Journal of Informetrics 17 (2023) 101411

Contents lists available at ScienceDirect

Journal of Informetrics

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

Disruptive coefficient and 2-step disruptive coefficient: Novel measures for identifying vital nodes in complex networks

Alex J. Yang [a,b,c], Sanhong Deng [a,b,c,*], Hao Wang [a,b,c], Yiqin Zhang [a,b,c], Wenxia Yang [a,b,c]

[a] School of Information Management, Nanjing University, Nanjing, 210023, China
[b] Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing, 210023, China
[c] International Joint Informatics Laboratory, Nanjing University, Nanjing, 210023, China

ARTICLE INFO

Keywords:
Complex networks
Vital nodes
Power law
Science of science
Disruption

ABSTRACT

The identification and ranking of vital nodes in complex networks have been a critical issue for a long time. In this paper, we present an extension of existing disruptive metrics and introduce new ones, namely the disruptive coefficient (D) and 2-step disruptive coefficient (2-step D), as innovative tools for identifying critical nodes in complex networks. Our approach emphasizes the importance of disruptiveness in characterizing nodes within the network and detecting their criticality. Our new measures take into account both prior and posterior information of the focal nodes, by evaluating their ability to disrupt the previous network paradigm, setting them apart from traditional measures. We conduct an empirical analysis of four real-world networks to compare the rankings or identification of nodes using D and 2stepD with those obtained from four renowned benchmark measures, namely, degree, h-index, PageRank, and the CD index. Our analysis reveals significant differences between the nodes identified by D and 2stepD and those identified by the benchmark measures. We also examine the correlation coefficient and efficiency of the metrics and find that D and 2stepD have significant correlations with the CD index, but have weak correlations with the benchmark measures. Furthermore, we show that D and 2stepD outperform CD index and random ways in intentional attacks. We find power law distributions for D, 2stepD, and CD, indicating a small number of highly disruptive nodes and a large number of less disruptive nodes in the networks. Our results suggest that D and 2stepD are capable of providing valuable and distinct insights for identifying critical nodes in complex networks.

1. Introduction

The task of ranking and detecting vital nodes in complex networks (Newman, 2012) has attracted extensive attention across various fields, such as physics, bioinformatics, sociology, management science, science of science, etc. (Fan et al., 2021; Fortunato et al., 2018; Gysi et al., 2021; Iniguez et al., 2022; Lu et al., 2016; Park et al., 2023; Ullah et al., 2021; Zhang et al., 2019). Identifying vital nodes is a crucial and fundamental task that plays a prominent role in building recommendation systems (Cai et al., 2022), finding targeted proteins or drugs (Gysi et al., 2021), selecting influential spreaders (Iniguez et al., 2022), quantifying academic influences (Fortunato et al., 2018; Jo et al., 2022), and predicting success or failures in careers (Yin et al., 2019).

* Corresponding author.
E-mail addresses: alexjieyang@outlook.com (A.J. Yang), sanhong@nju.edu.cn (S. Deng).

https://doi.org/10.1016/j.joi.2023.101411
Received 1 October 2022; Received in revised form 29 March 2023; Accepted 16 April 2023
1751-1577/© 2023 Elsevier Ltd. All rights reserved.

Many real-world networks are scale-free networks (Barabasi & Albert, 1999), which means that the effect and function of different nodes can vary significantly (Lu et al., 2016). To accurately portray the prominence of different nodes, scientists have explored a pool of node detection algorithms from various perspectives (Fan et al., 2021; Gleich, 2015; Jo et al., 2022; Kitsak et al., 2010; Lü et al., 2016; Zhai et al., 2018). In this paper, we propose a set of novel measures for critical node identification from the perspective of disruptiveness (Funk & Owen-Smith, 2017; Leydesdorff & Bornmann, 2021; Leydesdorff et al., 2021; Park et al., 2023; Wu et al., 2019). These measures are inherently different from conventional measures in complex networks and aim to comprehensively and accurately identify critical nodes.

The concept of disruptiveness is closely related to novelty (Uzzi et al., 2013; Wang et al., 2017), innovation (Hofstra et al., 2020), and originality (Dumas et al., 2021). In the realm of science of science, the disruptiveness of publications (Bornmann et al., 2020; Wu et al., 2019), patents (Funk & Owen-Smith, 2017; Park et al., 2023), scientists (Bedru et al., 2022), and research themes (Wang et al., 2022) has been widely studied, with a focus on how these entities disrupt previous paradigms and create new insights (Leydesdorff & Bornmann, 2021). We believe that these ideas can be extended and applied to any directed network.

The main objective of this paper is to propose novel metrics for quantifying the disruptiveness of nodes in complex networks and to compare them with traditional measures through extensive experimentation. Our metrics aim to capture the extent to which a node can disrupt the previous network paradigm, which sets them apart from conventional measures. By exploring disruptiveness in complex networks, we aim to provide insights into the critical roles that nodes can play in shaping and transforming network behavior. Our paper primarily focuses on investigating the correlation of the new measures with the traditional ones and analyzing the power-law distribution of these measures in real-world networks.

The article is structured as follows: (1) We begin with a comprehensive literature review of measures for vital node detection and disruptive perspectives, establishing the research foundations. (2) Next, we expand on the concept of disruptiveness and introduce the 2-step version (2-step D) of the disruption coefficient (D) as a measure for identifying critical nodes in complex networks from the perspective of disruptiveness. We explain the fundamental properties of these measures. (3) We then conduct an empirical study on four real-world networks to compare the performance of our proposed measures with traditional benchmarks. (4) Subsequently, we analyze the results using various techniques, including Spearman correlation analysis, node ranking analysis, vital node analysis, percolation analysis, and power law distribution analysis, to thoroughly evaluate the consistency, correlations, and distributions of the new metrics and traditional benchmarks. (5) Finally, we discuss and summarize our findings and suggest possible directions for future research.

2. Literature review

2.1. Vital node identification

The identification of critical nodes in complex networks has become an important area of research, as it can help to optimize the use of limited resources, facilitate information dissemination (Ullah et al., 2021), and improve access to information in networks (Fan et al., 2021). Since Freeman's (1977) introduction of centrality metrics based on betweenness, the field of critical node detection methods in complex networks has rapidly expanded.

Conventional node detection algorithms, as outlined by Lu et al. (2016), typically fall into three categories: (i) neighbor node-based measures, such as degree (Kitsak et al., 2010), and h-index (Lü et al., 2016); (ii) path-based measures, such as closeness, betweenness (Freeman, 1977), and information index (Poulin et al., 2000); and (iii) iterative refinement measures, such as PageRank (Gleich, 2015), HITs, and SALSA.

In addition to these conventional measures, researchers have proposed other critical node detection measures based on the network topology of nodes, such as the Giant index (Jo et al., 2022), the cycle structure index (Fan et al., 2021), and the bi-directional h-index (Zhai et al., 2018). To date, various measures of nodes' impact have been proposed in complex networks, but none of them have explicitly taken into account the perspective of disruptiveness. Our study aims to bridge this gap by proposing novel measures for identifying critical nodes based on their disruptiveness.

2.2. Disruptive perspectives

The notion of disruptiveness has its origins in the paradigm-shifting models put forward by Popper (1934) and Kuhn (1962), which highlight that the development of science is a never-ending process established in revolution. Disruptive research tends to disrupt the paradigms of science with novel ideas. The first widely recognized disruptiveness-related notion was the concept of disruptive technology with business models (Bower, 1995). More recently, the notion of disruptiveness in citation networks, put forward by Funk and Owen (2017), has aroused extensive studies and further discussion. Wu et al. (2019) have proposed disruption indices in bibliometrics to represent whether a paper destabilizes or consolidates existing documents in an informetric network. It has been argued that disruptive papers "represent punctuated advances beyond previous theory, methods or findings" (Bornmann et al., 2020). The concept of disruptiveness refers to the disapproval or reformation of previous science paradigms, and its essence is the proportion relation for cited and citing documents of the focal paper in citation networks (Funk & Owen-Smith, 2017; Park et al., 2023; Wu et al., 2019).

Disruption represents the notion of knowledge discontinuity in information networks (Leydesdorff & Bornmann, 2021). Numerous scholars have explored the application of disruption indices in detecting milestone publications (Bornmann et al., 2020; Wu et al., 2019), essential patents (Funk & Owen-Smith, 2017; Park et al., 2023), exceptional scientists (Bedru et al., 2022), and novel research

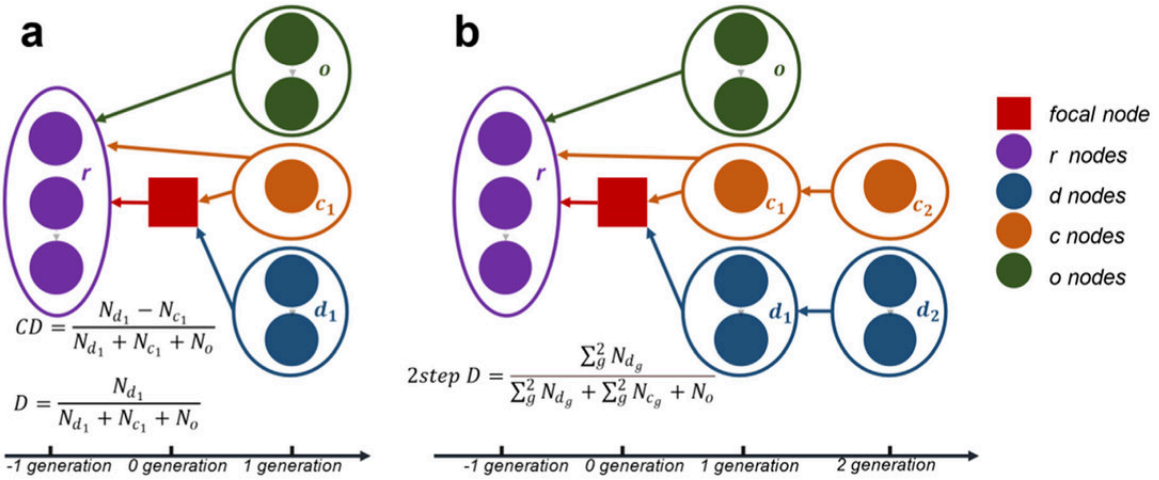


Fig. 1. Illustrations of the CD, D, and 2stepD measures in directed networks. (a) Adjacent nodes within the first generation of the focal node are divided into four categories: r nodes, o nodes, c nodes, and d nodes. (b) The topological structures of the 2stepD measure, wherein 2-order neighbor nodes are taken into consideration.

themes (Wang et al., 2022). However, the concept of disruptiveness has not yet been thoroughly investigated in identifying critical nodes in complex networks.

Therefore, building on these ideas, we extend and propose new measures for detecting vital nodes in complex networks that take into account the disruptiveness connotations of nodes. In the following sections, we provide a theoretical analysis of these new measures, including their definitions and algorithms.

3. Theoretical aspects

3.1. Disruption coefficients

The objective of this paper is to introduce a new set of metrics for ranking and identifying critical nodes in complex networks. Our proposed metrics go beyond traditional measures and emphasize the importance of disruptiveness in determining the significance of nodes within the network. Our study primarily focuses on unweighted directed connected networks. However, the related notions we present can also be extended to weighted networks. In a directed network, the node being evaluated is referred to as the focal node, as shown in Fig. 1. The prior neighboring nodes of the focal node are denoted by the symbol *r*. The posterior neighboring nodes of the focal node are divided into two types: *c* nodes (consolidating nodes) and *d* nodes (disruptive nodes). The *c* nodes are connected to both the focal node and at least one of the *r* nodes, while *d* nodes are connected to the focal node but not to any of the *r* nodes. We also define *o* nodes (posterior nodes) as nodes that are connected to at least one of the *r* nodes but not to the focal node.

To measure the disruptiveness of a focal paper in publication citation networks, the CD index (Park et al., 2023) is used and is calculated:

$$CD = \frac{N_d - N_c}{N_d + N_c + N_o} \tag{1}$$

To identify and rank vital nodes in more general directed networks, not limited to citation networks, we introduce new metrics, the disruptive coefficient (D) and 2-step disruptive coefficient (2stepD), by extending the CD index to complex networks and considering multiple generations.

The Disruptive coefficient (D) is defined as

$$D = \frac{N_{d_1}}{N_{d_1} + N_{c_1} + N_o} \tag{2}$$

where N_d, N_c and N_o represent the numbers of d_1 nodes, c_1 nodes, and o_1 nodes in the network, respectively. This equation is similar to the CD index formula, but the numerator is N_d instead of (N_c - N_d). This modification provides a more precise and consistent description of the level of disruptiveness (Leydesdorff & Bornmann, 2021) and is identical to DI^* proposed by Leydesdorff et al. (2021) in citation networks.

The 2-step disruptive coefficient is defined as:

$$2step\ D = \frac{\sum_g^2 N_{d_g}}{\sum_g^2 N_{d_g} + \sum_g^2 N_{c_g} + N_o} \tag{3}$$

The posterior nodes of the focal node in the first generation are divided into two groups: c_1 nodes and d_1 nodes, and in the second generation are divided into c_2 nodes and d_2 nodes. The c_1 nodes are connected to both the focal node and at least one of the r nodes, while d_1 nodes are connected to the focal node but not to any of the r nodes. Note that the 2-step measure can be extended to k-step measures, but it is not elaborated on due to space constraints.

3.2. Theoretical sense

The D formula and the CD index exhibit a high degree of similarity, suggesting that disruptiveness can be seen as a driving factor for the focal node that disrupts the existing network paradigm and introduces novel information. Conversely, consolidation can be seen as the opposite effect, where the focal node mainly transfers existing information (Funk & Owen-Smith, 2017; Leydesdorff & Bornmann, 2021; Wu et al., 2019). We believe that the CD index combines both disruption and consolidation, making it less consistent with our aim to focus solely on the disruptiveness of nodes. Both D and CD leverage both prior and posterior information of the focal node's neighboring nodes. We propose that d nodes can be viewed as disruptive, and c nodes as consolidating. With more d nodes and fewer c nodes, the focal node is likely to have a greater disruptive effect on the network. Consolidation and disruption can be thought of as a type of balance on a particular dimension, where less disruption implies more consolidation, and vice versa.

The 2stepD measure builds upon the general idea of D but considers a broader scope of information by examining more nodes and multiple generations of neighbors in the network, even though it is more complex and computationally intensive. Both D and 2stepD range from 0 to 1. We also propose several properties of D and 2stepD in the context of directed networks.

Proposition 1. In a directed connected network, if a node has out-degree 0, then its D equals 0. If a node has in-degree 0, then its D equals 1.

Proof. For a node with out-degree 0, $N_c + N_o = 0$, so $D = N_d/(N_d + N_c + N_o) = 1$. For a node with in-degree, $N_d = 0$, so $N_d/(N_d + N_c + N_o) = 0$.

Proposition 2. The range of D is [0,1] for all nodes in any directed acyclic network.

Proof. A directed acyclic graph contains a node with out-degree 0 (the starting node) and a node with in-degree 0 (the ending node), by definition. According to Proposition 2, these nodes have $D = 0$ and $D = 1$, respectively. Since D is a ratio between 0 and the total number of neighbors, its range is [0,1] for all other nodes in the network.

3.3. Benchmark measures

We selected five renowned conventional node detection algorithms in complex networks as benchmark measures: the CD index, the degree, the h-index (Lü et al., 2016), and the PageRank (Gleich, 2015) as benchmark measures.

For a focal node v_i, degree is the number of adjacent nodes. The h-index of v_i equals h if h adjacent nodes connected to v_i each have a degree of at least h (Lü et al., 2016). PageRank is a random walk model based on eigenvector in directed networks and is defined as (Brin & Page, 1998; Pinski & Narin, 1976):

$$\mathrm{PageRank}(v_i) = \frac{(1-d)}{N} + d\sum_{i=1}^{n}\frac{PageRank(v_j)}{outdegree(v_j)} \tag{4}$$

where $indegree(v_j)$ indicates the in-degree of v_j, and d is the damping factor, we choose d=0.85 in this study.

We selected CD as it is a well-established indicator for measuring node disruption and consolidation. Degree, h-index, and PageRank are classic indicators commonly used to identify vital nodes. Degree and h-index are local measures as they only consider adjacent nodes, while PageRank is a global metric (Ullah et al., 2021). We contend that D, 2stepD, and CD are also local measures.

4. Empirical studies

4.1. Data and preprocessing

We have selected four real-world networks for our empirical studies. (1) The first network is the Wikipedia Network (Sun et al., 2016), which consists of 1182 nodes and 13,754 edges. It represents the communication network of the Breton Wikipedia, where nodes correspond to users, and edges indicate that a user wrote a message on the talk page of another user at a certain time. (2) The second network is the Blog Network (Kunegis, 2013) with 1125 nodes and 19,025 edges. This network includes front-page hyperlinks between blogs related to the US election, where nodes represent blogs, and edges represent hyperlinks between them. (3) The third network is the Email Network (Kunegis, 2013) with 2030 nodes and 39,264 edges. This directed network consists of emails related to the democratic national committee email leak, where a node in the Email Network represents an email user, and edges represent the emails that were sent between users. (4) The last network is the APS Network, which is a large-scale citation network based on citation data provided by the American Physical Society (APS). The network consists of 463,348 research papers and 4745,622 citation links and covers all literature and citation data from 1893 to 2010, including 19 journals from the Physical Review series.

Table 1
Structure overview of four real-world networks.

Network	Nodes	Edges	Mean degree	Mean distance	Reciprocity
Wikipedia network	1 182	13 754	23	3	0.33
Blog network	1 225	19 025	31	2.7	0.42
Email network	2 030	39 264	39	3.4	0.24
APS network	463 348	4 745 622	21	9.3	0.00

Table 2
Descriptive statistics of D, 2stepD and benchmark measures in four networks.

Network	Variable	Mean	Median	S. D.	Min	Max
Wikipedia	D	0.547	1	0.497	0	1
	2stepD	0.545	1	0.498	0	1
	CD	0.401	1	0.723	−1	1
	degree	23.272	3	165.670	0	3297
	h-index	2.744	2	4.300	0	45
	PageRank	8.46E-4	2.83E-4	1.72E-3	2.61E-4	3.43E-2
Blog	D	0.218	0.008	0.389	0	1
	2stepD	0.195	0.002	0.397	0	1
	CD	0.168	−0.006	0.418	−0.83	1
	degree	31.061	14	45.335	0	467
	h-index	17.317	12	16.599	0	69
	PageRank	8.16E-4	2.68E-4	1.57E-3	1.97E-4	1.88E-2
Email	D	0.261	0	0.435	0	1
	2stepD	0.258	0	0.466	0	1
	CD	0.237	0	0.467	−1	1
	degree	38.684	2	239.253	0	5813
	h-index	3.114	1	7.011	0	74
	PageRank	4.93E-4	2.03E-4	2.32E-3	1.84E-4	6.33E-2
APS	D	0.048	0.001	0.173	0	1
	2stepD	0.033	0.012	0.162	0	1
	CD	0.023	−0.002	−0.002	−1	1
	degree	20.484	15	32.529	0	5474
	h-index	3.329	2	4.142	0	106
	PageRank	2.16E-6	6.48E-7	1.14E-5	4.29E-7	2.56E-3

We developed Python-3.9.10 programs to modify the selected networks to meet the requirements of igraph-0.10.2 (Ju et al., 2016) and networkx-2.8.8 (Hagberg et al., 2008), calculate the related measures, and visualize the networks. We then used scipy-1.10.1 (Virtanen et al., 2020) to perform Spearman correlation analysis between D, 2stepD, and benchmark measures.

Table 1 presents the meta information and basic statistics of the four real-world networks used in the study, including the number of nodes, edges, average degree, average shortest distance, and reciprocity of each network. We note that the sizes of the networks are within the same order of magnitude, except for the APS Network, which is significantly larger.

4.2. Descriptive statistics

To initiate the analysis, we computed D, 2stepD, and four benchmark measures in four different networks. The descriptive statistics, including mean values, median values, standard deviation, and the extremums of these measures, are presented in Table 2. The range of D and 2stepD in all four networks was observed to be between 0 and 1, while the range of CD was between −1 and 1. Notably, the benchmark measures varied greatly in the four networks. For example, the degree of Wikipedia Network ranged from 0 to 3297, while that of APS Network ranged from 0 to 151. The pairwise distributions of metrics in the APS Network are shown in Appendix Fig. A1, indicating all measures are heterogeneous.

We also observed that the distributions of these measures were heterogeneous across the four networks. The measures in all four networks were not uniformly or normally distributed, as evidenced by the significant differences between mean and median values. To further investigate the distributions of these measures, we explored their power-law properties in the following sections. These findings suggest that D, 2stepD, and the benchmark measures exhibit diverse characteristics across different networks, highlighting the need to carefully choose and compare measures based on the specific network under study.

5. Results and discussion

5.1. Correlation

In order to investigate the consistency of the measures, we computed the Spearman correlation coefficient between the measures proposed in this study and the benchmark measures. Fig. 2 displays the Spearman correlation coefficients for each pair of distinct mea-

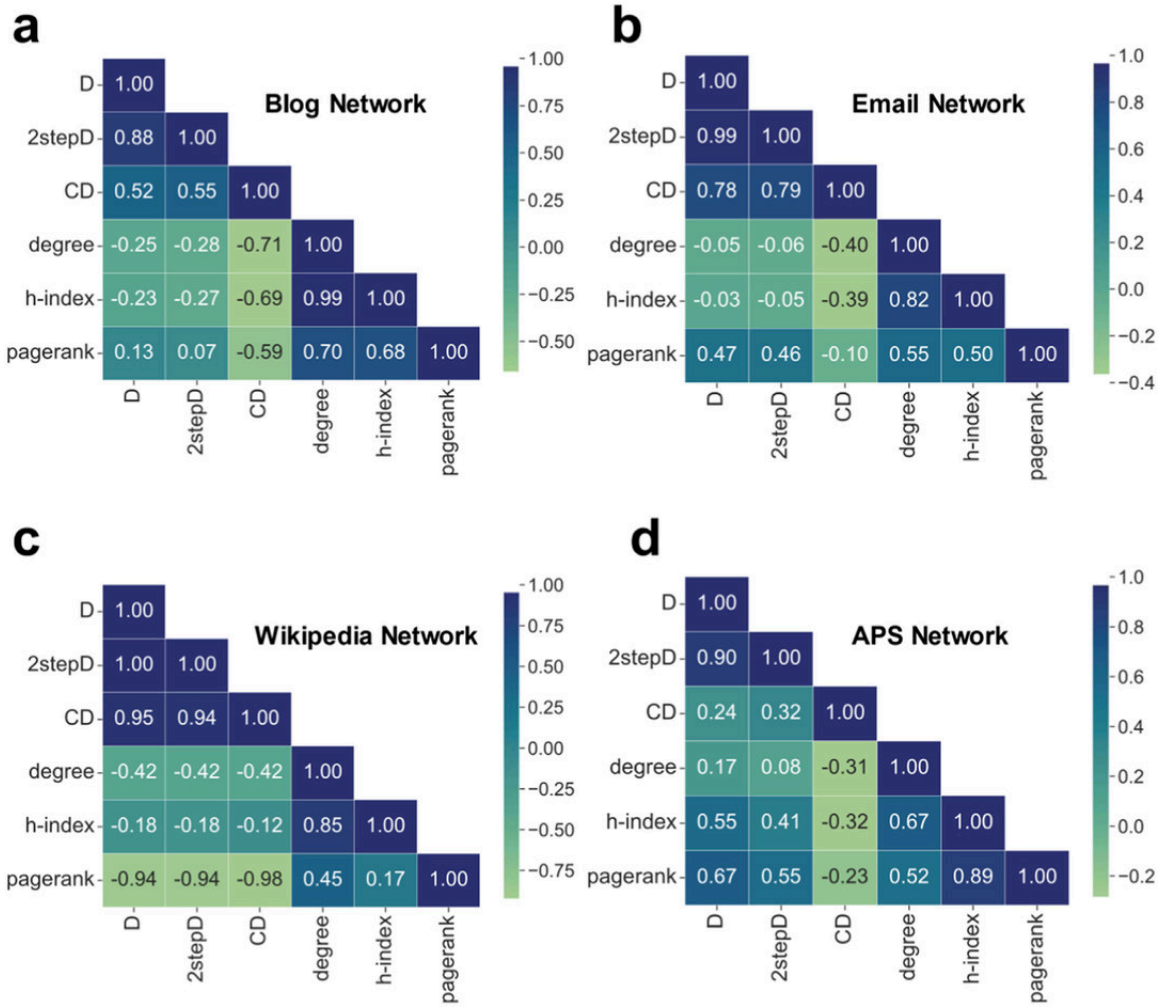


Fig. 2. Spearman Correlation Analysis of D, 2stepD, CD, Degree, h-index, and PageRank for the Blog Network (a), the Email Network (b), the Wikipedia Network (c), and the APS Network (d), Respectively. The Analysis is Based on Raw Data Using Two-Tailed Spearman Correlation.

sures, namely D, 2stepD, CD, degree, h-index, and PageRank, respectively. Furthermore, we plotted scatter plots for some noteworthy pairs using the APS Network as an example, as shown in Fig. A1.

Degree and h-index exhibit markedly high correlations with each other in all four networks, with Spearman correlation coefficients ranging from 0.67 to 0.99, which is in line with the fact that the h-index is essentially based on degree (Fan et al., 2021), and hence they share high similarities. On the other hand, the correlation coefficients between D and 2stepD are very high, exceeding 0.88 in all four networks and even reaching 1 in the Wikipedia Network (Fig. 2C), indicating that these two measures are highly similar.

Interestingly, the correlation between D and CD exhibits significant variation across different networks. While they are highly consistent in the Wikipedia Network (Fig. 2C), in the APS Network their correlation coefficient is very low (Fig. 2D). This result is not surprising, as the noticeable difference between D and CD in the APS Network reflects the distinction between pure disruption and consolidation-disruption, while their high consistency in the Wikipedia Network suggests that this difference may not be significant in certain networks.

The correlations between D and degree as well as h-index exhibit high heterogeneity across the four networks. Specifically, the correlations are negatively low in the Blog Network (Fig. 2A), Email Network (Fig. 2B), and the Wikipedia Network (Fig. 2C), but positively weak in the APS Network (Fig. 2D). This finding implies that the relationship between D and these traditional measures of node importance differs depending on the network context. One possible explanation for the negative correlations in the first three networks is that D captures a different aspect of node importance than degree and h-index, which are both based solely on node degree. In contrast, the positive weak correlation in the APS Network suggests that degree and h-index are somewhat indicative of node disruption in this network, but not as strongly as in the other three networks. It is worth noting that the correlations between D and other traditional measures, such as PageRank and CD, also vary across the four networks, indicating the importance of considering multiple measures of node importance when analyzing network structure and dynamics.

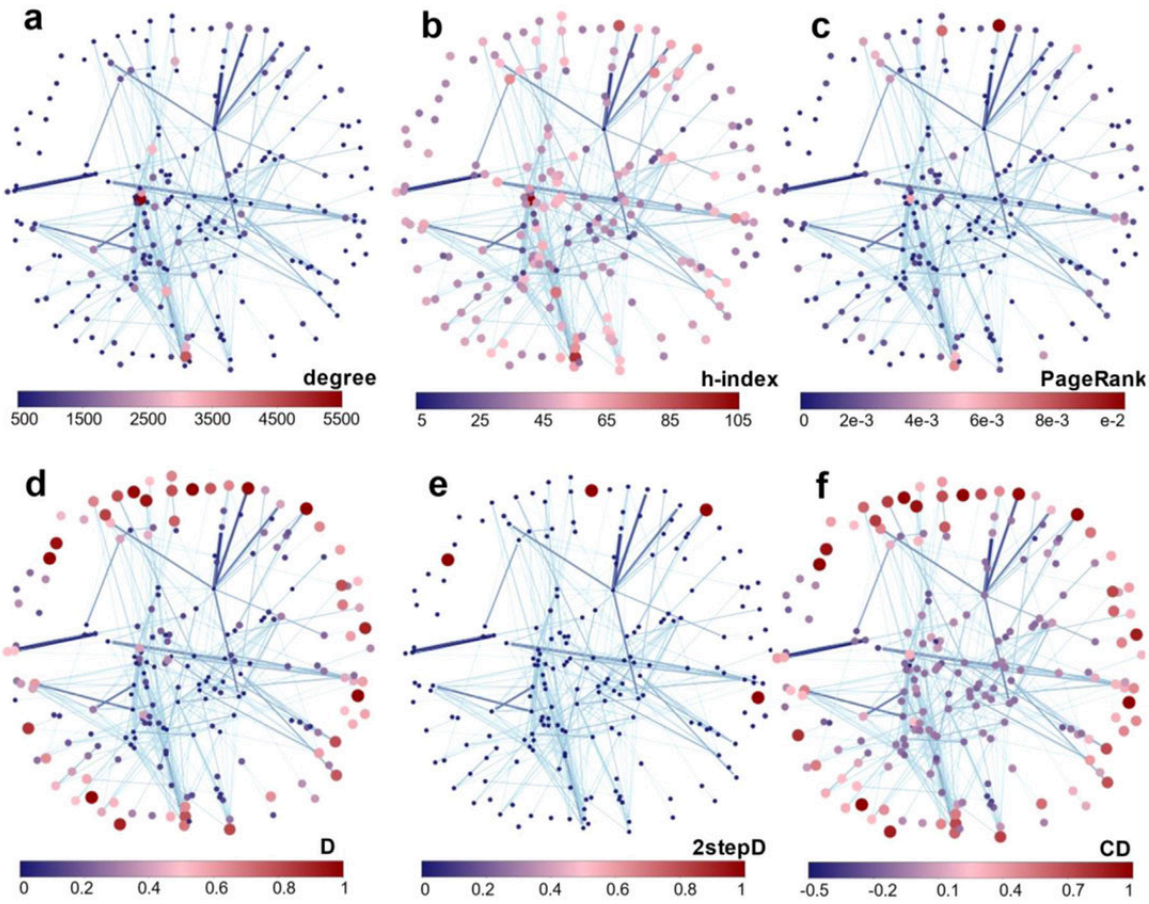


Fig. 3. Network structure and rankings of critical nodes in the APS Network using different node detection algorithms. Node sizes and colors indicate the ranking of nodes based on (a) degree, (b) h-index, (c) PageRank, (d) D, (e) 2stepD, and (f) CD index. The network layout is fixed, and edge widths represent edge-betweenness. Only nodes with a degree greater than 500 are shown.

This finding illustrates that D and 2stepD cannot be replaced by other measures, as their low correlations with benchmark measures demonstrate that they capture distinct information (Zhai et al., 2018). Therefore, we argue that D and 2stepD are irreplaceable for measuring nodes, as they provide valuable insights that differ from those obtained by benchmark measures.

5.2. Rank nodes

According to Zhang et al. (2019), each algorithm allows for the ranking of all nodes within a given network, while novel ones exhibit different patterns. To demonstrate this, we present the ranking outcomes of degree, h-index, PageRank, D, 2stepD, and CD for the APS Network (Fig. 3). Specifically, we limit the nodes with a degree greater than 500, resulting in 206 nodes, and fix the layout and positions of all nodes within each subplot. This allows for the focal nodes to be distinguished by their position in the network.

Nodes highly ranked by D (Fig. 3d) and CD (Fig. 3f) appear to be located on the periphery of the network. In contrast, nodes ranked highly 2stepD (Fig. 3e) exhibits a markedly different pattern. These ranking outcomes differ substantially from degree, h-index and PageRank. Highly ranked nodes according to D and 2stepD exhibit unique and destructive positions in the network, which are not necessarily associated with high degree or PageRank. This suggests that the disruptiveness captured by D and 2stepD is distinct from other node properties, providing valuable insights beyond traditional measures.

The analysis presented in this part highlights the distinctive nature of the disruptiveness measures (D and 2stepD) in comparison to the traditional measures such as degree, h-index, and PageRank. The clustering of nodes with high rankings by degree, h-index, and PageRank in the middle of the network, consistent with the rich club phenomenon (Pedreschi et al., 2022), is an expected outcome. However, the observation that nodes highly ranked by D and 2stepD exhibit unique and destructive positions in the network, which are not necessarily associated with high degree or PageRank, suggests that these measures capture novel and valuable information about nodes' roles in the network.

Furthermore, the observation that nodes highly ranked by CD share similarities with those ranked by D and 2stepD supports the notion that CD is a useful measure for capturing disruption and consolidation-disruption patterns in the network. Overall, this analysis

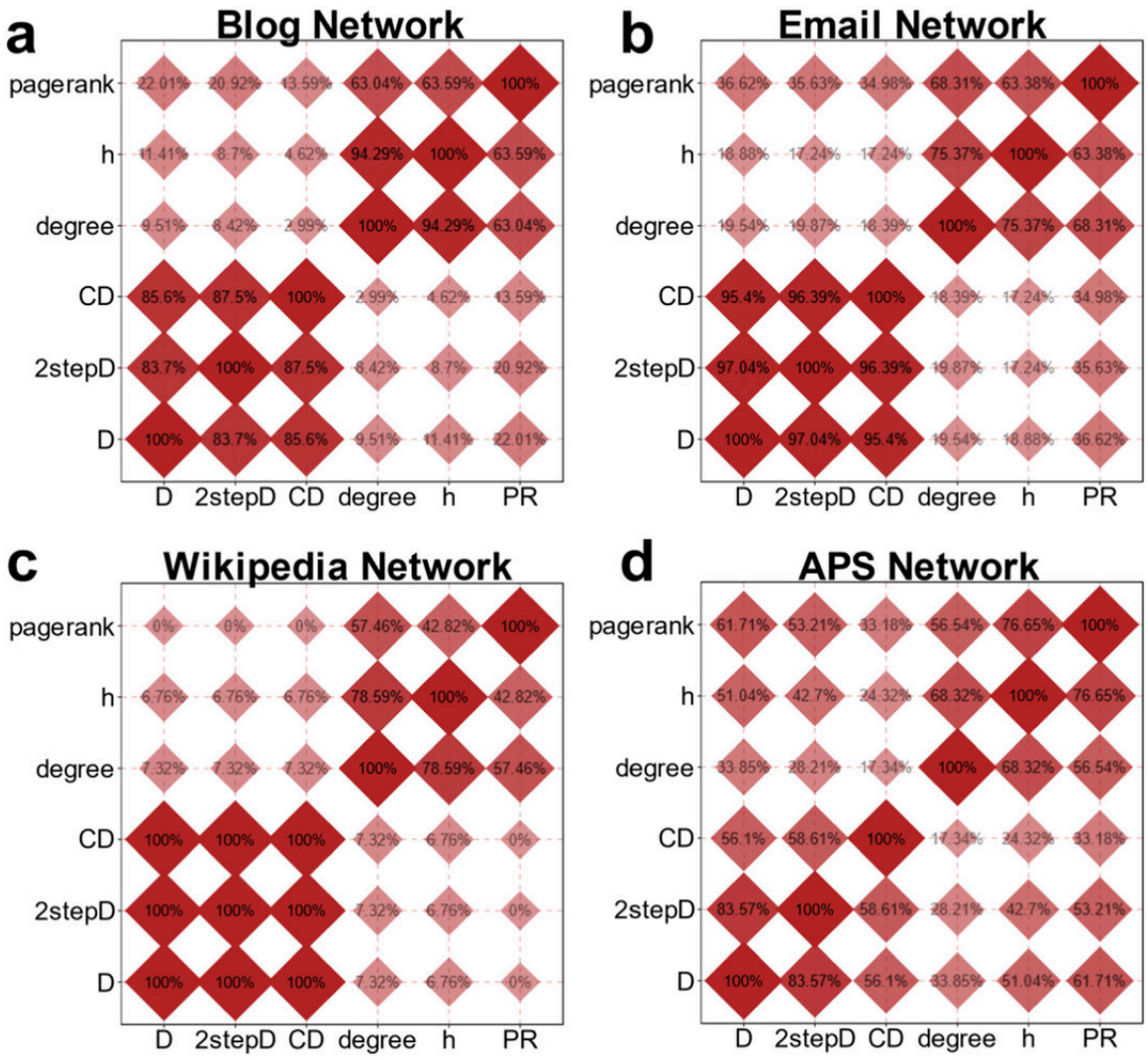


Fig. 4. Bubble matrix of the overlap ratio of top vital nodes identified by different measures in four networks. Bubble matrix shows the overlap ratio of top 30% vital nodes identified by D, 2stepD, CD, degree, h-index and PageRank in (a) Wikipedia Network, (b) Blog Network, (c) Email Network, and (d) APS Network. The color depth and sizes of bubbles are proportional to the overlap ratio within each network.

underscores the distinct and informative nature of the disruptiveness measures and their potential to uncover novel insights beyond traditional measures.

5.3. Overlap ratio

To gain a deeper understanding of the differences between the vital nodes identified by D, 2stepD, CD, and other benchmark measures, we calculated the overlap ratios of nodes selected by these measures in the four networks. Fig. 4 shows the results at the top 30% level. The findings demonstrate that the vital nodes identified by D and 2stepD exhibit a high overlap ratio with each other, ranging from 83% to 95% in all four networks. Furthermore, they also share a relatively high overlap ratio with the vital nodes selected by CD (more than 85%) in the Blog Network, the Email Network, and the Wikipedia Network, but a lower overlap ratio of 56.1% in the APS Network. In contrast, the vital nodes identified by D, 2stepD, and CD exhibit great divergence from nodes selected by other benchmarks, such as degree, h-index, and PageRank, in all four networks. The overlap ratios between these measures were consistently lower than 10% in the Wikipedia Network (Fig. 4c). We also performed a robustness check at the 10% level (Appendix Fig. A2), which yielded similar results and confirmed that D, 2stepD, and CD have even lower overlap ratios with other benchmarks. Overall, these findings demonstrate that D and 2stepD exhibit significantly different patterns in identifying the most critical nodes compared to other benchmark measures.

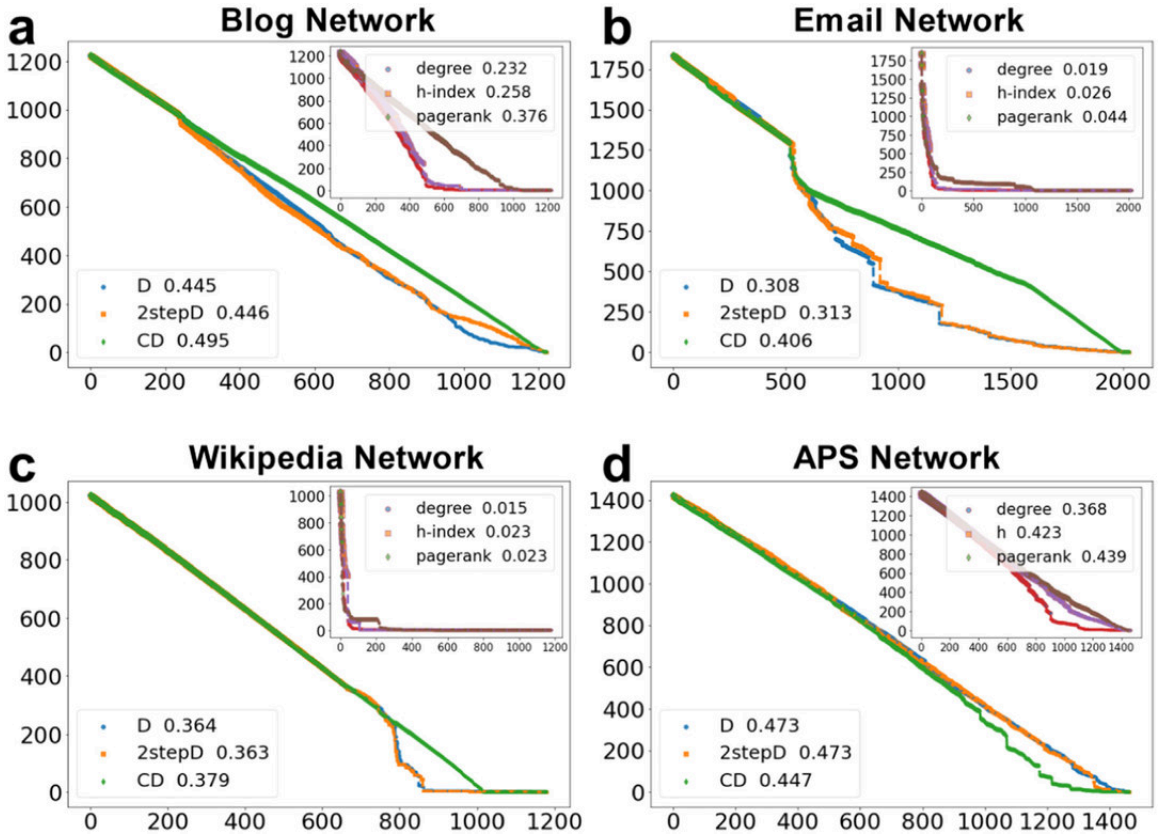


Fig. 5. Percolation outcomes and network robustness under intentional attacks by different algorithms. The x-axis represents the number of dropped nodes, and the y-axis represents the size of the largest connected subgraph after dropping vital nodes in (a) the Blog Network, (b) the Email Network, (c) the Wikipedia Network, and (d) the APS Network. The robustness of D, 2stepD, and CD is shown in each plot, and each inset shows the robustness of other benchmark measures. We only calculate the nodes with degree larger than 200 in the APS Network due to the computational power limitation.

The results suggest that D and 2stepD share a high degree of overlap in their selections. And in some real-world networks, the high overlap ratio between D, 2stepD, and CD indicates that these measures may capture similar network properties related to disruption and consolidation-disruption. Additionally, the low overlap ratio between D and other benchmark measures, such as degree, h-index, and PageRank, suggests that they identify different types of critical nodes in the network. This finding highlights the importance of considering multiple measures when identifying critical nodes in complex networks.

5.4. Percolation

To assess the importance of the vital nodes identified by each algorithm in maintaining network connectivity, we performed the percolation analysis (Li et al., 2021) with intentional attacks, by removing the top-ranked nodes according to each measure and tracked the resulting size of the largest connected subnetwork. We measured the effectiveness of the intentional attacks using the robustness metric (Gross & Barth, 2022), defined as the average size of the largest connected subnetwork after removing nodes, normalized by the total number of nodes in the network (i.e., $\text{Robustness} = \frac{1}{n}\sum_{i=1}^{n} size(g_i)$). A higher reduction in robustness indicates a greater impact of the removed nodes on network connectivity.

Fig. 5 shows the percolation results and robustness for each measure in the four real-world networks. Notably, intentional attacks using D and 2stepD exhibit significantly lower efficiency in disrupting network connectivity compared to degree, h-index, and PageRank, but outperform CD and random attack. This implies that nodes identified as vital by D and 2stepD play a unique and distinct role in maintaining network connectivity, which is not captured by traditional measures like degree and PageRank, but is similar to the role of nodes identified by CD.

In addition, we observe that the robustness of D and 2stepD is relatively stable in all four networks, while the robustness of other benchmark measures varies significantly depending on the network. For example, the robustness of degree, h-index, and PageRank is high in the Blog Network, but much lower in the Email Network and Wikipedia Network. Overall, these findings suggest that D, 2stepD, and CD perform not well in identifying critical nodes with respect to network robustness under intentional attacks.

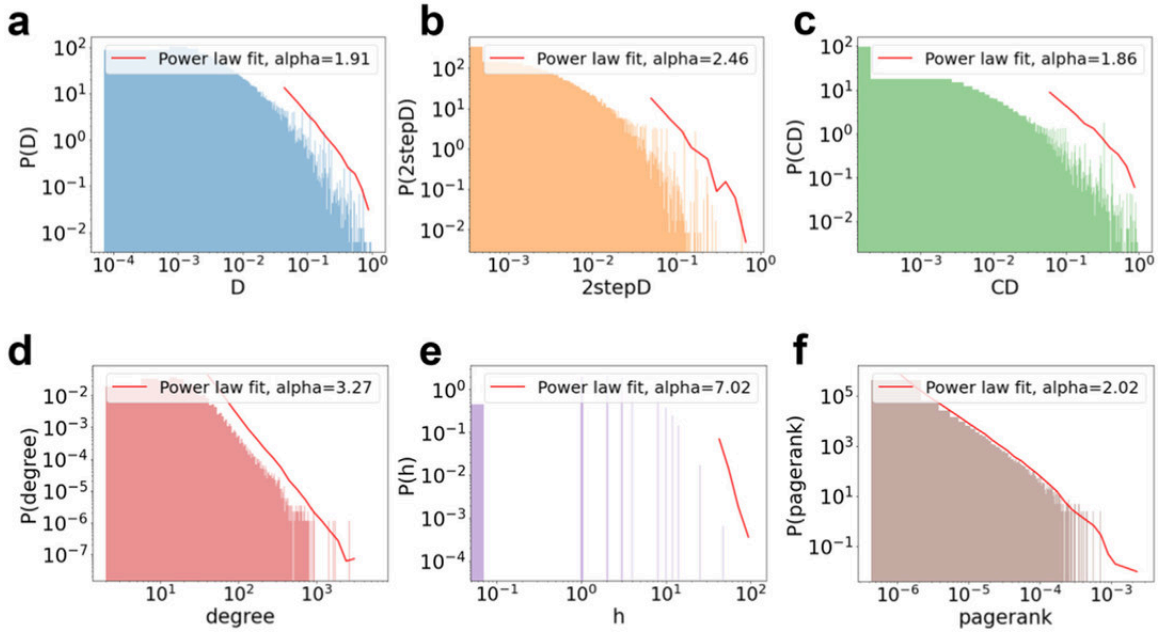


Fig. 6. Power law distribution fitting results for D, 2stepD, CD, degree, h-index, and PageRank in the APS Network. The histograms and fitting curves for each measure are displayed in (a) D, (b) 2stepD, (c) CD, (d) degree, (e) h-index, and (f) PageRank. All measures exhibit significant power law distributions.

5.5. Power law

The power law distribution of metrics in complex networks is universal and crucial (Barabasi & Albert, 1999; Peterson et al., 2010), as it shows the preferential attachment mechanism (Barabasi & Albert, 1999) as well as the Matthew effect (Merton, 1968) in science.

In this study, we employed the function $N(x) \sim x^{-\alpha}$ to fit the power law distributions of nodes under each metric, and we probed into the power law distribution of measures. We utilized the powerlaw package (Alstott et al., 2014) to fit the power law distributions of D, 2stepD, CD, degree, h-index, and PageRank. Based on the power law distribution fitting results presented in Fig. 6 and Appendix Fig. A3, the degree, h-index and PageRank follow a power law distribution, revealing the presence of a small number of highly connected or influential nodes and a large number of less connected or less influential nodes in the networks.

In the APS Network (Fig. 6), the fitting alpha for D is 1.91, for 2stepD is 2.46, for CD is 1.86, for degree is 3.27, for h-index is 7.02, and for PageRank is 2.02. The alpha values are indicators of the level of heterogeneity in the distribution of node importance measures, with smaller values indicating a more heterogeneous distribution. The results suggest that the distributions of D, 2stepD, and CD are more heterogeneous than that of degree and h-index, indicating that these metrics are better at capturing the characteristic of highly disruptive nodes in the networks. Similar results were obtained for other networks (Appendix Fig. A3). Overall, these power law distribution fitting results provide valuable insights into the underlying structure of the networks and the distribution of node disruptiveness measures.

6. Discussion

6.1. Conclusion

In the current study, we introduced an extension of the disruption coefficient (D) and proposed a 2-step version of it (2stepD) as fundamental measures for nodes in directed networks, emphasizing the role of disruptiveness in identifying critical nodes. The construction of new measures was inspired by the CD index proposed recently (Funk & Owen-Smith, 2017; Wu et al., 2019). An in-depth empirical study was operationalized in comparison with four well-known benchmark measures in four real-world networks.

Taken together, this study extends the disruptive measures for identifying critical nodes in complex networks from a disruptiveness perspective. We demonstrate the effectiveness and applicability of these measures in both synthetic and real-world networks. Specifically, our findings suggest that D and 2stepD are valuable and effective measures for identifying critical nodes, providing a distinct approach from conventional node identification measures. The power law distribution fitting results show that all measures, including D, 2stepD and CD, follow a power law distribution. This suggests that there is a small number of highly disruptive nodes, and a large number of less disruptive or more consolidating nodes in the networks. Our study reveals that D and 2stepD consistently identify critical nodes, with high overlap ratios in all four networks, while differing significantly from benchmark measures such

as degree, h-index, and PageRank. Furthermore, they exhibit superior performance in network robustness under intentional attacks compared to CD index.

6.2. Theoretical implications

Theoretical implications of our study are significant in the field of network science and science of science. First, we introduce a new perspective of measuring the disruptiveness of nodes, which is fundamentally different from traditional node-detecting measures (Lu et al., 2016). Second, we reveal that the new measures, D and 2stepD, are capable of identifying totally different critical nodes in networks compared to traditional benchmarks. Third, our study sheds light on the importance of considering the role of nodes in network connectivity and the effects of intentional attacks on network robustness. Fourth, our study highlights the importance of considering both prior and posterior information when identifying critical nodes in complex networks, which is a departure from traditional measures that only focus on the current state of the network (Li et al., 2021). These findings have significant implications for the identification of critical nodes in real-world systems, such as transportation networks (Diao et al., 2021; Erhardt et al., 2019), social networks (Kim & Hastak, 2018), and biological networks (Gysi et al., 2021). Our study highlights the importance of developing new measures and approaches for node identification in complex networks, which can lead to a better understanding of network structure and function.

Our study also sheds light on the power law distribution of node importance measures in complex networks, providing important insights into the underlying the scale-free structure of the networks. The power-law distribution fitting results of D, 2stepD and CD highlight the presence of highly disruptive nodes in the networks, along with a large number of less disruptive or more consolidating nodes. This pattern is indicative of the preferential attachment mechanism (Barabasi & Albert, 1999) as well as the Matthew effect (Merton, 1968) in science.

Our new measures of D and 2stepD can fill in the gaps in the field of critical node identification in complex networks from the perspective of disruptiveness, and inspires further research on the identification of vital nodes in directed networks and contribute to the development of more efficient network disruption strategies in various fields.

6.3. Practical implications

Our study has practical implications in various fields such as bioinformatics (Gysi et al., 2021), social network analysis (Kim & Hastak, 2018), knowledge management (Wang et al., 2022), and transportation planning (Diao et al., 2021; Erhardt et al., 2019). The power law distribution fitting results of D, 2stepD, CD, degree, h-index, and PageRank can guide the selection of appropriate measures for identifying critical nodes in different types of networks. The D and 2stepD measures can be used to identify important proteins in bioinformation networks that are not necessarily highly connected, as well as to identify key individuals in social networks who may have low degree centrality but high disruptiveness. In knowledge networks, our measures can help to identify disruptive ideas or concepts that are not widely adopted but have the potential to reshape the entire field. In transportation networks, the measures can be used to identify pivotal transit nodes that are critical for maintaining the connectivity and efficiency of the network. The potential applications and extensions of D and 2stepD in various domains are extensive, and our study lays the groundwork for future research to explore their full practical potential.

6.4. Limitation

The present study has several limitations that need to be addressed in future research. First, while we proposed the 2stepD measure, which considers two generations of neighbor nodes, we did not examine measures that consider three, four, or more generations of neighbor nodes. Future research could investigate the effectiveness of these higher-order measures. Second, our analysis is limited to unweighted networks, and it would be interesting to explore how our measures perform in weighted networks. Third, it is worth noting that D and 2stepD are designed for directed connected networks and may not be applicable in other types of networks.

6.5. Prospects

We believe that there are numerous expansions on the basis of this research, including (1) the convergent validity of nodes' disruptiveness in real-world networks, especially in information networks, (2) the properties of D and how to extract subgraphs based on D, (3) extending local measures like D and 2stepD to global measures (k-step-D) by considering many generations of neighbors, (4) epidemic spreading and pinning control analysis for D and 2stepD, (5) studying D and 2stepD in weighted networks, and (6) exploring the mechanisms of the power law distributions of D, 2stepD and CD in real-world networks. Additionally, future research can explore the application of D and 2stepD in various domains, such as bioinformation networks, social networks, knowledge networks, and traffic networks, and compare their effectiveness with existing measures. Overall, we believe that our proposed measures have great potential in identifying vital nodes in various complex networks, and we look forward to exploring their applications and extensions in future research.

Data availability

The link data for the Wikipedia Network, the Blog Network, and the Email Network were obtained from Networks (konect.cc), and were downloaded on January 1, 2023. We provide the node data with each metric on our GitHub repository at https://github.com/AlexJieYang/D-2stepD.

The APS dataset was obtained from APS Data Sets for Research. Node data with each metric is kept in our local database and can be accessed upon reasonable request.

Code availability

We have made all the codes used to preprocess and redraw the figures in our paper publicly available at https://github.com/AlexJieYang/D-2stepD.

Declaration of Competing Interest

No other author has reported a potential conflict of interest relevant to this article.

CRediT authorship contribution statement

Alex J. Yang: Conceptualization, Methodology, Visualization, Writing – original draft, Software. **Sanhong Deng:** Data curation, Writing – review & editing. **Hao Wang:** Supervision, Validation. **Yiqin Zhang:** Data curation, Investigation, Validation. **Wenxia Yang:** Investigation, Validation.

Acknowledgments

The authors would like to thank three anonymous reviewers for valuable advises. This work was supported by the National Social Science Fund of China (19BTQ062), the International Joint Informatics Laboratory Fund, and the Fundamental Research Funds for the Central Universities.

Appendix

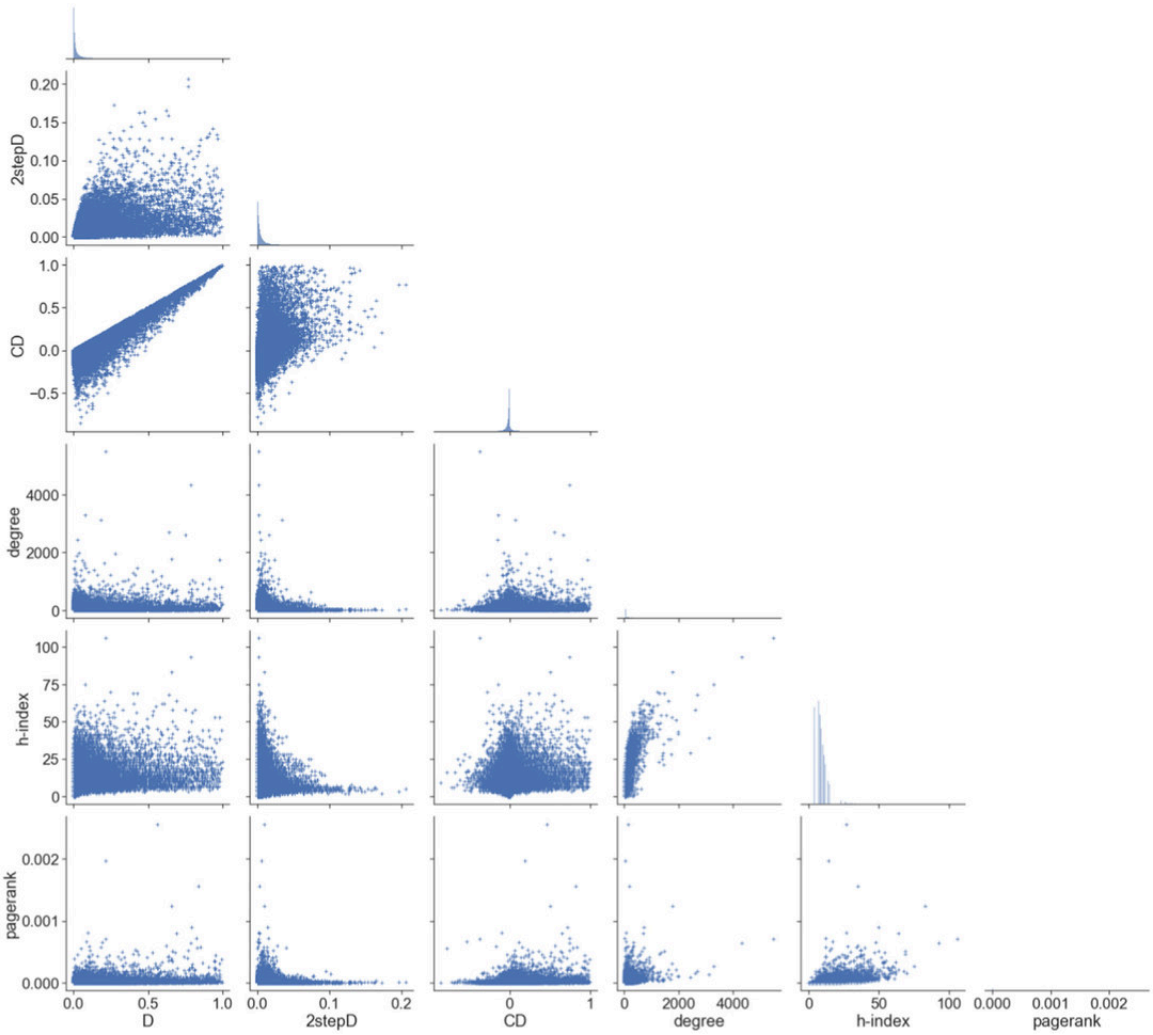


Fig. A1. Taking the APS Network as an example, the scatter plots and distribution histograms of D, 2stepD and four benchmark measures (raw data).

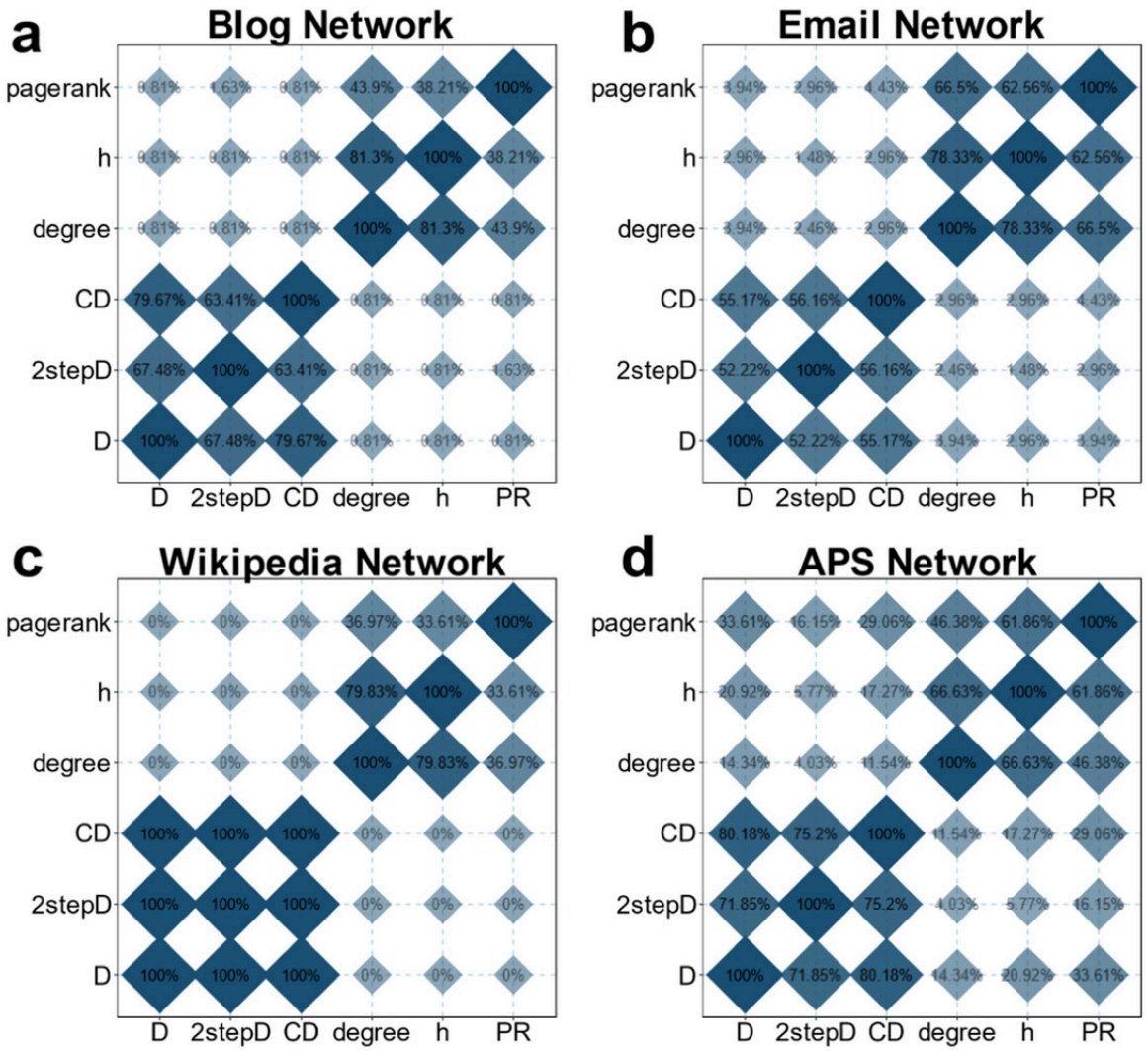


Fig. A2. Bubble matrix of the overlap ratio of top vital nodes identified by different measures in four networks. Bubble matrix shows the overlap ratio of top 10% vital nodes identified by D, 2stepD, CD, degree, h-index and PageRank in (a) Wikipedia Network, (b) Blog Network, (c) Email Network, and (d) APS Network. The color depth and sizes of bubbles are proportional to the overlap ratio within each network.

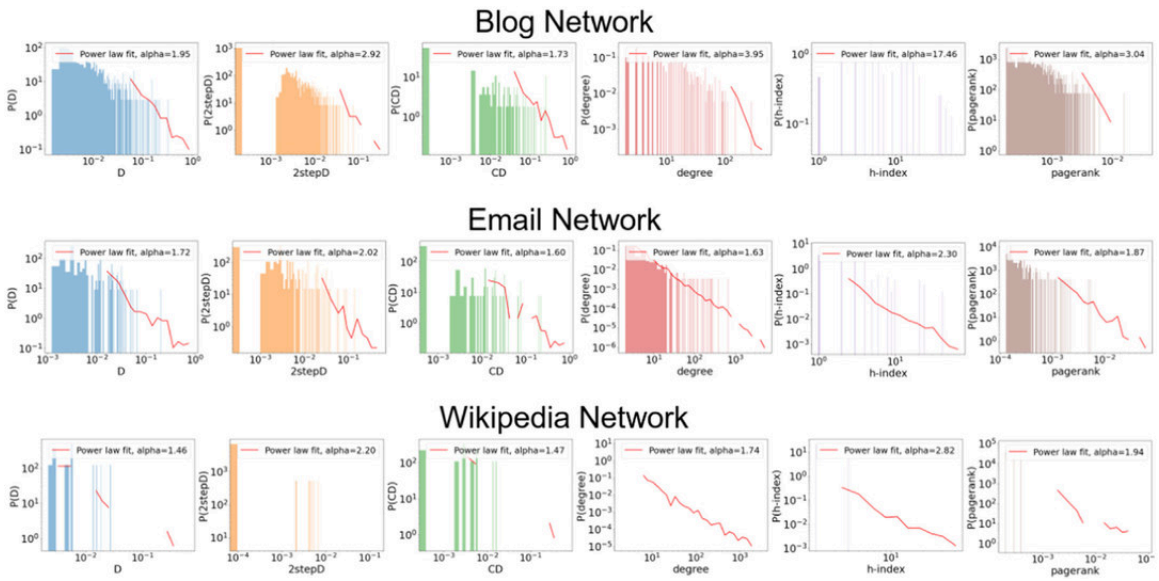


Fig. A3. Power law distribution fitting results for D, 2stepD, CD, degree, h-index, and PageRank in the Blog Network, Email Network and Wikipedia network.

References

Alstott, J., Bullmore, E., & Plenz, D. (2014). Powerlaw: A python package for analysis of heavy-tailed distributions. *PloS One, 9*(1), e85777. 10.1371/journal.pone.0085777.

Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science (New York, N.Y.), 286*(5439), 509–512. 10.1126/science.286.5439.509.

Bedru, H. D., Zhang, C., Xie, F., Yu, S., & Hussain, I. (2022). CLARA: Citation and similarity-based author ranking. *Scientometrics*. 10.1007/s11192-022-04590-5.

Bornmann, L., Devarakonda, S., Tekles, A., & Chacko, G (2020). Disruptive papers published in Scientometrics: Meaningful results by using an improved variant of the disruption index originally proposed by Wu, Wang, and Evans (2019). *Scientometrics, 123*(2), 1149–1155. 10.1007/s11192-020-03406-8.

Bower, J. L., & Christensen, C. M. (1995). Disruptive technologies: Catching the wave *(Vol. 73(1))*. Harvard Business Review.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117. 10.1016/s0169-7552(98)00110-x.

Cai, X., Wang, N., Yang, L., & Mei, X. (2022). Global-local neighborhood based network representation for citation recommendation [Article]. *Applied Intelligence, 52*(9), 10098–10115. 10.1007/s10489-021-02964-5.

Diao, M., Kong, H., & Zhao, J. (2021). Impacts of transportation network companies on urban mobility. *Nature Sustainability, 4*(6), 494–500. 10.1038/s41893-020-00678-z.

Dumas, D., Organisciak, P., & Doherty, M. (2021). Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics Creativity and the Arts, 15*(4), 645–663. 10.1037/aca0000319.

Erhardt, G. D., Roy, S., Cooper, D., Sana, B., Chen, M., & Castiglione, J. (2019). Do transportation network companies decrease or increase congestion? *Science Advances, 5*(5) eaau2670. 10.1126/sciadv.aau2670.

Fan, T., Lü, L., Shi, D., & Zhou, T. (2021). Characterizing cycle structure in complex networks. *Communications Physics, 4*(1), 272. 10.1038/s42005-021-00781-3.

Fortunato, S., Bergstrom, C. T., Boerner, K., Evans, J. A., Helbing, D., Milojevic, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabasi, A.-L. (2018). Science of science. *Science (New York, N.Y.), 359*(6379) eaao0185. 10.1126/science.aao0185.

Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry, 40*, 35–41. 10.2307/3033543.

Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science, 63*(3), 791–817. 10.1287/mnsc.2015.2366.

Gleich, D. F. (2015). PageRank beyond the web. *Siam Review, 57*(3), 321–363. 10.1137/140976649.

Gross, T., & Barth, L. (2022). Network robustness revisited. *Frontiers in Physics, 10*, Article 823564. 10.3389/fphy.2022.823564.

Gysi, D. M., do Valle, I., Zitnik, M., Ameli, A., Gan, X., Varol, O., Ghiassian, S. D., Patten, J. J., Davey, R. A., Loscalzo, J., & Barabasi, A.-L. (2021). Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proceedings of the National Academy of Sciences of the United States of America, 118*(19), Article e2025581118. 10.1073/pnas.2025581118.

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkX. *Proceedings of the 7th Python in Science Conference, Pasadena, CA USA*.

Hofstra, B., Kulkarni, V. V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., & McFarland, D. A. (2020). The diversity-innovation paradox in science. *Proceedings of the National Academy of Sciences, 117*(17), 9284–9291. 10.1073/pnas.1915378117.

Iniguez, G., Pineda, C., Gershenson, C., & Barabasi, A. L. (2022). Dynamics of ranking. *Nature Communications, 13*(1), 1646. 10.1038/s41467-022-29256-x.

Jo, W. S., Liu, L., & Wang, D. (2022). See further upon the giants: Quantifying intellectual lineage in science. *Quantitative Science Studies, 3*(2), 319–330. 10.1162/qss_a_00186.

Ju, W. Y., Li, J. X., Yu, W. R., & Zhang, R. C. (2016). iGraph: An incremental data processing system for dynamic graph. *Frontiers of Computer Science, 10*(3), 462–476. 10.1007/s11704-016-5485-7.

Kim, J., & Hastak, M. (2018). Social network analysis: Characteristics of online social networks after a disaster [Article]. *International Journal of Information Management, 38*(1), 86–96. 10.1016/j.ijinfomgt.2017.08.003.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature Physics, 6*(11), 888–893. 10.1038/nphys1746.

Kuhn, T. S. (1962). Historical structure of scientific discovery. *Science (New York, N.Y.), 136*(3518), 760–764.

Kunegis, J. (2013). *KONECT: The Koblenz network collection.* https://doi.org/10.1145/2487788.2488173

Leydesdorff, L., & Bornmann, L. (2021). Disruption indices and their calculation using web-of-science data: Indicators of historical developments or evolutionary dynamics? *Journal of Informetrics, 15*(4). 10.1016/j.joi.2021.101219.

Leydesdorff, L., Tekles, A., & Bornmann, L. (2021). A proposal to revise the disruption index. *El Profesional de la Información*. 10.3145/epi.2021.ene.21.
Li, M., Liu, R.-. R., Lü, L., Hu, M.-. B., Xu, S., & Zhang, Y.-. C. (2021). Percolation on complex networks: Theory and application. *Physics Reports, 907*, 1–68. 10.1016/j.physrep.2020.12.003.
Lu, L., Chen, D., Ren, X.-. L., Zhang, Q.-M., Zhang, Y.-C., & Zhou, T. (2016). Vital nodes identification in complex networks [Review]. *Physics Reports-Review Section of Physics Letters, 650*, 1–63. 10.1016/j.physrep.2016.06.007.
Lü, L., Zhou, T., Zhang, Q.-. M., & Stanley, H. E. (2016). The H-index of a network node and its relation to degree and coreness. *Nature Communications, 7*(1), 10168. 10.1038/ncomms10168.
Merton, R. K. (1968). The matthew effect in science. *Science (New York, N.Y.), 159*(3810), 56–63 doi:. 10.1126/science.159.3810.56.
Newman, M. E. J. (2012). Communities, modules and large-scale structure in networks. *Nature Physics, 8*(1), 25–31. 10.1038/nphys2162.
Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature, 613*(7942), 138–144. 10.1038/s41586-022-05543-x.
Pedreschi, N., Battaglia, D., & Barrat, A. (2022). The temporal rich club phenomenon. *Nature Physics, 18*(8) 931-+. 10.1038/s41567-022-01634-8.
Peterson, G. J., Pressé, S., & Dill, K. A. (2010). Nonuniversal power law scaling in the probability distribution of scientific citations. *Proceedings of the National Academy of Sciences, 107*(37), 16023–16027. 10.1073/pnas.1010757107.
Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management, 12*(5), 297–312. 10.1016/0306-4573(76)90048-0.
Popper, K. R. (1934). *The logic of scientific discovery*. Basic Books.
Poulin, R., Boily, M. C., & Mâsse, B. R. (2000). Dynamical systems to define centrality in social networks. *Social Networks, 22*(3), 187–220. 10.1016/S0378-8733(00)00020-4.
Sun, J., Kunegis, J., & Staab, S. (2016). *Predicting User Roles in Social Networks Using Transfer Learning with Feature Transformation*. https://doi.org/10.1109/ICDMW.2016.0026
Ullah, A., Wang, B., Sheng, J., Long, J., Khan, N., & Sun, Z. (2021). Identifying vital nodes from local and global perspectives in complex networks. *Expert Systems with Applications, 186*, Article 115778. 10.1016/j.eswa.2021.115778.
Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical Combinations and Scientific Impact. *Science (New York, N.Y.), 342*(6157), 468–472. 10.1126/science.1240474.
Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., . . . SciPy, C. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*(3), 261–272. 10.1038/s41592-019-0686-2.
Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy, 46*(8), 1416–1436. 10.1016/j.respol.2017.06.006.
Wang, S., Ma, Y., Mao, J., Bai, Y., Liang, Z., & Li, G. (2022). Quantifying scientific breakthroughs by a novel disruption indicator based on knowledge entities. *Journal of the Association for Information Science and Technology*. 10.1002/asi.24719.
Wu, L. F., Wang, D. S., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature, 566*(7744) 378-+. 10.1038/s41586-019-0941-9.
Yin, Y., Wang, Y., Evans, J. A., & Wang, D. (2019). Quantifying the dynamics of failure across science, startups and security. *Nature, 575*(7781), 190–194. 10.1038/s41586-019-1725-y.
Zhai, L., Yan, X., & Zhang, G. (2018). Bi-directional h-index: A new measure of node centrality in weighted and directed networks. *Journal of Informetrics, 12*(1), 299–314. 10.1016/j.joi.2018.01.004.
Zhang, S., Medo, M., Lü, L., & Mariani, M. S. (2019). The long-term impact of ranking algorithms in growing networks. *Information Sciences, 488*, 257–271. 10.1016/j.ins.2019.03.021.