Chaos, Solitons and Fractals 166 (2023) 112952

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

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

Efficient network intervention with sampling information ☆

Qi Mingze [a], Tan Suoyi [b], Chen Peng [a], Duan Xiaojun [a,*], Lu Xin [b,*]

[a] College of Science, National University of Defense Technology, Changsha, Hunan, 410073, PR China
[b] College of Systems Engineering, National University of Defense Technology, Changsha, Hunan, 410073, PR China

ARTICLE INFO

Keywords:
Network intervention
Network attack
Network immunization
Graph sampling
Incomplete information

ABSTRACT

Most existing studies assume that the network topology is already known when designing intervention strategies, which is difficult to achieve in practice. This paper focuses on network intervention with sampling information and assumes that the nodes are obtained by three typical graph sampling algorithms. The characteristics of sampling nodes' degrees and its influence on the design of intervention strategies are analyzed. Moreover, we propose a cutoff degree-based method for utilizing sampling information. Experiments in synthetic and real networks show that our method could effectively disintegrate networks by estimating networks' mean degrees with sampling information. The results depend on the degree preference of sampling algorithms and the accuracy of the average degree estimation. For sampling algorithms with high degree preference, the intervention effect of sampling partial data could approach that of complete data when selecting the appropriate cutoff degree value.

1. Introduction

Networks can effectively represent the structure and dynamics of experiential systems through interacting entities. Typical examples include the power grid, the Internet, social networks and biological networks [1–3]. Structural connectivity can greatly influence the function and dynamic of such networks, so many researchers focus on the robustness of networks in maintaining connectivity against random failures and targeted attacks [4,5]. Meanwhile, much attention is directed to another side of the problem to identify critical nodes (or links) that disproportionately influence networks [6]. Targeting these nodes for intervention can effectively disintegrate the network functions and prevent the epidemic dissemination of infectious diseases or malicious rumors [7–9]. Similar studies have also used terms such as network attack [4], network immunization [10], optimal percolation [6], network dismantling [11] and network disintegration [12].

Most studies of network intervention assume that information about the global structures of networks is known. In the beginning, researchers attack the nodes according to their importance as defined by structural centrality and improve the attack effect by recalculating the centrality scores of nodes during the removal process [13]. Furthermore, many heuristic algorithms have been proposed to identify the minimal set of nodes whose removal would disintegrate the network into many separate pieces [6,14]. Recently, some combinatorial optimization-based and machine learning-based algorithms were also introduced to develop effective network intervention strategies [15–17].

Although the above network intervention strategies are efficient, the unavoidable problem is that we hardly obtain complete network information. Therefore, many studies pay attention to network intervention under incomplete information in different ways [18]. According to the assumptions about network information, the research could be divided into incomplete global information-based methods and local information-based methods. The formers always occur in network attacks or robustness problems [18,19], and the latters are usually represented in network immunization or vaccination [20].

Incomplete global information research started from the robustness analysis of centrality measures with imperfect data [21]. The generating function methods were introduced to study the intentional attack under incomplete information [18,22], the optimal attack strategies

☆ S. Tan is supported by the National Natural Science Foundation of China (NSFC) [72001211] and the Hunan Science and Technology Plan Project, PR China (2020JJ5679). X. Lu is supported by the National Natural Science Foundation of China (NSFC) [72025405, 71790615, 91846301, 72088101], National Social Science Foundation of China (22ZDA102), the Shenzhen Basic Research Project for Development of Science and Technology (JCYJ20200109141218676, 202008291726500001), and the Innovation Team Project of Colleges in Guangdong Province (2020KCXTD040). X. Duan and M. Qi are supported by the National Natural Science Foundation of China (NSFC) [12101608, 62103422] and the Postgraduate Scientific Research Innovation Project of Hunan Province (CX20200001).

* Corresponding authors.
E-mail addresses: qimingze17@nudt.edu.cn (M. Qi), tansuoyi@nudt.edu.cn (S. Tan), pengchen19@nudt.edu.cn (P. Chen), xjduan@nudt.edu.cn (X. Duan), xin.lu@flowminder.org (X. Lu).

https://doi.org/10.1016/j.chaos.2022.112952
Received 7 July 2022; Received in revised form 10 October 2022; Accepted 23 November 2022
Available online 6 December 2022

by adjusting the attack proportions of different areas [23] and the robustness of subgraph obtained from uniform and nonuniform random sampling [19]. Moreover, considering random link missing, the link prediction methods are introduced to improve the network disintegration effect [24]. It is worth noting that these studies mainly consider random loss of network data and pay less attention to the ways and causes of data loss.

Local information-based methods focus on developing efficient immunization or vaccination strategies when each removal is based on local topology or neighbor information. Typical methods include acquaintance immunization and its variants, which immunize the random neighbors of randomly selected nodes [25]. The respondent-driven sampling method was also introduced to network immunization and specially adapted to hidden populations [26]. Moreover, researchers focused on reducing the cost of collecting information and improving the effectiveness of interventions by selectively data collection and vaccination [20,27]. Recently, a percolation framework was proposed to analyze of immunization under incomplete information where a group of nodes is observed at a time and the node with the highest degree is immunized [28], which was also expanded to other different attack models and conditions [29,30]. Although these models and algorithms are theoretically feasible, they ignore the global cost of sampling local information. In other words, in some algorithms, the total number of sampling nodes approaches or even exceeds the total number of nodes.

The acquisition of network data could be regarded as a sampling process from the underlying real network [31,32]. Due to limited sampling resources and individuals' accessibility, it is difficult to obtain all information on nodes and links. A sampling network is generally the subgraph consisting of the subsets of the nodes and links in the original network [33–35]. The related issue focuses on the property preservation and estimation of the actual network through different sampling algorithms [36,37]. Relevant conclusions show that the characteristics of incomplete network topology are closely related to sampling algorithms [38]. So the sampling information obtained from different sampling algorithms will largely influence the intervention strategies. At the same time, how to use sample information to improve the effectiveness of network intervention is also an important issue. Therefore, focusing on the above two aspects, we study effective network intervention with sampling information in this paper. On the one hand, we investigate the influence of network data obtained by different sampling algorithms on the intervention effects. On the other hand, we design effective intervention methods to maximize the use of sampling information.

The paper is organized as follows. We define the above problem and introduce three classical sampling algorithms for research in Section 2. The characteristics of the sampling algorithms and the available auxiliary information are also analyzed. In Section 3, an effective cutoff-degree based method for network intervention based on global mean estimation is presented. Experiments in artificial and real networks examine the selection of cutoff degree value and the method's effectiveness. At last, the article is summarized and discussed.

2. Model and materials

Let us denote an unweighted and undirected network $G = \langle V, E\rangle$. $V = \{v_1, v_2, \ldots, v_n\}$ is the set of nodes and $E \subseteq \{(v_i, v_j)|v_i \in V, v_j \in V, i \neq j\}$ is the set of edges between them, where (v_i, v_j) is an unordered pair. $N = |V|$ and $W = |E|$ are the numbers of nodes and edges, respectively. Let the degree k_i of the node v_i be the number of edges incident with the node and the average degree $\langle k\rangle = \sum_{i=1}^{N} k_i/N$.

Given the input network, we assume that the sampling network $G_s = \langle V_s, E_s\rangle$ is obtained through different graph sampling algorithms, in which $V_s \subseteq V$ is the set of sampling nodes, $E_s = \{e_{ij} = (v_i, v_j), e_{ij} \in E|v_i, v_j \in V_s\}$ and the sampling fraction $\alpha = |V_s|/|V|$. That is, G_s is the induced subgraph of G over the sampling nodes V_s. As shown in Fig. 1 (a), only a fraction of a node's neighbors may be selected for inclusion. On the other hand, we assume that sampling a node means obtaining its neighborhood information. For example, we could obtain the buddy list of one user in the network crawling, which is also the essential condition of many sampling algorithms. Meanwhile, the IDs of neighbors are difficult to obtain due to probing costs or accessibility. Therefore, during the sampling process, only the real degree of sampling nodes $K_s = \{k_i|v_i \in V_s\}$ could be obtained (shown in Fig. 1 (b)). The main problem we focus on is how to develop an effective intervention strategy according to the above sampling information. This is also known in some studies as the network attack or network immunization problem [32].

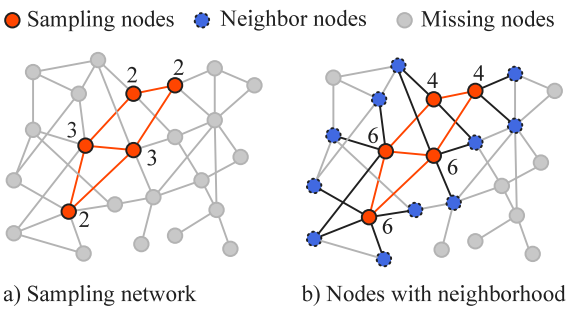

Fig. 1. Incomplete sampling information. The sampling network induced from sampling nodes (a) and the sampling nodes with a neighborhood (b). The sampling nodes, neighbor nodes, and missing nodes are given in different colors, and the number is the observed degree of nodes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For the incomplete network obtained by the sampling algorithms in the previous sections, how to formulate effective intervention strategies is our main concern in this paper. We assume that we can intervene all the nodes with limited sampling information, which means the removal of all connected links of nodes. For the sampling nodes in V_s, we could preferentially remove the most important nodes with the highest degree. But for the missing nodes in $\widetilde{V} = V - V_s$ without any attack information, we could only remove them in random order.

We take the size of the largest connected components P_∞ as the measure of network connectivity and use the critical removal fraction of nodes f_c to characterize the effect of intervention. We record the number of nodes N_r removed up to the point that $P_\infty < \sqrt{N}$, in which the network is almost completely disconnected and its spreading capability is severely limited. The threshold f_c is calculated as $f_c = N_r/N$. A smaller f_c implies more efficient intervention. In particular, for the uncorrelated networks, $\kappa \equiv \langle k^2\rangle/\langle k\rangle = 2$ is often also used as the general criterion, where $\langle k^2\rangle = \sum_{i=1}^{N} k_i^2/N$. The network is fragmented when $\kappa < 2$. For the networks whose scale is larger than 1000, we remove 1% of the total number of nodes at a time in the experiment.

2.1. Network sampling algorithms

The sampling information in the above problems is greatly influenced by the sampling algorithms. Classical network sampling techniques could be classified as node sampling, edge sampling and traversal-based sampling, in which we choose the simplest random algorithms [36].

Random node sampling (RNS): nodes in $V_s \subseteq V$ are chosen independently and uniformly at random. All edges among the sampling nodes are then added to form the sampling network $E_s = \{e_{ij} = (v_i, v_j), e_{ij} \in E|v_i, v_j \in V_s\}$.

Random edge sampling (RES): we select nodes in pairs by randomly sampling edges and including both endpoints. The sampling process stops when the target fraction ϕ of nodes is collected. Furthermore, the sampling network could be obtained by adding the other edges between sampling edges through the graph induction process. The sampling

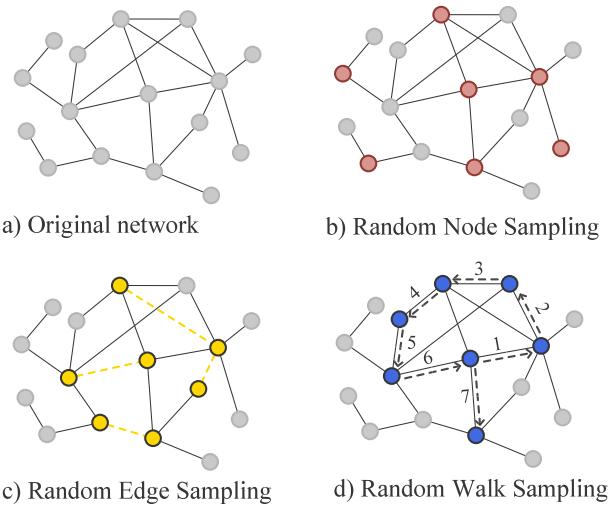

Fig. 2. Three graph sampling algorithms. For the target network (a), the three algorithms sample nodes by random node selection (b), random link selection (c) and random walk (d), respectively.

method introduced as the graph induction process is also called totally induced edge sampling [36].

The samples in the above two algorithms are uncorrelated, making them suitable for theoretical analysis. However, we could not perform them in most real applications due to different limitations, and traversal based sampling (TBS) is more practical. TBS starts from a small starting topology (such as a set of initial nodes) and expands the sample based on current observations. This paper considers a simple random walk sampling method, which is widely used for hidden populations and better simulates the network crawling.

Random walk sampling (RWS): start from a random initial node $v^{(0)}$. Then choose one of its neighboring nodes u uniformly and randomly and let $v^{(i)} = u$ at step i. Repeat t steps and until the expected fraction ϕ of nodes is collected. The sampling network $G_s = \langle V_s, E_s \rangle$ consists of $V_s = \{v^{(0)}, v^{(1)}, \ldots, v^{(t)}\}$ and $E_s = \{e_i | (v^{(i-1)}, v^{(i)}), i = 1, \ldots, t\}$. In particular, the RWS is memoryless and one can revisit some vertices, which make it appealing for theoretical analysis. Similarly, the graph-induced process could be introduced to obtain the additional edges between sampled nodes, and the overall process is also named induced subgraph random walk sampling.

The diagrams of the above algorithms are given in Fig. 2. More graph sampling methods can be found in the review articles [37,39]. According to the assumption of the last section, the main information we used for forming the intervention strategy is the real degree of sampling nodes K_s. Thus, what affects the effect of the intervention strategy is the distribution of the real degree of sampling nodes and the removing sequences of nodes in different areas (i.e., V_s and $\widetilde{V}$). Before forming a intervention strategy, we could first estimate the global degree distribution and average degree of the original network according to the degree distribution of the sampling nodes, which could help us better determine the removing sequence.

2.2. Network inference with sampling information

We consider the degree distribution $p(k)$ ($k_{min} \le k \le k_{max}$) of network G, where k_{min} is the minimum degree and k_{max} is the maximum degree. The mean degree $\langle k \rangle$ could be calculated through $\langle k \rangle = \sum_{k=1}^{k_{max}} k \cdot p(k)$. Similarly, assuming that $q(k)$ is the degree distribution of the sampling nodes, the degree distribution $p(k)$ and the average degree $\langle k \rangle$ of the original network G could be estimated by $q(k)$.

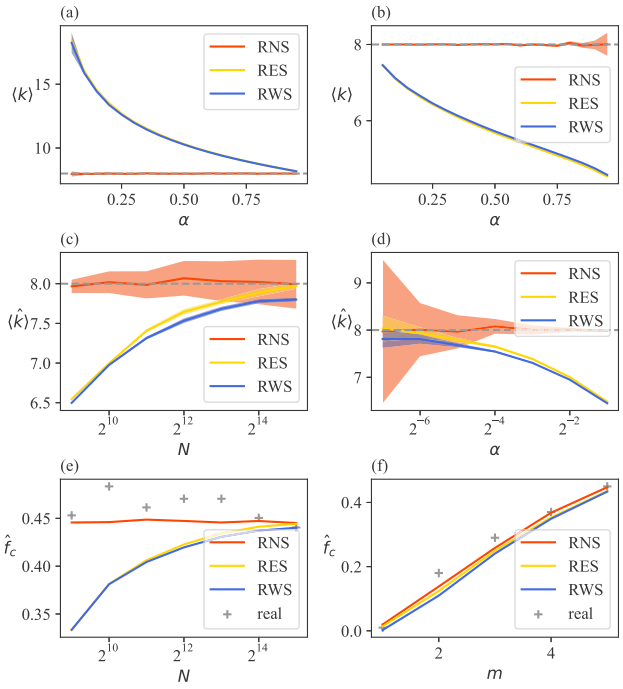

Fig. 3. Estimation of the average degree and the critical removal fraction under the target intervention. (a–d) The average degree of sampling nodes (a) and the missing nodes (b) varies with the sampling proportion α. (c–d) The estimate of the average degree $\langle \hat{k} \rangle$ in BA networks with different sizes N (c) and sampling proportions α (d). The results in (a–d) are presented for three sampling algorithms in the BA network with $\langle k \rangle \approx 8$ (gray dashed lines). (e–f) The estimated critical removal fraction $\hat{f}_c$ under the target intervention in BA networks with different sizes N (e) and m (f). The plus signs represent the real f_c of the BA networks under target intervention. Each line corresponds to an average over 100 independent realizations of the sampling algorithms, and the error bars are displayed as shadows. The numbers of sampling nodes in (c), (e) and (f) are both 500. We generated BA networks with $m = 4$ in (a–e) and $N = 10000$ in (f).

For the RNS, we could approximatively take $p(k) = q(k)$ and obtain an unbiased estimator for $\langle k \rangle$

$$\langle \hat{k} \rangle = \sum_{k=1}^{k_{max}} k \cdot q(k), \tag{1}$$

which is also the average of the degrees of sampling nodes. However, for RES and RWS, the nodes are sampled with a probability proportional to their degree, and the algorithm is biased towards high degree nodes. Thus, $p(k)$ could be estimated as

$$\hat{p}(k) = \frac{\frac{1}{k} \cdot q(k)}{\sum_{k=1}^{k_{max}} \frac{1}{k} \cdot q(k)}. \tag{2}$$

In addition, the average degree could be estimated as

$$\langle \hat{k} \rangle = \sum_{k=1}^{k_{max}} k \cdot \hat{p}(k), \tag{3}$$

which could also be presented by the harmonic mean of the degrees of the sampling nodes [39,40].

Notably, the above conclusion is theoretically valid when the network is infinitely large and degree uncorrelated. For a finite network, the estimated effect would be affected by the network scale N and sampling proportion α. As shown in Fig. 3(a) and (b), we give the average degree $\langle k \rangle$ of sampling nodes and missing nodes with different sampling proportions, respectively. The results show that the $\langle k \rangle$ of sampling nodes in the RES and RWS algorithms decreases from a higher value to the average degree of the network G. From Eqs. (1) and (3), we can obtain the estimate of the average degree through

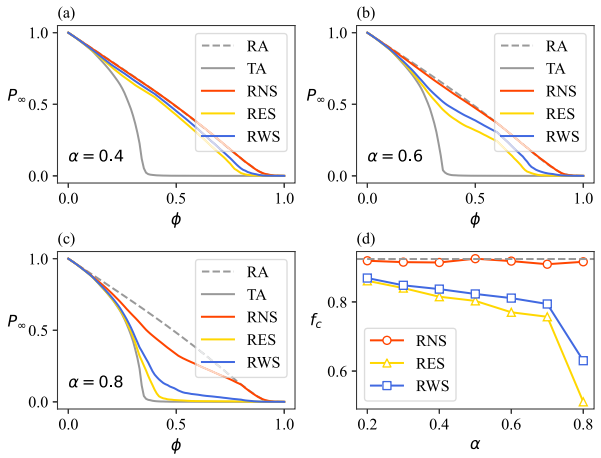

Fig. 4. Network intervention with sampling information. (a–c) The size of largest connected components P_∞ vs. the nodes removal fraction ϕ, where the sampling proportion $\alpha = 0.4$ (a), 0.6 (b) and 0.8 (c), respectively. (d) The critical removal fraction f_c vs. the sampling fraction α. Each point corresponds to an average over 100 independent realizations.

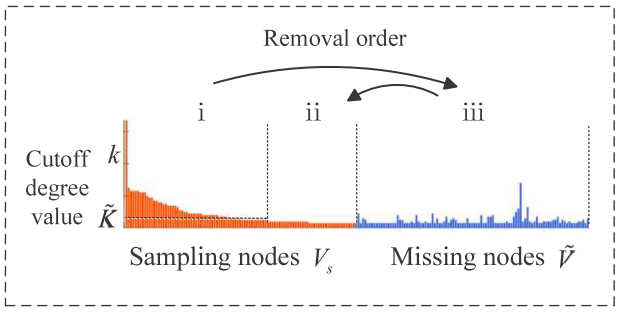

Fig. 5. The illustration of the cutoff degree-based method. The sampling nodes V_s are obtained by RES algorithm with $\alpha = 0.5$. The nodes are divided into three parts by the cutoff degree value $\tilde{K}$ and the purpose is to make the nodes in part *ii* less important than the nodes in part *iii*.

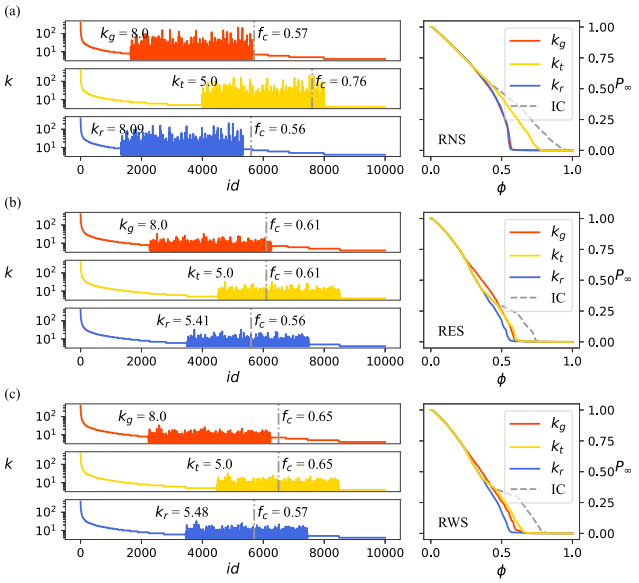

Fig. 6. The cutoff degree-based method for three sampling algorithms. For the networks obtained from RNS (a), RES (b) and RWS (c), we give the nodes' degrees ordered by attack sequences in the cutoff degree-based method with different cutoff degree values and the size of the largest connected components P_∞ varying with nodes removal fraction ϕ. The dash–dot lines in each subgraph on the left present the place of the critical removal fraction of nodes f_c. The dotted lines represent the intervention presenting the network intervention results without cutoff degree-based methods (IC). The different cutoff degree values and their f_c are also given in subgraphs on the left. The sampling proportion $\alpha = 0.6$.

the degree distribution of sampling nodes. In Fig. 3(c) and (d), we give the estimated results with a fixed sampling number in different scale networks and different sampling proportions in the same network. We found that the estimate accuracy in RES and RWS increases with the network scale but decreases with the sampling proportion. On the contrary, the estimated effect in RNS is more robust when there is a larger sampling proportion. Meanwhile, we found that when $N > 2^{13} = 8192$, the sampling number is 500, and $\alpha \approx 2^{-4} = 0.0625$, there are good estimate effects in the three algorithms. Therefore, we take the sampling proportion $\alpha = 0.05$ to estimate the degree distribution and average degree in the following.

The networks in Fig. 3 are generated according to the Barabási–Albert preferential attachment model (BA network) [1], in which the network is grown by attaching new nodes each with m edges that are preferentially attached to existing nodes with a high degree. The average degree of the network $\langle k \rangle \approx 2 \times m$ when N is large. It has been proven that the breakdown or immunization threshold of such networks can be estimated by their average degree $\langle k \rangle$ [41,42]. Similarly, we could estimate the critical removal fraction f_c^a of the BA network ($\kappa = 2$) under the target intervention through sampling information by

$$\hat{f}_c^a = \exp(-8/\langle \hat{k} \rangle). \tag{4}$$

The results in Fig. 3(e) and (f) show that we can better estimate the f_c of BA networks under the target intervention, and the estimation accuracy depends on the estimation accuracy of the average degree $\langle k \rangle$.

3. Efficient network intervention with sampling information

3.1. Network intervention with sampling information

According to the analysis in the above sections, different network sampling algorithms have different preferences for nodes with higher degree values, which would have a large impact on the intervention's effectiveness with sampling information. Therefore, the simplest intervention method is to remove the sampling nodes first in descending order of the known degree and then remove the remaining missing nodes randomly. We name it the incomplete sampling information intervention (IC) strategy. In Fig. 4, we display the intervention experiment's results of the IC strategy with different sampling information obtained from the RNS, RES, and RWS algorithms. Meanwhile, we introduce random intervention (RI) and target intervention (TI) strategies for comparison. RI and TI strategies remove nodes in random order and the descending order of the actual degree and correspond to the situation where the network information is entirely unknown ($\alpha = 0$) and known ($\alpha = 1$), respectively.

An important characteristic of the BA network is that it is vulnerable to target intervention and robust to random intervention. Therefore, the preference of sampling nodes for high degree nodes is conducive to making better intervention decisions. According to Fig. 4, the incomplete information greatly impacts the attack's effect. Especially for RNS, the sampling information is not helpful for reducing the f_c before the proportion arrives at the critical removal fraction of RA, which was fully discussed in Ref. [18]. For RES and RWS, the degree preference of the sampling information makes the sampling nodes contain more hub nodes. Thus, the intervention effect has some improvement. In addition, result of RES is better than that of RWS due to the more independent distribution of sampling nodes.

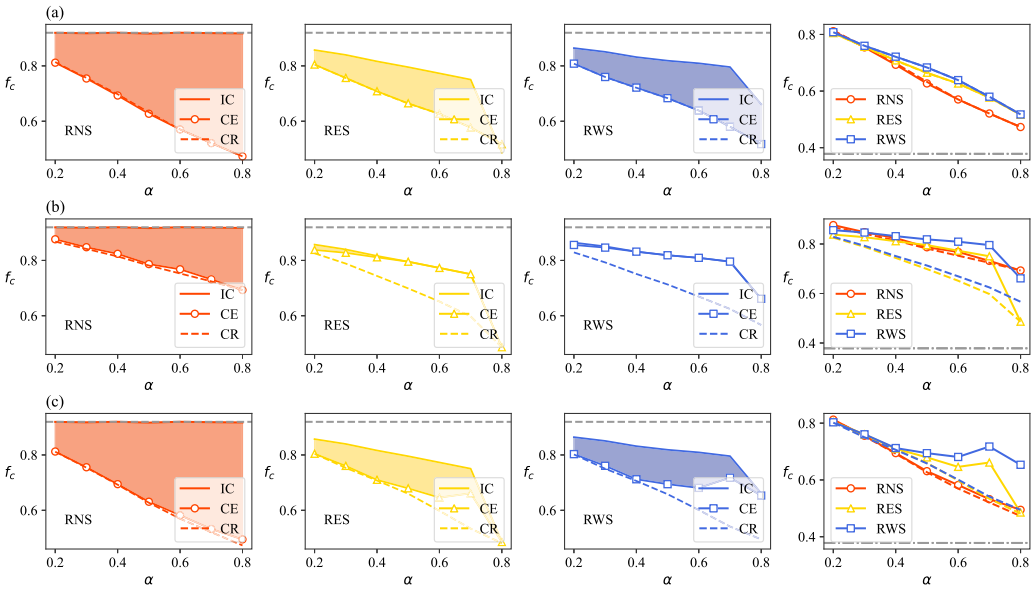

Fig. 7. Intervention experiments in BA networks with different cutoff degree values. We offer the critical removal ratio f_c as a function of the sampling fraction α with RNS, RES and RWS, in which the $\widetilde{K} = \hat{k}_g$ (a), $\hat{k}_t$ (b) and $\hat{k}_r$ (c), respectively. The results of the cutoff degree-based method based on the estimated average degree (CE) are given in different colors and symbols. The solid and dashed lines in the same color represent the results without the cutoff degree (IC) and with the real cutoff degree (CR), respectively. The shadow represents the improvement of CE over IC. The results of CE and CR in the three sampling algorithms are compared together in the fourth subgraph. In addition, the gray dashed and dash–dot lines present the critical removal ratio f_c of networks under random intervention (RI) and target intervention (TI), respectively. Each point or line corresponds to an average of over 100 independent realizations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2. Cutoff degree-based method

Though the sampling algorithms have degree preference, there are still many nodes with small degree values in the sampling nodes. The IC strategy can only adjust the removal sequence of sampling nodes. In the extreme case, when removing all sampling nodes does still not disintegrate the network, the adjustment of the removal order is meaningless. The effect of intervention depends on the degree preference of sampling algorithms. That is why the IC strategy underperformed with a smaller sampling proportion. Sorting all the $N\alpha$ sampling nodes in the decreasing order of degree, there exists a node with degree $\widetilde{K}$ behind which the nodes is less important than the missing nodes.

We name $\widetilde{K}$ as the cutoff degree value and propose a cutoff degree-based method for intervening in the network based on sampling information. As shown in Fig. 5, we remove the sampling nodes V_s and the missing nodes $\widetilde{V}$ in the following order:

(a) We remove the nodes in V_s by the descending order of their degree K_s until reaching the first node whose degree is smaller than $\widetilde{K}$ (nodes in part *i*).

(b) We remove the nodes in $\widetilde{V}$ (nodes in part *iii*) randomly.

(c) We remove the rest of the nodes in V_s (nodes in part *ii*) by the decreasing degree order.

For the networks whose scale is larger than 1000, we similarly remove 1% of the total number of nodes at a time and determine whether the degree of a boundary node is smaller than $\widetilde{K}$.

In this article, we consider three different cutoff degree values, i.e., the global average degree k_g, the critical degree of the network under target intervention k_t and the average degree of the remaining nodes k_r. Because the original network's average degree can be estimated through the degree distribution of the sampling nodes, the above cutoff degree values can be estimated by the sampling nodes.

Based on the estimated global average degree $\langle \hat{k} \rangle$ from Eqs. (1) or (3), we obtain

$$\hat{k}_g = \langle \hat{k} \rangle. \tag{5}$$

Meanwhile, in the BA networks, we further combine Eq. (4) and obtain

$$\hat{k}_t = max\{k | \hat{p}_c(k) \le 1 - \hat{f}_c\}, \tag{6}$$

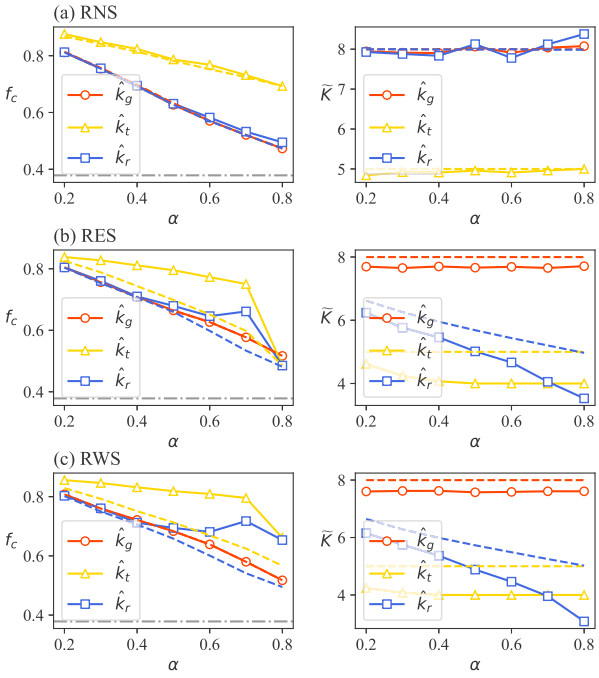

Fig. 8. Different cutoff degree values in three sampling algorithms. With the sampling information obtained by RNS (a), RES (b) and RWS (c), we give the critical removal fraction f_c and the estimated $\widetilde{K}$ used in the cutoff degree-based method. The dashed lines in the same color represent the results with the real cutoff degree (CR). The results are obtained from the same experiments in Fig. 7. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where $\hat{p}_c(k)$ is the cumulative degree distribution obtained from $\hat{p}(k)$. Moreover, introducing the average degree of sampling nodes $\langle k \rangle_s$ and the sampling proportion α, we can obtain

$$\hat{k}_r = \frac{\langle \hat{k} \rangle N - \langle \hat{k} \rangle_s N\alpha}{N(1-\alpha)} = \frac{\langle \hat{k} \rangle - \langle k \rangle_s \alpha}{(1-\alpha)}. \tag{7}$$

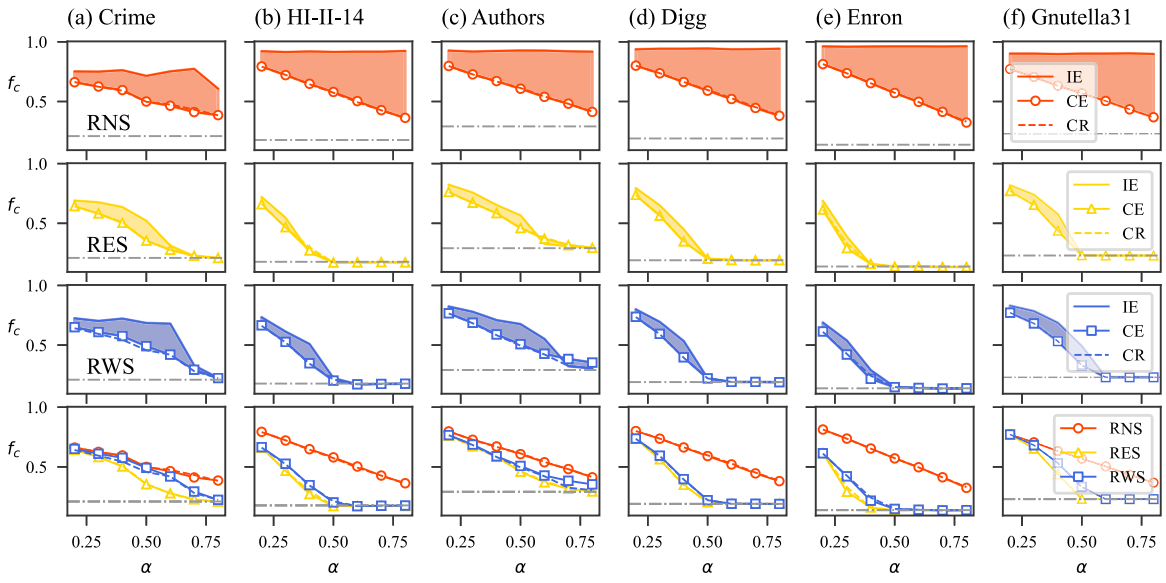

Fig. 9. Intervention experiments in real networks with $\widetilde{K} = \hat{k}_g$. For each network in Table 1, we show the critical removal ratio f_c as a function of the sampling fraction α and compare CE and CR in three sampling algorithms. The definitions of the symbols and colors are the same as in Fig. 7. In addition, the gray dash–dot lines represent the critical removal ratio f_c of networks under target intervention (TI). Each point or line corresponds to an average of 10 independent realizations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

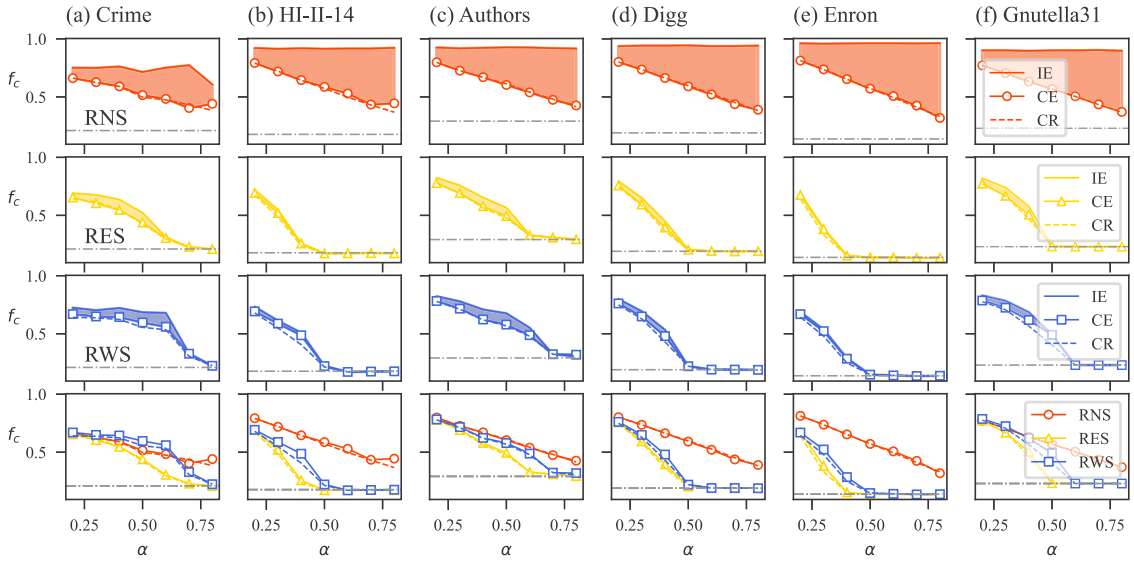

Fig. 10. Intervention experiments in real networks with $\widetilde{K} = \hat{k}_r$. The other settings are the same as those in Fig. 9.

To further explore the effectiveness of the above three cutoff degree values, for the BA network used in Fig. 4, we show the degree of nodes according to the attack order with different $\widetilde{K}$ in Fig. 6 and the change in the size of the largest component with different remove orders. To exclude the influence of estimate error, the cutoff degree values are both calculated by the real average degree distribution $p(k)$ and the average degree $\langle k \rangle$ in this figure. To better compare the differences in cutoff degree values, we keep the random order of nodes in $\widetilde{V}$ unchanging but only change the place of the cutoff degree values in the experiments.

In Fig. 6 we can see that the cutoff degree-based method could effectively order the nodes with high degrees in $\widetilde{K}$ before the nodes with lower degrees in V_s and then improve the attack effect. At the same time, the effect of k_g and k_r is significantly better than that of k_t.

Furthermore, in the same BA network, we change the sampling proportion α and compare the critical removal ratio f_c of applying the cutoff degree-based method or not in three sampling algorithms. For each sampling algorithm, we remove the nodes with orders obtained from incomplete information (IC) as well as the cutoff degree-based methods based on the estimated degree distribution (CE) and real degree distribution (CR). At last, we compare the results of the cutoff degree-based methods in three sampling algorithms with different cutoff degree values. As can be seen in Fig. 7, the cutoff degree-based method can effectively improve the attack effect in the BA network with different sampling proportions α. However, in RES and RWS, the method fails to achieve the most effective effect due to the poor estimation effect of the cutoff degree values $\hat{k}_t$ and $\hat{k}_r$.

To better compare the difference in cutoff degree values, in Fig. 8, we present the intervention results with different $\widetilde{K}$ as well as the change in different cutoff degree values with the sampling proportion α. The results show that, in RNS, the cutoff degree values calculated by sampling information (estimated values) are similar to those of real

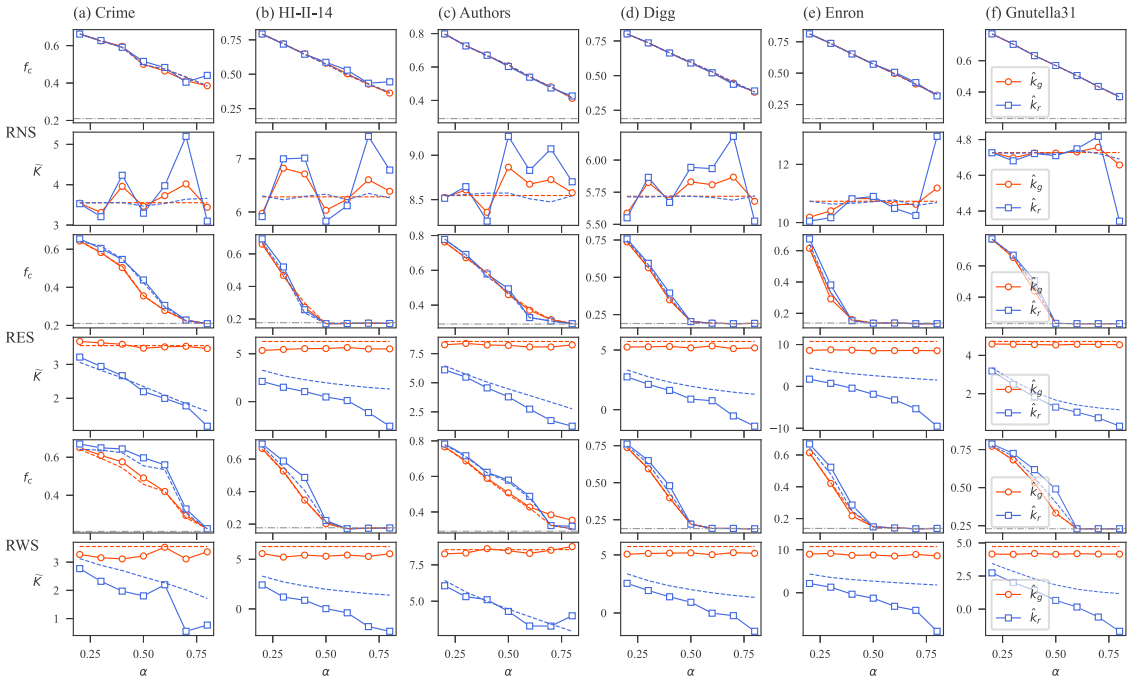

Fig. 11. Comparison of cutoff degree values $\hat{k}_g$ and $\hat{k}_r$ in real networks. For each network in Table 1, we give the critical removal fraction f_c and the estimated $\widetilde{K}$ used in the cutoff degree-based method. The dashed lines in the same color represent the results with the real cutoff degree (CR). The results are obtained from the same experiments in Figs. 9 and 10. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

information (real values). In RES and RWS, the estimated values are significantly lower than the real values, especially in $\hat{k}_t$ and $\hat{k}_r$. k_g and k_r have better effects than k_t when the real degree distribution is used to calculate the cutoff degree value, which is similar to the conclusion in Fig. 6. However, compared with k_g and $\hat{k}_g$, there is a larger error between k_r and $\hat{k}_r$. This makes the estimated cutoff degree values less effective than the true value, especially when the sampling proportion α is high. The above results also show that the k_t based on the estimated target intervention threshold is not an effective cutoff degree value and the threshold of networks with other degree distributions is more difficult to estimate accurately through k_t. Therefore, k_g and k_r are mainly used for cutoff degree-based methods in what follows.

3.3. Experiments in real networks

In the previous sections, we mainly experimented in BA networks. Due to the scale-free degree distribution, BA networks are robust against random intervention but vulnerable to target intervention, which makes it easy to reflect the differences of attack ranking in the intervention effect. On the other hand, many systems taking the form of networks in the real world are more complicated than synthetic networks. To further analyze the difference of sampling algorithms and the selection of $\widetilde{K}$ in the cutoff degree-based method, we experiment in more real networks in this section. The information on the real networks is shown in Table 1. Specifically, we present the average clustering coefficient and assortativity coefficient [43]. The data can be obtained from the website SNAP [44].

Similar to Fig. 7, we give the results of network intervention in real networks for $\widetilde{K} = \hat{k}_g$ and $\hat{k}_r$ in Figs. 9 and 10, respectively. We can see that, in most real networks, the effect of intervention under partial sampling information obtained by RES and RWS can reach that of target interventions (dash–dot lines) before the sampling proportion α arrives at a high value.

Meanwhile, the cutoff degree-based method can also effectively take advantage of sampling information to improve the intervention effect. Unlike the error caused by the estimated average degree in the BA network, in most real networks, the results obtained by the estimated average degree (CE) are close to the results obtained by the real average degree (CR). Especially for RES and RWS, the attack effect is mainly determined by the degree distribution of the sampling nodes, and the cutoff degree-based method has little effect. That is because the main "hub" nodes have been sampled due to the algorithms' degree preference. Moreover, the results of RES are also better than those of RWS due to the independent distribution of the sampling nodes, which is similar to the results in Fig. 4.

Table 1
The characteristics of the real networks analyzed in this paper.

Name	N	M	$\langle k \rangle$	C	AC
Crime	829	1473	3.554	0.006	−0.165
HI-II-14	4165	13087	6.284	0.044	−0.202
Authors	21363	91286	8.546	0.642	0.125
Digg	29652	84781	5.718	0.005	0.003
Enron	33696	180811	10.732	0.509	−0.116
Gnutella31	62561	147878	4.727	0.005	−0.093

For each network, we show its name, the number of nodes (N) and edges (M), the average degree, the average clustering coefficient C, and the assortativity coefficient AC.

To further analyze the influence of cutoff degree values, we compare the results of cutoff degree-methods for $\widetilde{K} = \hat{k}_g$ and $\hat{k}_r$ in Fig. 11 and give the variation in estimated cutoff degree values with sampling proportion α. As opposed to the results in BA networks, in most real networks, the effect of real values k_g is better than that of k_r. At the same time, the estimated error of $\hat{k}_r$ is larger than that of $\hat{k}_g$. Therefore, in most cases, it is better to use $\hat{k}_g$ as the cutoff degree value in network intervention.

4. Conclusion and discussion

Network intervention aims to destroy network function, where incomplete network information will greatly affect the effect. Considering the access to network information, we study network intervention with sampling information in this paper. Based on three classical sampling algorithms, i.e., random node sampling, random edge sampling and

random walking sampling, we analyzed the degree preferences of algorithms and their influence on intervention according to incomplete information. Furthermore, we propose the cutoff degree-based method, which could effectively use the degree distribution of sampling nodes and improve the intervention effect by estimating the average degree of the original network. The selection of different cutoff degree values and their effects are discussed in models and natural network experiments.

As opposed to the other network intervention research which mainly focuses on random information loss, we consider how the network information is obtained in this paper. We found that the degree distribution of known nodes determines the intervention effect. On the other hand, we synthesize the global sampling information and avoid the repeated acquisition of nodes' information. The results indicate that we could obtain the approximate target intervention effect by partial sampling information when the appropriate cutoff-degree value is estimated through the sampling information.

It is worth mentioning that we make some ideal assumptions, such as the known neighborhood and real degree of nodes. Whereas, in this paper, we did not consider the network topology between sampling nodes during the process of developing the intervention strategy. More scenarios of incomplete sampling information could be expanded in the future. In addition, most of the results in this paper are obtained from the simulation experiments in representative networks. We could also analyze the intervention effect in different sampling algorithms based on percolation theory for networks with specific degree distributions.

CRediT authorship contribution statement

Qi Mingze: Conceptualization, Writing – original draft, Methodology, Software, Visualization. **Tan Suoyi:** Writing – review & editing, Software, Visualization. **Chen Peng:** Methodology, Visualization. **Duan Xiaojun:** Conceptualization, Methodology, Writing – review & editing. **Lu Xin:** Conceptualization, Methodology, Writing – review.

Declaration of competing interest

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

Data availability

Data will be made available on request.

References

[1] Barabasi AL, Albert R. Emergence of scaling in random networks. Science 1999;286(5439):509–12.
[2] Newman MEJ. The structure and function of complex networks. SIAM Rev 2003;45(2):167–256.
[3] Coscia M. The atlas for the aspiring network scientist. 2021, arXiv preprint arXiv:2101.00863.
[4] Holme P, Kim BJ, Yoon CN, Han SK. Attack vulnerability of complex networks. Phys Rev E 2002;65(5).
[5] Freitas S, Yang D, Kumar S, Tong H, Chau DH. Graph vulnerability and robustness: A survey. IEEE Trans Knowl Data Eng 2022;1.
[6] Morone F, Makse HA. Influence maximization in complex networks through optimal percolation. Nature 2015;524(7563). 65–U122.
[7] Clusella P, Grassberger P, Pérez-Reche FJ, Politi A. Immunization and targeted destruction of networks using explosive percolation. Phys Rev Lett 2016;117(20). 0031-9007 1079-7114.
[8] Pósfai M, Braun N, Beisner BA, McCowan B, D'Souza RM. Consensus ranking for multi-objective interventions in multiplex networks. New J Phys 2019;21(5).
[9] Bak-Coleman JB, Kennedy I, Wack M, Beers A, Schafer JS, Spiro ES, Starbird K, West JD. Combining interventions to reduce the spread of viral misinformation. Nat Hum Behav 2022.
[10] Chen Y, Paul G, Havlin S, Liljeros F, Stanley HE. Finding a better immunization strategy. Phys Rev Lett 2008;101(5).
[11] Braunstein A, Dall'Asta L, Semerjian G, Zdeborova L. Network dismantling. Proc Natl Acad Sci 2016;113(44):12368–73, 1091-6490 (Electronic) 0027-8424 (Linking).
[12] Qi M, Deng Y, Deng H, Wu J. Optimal disintegration strategy in multiplex networks. Chaos 2018;28(12).
[13] Lu LY, Chen DB, Ren XL, Zhang QM, Zhang YC, Zhou T. Vital nodes identification in complex networks. Phys Rep 2016;650:1–63.
[14] Zdeborová L, Zhang P, Zhou H-J. Fast and simple decycling and dismantling of networks. Sci Rep 2016;6.
[15] Fan C, Zeng L, Sun Y, Liu Y-Y. Finding key players in complex networks through deep reinforcement learning. Nat Mach Intell 2020;2(6):317–24.
[16] Grassia M, De Domenico M, Mangioni G. Machine learning dismantling and early-warning signals of disintegration in complex systems. Nature Commun 2021;12(1):5190, 2041-1723 (Electronic) 2041-1723 (Linking).
[17] Chen P, Qi M, Lu X, Duan X, Kurths J. Efficient network immunization strategy based on generalized Herfindahl–Hirschman index. New J Phys 2021;23(6).
[18] Wu J, Deng HZ, Tan YJ, Zhu DZ. Vulnerability of complex networks under intentional attack with incomplete information. J Phys A 2007;40(11):2665–71, 1751-8113 1751-8121.
[19] Shang Y. Subgraph robustness of complex networks under attacks. IEEE Trans Syst Man Cybern 2019;49(4):821–32.
[20] Rosenblatt SF, Smith JA, Gauthier GR, Hebert-Dufresne L. Immunization strategies in networks with missing data. PLoS Comput Biol 2020;16(7):e1007897, 1553-7358 (Electronic) 1553-734X (Linking).
[21] Borgatti SP, Carley KM, Krackhardt D. On the robustness of centrality measures under conditions of imperfect data. Social Networks 2006;28(2):124–36.
[22] Gallos LK, Cohen R, Argyrakis P, Bunde A, Havlin S. Stability and topology of scale-free networks under attack and defense strategies. Phys Rev Lett 2005;94(18).
[23] Li J, Wu J, Li Y, Deng H-Z, Tan Y-J. Optimal attack strategy in random scale-free networks based on incomplete information. Chin Phys Lett 2011;28(6). 0256-307X 1741-3540.
[24] Tan S-Y, Wu J, Lü L, Li M-J, Lu X. Efficient network disintegration under incomplete information: the comic effect of link prediction. Sci Rep 2016;6(1).
[25] Cohen R, Havlin S, ben-Avraham D. Efficient immunization strategies for computer networks and populations. Phys Rev Lett 2003;91(24).
[26] Chen S, Lu X. An immunization strategy for hidden populations. Sci Rep 2017;7(1):3268, 2045-2322 (Electronic) 2045-2322 (Linking).
[27] Yang YR, McKhann A, Chen SX, Harling G, Onnela JP. Efficient vaccination strategies for epidemic control using network information. Epidemics 2019;27:115–22.
[28] Liu Y, Sanhedrai H, Dong G, Shekhtman LM, Wang F, Buldyrev SV, Havlin S. Efficient network immunization under limited knowledge. Natl Sci Rev 2021;8(1).
[29] Shang Y. Immunization of networks with limited knowledge and temporary immunity. Chaos 2021;31(5).
[30] Shang Y. Generalized k-cores of networks under attack with limited knowledge. Chaos Solitons Fractals 2021;152.
[31] Lu X, Bengtsson L, Britton T, Camitz M, Kim BJ, Thorson A, Liljeros F. The sensitivity of respondent-driven sampling. J Roy Statist Soc Ser A 2012;175:191–216.
[32] Papagelis M, Das G, Koudas N. Sampling online social networks. IEEE Trans Knowl Data Eng 2013;25(3):662–76.
[33] Smith JA, Moody J, Morgan J. Network sampling coverage II: The effect of non-random missing data on network measurement. Social Networks 2017;48:78–99, 0378-8733 (Print) 0378-8733 (Linking).
[34] Chen SR, Lu X, Liu Z, Jia ZW. Inferring the population mean with second-order information in online social networks. Entropy 2018;20(6).
[35] Smith JA, Morgan JH, Moody J. Network sampling coverage III: Imputation of missing network data under different network and missing data conditions. Social Networks 2022;68:148–78, 0378-8733 (Print) 0378-8733 (Linking).
[36] Ahmed N, Neville J, Kompella R. Network sampling via edge-based node selection with graph induction. 2011.
[37] Ahmed NK, Neville J, Kompella R. Network sampling: From static to streaming graphs. ACM Trans Knowl Discov Data 2014;8(2).
[38] Chen SR, Lu X, Liu Z, Jia ZW. Sampling on bipartite networks: a comparative analysis of eight crawling methods. J Stat Mech Theory Exp 2018.
[39] Hu P, Lau WC. A survey and taxonomy of graph sampling. Comput Sci 2013.
[40] Salganik MJ, Heckathorn DD. Sampling and estimation in hidden populations using respondent-driven sampling. Sociol Methodol 2004;34(1):193–240.
[41] Cohen R, Erez K, Ben-Avraham D, Havlin S. Breakdown of the internet under intentional attack. Phys Rev Lett 2001;86(16):3682–5.
[42] Pastor-Satorras R, Vespignani A. Immunization of complex networks. Phys Rev E 2002;65(3).
[43] Newman MEJ. Mixing patterns in networks. Phys Rev E 2003;67(2).
[44] Leskovec J, Krevl A. SNAP datasets: Stanford large network dataset collection. 2014, http://snap.stanford.edu/data.