Subgraph Robustness of Complex Networks under Attacks

Yilun Shang

***Abstract*—Network measures derived from empirical observations are often poor estimators of the true structure of system as it is impossible to observe all components and all interactions in many real world complex systems. Here, we study attack robustness of complex networks with data missing caused by (i) a uniform random sampling and (ii) a non-uniform random sampling. By introducing the subgraph robustness problem, we develop analytically a framework to investigate robustness properties of the two types of subgraphs under random attacks, localized attacks, and targeted attacks. Interestingly, we find that the benchmark models such as Erdős-Rényi graphs, random regular networks, and scale-free networks possess distinct characteristic subgraph robustness features. We show that the network robustness depends on several factors including network topology, attack mode, sampling method and the amount of data missing, generalizing some well-known robustness principles of complex networks. Our results offer insight into the structural effect of missing data in networks and highlight the significance of understanding different sampling processes and their consequences on attack robustness, which may be instrumental in designing robust systems.**

***Index Terms*—Complex networks, complex systems, sampling, attack robustness.**

I. Introduction

COMPLEX networks, such as social networks, the World Wide Web, and gene regulatory networks, provide a compact and powerful representation of the interaction structure of a wide range of complex systems, where nodes represent entities (e.g., people, web sites, genes) and edges represent some type of connections (e.g., friendship, communication, regulation) [1], [2]. To study the networks one needs to first collect reliable large scale network data. Even with the emergence of the Internet, social media, and high-throughput gene expression analysis, in most cases data collected for complex networks are incomplete with nodes and edges missing. In social network analysis, this is often due to the so-called boundary effects or respondent inaccuracy in network surveys [3]–[5]. For example, networks arising from the popular social network platforms are not completely mapped because of the boundary effects; namely, there are people who do not actually use the social networking service except setting up an account (so-called "zombie accounts") and so we cannot observe their connections. Anonymous purchase in online shopping sites also induces a similar boundary effect. Likewise, respondents may be absent on the day of survey administration or have opted-out of the survey for privacy reasons, leading to unobserved nodes in the network. In other empirical studies, having access to all the nodes may be virtually impossible due to huge network size or limited resources. All these situations give rise to a sampling of the network nodes, i.e., a partially observed subgraph of the network [6], [7].

Networks with incomplete or missing data have been probed mainly in two directions of research. Broadly speaking, one line of work has focused on prediction/inference of missing edges or nodes with a view to determine the full network structure, which is a common requirement in many graph-mining tasks, such as community detection, belief propagation, and influence maximization, etc. The edge oriented version is commonly known as the link prediction problem [8]–[10], which has various applications ranging from recommender systems to computational biology; see the survey [11]. The node oriented version, referred to as missing node identification problem, has been studied recently in [12]–[15]. Important applications in the security community, for example, include identification of missing person in a family tree or people wanted by the police as suspects in a crime. A node-based incident prediction approach is proposed in [16], which has applications in industrial control systems. This problem is significantly more difficult than the link prediction problem as neither the nodes nor their edges are known with certainty [17].

On the other hand, a different vibrant line of work concerning missing data deals with structural effects of missing edges or nodes on varied measurable properties of networks, such as degree distribution [18], [19], average degree, average path length, assortativity, clustering coefficient [4], [20], centrality [7], [21], community structure [22], and the number of small fixed subgraphs [23], to name just a few. These works basically address the question "what happens to network measures when some edges or nodes are missing?" based upon Monte-Carlo simulations. Various sampling procedures, such as uniform random sampling [4], [7], snowball sampling [24], respondent driven sampling [3], [25], and random walks [22], [23], have also been developed to generate the partially observed subnetwork. For example, the work [18] warns that randomly sampled subnetworks of scale-free networks no longer show scale-free properties. In general, however, the effect of induced bias and how does it correlate with different network topologies or levels of missing data are often not fully understood.

In this paper, we follow the second line of research and focus on another important network property, namely, the

Y. Shang is with the School of Mathematical Sciences, Tongji University, Shanghai 200092, China (e-mail: shylmath@hotmail.com).

Manuscript received April xx, 2017; revised June xx, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 11505127 and in part by the Shanghai Pujiang Program under Grant 15PJ1408300.

network robustness against attacks [26]–[29]. As is known, the function and stability of networks rely crucially on the interconnections between nodes in which failed nodes will disable others connecting through them to the network and may destroy or cripple the entire network. The vast majority of previous work on network robustness assumes a complete network, namely, all nodes and edges in the network are observed; see, e.g. [1], [26], [30]–[33]. Motivated by the above consideration, the goal of this paper is to investigate analytically and by simulations attack robustness on networks with missing nodes and edges, which we refer to as the subgraph robustness problem.

We mention that the question addressed here is related in concept to some previous works on attack robustness with incomplete information. The work [34] examines the optimal attack strategy on scale-free networks, in which a fixed portion of nodes are unobserved. Similarly, efficient attack strategies with missing edges are examined in [35] with the aid of link prediction techniques. Attack robustness of networks with uncertain or local knowledge has been investigated by some researchers; see e.g. [36]–[38]. In these works, all edges and nodes in the network are present, while only some information (such as the node degree) is not fully/precisely known.

II. Model description and methodologies

In studying the subgraph robustness problem, we formally consider two types of sampling by sampling the nodes either uniformly at random, leading to a uniform subgraph (US) or in a non-uniform manner, leading to a non-uniform subgraph (NS). Using percolation theory [1], we investigate the robustness of US and NS under different attacks in terms of the relative size of giant component and the critical percolation threshold at which the giant component first collapses. The network attacks considered here include:

- Random attack (RA), where randomly chosen nodes are removed from the network, meaning that each node in the network is attacked with equal probability. RA describes random errors, system decay, or attacks without prior knowledge of the network architecture; see e.g. [1], [26], [32], [33], [39].
- Localized attack (LA), where nodes surrounding a seed node are removed layer by layer, causing aggregated damage of adjacent components limited to a specific area. LA can be caused by natural disasters such as earthquakes and floods, as well as mass attacks including bomb blasts and malware infection; see e.g. [40]–[43].
- Targeted attack (TA), where nodes with a higher degree are more vulnerable, meaning that nodes are attacked in decreasing order of their connectivity. TA captures sabotage on the Internet and some malicious attacks against transportation hubs, important power stations, etc.; see e.g. [26], [32], [44], [45].

We apply our derived theoretical frameworks to three types of network models including Erdős-Rényi (ER) networks [46] with a Poisson degree distribution, random regular (RR) networks following a degenerated degree distribution, and scale-free (SF) networks [26], [47] characterized by a power-law degree distribution. Formally, consider a random network captured by an arbitrary degree distribution $P(k)$, which is the probability that a randomly chosen node has k neighbors. The generating function of the degree distribution is defined as $G_0(x) = \sum_{k=0}^{\infty} P(k)x^k$ [1], [48]. Here, we are interested in networks with missing data generated in the following two types of sampling processes (see Fig. 1):

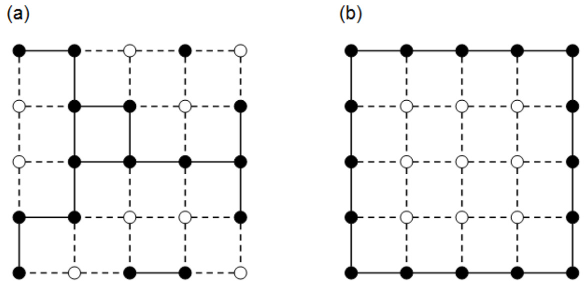

Fig. 1. Schematic illustration of (a) US and (b) NS on square lattices. Solid nodes and lines represent the observed subgraphs.

- Uniform random sampling, meaning that a fraction q of nodes deployed uniformly at random in the network are observed. This is a natural setting commonly used in other work; see e.g. [4], [6], [19]. The induced subgraph on the observed nodes is said to be the uniform subgraph (US). Namely, US is constructed by removing a fraction $1 - q$ of unobserved nodes as well as their contributing edges;
- Non-uniform random sampling, where a fraction $1 - q$ of nodes sitting in some multi-hop neighborhood of a random selected node cannot be observed. In other words, the observed subgraph, referred to as the non-uniform subgraph (NS), is obtained by removing a seed node, its nearest neighbors, its second nearest neighbors and so on until a fraction $1 - q$ of nodes in the entire network are removed. This situation reflects networks suffering from a single-source spreading data contamination [5], [49] or a diffusive non-respondent bias in network surveys [3], [21], which have been studied extensively in social network analysis.

As we have mentioned, there are multiple ways in which data missing can be biased. The non-uniform random sampling considered here not only provides a comparison for the typical random sampling but is amenable to analytical treatment for all attack strategies we are interested in this paper. We assume that attack is launched against the observed subgraph, i.e., US and NS, until a fraction $1 - p$ of nodes in the subgraph are attacked. A major characteristic of network functionality is the relative size of the giant component, denoted by P_∞, consisting of all remaining nodes that survive the attack. The critical threshold at which the giant component first collapses, i.e. $P_\infty \sim 0$, is denoted by p_c. Evidently, when $q = 1$, we are reduced to the usual percolation settings where all nodes in the network are observed [1]. We will focus on the effects of network topologies (ER, RR, SF), attack strategies (RA, LA, TA), amount of data missing (q), and sampling methods (US, NS) on the two measures p_c and P_∞.

Obviously, irrespective of the network topology, a given subnetwork is more vulnerable than the complete network in the sense that a subnetwork always collapses prior to the complete network. Hence, interpreting the attack robustness results when only a subgraph is observed, as is often the case in the real world, could make us, on one hand, over-optimistic in the situations of beneficial attacks such as the regulation of cancer stem cells or containing pandemic diseases, while on the other hand, over-pessimistic when the attacks are malicious, e.g., sabotage on the Internet and damage on the infrastructures. Interestingly, we find that such illusion could be significant in some situations while negligible in others, depending on the interplay between the network topology, the attack mode, the sampling method, as well as the amount of data missing. Our extensive simulations are in good agreement with analytical calculations. Simulation results are based on synthetic network models with $N = 10^6$ nodes and averages over 100 realizations. In addition to the model networks, simulations on real-world networks, including social, technical, biological and infrastructural ones are also performed.

It is worth noting that, in addition to p_c and P_∞ considered here, there have been a number of other robustness metrics reported in the literature, but mostly based upon these two measures as well as shortest path lengths in the networks; see e.g. [27], [50], [51]. As such, these measures are computationally more involved and not analytically tractable in general.

III. Theoretical framework on subgraph robustness

In this section, we perform the analytical study on subgraph robustness under three types of attacks, RA, LA, and TA, respectively. We mention that the equations derived below for finding robustness in RA, LA, and TA have more or less been studied in previous works, particularly, [32], [40], [44], in the case of fully observed networks (i.e., $q = 1$). We show below how these techniques can be adapted in networks with any amount of data missing to paint a larger picture in the framework of subgraph robustness. The results reveal non-trivial phenomena which have not been observed in an entire network with given degree distribution (see Section IV for details).

A. Subgraph robustness under RA

We begin with the robustness of US under RA. In a random attack, each node of US is occupied, i.e., remains intact, with probability p. Therefore, RA launched on US is equivalent to the classical node percolation on the entire network with occupation probability pq [1]. Recall that the generating function of the degree distribution is $G_0(x) = \sum_{k=0}^{\infty} P(k)x^k$. The generating function $H_1(x)$ of the size distribution of the clusters that can be reached following a randomly chosen edge satisfies a self-consistency equation $H_1(x) = 1 - pq + pqxG_1(H_1(x))$, where $G_1(x) = G_0'(x)/G_0'(1)$ [32]. Likewise, the generating function for the size of the cluster to which a randomly chosen node belongs is generated by $H_0(x) = 1 - pq + pqxG_0(H_1(x))$. Therefore, the mean size of small clusters is

$$H_0'(1) = pq\left[1 + \frac{pqG_0'(1)}{1 - pqG_1'(1)}\right], \tag{1}$$

which diverges when $1 = pqG_1'(1)$ marking the critical value p_c at which the giant component collapses. Noting that $q > 1/G_1'(1)$ guarantees the existence of a giant component in US, we have

$$p_c(\mathrm{RA}) = \min\left\{1, \frac{1}{qG_1'(1)}\right\}. \tag{2}$$

The fraction of the giant component in the original network, denoted by $S(\mathrm{RA})$, is given by

$$S(\mathrm{RA}) = 1 - H_0(1) = pq[1 - G_0(u)], \tag{3}$$

where $u = H_1(1)$ satisfies $u = 1 - pq + pqG_1(u)$. We define P_∞ as the relative size of the giant component as a fraction of the entire network. By definition, we have $P_\infty(\mathrm{RA}) = S(\mathrm{RA})$. Clearly, when all nodes are observable, i.e., $q = 1$, Eqs. (2) and (3) reduce to the usual site percolation framework [32].

Next, we turn to the robustness of NS under RA. A key observation here is that the non-uniform sampling can be described by the so-called localized attack procedure, where nodes are attacked shell by shell from a random root node until a certain fraction of nodes are removed [40], [52]. Following [40], the generating function of the degree distribution of NS becomes

$$\hat{G}_0(x) = \frac{1}{G_0(f)}G_0\left(f + \frac{G_0'(f)}{G_0'(1)}(x-1)\right), \tag{4}$$

where $f = G_0^{-1}(q)$. Let $\hat{G}_1(x) = \hat{G}_0'(x)/\hat{G}_0'(1)$. By defining the two generating functions $\hat{H}_0(x)$ and $\hat{H}_1(x)$ for the size distributions of the clusters similarly, and following the above site percolation procedure with occupation probability p, we are led to the critical equation which determines the break-up point of the giant component $1 = p\hat{G}_1'(1)$. Hence, by using (4) we obtain

$$p_c(\mathrm{RA}) = \min\left\{1, \frac{G_0'(1)}{G_0''(f)}\right\}, \tag{5}$$

where again $f = G_0^{-1}(q)$. Note that when $\hat{G}_1'(1) = 1$, namely, q satisfies $G_0'(1) = G_0''(f)$, we have $p_c = 1$, which is precisely the time when a giant component first forms in NS [40].

The fraction of the giant component in NS, denoted by $S(\mathrm{RA})$, is given by

$$S(\mathrm{RA}) = 1 - \hat{H}_0(1) = p[1 - \hat{G}_0(u)], \tag{6}$$

where $u = \hat{H}_1(1)$ satisfies $u = 1 - p + p\hat{G}_1(u)$. By definition, we have $P_\infty(\mathrm{RA}) = qS(\mathrm{RA})$. Note that when $q = 1$, i.e., all nodes are observed, we readily reproduce the framework in [40] since $\hat{G}_0(x) = G_0(x)$.

B. Subgraph robustness under LA

In this section, we investigate another popular type of attack, LA, which is first introduced in [40] and further developed by some other researchers; see e.g., [41]–[43].

First, we consider the robustness of US under LA. US is a random subgraph obtained by occupying each node with probability q in the original network. Following the method introduced in [32], [44], we find the generating function for the degree distribution of US to be

$$\tilde{G}_0(x) = \sum_{k=0}^{\infty} P(k)(1-q+qx)^k. \tag{7}$$

We then perform LA on US until a fraction $1-p$ of the nodes are removed. The generating function of the degree distribution of the remaining nodes in US can be derived as [40]

$$\tilde{G}_{0,p}(x) = \frac{1}{\tilde{G}_0(g)} \tilde{G}_0\left(g + \frac{\tilde{G}_0'(g)}{\tilde{G}_0'(1)}(x-1)\right), \tag{8}$$

where $g = \tilde{G}_0^{-1}(p)$. Let $\tilde{G}_{1,p}(x) = \tilde{G}_{0,p}'(x)/\tilde{G}_{0,p}'(1)$. By combining (7), (8) and the criterion for the network to collapse, $\tilde{G}_{1,p}'(1) = 1$ [1], [32], we find that

$$p_c(\mathrm{LA}) = \min\{1, \tilde{p}_c\}, \tag{9}$$

where $\tilde{p}_c$ satisfies $\tilde{G}_0''(\tilde{G}_0^{-1}(p)) = \tilde{G}_0'(1)$. Note that when $q = 1/G_1'(1)$, we have $\tilde{G}_0''(1) = \tilde{G}_0'(1)$ and hence $\tilde{p}_c = 1$, which is precisely the time when a giant component first forms in US.

The fraction of the giant component in US can be expressed by

$$S(\mathrm{LA}) = 1 - \tilde{G}_{0,p}(u), \tag{10}$$

where $u = \tilde{G}_{1,p}(u)$ [48]. By definition, we have $P_\infty(\mathrm{LA}) = pqS(\mathrm{LA})$. Clearly, when $q = 1$, we reproduce the framework in [40] since $\tilde{G}_0(x) = G_0(x)$ by (7).

Next, we study the robustness of NS under LA. The degree distribution of NS is generated by (4). We now perform LA on NS until a fraction $1-p$ of the nodes are removed. As in the above case of US, the generating function of the degree distribution of the remaining nodes in NS is shown to be given by

$$\hat{G}_{0,p}(x) = \frac{1}{\hat{G}_0(h)} \hat{G}_0\left(h + \frac{\hat{G}_0'(h)}{\hat{G}_0'(1)}(x-1)\right), \tag{11}$$

where $h = \hat{G}_0^{-1}(p)$. Define $\hat{G}_{1,p}(x) = \hat{G}_{0,p}'(x)/\hat{G}_{0,p}'(1)$. Combining (4), (11) and the criterion for the network to collapse, $\hat{G}_{1,p}'(1) = 1$ [32], we find that

$$p_c(\mathrm{LA}) = \min\{1, \hat{p}_c\}, \tag{12}$$

where $\hat{p}_c$ satisfies $\hat{G}_0''(\hat{G}_0^{-1}(p)) = \hat{G}_0'(1)$. Note that when $\hat{G}_1'(1) = 1$, namely, q satisfies $G_0'(1) = G_0''(G_0^{-1}(q))$, we have $\hat{G}_0''(h) = \hat{G}_0'(1) = \hat{G}_0''(1)$ and hence $\hat{p}_c = 1$, which is precisely the time when a giant component first forms in NS.

The fraction of the giant component in NS is given by

$$S(\mathrm{LA}) = 1 - \hat{G}_{0,p}(u), \tag{13}$$

where $u = \hat{G}_{1,p}(u)$. By definition, we have $P_\infty(\mathrm{LA}) = pqS(\mathrm{LA})$. Noting that $\hat{G}_0(x) = G_0(x)$ by (4) when $q = 1$, we again reproduce the usual LA attacks on the entire network.

C. Subgraph robustness under TA

In a targeted attack, a fraction $1-p$ of nodes are attacked and removed according to their degrees. Following [37], [44], we assign to each node in the observed subgraph a value

$$W_\alpha(k_i) = \frac{k_i^\alpha}{\sum_{i=1}^N k_i^\alpha}, \tag{14}$$

to indicate the probability that a node i with degree k_i is attacked, where α is a real and N is the number of nodes in the subnetwork in question. When $\alpha > 0$, nodes with higher degree have a higher probability to be removed; pushing it to the limit $\alpha \to \infty$ yields the attack strategy that nodes are removed strictly in the decreasing order of connectivity. The case $\alpha < 0$ implies the opposite strategies. Note that TA with $\alpha = 0$ is equivalent to RA with equal probability. In fact, we have $p_c(\mathrm{TA}) = p_c(\mathrm{RA})$ and $P_\infty(\mathrm{TA}) = P_\infty(\mathrm{RA})$ for both US and NS when $\alpha = 0$; see below and Appendix A for a proof.

Fix a value of α. We begin with the robustness of US under TA. The generating function for the degree distribution of US, denoted by $\tilde{P}(k)$, is given by (7). In other words, we have $\tilde{G}_0(x) = \sum_{k=0}^\infty \tilde{P}(k)x^k = \sum_{k=0}^\infty P(k)(1-q+qx)^k$ In fact, $\tilde{P}(k)$ can be explicitly calculated as $\tilde{P}(k) = \frac{1}{k!}\frac{d^k\tilde{G}_0(x)}{(dx)^k}\big|_{x=0} = q^k \sum_{l=k}^\infty P(l)P_l^k(1-q)^{l-k}$ for $k \geq 0$. Following [44], [52], [53], we define $\tilde{G}_\alpha(x) = \sum_{k=0}^\infty \tilde{P}(k)x^{k^\alpha}$ and $t = \tilde{G}_\alpha^{-1}(p)$, and the degree distribution of the remaining nodes in US after TA (but keeping the edges connecting to those removed nodes) is generated by $\tilde{G}_t(x) = p^{-1}\sum_{k=0}^\infty \tilde{P}(k)t^{k^\alpha}x^k$. Performing another bond percolation by using the same approach as in [1], [44], we obtain the generating function of the remaining network as

$$\tilde{G}_{0,t}(x) = \tilde{G}_t(1 - p_t + p_t x), \tag{15}$$

where $p_t = \left[\sum_{k=0}^\infty \tilde{P}(k)kt^{k^\alpha}\right] \Big/ \left[\sum_{k=0}^\infty \tilde{P}(k)k\right]$. Define $\tilde{G}_{1,t}(x) = \tilde{G}_{0,t}'(x)/\tilde{G}_{0,t}'(1)$. Combining (7), (15) and the criterion for the network to collapse, $\tilde{G}_{1,t}'(1) = 1$, we find that

$$p_c(\mathrm{TA}) = \min\{1, p_{c,t}\}, \tag{16}$$

where $p_{c,t}$ satisfies $t = \tilde{G}_\alpha^{-1}(p)$ and $\sum_{k=0}^\infty \tilde{P}(k)k = \sum_{k=0}^\infty \tilde{P}(k)t^{k^\alpha}k(k-1)$. Note that when $q = 1/G_1'(1)$, we have $\tilde{G}_0''(1) = \tilde{G}_0'(1)$ and hence $p_{c,t} = t = 1$, which is precisely the time when a giant component first forms in US.

The fraction of the giant component in US can be expressed by

$$S(\mathrm{TA}) = 1 - \tilde{G}_{0,t}(u), \tag{17}$$

where $u = \tilde{G}_{1,t}(u)$. By definition, we have $P_\infty(\mathrm{TA}) = pqS(\mathrm{TA})$. Clearly, when $q = 1$, we reproduce the usual targeted attack framework in [44], [48] since $\tilde{G}_0(x) = G_0(x)$.

Finally, we consider the robustness of NS under TA. The generating function for the degree distribution of NS, denoted by $\hat{P}(k)$, is given by (4). Therefore, $\hat{P}(k)$ can be explicitly calculated as $\hat{P}(k) = \frac{1}{k!}\frac{d^k\hat{G}_0(x)}{(dx)^k}\big|_{x=0} = [k!G_0(f)]^{-1}G_0^{(k)}\left(f - G_0'(f)/G_0'(1)\right)\left[G_0'(f)/G_0'(1)\right]^k$ for $k \geq 0$. Similarly, following [44], [52], [53], we define $\hat{G}_\alpha(x) = \sum_{k=0}^\infty \hat{P}(k)x^{k^\alpha}$ and $s = \hat{G}_\alpha^{-1}(p)$, and the degree

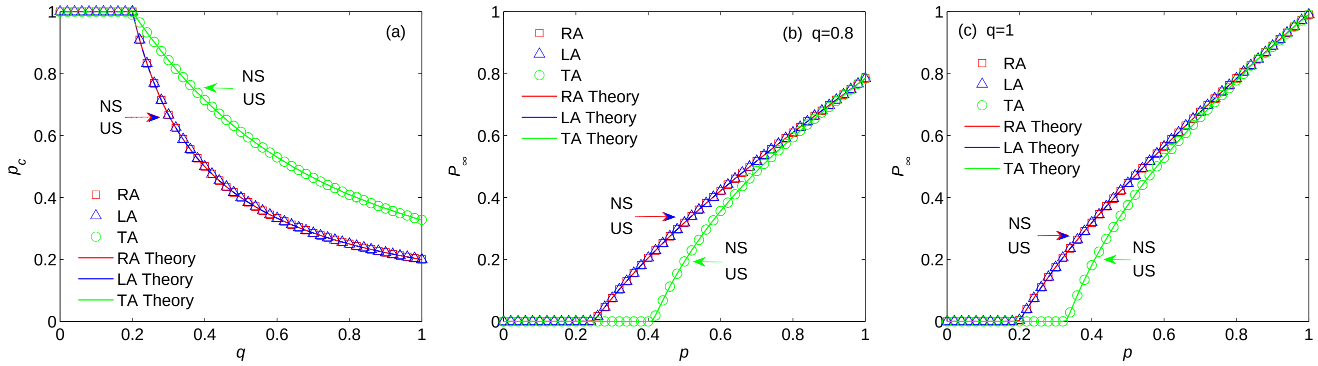

Fig. 2. (a) Percolation threshold p_c as a function of relative subgraph size q for ER networks with size $N = 10^6$ and $\lambda = 5$. Corresponding fraction of giant component P_∞ as a function of p is presented for (b) $q = 0.8$ and (c) $q = 1$. Theoretical predictions (solid lines) and simulations (symbols) for US and NS, and for RA, LA, and TA (with $\alpha = 1$), respectively, agree well with each other, where averages are taken over 100 realizations.

distribution of the remaining nodes in NS after TA (but keeping the edges connecting to those removed nodes) is generated by $\hat{G}_s(x) = p^{-1}\sum_{k=0}^{\infty}\hat{P}(k)s^{k^{\alpha}}x^k$. Performing a bond percolation by using the same approach as in [1], [44], we obtain the generating function of the remaining network as

$$\hat{G}_{0,s}(x) = \hat{G}_s(1 - p_s + p_s x), \tag{18}$$

$p_s = \left[\sum_{k=0}^{\infty}\hat{P}(k)ks^{k^{\alpha}}\right] \Big/ \left[\sum_{k=0}^{\infty}\hat{P}(k)k\right]$. Define $\hat{G}_{1,s}(x) = \hat{G}'_{0,s}(x)/\hat{G}'_{0,s}(1)$. Combining (4), (18) and the criterion for the network to collapse, $\hat{G}'_{1,s}(1) = 1$, we find that

$$p_c(\mathrm{TA}) = \min\{1, p_{c,s}\}, \tag{19}$$

where $p_{c,s}$ satisfies $s = \hat{G}_{\alpha}^{-1}(p)$ and $\sum_{k=0}^{\infty}\hat{P}(k)k = \sum_{k=0}^{\infty}\hat{P}(k)s^{k^{\alpha}}k(k-1)$. Note that when $\hat{G}'_1(1) = 1$, namely, q satisfies $G'_0(1) = G''_0(G_0^{-1}(q))$, we have $\hat{G}''_0(1) = \hat{G}'_0(1)$ and hence $p_{c,s} = s = 1$, which is precisely the time when a giant component first forms in NS.

Similarly, the fraction of the giant component in NS is given by

$$S(\mathrm{TA}) = 1 - \hat{G}_{0,s}(u), \tag{20}$$

where $u = \hat{G}_{1,s}(u)$. By definition, we have $P_\infty(\mathrm{TA}) = pqS(\mathrm{TA})$. When $q = 1$, we again reproduce the usual targeted attack framework because $f = 1$ and $\hat{G}_0(x) = G_0(x)$.

IV. Numerical results

In this section, we calculate numerical solutions of the analytical expressions and compare our theoretical results with simulations on three types of complex network benchmarks including ER, RR, and SF networks. An ER network follows a Poisson degree distribution $P(k) = e^{-\lambda}\lambda^k/k!$ ($k \geq 0$) with average degree $\langle k\rangle = \lambda$. An RR network has a degenerated degree distribution $P(k) = \delta_{k,k_0}$, meaning that each node is connected to the same number k_0 of neighbors. A SF network follows a power-law degree distribution $P(k) \sim k^{-\gamma}$ ($k_{\min} \leq k \leq k_{\max}$), where $\gamma > 0$ is the scaling exponent, $k_{\min}$ and $k_{\max}$ indicate the minimum and maximum degrees, respectively. All the simulation results are obtained for networks with $N = 10^6$ nodes. We also instantiate the general formula obtained in Section 3 in the special cases of ER and RR networks in Appendices B and C, respectively, for reference.

A. ER networks

The subgraph robustness results gathered in Fig. 2 for ER network allow us to draw several interesting comments. First, an increase in the relative subgraph size q systematically yields an decrease in p_c as well as an increase in P_∞ for both US and NS, and for all attack strategies. This means that the larger the observed subgraph is, the longer it takes to break it, as one would expect. Furthermore, the bias of p_c caused by the data missing is not linearly correlated with q. When q is small, e.g., $q \in [0.2, 0.4]$, p_c changes dramatically, while $p_c(q)$ is relatively close to $p_c(1)$ for $q > 0.8$. This suggests that ER networks have a tolerance for mild data missing (esp. under RA and LA), lending support to the qualitatively similar observations for other topological properties [7], [19], [20].

Second, as predicted in Appendix B, we have $p_c^{\mathrm{US}}(\mathrm{RA}) = p_c^{\mathrm{NS}}(\mathrm{RA}) = p_c^{\mathrm{US}}(\mathrm{LA}) = p_c^{\mathrm{NS}}(\mathrm{LA})$ and $P_\infty^{\mathrm{US}}(\mathrm{RA}) = P_\infty^{\mathrm{NS}}(\mathrm{RA}) = P_\infty^{\mathrm{US}}(\mathrm{LA}) = P_\infty^{\mathrm{NS}}(\mathrm{LA})$ for all p and q. This is an extension of the phenomenon discovered originally in [40] that the two competitive factors behind LA, namely, the factor due to heterogeneity that hubs are more likely within the attacked area accelerating the network fragmentation and the factor due to localization that only nodes on the surface of the attacked area contribute to the breakdown mitigating the fragmentation process, compensate exactly for each other in ER networks. Our theoretical calculations (see Appendix B) indicate that both US and NS possess the same thinned Poisson degree distribution (namely, with smaller event rate), which explains the equivalent effect of RA and LA on them. In other words, the uniform sampling and non-uniform sampling considered here for ER networks are essentially equivalent.

Third, among the three types of attacks, TA is always the most powerful for both US and NS, as well as all p and q. Moreover, the observed equivalence, namely, $p_c^{\mathrm{US}}(\mathrm{TA}) = p_c^{\mathrm{NS}}(\mathrm{TA})$ and $P_\infty^{\mathrm{US}}(\mathrm{TA}) = P_\infty^{\mathrm{NS}}(\mathrm{TA})$, holds for all α, which is again due to the second point mentioned above. Fourth, from Fig. 2(b) and (c) we observe second-order percolation transition behaviors as expected. The critical threshold at

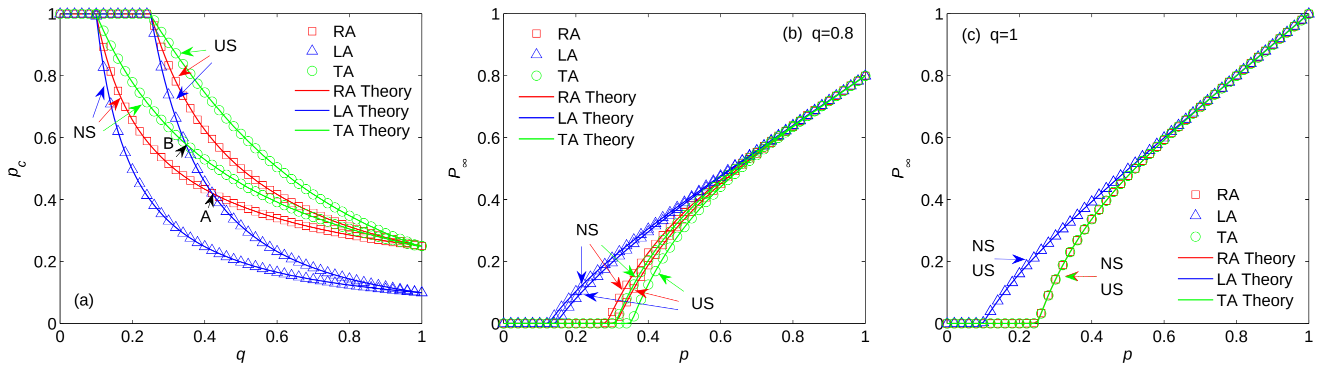

Fig. 3. (a) Percolation threshold p_c as a function of relative subgraph size q for RR networks with size $N = 10^6$ and $k_0 = 5$. Two points of intersection are indicated by A and B. Corresponding fraction of giant component P_∞ as a function of p is presented for (b) $q = 0.8$ and (c) $q = 1$. Theoretical predictions (solid lines) and simulations (symbols) for US and NS, and for RA, LA, and TA (with $\alpha = 1$), respectively, agree well with each other, where averages are taken over 100 realizations.

$P_\infty = 0$ coincides with the critical probability p_c in Fig. 2(a) for all attack strategies and all q considered. For instance, both Fig. 2(a) and (b) indicate $p_c \approx 0.4$ for TA (with $\alpha = 1$) when $q = 0.8$, i.e., 80% of the nodes are observed.

B. RR networks

In Fig. 3 we show the subgraph robustness for RR networks under various attack schemes. The behaviors observed deviate largely from those in ER networks. First, we have $p_c^{\rm US} \geq p_c^{\rm NS}$ for any relative subgraph size q and attack strategies considered, where the equality is attained at $q = 1$ (see Fig. 3(a)). This means that US of RR networks is more vulnerable against attacks than NS of the same size. Note that one of the fundamental observations in [40], [43], putting it in our language, that the giant component in NS of RR networks is larger than that in US of RR networks does not imply our result since p_c is not linearly correlated with P_∞.

Second, for both US and NS, TA is the most powerful one among these attacks as one would expect. Moreover, we observe $p_c^{\rm US}({\rm RA}) \geq p_c^{\rm US}({\rm LA})$ and $p_c^{\rm NS}({\rm RA}) \geq p_c^{\rm NS}({\rm LA})$, and the analogous inequalities for $P_\infty^{\rm US}$ and $P_\infty^{\rm NS}$ hold for all p, meaning that RA is always more powerful than LA. (The rigorous quantitative relationship between them can be found in Appendix C.) When $q = 1$, i.e., the entire RR network is observed, this phenomenon can be attributed to the disappearance of heterogeneity factor behind LA, with only the localization factor mitigating the fragmentation of the network leading to the lower efficiency of LA [40]. Our finding highlights that US and NS for any $q \leq 1$ are still quite homogeneous to the extent that localization factor becomes dominant and the subnetwork in question becomes more robust against LA than against RA.

Third, we find interestingly from Fig. 3(a) two crossover points, namely, A of $p_c^{\rm US}({\rm LA})$ and $p_c^{\rm NS}({\rm RA})$, and B of $p_c^{\rm US}({\rm LA})$ and $p_c^{\rm NS}({\rm TA})$. It is easy to verify that such crossover points exist for all RR networks with $k_0 > 2$ by our theoretical derivation in Appendix C. The existence of these crossover points indicates that, although RR networks are more resilient against LA than against RA for a given type of subgraph sampling, US under LA can be more fragile than NS under RA when $q < A_q$, where A_q represents the value of q corresponding to the point A. Similar results hold when comparing LA and TA (with any given α). These phenomena highlight that LA is not always the least harmful strategy for RR networks when subgraph robustness is taken into account, in sharp contrast to the robustness behavior of the entire network [40], [43].

Fourth, comparing Fig. 3(b), (c) and Fig. 3(a), we see that the critical threshold at $P_\infty = 0$ again coincides with the critical probability p_c for all attack strategies and all q considered. The behaviors of P_∞ and p_c are generally consistent with each other. As in ER networks, $p_c(q)$ (and $P_\infty(p)$ at q, resp.) is quite close to $p_c(1)$ (and $P_\infty(p)$ at $q = 1$, resp.) when q is large, say, $q \geq 0.8$. This implies that that RR networks also have a tolerance for mild data missing.

C. SF networks

We show in Fig. 4 the subgraph robustness for SF networks with an archetypal scaling exponent $\gamma = 2.36$ under various attack schemes. The behaviors observed differ markedly from those for homogeneous ones such as ER and RR networks. First, we observe that $p_c^{\rm US} \leq p_c^{\rm NS}$ for any relative subgraph size q and attack strategies considered (see Fig. 4(a)), where the equality is attained at $q = 1$ implying that US of SF networks is more resilient against attack than NS of the same size. Similarly as commented above, this goes beyond the basic observation in [40], [43] that the giant component in US of SF networks is larger than that in NS of SF networks. It is worth mentioning that the difference between the two curves $p_c^{\rm NS}$ and $p_c^{\rm US}$ hinges on the degree of heterogeneity of the underlying network. For instance, when the network in question becomes more homogeneous, i.e., when it possesses a larger γ, the inequality may reverse (the corresponding percolation thresholds will look like Fig. 3(a) for an RR network). We verified this by simulations for SF networks with $\gamma = 5$. Similar crossover phenomena for SF networks have been reported recently in [40], [42], [43] due to the competition of the two factors behind LA.

Second, TA is again the most powerful one among these three types of attack for both US and NS. However, in contrast

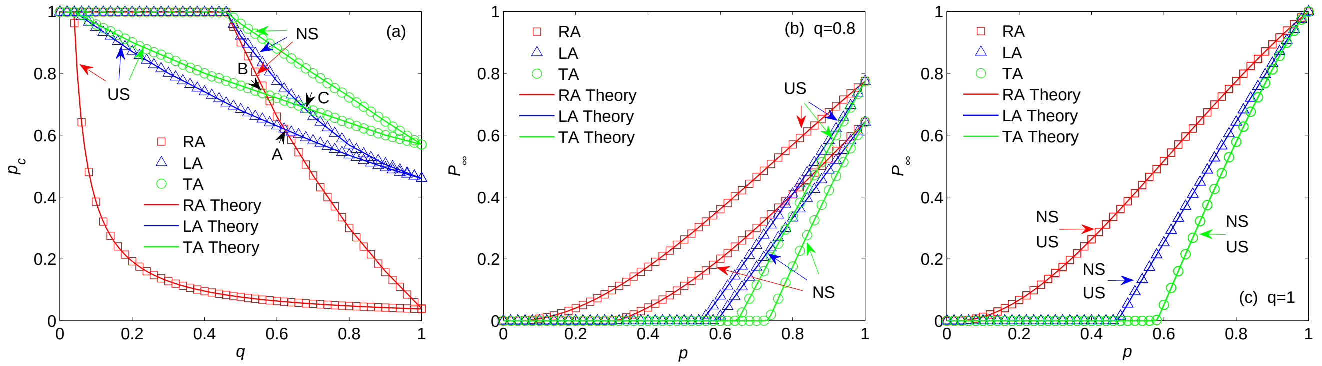

Fig. 4. (a) Percolation threshold p_c as a function of relative subgraph size q for SF networks with size $N = 10^6$, $\gamma = 2.36$, $k_{\min} = 2$, and $\langle k \rangle = 5$. Three points of intersection are indicated by A, B and C. Corresponding fraction of giant component P_∞ as a function of p is presented for (b) $q = 0.8$ and (c) $q = 1$. Theoretical predictions (solid lines) and simulations (symbols) for US and NS, and for RA, LA, and TA (with $\alpha = 1$), respectively, agree well with each other, where averages are taken over 100 realizations.

to ER and RR networks, we see that $p_c^{\rm US}({\rm LA}) \geq p_c^{\rm US}({\rm RA})$ and $p_c^{\rm NS}({\rm LA}) \geq p_c^{\rm NS}({\rm RA})$, and the analogous inequalities for $P_\infty^{\rm US}$ and $P_\infty^{\rm NS}$ hold for all p, suggesting that LA is always more powerful than RA. In the special case of $q = 1$, this phenomenon is first reported in [40], attributing to the dominance of heterogeneity factor of LA, which accelerates the breakdown of the network under LA. Our finding reveals that US and NS for any $q \leq 1$ are still heterogenous to the extent that heterogeneity factor remains dominant. This phenomenon, nevertheless, again relies on the heterogeneity of the SF network in question; namely, for SF networks with large γ, the robustness behaviors will more or less like those for RR networks. This is confirmed by simulations for SF networks with $\gamma = 5$.

Third, Fig. 4(a) displays three crossover points, i.e., A of $p_c^{\rm US}({\rm LA})$ and $p_c^{\rm NS}({\rm RA})$, B of $p_c^{\rm US}({\rm TA})$ and $p_c^{\rm NS}({\rm RA})$, and C of $p_c^{\rm US}({\rm TA})$ and $p_c^{\rm NS}({\rm LA})$. It is easy to verify numerically that such crossover points exist for SF networks with relatively small γ, namely, for typical heterogenous SF networks. The existence of these crossover points suggests that, although a typical SF network is more resilient against RA than against LA or TA for a given type of subgraph sampling, NS under RA can be more vulnerable than US under LA when $q < A_q$, and resp., US under TA when $q < B_q$, where A_q and B_q are defined as before. This means that RA is not always the least harmful strategy for a typical SF network when subgraph robustness is taken into account, Furthermore, TA is not always the most powerful attack either when it comes to different subgraph robustness. For example, NS under LA is more fragile than US under TA (with $\alpha = 1$) when $q < C_q$. These phenomena highlight the importance of understanding subgraph robustness in predicting network robustness as well as designing resilient infrastructures.

Fourth, comparing Fig. 4(b), (c) and Fig. 4(a), we observe that the critical threshold at $P_\infty = 0$ coincides with the critical probability p_c for all attack strategies and all q considered. Similarly as in ER and RR networks, the behaviors of P_∞ and p_c are generally consistent with each other.

Finally, distinct from ER and RR networks, we notice that the curve $p_c^{\rm NS}({\rm RA})$ (as well as $P_\infty^{\rm NS}({\rm RA})$ at q) changes much more prominently as compared to the other five curves for relatively large values of q, e.g., $q \geq 0.8$ (see Fig. 4(a)). This can be intuitively explained as follows. When q decreases gradually starting from 1 in the non-uniform way, the observed subnetwork (NS) undergoes a change by missing possibly a handful of hubs but in a localized way. The subgraph robustness for LA and TA is not very sensitive to q because the attacked nodes under these two strategies can be far away from the missing nodes. However, for RA, the attacked nodes are likely to escalate the damage caused by the missing nodes, producing an evident change of the subgraph robustness with respect to q. On the other hand, as one would expect, $p_c^{\rm US}$ is not sensitive for large q since the missing nodes are mostly small degree nodes having limited contribution to the network robustness. Our result indicates that a typical SF network may have very poor error tolerance for mild data missing in the non-uniform way while at the same time hold relatively strong tolerance for mild data missing in the uniform way, complementing the celebrated robustness characteristics of SF networks, namely, they are resilient against random error but fragile to targeted attack [26].

D. A closer look at US and NS with data missing

The quantitative difference between uniform and non-uniform random sampling for different network topologies and attack strategies can be better fathomed using the ratio $p_c^{\rm US}/p_c^{\rm NS}$ in Fig. 5 and the discrepancy $P_\infty^{\rm NS} - P_\infty^{\rm US}$ in Fig. 6. Marked signatures of ER, RR, and SF networks can be observed. First of all, US and NS of ER networks have precisely the same robustness in terms of p_c and P_∞ under all kinds of attack strategies as discussed above.

For RR networks, NS turns out to be always more robust than US with the peak of the ratio $p_c^{\rm US}/p_c^{\rm NS}$ attained at $q = (k_0 - 1)^{-1} = 0.25$ (c.f. Fig. 3(a)). The peak of the ratio for LA is over 2.5, which is more prominent than those for RA and TA. However, with relatively mild data missing, say, $q \geq 0.6$, the ratio for RA, LA, and TA are similar. Therefore, if we have a subnetwork of an RR network without knowing the sampling process, we may perform LA on it and establish

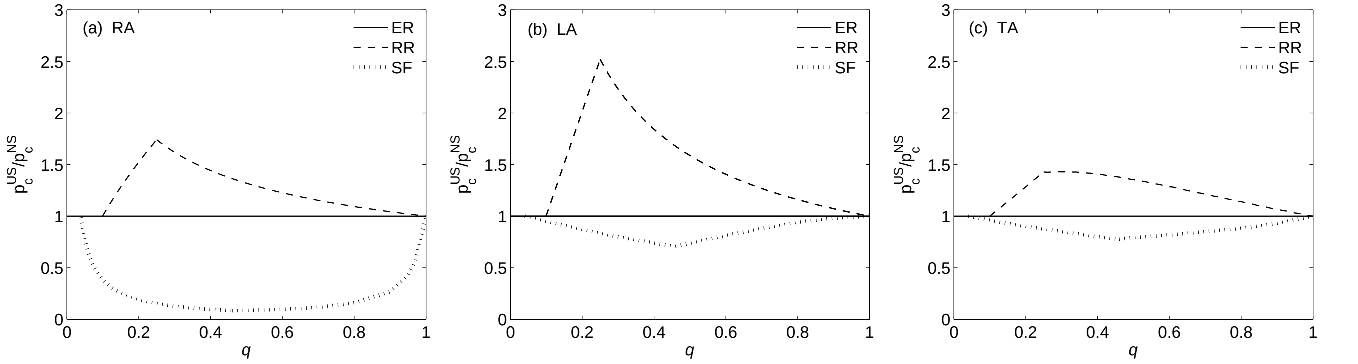

Fig. 5. The ratio $p_c^{\mathrm{US}}/p_c^{\mathrm{NS}}$ as a function of relative subgraph size q under (a) RA, (b) LA, and (c) TA (with $\alpha = 1$). The same ER, RR, and SF networks as above are used here.

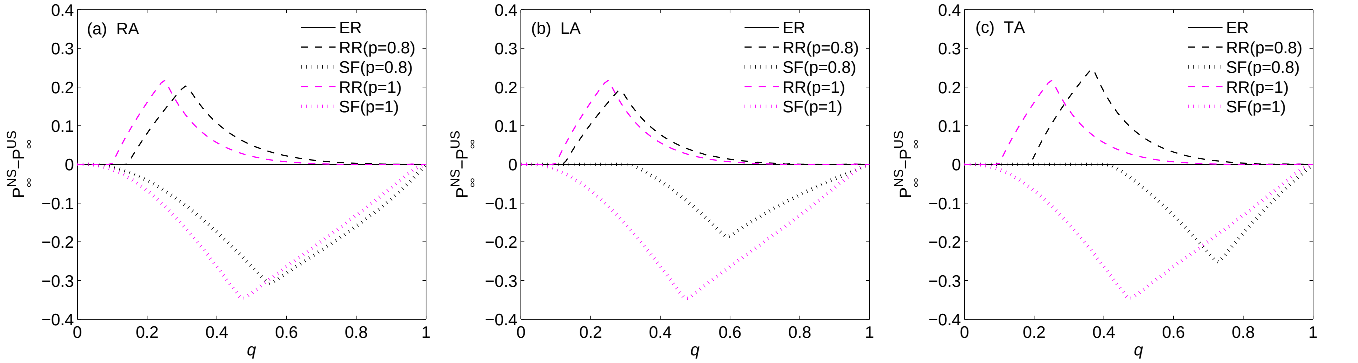

Fig. 6. The difference $P_\infty^{\mathrm{NS}} - P_\infty^{\mathrm{US}}$ as a function of relative subgraph size q under (a) RA, (b) LA, and (c) TA (with $\alpha = 1$) for $p = 0.8$ and 1. The same ER, RR, and SF networks as above are used here.

Bayesian statistical tests to determine whether uniform or non-uniform sampling is closer to the truth. It follows from Fig. 6 that all the three attacks are of similar efficacy on RR networks when $P_\infty^{\mathrm{NS}} - P_\infty^{\mathrm{US}}$ is taken into consideration. They are good estimators for differentiating between uniform and non-uniform random samplings again around $q = (k_0 - 1)^{-1}$ (where the peaks appear in Fig. 6) with a shift towards the larger q as the attack goes on. It is also noteworthy that, in the event of mild data missing, e.g., $q \geq 0.7$, the difference between P_∞^{NS} and P_∞^{US} appears negligible when the loss of the nodes is minor, e.g., does not exceed 20% (c.f. Fig.3 (b), (c)).

For SF networks, US is always more robust than NS as we have explained above. The ratios $p_c^{\mathrm{US}}/p_c^{\mathrm{NS}}$ for LA and TA are very similar with a negative peak at $q \approx 0.45$ corresponding to the critical value at which the giant component of NS collapses (c.f. Fig. 4(a)). Strikingly, the ratio for RA behaves rather differently: it has a flat and deep bottom in the interval (approximately) $q \in [0.2, 0.8]$ and increases rapidly when q approaches 1. Due to the displayed remarkable difference between p_c^{US} and p_c^{NS}, RA can be exploited to effectively distinguish between US and NS for both mild and severe data missing scenarios in SF networks. We observe from Fig. 6 that the discrepancy $P_\infty^{\mathrm{NS}} - P_\infty^{\mathrm{US}}$ displays similar patterns for all three attacks on SF networks; namely, a negative peak appears at $q \approx 0.45$ with a shift towards the larger q as the attack continues. In contrast to RR networks, we find that $P_\infty^{\mathrm{NS}} - P_\infty^{\mathrm{US}}$ for SF networks is the most sensitive to TA while the least sensitive to RA among the three attacks considered. This sheds light on the essential difficulty in dealing with TA on SF networks, which perhaps is the most common real-life attack situation [26]; namely, the robustness of SF networks under TA distinctively associates to varied factors, including the amount of data missing, the sampling methods, and the different stages of attack.

The key contributions of this work are summarized in Table I below.

V. Applications on Real Networks

The study of subgraph robustness is important for understanding appropriately the resilience of many real-world networks since data missing is prevalent in such systems. To illustrate the availability of our framework, we investigate four real-world networks: (i) a friendship network based on Brightkite social network (Friend) on $N = 56739$ nodes, where nodes represent users and edges indicate the friendship between them [54]; (ii) a road network of Pennsylvania (Road) on $N = 1087562$ nodes, where nodes represent intersections between roads and edges mean road segments [55]; (iii) a metabolic network of the Reactome project (Metabolism) on $N = 5973$ nodes, where nodes are proteins and edges are interactions between them [56]; (iv) a computer network of

TABLE I
SUBGRAPH ROBUSTNESS CHARACTERIZATION OF NETWORKS UNDER ATTACKS.

	Main findings	
ER networks	1. US and NS are equivalent; 2. Both US and NS have good tolerance for mild data missing under all attack strategies; 3. No crossover phenomenon occurs between US and NS;	1. Attack robustness increases with the relative subgraph size for both US and NS; 2. TA is the most harmful attack strategy for both US and NS; 3. P_∞ shows second-order phase transition for both US and NS under all attack strategies;
RR networks	1. US is more vulnerable against attacks than NS of the same size under all attack strategies; (The difference is maximized in general under LA); 2. RA is more harmful than LA for both US and NS; 3. Both US and NS have good tolerance for mild data missing under all attack strategies; 4. Two crossover points exist between US and NS with respect to relative subgraph size;	
SF networks	1. US is more resilient against attacks than NS of the same size under all attack strategies; (The difference is maximized in general under RA); 2. LA is more harmful than RA for both US and NS; 3. NS has very poor tolerance for mild data missing under RA, while US and NS have good tolerance for mild data missing under other attack strategies; 4. Three crossover points exist between US and NS with respect to relative subgraph size;	

the Skitter project (Computer) on $N = 1694616$ nodes, where nodes are autonomous system on the Internet and edges are connections [57].

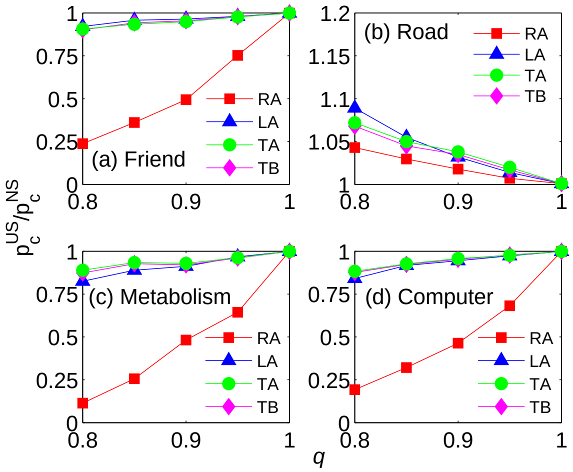

Fig. 7. The ratio $p_c^{\rm US}/p_c^{\rm NS}$ as a function of relative subgraph size q for four real networks (a) Friend (b) Road (c) Metabolism (d) Computer under RA, LA, TA (with $\alpha = 1$), and TB. For each q, the result is averaged over 100 simulation runs.

In addition to RA, LA, TA, we here study the effect on network robustness of targeted removal of nodes according to betweenness centrality (TB for short). The betweenness centrality of a node is defined as the number of geodesic paths that pass through this node [1]. Different from degree centrality, betweenness centrality accounts for non-local structure of the network, where nodes with high betweenness centrality play a key role in governing the information flow.

We simulate the ratio $p_c^{\rm US}/p_c^{\rm NS}$ for the four types of attack strategies on these networks with mild data missing, i.e., $q \in [0.8, 1]$. The results are gathered in Fig. 7. Apparently, Friend, Metabolism, and Computer networks show the signature of SF networks; the curve for RA is much steeper than those for LA and TA when q approaches 1. This agrees with our theory since the empirical statistics obtained in [54], [56], [57] show power-law degree distributions for all these three networks with scaling exponents $\gamma \approx 2.5, 1.7$, and 2.3, respectively. On the other hand, note that Road network is a highly regular one with maximum degree 9 [55]. The curves in Fig. 7(b) decreases gradually with respect to q. It is interesting to observe that the ratio under LA begins to surpass that under TA when q is getting smaller than, say, 0.85, which is in line with our above derived result for RR.

One somewhat unexpected result is that targeting nodes according to either degree (TA) or betweenness (TB) has much the same effect. We contend that the similarity in effect may find its origins in the lack of specific structural properties that would favor betweenness centrality to be superior robustness indicators than degree, as is observed in [57] for Computer network. For example, although the networks considered here have power-law or exponential degree distributions with changing q, they are essentially random by nature and are lack of low degree nodes acting as “bridges” connecting highly connected parts of the networks. The correlation between TA and TB has also been discussed and compared numerically in [31], [58] in terms of giant component size and diameter.

VI. CONCLUSIONS

Complex networks underlying a variety of technical, biological, social, and physical systems are confronted with data missing constantly. In this paper, we introduce the subgraph robustness problem created from the uniform random sampling as well as the non-uniform random sampling. We develop a theoretical framework to investigate robustness properties of the two types of subnetworks under random attacks, localized attacks, and targeted attacks. We show that ER, RR, and SF networks have their own characteristic subgraph robustness features, which are distinct from the robustness of the entire networks. Our results underscore the importance of understanding the different sampling processes and their consequences on attack robustness of various network structures (see Table I in Section IV).

In the present study, we have shown that the evaluation of the impact of failures on the network connectivity can be biased by the lack of information based on a priori knowledge of the degree distribution. Nevertheless, in most practical cases, it is unfeasible to estimate the network properties in advance as a means of supporting the robustness to failure analysis. Techniques tailored for specific applications, such as immunization and communication systems, are to be proposed to deal with the lack of information despite the network topology properties. On the other hand, the importance of looking at the entire history of the disintegration process to understand the network robustness is argued in [51]. How data missing influences the structural robustness in this context is to be understood. The present work may provide a useful theoretical reference for these future quest and exploitation. It is hoped that our results will stimulate further research efforts on the subgraph robustness problem and other related interesting and challenging questions.

Appendix A
Proof for equivalence of RA and TA when $\alpha = 0$

Let $\alpha = 0$ in (14). We first prove the equivalence of RA and TA in US. It follows from (7) and $\sum_{k=0}^{\infty}\tilde{P}(k)k = \sum_{k=0}^{\infty}\tilde{P}(k)tk(k-1)$ (see the equation below Eq.(16)) that

$$t = \frac{\sum_{k=0}^{\infty}\tilde{P}(k)k}{\sum_{k=0}^{\infty}\tilde{P}(k)k(k-1)} = \frac{\tilde{G}_0'(1)}{\tilde{G}_0''(1)} = \frac{G_0'(1)}{qG_0''(1)}. \tag{21}$$

Noting that $\tilde{G}_{\alpha=0}(x) = x$, we have $p_{c,t} = t = 1/[qG_1'(1)]$ and $p_c(\mathrm{TA}) = p_c(\mathrm{RA})$ by (16) and (2). Now, we turn to the relative size of the giant component. Note that $t = p$. It then follows from (15) that $\tilde{G}_{0,t}(x) = \tilde{G}_0(1-t+tx) = \sum_{k=0}^{\infty}P(k)(1-pq+pqx)^k$. Therefore, (17) reduces to

$$\begin{cases} P_\infty(\mathrm{TA}) = pq\left[1-\sum_{k=0}^{\infty}P(k)(1-pq+pqv)^k\right], \\ v = \tilde{G}_{1,t}(u) = \frac{\sum_{k=0}^{\infty}(1-pq+pqv)^{k-1}}{\sum_{k=0}^{\infty}kP(k)}. \end{cases} \tag{22}$$

It is easy to see that $P_\infty(\mathrm{TA}) = P_\infty(\mathrm{RA})$ by comparing (22) and (3) and employing the transformation $u = 1-pq+pqv$.

Next, we prove the equivalence of RA and TA in NS. It follows from (4) and $\sum_{k=0}^{\infty}\hat{P}(k)k = \sum_{k=0}^{\infty}\hat{P}(k)sk(k-1)$ (see the equation below Eq.(19)) that

$$s = \frac{\sum_{k=0}^{\infty}\hat{P}(k)k}{\sum_{k=0}^{\infty}\hat{P}(k)k(k-1)} = \frac{\hat{G}_0'(1)}{\hat{G}_0''(1)} = \frac{1}{\hat{G}_1'(1)}. \tag{23}$$

Noting that $\hat{G}_{\alpha=0}(x) = x$, we have $p_{c,s} = s = 1/\hat{G}_1'(1)$ and hence $p_c(\mathrm{TA}) = p_c(\mathrm{RA})$ by (19) and (5). Note that $s = p$. Using (4) and (18), we observe that (20) can be recast as

$$\begin{cases} P_\infty(\mathrm{TA}) = pq[1-\hat{G}_s(1-p+pv)] \\ = pq\left[1-\sum_{k=0}^{\infty}\hat{P}(k)(1-p+pv)^k\right], \\ v = \hat{G}_{1,s}(v) = \frac{\hat{G}_{0,s}'(v)}{\hat{G}_{0,s}'(1)} = \frac{\hat{G}_0'(1-p+pv)}{\hat{G}_0'(1)} \\ = \hat{G}_1(1-p+pv). \end{cases} \tag{24}$$

It is then direct to verify that $P_\infty(\mathrm{TA}) = P_\infty(\mathrm{RA})$ by comparing (24) and (6) and applying the transformation $u = 1-p+pv$.

Appendix B
Derivation for subgraph robustness of ER networks

Recall that an ER network follows a Poisson degree distribution $P(k) = e^{-\lambda}\lambda^k/k!$ for $k \geq 0$. Hence, $G_0(x) = G_1(x) = e^{\lambda(x-1)}$. By directed calculations based on the theoretical framework developed in Section 2, we derive the following.

For US under RA, we have $p_c^{\mathrm{US}}(\mathrm{RA}) = \min\{1, (\lambda q)^{-1}\}$. (Here, and in what follows, we will often use the superscripts US and NS to avoid ambiguity.) $P_\infty^{\mathrm{US}}(\mathrm{RA}) = pq[1-e^{\lambda(u-1)}]$, where $u = 1-pq+pqe^{\lambda(u-1)}$. Note that $\hat{G}_0(x) = \hat{G}_1(x) = e^{\lambda q(x-1)}$. Hence, for NS under RA, we have $p_c^{\mathrm{NS}}(\mathrm{RA}) = \min\{1, (\lambda q)^{-1}\}$ and $P_\infty^{\mathrm{NS}}(\mathrm{RA}) = pq[1-e^{\lambda q(u-1)}]$, where $u = 1-p+pe^{\lambda q(u-1)}$.

For US under LA, we obtain $p_c^{\mathrm{US}}(\mathrm{LA}) = \min\{1, (\lambda q)^{-1}\}$ and $P_\infty^{\mathrm{US}}(\mathrm{LA}) = pq[1-e^{\lambda pq(u-1)}]$, where $u = e^{\lambda pq(u-1)}$ by noting that $\tilde{G}_0(x) = e^{\lambda q(x-1)}$ and $\tilde{G}_{0,p}(x) = \tilde{G}_{1,p}(x) = e^{\lambda pq(x-1)}$. Similarly, we have $\hat{G}_{0,p}(x) = \hat{G}_{1,p}(x) = e^{\lambda pq(x-1)}$. Therefore, for NS under LA, we have $p_c^{\mathrm{NS}}(\mathrm{LA}) = \min\{1, (\lambda q)^{-1}\}$ and $P_\infty^{\mathrm{NS}}(\mathrm{LA}) = pq[1-e^{\lambda pq(u-1)}]$, where $u = e^{\lambda pq(u-1)}$. It is easy to see that $p_c^{\mathrm{US}}(\mathrm{RA}) = p_c^{\mathrm{NS}}(\mathrm{RA}) = p_c^{\mathrm{US}}(\mathrm{LA}) = p_c^{\mathrm{NS}}(\mathrm{LA})$ and

$$P_\infty^{\mathrm{US}}(\mathrm{RA}) = P_\infty^{\mathrm{NS}}(\mathrm{RA}) = P_\infty^{\mathrm{US}}(\mathrm{LA}) = P_\infty^{\mathrm{NS}}(\mathrm{LA}) \tag{25}$$

holds for all p and q.

To better appreciate the subgraph robustness under TA, we here focus on the special case $\alpha = 1$, that is, nodes are deleted linearly depending on their degrees. Note that $\tilde{P}(k) = e^{-\lambda q}(\lambda q)^k/k!$ for $k \geq 0$, $p_t = tp$, and $\tilde{G}_t(x) = p^{-1}e^{\lambda q(tx-1)}$. For US under TA, we obtain $p_c^{\mathrm{US}}(\mathrm{TA}) = \min\{1, p_{c,t}\}$, where $p_{c,t}$ is determined by $1 = \lambda qt^2e^{\lambda q(t-1)}$ and $p = e^{\lambda q(t-1)}$. We have $\tilde{G}_{0,t}(x) = \tilde{G}_{1,t}(x) = e^{\lambda pqt^2(x-1)}$. Hence, $P_\infty^{\mathrm{US}}(\mathrm{TA}) = pq[1-e^{\lambda pqt^2(u-1)}]$, where t is determined by $p = e^{\lambda q(t-1)}$ and u is determined by $u = e^{\lambda pqt^2(u-1)}$. Note that $\hat{P}(k) = \tilde{P}(k)$, $s = t$, and $\hat{G}_s(x) = p^{-1}e^{\lambda q(sx-1)}$. Accordingly, we have $p_c^{\mathrm{NS}}(\mathrm{TA}) = p_c^{\mathrm{US}}(\mathrm{TA})$ and

$$P_\infty^{\mathrm{NS}}(\mathrm{TA}) = P_\infty^{\mathrm{US}}(\mathrm{TA}) \tag{26}$$

for all p and q. It is not difficult to see that (26) also holds for all α.

Appendix C
Derivation for subgraph robustness of RR networks

An RR network has a degenerated degree distribution $P(k) = \delta_{k,k_0}$. Hence, $G_0(x) = xG_1(x) = x^{k_0}$. For US under RA, we have $p_c^{\mathrm{US}}(\mathrm{RA}) = \min\{1, [q(k_0-1)]^{-1}\}$ and $P_\infty^{\mathrm{US}}(\mathrm{RA}) = pq(1-u^{k_0})$, where $u = 1-pq+pqu^{k_0-1}$. Note that $\hat{G}_0(x) = \left[1+q^{\frac{k_0-2}{k_0}}(x-1)\right]^{k_0}$ and $\hat{G}_1(x) = \left[1+q^{\frac{k_0-2}{k_0}}(x-1)\right]^{k_0-1}$. For NS under RA, we have $p_c^{\mathrm{NS}}(\mathrm{RA}) = \min\left\{1, \left[(k_0-1)q^{\frac{k_0-2}{k_0}}\right]^{-1}\right\}$ and $P_\infty^{\mathrm{NS}}(\mathrm{RA}) = pq\left\{1-\left[1+q^{\frac{k_0-2}{k_0}}(u-1)\right]^{k_0}\right\}$, where $u = 1-p+p\left[1+q^{\frac{k_0-2}{k_0}}(u-1)\right]^{k_0-1}$.

For US under LA, we have $p_c^{\mathrm{US}}(\mathrm{LA}) = \min\left\{1, [q(k_0-1)]^{-\frac{k_0}{k_0-2}}\right\}$ by noting $\tilde{G}_0(x) = (1-q+qx)^{k_0}$ and $p = (1-q+qg)^{k_0}$. $P_\infty^{\mathrm{US}}(\mathrm{LA}) = pq\left\{1-\left[1+qp^{\frac{k_0-2}{k_0}}(u-1)\right]^{k_0}\right\}$, where u is given by $u = \left[1+qp^{\frac{k_0-2}{k_0}}(u-1)\right]^{k_0-1}$. For NS under LA, we have $p_c^{\mathrm{NS}}(\mathrm{LA}) = \min\left\{1, q^{-1}(k_0-1)^{-\frac{k_0}{k_0-2}}\right\}$ and $P_\infty^{\mathrm{NS}}(\mathrm{LA}) = pq\left\{1-\left[1+(pq)^{\frac{k_0-2}{k_0}}(u-1)\right]^{k_0}\right\}$, where u is given by $u = \left[1+(pq)^{\frac{k_0-2}{k_0}}(u-1)\right]^{k_0-1}$. We find that for all q the following relations between RA and LA hold:

$$\begin{aligned} p_c^{\mathrm{US}}(\mathrm{RA})^{k_0} &= p_c^{\mathrm{US}}(\mathrm{LA})^{k_0-2}, \\ p_c^{\mathrm{NS}}(\mathrm{RA})^{k_0} &= p_c^{\mathrm{NS}}(\mathrm{LA})^{k_0-2}. \end{aligned} \tag{27}$$

For US under TA with $\alpha = 1$, we have $p_c^{\mathrm{US}}(\mathrm{TA}) = \min\left\{1, [t^2 q(k_0-1)]^{-\frac{k_0}{k_0-2}}\right\}$, where t satisfies $1 = t^2(k_0-1)q[1+q(t-1)]^{k_0-2}$; and $P_\infty^{\mathrm{US}}(\mathrm{TA}) = pq\left\{1-\left[1+qt^2p^{\frac{k_0-2}{k_0}}(u-1)\right]^{k_0}\right\}$, where u satisfies $u = \left[1+qt^2p^{\frac{k_0-2}{k_0}}(u-1)\right]^{k_0-1}$ and t is given by $p = [1+q(t-1)]^{k_0}$. For NS under TA with $\alpha = 1$, we derive similarly that $p_c^{\mathrm{NS}}(\mathrm{TA}) = \min\left\{1, q^{-1}[s^2(k_0-1)]^{-\frac{k_0}{k_0-2}}\right\}$, where s satisfies $1 = s^2(k_0-1)q^{\frac{k_0-2}{k_0}}\left[1+q^{\frac{k_0-2}{k_0}}(s-1)\right]^{k_0-2}$; and $P_\infty^{\mathrm{NS}}(\mathrm{TA}) = pq\left\{1-\left[1+s^2(pq)^{\frac{k_0-2}{k_0}}(u-1)\right]^{k_0}\right\}$, where u satisfies $u = \left[1+s^2(pq)^{\frac{k_0-2}{k_0}}(u-1)\right]^{k_0-1}$ and s is given by $p = \left[1+q^{\frac{k_0-2}{k_0}}(s-1)\right]^{k_0}$.

ACKNOWLEDGMENT

The author is grateful to the reviewers and the editor for their constructive and insightful comments that helped improve the paper significantly.

REFERENCES

[1] M. Newman, *Networks: An Introduction*. New York: Oxford University Press, 2010.

[2] K. S. K. Chung, M. Piraveenan, and L. Hossain, "Topology of online social networks," in *Encyclopedia of Social Network Analysis and Mining*. New York: Springer, 2014, pp. 2191–2202.

[3] G. Robins, P. Pattison, and J. Woolcock, "Missing data in networks: exponential random graph ($\rho*$) models for networks with non-respondents," *Soc. Netw.*, vol. 26, pp. 257–283, 2004.

[4] G. Kossinets, "Effects of missing data in social networks," *Soc. Netw.*, vol. 28, pp. 247–268, 2006.

[5] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. Hoboken, NJ: Wiley-Interscience, 2002.

[6] M. Papagelis, G. Das, and N. Koudas, "Sampling online social networks," *IEEE Trans. Know. Data Eng.*, vol. 25, pp. 662–676, 2013.

[7] J. A. Smith and J. Moody, "Structural effects of network sampling coverage i: nodes missing at random," *Soc. Netw.*, vol. 35, pp. 652–668, 2013.

[8] D. Liben-Nowell and J. M. Kleinberg, "The link-prediction problem for social networks," *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, pp. 1019–1031, 2007.

[9] A. Clauset, C. Moore, and M. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, pp. 98–101, 2008.

[10] A. F. AlEroud and G. Karabatis, "Queryable semantics to detect cyber-attacks: a flow-based detection approach," *IEEE Trans. Syst. Man Cy. Syst.*, 2016, doi:10.1109/TSMC.2016.2600405.

[11] L. Lü and T. Zhou, "Link prediction in complex networks: a survey," *Physica A*, vol. 390, pp. 1150–1170, 2011.

[12] Y. Ma, G. Cheng, Z. Liu, and F. Xie, "Fuzzy nodes recognition based on spectral clustering in complex networks," *Physica A*, vol. 465, pp. 792–797, 2017.

[13] F. Masrour, I. Barjesteh, R. Forsati, A.-H. Esfahanian, and H. Radha, "Network completion with node similarity: a matrix completion approach with provable guarantees," in *Proc. IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining*. Paris, France: ACM, 2015, pp. 302–307.

[14] D. Hric, T. P. Peixoto, and S. Fortunato, "Network structure, metadata, and the prediction of missing nodes and annotations," *Phys. Rev. X*, vol. 6, p. 031038, 2016.

[15] M. Kim and J. Leskovec, "The network completion problem: inferring missing nodes and edges in networks," in *Proc. SIAM Int. Conf. on Data Mining*. Philadelphia, USA: SIAM, 2011, pp. 47–58.

[16] Q. Zhang, C. Zhou, N. Xiong, Y. Qin, and S. Huang, "Multimodel-based incident prediction and risk assessment in dynamic cybersecurity protection for industrial control systems," *IEEE Trans. Syst. Man Cy. Syst.*, vol. 46, pp. 1429–1444, 2016.

[17] R. Eyal, A. Rosenfeld, S. Sina, and S. Kraus, "Predicting and identifying missing node information in social networks," *ACM Trans. Knowl. Discov. Data*, vol. 8, p. art. 14, 2014.

[18] M. P. H. Stumpf, C. Wiuf, and R. M. May, "Subnets of scale-free networks are not scale-free: sampling properties of networks," *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 4221–4224, 2005.

[19] S. H. Lee, P. J. Kim, and H. Jeong, "Statistical properties of sampled networks," *Phys. Rev. E*, vol. 73, p. 016102, 2006.

[20] C. A. Bliss, C. M. Danforth, and P. S. Dodds, "Estimation of global network statistics from incomplete data," *PLoS ONE*, vol. 9, p. e108471, 2014.

[21] J. A. Smith, J. Moody, and J. H. Morgan, "Network sampling coverage ii: the effect of non-random missing data on network measurement," *Soc. Netw.*, vol. 48, pp. 78–99, 2017.

[22] D. R. Amancio, O. N. O. Jr, and L. F. Costa, "Robustness of community structure to node removal," *J. Stat. Mech. Theory Exp.*, vol. 2015, p. P03003, 2015.

[23] C. Cooper, T. Radzik, and Y. Siantos, "Fast low-cost estimation of network properties using random walks," *Internet Math.*, vol. 12, pp. 221–238, 2016.

[24] A. D. Stivala, J. H. Koskinen, D. A. Rolls, P. Wang, and G. L. Robins, "Snowball sampling for estimating exponential random graph models for large networks," *Soc. Netw.*, vol. 47, pp. 167–188, 2016.

[25] L. E. Rocha, A. E. Thorson, R. Lambiotte, and F. Liljeros, "Respondent driven sampling bias induced by community structure and response rates in social networks," *J. R. Statist. Soc. A*, 2016. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/rssa.12180

[26] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, pp. 378–382, 2000.

[27] J. Wu, M. Barahona, Y.-J. Tan, and H.-Z. Deng, "Spectral measure of structural robustness in complex networks," *IEEE Trans. Syst. Man Cy. Part A*, vol. 41, pp. 1244–1252, 2011.

[28] E. Zio and G. Sansavini, "Vulnerability of smart grids with variable generation and consumption: a system of systems perspective," *IEEE Trans. Syst. Man Cy. Syst.*, vol. 43, pp. 477–487, 2013.

[29] M. A. Suresh, R. Stoleru, E. M. Zechman, and B. Shihada, "On event detection and localization in acyclic flow networks," *IEEE Trans. Syst. Man Cy. Syst.*, vol. 43, pp. 708–723, 2013.

[30] Y. Shang, "Unveiling robustness and heterogeneity through percolation triggered by random-link breakdown," *Phys. Rev. E*, vol. 90, p. 032820, 2014.

[31] S. Iyer, T. Killingback, B. Sundaram, and Z. Wang, "Attack robustness and centrality of complex networks," *PLoS ONE*, vol. 8, p. e59613, 2013.

[32] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts, "Network robustness and fragility: Percolation on random graphs," *Phys. Rev. Lett.*, vol. 85, no. 25, pp. 5468–5471, Dec 2000.

[33] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, "Catastrophic cascade of failures in interdependent networks," *Nature*, vol. 464, pp. 1025–1028, 2010.

[34] J. Li, J. Wu, Y. Li, H. Z. Deng, and Y. J. Tan, “Optimal attack strategy in random scale-free networks based on incomplete information,” *Chin. Phys. Lett.*, vol. 28, p. 068902, 2011.

[35] S. Y. Tan, J. Wu, L. Lü, M.-J. Li, and X. Lu, “Efficient network disintegration under incomplete information: the comic effect of link prediction,” *Sci. Rep.*, vol. 6, p. 22916, 2016.

[36] Y. Shang, “Robustness of scale-free networks under attack with tunable grey information,” *Europhys. Lett.*, vol. 95, p. 28005, 2011.

[37] L. K. Gallos, R. Cohen, P. Argyrakis, A. Bunde, and S. Havlin, “Stability and topology of scale-free networks under attack and defense strategies,” *Phys. Rev. Lett.*, vol. 94, p. 188701, 2005.

[38] Q. Zhu, Z. Zhu, Y. Wang, and H. Yu, “Fuzzy-information-based robustness of interconnected networks against attacks and failures,” *Physica A*, vol. 458, pp. 194–203, 2016.

[39] Y. Shang, “Vulnerability of networks: fractional percolation on random graphs,” *Phy. Rev. E*, vol. 89, p. 012813, 2014.

[40] S. Shao, X. Huang, H. E. Stanley, and S. Havlin, “Percolation of localized attack on complex networks,” *New J. Phys.*, vol. 17, p. 023049, 2015.

[41] Y. Berezin, A. Bashan, M. M. Danziger, D. Li, and S. Havlin, “Localized attacks on spatially embedded networks with dependencies,” *Sci. Rep.*, vol. 5, p. 8934, 2015.

[42] X. Yuan, Y. Dai, H. E. Stanley, and S. Havlin, “k-core percolation on complex networks: comparing random, localized, and targeted attacks,” *Phys. Rev. E*, vol. 93, p. 062302, 2016.

[43] Y. Shang, “Localized recovery of complex networks against failure,” *Sci. Rep.*, vol. 6, p. 30521, 2016.

[44] X. Huang, J. Gao, S. V. Buldyrev, S. Havlin, and H. E. Stanley, “Robustness of interdependent networks under targeted attack,” *Phys. Rev. E*, vol. 83, p. 065101(R), 2011.

[45] C. M. Schneider, A. A. Moreira, J. J. S. Andrade, S. Havlin, and H. J. Herrmann, “Mitigation of malicious attacks on networks,” *Proc. Natl. Acad. Sci. USA*, vol. 108, pp. 3838–3841, 2011.

[46] B. Bollobás, *Random Graphs*. Cambridge: Cambridge University Press, 2001.

[47] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, pp. 509–512, 1999.

[48] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, “Random graphs with arbitrary degree distributions and their applications,” *Phys. Rev. E*, vol. 64, no. 2, p. 026118, Jul 2001.

[49] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, “Cost-effective outbreak detection in networks,” in *Proc. 13th ACM SIGKDD*. New York: ACM, 2007, pp. 420–429.

[50] L. D. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas, “Characterization of complex networks: a survey of measurements,” *Adv. Phys.*, vol. 56, pp. 167–242, 2007.

[51] M. Piraveenan, G. Thedchanamoorthy, S. Uddin, and K. S. K. Chung, “Quantifying topological robustness of networks under sustained targeted attacks,” *Soc. Netw. Anal. Min.*, vol. 3, pp. 939–952, 2013.

[52] J. Shao, S. V. Buldyrev, L. A. Braunstein, S. Havlin, and H. E. Stanley, “Structure of shells in complex networks,” *Phys. Rev. E*, vol. 80, p. 036105, 2009.

[53] Y. Shang, “Degree distribution dynamics for disease spreading with individual awareness,” *J. Syst. Sci. Complex.*, vol. 28, pp. 96–104, 2015.

[54] E. Cho, S. A. Myers, and J. Leskovec, “Friendship and mobility: user movement in location-based social networks,” in *Proc. Int. Conf. Knowledge Discovery and Data Mining*. San Diego, CA: ACM, 2011, pp. 1082–1090.

[55] J. Leskovec, K. Lang, A. Dasgupta, and M. W. Mahoney, “Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters,” *Internet Math.*, vol. 6, pp. 29–123, 2009.

[56] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D’Eustachio, E. Schmidt, B. de Bono, B. Jassai, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein, “Reactome: a knowledgebase of biological pathways,” *Nucleic Acids Res.*, vol. 33, pp. D428–D432, 2005.

[57] J. Leskovec, J. Kleinberg, and C. Faloutsos, “Graph evolution: densification and shrinking diameters,” *ACM Trans. Knowl. Discov. Data*, vol. 1, pp. 1–40, 2007.

[58] T. Nie, Z. Guo, K. Zhao, and Z. M. Lu, “The dynamic correlation between degree and betweenness of complex network under attack,” *Physica A*, vol. 457, pp. 129–137, 2016.

PLACE
PHOTO
HERE

Yilun Shang received the B.S. and Ph.D. degrees in mathematics from Shanghai Jiao Tong University, Shanghai, China, in 2005 and 2010, respectively. He was a Postdoctoral Fellow with the Institute for Cyber Security, Department of Computer Science, University of Texas at San Antonio, San Antonio, TX, USA, from 2010 to 2013, the SUTD-MIT International Design Centre, Engineering Product Development Pillar, Singapore University of Technology and Design, Singapore, from 2013 to 2014, and the Einstein Institute of Mathematics, Hebrew University of Jerusalem, Jerusalem, Israel, in 2014. He was an International Visiting Fellow with the Department of Mathematical Sciences, University of Essex, Colchester, UK, in 2017. Currently, he is an Associate Professor with the School of Mathematical Sciences, Tongji University, Shanghai, China.

Dr. Shang’s research interests include the structure and dynamics of complex networks, multi-agent systems, applied probability, combinatorics, algorithms and computation. He received the 2016 Dimitrie Pompeiu Prize from the Section of Mathematics of Romanian Academy. He is an Invited Reviewer for Zentralblatt MATH (Germany) and Mathematical Reviews (USA). He is on the Editorial Boards of IEEE Access and PLoS ONE.