooten E and Dawson S M 2003 The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations—can geographic isolation explain this unique trait? *Behav. Ecol. Sociobiol.* **54** 396–405
- [50] Rossi R and Ahmed N 2015 The network data repository with interactive graph analytics and visualization *Proc. 29th Aaai Conf. on Artificial Intelligence (Assoc Advancement Artificial Intelligence)* pp 4292–3
- [51] Newman M E J 2006 Finding community structure in networks using the eigenvectors of matrices *Phys. Rev. E* **74** 036104
- [52] Gleiser P M and Danon L 2003 Community structure in jazz *Adv. Complex Syst.* **6** 565–73
- [53] Guimera R, Danon L, Diaz-Guilera A, Giralt F and Arenas A 2003 Self-similar community structure in a network of human interactions *Phys. Rev. E* **68** 065103

- [54] Bu D B *et al* 2003 Topological structure analysis of the protein-protein interaction network in budding yeast *Nucleic Acids Res.* **31** 2443–50
- [55] Kermack W O and Mckendrick A 1927 A contribution to the mathematical theory of epidemics *Proc. R. Soc. A* **115** 700–21
- [56] Simsek A 2022 Lexical sorting centrality to distinguish spreading abilities of nodes in complex networks under the susceptible-infectious-recovered (SIR) model *J. King Saud Univ. Comput. Inf. Sci.* **34** 4810–20
- [57] Zhao Z, Li D, Sun Y, Zhang R and Liu J 2023 Ranking influential spreaders based on both node k-shell and structural hole *Knowl.-Based Syst.* **260** 110163
- [58] Hebert-Dufresne L, Allard A, Young J-G and Dube L J 2013 Global efficiency of local immunization on complex networks *Sci. Rep.* **3** 2171
- [59] Zareie A, Sheikahmadi A, Jalili M and Fasaie M S K 2020 Finding influential nodes in social networks based on neighborhood correlation coefficient *Knowl.-Based Syst.* **194** 105580
- [60] Mahadevan P, Hubble C and Krioukov D 2007 Orbis: rescaling degree correlations to generate annotated internet topologies *ACM SIGCOMM Computer Communication Review* vol 37 pp 325–36
- [61] Gjoka M, Kuran M and Markopoulou A 2013 2.5K-graphs: from sampling to generation 2013 *Proc. IEEE INFOCOM* pp 1968–76