frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 11 October 2023
DOI 10.3389/fphy.2023.1245564

OPEN ACCESS

EDITED BY
Nuno Crokidakis,
Fluminense Federal University, Brazil

REVIEWED BY
Divya Sindhu Lekha,
Indian Institute of Information Technology, India
Jihui Han,
Zhengzhou University of Light Industry, China

*CORRESPONDENCE
Michele Bellingeri,
michele.bellingeri@unipr.it

RECEIVED 23 June 2023
ACCEPTED 22 September 2023
PUBLISHED 11 October 2023

CITATION
Bellingeri M, Turchetto M, Scotognella F, Alfieri R, Nguyen N-K-K, Nguyen Q and Cassi D (2023), Forecasting real-world complex networks' robustness to node attack using network structure indexes. *Front. Phys.* 11:1245564. doi: 10.3389/fphy.2023.1245564

Forecasting real-world complex networks' robustness to node attack using network structure indexes

Michele Bellingeri[1,2]*, Massimiliano Turchetto[1,2], Francesco Scotognella[3,4], Roberto Alfieri[1,2], Ngoc-Kim-Khanh Nguyen[5], Quang Nguyen[6,7] and Davide Cassi[1,2]

[1]Dipartimento di Scienze Matematiche, Fisiche e Informatiche, Università di Parma, Parma, Italy, [2]Istituto Nazionale di Fisica Nucleare, Gruppo Collegato di Parma, Parma, Italy, [3]Dipartimento di Fisica, Politecnico di Milano, Milano, Italy, [4]Center for Nano Science and Technology@PoliMi, Istituto Italiano di Tecnologia, Milan, Italy, [5]Faculty of Fundamental Sciences, Van Lang University, Ho Chi Minh City, Vietnam, [6]Department of Physics, International University, Ho Chi Minh City, Vietnam, [7]Vietnam National University, Ho Chi Minh City, Vietnam

In this study, we simulate the degree and betweenness node attack over a large set of 200 real-world networks from different areas of science. We perform an initial node attack approach, where the node centrality rank is computed at the beginning of the simulation, and it is not updated along the node removal process. We quantify the network damage by tracing the largest connected component (*LCC*) and evaluate the network robustness with the "percolation threshold q_c," i.e., the fraction of nodes removed, for which the size of the *LCC* is quasi-zero. We correlate q_c with 20 network structural indicators (NSIs) from the literature using single linear regression (SLR), multiple linear regression (MLR) models, and the Pearson correlation coefficient test. The NSIs cover most of the essential structural features proposed in network science to describe real-world networks. We find that the Estrada heterogeneity (*EH*) index, evaluating the degree difference of connected nodes, best predicts q_c. The *EH* index measures the network node degree heterogeneity based on the difference of functions of node degrees for all pairs of linked nodes. We find that the q_c value decreases as a function of the *EH* index, unveiling that heterogeneous real-world networks with a higher variance in the degree of connected nodes are more vulnerable to node attacks.

KEYWORDS

complex network, network robustness and resilience, machine learning, node attack sequence, statistical physics

1 Introduction

Networks can model many real-world complex systems, where nodes (vertices) represent the constituent components and links (edges) describe the relationships among the node components [1, 2]. A paramount issue in complex network science is to determine the robustness of the overall system to the failure or attack of its nodes [3–10]. On the other hand, the robustness in complex networks is a problem closely related to understanding which kind of node removal (attack) strategy is the most effective in damaging the network [3, 11–14]. The node attack may model different real-world problems of high interest, such

TABLE 1 Network structural indicator (NSI) list with a short definition and reference.

ID	Key	Full name	Formula	Definition	Reference
1	N	Node number		N is the number of nodes in the network	
2	L	Link number		L is the number of links in the network	
3	C	Connectance	$C = \frac{2L}{N(N-1)}$	L is the number of links, and N is the number of nodes	[15]
4	$\bar{k}$	Average node degree	$\bar{k} = \frac{1}{N}\sum_{i=1}^{i=N} k_i$	k_i is the degree of the node i, and N is the nodes' number	[1]
5	σ_k	Node degree standard deviation	$\sigma_k = \sqrt{\frac{\sum_{i=1}^{i=N}(k_i-\bar{k})^2}{N-1}}$	k_i is the degree of the node i, $\bar{k}$ is the average node degree, and N is the nodes' number	[49]
6	AH	Albertson index	$AH = \sum_{i,j\in L} \lvert k_i - k_j \rvert$	i, j is the link connecting nodes i and j, k_i is the degree of the node i, k_j is the degree of the node j, and L is the network link set.	[46]
7	nAH	Normalized Albertson index	$nAH = \frac{AH}{L}$	AH is the Albertson index, and L is the number of links	[49]
8	EH	Estrada heterogeneity index	$EH = \frac{\sum_{i,j\in L}(k_i^{-1/2}-k_j^{-1/2})^2}{N-2\sqrt{N-1}}$	i, j is the link connecting nodes i and j, k_i is the degree of the node i and k_j is the degree of the node j, L is the network link set, and N is the node number	[30]
9	A	Network assortativity	$A = \frac{1}{\sigma_q^2}\sum_{j,k\in N} jk(e_{jk} - q_j q_k)$	σ_q is the standard deviation of the excess degree distribution, e_{jk} is the fraction of links connecting nodes of degree j and k, and q_j and q_k are the excess degree of nodes of degrees j and k, respectively	[39]
10	$\bar{d}$	Average node distance	$\bar{d} = \frac{1}{N(N-1)}\sum_{i,j\in N, i\neq j} d_{ij}$	d_{ij} is the distance between nodes i and j, and N is the node number	[41]
11	Φ	Network eccentricity	$\Phi = \frac{1}{N}\sum_{i=1}^{i=N}\varepsilon(i)$	$\varepsilon(i)$ is the eccentricity of the node i, and N is the node number	[41]
12	D	Network diameter	$D = \max_{i,j\in N, i\neq j}(d_{ij})$	d_{ij} is the distance between i and j, and N the node number	[41]
13	π	Network radius	$\pi = \min_{i\in G}(\varepsilon(i))$	$\varepsilon(i)$ is the eccentricity of the node i	[41]
14	Eff	Network efficiency	$Eff = \frac{1}{N(N-1)}\sum_{i\neq j}\frac{1}{d_{ij}}$	d_{ij} is the distance between node i and node j, and N is the node number	[52]
15	T	Average node transitivity	$T = \frac{1}{N}\sum_{i=1}^{i=N}\tau_i$	τ_i is the transitivity of the node i, and N is the node number	[3]
16	B	Average node betweenness	$B = \frac{1}{N}\sum_{i=1}^{i=N} g(i)$	N is the number of nodes and $g(i)$ the betweenness of the node i	[53]
17	nB	Average normalized node betweenness	$nB = \frac{1}{N}\sum_{i=1}^{i=N} {}^n g(i)$	N is the number of nodes, and ${}^n g(i)$ is the normalized betweenness of the node i	[53]
18	Clo	Average node closeness	$Clo = \frac{1}{N}\sum_{i=1}^{i=N} C_i$	C_i is the closeness of the node i, and N is the node number	[54]
19	$nClo$	Average normalized node closeness	$nClo = \frac{1}{N}\sum_{i=1}^{i=N} nC_i$	nC_i is the normalized closeness of the node i, and N is the node number	[49]
20	Q	Network modularity	$Q = \frac{1}{2L}\sum_{i,j}(a_{ij} - \frac{k_i k_j}{2L})\delta(c_i c_j)$	L is the total number of links in the network; a_{ij} is the element i, j of the adjacency matrix, equal to 1 if i and j are connected, and 0 otherwise; k_i and k_j are the degrees of i and j, respectively; c_i and c_j are the modules (or community) of nodes i and j, respectively; and $\delta(x, y)$ is 1 if $x = y$ and 0 otherwise	[36]

as the nodes/species extinction in ecological networks [15–17], the aging of nodes/chromophores in the photosynthetic network [18], the vaccination of nodes/individuals in social networks [19–22], or the malfunctioning of nodes/routers in computer networks [23, 24].

Network robustness to node attack may change in real-world networks with different structures [11]. Iyer et al. [3] studied network robustness as a function of the node clustering coefficient (or node transitivity). This study demonstrates that networks with higher clustering coefficients are more robust, with the most critical effect for the node degree and node betweenness attack. Nguyen and Trang [25] studied the Facebook social network. They found that those networks with higher modularity, i.e., networks presenting communities of nodes that are highly connected among them, have lower robustness to node removal. Zhou et al. [26] observed that increasing the assortativity of a network makes the network more robust against node removal and the network less stable. Nguyen et al. [27] showed that machine learning approaches

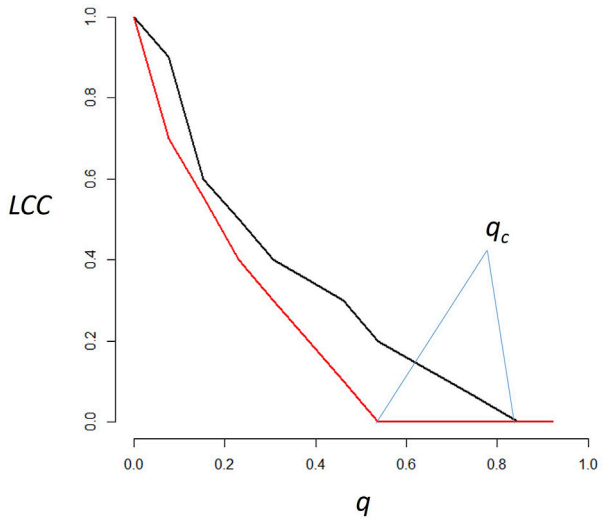

FIGURE 1
LCC as a function of the node removal fraction (q). The percolation threshold q_c value corresponds to the q-value at which *LCC* is quasi-zero. A higher percolation threshold q_c denotes a slower *LCC* decrease. Consequently, a higher percolation threshold q_c denotes a more robust network. The red line presents lower q_c, describing a more vulnerable network response to node attack than the black strategy. In other words, the black line denotes a more robust network response to a node attack.

unveil the degree assortativity, global closeness, and average node degree as the most critical factors in predicting the robustness (R) of real-world social networks.

Network science research shows contrasting outcomes about the role of the network structure in affecting its robustness to node attacks. On one hand, these studies are often based on small datasets of real-world networks, and they need more (robust) statistical analyses. On the other hand, research outcomes generally restrict the investigation, focusing on a few structural features of the networks, thus lacking a wide comparison of network structural indicators (NSIs) to forecast network robustness. For these reasons, understanding which structural features of real-world networks affect their robustness to node removal is still an urgent problem in network science.

In this research, we implement two well-known node attack strategies, i.e., the degree and betweenness node removal over a large set of 200 real-world networks from different areas of science.

We quantify the network functioning damage along the node attack sequence using the largest connected component (*LCC*) indicator [3, 11, 28]. To evaluate the network robustness against the node attack, we adopt the "percolation threshold" (q_c), i.e., the fraction of nodes removed at which the network becomes disconnected or, in other terms, the fraction of nodes removed for which the size of the *LCC* is *quasi*-zero [29].

Then, to understand how the network structure affects the network robustness (and the node attack efficacy), we correlate q_c with 20 NSIs from the literature. To study this correlation, we performed regression analysis, single linear regression (SLR), multiple linear regression (MLR) models, and the Pearson correlation coefficient test to find the best NSI predictors of the target variable q_c.

We find that the Estrada heterogeneity (*EH*) index [30] best predicts q_c in both the SLR and MLR models. The q_c value decreases as a function of the *EH* index. The *EH* index measures network degree heterogeneity based on the difference in functions of node degrees for all pairs of linked nodes [30]. This result indicates that the degree heterogeneity of linked nodes may negatively affect the real-world network robustness to node attack, specifically the network robustness against removing the most connected and highest betweenness nodes. Our outcomes shed light on the role of the real-world network structure in shaping their robustness and can help assemble more robust network structures.

2 Methods

2.1 The node attack strategies

We simulated two classic node attack (removal) strategies. The first is the removal of nodes according to their degree (*DEG*), i.e., the number of links to the node [3, 4, 31]. The *DEG* strategy removes nodes in decreasing order of connectivity, i.e., the most connected nodes (hubs) are removed first. The second node attack strategy removes nodes in decreasing order of betweenness centrality (*BET*) [3, 7, 32]. The betweenness centrality is a node centrality based on the shortest paths between node pairs (also called geodesic paths). The shortest path between two nodes is the minimum number of links required to travel from one node to another [33]. The betweenness centrality of a node returns the number of shortest paths from every node pair of the network passing along that node. The betweenness $g(i)$ of the node i is $(i) = \sum_{s,t=1}^{N} \frac{\sigma_{st}(i)}{\sigma_{st}}$, where σ_{st} is the total number of shortest paths between nodes s and t and $\sigma_{st}(i)$ is the number of these shortest paths passing through the node i, and N is the number of nodes.

We perform an "initial node attack approach," i.e., the node centrality rank is computed at the beginning of the simulation, and it is not updated along the node removal process [11]. The "initial node attack approach" differs from the recalculated (also named adaptive) node attack, in which node centralities are updated after node removals [11, 28]. The initial node attack describes the case where it is not possible to collect information about node features during the node removal process, such as vaccinating nodes/individuals in a social contact network with limited resources (limited time or vaccines) [34] or attacking nodes/routers in a computer network with a simultaneous node attack [28].

For both the node attack strategies, in the case of ties, i.e., nodes with equal ranking, we randomly sort their sequence. We perform 10^3 simulations for each node attack strategy. We implemented the node attack simulations using the igraph package of the R program. The simulations are carried out on the high-performance computing (HPC) cluster of the "Università degli Studi di Parma."

2.2 Real-world networks

We analyzed a large dataset of real-world network systems composed of 200 networks from different fields of science. The

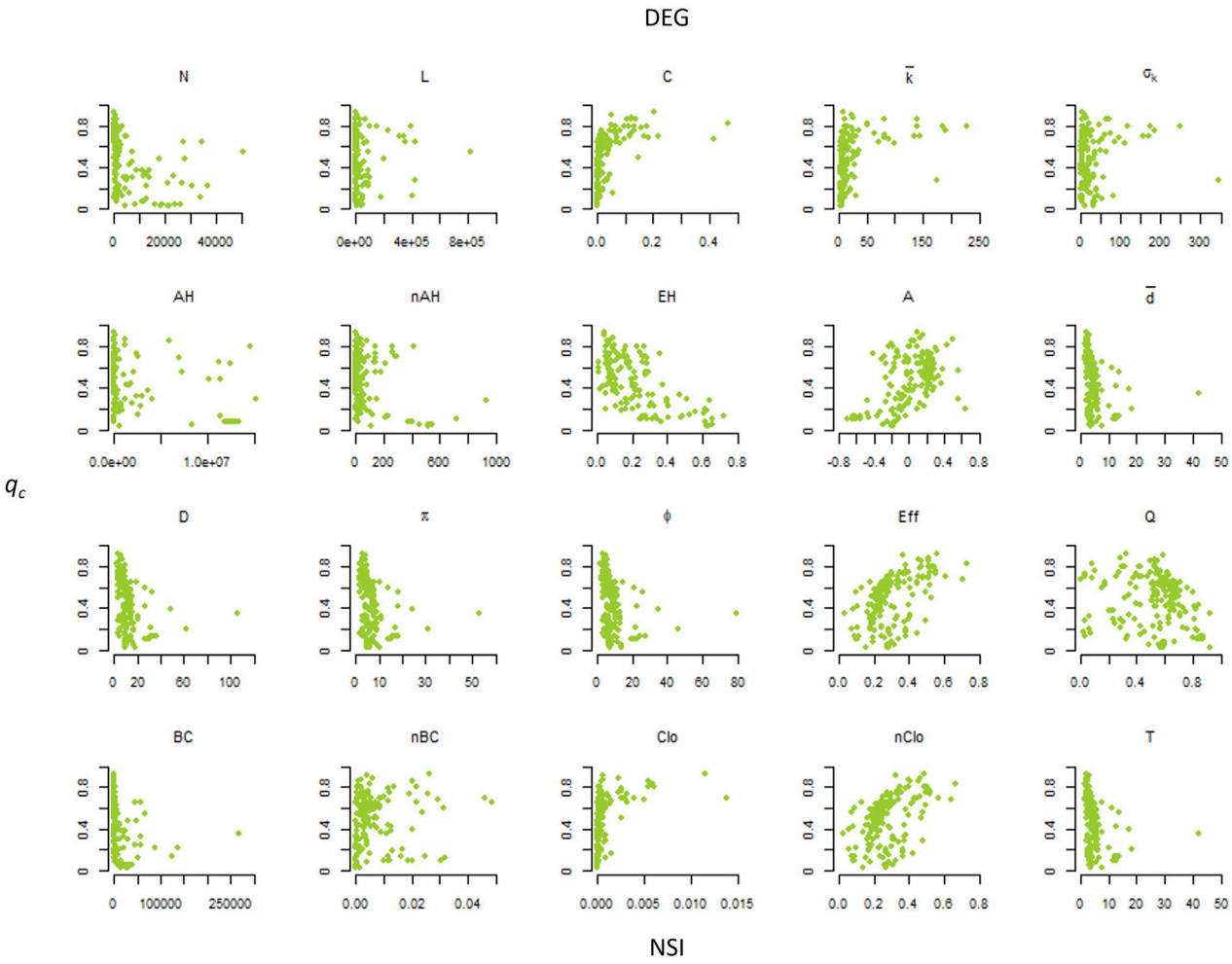

FIGURE 2
Scatterplots of the percolation threshold (q_c) vs. the network structural indicators (NSIs) for the *DEG* node attack strategy, removing nodes with higher degrees first.

real-world networks analyzed here come from social, biological, Internet, road, transportation, neuronal, and ecological networks. The networks analyzed here are undirected (i.e., do not account for link directionality) and unweighted (do not account for link weight). The number of network nodes ranges from $N = 25$ to $N = 75{,}811$; the average is $\bar{N} = 4{,}955.6$. The real-world network datasets analyzed in this study are available in the "Netzschleuder" repository [https://networks.skewed.de/], in the "Stanford Large Network Dataset Collection" repository [https://snap.stanford.edu/data/index.html], and in "the Colorado Index of Complex Networks (ICON)" repository [https://icon.colorado.edu/#!/]. The complete list of the real-world networks is provided in Supplementary Table A1 in Supplementary Appendix A1.

2.3 Network structure indexes

We considered 20 different NSIs from the network science literature, graph theory, and chemical graph theory to predict q_c in a large real-world network dataset. NSI adopted in this work covers most of the salient structural features of the real-world networks proposed in the network science literature, such as the node connectivity level [35], presence of a community structure [36, 37], degree heterogeneity [30, 38], node assortativity [39], node transitivity (or clustering) [3, 40], distance among nodes [41], and different notions of node centrality [42]. The list of NSIs is provided in Table 1.

2.4 The network robustness

To evaluate the networks' response to node attack, we trace *LCC* as a function of the fraction of nodes removed q. *LCC* (also named the giant component) is the maximum number of connected nodes [1]. In other terms, *LCC* is the maximal set of nodes in the network such that a path connects each node pair. *LCC* is the most commonly used measure to evaluate the network response to node removal [11]. Then, to evaluate the network robustness to node attack, we use q_c that represents the fraction of nodes to remove for reducing *LCC* to *quasi*-zero [29]. This work defines q_c as

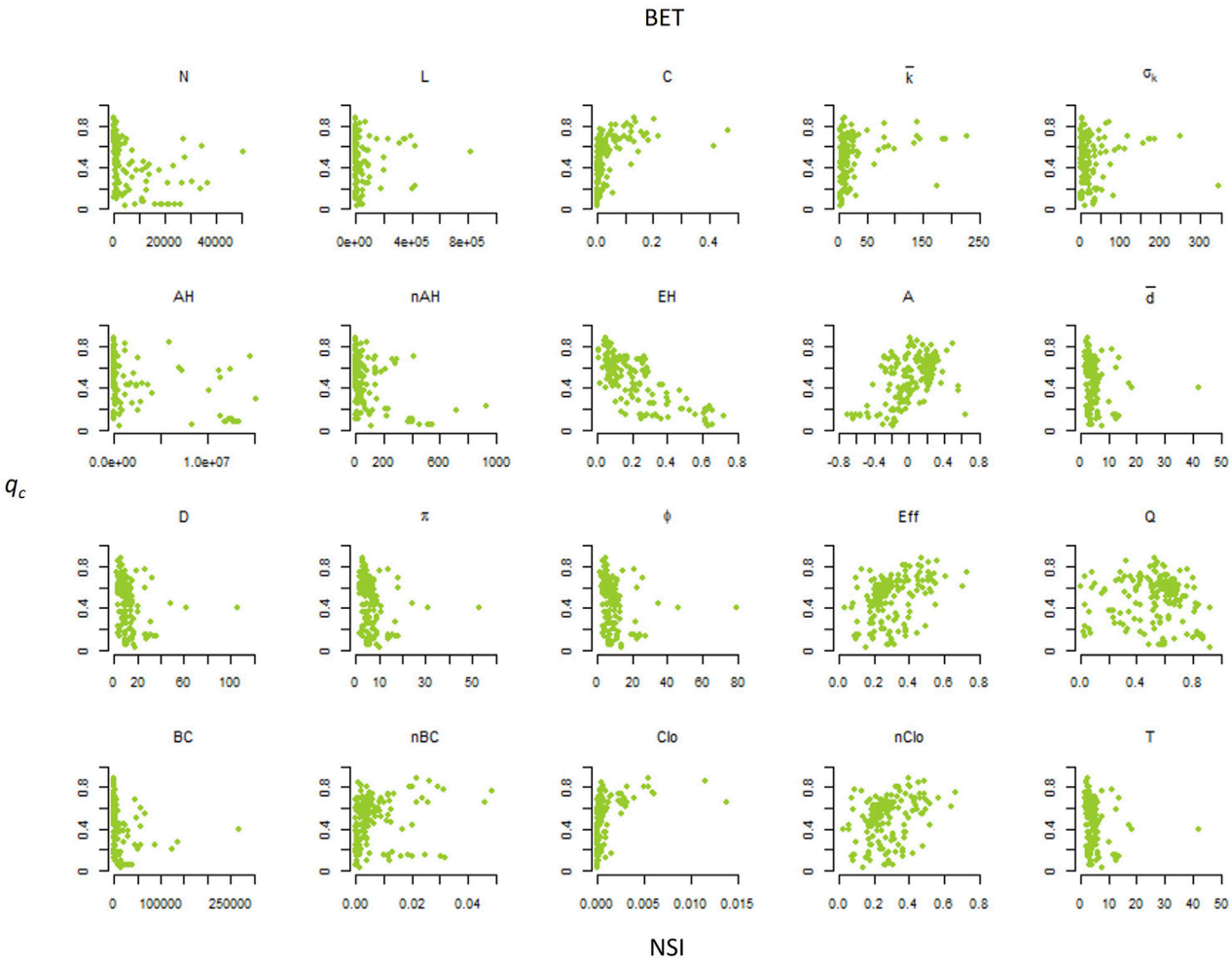

FIGURE 3
Scatterplots of the percolation threshold (q_c) vs. the network structural indicators (NSIs) for the *BET* node attack strategy, removing nodes with higher betweenness first.

the fraction q of nodes removed to reduce the *LCC* value equal to or lower to 0.05 of its initial size. The lower the q_c value, the lower the network robustness (Figure 1). Furthermore, the lower the q_c value, the higher the efficacy of the node attack strategies to dismantle the network [29].

2.5 The linear regression models

We perform regression model analyses to understand the relationship between NSI and the q_c value of the real-world networks. First, we perform SLR. The SLR model between q_c and an NSI x is expressed by the following linear equation:

$$q_c = a + b \cdot NSI, \tag{1}$$

where a is the intercept and b is the slope. We choose the one with the highest R-squared among the significant SLRs to evaluate the best SLR model and, consequently, the best predictor. In linear regression, R-squared (R^2), also named the coefficient of determination, measures how close the data points are to the fitted line. Higher R^2 denotes better regression fitting models [48].

Then, we perform MLR models. MLR is an extension of SLR for multi-dimension variables $x = (x_1, x_2, \ldots, x_n)$. The linear equation between the q_c value and NSIs becomes

$$q_c = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n, \tag{2}$$

where a_i are coefficients obtained performing the ordinary least square (OLS) method and $x_1, x_2, \ldots, x_n$ are NSIs. The a_i coefficients quantify the association between NSI (variable) and q_c (response). We interpret a_i as the average effect on q_c of a one-unit increase in NSI, holding all other NSI predictors fixed [48]. In practice, we often have more than one predictor, and the MLR model, differently from SLR, can directly accommodate multiple predictors. To evaluate the best predictor carried out by the MLR model, we choose the significant NSI with the highest absolute t-value. The t-value used in MLR is the t di-student statistic value from a two-sided t-test. The larger the absolute value of the t-test statistic, the less likely the results occurred by chance [48]. For this, larger absolute t-values are associated with better predictors (NSIs).

TABLE 2 Single linear regression model outcomes. The best significant predictor with the highest R^2 value is in bold.

	DEG				BET			
NSI	Intercept	Slope	*p*-value	R^2	Intercept	Slope	*p*-value	R^2
N	0.515	0.000	$<10^{-4}$ ***	0.149	0.519	0.000	$<10^{-4}$ ***	0.149
L	0.467	0.000	0.903	0.000	0.477	0.000	0.713	0.001
C	0.413	1.887	$<10^{-4}$ ***	0.217	0.431	1.502	$<10^{-4}$ ***	0.159
B	0.493	0.000	$<10^{-4}$ ***	0.079	0.497	0.000	<0.001**	0.070
nB	0.446	4.184	0.05	0.019	0.443	5.792	<0.05*	0.041
Clo	0.429	56.559	$<10^{-4}$ ***	0.146	0.441	49.661	$<10^{-4}$ ***	0.130
$nClo$	0.213	0.939	$<10^{-4}$ ***	0.195	0.305	0.628	$<10^{-4}$ ***	0.101
$\bar{d}$	0.558	−0.019	$<10^{-4}$ ***	0.089	0.536	−0.013	<0.05*	0.048
D	0.560	−0.008	$<10^{-4}$ ***	0.105	0.539	−0.005	<0.001**	0.059
π	0.565	−0.015	$<10^{-4}$ ***	0.094	0.540	−0.010	<0.05*	0.049
T	0.290	0.866	$<10^{-4}$ ***	0.330	0.330	0.703	$<10^{-4}$ ***	0.251
$\bar{k}$	0.424	0.002	$<10^{-4}$ ***	0.114	0.446	0.002	<0.001**	0.059
σ_k	0.460	0.000	0.401	0.004	0.479	0.000	0.642	0.001
A	0.454	0.520	$<10^{-4}$ ***	0.294	0.460	0.532	$<10^{-4}$ ***	0.355
Q	0.595	−0.234	<0.05*	0.037	0.530	−0.100	0.214	0.008
AH	0.471	0.000	0.449	0.003	0.480	0.000	0.185	0.009
EH	**0.665**	**−0.940**	**$<10^{-4}$** ***	**0.567**	**0.674**	**−0.952**	**$<10^{-4}$** ***	**0.671**
Φ	0.556	−0.010	$<10^{-4}$ ***	0.097	0.535	−0.007	$<10^{-4}$ ***	0.053
Eff	0.183	0.963	$<10^{-4}$ ***	0.237	0.274	0.679	$<10^{-4}$ ***	0.136
nAH	0.524	−0.001	$<10^{-4}$ ***	0.200	0.534	−0.001	0.000	0.261

We use the *lm* function of the R program to perform the SLR and MLR models. The fitting process is computed using the OLS method, which estimates the coefficients by minimizing an appropriate loss function [49].

Last, we perform the Pearson correlation coefficient (r) to test the goodness of the correlation between NSI and q_c. The r coefficient is the most common way of measuring the strength of a linear correlation [50]. It is a number between −1 and 1 that measures the strength and direction of the relationship between two variables. To evaluate the best correlation performed by the r coefficient, we choose the significant NSI with the highest absolute t-value. Last, we furnish the p-value to show the statistical significance of each model.

3 Results

Figure 2 shows the scatterplots of q_c vs. NSIs for the *DEG* node attack strategy. Figure 3 shows the scatterplots of q_c vs. NSIs for the *BET* node attack strategy.

Table 2 shows the outcomes of the SLR model. The best NSI to fit an SLR model with q_c is the *EH* index for both *DEG* (p-value $<10^{-4}$, R^2 = 0.567) and *BET* (p-value $<10^{-4}$, R^2 = 0.671) strategies. SLR $q_c \sim EH$ returns the lowest p-values and the highest R^2 for both node attack strategies (Table 2). The $q_c \sim EH$ fitting slopes are negative, indicating that q_c decreases as a function of *EH*, i.e., the robustness of the network is negatively correlated with *EH* for both node attack strategies (Figures 2, 3).

Table 3 shows the outcomes of the MLR model. The best NSI to predict q_c with the MLR model is the *EH* index for both *DEG* (t-value = −11.9, p-value $<10^{-23}$) and *BET* (t-value = −11.8, p-value $<10^{-23}$) strategies. MLR estimates a negative correlation between q_c and *EH* for both node attack strategies (negative correlation estimate, Table 3).

Table 4 summarizes the r coefficient test outcomes. The best NSI to correlate q_c is the *EH* index for both *DEG* (t-value = −11.9, p-value $<10^{-23}$) and *BET* (t-value = −11.8, p-value $<10^{-23}$) strategies. The r coefficient estimates a negative correlation between q_c and *EH* for both node attack strategies (−16.063 for *DEG* and −20.035 for *BET*, Table 4).

4 Discussion

The *EH* index is the best predictor of n q_c in our NSI set. Estrada [30] proposed the *EH* index as a unique characterization of network degree heterogeneity based on the difference in functions of node

TABLE 3 Multiple linear regression model outcomes. The best significant predictor with the highest absolute *t*-value is in bold.

	DEG			BET		
NSI	Estimate	t-value	*p*-value	Estimate	t-value	*p*-value
N	$1.36{\cdot}10^{-06}$	0.574	0.567	$1.326{\cdot}10^{-06}$	0.589	0.556
L	$6.537{\cdot}10^{-07}$	3.475	<0.001**	$6.145{\cdot}10^{-07}$	3.439	<0.001**
C	−0.767	−3.422	<0.001**	−0.614	−2.885	<0.05*
B	$-6.725{\cdot}10^{-07}$	−1.029	0.305	$-8.068{\cdot}10^{-07}$	−1.299	0.196
nB	−1.387	−0.975	0.331	−1.016	−0.752	0.453
Clo	−5.647	−0.927	0.355	−9.027	−1.560	0.121
$nClo$	−3.551	−3.552	<0.001**	−4.915	−5.173	$<10^{-4}$ ***
$\bar{d}$	0.036	3.188	<0.05*	0.029	2.788	<0.05*
D	−0.02	−2.705	<0.05*	−0.017	−2.378	<0.05*
π	0.0006	0.049	0.961	0.015	1.384	0.168
T	−0.199	−3.179	<0.05*	−0.315	−5.290	$<10^{-4}$ ***
$\bar{k}$	−0.0006	−0.757	0.449	−0.001	−1.493	0.137
σ_k	0.001	1.257	0.210	0.003	1.752	0.081
A	0.203	5.023	$<10^{-4}$ ***	0.222	5.789	$<10^{-4}$ ***
Q	0.015	0.283	0.778	0.004	0.085	0.932
AH	$-1.33{\cdot}10^{-09}$	−2.635	<0.05*	$-1.243{\cdot}10^{-09}$	−2.597	<0.05*
EH	**−0.915**	**−11.836**	**$<10^{-4}$ *** **	**−0.874**	**−11.904**	**$<10^{-4}$ *** **
Φ	0.0121	0.934	0.352	0.002	0.125	0.900
Eff	4.945	5.254	$<10^{-4}$ ***	6.033	6.747	$<10^{-4}$ ***
nAH	$-1.47{\cdot}10^{-07}$	−0.001	0.999	$-4.193{\cdot}10^{-05}$	−0.432	0.666
Intercept	$1.505{\cdot}10^{-01}$		0.089	$2.13{\cdot}10^{02}$		0.012
Outcome		RSE: 0.06 multiple R^2: 0.94 *p*-value: <0.001			RSE: 0.06 multiple R^2: 0.94 *p*-value	

degrees for all pairs of linked nodes. EH quantifies the degree heterogeneity of the network as a quadratic form of the Laplacian matrix of the network. It takes the value of zero if all nodes have the same degree as it happens in regular networks, and it is maximized when the difference of both degrees increases. The EH index has two bound or limit structures, i.e., it is equal to zero for any regular network (where all nodes present the same degree) and equal to one only for star graphs, i.e., networks in which N-1 nodes are directly connected to a single central node [30]. We find that q_c decreases as a function of the EH index (Figures 2, 3). This finding indicates that heterogeneous real-world networks with a higher variance in the degree of connected nodes are more vulnerable to node attacks.

EH is conceived as a refining of the Albertson index (AH), which computes the sum of the absolute value of the degree difference of the connected nodes [44]. The AH index, its normalized version nAH, and the node degree standard deviation σ_k are all indicators we used to quantify the network degree heterogeneity. The statistical analyses we performed, both SLR and MLR and the r coefficient test, indicate that these NSIs are not good predictors of q_c. σ_k did not return significant fittings for all statistical models (Tables 2–4). σ_k evaluates the whole node degree heterogeneity, neglecting whether the node degree variance is among connected nodes. Differently, the EH index measures the degree difference among connected nodes [30]. For this reason, we can argue that the node degree heterogeneity would play a significant role in affecting the network robustness only if the node degree heterogeneity is located (and evaluated) among connected nodes.

The third and fourth ring roads of Beijing City, the capital of China, are the real-world networks of the lowest EH in our dataset (EH = 0.008 and 0.009). In these networks, nodes represent the road intersections and links depict the roads connecting nodes [51]. The connected nodes present homogenous degrees, and for this reason, removing higher-degree road intersections would cause a slower network fragmentation with very high q_c values (q_c = 0.6 and 0.56), indicating lower network damage. On the contrary, the academia US faculty hiring network shows the highest EH value (EH = 0.73). In

TABLE 4 Pearson correlation coefficient test outcomes. The best significant predictor with the highest absolute *t*-value is in bold.

	DEG			BET		
NSI	Estimate	t-value	*p*-value	Estimate	t-value	*p*-value
N	−0.386	−5.876	$<10^{-4}$ ***	−0.386	−5.88	$<10^{-4}$ ***
L	0.009	0.123	0.903	−0.026	−0.369	0.713
C	0.466	7.388	$<10^{-4}$ ***	0.398	6.092	$<10^{-4}$ ***
B	−0.281	−4.102	$<10^{-4}$ ***	−0.264	−3.841	0.001
nB	0.137	1.928	0.06	0.203	2.896	0.004
Clo	0.382	5.805	$<10^{-4}$ ***	0.36	5.42	$<10^{-4}$ ***
$nClo$	0.442	6.917	$<10^{-4}$ ***	0.318	4.7	$<10^{-4}$ ***
$\bar{d}$	−0.299	−4.398	$<10^{-4}$ ***	−0.218	−3.141	0.002
D	−0.324	−4.807	$<10^{-4}$ ***	−0.243	−3.512	0.001
π	−0.307	−4.527	$<10^{-4}$ ***	−0.22	−3.17	0.002
T	0.575	9.859	$<10^{-4}$ ***	0.501	8.128	$<10^{-4}$ ***
$\bar{k}$	0.337	5.022	$<10^{-4}$ ***	0.243	3.514	0.001
σ_k	0.06	0.842	0.401	−0.033	−0.466	0.642
A	0.542	9.059	$<10^{-4}$ ***	0.596	10.408	$<10^{-4}$ ***
Q	−0.192	−2.752	0.007	−0.088	−1.246	0.214
AH	−0.054	−0.759	0.448	−0.094	−1.329	0.185
EH	**−0.753**	**−16.063**	**$<10^{-4}$ *** **	**−0.819**	**−20.035**	**$<10^{-4}$ *** **
Φ	−0.311	−4.599	$<10^{-4}$ ***	−0.23	−3.312	0.001
Eff	0.487	7.825	$<10^{-4}$ ***	0.369	5.569	$<10^{-4}$ ***
nAH	−0.447	−7.02	$<10^{-4}$ ***	−0.511	−8.349	$<10^{-4}$ ***

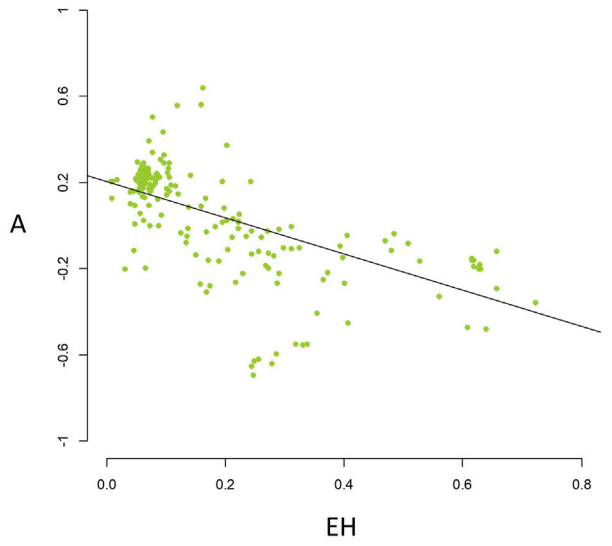

FIGURE 4
Scatterplot of the assortativity coefficient (*A*) vs. the Estrada heterogeneity (*EH*) index. The black line represents the significant linear regression $A = 0.2 - 0.84 \cdot EH$ (p-value < 0.001).

this network, a node is a Ph.D.-granting institution, and a link from node i to node j indicates that a person received their Ph.D. from node i and was tenure-track faculty at node j [52]. This network presents the highest degree heterogeneity of connected nodes, i.e., famous higher-degree nodes/institutions are connected with many lower-degree institutions. Therefore, the removal of the highest degree nodes, i.e., the removal of famous institutions sending many Ph.D. to other institutions, can cause a quick network disconnection. Therefore, the academia US faculty hiring network returns a lower q_c value (q_c = 0.13), indicating more significant network damage.

The nAH index is computed by averaging the original AH index over the number of links in the network. It can be viewed as the average degree difference among connected nodes [43]. nAH shows significant fitting for SLR (Table 2). Nonetheless, R^2 of SLR is much higher for EH than that for the nAH index (0.567 for EH and 0.200 for nAH, Table 2), indicating that the EH index can better explain the data. AH returns significant fitting for MLR (Table 3), but the absolute t-value for AH is much lower than that for EH (Table 3). Furthermore, AH did not return a significant r coefficient test (Table 4). These statistical results indicate that only EH correlates the nodes' degree heterogeneity of the networks with their robustness to the attack of connected nodes. On the other hand, these results suggest that networks presenting, on average, similar node degrees of the connected nodes should be robust to node attack. For this reason, networks of lower EH should show higher robustness to node attack and higher q_c.

The assortativity coefficient A (Table 1) measures how nodes tend to be connected with nodes of similar degrees [39]. "Assortative networks" present a preference for a network's nodes to attach to others with similar node degrees [39, 53]. On the contrary, a network is "disassortative" when, on average, high-degree nodes are connected to nodes with a lower degree, and on average, low-degree nodes are connected to nodes with a higher degree. Positive values of A indicate a correlation between nodes of similar degrees, while negative values indicate relationships between nodes of different degrees [39].

Given a certain node degree heterogeneity, assortative networks should have, on average, lower EH than disassortative networks. The linear regression $A \sim EH$ indicates a negative correlation (p-value < 0.001) in our real-world network dataset and confirms this hypothesis, i.e., higher values of A are associated with lower EH (Figure 4).

Consequently, assortative networks should show higher robustness to node attack and higher q_c. According to this hypothesis, we find that q_c increases as a function of A (Figures 2, 3), and all models SLR, MLR (Tables 2, 3 respectively), and the r coefficient (Table 4) return a positive significant fitting between A and q_c. The literature research results corroborate this finding, unveiling that increasing the assortativity of a network makes the network more robust against node removal [26], and a moderate assortativity increase positively affects the network's robustness against targeted node attacks [54]. Therefore, real-world networks with higher-degree differences of connected nodes are likely to present lower q_c.

To further investigate the relationship between node degree heterogeneity and network robustness, we perform an MLR model

TABLE 5 Multiple linear regression model $q_c \sim EH + Asso$

	DEG			BET		
NSI	Estimate	t-value	*p*-value	Estimate	t-value	*p*-value
A	0.093	1.584	0.115	0.103	2.175	0.031 *
EH	−0.862	−11.300	$<10^{-4}$ ***	−0.865	−14.051	$<10^{-4}$ ***
Outcome	RSE: 0.16 multiple R^2: 0.57 *p*-value: <0.001			RSE: 0.12 multiple R^2: 0.68 *p*-value: <0.001		

holding only *EH* and *A* as predictors of q_c, i.e., we fit the model $q_c \sim EH + A$. The outcomes of this analysis are shown in Table 5. *EH* is highly significant for the *DEG* strategy and presents the lowest *t*-value, whereas *A* is not a significant predictor. *EH* is highly significant for the *BET* strategy and presents a smaller *t*-value than *A*. This finding supports *EH* as NSI that can correlate with the real-world networks q_c.

5 Conclusion

Investigating node attack strategies provides valuable insights into enhancing network robustness by anticipating potential threats and identifying components that need protection. On the other side of the coin, node attack research plays a crucial role when the aim is to perform a fast network disruption, such as halting the spread of a disease or stopping the diffusion of a computer virus. Here, we investigate the relationship between the network structure and its robustness to node attack in a large dataset of real-world networks. Our results indicate that the degree heterogeneity of connected nodes negatively affects the network robustness. Specifically, the *EH* index evaluates the node degree heterogeneity, and it is the best predictor of q_c in our NSI set. This result unveils that heterogeneous real-world networks presenting higher differences in the degree of connected nodes are more vulnerable to node attacks. These results may help quantify real-world networked systems' robustness and build more robust networks.

This paper presents some limitations that may open new lines of research. First, we perform linear regression models only. The relationship between NSIs and the percolation threshold qc of the real-world networks may follow nonlinear models. Therefore, a natural extension of this research may consider nonlinear regression models, such as logistic, monomolecular, or exponential functions, to describe the relationship between the structure and the percolation threshold of real-world networks. Then, we adopt an initial node attack approach to study network robustness. Future research may analyze the robustness of real-world networks using recalculated node attacks, in which node ranking is updated after each node removal. Last, it would be interesting to investigate how NSIs correlate with other robustness indexes besides q_c, such as, for example, the network robustness index *R* robustness proposed by Schneider et al [55]. The *R* measurement considers the size of *LCC* during the whole node attack process not only at the point the network collapses. Therefore, adopting R may unveil a new correlation pattern between NSIs and network robustness.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Author contributions

MB, RA, and DC conceived the research. MB wrote the simulation codes. MB and RA performed the simulations. MB performed statistical analyses. All authors contributed to the article and approved the submitted version.

Funding

This research is funded by a grant from the Italian Ministry of Foreign Affairs and International Cooperation, by the Ecosister project, funded under the National Recovery and Resilience Plan (NRRP), and Mission 4 Component 2 Investment 1.5—Call for tender No. 3277 of 30/12/2021 of Italian Ministry of University and Research funded by the European Union—NextGenerationEU Award Number: Project code ECS00000033, Concession Decree No. 1052 of 23/06/2022 adopted by the Italian Ministry. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. (816313)). This work is supported by the Vietnam's Ministry of Science and Technology (MOST) under the Vietnam-Italy scientific and technological cooperation program for the period 2021–2023. This work is supported by the Vietnam National University Ho Chi Minh City (VNU-HCM), Ho Chi Minh City, Vietnam, under grant number B2018-42-01. This research is funded by a grant from the Italian Ministry of Foreign Affairs and International Cooperation.

Acknowledgments

MB, MT, DC, and RA acknowledge the Italian Ministry of Foreign Affairs and International Cooperation. The authors are greatly thankful to Van Lang University, Vietnam, for providing the budget for this study. This research has benefited from the high-performance computing (HPC) cluster of the Università degli Studi di Parma. They thank Fabio Sartori for the revision of the first manuscript draft. They also thank Prof. Stefano Poletti for the intriguing discussions about this research.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphy.2023.1245564/full#supplementary-material

References

1. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU. Complex networks: Structure and dynamics. *Phys Rep* (2006) 424:175–308. doi:10.1016/j.physrep.2005.10.009

2. Bellingeri M, Bevacqua D, Scotognella F, Alfieri R, Nguyen Q, Montepietra D, et al. Link and node removal in real social networks: A review. *Front Phys* (2020) 8:8. doi:10.3389/fphy.2020.00228

3. Iyer S, Killingback T, Sundaram B, Wang Z. Attack robustness and centrality of complex networks. *PLoS One* (2013) 8:e59613. doi:10.1371/journal.pone.0059613

4. Nguyen Q, Vu T, Dinh H, Cassi D, Scotognella F, Alfieri R, et al. Modularity affects the robustness of scale-free model and real-world social networks under betweenness and degree-based node attack. *Appl Netw Sci* (2021) 6:82. doi:10.1007/s41109-021-00426-y

5. Da Cunha BR, González-Avella JC, Gonçalves S. Fast fragmentation of networks using module-based attacks. *PLoS One* (2015) 10. doi:10.1371/journal.pone.0142824

6. Cerqueti R, Ciciretti R, Dalò A, Nicolosi M. A new measure of the resilience for networks of funds with applications to socially responsible investments. *Phys A Stat Mech Its Appl* (2022) 593:126976. doi:10.1016/j.physa.2022.126976

7. Lekha DS, Balakrishnan K. Central attacks in complex networks: A revisit with new fallback strategy. *Phys A Stat Mech Its Appl* (2020) 549:124347. doi:10.1016/j.physa.2020.124347

8. Shang Y. *Random lifts of graphs: Network robustness based on the Estrada index* (2012). Available at: http://www.math.nthu.edu.tw/amen/.

9. Shang YL. Local natural connectivity in complex networks. *Chin Phys Lett* (2011) 28:068903. doi:10.1088/0256-307X/28/6/068903

10. Shang Y. Biased edge failure in scale-free networks based on natural connectivity. *Indian J Phys* (2012) 86:485–8. doi:10.1007/s12648-012-0084-4

11. Wandelt S, Sun X, Feng D, Zanin M, Havlin S. A comparative analysis of approaches to network-dismantling. *Sci Rep* (2018) 8:13513. doi:10.1038/s41598-018-31902-8

12. Tian L, Bashan A, Shi DN, Liu YY. Articulation points in complex networks. *Nat Commun* (2017) 8:14223. doi:10.1038/ncomms14223

13. Bellingeri M, Lu ZM, Cassi D, Scotognella F. Analyses of the response of a complex weighted network to nodes removal strategies considering links weight: The case of the Beijing urban road system. *Mod Phys Lett B* (2018) 32:1850067–11. 1850067. doi:10.1142/S0217984918500677

14. Cuadra L, Salcedo-Sanz S, Del Ser J, Jiménez-Fernández S, Geem ZW. A critical review of robustness in power grids using complex networks concepts. *Energies* (2015) 8:9211–65. doi:10.3390/en8099211

15. Bellingeri M, Vincenzi S. Robustness of empirical food webs with varying consumer's sensitivities to loss of resources. *J Theor Biol* (2013) 333:18–26. doi:10.1016/j.jtbi.2013.04.033

16. Calizza E, Costantini ML, Rossi L. Effect of multiple disturbances on food web vulnerability to biodiversity loss in detritus-based systems. *Ecosphere* (2015) 6:art124–20. doi:10.1890/ES14-00489.1

17. Dunne Ja., Williams RJ, Martinez ND. Network structure and biodiversity loss in food webs: Robustness increases with connectance. *Ecol Lett* (2002) 5:558–67. doi:10.1046/j.1461-0248.2002.00354.x

18. Montepietra D, Bellingeri M, Ross AM, Scotognella F, Cassi D. Modelling photosystem I as a complex interacting network. *J R Soc Interf* (2020) 17:20200813. doi:10.1098/rsif.2020.0813

19. Sartori F, Turchetto M, Bellingeri M, Scotognella F, Alfieri R, Nguyen NKK, et al. A comparison of node vaccination strategies to halt SIR epidemic spreading in real-world complex networks. *Sci Rep* (2022) 12:21355. doi:10.1038/s41598-022-24652-1

20. Wang Z, Zhao DW, Wang L, Sun GQ, Jin Z. Immunity of multiplex networks via acquaintance vaccination. *EPL* (2015) 112:48002. doi:10.1209/0295-5075/112/48002

21. Gallos LK, Liljeros F, Argyrakis P, Bunde A, Havlin S. Improving immunization strategies. *Phys Rev E - Stat Nonlinear, Soft Matter Phys* (2007) 75:045104. doi:10.1103/PhysRevE.75.045104

22. Hartnett GS, Parker E, Gulden TR, Vardavas R, Kravitz D. Modelling the impact of social distancing and targeted vaccination on the spread of COVID-19 through a real city-scale contact network. *J Complex Networks* (2021) 9:cnab042. doi:10.1093/comnet/cnab042

23. Wang J, Jiang C, Qian J. Robustness of Internet under targeted attack: A cascading failure perspective. *J Netw Comput Appl* (2014) 40:97–104. doi:10.1016/j.jnca.2013.08.007

24. Doyle JC, Alderson DL, Li L, Low S, Roughan M, Shalunov S, et al. The "robust yet fragile" nature of the Internet. *Proc Natl Acad Sci U S A* (2005) 102:14497–502. doi:10.1073/pnas.0501426102

25. Nguyen Q, Le T-T. Structure and robustness of Facebook's pages networks. In: Proceeding of the 2019 the 10th conference on network modeling and analysis; November 06 - 08, 2019; Dijon, France (2019).

26. Zhou D, Stanley HE, D'Agostino G, Scala A. Assortativity decreases the robustness of interdependent networks. *Phys Rev E - Stat Nonlinear, Soft Matter Phys* (2012) 86:066103. doi:10.1103/PhysRevE.86.066103

27. Nguyen N-K-K, Nguyen Q, Pham H-H, Le T-T, Nguyen T-M, Cassi D, et al. Predicting the robustness of large real-world social networks using a machine learning model. *Complexity* (2022) 2022:1–16. doi:10.1155/2022/3616163

28. Bellingeri M, Cassi D, Vincenzi S. Efficiency of attack strategies on complex model and real-world networks. *Phys A Stat Mech Its Appl* (2014) 414:174–80. doi:10.1016/j.physa.2014.06.079

29. Holme P, Jun Kim B, No Yoon C, Kee Han S. *Attack vulnerability of complex networks* (2002).

30. Estrada E. Quantifying network heterogeneity. *Phys Rev E - Stat Nonlinear, Soft Matter Phys* (2010) 82:066102. doi:10.1103/PhysRevE.82.066102

31. Nie T, Guo Z, Zhao K, Lu ZM. New attack strategies for complex networks. *Phys A Stat Mech Its Appl* (2015) 424:248–53. doi:10.1016/j.physa.2015.01.004

32. Nguyen Q, Pham HD, Cassi D, Bellingeri M. Conditional attack strategy for real-world complex networks. *Phys A Stat Mech Its Appl* (2019) 530:121561. doi:10.1016/j.physa.2019.121561

33. Freeman HE. A set of measures of centrality based on betweenness. *Sociometry* (1977) 40:35. doi:10.2307/3033543

34. Sartori F, Turchetto M, Bellingeri M, Scotognella F, Alfieri R, Nguyen N-K-K, et al. A comparison of node vaccination strategies to halt SIR epidemic spreading in real-world complex networks. *Res Sq* (2022). doi:10.21203/rs.3.rs-1870717/v1

35. Albert R, Barabási A. Statistical mechanics of complex networks. *Rev Mod Phys* (2002) 74:47–97. doi:10.1103/revmodphys.74.47

36. Clauset C, Newman MJ, Moore C. Finding community structure in very large networks. *Phys Rev E* (2004) 70:066111. doi:10.1103/physreve.70.066111

37. Salathe M, James J. Dynamics and control of diseases in networks with community structure. *Plos Comput Biol* (2010) 6:e1000736. doi:10.1371/journal.pcbi.1000736

38. Estrada E. The many facets of the Estrada indices of graphs and networks. *Sema J* (2022) 79:57–125. doi:10.1007/s40324-021-00275-w

39. Noldus R, Mieghem PV. Assortativity in complex networks. *J Complex Networks* (2014) 3:507–42. doi:10.1093/comnet/cnv005

40. Gleeson JP, Melnik S, Hackett A. How clustering affects the bond percolation threshold in complex networks. *Phys Rev E - Stat Nonlinear, Soft Matter Phys* (2010) 81:066114. doi:10.1103/PhysRevE.81.066114

41. Buckley F, Harary F. *Distance in graphs*. Redwood City, CA: Addison-Wesley Publishing Company (1990). doi:10.1201/b16132-64

42. Lü L, Chen D, Ren XL, Zhang QM, Zhang YC, Zhou T. Vital nodes identification in complex networks. *Phys Rep* (2016) 650:1–63. doi:10.1016/j.physrep.2016.06.007

43. Bellingeri M, Bevacqua D, Turchetto M, Scotognella F, Alfieri R, Nguyen NKK, et al. Network structure indexes to forecast epidemic spreading in real-world complex networks. *Front Phys* (2022) 10:10. doi:10.3389/fphy.2022.1017015

44. Albertson MO. The irregularity of a graph. *Ars Comb* (1997) 46:219–25.

45. Latora V, Marchiori M. Efficient behavior of small-world networks. *Phys Rev Lett* (2001) 87:198701––4. doi:10.1103/PhysRevLett.87.198701

46. Barthélemy M. Betweenness centrality in large complex networks. *Eur Phys J B* (2004) 163–8. doi:10.1140/epjb/e2004-00111-4

47. Rochat Y. Closeness centrality extended to unconnected graphs: The harmonic centrality index. *Appl Soc Netw Anal* (2009) 117.

48. James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning with applications in R*. Cham: Springer (2013).

49. Goldberger AS. *Econometric theory*. New Jersey, United States: Wiley (1964).

50. Schober P, Schwarte LA. Correlation coefficients: Appropriate use and interpretation. *Anesth Analg* (2018) 126:1763–8. doi:10.1213/ANE.0000000000002864

51. Bellingeri M, Bevacqua D, Scotognella F, Lu Z-M, Cassi D. Efficacy of local attack strategies on the Beijing road complex weighted network. *Phys A Stat Mech Its Appl* (2018) 510:316–28. doi:10.1016/j.physa.2018.06.127

52. Wapman KH, Zhang S, Clauset A, Larremore DB. Quantifying hierarchy and dynamics in US faculty hiring and retention. *Nature* (2022) 610:120–7. doi:10.1038/s41586-022-05222-x

53. Newman MEJ. Mixing patterns in networks. *Phys Rev E - Stat Physics, Plasmas Fluids Relat Interdiscip Top* (2003) 67:026126. doi:10.1103/PhysRevE.67.026126

54. Trajanovski S, Martín-Hernández J, Winterbach W, Van Mieghem P. Robustness envelopes of networks. *J Complex Networks* (2013) 1:44–62. doi:10.1093/comnet/cnt004

55. Schneider CM, Moreira AA, Andrade JS, Jr, Havlin S, Herrmann HJ. Mitigation of malicious attacks on networks. *Proc Natl Acad Sci* (2011) 108(10):3838–41. doi:10.1073/pnas.1009440108