SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

September 2023, Vol. 66 192203:1–192203:10
https://doi.org/10.1007/s11432-022-3715-4

Cost effective approach to identify multiple influential spreaders based on the cycle structure in networks

Wenfeng SHI[1], Shuqi XU[2,1*], Tianlong FAN[3,4] & Linyuan LÜ[1*]

[1]*Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu* 611731*, China;*
[2]*Institute of Dataspace, Hefei Comprehensive National Science Center, Hefei* 230088*, China;*
[3]*Department of Physics, University of Fribourg, Fribourg* 1700*, Switzerland;*
[4]*Alibaba Research Center for Complexity Sciences, Hangzhou Normal University, Hangzhou* 311121*, China*

Received 11 October 2022/Revised 27 December 2022/Accepted 16 February 2023/Published online 28 August 2023

Abstract Identifying influential spreaders has theoretical and practical significance in complex networks. Traditional centrality methods can efficiently find a single spreader, but it could lead to influence redundancy and high initializing costs when used to identify a set of multiple spreaders. A cycle structure is one of the most crucial reasons for the complexity of a network and the cornerstone of the feedback effect. From this novel perspective, we propose a new method based on basic cycles in networks to identify multiple influential spreaders with superior spreading performance and low initializing costs. Experiments on six empirical networks show that the spreaders selected by the proposed method are more scattered in the network and yield the best spreading performance compared with those on seven well-known methods. Importantly, the proposed method is the most cost effective under the same spreading performance. The cycle-based method has the advantage of generating multiple solutions. Our work provides new insights into identifying multiple spreaders and hence can benefit wide applications in practical scenarios.

Keywords multiple spreaders, basic cycle, complex network, cycle structure, spreading dynamics

Citation Shi W F, Xu S Q, Fan T L, et al. Cost effective approach to identify multiple influential spreaders based on the cycle structure in networks. Sci China Inf Sci, 2023, 66(9): 192203, https://doi.org/10.1007/s11432-022-3715-4

1 Introduction

Spreading is a universal dynamical process in complex networks, and typical examples include disease transmission [1] and information propagation [2]. Identifying influential spreaders in networks has theoretical and practical significance and can help effectively control epidemic outbreaks [3], limit the spread of rumors [4], and promote new products [5]. At present, several common node centralities can be used to address the problem of identifying important nodes, which can be classified into four categories [6]: (1) structure-based centralities, such as degree centrality (DC) [7], H-index (HI) [8], and betweenness [9]; (2) iteration-based centralities, such as eigenvector centrality (EC) [10] and PageRank [11]; (3) operation-based centralities, such as connection-sensitive method [12]; and (4) dynamics-based centralities, such as collective influence (CI) [13, 14], explosive-percolation-based centrality [15], and dynamics-sensitive centrality [16]. Although the above methods have outstanding performance in identifying a single influential spreader, they will cause problems when used to identify multiple spreaders. The specific problems include the following. (i) Influence redundancy: Nodes with large centrality values tend to be close to one another. Thus, their influence could be overlapped to a large extent [17, 18]. (ii) High initializing cost: Initializing (e.g., control or employ) the spreaders identified by the centrality methods usually needs huge expenses [19]. For example, using Internet celebrities for endorsement costs much more than using ordinary people.

* Corresponding author (email: xushuqi@idata.ah.cn, linyuan.lv@ustc.edu.cn)

Recently, scholars have revealed that the cycle structure plays an indispensable role in complex networks [20–23], which provides a new perspective in studying nodes' impact. Fan et al. [24] found that the impact of a node depends not only on the influence of its directly connected neighbors but also on the influence of its co-cycle neighbors. Accordingly, the authors proposed a cycle-based indicator, cycle ratio (CR), to measure nodes' importance according to their participation in other nodes' shortest cycles. The CR can determine vital nodes scattered throughout the whole network and perform effectively in maintaining network connectivity, promoting synchronization, and spreading. These studies encourage us to start with cycle structures to identify multiple spreaders with large aggregate influence and less initializing cost.

In this study, we consider a significant class of network cycles, i.e., basic cycles, which are a representative set of all cycles in a network. The size of basic cycles can vary in a large range, which is different from the shortest cycles considered in the work of Fan et al. [24]. We first examined the topological properties of basic cycles in six empirical networks and three model networks. Based on the basic cycle-related statistics, we propose a new node centrality indicator, the number of basic cycles (NC), to quantify the importance of each node. The underlying idea of this method is to consider the important role that cycles play in network structures and dynamics. As such, a node involved in more basic cycles is likely to be critical in spreading. Experiments on six empirical networks show that compared with seven commonly used centrality indicators, namely, DC [7], coreness (KC) [17], CI [14], HI [25, 26], closeness centrality (CC) [27], EC [10], and a cycle-based indicator, i.e., CR [24], the new indicator can identify multiple spreaders with the best spreading performance. Furthermore, the spreaders selected by the proposed method have the lowest cost while being initiated under the same spreading performance. Finally, the proposed method can provide several alternatives that have a similar performance. Our study explores the novel perspective of the cycle structure in complex networks and proves its vital role in identifying multiple influential spreaders. The proposed method provides efficient solutions for practical problems, such as Internet advertisement placement, online rumor control, and cognition shaping.

2 Methods

2.1 Basic cycles in networks

A cycle in a network can be simply defined as a closed path whose edges are different from one another. Common examples of network cycles include groups on WeChat and Facebook and neural feedback loops in brain networks [21].

A cycle basis (denoted by B) of a network is a minimal set of simple cycles, such that any cycle in the network can be formed by the operation of the symmetric difference of the cycle in the cycle basis, where a simple cycle is a cycle with no repeated nodes except for the beginning and ending nodes. The cycle basis can be captured from any spanning tree of a given network, which is not necessarily unique because the network may have multiple spanning trees. The cycles in B are called basic cycles. Any two basic cycles in B are linearly independent. By observing basic cycles, we can capture the whole picture of network cycles. Given a network $G(V, E)$, where V represents the set of nodes ($|V| = N$) and E represents the set of edges ($|E| = M$) and its cycle basis B, we define the number of basic cycles passing through a node (denoted by NC) as a node centrality index to identify influential nodes. The computation process is as follows.

For an undirected network G, we initialized a spanning tree $T = \emptyset$ and an unchecked node set $U = V$, in which all nodes in the network are initially set as the "unchecked" status.

Step 1. Randomly select a node a from U as the root of the spanning tree, and add it to T, i.e., $T = \{a\}$.

Step 2. Perform the following process: (i) Choose a node $z \in T \cap U$ following the rule that the node that enters T last should be selected first. If $T \cap U = \emptyset$, skip to step 3. (ii) Traverse each neighbor n of z; if $n \notin T$, add edge (z, n) and node n to the tree and T, respectively. Putting $n \in T$, (z, n) together with the unique path in the tree linking z to n forms a basic cycle c. (iii) After all possible neighbors n of node z have been considered, remove the current node z from U.

Step 3. Repeat step 2 until the algorithm ends when $U = \emptyset$, and then we derive a cycle basis B. We calculate the number of involved basic cycles for each node i, NC_i. If node i is not in any basic cycle, $\mathrm{NC}_i = 0$. The larger the value of NC_i is, the more important node i is. The pseudocode of this algorithm

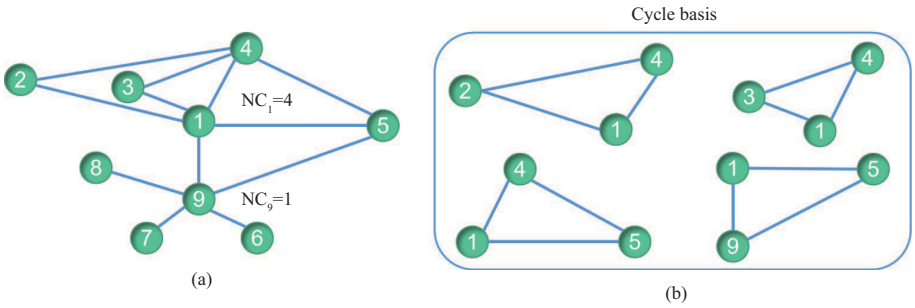

Figure 1 (Color online) (a) Sample network and (b) its cycle basis.

Table 1 Basic characteristics of the studied networks[a)]

Networks	N	M	C	D	$\langle k\rangle$	β_c
C. elegans [33]	297	2148	0.292	0.0488	14.464	0.041
USA airports [34]	1574	17215	0.504	0.0139	21.874	0.009
Yeast [35]	2375	11693	0.306	0.0041	9.846	0.031
Soc-hamsterster [36]	2426	16630	0.538	0.0056	13.710	0.024
Asian-last.fm [37]	7624	27806	0.219	0.0009	7.294	0.042
Router [38]	5022	6258	0.011	0.0005	2.492	0.085

a) N: number of nodes; M: number of links; C: average clustering coefficient; D: density; $\langle k\rangle$: average degree; $\beta_c = \langle k\rangle/(\langle k^2\rangle - 2\langle k\rangle)$: epidemic threshold of the SIR model [39].

is provided in Appendix A, and the time complexity of the method is $O(M\langle|c|\rangle)$.

Different spanning trees may lead to different cycle bases, but this not only affects the performance of NC (see Subsection 3.5 for a detailed discussion). In Figure 1, we provide an example network and its cycle basis. In this network, nodes 1 and 9 have the same degree, whereas $\mathrm{NC}_1 = 4$, $\mathrm{NC}_9 = 1$, which suggests that according to our methods, node 1 is more influential than node 9.

2.2 Network spreading model

To evaluate the spreading performance of the identified nodes, we used a well-established spreading model, the susceptible-infectious-recovered (SIR) model [28]. The SIR model assumes that the nodes in a network are in one of three states: susceptible, infected, and recovered. The infected nodes will infect their neighboring susceptible nodes with probability β, the infected nodes will recover with probability μ, and the recovered nodes will no longer be infected. In the following experiments, we set $\mu = 0.5$ and $\beta = \beta_c = \langle k\rangle/(\langle k^2\rangle - 2\langle k\rangle)$ (see Table 1), where $\langle k\rangle$ and $\langle k^2\rangle$ denote the average degree and average square degree of the network, respectively. We set the top-ranked nodes by indicators as seed infected nodes and others as susceptible nodes. When the spreading process starts, nodes update their states at each time step according to the above rules until there is no infected node in the network. To quantify the spreading performance of the top-ranked spreaders, we introduce a metric, named overall spreading ability (R) [29,30], which is defined as the ratio of the recovered nodes when the spreading process ends. We also measured the spreading performance in each time step t by the ratio of the infected and recovered nodes (R_t).

2.3 Benchmark centrality indicators

We consider six common-used indicators with different mechanisms and a newly proposed cycle-based metric as benchmark indicators, and their definitions are provided below. Let A denote the adjacency matrix of an undirected network, in which $A_{ij} = 1$ means that there exists a link between nodes i and j, and 0 otherwise.

DC. DC is the simplest measure of node centrality in networks [7], which is defined as $\mathrm{DC}_i = \sum_{j=1}^{N} A_{ij}$.

KC. The computation of the coreness of nodes in a network is through the K-shell decomposition process [17], which is like peeling an onion layer by layer. In this process, first, all nodes with degrees equal to one are removed; this step is repeated until the degrees of all left nodes are larger than one. All the removed nodes belong to 1-shell, and their coreness is equal to 1. Then, removing nodes is continued to a higher degree until all nodes are removed. Finally, each node is associated with a coreness score.

CI. The CI algorithm assumes that the importance of a node depends not only on its degree but also on the degree of its l-th order neighbors [14]. The CI score of a node is defined as

$$\mathrm{CI}_i = (k_i - 1) \sum_{d_{ij}=l} (k_j - 1), \tag{1}$$

where k_i is the degree of node i and d_{ij} is the shortest distance between nodes i and j. In this paper, we set $l = 2$.

HI. HI is originally used to measure the academic impact of researchers. The HI of a researcher is defined as the maximum value h, such that he/she has at least h papers, each of which with citations no less than h [31]. Recently, HI was extended to networks to measure the influence of nodes [25, 26], which is defined as the largest h, satisfying that node i has at least h neighbors for each with a degree no less than h.

CC. CC lies on the idea that a node with a smaller average shortest distance to other nodes is in the center of the network and thus is more influential [27]. The CC score of a node is defined as

$$\mathrm{CC}_i = \frac{1}{N-1} \sum_{j(\neq i)} \frac{1}{d_{ij}}. \tag{2}$$

If there is no path connected to i and j, $1/d_{ij}$ is set to 0.

EC. Considering the importance of a node and its neighbors, Bonacich [10] proposed the EC indicator, which is defined as

$$\mathrm{EC}_i = q_i = c \sum_{j=1}^{N} A_{ij} q_j, \tag{3}$$

where $c = 1/\lambda_{\max}$ and $\lambda_{\max}$ is the largest eigenvalue of A. EC can be calculated by an iteration process [32] in which nodes' EC scores are obtained when the steady state is reached. Eq. (3) can be written in the following matrix form: $Q = cAQ$. Here, $Q = (q_1, q_2, \ldots, q_n)^{\mathrm{T}}$, which is the eigenvector of A corresponding to $\lambda_{\max}$.

CR. Fan et al. [24] defined the shortest cycles of node i as the cycles containing i with the smallest size, which is defined as S_i. Based on this concept, they proposed a cycle number matrix C, in which

$$c_{ij} = \begin{cases} \text{the number of cycles in } S_i \text{ that pass through } i \text{ and } j, & i \neq j, \\ \text{the number of cycles in } S_i, & i = j. \end{cases} \tag{4}$$

Then, the CR is put forward to compute the level of node i participating in the shortest cycles of other nodes. In the formula, it is

$$\mathrm{CR}_i = \begin{cases} 0, & c_{ii} = 0, \\ \sum_{j, c_{ij}>0} \frac{c_{ij}}{c_{jj}}, & c_{ii} > 0. \end{cases} \tag{5}$$

The authors showed that the CR overall outperforms DC, HI, and KC in the early stages of spreading.

2.4 Empirical data

To evaluate the performance of the NC method proposed in this paper, we conducted experiments on six real networks, which are undirected and unweighted. C. elegans [33] is a neural network of the Caenorhabditis elegans nematode, with nodes representing neurons and links representing synapses or gap junctions among two neurons. USA airports [34] is a flight network among all commercial airports in the United States. The original network is directed, and we simply transferred it into undirected in the analysis. Its nodes represent airports, and a link between two nodes indicates that there exists an airline between two airports. Yeast [35] is a protein-protein interaction network in budding yeast. Soc-hamsterster [36] is a friendship network from the Hamsterster website in which nodes represent users and links denote friend or family relationships. Asian-last.fm [37] is a social network of Last.FM users, which was collected from the public API in March 2020. This network describes the mutual follower relationships between users from Asian countries. Router [38] is a communication network where nodes represent autonomous systems and a link connecting two nodes indicates traffic exchanges between the two systems. The basic statistical properties of the six networks are shown in Table 1 [39].

Table 2 Basic cycle statistics of six empirical networks[a)]

Networks	N_b	$b_{\text{s\_avg}}$	$b_{\text{s\_max}}$	$n_{\text{b\_avg}}$	$n_{\text{b\_max}}$	γ
C. elegans	1852	7.314	45	48.039	444	True
USA airports	15643	4.972	51	62.834	6229	False
Yeast	9319	13.612	160	70.669	2676	True
Soc-hamsterster	14352	13.110	139	89.516	2088	False
Asian-last.fm	20183	75.232	656	266.574	3549	True
Router	1237	5.211	43	8.867	321	True

a) N_b: number of basic cycles; $b_{\text{s\_avg}}$: average size of all basic cycles; $b_{\text{s\_max}}$: maximum size of all basic cycles; $n_{\text{b\_avg}}$: average number of basic cycles of nodes; $n_{\text{b\_max}}$: maximum number of basic cycles of nodes; Boolean value γ indicates whether the network is connected.

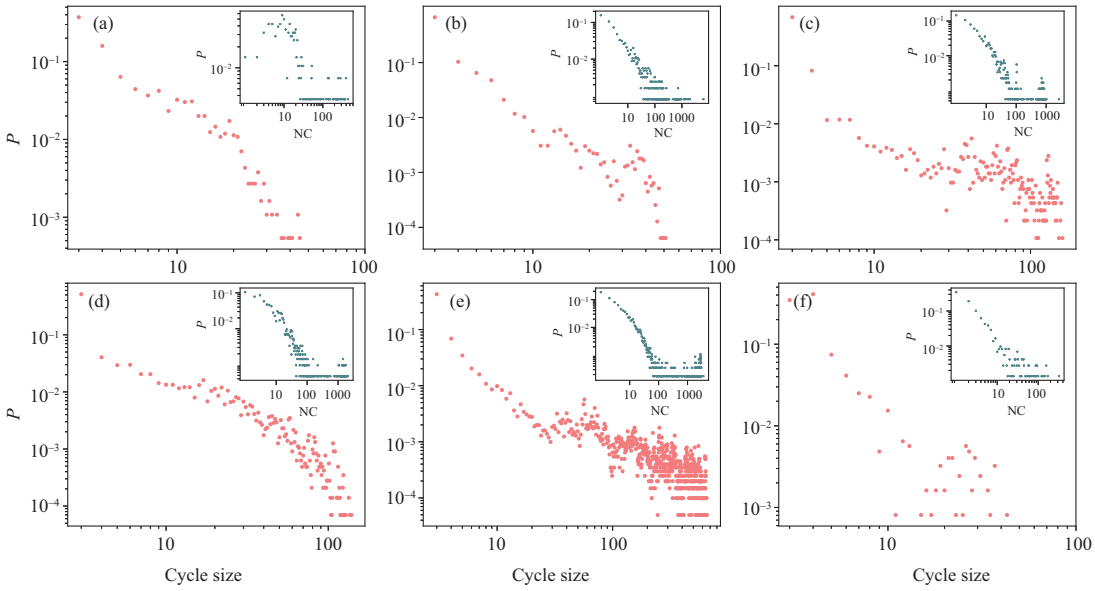

Figure 2 (Color online) Distribution of the size of all basic cycles in the six empirical networks. The x-axis represents the size of basic cycles, and the y-axis represents the probability. Inset: distribution of nodes' number of basic cycles (NC) of the corresponding network. (a) C. elegans; (b) USA airports; (c) Yeast; (d) Soc-hamsterster; (e) Asian-last.fm; (f) Router.

3 Results

3.1 Statistics of basic cycles in empirical networks

We first analyzed the statistics of basic cycles in the six empirical networks. Table 2 shows the number of basic cycles (N_b) in each network. N_b is equal to $M - N + 1$ if the network is connected. In terms of the size of basic cycles, Table 2 provides the average ($b_{\text{s\_avg}}$) and maximum ($b_{\text{s\_max}}$) size of basic cycles in each network. Figure 2 shows the distributions. We find that most basic cycles are small. On average, the proportion of basic cycles with a size not greater than 5 is 69.2%, and around half of the basic cycles have the minimum size, i.e., 3. Moreover, Tables 1 and 2 demonstrate that the maximum size of the basic cycles is negatively related to the density of the network (D); namely, the largest basic cycle of a denser network tends to be smaller. For example, the C. elegans network has the highest density ($D = 0.0488$), and its largest basic cycle is the smallest ($b_{\text{s\_max}} = 45$) among all networks except the Router network. The opposite is true for the Asian-last.fm network. A special case is the Router network, whose density is the smallest among all networks, while its $b_{\text{s\_max}}$ value is also the smallest ($b_{\text{s\_max}} = 43$). The reason is that the Router network is likely a tree structure, resulting in a small probability of forming a large basic cycle.

We further demonstrate the distributions of the nodes' number of basic cycles in six networks (see the inset in Figure 2). It shows that the distributions are close to the long-tailed distribution, which means that a large fraction of nodes have a few basic cycles, whereas there are few nodes with many basic cycles. Table 2 reports the average ($n_{\text{b\_avg}}$) and maximum ($n_{\text{b\_max}}$) number of involved basic cycles of nodes. Nodes in the Router network have the smallest $n_{\text{b\_avg}}$ and $n_{\text{b\_max}}$ ($n_{\text{b\_avg}} = 8.867$, $n_{\text{b\_max}} = 321$) compared with those in other networks because the Router network has few basic cycles and their average

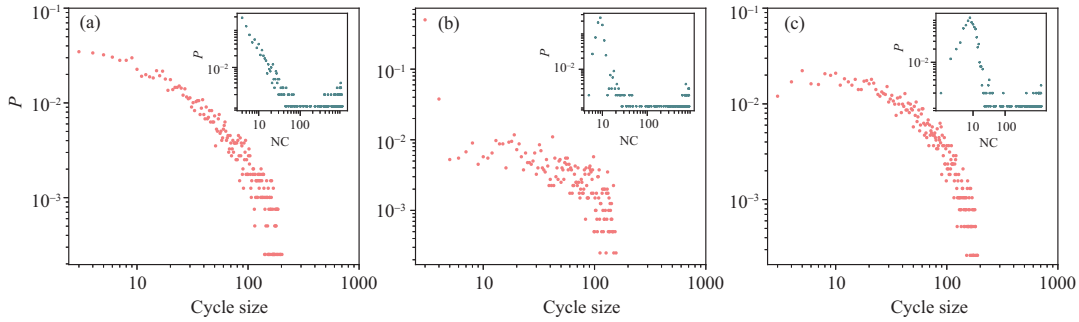

Figure 3 (Color online) Distribution of the size of all basic cycles in the three model networks. Inset: distribution of the nodes' number of basic cycles (NC) of the corresponding network. (a) BA; (b) WS; (c) ER.

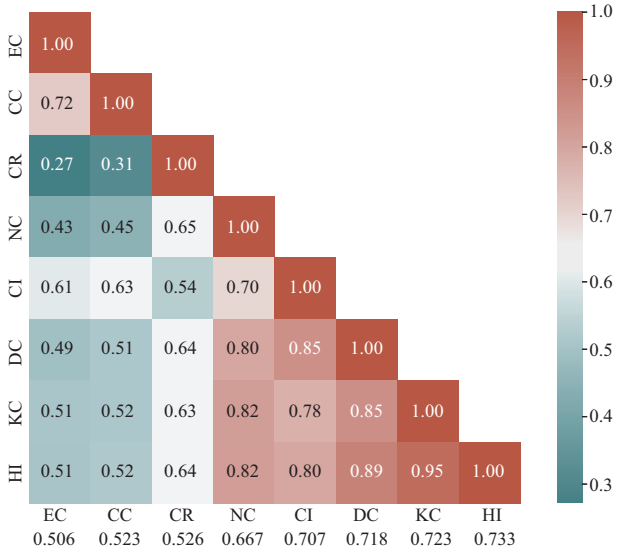

Figure 4 (Color online) Average Kendall's correlation coefficient among the eight indicators. The values below denote the average correlation coefficients between each indicator and other indicators. The colorbar denotes the average Kendall's correlation coefficient, which is averaged over six empirical networks.

sizes are small. Furthermore, the Asian-last.fm and Soc-hamsterster networks have a large number of basic cycles with a large size, so their $n_{\mathrm{b\_avg}}$ values are leading the list ($n_{\mathrm{b\_avg}} = 266.574$ and 89.516, respectively).

To determine the difference between empirical networks and model networks in terms of basic cycles, we employed the Barabási-Albert (BA), Watts-Strogatz (WS), and Erdös-Rényi (ER) networks as three types of model networks, with the same number of nodes, 1000, and average degree, 10. Specifically, for the BA network, each time, we introduce a new node and connect it to five existing nodes based on the preferential attachment rule [40]. For the WS network, we first initialize a nearest-neighbor coupling network, in which each node is connected with its ten nearest neighbors. Then, we set the rewiring probability of each edge as 0.2. For the ER network, we set the probability of connecting any two nodes as 0.01. As shown in Figure 3, the distributions of the basic cycle size of model networks are similar. For nodes' NC (see the inset in Figure 3), the distributions of the WS and ER networks are similar to that of the C. elegans network, whereas the result of the BA network is more likely those of the other five empirical networks. The reason could be that the degree distributions of WS, ER, and C. elegans networks are normal distribution, while for others, they follow a power-law distribution (see Figure B1 in Appendix B).

3.2 Correlations between NC and benchmark indicators

Before examining the performance of spreaders identified by NC, we analyzed its correlation with other centrality indicators. To this end, we compute Kendall's correlation coefficient τ [41] between the rankings of nodes obtained by every two indicators. As shown in Figure 4, the correlations among the four

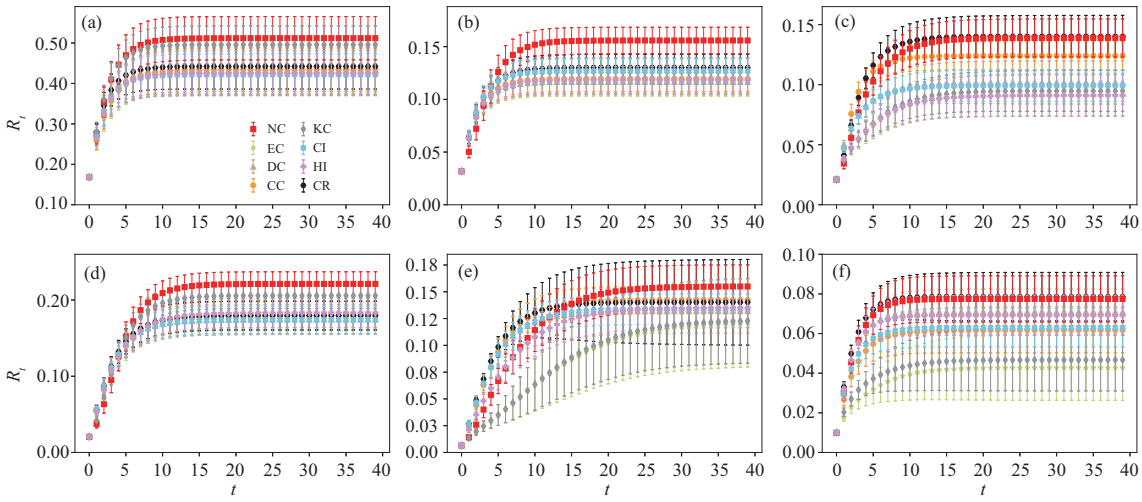

Figure 5 (Color online) Comparison of the spreading ability R_t of the top-50 nodes ranked by NC and seven benchmark indicators in the six empirical networks. We set the recovery rate to $\mu = 0.5$ and the spreading rate to $\beta = \beta_c$. Error bars denote the standard deviation among 1000 realizations. (a) C. elegans; (b) USA airports; (c) Yeast; (d) Soc-hamsterster; (e) Asian-last.fm; (f) Router.

degree-based indicators (DC, KC, CI, and HI) and our proposed method NC are relatively high, which is demonstrated by the red area in the bottom right of Figure 4. Moreover, EC, CC, and CR show low correlations with other indicators because they are iteration-based, path-based, and cycle-based, respectively. Figure 4 shows that although CR and NC are both cycle-based, their correlation is not high ($\tau = 0.65$). Moreover, we calculated the average correlation value between each indicator and other indicators. The value of NC is 0.667, which is at the medium level among all indicators. This finding indicates that NC can identify more different influential spreaders. The detailed correlation results on the six empirical networks are provided in Figure C1 in Appendix C.

3.3 Spreading performance of the NC indicator

Then, we examine the spreading performance of multiple spreaders identified by NC and other benchmark indicators. We chose the top-w ($w = 50$) nodes ranked by each method as the seed spreaders and computed their spreading ability (R_t) at a given time step t. Figure 5 demonstrates that NC has the best spreading ability among all considered indicators on six networks. The performance advantage of NC over CR is attributed to NC's consideration of a more diverse and representative set of cycles. The results with $w = 20$ and 100 are shown in Figures D1 and D2 in Appendix D, which are consistent with those in Figure 5. These results prove that the multiple spreaders identified by NC have a large influence on networks. To verify the robustness of this finding, we adjusted $\beta = \beta_c$, $2\beta_c$, $3\beta_c$, and $4\beta_c$ and computed the R of the top-50 spreaders accordingly. The results in Figure D3 in Appendix D show that under four different spreading rates, NC still outperforms all other benchmarks on the six networks, which reveals its superiority.

Next, we attempted to explain the origin of the advantage of NC. The distance among the initial spreaders is considered a critical factor in determining the spreading performance [42–45]. For example, Hu et al. [46] proved that in regular networks, the larger the distance between two spreaders, the more influential the spreading. Moreover, Kitsak et al. [17] found that requiring that any two spreaders are not linked with each other would lead to a good spreading strategy. Based on this perspective, we further analyzed the average shortest distance d [47] among the top-w spreaders selected by various indicators. The results in Figure 6 show that for almost all values of $w \in [10, 100]$, the average shortest distance among the top-w spreaders identified by NC is the largest in all studied networks. Hence, compared with other indicators, nodes with large NC are usually far away from one another and thus tend to scatter across the network (see Figure D4 in Appendix D). In particular, in the USA airports network (Figure 6(b)), the gap between NC and other indicators in terms of d is the largest among all networks, which is consistent with the outperformance of NC in terms of the spreading ability in Figure 5(b). This advantage of NC stems from the dispersity of the cycle basis, which ensures that all cycle-related nodes are included by it. In addition, in Figure 6, the d of the top-w spreaders identified by the two cycle-based

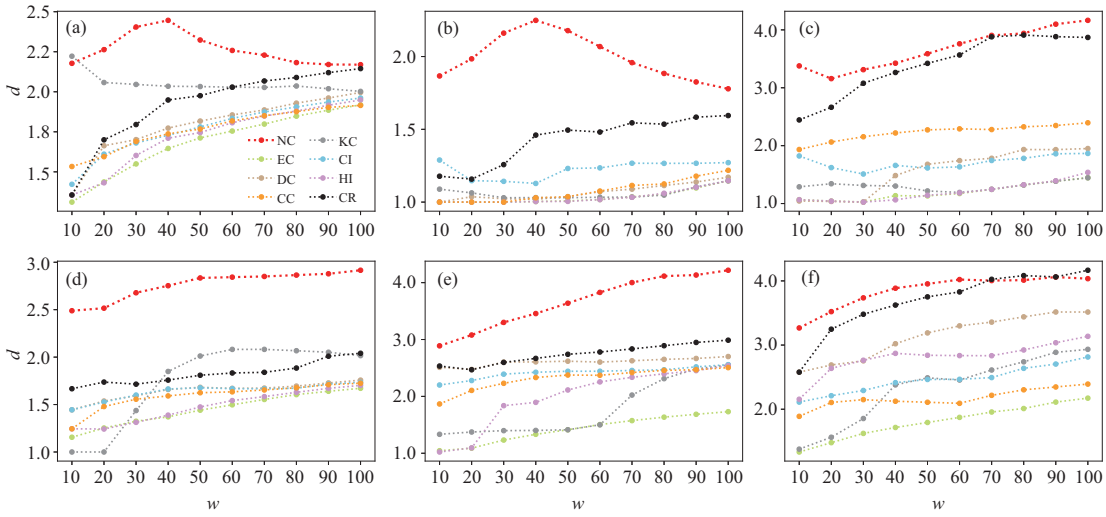

Figure 6 (Color online) Comparison of the average shortest distance d among the top-w nodes selected by eight indicators in the six empirical networks. (a) C. elegans; (b) USA airports; (c) Yeast; (d) Soc-hamsterster; (e) Asian-last.fm; (f) Router.

indicators (NC and CR) are generally larger than those of the other indicators, showing the distinct advantage of the cycle structure in spreading.

3.4 Initializing cost of the NC indicator

In practical scenarios, initializing the source spreaders is always costly, such as paying celebrities for posting or forwarding an advertisement. Thus, people always seek to trigger a wider spread with less cost. We assume that the cost of initializing a spreader depends on two factors: its influence and scarcity. The former is measured by its degree k (e.g., the number of followers on a social media platform), and the latter is represented by the probability of finding a spreader with degree k in the network, i.e., $p(k)$ [19]. The greater the k and the smaller the $p(k)$, the higher the cost. The total cost λ of initializing the top-w spreaders is defined as summing the cost of initializing each selected spreader:

$$\lambda = \sum_{i=1}^{w} \frac{k_i}{p(k_i)}, \tag{6}$$

in which k_i refers to the degree of the top i-th spreader and $p(k_i)$ denotes the probability of the spreader with degree k_i in the network. Based on the top spreaders ranked by different methods, we computed the total cost λ of the top-w ($w = 2, 3, \ldots, 100$) nodes and their corresponding overall spreading ability R, as shown in Figure 7. The results demonstrate that NC is the most cost effective for a given spreading performance. Moreover, the set of the top spreaders selected by NC has the maximum overall spreading ability at the same initializing cost on six networks. Considering the size of the Yeast, Soc-hamsterster and Asian-last.fm networks are relatively large, and the top-100 spreaders are too few to fully demonstrate the changes in initializing cost and spreading ability of indicators. Thus, we expanded the number of seed spreaders for the three networks. The results in Figures 7(c)–(e) show that the superiority of NC is robust. The results prove that the spreaders identified by NC not only have an excellent spreading ability but also have a low cost, which is of great significance in designing spreading strategies on networks, such as implementing viral marketing on social media.

3.5 Multi-solution advantage of the NC indicator

Considering that the set of basic cycles depends on the spanning tree of the network, the top spreaders selected by NC may differ in each realization because of the changing of the selected spanning tree. To prove the stability of the proposed method, we randomly generated 100 realizations of NC based on different spanning trees and further calculated the overall spreading ability of the top-20 spreaders in every realization. The average result among 100 realizations is shown in Table 3. The result reveals that the average R of NC still keeps advantages among all indicators on five of all six networks, which means that the performance of NC is robust to random selections of spanning trees. In other words, the NC

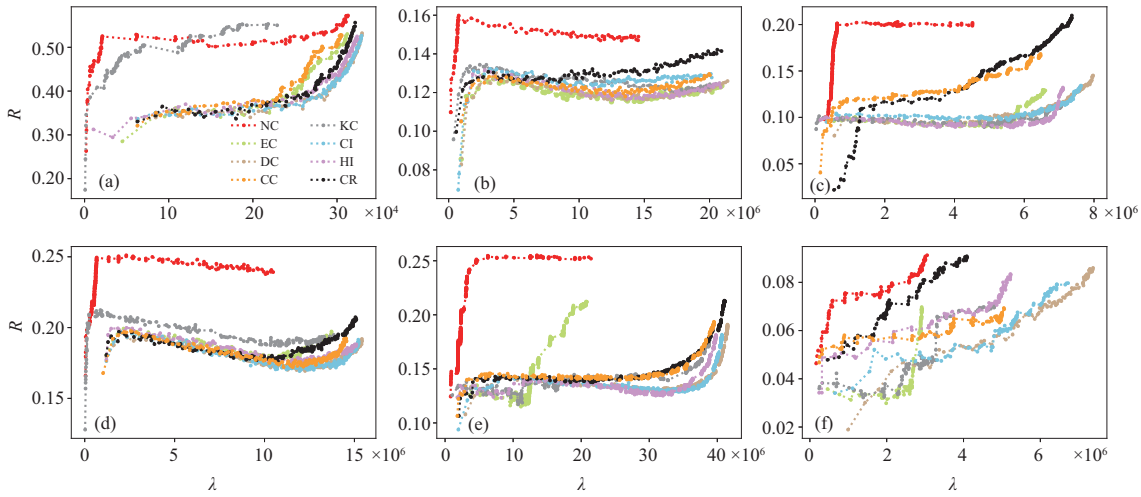

Figure 7 (Color online) Overall spreading ability (R) and cost (λ) of the six networks. The x-axis represents the total cost of the top-w ($w = 2, 3, \ldots, 100$) nodes, and the y-axis represents their corresponding overall spreading ability. In the Yeast and Soc-hamsterster networks, the number of seed spreaders w ranges from 2 to 200. In the Asian-last.fm network, w ranges from 2 to 700. (a) C. elegans; (b) USA airports; (c) Yeast; (d) Soc-hamsterster; (e) Asian-last.fm; (f) Router.

Table 3 Overall spreading ability (R) of the top-20 nodes of eight methods on the six networks. For the NC indicator, each value is the average overall spreading ability among 100 realizations, which are based on 100 different spanning trees[a)]

Networks	NC	EC	DC	CC	KC	CI	HI	CR
C. elegans	**0.4621**	0.3654	0.3673	0.3770	0.4505	0.3632	0.3698	0.3706
USA airports	**0.1431**	0.1224	0.1231	0.1232	0.1322	0.1274	0.1232	0.1268
Yeast	0.1150	0.0956	0.0959	**0.1174**	0.0987	0.1002	0.0969	0.1101
Soc-hamsterster	**0.2096**	0.1917	0.1865	0.1875	0.2007	0.1867	0.1939	0.1890
Asian-last.fm	**0.1496**	0.1266	0.1395	0.1458	0.1297	0.1395	0.1247	0.1408
Router	**0.0671**	0.0319	0.0546	0.0565	0.0346	0.0508	0.0600	0.0652

a) The bold value in each line represents the best performing index.

method can obtain alternative top spreaders to achieve a similar spreading performance when some nodes cannot be chosen as spreading sources.

4 Conclusion

Cycles have been increasingly acknowledged to play an important role in network structures and functions, but their potential in identifying influential spreaders has remained largely unexplored. In this study, the characteristics of basic cycles, a typical type of cycles in networks, are analyzed, and a method to identify multiple influential spreaders based on the basic cycles is proposed. We first illustrate the distributions of the basic cycle size and the number of involved basic cycles of nodes on six empirical networks and three model networks. Based on the analysis, we propose the NC indicator to assess nodes' influence in spreading. The correlation analysis between the rankings by NC and seven benchmark indicators indicates that NC contains relatively more information beyond degree-based indicators. The experimental results reveal that the proposed NC has the best spreading performance compared with well-known centrality indicators. NC also performs better than a newly proposed cycle-based method, which merely considers the shortest cycles. We further investigated the average distance among spreaders identified by NC and found that they are more scattered in networks than those identified by benchmarks, which can effectively avoid the influence redundancy of the top spreaders. Most importantly, we reveal that the multiple spreaders identified by NC have the lowest cost under the same spreading performance, and the proposed method can produce alternative choices of multiple spreaders.

Nonetheless, this work has two limitations: First, we treated each basic cycle independently and equally, without considering the underlying features of each basic cycle, such as length, weight, and direction. Addressing them would extend the research scope to weighted or directed networks and differentiate different roles of every cycle and thus provide a deepening understanding of network cycles in the influential spreader identification and influence maximization problem. Second, we focused on the first-order cycles

in networks, which are formed by links. Extending the analysis to high-order cycles, which are formed by triangles, tetrahedrons, or other high-order simplices, will help enrich the studies on network science and bring up more opportunities for the study of network spreading and other dynamics.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. T2293771), STI 2030-Major Projects (Grant No. 2022ZD0211400), and the New Cornerstone Science Foundation through the XPLORER PRIZE.

Supporting information Appendixes A–D. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Goh K I, Cusick M E, Valle D, et al. The human disease network. Proc Natl Acad Sci USA, 2007, 104: 8685–8690
2 Guille A, Hacid H, Favre C, et al. Information diffusion in online social networks: a survey. ACM SIGMOD Rec, 2013, 42: 17–28
3 Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, 2003. 137–146
4 Yang L, Li Z, Giua A. Containment of rumor spread in complex social networks. Inf Sci, 2020, 506: 113–130
5 Lü L, Medo M, Yeung C H, et al. Recommender systems. Phys Reports, 2012, 519: 1–49
6 Lü L, Chen D, Ren X L, et al. Vital nodes identification in complex networks. Phys Reports, 2016, 650: 1–63
7 Newman M. Networks. Oxford: Oxford University Press, 2018
8 Fan T, Li H, Ren X L, et al. The rise and fall of countries on world trade web: a network perspective. Int J Mod Phys C, 2021, 32: 2150121
9 Freeman L C. A set of measures of centrality based on betweenness. Sociometry, 1977, 40: 35–41
10 Bonacich P. Factoring and weighting approaches to status scores and clique identification. J Math Sociol, 1972, 2: 113–120
11 Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Comput Networks ISDN Syst, 1998, 30: 107–117
12 Li P X, Ren Y Q, Xi Y M. An importance measure of actors (set) within a network. Syst Eng, 2004, 22: 13–20
13 Morone F, Makse H A. Influence maximization in complex networks through optimal percolation. Nature, 2015, 524: 65–68
14 Morone F, Min B, Bo L, et al. Collective influence algorithm to find influencers via optimal percolation in massively large social media. Sci Rep, 2016, 6: 1–11
15 Qiu Z, Fan T, Li M, et al. Identifying vital nodes by Achlioptas process. New J Phys, 2021, 23: 033036
16 Liu J G, Lin J H, Guo Q, et al. Locating influential nodes via dynamics-sensitive centrality. Sci Rep, 2016, 6: 21380
17 Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks. Nat Phys, 2010, 6: 888–893
18 Wang X, Zhang X, Zhao C, et al. Effectively identifying multiple influential spreaders in term of the backward-forward propagation. Phys A-Stat Mech Its Appl, 2018, 512: 404–413
19 Ji S, Lü L, Yeung C H, et al. Effective spreading from multiple leaders identified by percolation in the susceptible-infected-recovered (SIR) model. New J Phys, 2017, 19: 073020
20 Shi D, Chen G, Thong W W K, et al. Searching for optimal network topology with best possible synchronizability. IEEE Circuits Syst Mag, 2013, 13: 66–75
21 Sizemore A E, Giusti C, Kahn A, et al. Cliques and cavities in the human connectome. J Comput Neurosci, 2018, 44: 115–145
22 Lizier J T, Atay F M, Jost J. Information storage, loop motifs, and clustered structure in complex networks. Phys Rev E, 2012, 86: 026110
23 Petermann T, Rios P D L. Role of clustering and gridlike ordering in epidemic spreading. Phys Rev E, 2004, 69: 066116
24 Fan T, Lü L, Shi D, et al. Characterizing cycle structure in complex networks. Commun Phys, 2021, 4: 272
25 Korn A, Schubert A, Telcs A. Lobby index in networks. Phys A-Stat Mech Its Appl, 2009, 388: 2221–2226
26 Lü L, Zhou T, Zhang Q M, et al. The H-index of a network node and its relation to degree and coreness. Nat Commun, 2016, 7: 10168
27 Freeman L C. Centrality in social networks conceptual clarification. Soc Networks, 1978, 1: 215–239
28 Anderson R M, May R M. Infectious Diseases of Humans: Dynamics and Control. Oxford: Oxford University Press, 1992
29 Liu J G, Wang Z Y, Guo Q, et al. Identifying multiple influential spreaders via local structural similarity. Europhys Lett, 2017, 119: 18001
30 Lü L, Zhang Y C, Yeung C H, et al. Leaders in social networks, the delicious case. Plos One, 2011, 6: e21202
31 Hirsch J E. An index to quantify an individual's scientific research output. Proc Natl Acad Sci USA, 2005, 102: 16569–16572
32 Hotelling H. Simplified calculation of principal components. Psychometrika, 1936, 1: 27–35
33 Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks. Nature, 1998, 393: 440–442
34 Opsahl T, Agneessens F, Skvoretz J. Node centrality in weighted networks: generalizing degree and shortest paths. Soc Networks, 2010, 32: 245–251
35 Jeong H, Mason S P, Barabási A L, et al. Lethality and centrality in protein networks. Nature, 2001, 411: 41–42
36 Rossi R, Ahmed N. The network data repository with interactive graph analytics and visualization. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence, Austin, 2015. 29: 4292–4293
37 Rozemberczki B, Sarkar R. Characteristic functions on graphs: birds of a feather, from statistical descriptors to parametric models. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020. 1325–1334
38 Spring N, Mahajan R, Wetherall D. Measuring ISP topologies with rocketfuel. ACM SIGCOMM Comput Commun Rev, 2002, 32: 133–145
39 Pastor-Satorras R, Castellano C, van Mieghem P, et al. Epidemic processes in complex networks. Rev Mod Phys, 2015, 87: 925–979
40 Newman M E J. Clustering and preferential attachment in growing networks. Phys Rev E, 2001, 64: 025102
41 Kendall M G. A new measure of rank correlation. Biometrika, 1938, 30: 81–93
42 Ma L, Ma C, Zhang H F, et al. Identifying influential spreaders in complex networks based on gravity formula. Phys A-Stat Mech Appl, 2016, 451: 205–212
43 Rodriguez A, Laio A. Clustering by fast search and find of density peaks. Science, 2014, 344: 1492–1496
44 Zhao X Y, Huang B, Tang M, et al. Identifying effective multiple spreaders by coloring complex networks. Europhys Lett, 2015, 108: 68005
45 Guo L, Lin J H, Guo Q, et al. Identifying multiple influential spreaders in term of the distance-based coloring. Phys Lett A, 2016, 380: 837–842
46 Hu Z L, Liu J G, Yang G Y, et al. Effects of the distance among multiple spreaders on the spreading. Europhys Lett, 2014, 106: 18002
47 Bondy J A, Murty U S R. Graph Theory With Applications. London: Macmillan Press, 1976