PHYSICAL REVIEW E **91**, 010801(R) (2015)

Predicting percolation thresholds in networks

Filippo Radicchi*
Center for Complex Networks and Systems Research, School of Informatics and Computing, Indiana University, Bloomington, Indiana 47405, USA
(Received 19 November 2014; published 15 January 2015)

We consider different methods, which do not rely on numerical simulations of the percolation process, to approximate percolation thresholds in networks. We perform a systematic analysis on synthetic graphs and a collection of 109 real networks to quantify their effectiveness and reliability as prediction tools. Our study reveals that the inverse of the largest eigenvalue of the nonbacktracking matrix of the graph often provides a tight lower bound for true percolation threshold. However, in more than 40% of the cases, this indicator is less predictive than the naive expectation value based solely on the moments of the degree distribution. We find that the performance of all indicators becomes worse as the value of the true percolation threshold grows. Thus, none of them represents a good proxy for the robustness of extremely fragile networks.

DOI: 10.1103/PhysRevE.91.010801 PACS number(s): 89.75.Hc, 64.60.aq

Percolation is one of the most studied processes in statistical physics [1]. The model assumes the presence of an underlying network structure, where nodes (site percolation) or edges (bond percolation) are independently occupied with probability p [2,3]. Nearest-neighbor occupied sites or bonds form clusters. For $p = 0$, only clusters of size one are present in the system. For $p = 1$ instead, a unique giant cluster spans the entire network. At intermediate values of p, the network can be found in two different phases: the nonpercolating regime, where clusters have microscopic size, in the sense that the number of nodes within each cluster is much smaller than the size of the network, and the percolating phase, where a single macroscopic cluster, whose size is comparable to the one of the entire network, is present. The value of p that separates the two phases is a network-dependent quantity called percolation threshold and it is usually denoted by p_c. Percolation models are commonly used to study network robustness against random failures [4,5] and the spreading of diseases or ideas [6,7]. In practical applications, knowing the value of the percolation threshold of a given network is thus extremely important. For example, threshold values of technological or infrastructural networks can be used to estimate the maximal number of local failures that these systems can tolerate before stopping to function [8]. In the case of networks of social contacts, the value of the percolation threshold can be interpreted as a proxy for the risk to observe disease outbreaks [9]. Unfortunately, there are only a few methods, which do not rely on direct numerical simulations of the percolation process, able to determine the value of the percolation threshold in a network [10]. In this paper we consider three different indicators that have been recently introduced to provide estimates of bond percolation thresholds in arbitrary networks. We first measure their performances in synthetic graphs and then test their predictive power in a large and heterogeneous set of real-world networks.

To this end, we first need to provide the right term of comparison, i.e., the best estimate of the true threshold by means of direct numerical simulations of the percolation process. Given an undirected and unweighted network with N nodes and E edges composed of a single connected component, we study bond percolation using the Monte Carlo method proposed by Newman and Ziff [11]. In each realization of the method, we start from a configuration with no connections. We then sequentially add edges in random order and monitor the evolution of the size of the largest cluster in the network $S(p)$ as a function of the bond occupation probability $p = e/E$, where e indicates the number of edges added from the initial configuration, i.e., $e = 0$. We repeat the entire process Q independent times and estimate the percolation strength as

$$P_\infty(p) = \frac{1}{NQ} \sum_{q=1}^{Q} S_q(p) \tag{1}$$

and the susceptibility as

$$\chi(p) = \frac{(1/N^2 Q) \sum_{q=1}^{Q} S_q(p) S_q(p) - [P_\infty(p)]^2}{P_\infty(p)}, \tag{2}$$

where $S_q(p)$ indicates the size of the largest cluster in the network observed, during the qth realization of the Monte Carlo algorithm, when the bond occupation probability equals p. We determine the best estimate of the percolation threshold p_c as the value of p where the susceptibility reaches its maximum

$$p_c = \arg \left\{ \max_p \chi(p) \right\}. \tag{3}$$

A concrete example of the application of the method described above to a real network is provided in Fig. 1. Note that Eq. (3) is not the only possible way to compute the best estimate of the true percolation threshold. Another very popular method is to determine p_c as the value of p where the size of the second largest cluster reaches its maximum [10]. Generally, the two definitions do not provide the same exact value for the best estimates, but values whose difference is within the tolerance required for the results of this paper.

As stated above, the main purpose of our paper is to understand the best way to predict the percolation threshold of Eq. (3) without performing direct numerical simulations. We consider three different indicators. Our first estimation is

*filiradi@indiana.edu

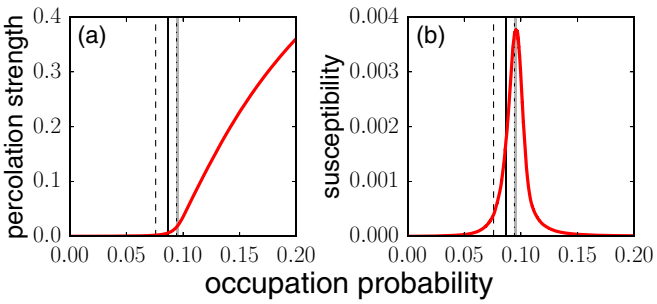

FIG. 1. (Color online) Bond percolation transition of the giant connected component of the *peer-to-peer Gnutella* network as of August 31, 2002 [12,13]. We consider the undirected version of the network, originally directed. The largest connected component is composed of $N = 62\,561$ nodes and $E = 147\,878$. (a) Percolation strength P_∞ as a function of the bond occupation probability p [Eq. (1)]. (b) Susceptibility χ as a function of the bond occupation probability p [Eq. (2)]. Here P_∞ and χ are represented as red thick lines and have been computed with $Q = 10\,000$ independent realizations of the Monte Carlo algorithm by Newman and Ziff [11]. The best estimate of the percolation threshold $p_c = 0.0955$ [Eq. (3), gray solid line] is compared with the estimations $\tilde{p}_c = 0.0943$ [Eq. (4), black dotted line], $\bar{p}_c = 0.0759$ [Eq. (5), dashed black line], and $\hat{p}_c = 0.0871$ [Eq. (6), black solid line].

based on the quantity

$$\tilde{p}_c = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}. \tag{4}$$

In Eq. (4), $\langle k \rangle$ and $\langle k^2 \rangle$ respectively represent the first and the second moments of the degree distribution of the graph and $\tilde{p}_c$ represents the value of the percolation threshold expected in uncorrelated graphs with prescribed degree sequence [5,14]. In general, such a prediction should fail in real networks where correlations are present. As a second prediction method of the percolation threshold we consider the inverse of the largest eigenvalue of the adjacency matrix A of the graph

$$\bar{p}_c = \left[\max_{\vec{v}} \frac{\vec{v}^T A \vec{v}}{\vec{v}^T \vec{v}} \right]^{-1}. \tag{5}$$

This is the expected value of the percolation threshold in graphs with a high density of connections [15]. Given that real networks are generally sparse, we expect therefore that this method is not able to produce very good prediction values of the percolation threshold. Finally, we provide an estimation of the percolation threshold as

$$\hat{p}_c = \left[\max_{\vec{w}} \frac{\vec{w}^T M \vec{w}}{\vec{w}^T \vec{w}} \right]^{-1}, \tag{6}$$

where the matrix M is defined as

$$M = \begin{pmatrix} A & \mathbb{1} - D \\ \mathbb{1} & \emptyset \end{pmatrix}. \tag{7}$$

Here M is a square matrix of dimension $2N \times 2N$ composed of four $N \times N$ blocks, A is the adjacency matrix of the graph, $\mathbb{1}$ is the identity matrix, and D is a diagonal matrix whose elements are equal to the degree of the nodes. The largest eigenvalue of M is identical to the largest eigenvalue of the so-called nonbacktracking or Hashimoto matrix associated with the graph [16]. Spectral properties of this matrix are relevant not just in percolation theory, but also for clustering algorithms and centrality measures [17,18]. The indicator of Eq. (6) is obtained in the approximation of locally treelike networks and it is therefore expected to provide good predictions for sparse networks [10,19].

First, we perform a systematic study of synthetic networks. Each network realization is obtained with the use of the so-called uncorrelated configuration model [20]. This is a variation of the model by Molloy and Reed to generate simple (i.e., without multiple or self-connections) random graphs with arbitrary degree distributions [21]. In our numerical experiments, we extract degrees from the power-law probability distribution $P(k) \sim k^{-\gamma}$ if $k \in [3, \sqrt{N}]$ and $P(k) = 0$ otherwise. Setting the degree cutoff at $\sqrt{N}$ ensures the absence of degree-degree correlations [22] and the correct behavior of the moments of the degree distribution [23]. We consider six different values of the degree exponent γ, ranging from 2.1 to 100.0, and study the performance of $\tilde{p}_c$, $\bar{p}_c$, and $\hat{p}_c$ as functions of the network size N (see Fig. 2). We note that for every network instance, all these quantities always provide a value smaller than p_c. We quantify the performance of a given predictor by measuring the difference between predicted value and best estimate p_c. The inverse of the largest eigenvalue of the adjacency matrix $\bar{p}_c$ provides reasonable predictions for the percolation threshold only for values of $\gamma < 3$. In that regime, the gap $p_c - \bar{p}_c$ goes to zero, as N increases, in a power-law fashion. For larger values of the degree exponent instead, the gap stabilizes to finite values even in the limit of infinitely large networks. Instead $\tilde{p}_c$ and $\hat{p}_c$ have very good performances for any value of γ. Their gaps with respect to p_c always tend to zero as the system size grows. The scaling of the gaps is well fitted by power-law functions having similar exponent values. Although $\hat{p}_c$ is always closer to p_c than $\tilde{p}_c$, the two measures essentially provide equivalent estimates of the percolation threshold on sparse random graphs. This type of result can be understood with an intuitive mathematical argument, according to which we rewrite the eigenvalue problem for the matrix M of Eq. (7) as $(\vec{w}_1^T, \vec{w}_2^T) M^T = \lambda (\vec{w}_1^T, \vec{w}_2^T)$, where $\vec{w}_1$ and $\vec{w}_2$ respectively contain the first and the second N components of the eigenvector $\vec{w}$. We can now write the eigenvalue problem as two coupled equations $A\vec{w}_1 + (\mathbb{1} - D)\vec{w}_2 = \lambda \vec{w}_1$ and $\vec{w}_1 = \lambda \vec{w}_2$. Multiplying the first equation by the row vector $\vec{u}^T = (1, \ldots, 1)$, we obtain

$$\lambda = \frac{\vec{d}^T \vec{w}_1}{\vec{u}^T \vec{w}_1} - 1, \tag{8}$$

where the components of the vector $\vec{d}$ are equal to the node degrees. In a random uncorrelated network, we should expect the nonbacktracking centrality of the nodes to be proportional to their degrees, i.e., $\vec{w}_1 \sim \vec{d}$, and so Eq. (6) reduces to Eq. (4) [18].

Second and more importantly, we perform a systematic study of the predictive power of $\tilde{p}_c$, $\bar{p}_c$, and $\hat{p}_c$ in a collection of 109 real networks. We consider graphs of heterogeneous nature, including biological, infrastructural, information, technological, social, and communication networks, and thus very variegated also in terms of structural properties (e.g.,

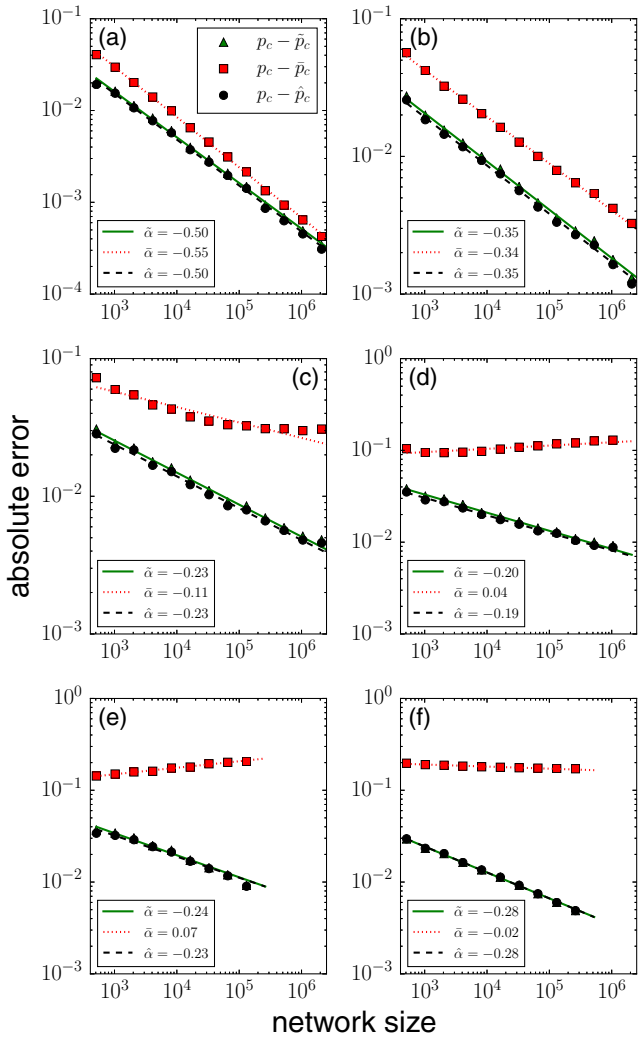

FIG. 2. (Color online) Difference between best estimates of the bond percolation threshold p_c and the predicted values $\tilde{p}_c$ [Eq. (4), green triangles], $\bar{p}_c$ [Eq. (5), red squares], and $\hat{p}_c$ [Eq. (6), black circles] in the uncorrelated configuration model [20]. Gaps between best estimates and predictions are plotted as a function of the network size N. We report results obtained for networks generated according to the power-law degree distribution $P(k) \sim k^{-\gamma}$, with support $[3, \sqrt{N}]$. Different panels correspond to different choices of the degree exponent γ: (a) $\gamma = 2.1$, (b) $\gamma = 2.5$, (c) $\gamma = 2.9$, (d) $\gamma = 3.5$, (e) $\gamma = 4.5$, and (f) $\gamma = 100.0$. Points are obtained by averaging over at least 50 independent network instances. Standard errors on the measures are smaller than the size of the symbols. For each network, the best estimate of the bond percolation threshold is obtained with $Q = 100$ Monte Carlo realizations of the Newman-Ziff algorithm [11]. In the various panels, lines correspond to best fits of the empirical points with the functions $p_c - \tilde{p}_c \sim N^{\tilde{\alpha}}$ (green solid lines), $p_c - \bar{p}_c \sim N^{\bar{\alpha}}$ (red dotted lines), and $p_c - \hat{p}_c \sim N^{\hat{\alpha}}$ (black dashed lines).

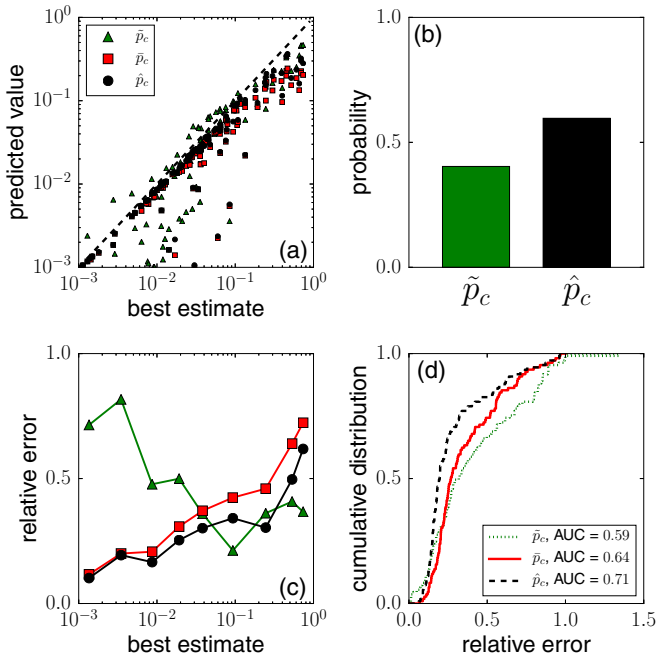

FIG. 3. (Color online) Comparison between best estimates and predicted values of the percolation threshold of 109 real networks. In each network, the best estimate p_c of the percolation threshold [Eq. (3)] is computed with $Q = 10\,000$ independent realizations of the Monte Carlo method by Newman and Ziff [11]. (a) For each network, we plot $\tilde{p}_c$ [Eq. (4), green triangles], $\bar{p}_c$ [Eq. (5), red squares], and $\hat{p}_c$ [Eq. (6), black circles] as functions of the best estimate p_c. The black dashed line represents perfect agreement between predicted values and best estimates. (b) Probability, computed over the entire sample of real networks, that either $\tilde{p}_c$ (green) or $\hat{p}_c$ assumes the closest value to p_c. (c) Relative error of the three predictors, with respect to the value of the best estimate, as a function of p_c. We use the same symbols and colors as those of (a). (d) Cumulative distribution of the relative error between predictions and best estimates. The green dotted line refers to $\tilde{p}_c$, the red solid line to $\bar{p}_c$, and the black dashed line to $\hat{p}_c$. Global performances of the prediction methods are measured in terms of the area under the curve (AUC).

degree distribution and correlations, clustering coefficient, and diameter). In our numerical study, we reduce, if necessary, weighted and/or directed networks to their unweighted and undirected projections. Also, we focus our attention only on their giant connected components. Results for all 109 real networks are provided in [24]. In Fig. 3 we summarize the main outcome of our analysis. As expected, both $\bar{p}_c$ and $\hat{p}_c$ always provide a lower bound for p_c [10]. Since we have by definition $\hat{p}_c \geqslant \bar{p}_c$, it is not a big surprise to find that $\hat{p}_c$ generates always better predictions than $\bar{p}_c$ [10]. It is however worth remarking that, whereas in synthetic networks the difference between the inverse of the largest eigenvalue of the adjacency matrices $\bar{p}_c$ and $\hat{p}_c$ can be very large, in real networks, the two indicators generate pretty similar estimations of the percolation threshold, suggesting relatively little advantage in using $\hat{p}_c$ in place of $\bar{p}_c$ (see [24]). It is interesting to observe that, in about the 40% of the real networks analyzed, the naive estimator $\tilde{p}_c$, based only on the fraction between first and second moments on the degree distribution, outperforms $\hat{p}_c$ in spite of using much less topological information. On the other hand, the use of $\tilde{p}_c$ as a prediction tool has the disadvantage of alternatively overestimating or underestimating the value of the percolation threshold. The relative error of $\hat{p}_c$ is an increasing function of the percolation threshold p_c; $\tilde{p}_c$ approximates better p_c as the percolation threshold increases. Also, the performances of the indicators do not seem to be

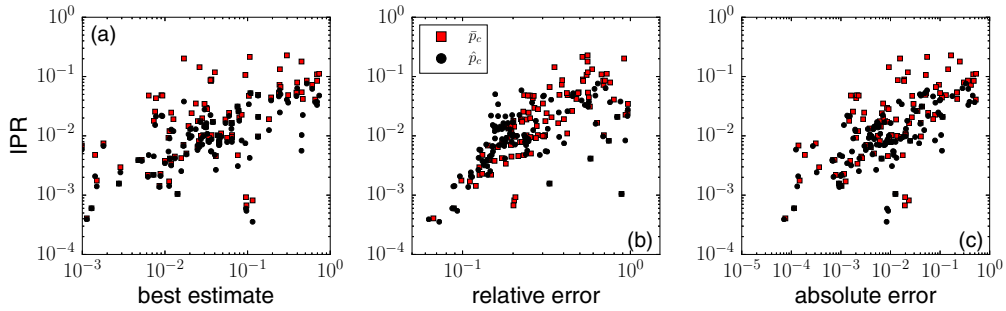

FIG. 4. (Color online) Inverse participation ratio (IPR) of the principal eigenvectors of the adjacency (red squares) and the nonbacktracking matrices [Eq. (7), black circles] as functions of the value of (a) the best estimate of the percolation threshold p_c, (b) the relative error, and (c) the absolute error. Each point refers to numerical results obtained on one of the 109 real networks considered in our analysis.

strongly related to the edge density of graph (see [24]). Overall, the predictive power of $\tilde{p}_c$ is lower than the one of $\hat{p}_c$, as $\tilde{p}_c$ often predicts too low values of the percolation threshold.

On the basis of our results, we can conclude that the inverse of the largest eigenvalue of the nonbacktracking matrix $\hat{p}_c$ represents the best measure currently available on the market, among those that do not rely on direct numerical simulations of the percolation process, to provide estimates of the bond percolation threshold p_c in an arbitrary network. Apparently, its predictive power is not much influenced by the density of edges, as long as the network is sufficiently sparse. Performances are instead strongly dependent on the value of the true percolation threshold. If threshold values are sufficiently small, $\hat{p}_c$ generates particularly good prediction values for bond percolation thresholds in networks. On the other hand, if the true percolation threshold is large, $\hat{p}_c$ fails by non-negligible amounts. This fact substantially happens for infrastructural networks, such as road networks [25] and power grids [26], off-line social networks [27–29], and biological networks [29] (see [24]). Also, this fact often happens in networks with double percolation transitions [30], where $\hat{p}_c$ tends to predict the lower threshold that does not necessarily coincide with the largest jump of the percolation strength.

The inadequacy of the indicator $\hat{p}_c$ to provide good predictions for p_c in the regime of large percolation thresholds may be related to the localization of the eigenvector associated with the principal eigenvalue of the matrix defined in Eq. (7). Since the second N components of any eigenvector of the matrix M are proportional to the first N, we quantify the localization of such eigenvector by normalizing the first N components (i.e., the sum of their squares equals one) and then computing its inverse participation ratio, defined as the sum of the fourth power of the components. Our measures performed on real networks indicate the presence of a positive correlation between the value of the inverse participation ratio and the value of the best estimate of the percolation threshold p_c [Fig. 4(a)]. They also show a positive correlation between the inverse participation ratio and the relative and absolute errors of $\hat{p}_c$ with respect to p_c [Figs. 4(b) and 4(c)]. Thus the reason for the bad performance of $\hat{p}_c$ seems analogous to the one valid for spectral estimators of epidemic thresholds in real networks [31]. Mathematically speaking, when the principal eigenvector is localized, the percolation transition placed at $\hat{p}_c$ involves in reality only a finite fraction of the nodes in the network and thus does not correspond to the true percolation transition located instead at p_c. The same type of considerations stressed for the principal eigenvector of the nonbacktracking matrix can be also extended to the principal eigenvector of the adjacency matrix to motivate why $\bar{p}_c$ suffers from the same limitations as those observed for $\hat{p}_c$ (Fig. 4).

As a final consideration, we would like to stress that the existence of a positive relation between the error committed by $\hat{p}_c$ to predict p_c and the value of p_c itself is an important limitation for the use of $\hat{p}_c$ as a proxy for network robustness. In this context, percolation is better thought of as removing edges rather than adding, so the robustness of a network is given by $1 - p_c$. The prediction $1 - \hat{p}_c$ always overestimates the true fraction of edge failures that a system can tolerate, therefore providing information not useful in preventing catastrophic events. Even more importantly, in the analysis of network resilience to random attacks, it is generally more meaningful to consider site percolation instead of bond percolation. Site percolation thresholds are larger than those valid for the edges (see [24]) and $1 - \hat{p}_c$ seems unable to provide sufficiently accurate predictions for the maximal number of vertices that can be removed from a network before reaching system failure. For all these reasons, we believe that additional theoretical efforts must be devoted to the search of good predictors able to determine tight upper bounds of percolation thresholds in networks.

The author thanks M. Boguña and S. Colomer-de-Simón for comments and suggestions. The author is indebted to A. Flammini and C. Castellano for fundamental discussions on the subject of this paper.

[1] D. Stauffer and A. Aharony, *Introduction to Percolation Theory* (Taylor & Francis, London, 1991).

[2] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[3] S. N. Dorogovtsev, A. V. Goltsev, and J. F. Mendes, Rev. Mod. Phys. **80**, 1275 (2008).

[4] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **406**, 378 (2000).

[5] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, Phys. Rev. Lett. **85**, 4626 (2000).

[6] R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. **86**, 3200 (2001).

[7] D. Kempe, J. Kleinberg, and É. Tardos, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2003), pp. 137–146.

[8] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, Phys. Rev. E **65**, 056109 (2002).

[9] L. Meyers, Bull. Am. Math. Soc. **44**, 63 (2007).

[10] B. Karrer, M. E. J. Newman, and L. Zdeborová, Phys. Rev. Lett. **113**, 208702 (2014).

[11] M. E. J. Newman and R. M. Ziff, Phys. Rev. Lett. **85**, 4104 (2000).

[12] J. Leskovec, J. Kleinberg, and C. Faloutsos, ACM Trans. Knowl. Discov. Data **1**, 2 (2007).

[13] M. Ripeanu, I. Foster, and A. Iamnitchi, IEEE Internet Comput. J. **6**(1), 50 (2002), special issue on peer-to-peer networking.

[14] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. Lett. **85**, 5468 (2000).

[15] B. Bollobás, C. Borgs, J. Chayes, O. Riordan *et al.*, Ann. Probab. **38**, 150 (2010).

[16] K.-i. Hashimoto, *Automorphic Forms and Geometry of Arithmetic Varieties* (Academic, Boston, 1989), pp. 211–280.

[17] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, Proc. Natl. Acad. Sci. USA **110**, 20935 (2013).

[18] T. Martin, X. Zhang, and M. E. J. Newman, Phys. Rev. E **90**, 052808 (2014).

[19] K. E. Hamilton and L. P. Pryadko, Phys. Rev. Lett. **113**, 208701 (2014).

[20] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras, Phys. Rev. E **71**, 027103 (2005).

[21] M. Molloy and B. Reed, Random Struct. Algor. **6**, 161 (1995).

[22] M. Boguná, R. Pastor-Satorras, and A. Vespignani, Eur. Phys. J. B **38**, 205 (2004).

[23] F. Radicchi, Phys. Rev. E **90**, 050801 (2014).

[24] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.91.010801 for detailed results for all real networks considered in the analysis.

[25] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, Internet Math. **6**, 29 (2009).

[26] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).

[27] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1977).

[28] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, Behav. Ecol. Sociobiol. **54**, 396 (2003).

[29] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, Science **303**, 1538 (2004).

[30] P. Colomer-de-Simón and M. Boguñá, Phys. Rev. X **4**, 041020 (2014).

[31] A. V. Goltsev, S. N. Dorogovtsev, J. G. Oliveira, and J. F. F. Mendes, Phys. Rev. Lett. **109**, 128702 (2012).