Journal of Computational Science 69 (2023) 102030

Contents lists available at ScienceDirect

Journal of Computational Science

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

Identifying influential nodes by leveraging redundant ties

Bitao Dai [a], Shuo Qin [b], Suoyi Tan [a], Chuchu Liu [a], Jianhong Mou [a], Hongzhong Deng [a], Fredrik Liljeros [c,d], Xin Lu [a,d,*,1]

[a] College of Systems Engineering, National University of Defense Technology, Changsha 410073, China
[b] State Key Laboratory on Blind Signal Processing, Chengdu 610041, China
[c] Department of Sociology, Stockholm University, Stockholm 10691, Sweden
[d] Department of Global Public Health, Karolinska Institutet, Stockholm 17177, Sweden

ARTICLE INFO

Keywords:
Complex networks
Influential nodes identification
Spanning tree
Redundant ties

ABSTRACT

Structure-based influential nodes identification is a long-term challenge in the study of complex networks. While global centrality-based approaches are generally considered to be more accurate and reliable, the requirements of complete network information and high computational complexity are hard to meet, limiting their applications in many practical scenarios. In addition, recent studies have highlighted the effect of cyclic structures introducing redundant paths in network connectivity and exaggerating the importance of traditional centrality measures. In this work, we develop a new centrality metric, called Multi-Spanning Tree-based Degree Centrality (MSTDC), to quantify node importance with linear complexity by leveraging redundant ties. MSTDC is calculated using the aggregation of degrees of a small number of spanning trees constructed with a few randomly selected root nodes. Experiments on synthetic and empirical networks reveal that MSTDC obtains superior performance than other benchmark network centralities in identifying influential nodes from the perspective of both maintaining network connectivity and maximizing spreading capacity. In addition, we find that MSTDC is extraordinarily effective in networks with high clustering coefficients. Our study provides novel insights into the role of redundant ties in network structural and functional analyses.

1. Introduction

The study of identifying influential nodes has attracted extensive attention from various scientific communities [1–7]. As one of the most significant topics in complex networks, identifying influential nodes has prominent practical applications such as suppressing epidemic spreading [8,9], stopping rumor propagation [10,11], promoting product adoption [12] and improving the robustness of route networks and power networks against malicious attacks [13,14].

Influential nodes are generally located in topological positions of great significance in a network [15–18]. Most existing studies concerning structure-based influential nodes identification can be classified as using either local information-based or global information-based methods. A variety of methods based on network topology have been proposed to identify influential nodes according to the nodes' positions, such as degree centrality [19], second order degree centrality [20], *k*-core [21], etc. Degree centrality, characterized as the number of direct neighbors of a node, is the most representative and widely used indicator. However, in some scenarios, it cannot effectively identify influential nodes. For instance, when a node has few direct neighbors, whereas the neighbors are opinion leaders, its influence may be more important than a node with abundant immediate neighbors. To solve this problem, [24] developed a centrality index to measure the importance of nodes by considering neighbors of second-order, i.e., the neighbors of neighbors. Furthermore, Kitsak et al. proposed the well-known *k*-shell decomposition method to split networks to search for core nodes, but this method performs poorly in some situations. For example, in a propagation tree [25,26], all nodes are in the same core and are regarded as equally important. However, some researchers have argued that the tree structure may be hierarchical, and the nodes closer to the root are more important [27]. Recently, local indices such as DWT and DSHC [22,23], which consider the connection between nodes and their neighbors, have been developed. While they presented satisfactory performances compared to traditional local metrics, they could not

* Corresponding author at: College of Systems Engineering, National University of Defense Technology, Changsha 410073, China.
E-mail address: xin_lyu@sina.com (X. Lu).
[1] ORCID: 0000-0002-3547-6493.

https://doi.org/10.1016/j.jocs.2023.102030
Received 11 August 2022; Received in revised form 20 March 2023; Accepted 8 April 2023
Available online 10 April 2023

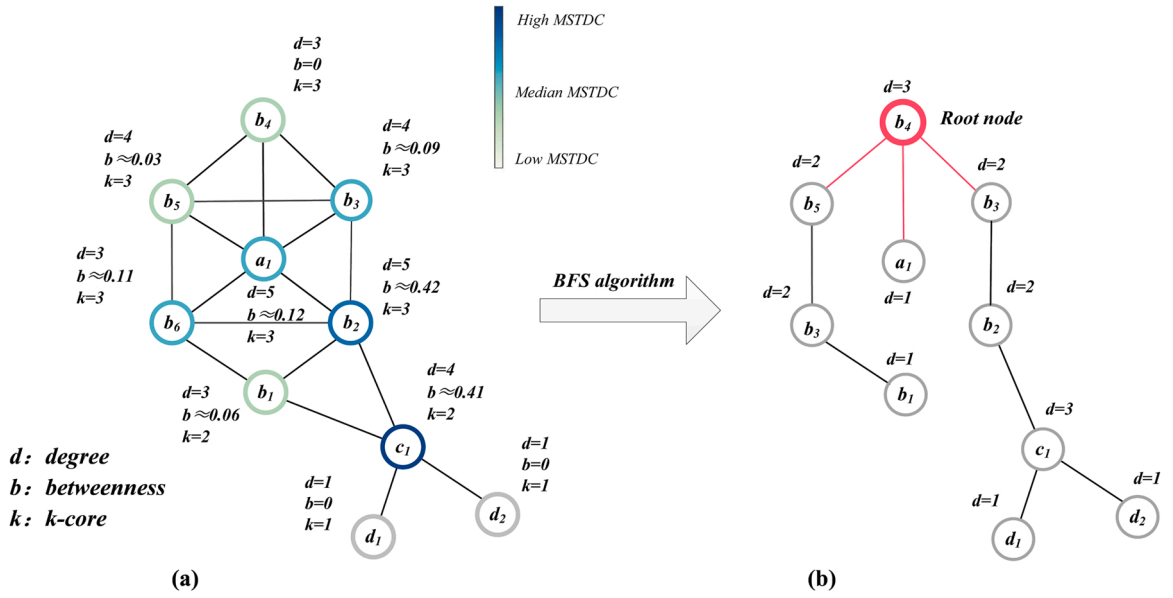


Fig. 1. Difference between the initial network and the spanning tree. (a) the initial network. (b) a spanning tree generated with the single root node b_4 according to BFS algorithm.

outpace global metrics with fast computing velocities.

Global metrics such as betweenness centrality [28], eigenvector centrality [29,30], and gravity models [31] have been proposed to combine several global or local centralities. Although these approaches are usually capable of identifying influential nodes with improved accuracy, they require complete network information and have high computational complexity, limiting their applications in many practical scenarios.

In recent years, the roles of clustering and redundancy concerning the identification of influential nodes in networks have attracted increasing attention. Generally, there is an enormous amount of interplay between nodes in a large-scale network. A node can contact other nodes directly or indirectly, (through several steps) and there may be substantial cyclic structures providing redundant paths to the involved node pairs. It has been validated that nodes with high clustering coefficients are generally not critical, while weak ties play a vital role in the propagation process [32–36]. Liu has confirmed that redundant ties could form locally densely connected core-like group, which may distort influential nodes identification result by traditional metrics [37]. Recently, Lü et al. have called for in-depth studies of cyclic structures, e.g., triangles and cycles, and their roles in network topology and dynamics [38]. In addition, Fan has proposed the Cycle Ratio enclosing the information of redundant ties from the perspective of cycles and confirmed that redundant ties strongly influence node importance [39]. Nevertheless, methods of considering redundant ties through cycles often come with high computational complexity. The structures formed by redundant ties are more complicated in reality and should be measured from multiple perspectives. There is still a prodigious gap in the research for comprehensively and effectively considering redundant ties in influential nodes detection.

To overcome these limitations, we propose a new centrality called Multi-Spanning Tree-based Degree Centrality (MSTDC), calculated by aggregating the degrees of a small number of spanning trees constructed with a few randomly selected root nodes. It considers neighbors' effects on nodes and interactions by incorporating local and global information. Furthermore, the proposed approach can quantify nodes' global importance by leveraging redundant ties' side effects. Experiments on synthetic and empirical networks reveal that MSTDC obtains superior performance compared to other benchmark network indices in terms of maintaining network connectivity and maximizing spreading capacity, especially in networks with high clustering coefficients. The rest of the paper is organized as follows: In Section 2, we implement the theoretical analysis and compute the complexity of our method. In Section 3, we describe the synthetic networks and empirical networks data as well as the evaluation metrics used in the study. In Section 4, we compare our method with some representative methods in different synthetic and real-world networks. Finally, in Section 5, we summarize the contributions of this study and discuss future work.

2. Methodology

2.1. Motivation

In this work, we develop an approach to eliminate the side effects of redundant ties and propose a metric to effectively identify influential nodes in networks. To illustrate the motivation of our approach, assuming that a virus starts from an arbitrary node in Fig. 1(a), it will transmit the virus to its neighbors in the next step. The critical step to suppress the spreading is to immunize or disintegrate the network through a set of key nodes. If we take measures based on network metrics, we find that the largest degree nodes are a_1 and b_2, the highest betweenness centrality nodes are b_2, while the highest k-core nodes are b_2, b_3, b_4, b_5, b_6, and a_1, respectively. We soon discover that any of these metrics cannot account for the fundamental importance of virus spreading, given any node being infected: the virus will spread further through c_1, because it has more outward-extending nodes [25].

Given any node except c_1 being immunized, the virus can still spread globally no matter which node is infected, with an average infection size of 9 nodes. Supposing c_1 is vaccinated, the virus cannot spread further if it starts from d_1 or d_2. In any case, the virus spreads locally with an average infection size of 7 nodes when it starts from an arbitrary node other than c_1, d_1, and d_2. The expected infection size is approximately 5.67 nodes in the case of immunizing c_1. Consequently, it is most efficient to immunize c_1 to suppress virus diffusion in this network, although c_1 is not the most influential node according to traditional network indices. The difficulty in characterizing node importance based on degree, betweenness, or k-core in this example can be attributed to the triangles and loops, which provide multiple paths to all involved node pairs and form a special structure called local coupling, a ubiquitous network structure, where nodes are tightly connected. On the one hand, the nodes in local coupling will form core-like [40] cliques, and they always stay in the same core, resulting in numerous pseudo-core

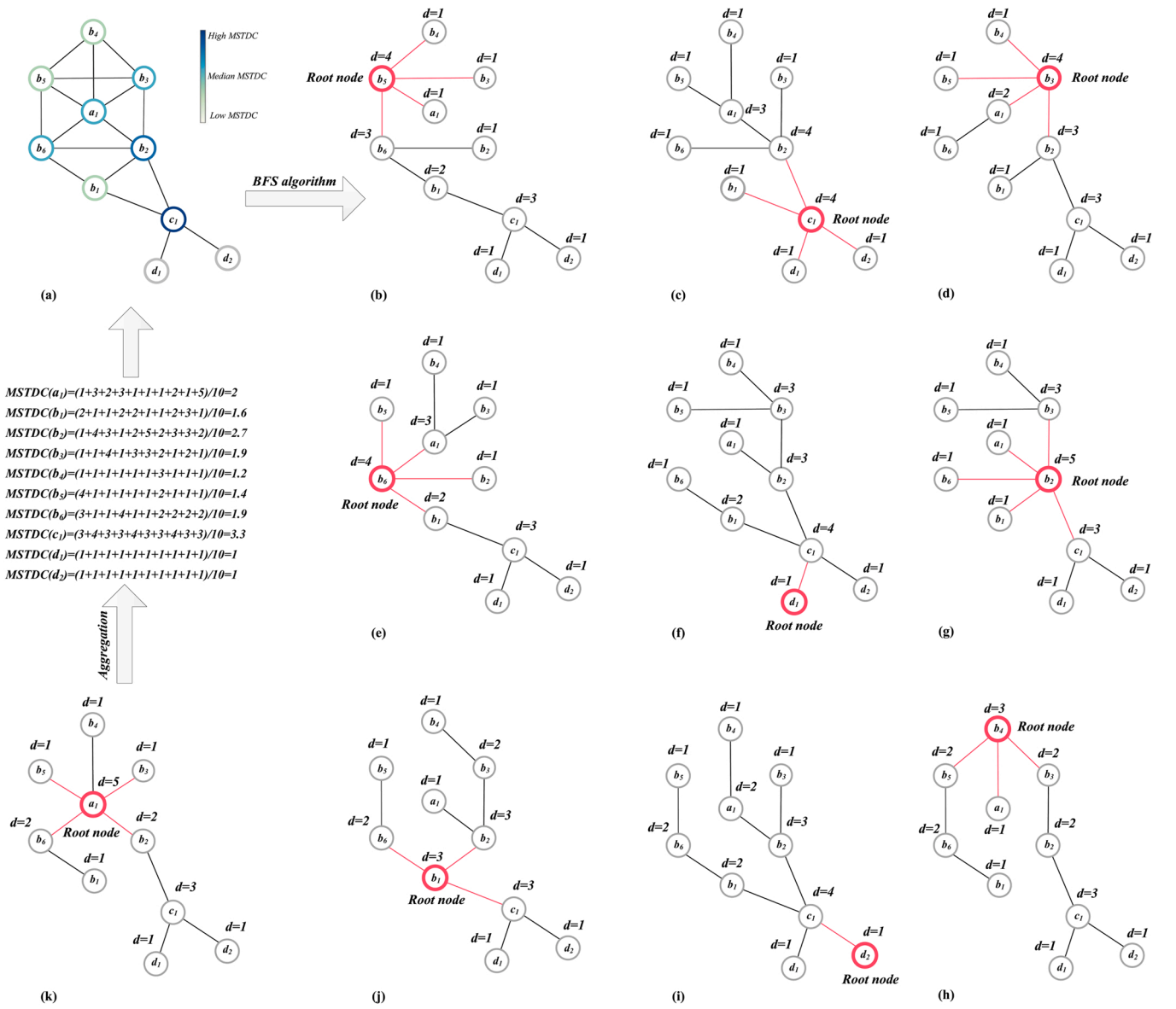


Fig. 2. Illustration of MSTDC calculation. (a) the initial network. (b)-(k) ten spanning trees generated with different root nodes.

nodes [25]. On the other hand, local coupling provides redundant paths to all involved nodes. For instance, if two nodes belong to a cycle, then there are at least two independent paths connecting them, which causes redundant ties. Consequently, the effects of redundant ties should be considered when measuring node importance.

2.2. Multi-spanning tree-based degree centrality

Based on the aforementioned discussions, we propose the construction of spanning trees [41] to reduce the effect of redundant ties and develop a new method, called MSTDC, to identify influential nodes. Let $G = (V, E)$ be a network with $N = |V|$ nodes and $M = |E|$ edges. In an undirected and unweighted network, a spanning tree is a connected subgraph of G with N nodes, $N-1$ edges, and no loops. According to these characteristics, there is only one path between any two nodes. In this work, we adopt breadth-first-search (BFS) algorithm [42] to construct spanning trees. In Fig. 1(b), we show the BFS spanning tree of the toy network, started from node b_4. Compared to Fig. 1(a), there are no cycles in Fig. 1(b) and the redundant ties and local coupling are appropriately eliminated. Then we propose the primary index referred to Spanning Tree-based degree centrality (STDC), represented by:

$$STDC(i) = \sum_{j \in \tau(i)} a_{ij}^{(i)} \tag{1}$$

where τ_i represents the neighbors of node i in the spanning tree started from node i, $a_{ij}^{(i)} = 1$ if there is an edge between i and j in the spanning tree, and otherwise, $a_{ij}^{(i)} = 0$.

It is worth noting that the branching process (i.e., determining which node is connected next) of the spanning tree and the selection of the root node could cause statistical fluctuations of STDC. In this work, we adopt equal probability as the branching rule to construct spanning trees. Each child node has the same probability of being connected first. Therefore, by aggregating the results of multiple experiments, the statistical fluctuations will be appropriately eliminated.

As the spanning trees may change substantially when the starting node is different, we randomly select a number of nodes as root nodes to construct multiple spanning trees and aggregate their topology information to eliminate the statistical fluctuations (Fig. 2). So, we propose a modified index called Multi-Spanning Tree-based degree centrality (MSTDC), calculated by:

$$MSTDC(i) = \frac{\sum_{k\in\sigma}\sum_{j\in\tau(i)} a_{ij}^{(k)}}{T} \tag{2}$$

where k represents the root node of one spanning tree, σ represents the randomly chosen node-set, and T represents the number of nodes in σ, i.e., the number of spanning trees.

As shown in Fig. 2, from the perspective of betweenness centrality, a node with high betweenness centrality also owns relatively high MSTDC, as does the degree. Nevertheless, we cannot see this phenomenon in terms of k-core. This means our method could reflect the characteristics of degree and betweenness centrality. b_2, along with the highest betweenness centrality, has a higher MSTDC value than a_1, although their degrees are equivalent and the best, which implies that our method incorporates more information of betweenness centrality. The highest MSTDC node is c_1, with a relatively high degree and betweenness centrality and it owns more outward-extending nodes than others, captured by our method.

The detailed procedure of MSTDC calculation is described in Algorithm 1. As the time complexity of BFS is $O(N + M)$, the complexity of MSTDC is $O(T(N + M))$. According to the results in Section 4.1, the value of T can be ignored in terms of network size. Consequently, the computational complexity of our method is $O(N + M)$, much lower than that of massive amounts of global network indices and much lower than indices considering the redundant ties through cycles, e.g., Cycle Ratio. For example, the complexity of betweenness centrality and closeness centrality is $O(MN^3)$ and $O(MN^2)$, respectively [24].

Algorithm 1. MSTDC Calculation.

Input: the giant connected component of original network $G = (V, E)$, number of spanning Trees T
Output: Ranking node list
Initialize a dictionary with key being the node ID and set all value to be zero, dic
Randomly choose T nodes to construct node set σ
For v in σ do:
  Adjacent edge list ← [ ]
  Q ← set(node)
  Visited node set ← set( )
  While Q is None do:
    v ← pop(Q)
    Visited node set = Visited node set $\bigcup$ v
    For i ← 1 to i ← length(neighbors set of v) do:
      Randomly visit one neighbor of v, denoted as u
      If u not in Visited node set:
        Q = Q $\bigcup$ u
        Adjacent edge list ←Adjacent edge list $\bigcup$ [v, u]
      End if
    End for
  End while
  $STDC_v$ ← dic{node: degree of node in network based on Adjacent edge list }
  For j in giant component node set do:
    $dic[j] \leftarrow dic[j] + STDC_v[j]$
  End for
End for
For key in dic.keys():
  $dic[key] \leftarrow \frac{dic[key]}{T}$
Ranking dic by value
Return index list of dic

3. Experimental setting

We evaluate the performance of MSTDC to identify influential nodes under two well-studied dynamical process, network disintegration [43–45] and epidemic spreading [46], whereby the importance of a node is measured according its criticality in maintaining the network connectivity or the outbreak size, respectively.

3.1. Data description

Synthetic networks. In order to explore the effects of redundancy on influential nodes identification, we use Watt-Strogatz (WS) model to

Table 1
Basic statistics of empirical networks. $\bar{k}$, $\bar{l}$, and C denote the average degree, the average shortest path lengths, and the average clustering coefficients of the network, respectively. $\beta^* = \bar{k}/\bar{k^2}$ is the epidemic threshold of a network and $\bar{k^2}$ is the second-order average degree. (These parameters correspond to the giant connected component of the original network.).

Networks	*N*	*M*	*Diameter*	$\bar{k}$	$\bar{l}$	*C*	β^*
CA-Grqc	4,158	26,580	17	12.915	6.049	0.557	0.056
CA-Hepph	11,204	117,649	13	20.996	4.673	0.62	0.008
Dolphin	62	159	6	5.129	3.111	0.248	0.147
Parsed Tap	1,373	6,833	12	9.953	5.224	0.529	0.061
Gnutella	62,561	147,878	11	4.728	5.936	0.005	0.086
Powergrid	4,941	6,594	46	2.67	18.99	0.10	0.258

generate various synthetic networks with varying clustering coefficients and average degrees. In the WS model, the network starts with a ring and each node is joined with $\bar{k}$ nearest neighbors. Then each edge $(u, v)$is rewired as $(u, w)$with probability p, where w is randomly chosen. By tuning p, the constructed networks are transiting between extremely regular $(p = 0)$and random $(p = 1)$. The average clustering coefficient (C) changes with p, where $C \approx 3/4$ for $p = 0$ and $C \approx \bar{k}/N$ for $p = 1$, respectively [47], i.e., a small p indicates a high clustering coefficient in WS networks. For the networks generated in this study, C ranges from 0.07 to 0.58 (see Table S1 for details).

Empirical networks. We use six empirical networks to validate the performance of MSTDC, including the collaboration network of general relativity and quantum cosmology from the e-print arXiv (CA-Grqc) [48], the collaboration network of high energy physics-phenomenology from the e-print arXiv (CA-Hepph) [49], the social network of bottlenose dolphins, observed between 1994 and 2001 (Dolphin) [50,51], the yeast protein-protein binding network generated by tandem affinity purification experiments (Parsed Tap) [52], the peer-to-peer file-sharing network of Gnutella [53], and the power grid network of the Western States of the United States of America (Powergrid) [54]. All networks mentioned above are considered to be undirected and unweighted, and only the giant connected components are extracted. Their basic topological features are summarized in Table 1.

3.2. Robustness measure

Network disintegration involves identifying the critical nodes or edges whose removal will lead to a network collapse. Given a fully connected network, if a number of nodes and their attached edges are removed, the network may be divided into many components, and the nodes in the giant connected component may decrease dramatically. We compute the ranking lists of nodes based on different indices, and the node with the largest index value is removed preferentially. As relevant research supports that dismantling 5–10 % of influential nodes is sufficient to meet most practical needs [56,57], we adopt the proportion of nodes in the giant connected component (s) after removing the Top-0.1 N nodes to reflect the capability of fast disintegration under a certain strategy. Besides, we use the area under the robustness curve R to capture the network's response when it suffers from attacks, which is defined as [55]:

$$R = \frac{1}{N}\sum_{Q=1}^{N} s(Q) \tag{3}$$

Where $s(Q)$ is the proportion of nodes in the giant connected component after removing Q nodes.

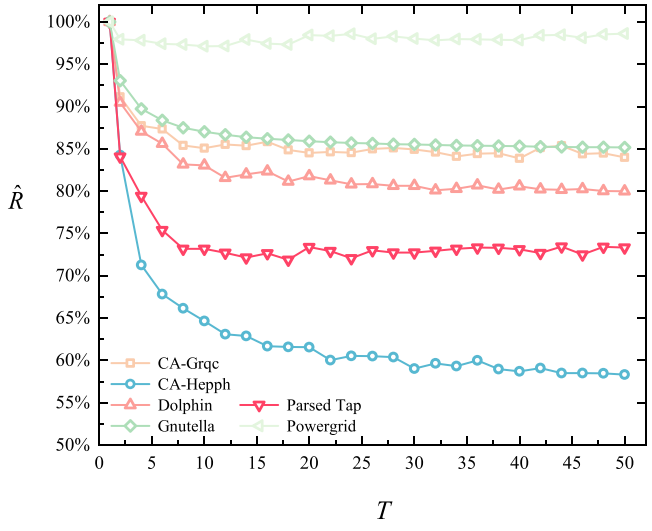


Fig. 3. The performance of MSTDC for network disintegration concerning $\widehat{R}$ when Tincreases. All results are averaged over 100 simulations.

3.3. Spreading capability

The spreading capability of nodes is widely used to evaluate the nodes' importance because the probability of accepting information or being infected relies on the source's influence [58]. We employ the standard susceptible- infected-recovered (SIR) model to simulate spreading dynamics. In the SIR model, each node belongs to either susceptible, infected, or recovered state. At each timestamp, an infected node will infect its neighbors with probability β and each infected node will transfer to recovered state with probability γ. For the sake of simplicity, γ is set as 1. As for β, we calculate the epidemic threshold of each network according to the mean-field approach (see Table 1 for details) and adjust the specific infection rate in the simulation to avoid the interference of inappropriate propagation probability. The spreading capability is quantified by the number of cumulative infected and recovered nodes denoted by F. The spreading process is terminated at step 100 or until no additional nodes can be infected.

4. Results

4.1. Comparison of STDC and MSTDC

In this section, we use empirical networks to explore the effects of the number of randomly sampled spanning trees, T, for the stability of MSTDC through network disintegration. In each experiment, we remove nodes and their attached edges in descending order according to a specific centrality, and then calculate R to measure the importance of nodes in maintaining network connectivity. When $T = 1$, MSTDC degenerates to STDC.

To compare the results on different networks, we analyze the relative performance gain of network disintegration $(\widehat{R})$ as a function of T, i.e., the number of spanning trees used to calculate MSTDC. $\widehat{R}$ is calculated as the ratio of R for when $T \geq 1$ and for when $T = 1$:

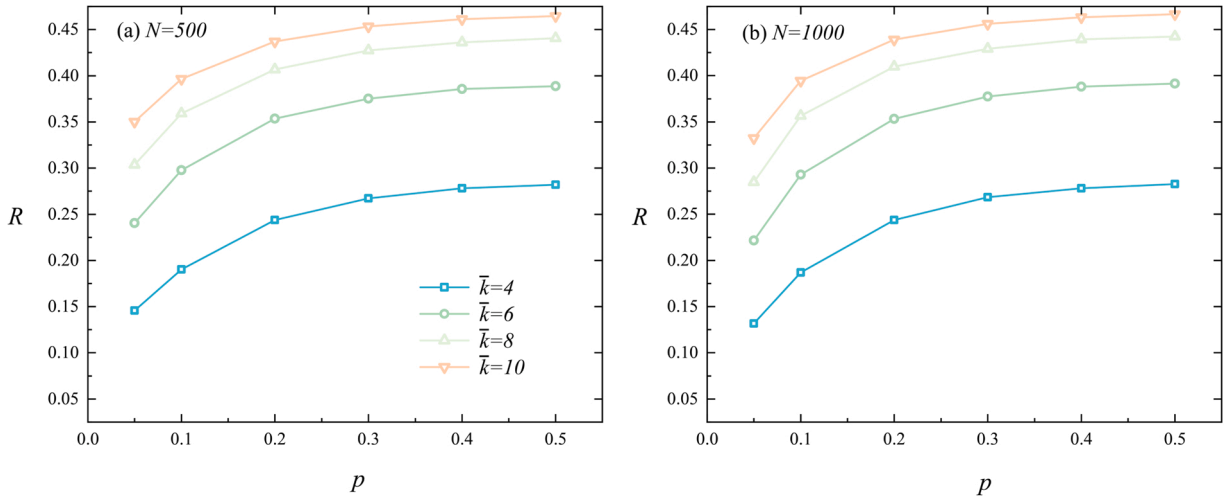


Fig. 4. R for WS networks disintegration with MSTDC. (a) $N = 500$; (b) $N = 1000$. All results are averaged over 100 simulations.

$$\widehat{R} = \frac{R_{MSTDC}}{R_{STDC}} \tag{4}$$

As shown in Fig. 3, $\widehat{R}$ changes rapidly when T increases from 1 to 10, and it becomes stable for $T \geq 15$ in most networks except Powergrid. The performance improvement of MSTDC in the tested networks ranges from 2 % to 40 %, respectively. Apparently, a small number of spanning trees used in MSTDC are capable of measuring the importance of nodes in maintaining network connectivity. Thus, the vital information missed by STDC can be integrated by MSTDC. For the remaining tests, we set $T = 30$ for the calculation of MSTDC to ensure that the results across different networks are comparable.

4.2. Network disintegration

In this section, we examine the performance of MSTDC in identifying influential nodes through network disintegration, by comparing it against a number of state-of-the-art methods: (1) degree centrality, (2) betweenness centrality (*Bet*), (3) eigenvector centrality (*Eig*), (4) *k*-core, (5) closeness centrality (*Clo*), (6) local clustering coefficients (*Clu*), (7) PageRank [59], (8) degree based on information entropy (Information Entropy) [60], and (9) Cycle Ratio.

4.2.1. Synthetic networks

For the synthetic networks, we first investigate the effects of network average degree $\bar{k}$ and clustering coefficient tunable parameter p on R. We generate 100 networks for each parameter setting in Table S1, and report the average values in Fig. 4. The results show that, when the average degree is fixed, R increase with p, i.e., higher average clustering coefficient implies lower R. For instance, when $\bar{k} = 4$, R increases from 0.14 to 0.28 and promotion rate varies from 31 % to 14 % as p increases from 0.05 to 0.5. R shows a strong correlation with $\bar{k}$ and their marginal benefits are asymptotically declined. For instance, when $p = 0.05$ and $N = 500$, R increases from 0.14 to 0.35 and the promotion rate ranges from 65 % to 15 %.

The performance of all disintegration methods on the WS networks are summarized and compared in Table S2. It is obvious that MSTDC generally outpaces other methods in most networks with a few

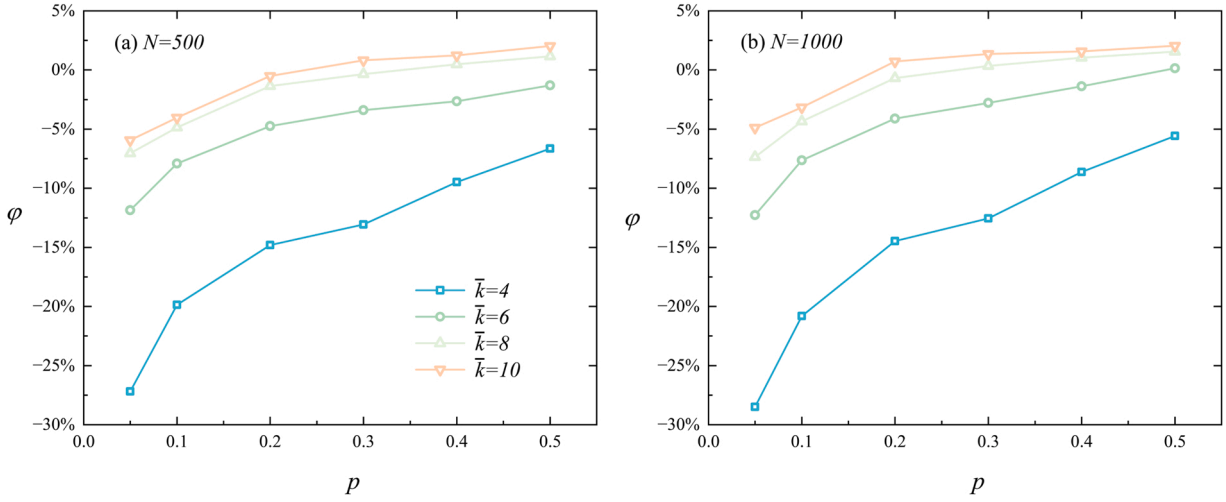


Fig. 5. Promotion rate φ for WS networks disintegration with MSTDC. (a) the network size is 500; (b) the network size 1000. All results are averaged over 100 simulations.

Table 2
Results of R on empirical networks.

Networks	*Information Entropy*	*Cycle Ratio*	*PageRank*	*Eig*	*Clo*	*K-core*	*Clu*	*Degree*	*Bet*	*MSTDC*
CA-Grqc	0.142	0.104	0.092	0.288	0.168	0.232	0.461	0.134	0.093	**0.090**
CA-Hepph	0.249	0.168	0.178	0.342	0.285	0.284	0.475	0.231	0.165	**0.161**
Dolphin	0.295	0.295	0.274	0.367	0.353	0.327	0.411	0.278	0.284	**0.262**
Parsed Tap	0.347	0.283	0.272	0.397	0.309	0.413	0.472	0.336	0.217	**0.209**
Gnutella	0.136	0.146	0.120	0.180	0.166	0.174	0.424	0.133	0.139	**0.114**
Powergrid	0.062	0.082	0.065	0.328	0.200	0.286	0.302	0.067	0.070	**0.056**

exceptions for $p = 0.4$ and $p = 0.5$, where MSTDC is relatively inferior to the optimal with average discrepancies of about 2 %. For each $\bar{k}$, the maximum discrepancy between MSTDC and the suboptimal is 13 %, for when $p = 0.05$. It implies that MSTDC is dramatically effective in high clustering networks compared to the state-of-art measures and the performance is also affected by the average degree.

To illustrate the superior performance of MSDTC, we select two classic centrality measures, i.e., degree and betweenness, for further comparison. The performance differences are measured through promotion rate φ (Eq. (5)), where the object of comparison is the optimal value of degree and betweenness centrality As shown in Fig. 5, for both $N = 500$ and $N = 1000$, φ increases constantly with p, i.e., MSDTC is more powerful for networks with higher redundancies. For each $\bar{k}$, φ reaches its maximum value (29 %) at $p = 0.05$, whereas the minimum is 6 %. Moreover, we find that when p is given, the promotion rate exhibits a strong correlation with $\bar{k}$ and their marginal benefits are asymptotically declined. For instance, when $p = 0.05$ and $N = 1000$, φ varies from 29 % to 5 % and the discrepancies between adjacent $\bar{k}$ varies from 16 % to 2.5 %.

$$\varphi = \frac{(R_{MSTDC} - R^{optimal}{}_{Degree\ and\ Betweenness})}{R^{optimal}{}_{Degree\ \text{and}\ Betweenness}} \times 100\% \tag{5}$$

4.2.2. Empirical networks

The R value of disintegration methods on the six empirical networks are presented in Table 2. It is clear that MSTDC shows better performance over other methods among all the benchmarks. Specifically, for networks with high clustering coefficients, i.e., the CA-Grqc, CA-Hephh and Parsed Tap networks, the promotion rates between MSTDC and the sub-optimal are fairly low (1.1 %, 2.42 %, and 3.69 % respectively). For networks with lower clustering coefficients, i.e., the Dolphin, Gnutella and Powergrid networks, the promotion rates between MSTDC and the sub-optimal are 4.38 %, 5 % and 9.68 %, respectively. These distinct differences can be explained by the joint effect of network density, for which the CA-Grqc, CA-Hepph and Parsed Tap have much larger average degree than the Dolphin, Gnutella and Powergrid. The maximum promotion rate reaches in Powergrid with the lowest average degree. Moreover, the result of Cycle Ratio exceeds that of degree in CA-Grqc, CA-Hepph and Parsed Tap, verifying that redundancy negatively impacts traditional centralities. It is worth noting that MSTDC outperforms

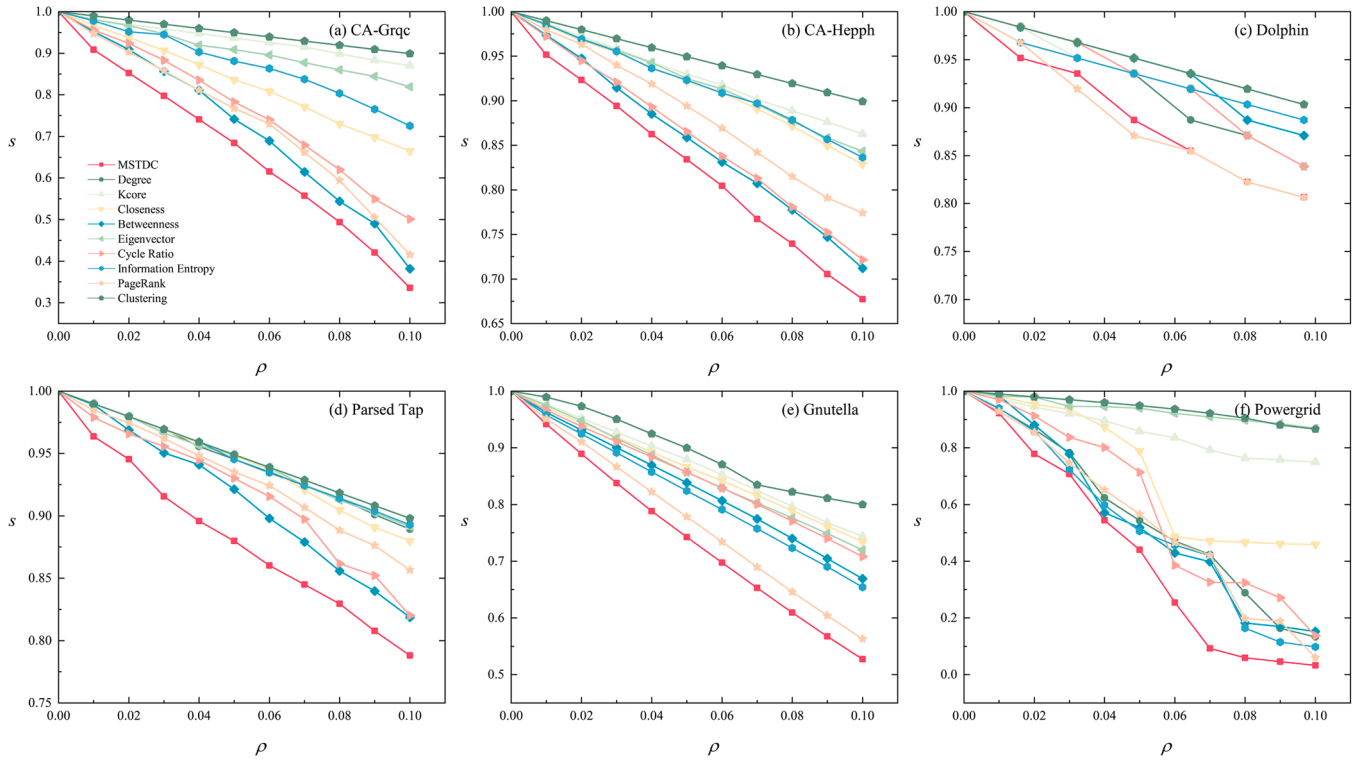


Fig. 6. The proportion of nodes in the giant connected component (s) when a proportion of nodes (ρ) is removed. (a) CA-Grqc; (b) CA-Hepph; (c) Dolphin; (d) Parsed Tap; (e) Gnutella; (f) Powergrid.

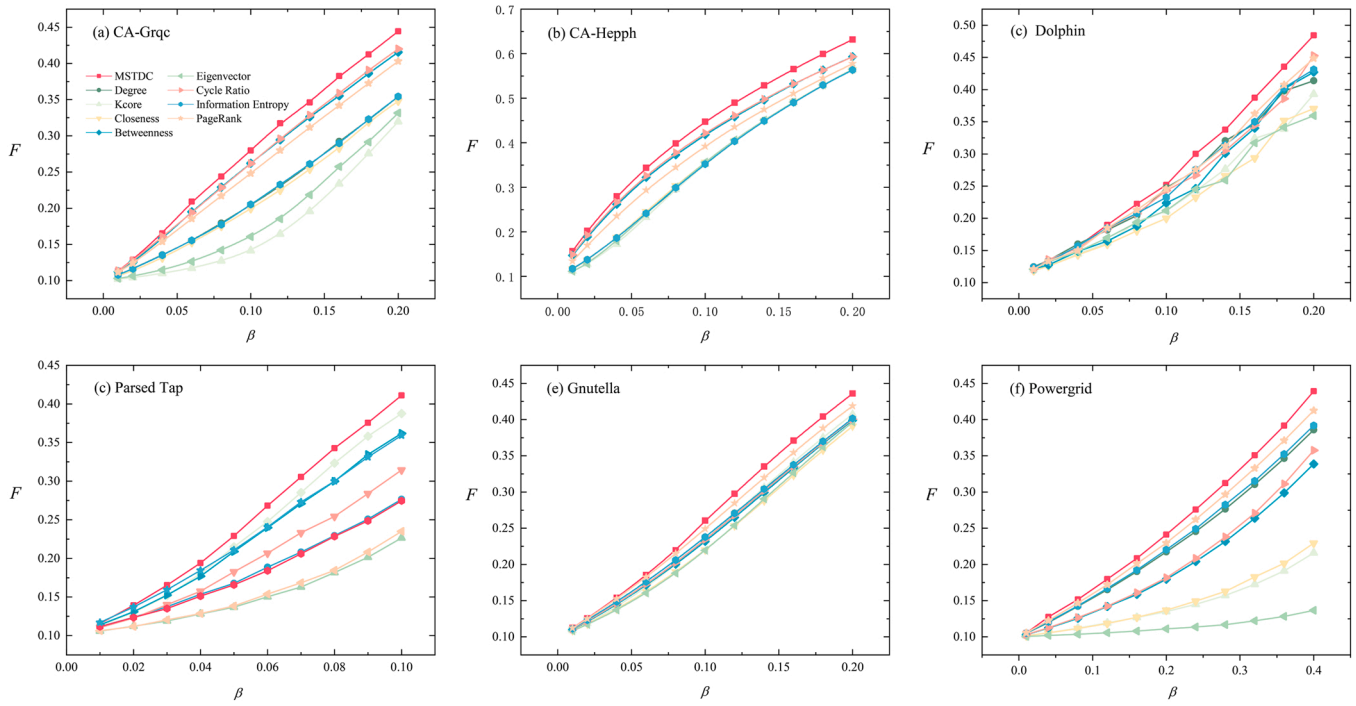


Fig. 7. Comparisons of the propagation capability of nine methods under different propagation probabilities. Each experiment is obtained by averaging over 100 independent runs. (a) CA-Grqc; (b) CA-Hepph; (c) Dolphin; (d) Parsed Tap; (e) Gnutella; (f) Powergrid.

the well-known method, PageRank. In CA-Hepph and Parsed Tap, the promotion rates are 10 % and 23 % respectively.

In addition, we investigate the dynamic process of network collapse by examining the curve of s after removing Top-0.1 N nodes. From the results presented in Fig. 6, we can see that MSTDC generally outpaces the other indices. Specifically, MSTDC achieves the best performance in CA-Grqc, with a 70 % reduction in GCC size after removing Top-0.1 N nodes. The performance of the local clustering coefficient is proven to be the worst in all tested networks. Specifically, removing $\text{Top} - \rho N$ nodes gives rise to the change of s less than the change of ρ, except in Gnutella. In Gnutella, with the average clustering coefficient of 0.005, removing the Top-0.1 N nodes according to the local clustering coefficients causes a change of approximately 20 % in s. This suggests that local clustering coefficients achieves better performance in sparse networks. These results confirm the theoretical analysis in Section 2.1 that clustering and local coupling negatively interfere with the identification of influential nodes.

Betweenness centrality or PageRank is the sub-optimal method. In high clustering coefficients networks, e.g., CA-Grqc, CA-Hepph, and Parsed Tap, betweenness centrality outpaces PageRank. In Dolphin and Gnutella, with relatively low clustering coefficients, PageRank exhibit better performance compared to betweenness centrality. Noticeably, the increment of MSTDC compared with the sub-optimal method among benchmarks is consistent with the result of R in empirical networks.

Combining the results of R and s in empirical networks, we can see that disintegrate networks with higher MSTDC is more effective, especially for networks with high clustering coefficients and low average degree.

4.3. Epidemic spreading

In this section, we measure the importance of nodes through spreading dynamics. The discrepancies among the spreading capabilities of different indices are fairly small when β is small and gradually grows larger with the increment of β.

As shown in Fig. 7, MSTDC is superior to the other indices among all empirical networks under different transmission probabilities. On average, the spreading with high MSTDC nodes can generate 6 % more widespread diffusion compared to the suboptimal. Specifically, in CA-Hepph, MSTDC causes the most widespread outbreak with 65 % of nodes infected when $\beta = 0.2$. It is also noteworthy that the curves of the other centralities except MSTDC become gradually coincident in CA-Hepph, Dolphin and Gnutella as β increases. Moreover, Cycle Ratio and betweenness centrality are the sub-optimal methods in CA-Grqc, and CA-Hepph. Cycle Ratio is relatively inferior to betweenness centrality in Parsed Tap. In sparse networks, i.e., Gnutella and Powergrid, PageRank is the sub-optimal method. Meanwhile, the discrepancies between betweenness centrality and degree centrality is relatively small.

In summary, combined with the comparison of network disintegration and epidemic spreading, MSTDC outperforms the other benchmark network indices in identifying influential nodes for maintaining network connectivity and maximizing spreading capacity and is extraordinarily effective in networks with high average clustering coefficients.

4.4. Correlation analysis

To further explore the relationship between MSTDC and other network indices, we measure the correlation between pairs of indices through Kendall's Tau (τ) [61]. From Fig. 8, we can see that the correlation between MSTDC and PageRank, degree and betweenness centrality are higher than that for the other indices (the average correlation coefficients are 0.65, 0.61 and 0.78 respectively). These attributes that spanning trees incorporate the global information of paths among nodes and the calculation of degree in a spanning tree integrates the local information. The correlations between MSTDC and the others are higher in networks with relatively low average clustering coefficients, but lower in high clustering networks. This explains why MSTDC is extraordinarily effective in networks with high average clustering coefficients. As suggested by the lower correlations, the node rankings produced by MSTDC contain more information than those produced by other indices in

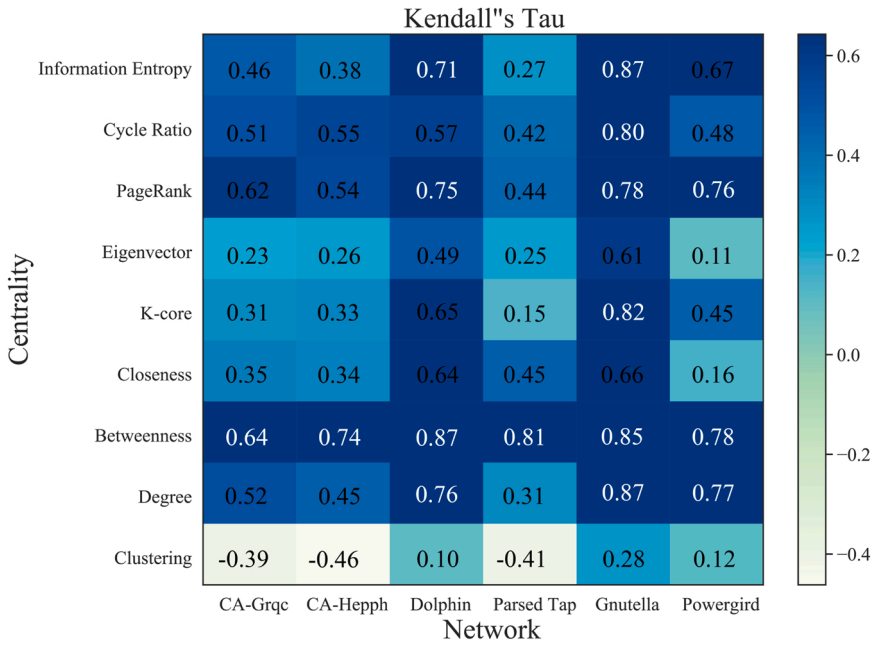


Fig. 8. The correlation matrix between MSTDC and other centralities over six networks.

networks with high average clustering coefficients.

MSTDC eliminates the redundant ties and local coupling by aggregating the information of multiple spanning trees. There are no triangles in spanning trees, so MSTDC is substantially different from local clustering coefficients. As exhibited in Fig. 8, the correlations between MSTDC and local clustering coefficients is negative in high clustering networks with $\overline{\tau} = -0.13$. This suggests that our method differs significantly from local clustering coefficients. Moreover, the correlations between MSTDC and Cycle Ratio are higher in low clustering networks and lower in high clustering networks. It indicates that MSTDC could reflect the effects of redundancy from another perspective.

5. Conclusion and discussion

In this work, we define a new node ranking method, called MSTDC, which can quantify the importance of an individual node by eliminating redundant ties and local coupling by aggregating the information of multiple spanning trees. Experiments on six empirical networks show that MSTDC obtains superior performance than the other indices in identifying influential nodes with regard to maintaining network connectivity and blocking network spreading. These attributes that spanning trees incorporate the global information of paths among nodes, and the calculation of degree in a spanning tree considers the local information. Furthermore, we have proven that our method is more suitable for networks with higher average clustering coefficients. While few studies have paid attention to the role of redundant ties in topology, the present study could extend our understanding of redundant ties and how they affect the identification of influential nodes. This study provides novel insights for considering the role of redundant ties by constructing spanning trees.

In this paper, we eliminate redundant ties by constructing spanning trees, which turns the network into a set of tree ensembles. As most empirical networks are rather complicated, eliminating all redundant ties may cause the loss of vital information, and the definition of redundancy can be different in practical scenarios. Future work is required to develop an appropriate index to quantify the degree of redundancy for edges. Future research should evaluate methodologies to set a threshold for redundancy dismantling for a more general framework of influential node identification. Furthermore, as the community structure is also one of the main properties for real-world networks [62, 63], the interacting effect of link redundancy, community structure and performance of MSTDC is yet to be explored.

CRediT authorship contribution statement

Bitao Dai: Methodology, Software, Writing, Visualization, Writing – review & editing. **Shuo Qin:** Visualization, Writing – review & editing. **Suoyi Tan:** Writing – review & editing. **Chuchu Liu:** Writing – review & editing. **Jianhong Mou:** Writing – review & editing. **Hongzhong Deng:** Writing – review & editing. **Fredrik Liljeros:** Conceptualization, Writing – review & editing. **Xin Lu:** Conceptualization, Methodology, Writing – review & editing.

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.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 72025405, 72088101, 72001211, the Hunan Science and Technology Plan Project (2020TP1013, 2020JJ4673, 2020JJ5679) the Shenzhen Basic Research Project for Development of Science and Technology (JCYJ20200109141218676, 20200829172650001) the Innovation Team Project of Colleges in Guangdong Province (2020KCXTD040) and the Postgraduate Scientific Research Innovation Project of Hunan Province (CX20220025).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jocs.2023.102030.

References

[1] D. Schoch, Centrality without indices: partial rankings and rank probabilities in networks, Soc. Netw. 54 (2018) 50–60.
[2] G. Maji, Influential spreaders identification in complex networks with potential edge weight based k-shell degree neighborhood method, J. Comput. Sci. 39 (2020), 101055.
[3] S. Aral, D. Walker, Identifying influential and susceptible members of social networks, Science 337 (2012) 337–341.
[4] S. Pei, F. Liljeros, J. Shaman, Identifying asymptomatic spreaders of antimicrobial-resistant pathogens in hospital settings, Proc. Natl. Acad. Sci. USA 118 (2021) e2111190118.
[5] H.-J. Li, Z. Bu, A. Li, Z. Liu, Y. Shi, Fast and accurate mining the community structure: integrating center locating and membership optimization, IEEE Trans. Knowl. Data Eng. 28 (2016) 2349–2362.
[6] T. Zhou, Z. Fu, B. Wang, Epidemic dynamics on complex networks, Prog. Nat. Sci. 16 (2006) 452–457.
[7] S. Chen, X. Lu, An immunization strategy for hidden populations, Sci. Rep. 7 (2017) 1–10.
[8] F.D. Sahneh, F.N. Chowdhury, C.M. Scoglio, On the existence of a threshold for preventive behavioral responses to suppress epidemic spreading, Sci. Rep. 2 (2012) 1–8.
[9] W. Wang, Q.-H. Liu, S.-M. Cai, M. Tang, L.A. Braunstein, H.E. Stanley, Suppressing disease spreading by using information diffusion on multiplex networks, Sci. Rep. 6 (2016) 1–14.
[10] A. Ibnoulouafi, M. El Haziti, Density centrality: identifying influential nodes based on area density formula, Chaos Solitons Fractals 114 (2018) 69–80.
[11] R.M. Tripathy, A. Bagchi, S. Mehta, A study of rumor control strategies on social networks, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 2010, pp. 1817–1820.
[12] J. Goldenberg, B. Libai, E. Muller, Talk of the network: a complex systems look at the underlying process of word-of-mouth, Mark. Lett. 12 (2001) 211–223.
[13] G. Chao, J. Shi-Hong, Z. Wang, Y. Dengf, Y. Fan, L. Xue-Long, A novel method to identify influential stations based on dynamic passenger flows (in chinese), Sci. China Inf. Sci. 51 (2021) 1490–1506, https://doi.org/10.1360/SSI-2020-0303.
[14] D.T. Nguyen, Y. Shen, M.T. Thai, Detecting critical nodes in interdependent power networks for vulnerability assessment, IEEE Trans. Smart Grid 4 (2013) 151–159.
[15] R. Guimerà, A. Díaz-Guilera, F. Vega-Redondo, A. Cabrales, A. Arenas, Optimal network topologies for local search with congestion, Phys. Rev. Lett. 89 (2002), 248701.
[16] P. Hu, T. Mei, Ranking influential nodes in complex networks with structural holes, Phys. A: Stat. Mech. Appl. 490 (2018) 624–631.
[17] F.D. Malliaros, M.-E.G. Rossi, M. Vazirgiannis, Locating influential nodes in complex networks, Sci. Rep. 6 (2016) 1–10.
[18] A. Namtirtha, A. Dutta, B. Dutta, Weighted kshell degree neighborhood: a new method for identifying the influential spreaders from a variety of complex network connectivity structures, Expert Syst. Appl. 139 (2020), 112859.
[19] P. Bonacich, Factoring and weighting approaches to status scores and clique identification, J. Math. Sociol. 2 (1972) 113–120.
[20] L.C. Freeman, Centrality in social networks conceptual clarification, Soc. Netw. 1 (1978) 215–239.
[21] M. Kitsak, L.K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H.E. Stanley, H.A. Makse, Identification of influential spreaders in complex networks, Nature, Physics 6 (2010) 888–893.
[22] J.-E. Wang, S.-Y. Liu, A. Aljmiai, Y.-G. Bai, Detection of influential nodes with multi-scale information, Chin. Phys. B 30 (2021), 088902.
[23] X.-L. Yan, Y.-P. Cui, S.-J. Ni, Identifying influential spreaders in complex networks based on entropy weight method and gravity law, Chin. Phys. B 29 (2020), 048902.
[24] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, T. Zhou, Identifying influential nodes in complex networks, Phys. A: Stat. Mech. Appl. 391 (2012) 1777–1787.
[25] F. Nian, S. Ren, Z. Dang, The propagation-weighted priority immunization strategy based on propagation tree, Chaos Solitons Fractals 99 (2017) 72–78.
[26] B. Zhou, S. Pei, L. Muchnik, X. Meng, X. Xu, A. Sela, S. Havlin, H.E. Stanley, Realistic modelling of information spread using peer-to-peer diffusion patterns, Nat. Hum. Behav. 4 (2020) 1198–1207.
[27] D.-B. Chen, H. Gao, L. Lü, T. Zhou, Identifying influential nodes in large-scale directed networks: the role of clustering, PLoS One 8 (2013), e77455.
[28] U. Brandes, On variants of shortest-path betweenness centrality and their generic computation, Soc. Netw. 30 (2008) 136–145.
[29] L. Fei, Y. Deng, A new method to identify influential nodes based on relative entropy, Chaos Solitons Fractals 104 (2017) 257–267.
[30] N. Zhao, J. Bao, N. Chen, Ranking influential nodes in complex networks with information entropy method, Complexity 2020 (2020) 1–15.
[31] J. Bi, J. Jin, C. Qu, X. Zhan, G. Wang, G. Yan, Temporal gravity model for important node identification in temporal networks, Chaos Solitons Fractals 147 (2021), 110934.
[32] K. Berahmand, A. Bouyer, N. Samadi, A new centrality measure based on the negative and positive effects of clustering coefficient for identifying influential spreaders in complex networks, Chaos Solitons Fractals 110 (2018) 41–54.
[33] D. Centola, M. Macy, Complex contagions and the weakness of long ties, Am. J. Sociol. 113 (2007) 702–734.
[34] M.S. Granovetter, The strength of weak ties, Am. J. Sociol. 78 (1973) 1360–1380.
[35] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, A.-L. Barabási, Structure and tie strengths in mobile communication networks, Proc. Natl. Acad. Sci. 104 (2007) 7332–7336.
[36] A. Keuchenius, P. Törnberg, J. Uitermark, Adoption and adaptation: a computational case study of the spread of granovetter's weak ties hypothesis, Soc. Netw. 66 (2021) 10–25.
[37] Y. Liu, M. Tang, T. Zhou, et al., Improving the accuracy of the k-shell method by removing redundant links: From a perspective of spreading dynamics, Sci. Rep. 5.1 (2015) 1–11.
[38] D. Shi, L. Lü, G. Chen, Totally homogeneous networks, Natl. Sci. Rev. 6 (2019) 962–969.
[39] T. Fan, L. Lü, D. Shi, T. Zhou, Characterizing cycle structure in complex networks, Commun. Phys. 4 (2021) 1–9.
[40] Y. Liu, M. Tang, T. Zhou, Y. Do, Core-like groups result in invalidation of identifying super-spreader by k-shell decomposition, Sci. Rep. 5 (2015) 1–8.
[41] R.C. Read, R.E. Tarjan, Bounds on backtrack algorithms for listing cycles, paths, and spanning trees, Networks 5 (1975) 237–252.
[42] C. Nash-Williams, Edge-disjoint spanning trees of finite graphs, J. Lond. Math. Soc. 1 (1961) 445–450.
[43] D.S. Callaway, M.E. Newman, S.H. Strogatz, D.J. Watts, Network robustness and fragility: Percolation on random graphs, Phys. Rev. Lett. 85 (2000) 5468.
[44] R. Cohen, K. Erez, D. Ben-Avraham, S. Havlin, Breakdown of the internet under intentional attack, Phys. Rev. Lett. 86 (2001) 3682.
[45] S.-Y. Tan, J. Wu, L. Lü, M.-J. Li, X. Lu, Efficient network disintegration under incomplete information: the comic effect of link prediction, Sci. Rep. 6 (2016) 1–9.
[46] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, A. Vespignani, Epidemic processes in complex networks, Rev. Mod. Phys. 87 (2015) 925.
[47] Xin Lu, Respondent-driven sampling: theory, limitations & improvements, Karolinska Inst. (2013).
[48] J. Leskovec, J. Kleinberg, C. Faloutsos, Graph evolution: Densification and shrinking diameters, ACM Trans. Knowl. Discov. Data (TKDD) 1 (2007), 2–es.
[49] B. Rozemberczki, R. Davies, R. Sarkar, C. Sutton, Gemsec: Graph embedding with self clustering, in: Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining, 2019, pp. 65–72.
[50] (a) D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, S.M. Dawson, The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations, Behav. Ecol. Sociobiol. 54 (2003) 396–405;
(b) M. Ripeanu, I. Foster, A. Iamnitchi, Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design, Comput. Sci. 6 (2002) 2002.
[51] Kunegis J. Konect: the koblenz network collection//Proceedings of the 22nd international conference on world wide web, 2013, pp. 1343–1350.
[52] Anne-Claude Gavin, et al., Functional organization of the yeast proteome by systematic analysis of protein complexes, Nature 415 (6868) (2002) 141–147.
[53] M. Ripeanu, I. Foster, A. Iamnitchi, Mapping the gnutella network: properties of large-scale peer-to-peer systems and implications for system design, Comput. Sci. 6 (2002) 2002.
[54] J.Watts Duncan, H.Strogatz Steven, Collective dynamics of 'small-world' networks, Nature 393 (1998) 440–442.
[55] Christian M. Schneider, et al., Mitigation of malicious attacks on networks, Proc. Natl. Acad. Sci. 108 (10) (2011) 3838–3841.
[56] R. Cohen, S. Havlin, D. Ben-Avraham, Efficient immunization strategies for computer networks and populations, Phys. Rev. Lett. 91 (2003), 247901.
[57] R. Pastor-Satorras, A. Vespignani, Immunization of complex networks, Phys. Rev. E 65 (2002), 036104.
[58] Zi-Ke Zhang, et al., Dynamics of information diffusion and its applications on complex networks, Phys. Rep. 651 (2016) 1–34.
[59] Lawrence Page, et al., The PageRank citation ranking: bringing order to the web, Stanf. Info (1999).
[60] Xiang Xu, et al., Identifying vital nodes in complex networks by adjacency information entropy, Sci. Rep. 10 (1) (2020) 1–12.
[61] M.G. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–93.
[62] Luan Yangyang, Xiaoqun Wu, Binghong Liu, Maximizing synchronizability of networks with community structure based on node similarity, Chaos: An Interdiscip. J. Nonlinear Sci. 32 (8) (2022), 083106.
[63] Stephany Rajeh, et al., Characterizing the interactions between classical and community-aware centrality measures in complex networks, Sci. Rep. 11 (1) (2021) 10088.

Bitao Dai is a Ph.D. candidate affiliated with the National University of Defense Technology, supervised by Xin Lu. His research interests include social network analysis, network disintegration, human behavior and information diffusion.

Shuo Qin received the M.E. degrees in National University of Defense Technology, Changsha, China, in 2019. She is currently working at State Key Laboratory on Blind Signal Processing, Chengdu, China. Her research interests include complex networks and data mining.

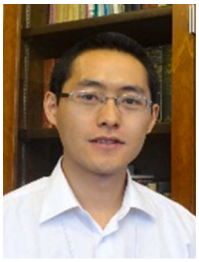
Suo-Yi Tan received the B.S degree, the M.S. degree, and the Ph.D. degree in management science and engineering from the National University of Defense Technology, Changsha, China (2012, 2014, and 2018). Now he is a lecturer from the College of System Engineering, National University of Defense Technology, China. From 2016–2017, He was a visiting Ph.D. student with the Center for Polymer Studies, Boston University, Boston, MA, USA. His current research is in the following areas: big data analytics, especially disaster response with big data analytics; other issues in big data, mobility, complex networks, social networks, human behaviors, etc.

Chuchu Liu is currently a Ph.D. candidate at National University of Defense Technology of China. Meanwhile, she is an intern at the School of Computing, NUS. Her research interests are complex networks, big data mining, social network analysis, and natural language processing.

Jianhong Mou received the B.E. and M.E. degrees in National University of Defense Technology, Changsha, China, in 2011 and 2017, respectively. He is currently working toward the Ph. D. degree in management science and engineering with the National University of Defense Technology, Changsha, China. Since December 2021. His research interests include complex networks, human behavior and information diffusion.

Hongzhong Deng is currently a Professor with the National University of Defense Technology, Changsha, China. His research interests include complex networks and network disintegration.

Fredrik Liljeros is a Professor of sociology with Stockholm University. His research is mainly focused on how human interaction patterns affect how social influences and diseases are spread in different populations. Fredrik did his post doc at the Swedish institute for infectious diseases and control, worked with pandemic planning, and studied, among other things, how travel restrictions affect the spread of diseases.



Xin Lu received the Ph.D. degree in medical science from the Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden, in 2013. He is currently a Professor with the National University of Defense Technology, Changsha, China. His research interests include big data analytics, complex networks, human behavior, and emergency management. He has been granted the National Science Fund for Distinguished Young Scholars.