Predicting percolation thresholds in networks

Filippo Radicchi

Center for Complex Networks and Systems Research, School of Informatics and Computing, Indiana University, Bloomington, USA

(Dated: January 16, 2015)

We consider different methods, that 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 non-backtracking 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 robustness of extremely fragile networks.

PACS numbers: 89.75.Hc, 64.60.aq

Percolation is one of the most studied processes in statistical physics. The model assumes the presence of an underlying network structure, where nodes (site percolation) or edges (bond percolation) are independently occupied with probability $p$. Nearest-neighboring 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 non-percolating 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; the percolating phase, where a single macroscopic cluster, whose size is comparable with 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 as $p_c$. Percolation models are commonly used to study network robustness against random failures and the spreading of diseases or ideas. 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. 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. Although the importance of this feature, there are only a few methods, that do not rely on direct numerical simulations of the percolation process, able to determine the value of the percolation threshold in a network. 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. The best estimate of the percolation threshold $p_c = 0.0955$ [Eq. (1), full gray line] is compared with the estimations $\bar{p}_c = 0.0943$ [Eq. (2), dotted black line], $\hat{p}_c = 0.0759$ [Eq. (3), dashed black line], and $\tilde{p}_c = 0.0871$ [Eq. (4), black full line].

Figure 1: (Color online) Bond percolation transition of the giant connected component of the peer-to-peer Gnutella network as of August 31, 2002. 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 $P_c$ as a function of the bond occupation probability $p$ [Eq. (1)]. b) Susceptibility $\chi$ as a function of the bond occupation probability $p$ [Eq. (2)]. $P_c$ and $\chi$ are represented as red thick lines and have been computed with $Q = 10,000$ independent realizations of the Montecarlo algorithm by Newman and Ziff. The best estimate of the percolation threshold $p_c = 0.0955$ [Eq. (1), full gray line] is compared with the estimations $\bar{p}_c = 0.0943$ [Eq. (2), dotted black line], $\hat{p}_c = 0.0759$ [Eq. (3), dashed black line], and $\tilde{p}_c = 0.0871$ [Eq. (4), black full line].

*Electronic address: filiradi@indiana.edu*
In Eq. (4),
\[
\bar{p}_c = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} .
\]
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. \(\bar{p}_c\) represents the value of the percolation threshold expected in uncorrelated graphs with prescribed degree sequence [5, 14]. In general, such 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
\[
\bar{p}_c = \left[ \max \frac{\bar{v}^T A \bar{v}}{\bar{v}^T \bar{v}} \right]^{-1} .
\]
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
\[
\hat{p}_c = \left[ \max \frac{\bar{w}^T M \bar{w}}{\bar{w}^T \bar{w}} \right]^{-1} ,
\]
where the matrix \(M\) is defined as
\[
M = \left( A \mathbb{I} - D \right) .
\]
\(M\) is a square matrix of dimension \(2N \times 2N\) composed of four \(N \times N\) blocks. \(A\) is the adjacency matrix of the graph, \(\mathbb{I}\) 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 non-backtracking or Hasimoto 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 tree-like networks, and it is therefore expected to provide good predictions for sparse networks [10, 19].

First, we perform a systematic study on 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 \(\frac{P(k) \sim k^{-\gamma}}{k \in [3, \sqrt{N}]}\), 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 \(\gamma\), ranging from 2.1 to 100.0, and study the performance of \(\hat{p}_c\) and \(\bar{p}_c\) 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 \(\hat{p}_c\) provides reasonable predictions for the percolation threshold only for values of \(\gamma < 3\). In that regime, the gap \(p_c - \hat{p}_c\) 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. \(\bar{p}_c\) and \(\hat{p}_c\) have instead very good performances for any value of \(\gamma\). 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 \(\bar{p}_c\) is always closer to \(p_c\), than \(\hat{p}_c\), the two measures essentially provide equivalent estimates of the percolation threshold on sparse random graphs. This type of results can be understood with an intuitive mathematical argument, according to which we rewrite the eigenvalue problem for
the matrix $M$ of Eq. (3) as $(\bar{w}_1^T, \bar{w}_2^T) M^T = \lambda (\bar{w}_1^T, \bar{w}_2^T)$, where $\bar{w}_1$ and $\bar{w}_2$ respectively contain the first and the second $N$ components of the eigenvector $\bar{w}$. We can now write the eigenvalue problem as two coupled equations: $A\bar{w}_1 + (1-D)\bar{w}_2 = \lambda \bar{w}_1$ and $\bar{w}_1 = \lambda \bar{w}_2$. Multiplying the first equation by the row vector $\bar{u}^T = (1, \ldots, 1)$, we obtain

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

where the components of the vector $\bar{d}$ are equal to the node degrees. In a random uncorrelated network, we should expect the non-backtracking centrality of the nodes to be proportional to their degrees, i.e., $\bar{w}_1 \sim \bar{d}$, and so Eq. (8) reduces to Eq. (4) [18].

Second and more importantly, we perform a systematic study of the predictive power of $\hat{p}_c$, $\bar{p}_c$ and $\tilde{p}_c$ in a collection of 109 real networks. We consider graphs of heterogeneous nature, including biological, infrastructural, informational, technological, social, and communication networks, and thus very variegated also in terms of structural properties (e.g., degree distribution and correlations, clustering coefficient, 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 to their giant connected components. Results for all 109 real networks are provided in the Supplemental Material. In Fig. 2 we summarize the main outcome of our analysis. As expected, both $\bar{p}_c$ and $\tilde{p}_c$ always provide a lower bound for $p_c$ [19]. Since we have by definition $p_c \geq \bar{p}_c$, it is not a big surprise to find that $\tilde{p}_c$ generates always better predictions than $\bar{p}_c$ [18]. It is however worth to remark that, whereas in synthetic networks the difference between the inverse of the largest eigenvalue of the adjacency matrix $\tilde{p}_c$ and $\bar{p}_c$ can be very large, in real networks, the two indicators generate pretty similar estimations of the percolation threshold, suggesting a relatively little advantage in using $\tilde{p}_c$ in place of $\bar{p}_c$ (see Supplemental Material). It is interesting to observe that, in about the 40% of the real networks analyzed, the naive estimator $\bar{p}_c$, based only on the fraction between first and second moments on the degree distribution, outperforms $\tilde{p}_c$ in spite of using much less topological information. On the other hand, the use of $\tilde{p}_c$ as prediction tool has the disadvantage of alternatively over- or under-estimate the value of the percolation threshold. The relative error of $\tilde{p}_c$ 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 strongly related to the edge density of the graph (see Supplemental Material). Overall, the predictive power of $\tilde{p}_c$ is lower than the one of $\bar{p}_c$, 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 non-backtracking matrix $\tilde{p}_c$ 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...
The relative error colation threshold [Eq. (3)] is computed with networks. In each network, the best estimate and predicted values of the percolation threshold of real networks. In each network, we plot [Eq. (4), green triangles], \( \hat{p}_c \) [Eq. (6), red squares], and \( \tilde{p}_c \) [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 \( \hat{p}_c \) assume the closest value to \( p_c \). c) Relative error of the three predictors, with respect to the value of the best estimate, as function of \( p_c \). We use the same symbols and colors as those of panel a. d) Cumulative distribution of the relative error between predictions and best estimates. Green dotted line refers to \( \tilde{p}_c \), red full line to \( p_c \), and black dashed line to \( \hat{p}_c \). Global performances of the prediction methods are measured in terms of the area under the curve (AUC).

Figure 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 Montecarlo method by Newman and Ziff [15]. a) For each network, we plot [Eq. (4), green triangles], \( \hat{p}_c \) [Eq. (6), red squares], and \( \tilde{p}_c \) [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 \( \hat{p}_c \) assume the closest value to \( p_c \). c) Relative error of the three predictors, with respect to the value of the best estimate, as function of \( p_c \). We use the same symbols and colors as those of panel a. d) Cumulative distribution of the relative error between predictions and best estimates. Green dotted line refers to \( \tilde{p}_c \), red full line to \( p_c \), and black dashed line to \( \hat{p}_c \). Global performances of the prediction methods are measured in terms of the area under the curve (AUC).

As a final consideration, we would like to stress that the existence of a positive relation between the error committed by \( \tilde{p}_c \) to predict \( p_c \) and the value of \( p_c \) itself is an important limitation for the use of \( \tilde{p}_c \) as a proxy for network robustness. This context, percolation is better thought by removing than adding edges, so that the robustness of a network is given by \( 1 - p_c \). The prediction \( 1 - \tilde{p}_c \) always overestimates the true fraction of edge failures that a system can tolerate, providing therefore an information not useful to prevent catastrophic events. Even more importantly, in the analysis of network resilience to random attacks, it is generally more meaningful to consider site instead of bond percolation. Site percolation thresholds are larger than those valid for the bonds (see Supplemental Material), and \( 1 - \tilde{p}_c \) seems not able 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.
Figure 4: (Color online) Inverse participation ratio (IPR) of the principal eigenvectors of the adjacency (red squares) and the non-backtracking matrices [Eq. 7, black circles] as functions of the value of the best estimate of the percolation threshold $p_c$ (panel a), the relative error (panel b), and the absolute error (panel c). Each point refers to numerical results obtained on one of the 109 real networks considered in our analysis.