The unreasonable effectiveness of tree-based theory for 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 and Statistics, University of Limerick, Ireland*

[2]*Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, Oxford OX1 3LB, United Kingdom*

[3]*CABDyN Complexity Centre, University of Oxford, Oxford OX1 1HP, United Kingdom*

[4]*Carolina Center for Interdisciplinary Applied Mathematics, Department of Mathematics, University of North Carolina, Chapel Hill, North Carolina 27599-3250, USA*

[5]*Institute for Advanced Materials, Nanoscience and Technology, University of North Carolina, Chapel Hill, North Carolina 27599-3216, USA*

(Received 11 January 2010; revised manuscript received 23 December 2010; published 23 March 2011)

We demonstrate that a tree-based theory for various dynamical processes operating on static, undirected networks 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—that is, 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 support this hypothesis numerically using both real-world networks from various domains and several classes of synthetic clustered networks. We present analytical calculations that further support our claim that tree-based theories can be accurate for clustered networks, provided that the networks are "sufficiently small" worlds.

DOI: 10.1103/PhysRevE.83.036112 PACS number(s): 89.75.Hc, 89.75.Fb, 64.60.aq, 87.23.Ge

I. 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. Most analyses assume that the network under study has a locally treelike structure, so that it possesses very few small cycles (or loops), whereas most real networks have significant clustering (and, in particular, possess numerous small cycles) [10]. Furthermore, 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. This has motivated a wealth of recent research concerning analytical results on networks with clustering [7,11–23].

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 (that is, it has exactly k neighbors) or using (ii) their degree distribution and their degree-degree correlations, which are defined by the joint degree-degree distribution $P(k,k')$, describing the probability that a random edge joins nodes of degrees k and k'. In the rest of this paper, we refer to case (i) as p_k-theory (the associated random graph ensemble is known as the configuration model [24]) 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 $N \to \infty$ [25], so these so-called *tree-based theories* generally cannot guarantee meaningful predictions for real-world networks with significant clustering.

We concentrate in this paper on static, undirected, unweighted real-world networks, each of which is completely described by an adjacency matrix. The adjacency matrix can be used directly to model various processes on the network that it represents. We refer to such calculations and results as *numerical*, because they do not involve any theory or assumptions about the network structure. Because there are no assumptions, such numerical results are the most accurate results; they are, however, computationally expensive when the network is large. From an analytical perspective, one can obtain the empirical distributions p_k and $P(k,k')$ from the network adjacency matrix and use them as respective inputs to tree-based analytical p_k- and $P(k,k')$-theories for dynamical processes. Such calculations are much less computationally expensive and can provide a deeper insight into dynamics of interest, but the results given by such theories might be inaccurate in the sense that they have the potential to differ significantly from the numerical results. One reason for this inaccuracy is that, unlike the adjacency matrix, the distributions p_k and $P(k,k')$ contain only partial information about the original network structure. For example, they cannot describe the loops that are present in the network. Therefore, such tree-based theories can guarantee accurate results only for random networks defined by these distributions (and in the limit of large network size).

In this paper, we consider real-world clustered networks and run several dynamical processes on these networks with a view to measuring the discrepancy between the analytical tree-based theories and the (true) numerical results for each of these dynamical processes. We investigate how the agreement between the tree-based theory and the corresponding (true) numerical result depends on the network structure. In other words, we assume that the dynamics on a given clustered network is similar to that on a random graph with the same distribution [p_k or $P(k,k')$] and determine the condition under which this assumption is adequate.

We demonstrate that analytical results derived using tree-based $P(k,k')$-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 [26] and the autonomous systems Internet graph [27]. Specifically, the analytical results for bond percolation, k-core sizes, and other processes accurately match numerical results for a given (clustered) network, provided that the mean intervertex distance in the network is sufficiently small. That is, it must be 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." In other words, how small must small-world networks be in order for $P(k,k')$-theory to give accurate results?

The remainder of this paper is organized as follows. In Sec. II, we consider several dynamical processes on highly clustered networks and show that tree-based theory adequately describes them on certain networks but not on others. In order to explain our observations, we introduce in Sec. III a measure of prediction quality E and develop a hypothesis, inspired by the well-known Watts-Strogatz example of small-world networks, regarding its dependence on the mean intervertex distance ℓ. We provide support for our hypothesis with analytical calculations in Appendix A and with numerical examination of a large range of networks in Appendix B. We discuss our conclusions in Sec. IV.

II. DYNAMICAL PROCESSES ON NETWORKS

A. Bond percolation

We begin by considering bond percolation, which has been studied extensively 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 use the terminology GCC for finite graphs as well; one can alternatively use the term "largest connected component" for finite graphs.) Bond percolation has been employed in simple models for epidemic dynamics. In such a context, p is related to the mean transmissibility of a disease, so the GCC is used to represent the size of an epidemic outbreak (and to give the steady-state infected fraction in a susceptible-infected-recovered model) [24].

Given the network adjacency matrix, we calculate the distributions p_k and $P(k,k')$ and then use them in the analytical expressions that predict the GCC size for a particular value of p. Analytical expressions for predicting GCC sizes using p_k-theory [28] can be found in Eq. (8.11) of Ref. [24], and analytical results for $P(k,k')$-theory are available in Eq. (12) of Ref. [29]. 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 [26], where nodes are people and links are friendships; (b) the Internet at the autonomous systems (AS) level [27], 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 [30–32]; and (d) the network representing the topology of the power grid of the western United States [33,34]. We treat all data sets as undirected, unweighted networks.

We perform numerical calculations of the GCC size by applying the algorithm of Ref. [35] to the adjacency matrices of our networks and plot the average results as black disks in Fig. 1. It is apparent from Figs. 1(a)–1(b) that $P(k,k')$-theory matches numerical results very accurately for the AS Internet and Oklahoma Facebook networks, and we obtain similar accuracy for all 100 single-university Facebook data sets available to us [36]. However, as shown in Figs. 1(c)–1(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 example networks in Fig. 1, even though it is accurately described by $P(k,k')$-theory.

The global clustering coefficients (defined as the mean of the local clustering coefficients over all nodes [33]) for the Oklahoma Facebook, AS Internet, PGP, and power grid networks are 0.23, 0.21, 0.27, and 0.08, respectively. (See Table I 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 numerical calculations in Figs. 1(c) and 1(d).

In considering other explanations, note that the discrepancy between theory and numerics in Figs. 1(c) and 1(d) does not arise from finite-size effects. To demonstrate this, we rewire the networks using an algorithm that preserves the $P(k,k')$ distribution but otherwise randomizes connections between the nodes [37]. Because this scheme preserves the degree correlation matrix $P(k,k')$, we call this the *P-rewiring* algorithm. Note that the $P(k,k')$-theory should be accurate for fully P-rewired networks, because 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 (not rewired) network.

We use the numerical algorithm of Ref. [35] again to calculate the GCC sizes for these rewired networks. We show the results averaged over 100 complete and independent rewirings with blue squares in Figs. 1(c)–1(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 numerical calculations 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 networks in (a) and (d) have smaller absolute values for the Pearson correlation coefficient r of the end-vertex degree of a random edge [24]. The value of r for the network in (a) is 0.074, and the mean over all 100 Facebook networks is 0.063; the value of r for the network in

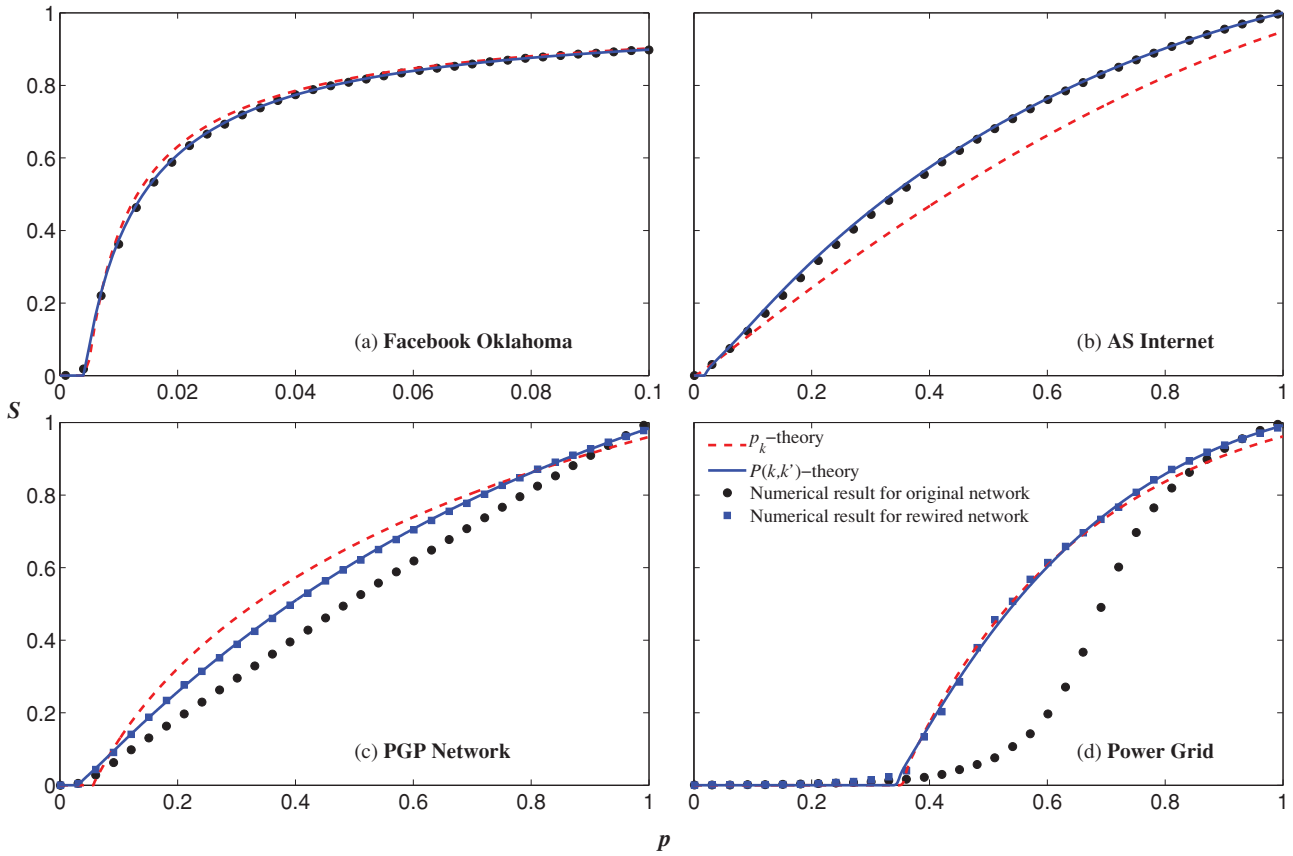

FIG. 1. (Color online) Bond percolation. Plots of giant connected component (GCC) size S versus bond occupation probability p for various real-world networks. These networks, which we also use as examples in other figures, are (a) the Facebook network for University of Oklahoma [26], (b) the Internet at the autonomous systems (AS) level [27], (c) the Pretty Good Privacy (PGP) network [30–32], and (d) the power grid for the western United States [33,34].

TABLE I. Basic summary statistics for the networks that we use 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 (that is, in a random network with the original joint degree-degree distribution); the mean intervertex distance ℓ_1^B predicted by Eq. (A2) using the branching matrix corresponding to a random network with the original joint degree-degree distribution; clustering coefficients C and $\widetilde{C}$ (whose respective definitions are given by Eqs. (3.6) and (3.4) of [24]); and the Pearson degree correlation coefficient r. The last column in the table gives the relevant citation number(s) in the bibliography.

	Network	N	z	ℓ	ℓ_1	ℓ_1^B	C	$\widetilde{C}$	r	Ref(s).
Real world	Power grid	4941	2.67	18.99	8.61	7.85	0.08	0.10	0.0035	[33,34]
	PGP network	10680	4.55	7.49	5.40	2.66	0.27	0.38	0.23	[30–32]
	AS Internet	28311	4.00	3.88	3.67	2.56	0.21	0.0071	−0.20	[27]
	RL Internet	190914	6.34	6.98	5.25	3.17	0.16	0.061	0.025	[38]
	Coauthorships	39577	8.88	5.50	4.45	2.93	0.65	0.25	0.19	[39,40]
	Airports 500	500	11.92	2.99	2.76	1.62	0.62	0.35	−0.278	[41,42]
	Interacting proteins	4713	6.30	4.22	4.05	2.96	0.09	0.062	−0.136	[43–45]
	C. Elegans metabolic	453	8.94	2.66	2.55	1.93	0.65	0.12	−0.226	[46,47]
	C. Elegans neural	297	14.46	2.46	2.33	1.84	0.29	0.18	−0.163	[33,48]
	Facebook Caltech	762	43.70	2.34	2.26	1.55	0.41	0.29	−0.066	[26]
	Facebook Georgetown	9388	90.67	2.76	2.55	1.79	0.22	0.15	0.075	[26]
	Facebook Oklahoma	17420	102.47	2.77	2.66	1.79	0.23	0.16	0.074	[26]
	Facebook UNC	18158	84.46	2.80	2.68	1.87	0.20	0.12	7×10^{-5}	[26]
Synthetic	γ-Theory [$\gamma(3,3)=1$]	1002	3	13.15	8.06	9.97	1/3	1/3	N/A	[14]
	γ-Theory [$\gamma(3,3)=1$]	10002	3	19.81	11.37	13.29	1/3	1/3	N/A	[14]
	Watts-Strogatz (WS)	1000	10	50.45	3.29	3.14	2/3	2/3	N/A	[33]
	Watts-Strogatz (WS)	10000	10	500.45	4.34	4.19	2/3	2/3	N/A	[33]

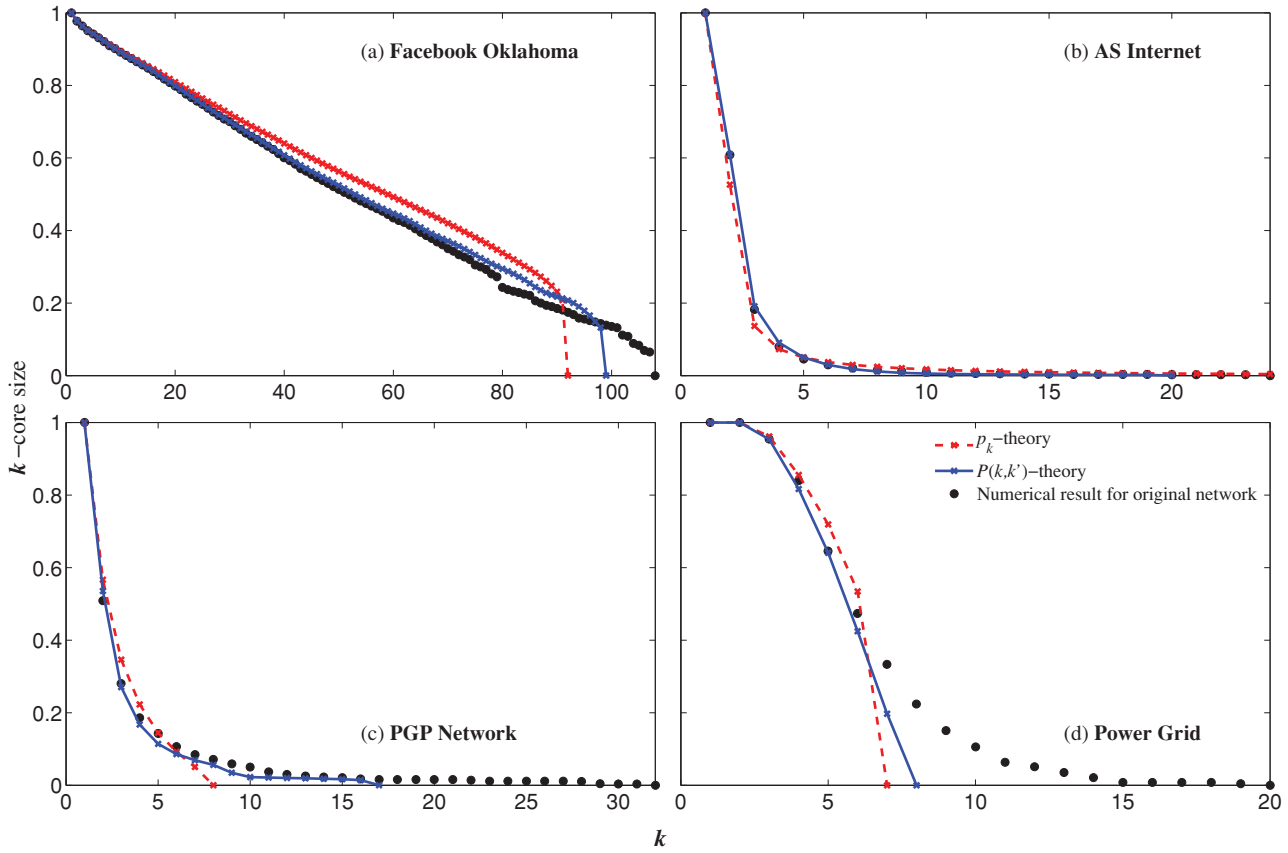

FIG. 2. (Color online) Plots of k-core sizes versus k for the real-world networks from Fig. 1. The highest values of k for which the k-core size is nonzero are (a) $K_{p_k} = 91$, $K_{P(k,k')} = 98$, $K_{\rm num} = 107$; (b) $K_{p_k} = 132$, $K_{P(k,k')} = 19$, $K_{\rm num} = 23$; (c) $K_{p_k} = 7$, $K_{P(k,k')} = 16$, $K_{\rm num} = 31$; and (d) $K_{p_k} = 6$, $K_{P(k,k')} = 7$, $K_{\rm num} = 19$.

(d) is 0.0035; the value for that in (b) is -0.20; and the value for that in (c) is 0.24.

B. k-Cores

Figures 2–4 show similar comparisons between analytical and numerical results for other well-studied processes on networks.

In Fig. 2, we plot the k-core sizes of the networks. The k-core is the largest subgraph whose nodes all have degree at least k within the subgraph. The p_k-theory for k-core sizes is given in Ref. [49], and the $P(k,k')$-theory is given by Eq. (32) of Ref. [50]. We compare these theoretical predictions with direct calculations of k-core sizes from the adjacency matrices. Although the direct calculation of k-cores is a measurement of the real network, we continue to use the term "numerical" in this subsection in order to contrast such calculations with theoretical predictions. As shown in Figs. 2(a) and 2(b), we again find very good agreement of $P(k,k')$-theory with numerical calculations on the AS Internet and Facebook networks and less accurate results for the other sample networks. This can be quantified by comparing the numerical (true) result for the highest value of k for which the k-core size is nonzero to the value that is predicted by $P(k,k')$-theory. (We use K to denote this maximal value of k.) For Figs. 2(a) and 2(b), we obtain $K_{P(k,k')}/K_{\rm num} \approx 0.916$ and $K_{P(k,k')}/K_{\rm num} \approx 0.826$, respectively. The corresponding values for Figs. 2(c) and 2(d) are $K_{P(k,k')}/K_{\rm num} \approx 0.516$ and $K_{P(k,k')}/K_{\rm num} \approx 0.368$.

C. Watts threshold model

Watts introduced a simple model for the spread of cultural fads [51]. It allows one to examine how a small initial fraction of early adopters can lead to a global cascade of adoption via a social network, distinguishing between "simple" and "complex" contagions [52,53]. The p_k-theory and $P(k,k')$-theory for the mean cascade size are given, respectively, in Refs. [54] and [50]. In Fig. 3, we compare these theories with numerical simulations on populations with Gaussian threshold distributions of mean μ and variance $\sigma^2 = 0.04$. The cascade size shows a sharp transition as μ is increased. As with the other processes discussed previously, the position of this transition is accurately captured by the theory for the Facebook and AS Internet networks but not for the other examples.

D. Susceptible-infected-susceptible model

In Fig. 4, we show a comparison between theory and numerical results for the time evolution of a susceptible-infected-susceptible (SIS) epidemic model on various networks. Unlike the other processes that we have discussed, the theory for this case—as given, for example, by Eq. (17) of Ref. [6]—is expected to apply accurately only to the early-time

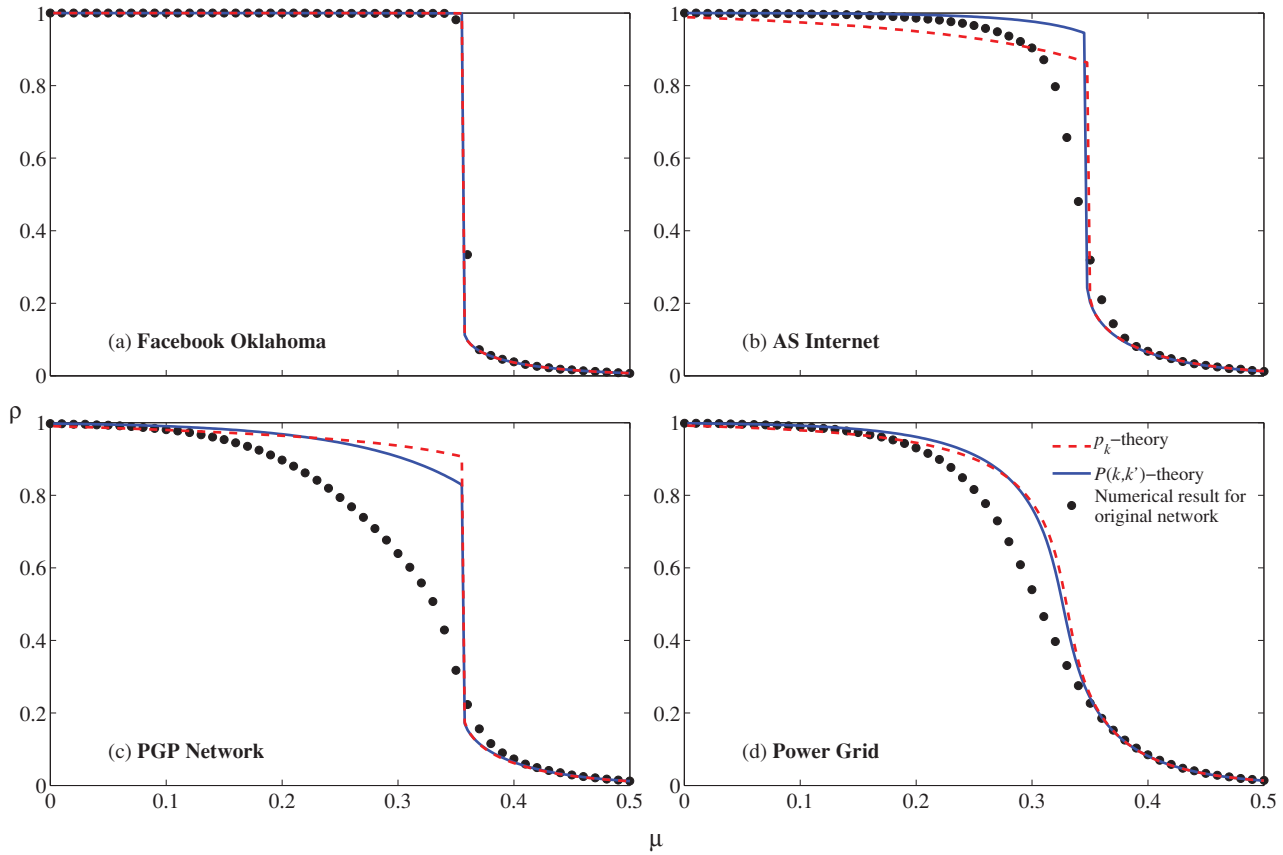

FIG. 3. (Color online) Watts threshold model, with threshold mean μ and variance $\sigma^2 = 0.04$, for the networks from Fig. 1. We use the seed fraction $\rho_0 = 0$ because the nodes with negative thresholds immediately turn on and act as seeds. In other words, the effective seed fraction is given by the value of the cumulative distribution function of the thresholds evaluated at zero. This fraction is $[1 + \mathrm{erf}\{-\mu/(\sigma\sqrt{2})\}]/2$.

development of the infection [55]. In view of this restriction, the results of Fig. 4 are consistent with those of Figs. 1–3. That is, the $P(k,k')$-theory once again provides accurate results for certain networks but is rather inaccurate for other networks.

III. 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. Because Figs. 1–4 demonstrate that this characterization holds for several processes, hereafter we concentrate primarily on the example of bond percolation.

A. Watts-Strogatz networks

Using the small-world networks introduced by Watts and Strogatz [33], one can conduct a systematic study of the effects of the clustering coefficient C and the mean intervertex distance ℓ. We start with a ring of $N = 10\,000$ nodes and connect each node to $z = 10$ nearest neighbors. We then randomly rewire a fraction f of the links in the network [56]. When $f = 0$, the values of C and ℓ are both high. When $f = 1$, the rewired network is randomized (that is, the node degrees are preserved, but everything else is random), which yields low C and ℓ values. For each value of f between 0 and 1, we numerically 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 $S_{\mathrm{th}}(p)$, gives a quantitative measure for the inaccuracy of the theory for this particular value of the rewiring parameter f. We define the error measure

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

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 mean vertical distance between the $S_{\mathrm{th}}(p)$ and $S_f(p)$ curves for $p \in [0,1]$.

In Fig. 5, 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 [57]. We highlight the similar scaling of ℓ_f and E_f in the inset of Fig. 5, in which we plot $\ell_f - \ell_1$ directly as a function of E_f for $f \geqslant 10^{-2}$. The approximately linear dependence that we observe contrasts to the clearly nonlinear relation between E_f and the clustering coefficient C_f that we show in the same inset. This strongly

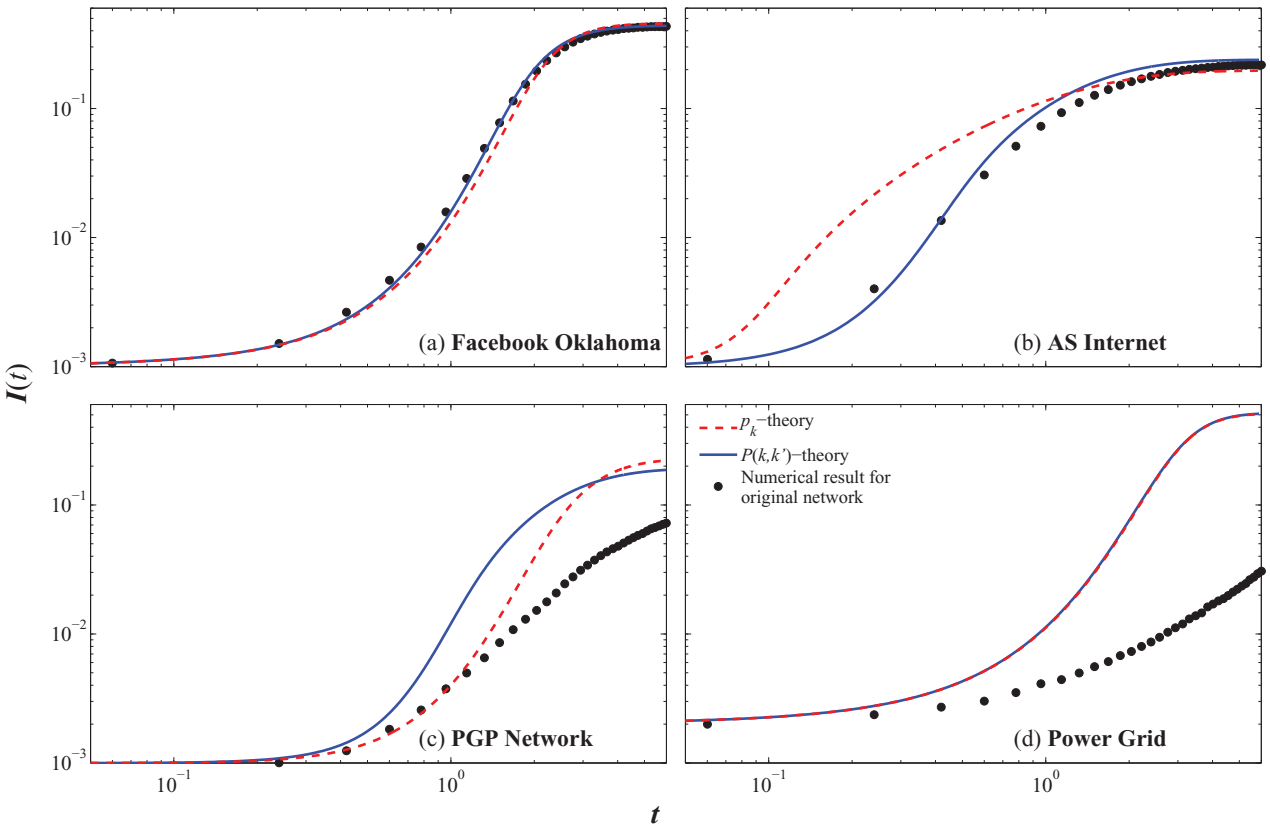

FIG. 4. (Color online) SIS dynamics, which we display as plots of infected fraction $I(t)$ versus time t for the networks from Fig. 1. The parameters in Eq. (17) of Ref. [6] are the recovery rate μ and the spreading rate λ. We use the value $\mu = 1$ in all figure panels; we use $I(0) = 10^{-3}$ in panels (a)–(c) and $I(0) = 0.002$ in panel (d); we use the value $\lambda = 0.02$ in panel (a); $\lambda = 0.2$ in panels (b) and (c); and $\lambda = 0.8$ in panel (d).

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

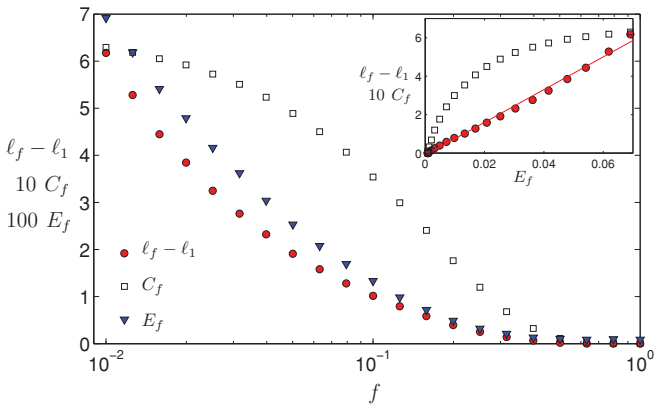

FIG. 5. (Color online) Watts-Strogatz small-world network: We plot $\ell_f - \ell_1$ (red circles), $10\ C_f$ (open squares), and $100\ E_f$ (blue triangles) as functions of rewiring fraction f. (See the main text for the definitions of these quantities.) The inset shows $\ell_f - \ell_1$ and C_f as functions of E_f for $f \geqslant 10^{-2}$. 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.

B. Real-world networks and additional examples

Our 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 coefficient 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 $S_{\text{num}}(p)$ and the theoretical prediction $S_{\text{th}}(p)$:

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

Essentially, E gives the mean distance between the numerics (black disks) and theory (solid blue curve) in Fig. 1. In Fig. 6(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. 6 in order to fully resolve the range of values for both variables.

We also include synthetic examples, such as Watts-Strogatz small-world networks and clustered random networks generated using the models described in Refs. [13,14], which we now briefly recall [58]. The fundamental quantity defining the γ-theory networks of Ref. [14] is the joint probability distribution $\gamma(k,c)$, which gives the probability that a randomly

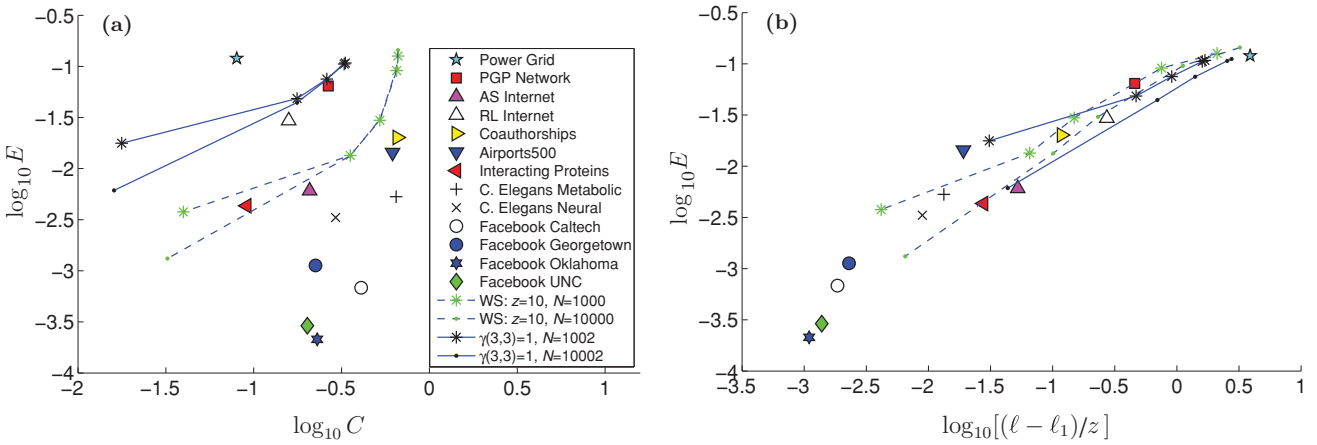

FIG. 6. (Color online) Scatter plots of $\log_{10} E$ versus (a) $\log_{10} C$ (with $R^2 \approx 0.087$) and (b) $\log_{10}[(\ell - \ell_1)/z]$ (with $R^2 \approx 0.94$).

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 $\gamma(k,c) = 0$ 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 Ref. [13], 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 do not belong to the triangles). In each synthetic network, we P-rewire a fraction f of the links and show our results for $f \in \{10^{-3}, 4 \times 10^{-3}, 0.04, 0.1, 0.4\}$.

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. 6(a), we calculate $R^2 \approx 0.087$ (using only the points and ignoring the connecting curves that help identify families of points). This relatively small value indicates that C is not a good predictor of the theory error for the set of networks that we tested (see Table I). After examining a wide range of possibilities (see the scatter plots in Appendix B), 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. 6(b)]. Recall that one can think of such fully P-rewired versions of a network as random networks with the same joint degree-degree distribution $P(k,k')$ and size (that is, number of nodes) 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 matrix $P(k,k')$. If the difference $\ell - \ell_1$ is sufficiently small [for example, if it is less than $z/10$, as is the case in Figs. 1(a)–1(b)], then the $P(k,k')$-theory can be expected to accurately give the GCC size, k-core sizes, and results for several dynamical processes (see Figs. 1–4). 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 dynamical processes on networks that are *not* locally treelike 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. We show in Appendix A that the error measure E depends linearly on $\ell - \ell_1$ in a certain class of networks with zero clustering. Although this theoretical result is restricted in its applicability, it lends weight to our claim that E depends primarily on $\ell - \ell_1$ rather than on the clustering coefficient C.

One possible explanation for the dependence of E on the mean intervertex distance ℓ is the following. The $P(k,k')$-theory assumes that the probability of connection between any two nodes depends only on their degrees and on nothing else. One can refer to this property and the networks that satisfy it as "mixed" to contrast them with real-world networks with community structure (where nodes belonging to different communities are much less likely to be connected than if they were within one community) or networks with a geographic component (with either explicit effects of geography, as in planar graphs, or implicit ones, as in the power grid network or the "sausage-like" networks of Ref. [59], where nodes that are situated far from each other are less likely to be linked together). The mean distance ℓ can be interpreted as a measure of how much the original network is mixed. If the network is well mixed, then ℓ is low (that is, it is similar to the value in a fully P-rewired version of the network) and the $P(k,k')$-theory will work well on such networks. If the network is poorly mixed, then the value of ℓ is higher. When poorly mixed networks are rewired, the decrease in ℓ is suggestive of what is happening: the network community structure or geographic dependence is gradually destroyed as the network becomes better mixed.

The fact that clustering apparently does not play a role [see Fig. 6(a)] might be related to the specific error measure that we define in Eq. (2) and use in this paper. For example, it is possible that clustering is crucial only near the percolation transition point (that is, the value of p at which the GCC emerges) and therefore does not significantly affect the mean vertical distance (2) between the curves for bond percolation. However, geographical or community structure potentially can

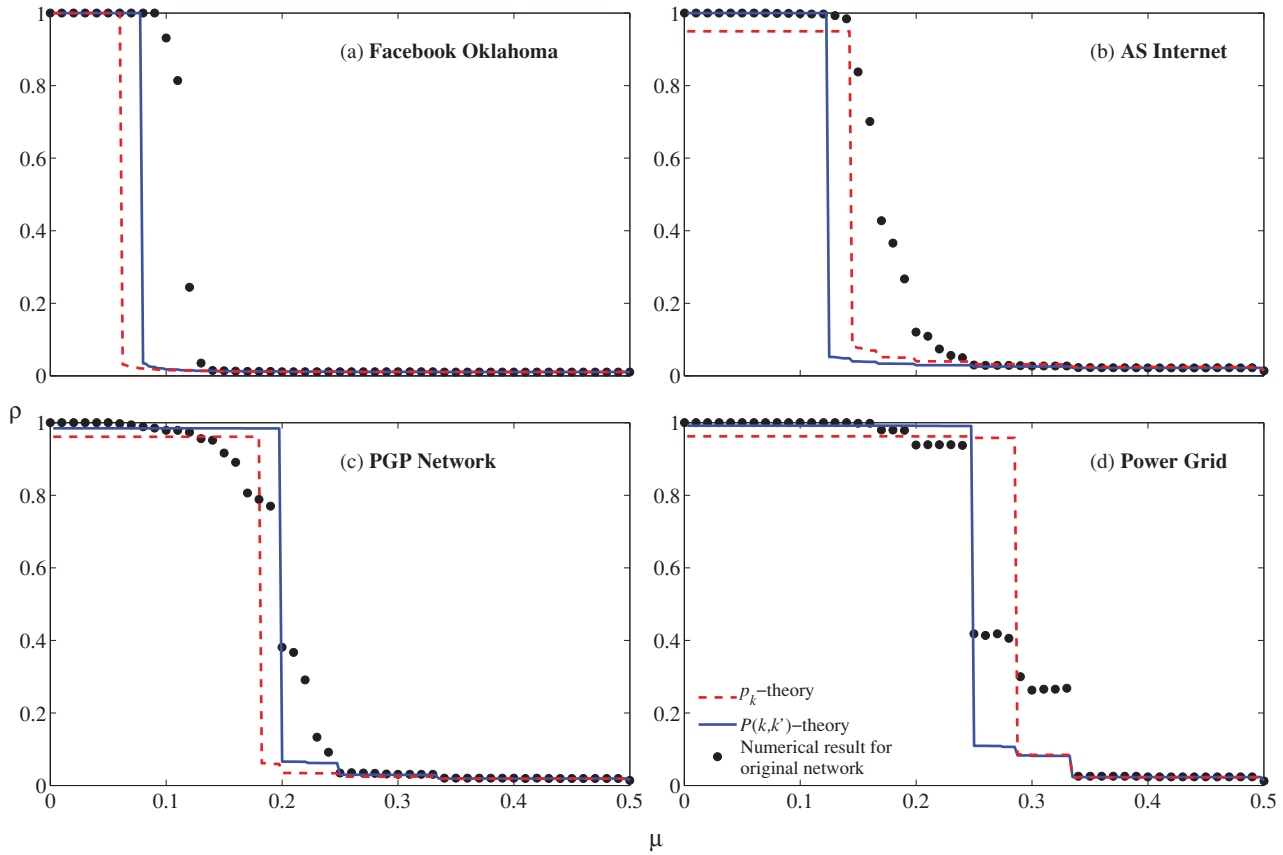

FIG. 7. (Color online) Watts threshold model, with threshold mean μ and variance $\sigma^2 = 0$ (that is, with uniform thresholds) for the networks from Fig. 1. We use a seed fraction of $\rho_0 = 10^{-2}$.

play a role throughout the entire range of p from 0 to 1, leading to a strong correlation between E and ℓ.

IV. 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 joint degree distribution $P(k,k')$ and number of nodes by no more than about 10% of the mean degree z. Surprisingly, the level of clustering seems to be much less important for the accuracy of $P(k,k')$-theory, which is why we found excellent matches between theory and numerical results, even in highly clustered graphs such as Facebook social networks and the AS Internet network.

Although we used bond percolation as our primary example, Figs. 1–4 suggest that if $P(k,k')$-theory is accurate for percolation, then it also works well for other processes. However, any measure of accuracy must, of course, depend on the process under scrutiny. For example, Fig. 7 shows a comparison between theory and numerical results for the Watts threshold model in which $\sigma = 0$, implying that all nodes have identical thresholds equal to μ (in contrast to Fig. 3). This example exhibits different results for theory and numerics even in the Facebook networks. This suggests that the $\sigma = 0$ case of the Watts model is particularly sensitive to deviations of the network from randomness and suggests that this case might provide a suitable testing ground for new analytically tractable models of networks that include clustering [13,14,22].

In summary, we have shown that for a variety of processes, including bond percolation and k-core size calculations, tree-based analytical theory yields highly accurate results for networks in which $\ell \approx \ell_1$ (i.e., when the value of the mean intervertex distance is close to that for an appropriate random network), even in the presence of significant clustering. Such graphs, which include the AS Internet network and Facebook social networks, are definitively not locally treelike, so the theory is working very well even in situations where the theory's fundamental hypothesis is known to fail utterly. The fact that analytical results for several dynamical processes can be expected to apply on "sufficiently small" small-world networks increases the value of existing theoretical work and highlights the types of processes for which improved analytical modeling of clustering effects should most profitably be targeted. We hope that the results of this paper will motivate further research on the underlying causes of this "unreasonable" effectiveness of tree-based theory for clustered networks.

ACKNOWLEDGMENTS

S.M., A.H., and J.P.G. acknowledge funding provided by Science Foundation Ireland under programs 06/IN.1/I366 and MACSI 06/MI/005. M.A.P. acknowledges a research award (no. 220020177) from the James S. McDonnell Foundation.

P.J.M. was funded by the National Science Foundation (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. We thank Alessandro Vespignani for useful comments.

APPENDIX A: THE RELATIONSHIP BETWEEN PREDICTION ERROR AND MEAN INTERVERTEX DISTANCE

We consider the class of networks for which one can define a branching matrix [60]. A branching matrix describes the connection probabilities in treelike networks with nontrivial structure—for example, in modular networks [61]. In this Appendix, we derive how the error measure E defined in Eq. (2) depends on $\ell - \ell_1$ for a network with a branching matrix when the network is close to fully P-rewired (that is, when it is close to a random network with the same joint degree-degree distribution). We give the final formula in Eq. (A6). Because clustering is negligible in these infinite networks, E cannot depend on the clustering coefficient C. In Fig. 6, we illustrate the relationship between E and C and between E and $(\ell - \ell_1)/z$ for real-world networks.

The branching matrix characterizes the mean intervertex distance ℓ in a network, and it also determines the bond-percolation behavior. The largest eigenvalue of the branching matrix, which we denote by λ, determines the percolation threshold:

$$p_{\text{th}} = \frac{1}{\lambda}\,. \tag{A1}$$

Additionally, an estimate of the mean intervertex distance can be written in terms of λ as [60]

$$\ell \approx \frac{\ln N}{\ln \lambda}\,, \tag{A2}$$

where N denotes the number of nodes in the network.

We now suppose that the network is almost fully P-rewired, and we consider how values of λ that differ from the fully P-rewired value (which we denote by λ_1) affect the values of ℓ and p_{th}. Note that it is easy to calculate λ_1, as the branching matrix of a fully P-rewired network is given in terms of the degree correlation matrix $P(k,k')$ by [60]

$$B_1(k,k') \equiv (k'-1)\frac{P(k,k')}{\sum_j P(k,j)} = (k'-1)\frac{P(k,k')}{kp_k/z}\,, \tag{A3}$$

and λ_1 is the largest eigenvalue of B_1. Moreover, for uncorrelated networks produced using the configuration model, $\lambda_1 = \sum_k k(k-1)p_k/z$. This implies that $\lambda_1 = z - 1$ for graphs in which all nodes have the same degree (such as P-rewired Watts-Strogatz networks and the special cases of γ-theory networks used in Sec. III).

Considering only small deviations from fully P-rewired values, we write $\lambda = \lambda_1 + \Delta\lambda$ and $\ell = \ell_1 + \Delta\ell$. Expanding to linear terms, we find from (A2) that the excess length is

$$\Delta\ell = -\frac{\Delta\lambda\,\ln N}{\lambda_1(\ln\lambda_1)^2}\,. \tag{A4}$$

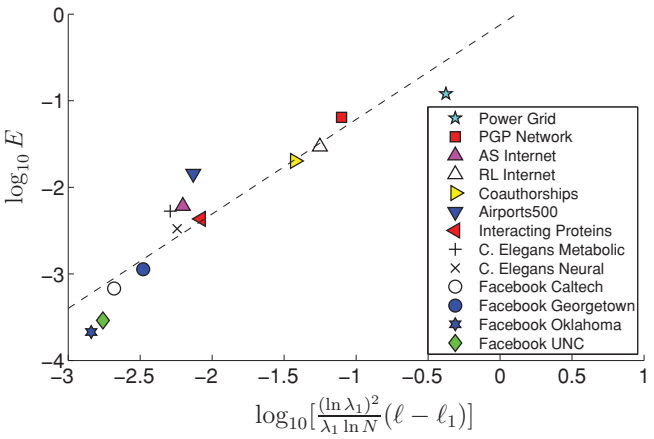

FIG. 8. (Color online) Log-log scatter plot of actual (numerical) values of E for real-world networks versus the values predicted by Eq. (A6), for which we numerically calculate ℓ and ℓ_1. We find that $R^2 \approx 0.87$; the slope of the fitted line is 1.09.

Similarly, we find from (A1) that the change in percolation threshold is

$$\Delta p_{\text{th}} = -\frac{\Delta\lambda}{\lambda_1^2}\,. \tag{A5}$$

If we now make the further assumption that Δp_{th} is approximately equal to the error E for the bond-percolation process [this approximation is exact if the effect of the perturbation is to shift the entire bond-percolation curve $S(p)$ to $S(p + \Delta p_{\text{th}})$], we obtain the relation

$$E \approx \frac{(\ln\lambda_1)^2}{\lambda_1 \ln N}(\ell - \ell_1)\,. \tag{A6}$$

Although the scope of our analysis is obviously limited by our assumptions, Eq. (A6) nevertheless supports our main claim that E depends primarily on the excess length $\ell - \ell_1$. Note that $C = 0$ for branching-matrix networks, so E is (trivially) independent of C. Compare this to the results for real-world networks that we show in Fig. 6(a). Moreover, the scatter plot of $\log_{10} E$ versus $\log_{10}[(\ln\lambda_1)^2(\ell - \ell_1)/(\lambda_1 \ln N)]$ in Fig. 8 indicates that Eq. (A6) gives a good fit ($R^2 \approx 0.87$) even for real-world networks.

APPENDIX B: SCATTER PLOTS

In this Appendix, we show scatter plots of $\log_{10} E$ versus a variety of possible predictors (see Fig. 9). Recall that E, which we defined in Eq. (2), gives an error measure for bond percolation. We test the possible dependence of E on various combinations of the mean degree z, mean intervertex distance ℓ, and clustering coefficients [62]. Recall again that ℓ_1 denotes the value taken by ℓ in a fully P-rewired version of a network (that is, in a random network with the same size and joint degree-degree distribution).

The scatter plots show data points for real-world networks, Watts-Strogatz small-world networks, and γ-theory networks (which are described in Sec. III B). 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

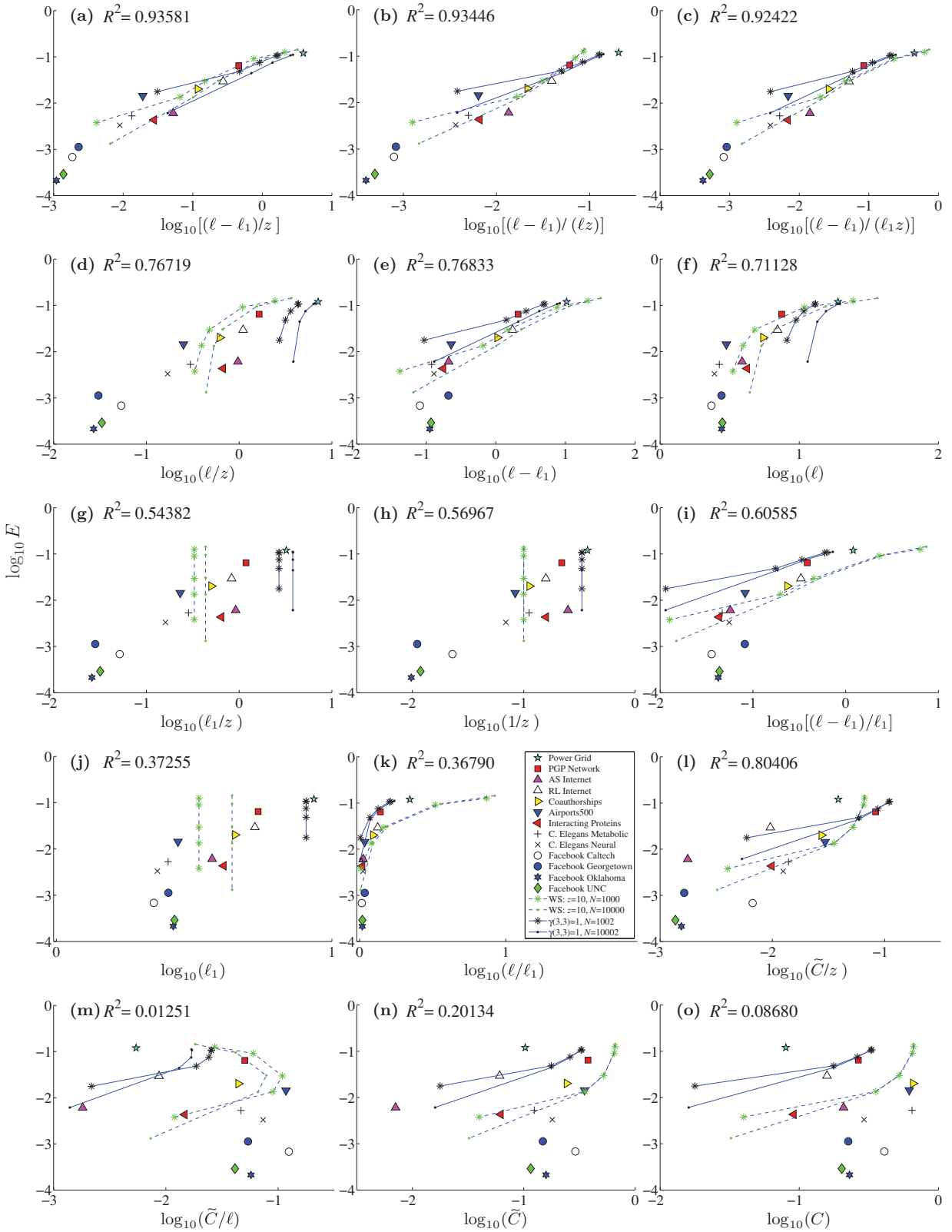

FIG. 9. (Color online) Scatter plots of error E versus various error predictors.

$R^2 < 0.3$). 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. 6(b) and the scaling function $F = \ln^2\lambda_1/(\lambda_1 \ln N)$ indicated by Eq. (A6) both give satisfactory fits.

[1] S. H. Strogatz, Nature (London) **410**, 268 (2001).

[2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Phys. Rep. **424**, 175 (2006).

[3] A. Barrat, A. Vespignani, and M. Barthelemy, *Dynamical Processes on Complex Networks* (Cambridge University Press, Cambridge, UK, 2008).

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

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

[6] M. Barthélemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, J. Theor. Biol. **235**, 275 (2005).

[7] K. T. D. Eames, Theor. Popul. Biol. **73**, 104 (2008).

[8] M. Garavello and B. Piccoli, *Traffic Flow on Networks* (American Institute of Mathematical Sciences, Springfield, MO, 2006).

[9] A. Arenas, A. Diaz-Guilera, J. Kurths, Y. Moreno, and C. Zhou, Phys. Rep. **469**, 93 (2008).

[10] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2010).

[11] M. E. J. Newman, Phys. Rev. E **68**, 026121 (2003).

[12] M. A. Porter, J.-P. Onnela, and P. J. Mucha, Not. Am. Math. Soc. **56**, 1082 (2009).

[13] M. E. J. Newman, Phys. Rev. Lett. **103**, 058701 (2009).

[14] J. P. Gleeson, Phys. Rev. E **80**, 036107 (2009).

[15] J. P. Gleeson and S. Melnik, Phys. Rev. E **80**, 046121 (2009).

[16] M. Ostilli and J. F. F. Mendes, Phys. Rev. E **80**, 011142 (2009).

[17] M. Á. Serrano and M. Boguñá, Phys. Rev. E **74**, 056114 (2006).

[18] M. Á. Serrano and M. Boguñá, Phys. Rev. E **74**, 056115 (2006).

[19] M. Á. Serrano and M. Boguñá, Phys. Rev. Lett. **97**, 088701 (2006).

[20] P. Trapman, Theor. Popul. Biol. **71**, 160 (2007).

[21] J. C. Miller, J. R. Soc. Interface **6**, 1121 (2009).

[22] J. C. Miller, Phys. Rev. E **80**, 020901(R) (2009).

[23] T. Britton, M. Deijfen, A. N. Lageras, and M. Lindholm, J. Appl. Probab. **45**, 743 (2008).

[24] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).

[25] We assume that the degree distribution has finite variance, as real-world networks necessarily have a finite cutoff in their degree sequence. Conditions when clustering tends to zero in configuration-model networks are discussed, for example, in the lecture notes by R. van der Hofstad, which are available at [http://www.win.tue.nl/∼rhofstad/NotesRGCN2010.pdf].

[26] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter, to appear in SIAM Review; available at e-print arXiv:0809.0690.

[27] 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].

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

[29] A. Vázquez and Y. Moreno, Phys. Rev. E **67**, 015101(R) (2003).

[30] X. Guardiola, R. Guimerà, A. Arenas, A. Diaz-Guilera, D. Streib, and L. A. N. Amaral, e-print arXiv:cond-mat/0206240.

[31] M. Boguñá, R. Pastor-Satorras, A. Diaz-Guilera, and A. Arenas, Phys. Rev. E **70**, 056122 (2004).

[32] Largest connected component of the network of users of the Pretty-Good-Privacy algorithm for secure information interchange, [http://deim.urv.cat/∼aarenas/data/xarxes/PGP.zip].

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

[34] An undirected, unweighted network representing the topology of the Western States Power Grid of the United States, [http://www-personal.umich.edu/∼mejn/netdata/power.zip].

[35] M. E. J. Newman and R. M. Ziff, Phys. Rev. E **64**, 016706 (2001).

[36] A. L. Traud, P. J. Mucha, and M. A. Porter, e-print arXiv:1102.2166.

[37] 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 (that is, it reduces the density of triangles). In applying this algorithm, we also take care to avoid multiple links and self-links.

[38] Internet router-level graph computed from ITDK0304 skitter and iffinder measurements, CAIDA's Internet Topology Data Kit no. 0304, San Diego Supercomputer Center, University of California, San Diego (2003), [http://www.caida.org/tools/measurement/skitter/router_topology/itdk0304_rlinks_undirected.gz].

[39] M. E. J. Newman, Proc. Natl. Acad. Sci. USA **98**, 404 (2001).

[40] Network of coauthorships between scientists posting preprints on the Condensed Matter E-Print Archive. This includes all preprints posted between 1 January 1995 and 31 March 2005, [http://www-personal.umich.edu/∼mejn/netdata/cond-mat-2005.zip].

[41] V. Colizza, R. Pastor-Satorras, and A. Vespignani, Nat. Phys. **3**, 276 (2007).

[42] A network obtained by considering the 500 U.S. airports with the largest amount of traffic from publicly available data, [http://sites.google.com/site/cxnets/US_largest500_airportnetwork.txt].

[43] V. Colizza, A. Flammini, A. Maritan, and A. Vespignani, Physica A **352**, 1 (2005).

[44] V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani, Nat. Phys. **2**, 110 (2006).

[45] Protein interaction network of the yeast *Saccharomyces cerevisae* extracted with different experimental techniques and collected at the Database of Interacting Proteins, [http://dip.doe-mbi.ucla.edu/]; [http://sites.google.com/site/cxnets/DIP.dat].

[46] J. Duch and A. Arenas, Phys. Rev. E **72**, 027104 (2005).

[47] A metabolic network of C. Elegans, [http://deim.urv.cat/∼aarenas/data/xarxes/celegans_metabolic.zip].

[48] A neural network of C. Elegans, [http://www-personal.umich.edu/∼mejn/netdata/celegansneural.zip].

[49] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Phys. Rev. Lett. **96**, 040601 (2006).

[50] J. P. Gleeson, Phys. Rev. E **77**, 046117 (2008).

[51] D. J. Watts, Proc. Natl. Acad. Sci. USA **99**, 5766 (2002).

[52] D. Centola, V. M. Eguiluz, and M. W. Macy, Physica A **374**, 449 (2007).

[53] D. Centola and M. Macy, Amer. J. Sociol. **113**, 702 (2007).

[54] J. P. Gleeson and D. J. Cahalane, Phys. Rev. E **75**, 056103 (2007).
[55] The mean-field theory of Ref. [6] implicitly assumes that contacts are broken and reformed at each moment, which is not the case here. However, as indicated in Refs. [63,64], such theories can work well in certain cases.
[56] We employ our P-rewiring algorithm that preserves the degree of each node. It is slightly different from the one used in [33], but this difference is not important here.
[57] When $f \ll 10^{-2}$, the quantity ℓ_f changes much more rapidly with f than E_f does. We focus on the range $f \geqslant 10^{-2}$ in Fig. 5, because for lower f, the values of the error E_f are much larger than those seen in any of the networks we study. (For example, the power grid network has $E_0 \approx 0.11$ and the PGP network has $E_0 \approx 0.065$, which should be compared to the maximum error of 0.07 seen in Fig. 5.)
[58] We could also have used the model in Ref. [22] for generating clustered random networks.
[59] A. Vázquez, M. Boguñá, Y. Moreno, R. Pastor-Satorras, and A. Vespignani, Phys. Rev. E **67**, 046111 (2003).
[60] A. V. Goltsev, S. N. Dorogovtsev, and J. F. F. Mendes, Phys. Rev. E **78**, 051105 (2008).
[61] S. N. Dorogovtsev, J. F. F. Mendes, A. N. Samukhin, and A. Y. Zyuzin, Phys. Rev. E **78**, 056106 (2008).
[62] We consider both common definitions of clustering coefficient. We use C to denote the coefficient defined by Eq. (3.6) of Ref. [24] and $\tilde{C}$ to denote that from Eq. (3.4) of Ref. [24].
[63] A. Baronchelli and V. Loreto, Phys. Rev. E **73**, 026103 (2006).
[64] J. P. Gleeson, S. Melnik, J. Ward, M. A. Porter, and P. J. Mucha, e-print arXiv:1011.3710.