The Unreasonable Effectiveness of Tree-Based Theory for Bond Percolation on Networks with Clustering

Sergey Melnik[1], Adam Hackett[1], Mason A. Porter[2,3], Peter J. Mucha[4,5], and James P. Gleeson[1]

[1] Department of Mathematics & Statistics, University of Limerick, Ireland
[2] Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, OX1 3LB, UK
[3] CABDyN Complexity Centre, University of Oxford, OX1 1HP, UK
[4] Carolina Center for Interdisciplinary Applied Mathematics, Department of Mathematics, University of North Carolina, Chapel Hill, NC 27599-3250, USA
[5] Institute for Advanced Materials, Nanoscience & Technology, University of North Carolina, Chapel Hill, NC 27599-3216, USA

Abstract. We demonstrate that a tree-based theory for bond percolation yields extremely accurate results for several networks with high levels of clustering. We find that such a theory works well as long as the mean intervertex distance ℓ is sufficiently small—i.e., as long as it is close to the value of ℓ in a random network with negligible clustering and the same degree-degree correlations. We confirm this hypothesis numerically using real-world networks from various domains and on several classes of synthetic clustered networks.

Key words: Random networks, clustering, shortest path, small-world networks

1 Introduction

One of the most important areas of network science is the study of dynamical processes on networks [1–4]. On one hand, research on this topic has provided interesting theoretical challenges for physicists, mathematicians, and computer scientists. On the other hand, there is an increasing recognition of the need to improve the understanding of dynamical systems on networks to achieve advances in epidemic dynamics [5–7], traffic flow in both online and offline systems [8], oscillator synchronization [9], and more [3].

Analytical results for complex networks are rather rare, especially if one wants to study a dynamical system on a network topology that attempts to incorporate even minimal features of real-world networks. If one considers a dynamical system on a real-world network rather than on a grossly simplified caricature of it, then theoretical results become almost barren. Furthermore, most analyses assume that the network under study has a locally tree-like structure,

so that they can only possess very few small cycles, whereas most real networks have significant clustering (and, in particular, possess numerous small cycles). This has motivated a wealth of recent research concerning analytical results on networks with clustering [10–21, 7, 22].

Most existing theoretical results for (unweighted) networks are derived for an ensemble of networks using (i) only their degree distribution p_k, which gives the probability that a random node has degree k (i.e., has exactly k neighbors) or using (ii) their degree distribution and their degree-degree correlations, which are defined by the joint degree distribution $P(k, k')$ describing the probability that a random edge joins nodes of degree k and k'. In the rest of this paper, we will refer to case (i) as "p_k-theory" (the associated random graph ensemble is known as the "configuration model" [23]) and to case (ii) as "$P(k, k')$-theory". The clustering in sample networks is low in both situations; it typically decreases as N^{-1} as the number of nodes $N \to \infty$[6].

We concentrate in this paper on undirected, unweighted real-world networks, which can be described completely using adjacency matrices. It is straightforward to calculate the empirical distributions p_k and $P(k, k')$, which can then be used as inputs to analytical theory for various well-studied processes. The results can subsequently be compared with large-scale numerical simulations using the original networks.

In the present paper, we demonstrate that analytical results derived using tree-based theory can be applied with high accuracy to certain networks despite their high levels of clustering. Examples of such networks include university social networks constructed using Facebook data [24] and the Autonomous Systems (AS) Internet graph [25]. The analytical results for bond percolation accurately match simulations on a given (clustered) network provided that the mean intervertex distance in the network is sufficiently small—i.e., that it is close to its value in a randomly rewired version of the graph. Recalling that a clustered network with a low mean intervertex distance is said to have the *small-world property*, we find that tree-based analytical results are accurate for networks that are "sufficiently small" small worlds. In discussing this result, we focus considerable attention on quantifying what it means to be "sufficiently small".

The remainder of this paper is organized as follows. In Sect. 2, we consider the bond percolation process on highly clustered networks and show that tree-based theory adequately describes it on certain networks but not on others. In order to explain our observations, we introduce in Sect. 3 a measure of prediction quality E and develop a hypothesis, inspired by the well-known Watts-Strogatz example of small-world networks, regarding the dependence of E on the mean intervertex distance ℓ. We provide support for our hypothesis by numerical examination of a large range of networks in Appendix, and discuss our conclusions in Sect. 4.

[6] We assume that the degree distribution has finite variance, as real-world networks necessarily have a finite cutoff in their degree sequence.

2 Bond Percolation on Networks

In bond percolation, network edges are deleted (or labeled as *unoccupied*) with probability $1 - p$, where p is called the *bond occupation probability*. One can measure the effect of such deletions on the aggregate graph connectivity in the limit of infinitely many nodes using $S(p)$, the fractional size of the giant connected component (GCC) at a given value of p. (In this paper, we will use the terminology GCC for finite graphs as well.) Bond percolation has been used in simple models for epidemiology. In such a context, p is related to the average transmissibility of a disease, so that the GCC is used to represent the size of an epidemic outbreak (and to give the steady-state infected fraction in an susceptible-infected-recovered model) [23].

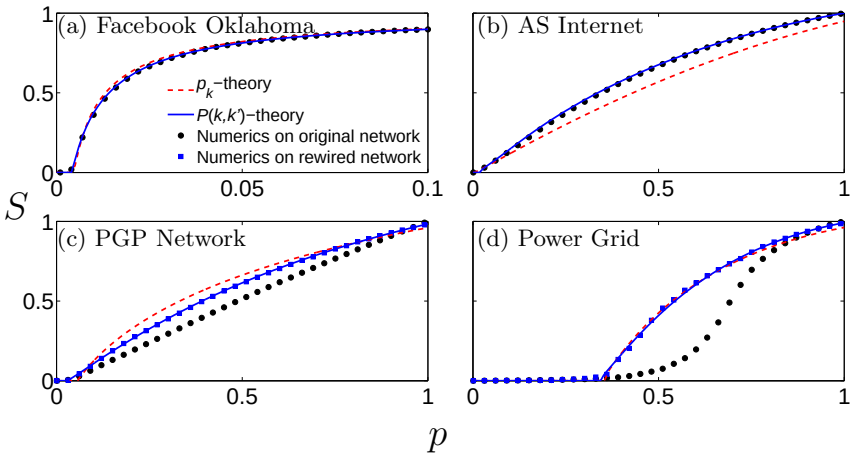


Fig. 1. Bond percolation. Plots of GCC size S versus bond occupation probability p for various real-world networks. These networks are (a) the Facebook network for University of Oklahoma [24], (b) the Internet at the AS level [25], (c) the PGP network [26, 27], and (d) the power grid for the western United States [28];

Analytical results for GCC sizes for p_k-theory [29] can be found in (8.11) of [23] and analytical results for $P(k, k')$-theory are available in (12) of [30]. We plot these theoretical predictions in Fig. 1 as dashed red and solid blue curves, respectively. In this figure, we use the following data sets as examples: (a) the September 2005 Facebook network for University of Oklahoma [24], where nodes are people and links are friendships; (b) the Internet at the Autonomous Systems (AS) level [25], where nodes represent ASs and links indicate the presence of a relationship; (c) the network of users of the Pretty-Good-Privacy (PGP) algorithm for secure information interchange [26, 27]; and (d) the network representing the topology of the power grid of the western United States [28]. We treat all data sets as undirected, unweighted networks.

We performed numerical calculations of the GCC size using the algorithm in [31] and plotted the results as black disks in Fig. 1. It is apparent from Fig. 1(a,b) that $P(k,k')$-theory matches numerical simulations very accurately for the AS Internet and Oklahoma Facebook networks, and we found similar accuracy for all 100 single-university Facebook data sets available to us. However, as shown in Fig. 1(c,d), the match between theory and numerics is much poorer on the PGP and Power Grid networks. The usual explanation for this lack of accuracy is that it is caused by clustering in the real-world network that is not captured by $P(k,k')$-theory. Note, however, that the Oklahoma Facebook network has one of the highest clustering coefficients of the four cases in Fig. 1 even though it is accurately described by its $P(k,k')$-theory.

Indeed, the global clustering coefficients (defined as the mean of the local clustering coefficient over all nodes [28]) for the Oklahoma Facebook, AS Internet, PGP, and Power Grid networks are 0.23, 0.21, 0.27, and 0.08, respectively. (See Table 1 for basic summary statistics for these networks.) The clustering coefficients for all 100 Facebook networks range from 0.19 to 0.41, and the mean value of these coefficients is 0.24. These observations suggest that one ought to consider other explanatory mechanisms for the discrepancy between theory and simulations in Fig. 1(c,d).

In considering other explanations, note that the discrepancy between theory and numerics in Fig. 1(c,d) does not arise from finite-size effects. To demonstrate this, we rewired the networks using an algorithm that preserves the $P(k,k')$ distribution but otherwise randomizes connections between the N nodes[7]. Because this scheme preserves the degree correlation matrix $P(k,k')$, we call this the *P-rewiring* algorithm. Note that the ensemble of fully P-rewired networks is in fact the ensemble of random networks defined by the $P(k,k')$ matrix of the original (unrewired) network.

We show numerical calculations of the GCC sizes for these rewired networks with blue squares in Fig. 1(c,d) and observe that they agree very well with the curves produced from $P(k,k')$-theory. We conclude that the structural characteristics of the original networks—rather than simply their sizes—must underlie the observed differences between simulations and analytics.

Also note that the agreement between $P(k,k')$- and p_k-theories in Fig. 1 is better in panels (a) and (d) than in panels (b) and (c). This is because the Pearson correlation coefficient r of the end-vertex degrees of a random edge [23] has smaller absolute values for the networks shown in panels (a) and (d) (0.074,

[7] We employ the following network rewiring algorithm: Choose an edge of the network at random. Denote its associated vertices by A and B and their corresponding degrees by k_A and k_B. From the set of edges that are connected to one vertex of degree k_A, choose another edge at random. This edge connects the vertices C and D, whose respective degrees are k_A and k_D. Now rewire the two chosen edges to obtain the edges AD and CB instead of AB and CD. This rewiring scheme does not affect the degrees of the rewired vertices, but applying it repeatedly significantly reduces the local clustering (i.e., the density of triangles). In applying this algorithm, we also take care to avoid multiple and self-links.

	Network	N	z	ℓ	ℓ_1	C	$\widetilde{C}$	Ref(s).
Real world	Power Grid	4941	2.67	18.99	8.61	0.08	0.10	[28]
	PGP Network	10680	4.55	7.49	5.40	0.27	0.38	[26, 27]
	AS Internet	28311	4.00	3.88	3.67	0.21	0.0071	[25]
	RL Internet	190914	6.34	6.98	5.25	0.16	0.061	[32]
	Coauthorships	39577	8.88	5.50	4.45	0.65	0.25	[33, 34]
	Airports500	500	11.92	2.99	2.76	0.62	0.35	[35, 36]
	Interacting Proteins	4713	6.30	4.22	4.05	0.09	0.062	[37, 38]
	C. Elegans Metabolic	453	8.94	2.66	2.55	0.65	0.12	[39, 40]
	C. Elegans Neural	297	14.46	2.46	2.33	0.29	0.18	[28, 41]
	Facebook Caltech	762	43.70	2.34	2.26	0.41	0.29	[24]
	Facebook Georgetown	9388	90.67	2.76	2.55	0.22	0.15	[24]
	Facebook Oklahoma	17420	102.47	2.77	2.66	0.23	0.16	[24]
	Facebook UNC	18158	84.46	2.80	2.68	0.20	0.12	[24]
Synthetic	γ-theory $[\gamma(3,3)=1]$	1002	3	13.15	8.06	1/3	1/3	[13]
	γ-theory $[\gamma(3,3)=1]$	10002	3	19.81	11.37	1/3	1/3	[13]
	Watts-Strogatz (WS)	1000	10	50.45	3.29	2/3	2/3	[28]
	Watts-Strogatz (WS)	10000	10	500.45	4.34	2/3	2/3	[28]

Table 1. Basic summary statistics for the networks that we used in this paper. We have treated all real-world data sets as undirected, unweighted networks and have computed the following properties: total number of nodes N; mean degree z; mean intervertex distance ℓ in original network; mean intervertex distance ℓ_1 in the corresponding fully P-rewired version of the network (i.e., in a random network with the original degree correlation); and clustering coefficients C and $\widetilde{C}$ (whose respective definitions are given by (3.6) and (3.4) of [23]). The last column in the table gives the citation number(s) for the data in the bibliography.

with the mean 0.063 over 100 Facebook networks, and 0.0035, respectively) than it does for the networks in (b) and (c) (-0.2 and 0.24, respectively).

3 Measure of Prediction Quality

We now aim to characterize the types of networks for which $P(k,k')$-theory can be expected to give good results.

Using the small-world networks introduced by Watts and Strogatz [28], one can conduct a systematic study of the effects of clustering C and the mean intervertex distance ℓ. We start with a ring of $N = 10000$ nodes and connect each node to $z = 10$ nearest neighbors. We then randomly rewire a fraction f of the links in the network[8]. When $f = 0$, the values of C and ℓ are both high. When $f = 1$, the rewired network is connected completely at random, which gives it low C and ℓ values. For each value of f between 0 and 1, we numerically

[8] We employ our P-rewiring algorithm that preserves the degree of each node, which is slightly different from the one used in [28], but this difference is not important for the phenomenon under study.

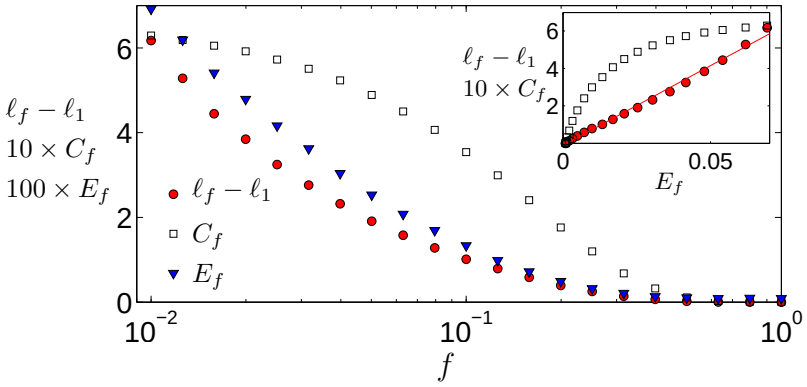


Fig. 2. Watts-Strogatz small-world network: $\ell_f - \ell_1$ (red circles), $10 \times C_f$ (open squares), and $100 \times E_f$ (blue triangles) as functions of rewiring fraction f. The inset shows $\ell_f - \ell_1$ and C_f as functions of E_f for $f \geq 10^{-2}$. Observe the linear relation between E_f and $\ell_f - \ell_1$, which suggests that $\ell_f - \ell_1$ might be a good indicator of how well the bond-percolation process on a network can be approximated by tree-based theory.

calculate the clustering coefficient C_f, the mean intervertex distance ℓ_f, and the GCC size $S_f(p)$ for all values of the bond occupation probability p between 0 and 1. The difference between $S_f(p)$ and the $P(k, k')$-theory curve, which we denote by $S_{\text{th}}(p)$, gives a quantitative measure for the inaccuracy of the theory for this particular value of the rewiring parameter f. We define the error measure

$$E_f = \frac{1}{M} \sum_{i=1}^{M} |S_{\text{th}}(p_i) - S_f(p_i)| \ , \tag{1}$$

where $p_i = i/M$ for $i = 1, 2, \ldots, M$ are uniformly-spaced values in the interval $[0, 1]$. Taking the spacing $1/M$ to be sufficiently fine (we use $1/M = 10^{-3}$) implies that the error measure E_f approaches the average vertical distance between the $S_{\text{th}}(p)$ and $S_f(p)$ curves for $p \in [0, 1]$.

In Fig. 2, we plot the values of $\ell_f - \ell_1$, C_f (scaled by a factor of 10 for ease of visualization), and E_f (scaled by a factor of 100) as functions of the rewiring parameter f. For values of f greater than 10^{-2}, the quantities ℓ_f and E_f exhibit similar behavior, whereas C_f remains near its $f = 0$ value of $2/3$ until f is much larger[9]. We highlight the similar scaling of ℓ_f and E_f in the inset of Fig. 2, in which we plot $\ell_f - \ell_1$ directly as a function of E_f for $f \geq 10^{-2}$. The approximately linear dependence that we observe contrasts to the clearly nonlinear relation between E_f and the clustering C_f that we show in the same

[9] When $f \ll 10^{-2}$, the quantity ℓ_f changes much more rapidly with f than E_f does. We focus on the range $f \geq 10^{-2}$ in Fig. 2 because for lower f, the values of the error E_f are much larger than those seen in any of the networks we study (e.g., the Power Grid network has $E \approx 0.11$ and the PGP network has $E \approx 0.065$, which should be compared to the maximum error of 0.07 seen in Fig. 2).

inset. This strongly suggests that differences between theory and numerics are related more directly to the mean intervertex distance than to the clustering coefficient.

The above results for Watts-Strogatz small-world networks motivate the examination of a range of real-world networks in order to seek a clear relationship between an error measure similar to (1) and some other characteristic of the network, such as clustering or mean intervertex distance. For each network, we calculate the inaccuracy of $P(k,k')$-theory in terms of the error E, which measures the distance between the actual (numerically calculated) GCC size curve $S_{\rm num}(p)$ and the theoretical prediction $S_{\rm th}(p)$:

$$E = \frac{1}{M}\sum_{i=1}^{M} |S_{\rm th}(p_i) - S_{\rm num}(p_i)| \ . \tag{2}$$

Essentially, E gives the average distance between the numerics (black disks) and theory (solid blue curve) in Fig. 1. In Fig. 3(a), we show a scatter plot of $\log_{10} E$ versus $\log_{10} C$, where C is the clustering coefficient of each network. We use logarithmic coordinates in Fig. 3 in order to fully resolve the range of values for both variables, as they vary by one or more orders of magnitude.

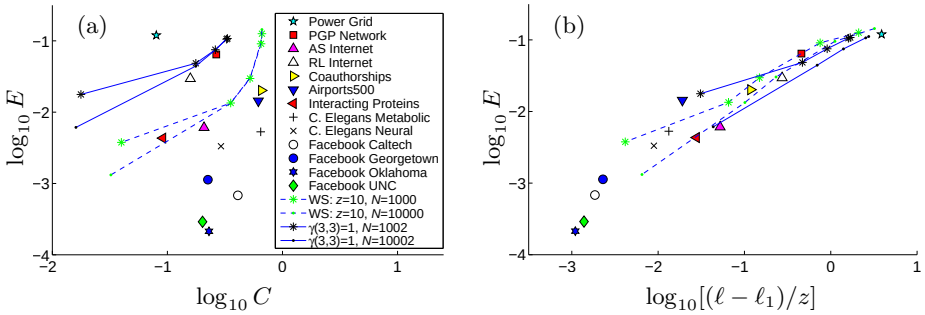


Fig. 3. Scatter plots of $\log_{10} E$ versus (a) $\log_{10} C$ (with $R^2 \approx 0.087$) and (b) $\log_{10}[(\ell-\ell_1)/z]$ (with $R^2 \approx 0.94$).

We also include synthetic examples, such as Watts-Strogatz small-world networks and clustered random networks generated using the recent models described in [13, 12], which we now briefly recall. The fundamental quantity defining the γ-theory networks of [13] is the joint probability distribution $\gamma(k,c)$, which gives the probability that a randomly chosen node has degree k and is a member of a c-clique (an all-to-all connected subgraph of c nodes). With $\gamma(3,3) = 1$ (and zero for other values of k and c), each node in such a network has degree 3 and is part of exactly one triangle. This is equivalent to the $p_{1,1} = 1$ case in the clustered random graph model of [12], where $p_{s,t}$ is the probability that a randomly chosen node is part of t different triangles and in addition has s single edges

(which don't belong to the triangles). In each synthetic network, we P-rewire a fraction f of links and show our results for $f = \{10^{-3}, 4 \times 10^{-3}, 0.04, 0.1, 0.4\}$.

In order to assess the strength of a relation between the theory error E and some characteristic of the network, we calculate the coefficient of determination R^2 using a linear regression. For the data in Fig. 3(a), we calculate $R^2 \approx 0.087$ (using the points only and ignoring the connecting curves which help identify families of points). This relatively small value indicates that C is not a good predictor of the theory error across the set of networks that we tested (see Table 1). After examining a wide range of possibilities (see the scatter plots in Appendix), we found that the network measure that best correlates with the error E (on logarithmic scales) is $(\ell - \ell_1)/z$ (which gives $R^2 \approx 0.94$), where z is the mean degree and ℓ_1 is the mean intervertex distance in the version of the network that has been fully rewired while preserving the joint degree distribution $P(k, k')$ [see Fig. 3(b)]. Recall that one can think of such fully P-rewired versions of a network as random networks with the same degree correlation $P(k, k')$ and size as the original network.

We can summarize our observations as follows. Given a network, we compare its mean intervertex distance ℓ with the value ℓ_1 in a random network of equal size and degree correlation $P(k, k')$. If the difference $\ell - \ell_1$ is sufficiently small—e.g., if it is less than $z/10$, as was the case in Fig. 1(a,b)—then the $P(k, k')$-theory can be expected to accurately give the GCC size. For example, the AS Internet graph has $(\ell - \ell_1)/z \approx 3.3 \times 10^{-2}$ and all 100 Facebook networks have values much smaller than this. However, the theory is not accurate for larger values of $\ell - \ell_1$. (For example, the PGP and Power Grid networks have $(\ell - \ell_1)/z$ values of approximately 0.45 and 3.9, respectively.)

Because the tree-based theory systematically gives accurate results for bond percolation on networks that are *not* locally tree-like when the intervertex distance is small, it seems that there must be a deeper argument than is currently known for the validity of such theories.

4 Conclusions

At the beginning of this paper, we posed the following question: "How small must small-world networks be in order for $P(k, k')$-theory to give accurate results?" Our heuristic answer is that they must have a value for the mean intervertex distance ℓ that differs from the mean intervertex distance in a random network with the same $P(k, k')$ and number of nodes by no more than about 10% of the mean degree z. Surprisingly, the level of clustering has much less of an impact on the accuracy of $P(k, k')$-theory, which is why we found excellent matches between theory and numerical simulations even in highly clustered graphs such as Facebook social networks and the AS Internet network.

Although our presentation used bond percolation as our primary example, we have shown in unpublished work [42] that on networks for which $P(k, k')$-theory is accurate for bond percolation, it also works well for some other processes, such as k-core sizes [43, 44] and susceptible-infected-susceptible (SIS) dynam-

ics [6]. However, an absolute measure of accuracy must, of course, depend on the process under scrutiny. Some processes (e.g. Watts' threshold model for the spread of cultural fads [45]) are particularly sensitive to deviations of the network from randomness and can provide a suitable testing ground for new analytically solvable models of networks that include clustering [12, 13].

In summary, we have shown that the tree-based analytical theory for bond percolation yields highly accurate results for networks in which $\ell \approx \ell_1$ even in the presence of significant clustering. Such graphs, which include the AS Internet network and Facebook social networks, are definitively not locally tree-like, so that the theory is working very well even in situations where the theory's fundamental hypothesis is known to fail utterly. We hope that the results of the present paper will motivate further research on the underlying causes of this "unreasonable" effectiveness of tree-based theory for clustered networks.

Acknowledgements

SM, AH, and JPG acknowledge funding provided by Science Foundation Ireland under programmes 06/IN.1/I366 and MACSI 06/MI/005. MAP acknowledges a research award (#220020177) from the James S. McDonnell Foundation. PJM was funded by the NSF (DMS-0645369). We thank Adam D'Angelo and Facebook for providing the Facebook data used in this study. We also thank Alex Arenas, Mark Newman, CAIDA, and Cx-Nets collaboratory for making publicly available other data sets used in this paper.

References

1. Strogatz, S.H.: Exploring complex networks. Nature (London) **410** (2001) 268–276
2. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex networks: Structure and dynamics. Phys. Rep. **424** (2006) 175–308
3. Barrat, A., Vespignani, A., Barthelemy, M.: Dynamical processes on complex networks. Cambridge University Press, Cambridge, UK (2008)
4. Dorogovtsev, S.N., Goltsev, A.V., Mendes, J.F.F.: Critical phenomena in complex networks. Rev. Mod. Phys. **80** (2008) 1275
5. Pastor-Satorras, R., Vespignani, A.: Epidemic spreading in scale-free networks. Phys. Rev. Lett. **86** (2001) 3200–3203
6. Barthélemy, M., Barrat, A., Pastor-Satorras, R., Vespignani, A.: Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. J. Theor. Biol. **235** (2005) 275–288
7. Eames, K.T.D.: Modelling disease spread through random and regular contacts in clustered populations. Theor. Pop. Biol. **73** (2008) 104–111
8. Garavello, M., Piccoli, B.: Traffic flow on networks. American Institute of Mathematical Sciences, Springfield, MO (2006)
9. Arenas, A., Diaz-Guilera, A., Kurths, J., Moreno, Y., Zhou, C.: Synchronization in complex networks. Phys. Rep. **469** (2008) 93–153
10. Newman, M.E.J.: Properties of highly clustered networks. Phys. Rev. E **68** (2003) 026121

11. Porter, M.A., Onnela, J.P., Mucha, P.J.: Communities in networks. Not. Amer. Math. Soc **9** (2009) 1082–1097, 1164–1166
12. Newman, M.E.J.: Random graphs with clustering. Phys. Rev. Lett. **103** (2009) 058701
13. Gleeson, J.P.: Bond percolation on a class of clustered random networks. Phys. Rev. E **80** (2009) 036107
14. Gleeson, J.P., Melnik, S.: Analytical results for bond percolation and k-core sizes on clustered networks. Phys. Rev. E **80** (2009) 046121
15. Ostilli, M., Mendes, J.F.F.: Communication and correlation among communities. Phys. Rev. E **80** (2009) 011142
16. Serrano, M.Á., Boguñá, M.: Clustering in complex networks. I. General formalism. Phys. Rev. E **74** (2006) 056114
17. Serrano, M.Á., Boguñá, M.: Clustering in complex networks. II. Percolation properties. Phys. Rev. E **74** (2006) 056115
18. Serrano, M.Á., Boguñá, M.: Percolation and epidemic thresholds in clustered networks. Phys. Rev. Lett. **97** (2006) 088701
19. Trapman, P.: On analytical approaches to epidemics on networks. Theor. Pop. Biol. **71** (2007) 160–173
20. Miller, J.C.: Spread of infectious disease through clustered populations. J. Roy. Soc. Interface **6** (2009) 1121
21. Miller, J.C.: Percolation and epidemics in random clustered networks. Phys. Rev. E **80** (2009) 020901
22. Britton, T., Deijfen, M., Lagerås, A.N., Lindholm, M.: Epidemics on random graphs with tunable clustering. J. Appl. Probab. **45** (2008) 743
23. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45** (2003) 167–256
24. Traud, A.L., Kelsic, E.D., Mucha, P.J., Porter, M.A.: Community structure in online collegiate social networks. arXiv (2008) 0809.0690
25. The CAIDA Autonomous System Relationships Dataset, 30-Jun-2008: `http://www.caida.org/data/active/as-relationships;http://as-rank.caida.org/data/2008/as-rel.20080630.a0.01000.txt`
26. Guardiola, X., Guimera, R., Arenas, A., Diaz-Guilera, A., Streib, D., Amaral, L.A.N.: Macro- and micro-structure of trust networks. arXiv (2002) 0206240
27. Boguñá, M., Pastor-Satorras, R., Diaz-Guilera, A., Arenas, A.: Models of social networks based on social distance attachment. Phys. Rev. E **70** (2004) 056122
28. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature (London) **393** (1998) 440–442
29. Callaway, D.S., Newman, M.E.J., Strogatz, S.H., Watts, D.J.: Network robustness and fragility: Percolation on random graphs. Phys. Rev. Lett. **85** (2000) 5468–5471
30. Vázquez, A., Moreno, Y.: Resilience to damage of graphs with degree correlations. Phys. Rev. E **67** (2003) 015101(R)
31. Newman, M.E.J., Ziff, R.M.: Fast Monte Carlo algorithm for site or bond percolation. Phys. Rev. E **64** (2001) 016706
32. Internet router-level graph computed from ITDK0304 skitter and iffinder measurements. "CAIDA's Internet Topology Data Kit #0304." San Diego Supercomputer Center, University of California, San Diego: `www.caida.org/tools/measurement/skitter/router_topology/itdk0304_rlinks_undirected.gz`
33. Newman, M.E.J.: The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. U.S.A. **98** (2001) 404–409

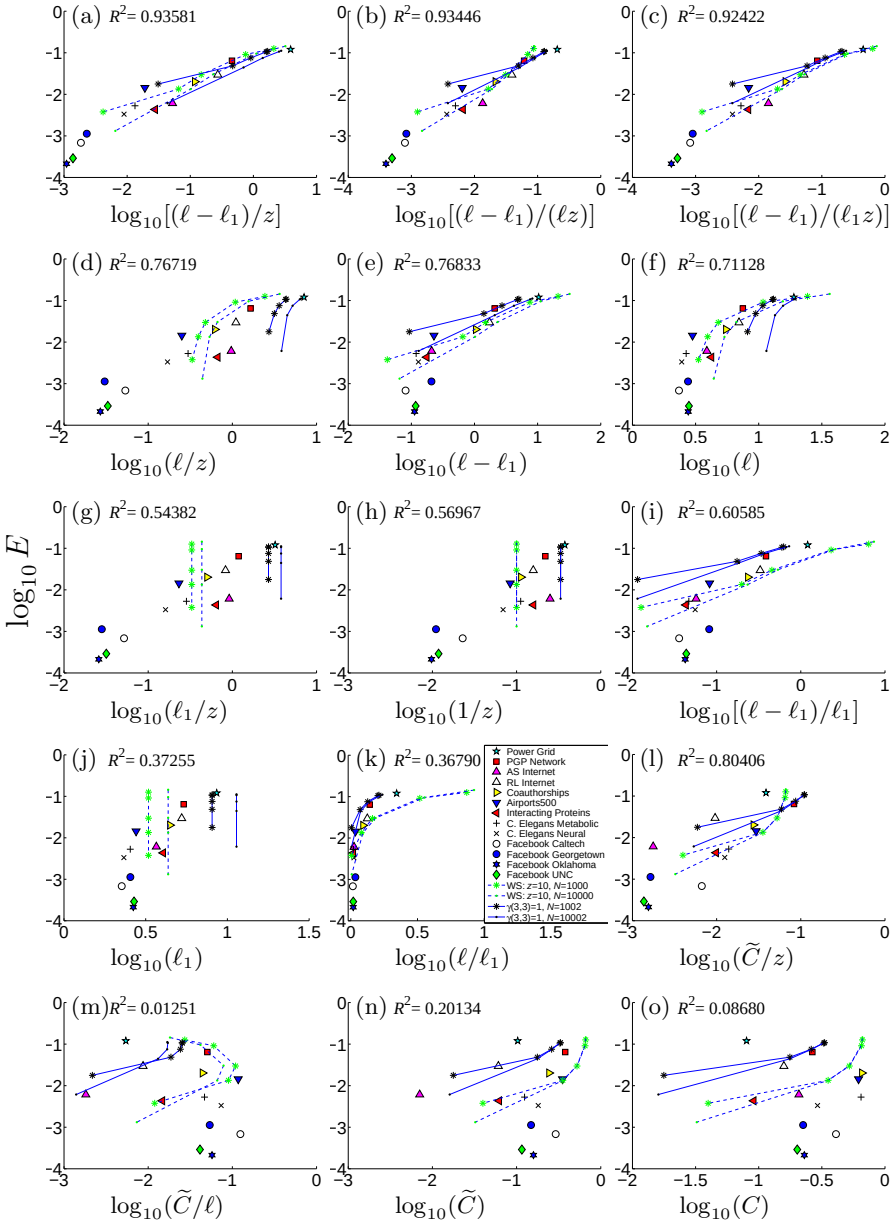
34. Network of coauthorships between scientists posting preprints on the Condensed Matter E-Print Archive, includes all preprints posted between 1-Jan-1995 and 31-Mar-2005: http://www-personal.umich.edu/~mejn/netdata/cond-mat-2005.zip
35. Colizza, V., Pastor-Satorras, R., Vespignani, A.: Reaction-diffusion processes and metapopulation models in heterogeneous networks. Nat. Phys. **3** (2007) 276
36. A network obtained by considering the 500 US airports with the largest amount of traffic from publicly available data: http://sites.google.com/site/cxnets/US_largest500_airportnetwork.txt
37. Colizza, V., Flammini, A., Maritan, A., Vespignani, A.: Characterization and modeling of protein-protein interaction networks. Physica A **352** (2005) 1
38. Colizza, V., Flammini, A., Serrano, M.A., Vespignani, A.: Detecting rich-club ordering in complex networks. Nat. Phys. **2** (2006) 110
39. Duch, J., Arenas, A.: Community detection in complex networks using extremal optimization. Phys. Rev. E **72** (2005) 027104
40. A metabolic network of C. Elegans: http://deim.urv.cat/~aarenas/data/xarxes/celegans_metabolic.zip
41. A neural network of C. Elegans: http://www-personal.umich.edu/~mejn/netdata/celegansneural.zip
42. Melnik, S., Hackett, A., Porter, M.A., Mucha, P.J., Gleeson, J.P.: The unreasonable effectiveness of tree-based theory for networks with clustering. arXiv (2010) 1001.1439
43. Dorogovtsev, S.N., Goltsev, A.V., Mendes, J.F.F.: k-core organization of complex networks. Phys. Rev. Lett. **96** (2006) 040601
44. Gleeson, J.P.: Cascades on correlated and modular random networks. Phys. Rev. E **77** (2008) 046117
45. Watts, D.J.: A simple model for global cascades on random networks. Proc. Natl. Acad. Sci. U.S.A. **99** (2002) 5766–5771

Appendix: Scatter Plots

In this appendix, we show scatter plots of $\log_{10} E$ versus a variety of possible predictors. Recall that E, which we defined in (2), gives an error measure for bond percolation. We test for the dependence of E on various combinations of the mean degree z, mean intervertex distance ℓ, and clustering coefficients[10]. Recall again that ℓ_1 denotes the value taken by ℓ in a fully P-rewired version of a network (i.e., in a random network with the same degree correlation and size).

The scatter plots show data points for real-world networks, and for synthetic Watts-Strogatz small-world networks and γ-theory networks, which are described in Sect. 3. The dependence of E on $\ell - \ell_1$ is clearly strong (see the top row of scatter plots, which all have $R^2 > 0.9$), whereas the dependence on clustering is weak (see the bottom row of scatter plots, which all have $R^2 < 0.3$). Given the relatively small number of available data sets, we cannot definitively select the best scaling function $F(z, \ell, \ldots)$ for the relation $E \approx F(z, \ell, \ldots)(\ell - \ell_1)$, but the simple choice $F = 1/z$ used in Fig. 3(b) gives satisfactory fits.

[10] We consider both common definitions of clustering coefficient. We use C to denote the coefficient defined by (3.6) of [23] and $\widetilde{C}$ to denote that from (3.4) of [23].


(a) R^2= 0.93581
(b) R^2= 0.93446
(c) R^2= 0.92422
(d) R^2= 0.76719
(e) R^2= 0.76833
(f) R^2= 0.71128
(g) R^2= 0.54382
(h) R^2= 0.56967
(i) R^2= 0.60585
(j) R^2= 0.37255
(k) R^2= 0.36790
(l) R^2= 0.80406
(m) R^2= 0.01251
(n) R^2= 0.20134
(o) R^2= 0.08680
$\log_{10} E$
$\log_{10}[(\ell-\ell_1)/z]$
$\log_{10}[(\ell-\ell_1)/(\ell z)]$
$\log_{10}[(\ell-\ell_1)/(\ell_1 z)]$
$\log_{10}(\ell/z)$
$\log_{10}(\ell-\ell_1)$
$\log_{10}(\ell)$
$\log_{10}(\ell_1/z)$
$\log_{10}(1/z)$
$\log_{10}[(\ell-\ell_1)/\ell_1]$
$\log_{10}(\ell_1)$
$\log_{10}(\ell/\ell_1)$
$\log_{10}(\widetilde{C}/z)$
$\log_{10}(\widetilde{C}/\ell)$
$\log_{10}(\widetilde{C})$
$\log_{10}(C)$
Power Grid
PGP Network
AS Internet
RL Internet
Coauthorships
Airports500
Interacting Proteins
C. Elegans Metabolic
C. Elegans Neural
Facebook Caltech
Facebook Georgetown
Facebook Oklahoma
Facebook UNC
WS: z=10, N=1000
WS: z=10, N=10000
γ(3,3)=1, N=1002
γ(3,3)=1, N=10002