Open access • Journal Article • DOI:10.1103/PHYSREVE.88.062820

Percolation on random networks with arbitrary k-core structure. — Source link

Laurent Hébert-Dufresne, Antoine Allard, Jean-Gabriel Young, Louis J. Dubé

Institutions: Laval University

Published on: 30 Dec 2013 - Physical Review E (American Physical Society)

Topics: Continuum percolation theory, Complex network, Random graph, Degree distribution and Percolation

Percolation on random networks with arbitrary *k*-core structure

Laurent Hébert-Dufresne,* Antoine Allard,* Jean-Gabriel Young, and Louis J. Dubé
Département de Physique, de Génie Physique, et d'Optique, Université Laval, Québec (Québec), Canada G1V 0A6
(Received 11 September 2013; published 30 December 2013)

The k-core decomposition of a network has thus far mainly served as a powerful tool for the empirical study of complex networks. We now propose its explicit integration in a theoretical model. We introduce a hard-core random network (HRN) model that generates maximally random networks with arbitrary degree distribution *and* arbitrary k-core structure. We then solve exactly the bond percolation problem on the HRN model and produce fast and precise analytical estimates for the corresponding real networks. Extensive comparison with real databases reveals that our approach performs better than existing models, while requiring less input information.

DOI: 10.1103/PhysRevE.88.062820 PACS number(s): 64.60.aq, 64.60.ah

I. INTRODUCTION

We address the challenge of designing a realistic model of complex networks while preserving its analytic tractability. The model should include the essential structural properties of real networks, and the theoretical framework should guarantee easy access to quantitative calculations. For the second aspect of this endeavour, we cast our analysis in terms of a percolation problem. This has been a topic of choice for some years since it can just as well represent the dynamics of a network as the dynamics on the network [1–8]. One might think of its growth, its robustness (to attacks or failures), and the propagation of emerging infectious agents (e.g., disease or information).

While the study of percolation models on idealized networks has led to a better understanding of both the processes they model and the networks that support them, the study of percolation on real networks has somewhat stagnated. Unfortunately, purely numerical approaches are time consuming, require a complete description of the networks under scrutiny, and lack the insights of an analytical description. Conversely, although analytical modeling provides a better understanding of the organization of real networks, they are limited at present to simplified random models [see Refs. [6,9], and references therein].

In this paper we demonstrate how the k-core structure of networks (hereafter simply core structure) plays a central role in the outcome of bond percolation, and how it acts as a proxy that captures the essential structural properties of real networks. The ensuing model, that we call the hard-core random network (HRN) model, creates maximally random networks with an arbitrary degree distribution *and* an arbitrary core structure. We also propose a Metropolis-Hastings algorithm to generate such random networks. The HRN model serves our purpose well since it is shown to be amenable to an exact solution for the size of the extensive "giant" component (in the limit of large network size). With less input information, it outperforms the current standard model [10] for precise prediction of percolation results on real networks.

The organization of this paper goes as follows. In Sec. II we introduce the bond percolation problem and briefly present the two models used for comparison. In Sec. III we present the HRN model, the equations used to solve the bond percolation problem, and the Metropolis-Hastings algorithm generating the corresponding random networks. We also compare the predictions of the HRN model and the ones of the two aforementioned models with the results obtained numerically using real network databases. Final remarks are collected in the last section.

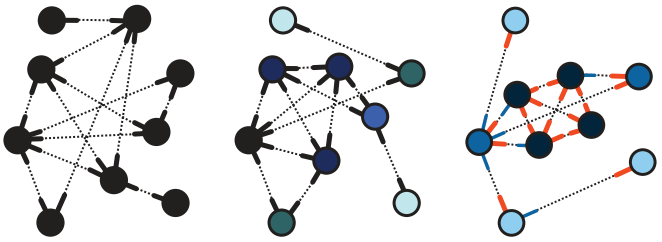
II. BOND PERCOLATION ON NETWORKS

The bond percolation problem concerns the connectivity of a network after the removal of a fraction $1 - T$ of its edges. More precisely, for a synthetic or empirical network, we are interested in the fraction S of nodes contained in the largest connected component—the giant component—after each edge has been removed independently with a probability $1 - T$. In the limit of large networks, this component undergoes a *phase transition* at a critical point T_c during which its size (the number of nodes it contains) becomes an extensive quantity that scales linearly with the number of nodes (N) of the whole network [11].

To compare and assert the precision of the predictions of our model, we use the *configuration model* (CM) and *correlated configuration model* (CCM) [12–15] as benchmarks [see Figs. 1(a) and 1(b)]. These models define maximally random network ensembles that are random in all respects other than the degree distribution (CM, CCM) and the degree-degree correlations (CCM). The degree distribution $\{P(k)\}_{k\in\mathbb{N}}$ is the distribution of the number of connections (the degree k) that nodes have. The degree-degree correlations are defined through the *joint degree distribution* $\{P(k,k')\}_{k,k'\in\mathbb{N}}$ giving the probability that a randomly chosen edge has nodes of degree k and k' at its ends.

For both models, the size of the giant component S and the percolation threshold T_c can be calculated in the limit $N \to \infty$ using probability generating functions (pgfs) [12–19]. To model bond percolation on a given network with these models, we simply extract the degree distribution and the joint degree distribution; the required information therefore scales as $k_{\max}$ and $k_{\max}^2$. The original network is then found within the random ensembles containing all possible networks that can be designed with the same degree distribution and/or degree-degree correlations. The readers unfamiliar with these models and/or the mathematics involved can get a brief overview of these subjects in Appendices A and B.

*These two authors contributed equally to this work.



(a) CM (b) CCM (c) HRN

FIG. 1. (Color online) Comparison of the three random network models considered. (a) The CM randomly connects stubs drawn from a given degree distribution $\{P(k)\}_{k\in\mathbb{N}}$. (b) The CCM distinguishes nodes according to their degree (colors) and randomly match stubs according to the joint degree distribution $\{P(k,k')\}_{k,k'\in\mathbb{N}}$. (c) The HRN model distinguishes nodes by their coreness (colors) and stubs by their contribution to a node's coreness (thicker red or smaller blue stubs). Stubs are then randomly matched according to the matrices **K** and **C**.

The degree distribution and the joint degree distribution can be seen as the one-point and two-point correlation functions of a network. The next logical step would therefore be to consider three-point correlations (i.e., clustering), and eventually to incorporate mesoscopic features such as motifs, cliques, and communities. Although many theoretical models have been proposed [19–29], a general, objective, and systematic method to tune these models in order to reproduce the features found in real networks as well as to predict the outcome of bond percolation is yet to be found [30].

III. HARD-CORE RANDOM NETWORKS (HRN)

We propose an alternative approach by considering a macroscopic measure of centrality: the *coreness* of nodes. This choice is motivated by the recent observation that a node's coreness is a better indicator of the likeliness for that node to be part of the giant component than its degree [31]. This measure also has the advantage of being general, objective, systematic, and easily calculated [32].

A. Network coreness

The coreness c of a node is specified through its position in the core decomposition of a network [33,34]. This decomposition assigns nodes to nested cores where nodes belonging to the nth core all share at least n edges with one another. A node has a coreness equal to c if it is found in the cth core, but not in the $(c+1)$th core. The set of nodes with a coreness equal to c forms the c shell.

This definition of the coreness may appear complicated to compute, but a simple algorithm allows us to do the decomposition very efficiently [32].

Input graph as lists of nodes $\mathcal{V}$ and neighbors $\mathcal{N}$
Output list $\mathcal{C}$ with coreness for each node
compute and list the degrees $\mathcal{D}$ of nodes;
sort $\mathcal{V}$ with increasing degree of nodes;
for each $v \in \mathcal{V}$ in the order of $\mathcal{V}$ **do**
  $\mathcal{C}(v) := \mathcal{D}(v)$;
  for each $u \in \mathcal{N}(v)$ **do**
    if $\mathcal{D}(u) > \mathcal{D}(v)$ **then**
      $\mathcal{D}(u) := \mathcal{D}(u) - 1$;
    end if
  end for
  re-sort $\mathcal{V}$ accordingly
end for

In short, this algorithm is similar to a *pruning* process which removes nodes in order of their effective degree, i.e., their number of links shared with nodes currently ranked higher in the process. In the end, the coreness of a node is simply given by its degree once the peeling process reaches this particular node. Hence, we know that a node of degree k and coreness c has c *contributing* edges and $k-c$ *noncontributing* edges. Based on this key observation, we develop a coreness-based random network model that defines a maximally random network ensemble with an arbitrary degree distribution *and* an arbitrary core structure.

B. The HRN model

The only two inputs of the HRN model are a **K** matrix whose elements K_{ck} correspond to the fraction of the nodes that have a coreness c and a degree k, and a matrix **C** whose elements $C_{cc'}$ give the fraction of edges that leave nodes of coreness c to nodes of coreness c'. As this model considers undirected networks, the matrix **C** is symmetric and each edge is counted twice to account for both directions.

The HRN model is a multitype version of the CM [18,19,35] in which each node is assigned to a type, its coreness, and in which edges are formed by randomly pairing stubs that either contribute to the node's coreness (say, *red* stubs) or do not contribute to it (say, *blue* stubs). Red stubs from nodes of coreness c may be paired with blue stubs from nodes of coreness $c' \geqslant c$, or with red stubs attached to nodes of coreness $c' = c$ (intrashell). Blue stubs stemming from nodes of coreness c may only be matched with red stubs stemming from nodes with a coreness $c' \leqslant c$. Blue stubs may never be paired together.

These rules enforce a minimal core structure, although random variations can bring nodes to a higher coreness than originally intended. For example, three nodes of original state $(k=2, c=1)$ could end up in the 2 shell in the unlikely event that they form a triangle. However, such random variations may never pull nodes to a lower coreness than intended, in addition to being extremely unlikely in the limit of large networks ($N \to \infty$). The matrices **K** and **C** (see Appendix C for consistency conditions) combined with the aforementioned stub pairing rules define a maximally random network ensemble with an arbitrary degree distribution and core structure [see Fig. 1(c)].

The **K** matrix encodes several useful quantities. For instance, the fraction of nodes of coreness c,

$$w_c = \sum_k K_{ck}, \tag{1}$$

and the associated joint degree distribution, i.e., the probability that a randomly chosen node of coreness c has k_r red stubs and k_b blue stubs,

$$P_c(\boldsymbol{k}) \equiv P_c(k_r,k_b) = \frac{\delta_{c,k_r}}{w_c} K_{k_r,k_r+k_b}, \tag{2}$$

where δ_{c,k_r} is the Kronecker delta. Furthermore, we can extract the average degree of nodes of coreness c,

$$\langle k\rangle_c = \frac{1}{w_c}\sum_k k K_{c,k}, \tag{3}$$

and the average degree of the whole network

$$\langle k\rangle = \sum_{c,k} k K_{ck}. \tag{4}$$

It follows from the above definition that a fraction $w_c\langle k\rangle_c/\langle k\rangle$ of stubs stems from nodes of coreness c, of which a fraction $w_c c/\langle k\rangle$ is red and a fraction $w_c(\langle k\rangle_c - c)/\langle k\rangle$ is blue.

The $\mathbf{C}$ matrix encodes the transition probability $R(c',j|c,i)$ that a node of coreness c through a stub of color i [red (r) or blue (b)] leads to a node of coreness c' through one of its stubs of color j. Since intershell edges can only be formed by matching a red with a blue stub, we readily obtain

$$R(c',b|c,r) = \frac{C_{cc'}}{w_c c/\langle k\rangle}, \tag{5a}$$

$$R(c,r|c',b) = \frac{C_{cc'}}{w_c(\langle k\rangle_c - c)/\langle k\rangle}, \tag{5b}$$

$$R(c',r|c,b) = R(c,b|c',r) = 0, \tag{5c}$$

for $c < c'$. Similarly, as the pairing of blue stubs is forbidden [$R(c',b|c,b) = 0$ for any c and c'], a blue stub stemming from a node of coreness c leads to a node belonging to the same shell (through its red stub) with probability

$$R(c,r|c,b) = \frac{w_c(\langle k\rangle_c - c)/\langle k\rangle - \sum_{c''<c} C_{cc''}}{w_c(\langle k\rangle_c - c)/\langle k\rangle}. \tag{5d}$$

This last result is computed by subtracting the number of blue stubs leading to outer shells (i.e., lower coreness) to the total number of blue stubs stemming from nodes of coreness c, and then by normalizing [$\sum_{c',j} R(c',j|c,i) = 1$ for $c \in \mathbb{N}$ and $i \in \{r,b\}$]. Finally, symmetry with Eq. (5d) implies that

$$R(c,b|c,r) = \frac{w_c(\langle k\rangle_c - c)/\langle k\rangle - \sum_{c''<c} C_{cc''}}{w_c c/\langle k\rangle}, \tag{5e}$$

and normalization leads to

$$R(c,r|c,r) = \frac{2w_c c/\langle k\rangle - C_{cc} - 2\sum_{c''>c} C_{cc''}}{w_c c/\langle k\rangle}, \tag{5f}$$

where we have used the fact that $\sum_{c''} C_{cc''} = w_c\langle k\rangle_c/\langle k\rangle$.

To compute the size of the giant component in the limit of large networks ($N \to \infty$), we define a probability generating function (pgf)

$$g_c(\boldsymbol{x}) = \sum_{\boldsymbol{k}} P_c(\boldsymbol{k}) \prod_i \left[(1-T) + T\sum_{c',j} R(c',j|c,i) x_{c'j}\right]^{k_i} \tag{6}$$

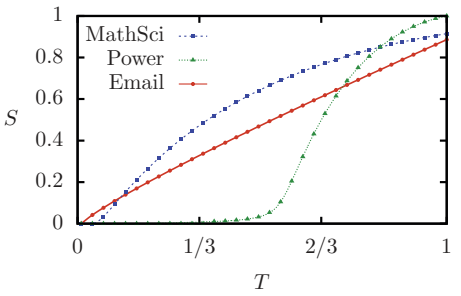


FIG. 2. (Color online) Validation of the HRN model. The predictions of Eqs. (9) and (10) (lines) are compared with the results obtained on networks generated with the Metropolis-Hastings algorithm described in Sec. III C (symbols). The matrices $\mathbf{K}$ and $\mathbf{C}$ were extracted from an email network, the MathSciNet co-authorship network, and a power grid chosen for their different behaviors (see Table I for data set details). Numerical results (symbols) represent the average value of over 5×10^5 simulations performed on networks with more than 3×10^5 nodes.

that generates the distribution of the number of nodes of each type (i.e., coreness c') that can be reached from a node of coreness c (the subscript j of the variable $x_{c'j}$ indicates the color of the stubs from which the node has been reached). To understand this last equation, consider a stub of color i stemming from a node of coreness c. This stub leads to an edge that has been removed with probability $1-T$, or leads to a node of coreness c' through one of its stubs of color j with probability $T R(c',j|c,i)$. Since both the stub pairing and the edge removal are done randomly and independently, the distribution of the number of nodes that are neighbors of a node of coreness c having k_i stubs of color i is generated by the pgf $[(1-T) + T\sum_{c',j} R(c',j|c,i)x_{c'j}]^{k_i}$, a multinomial distribution. Multiplying the pgfs for both stub colors (neighborhood from stubs of different colors are also independent) and averaging over the distribution of the number of stubs of each color that nodes of coreness c have, $P_c(\boldsymbol{k})$, leads to Eq. (6).

Similarly, if the node had previously been reached via one of its red stubs, the distribution of its neighbors reachable via its *other* stubs—its *excess* degree distribution—is simply generated by the pgf

$$f_{cr}(\boldsymbol{x}) = \sum_{\boldsymbol{k}} P_c(\boldsymbol{k}) \prod_i \left[1 - T + T\sum_{c',j} R(c',j|c,i) x_{c'j}\right]^{k_i-\delta_{ir}}. \tag{7}$$

Finally, if the node had been reached via one of its blue stubs instead, the distribution of its neighbors reachable via its other stubs is generated by the pgf

$$f_{cb}(\boldsymbol{x}) = \sum_{\boldsymbol{k}} \frac{k_b P_c(\boldsymbol{k})}{\langle k\rangle_c - c} \times \prod_i \left[1 - T + T\sum_{c',j} R(c',j|c,i) x_{c'j}\right]^{k_i-\delta_{ib}}. \tag{8}$$

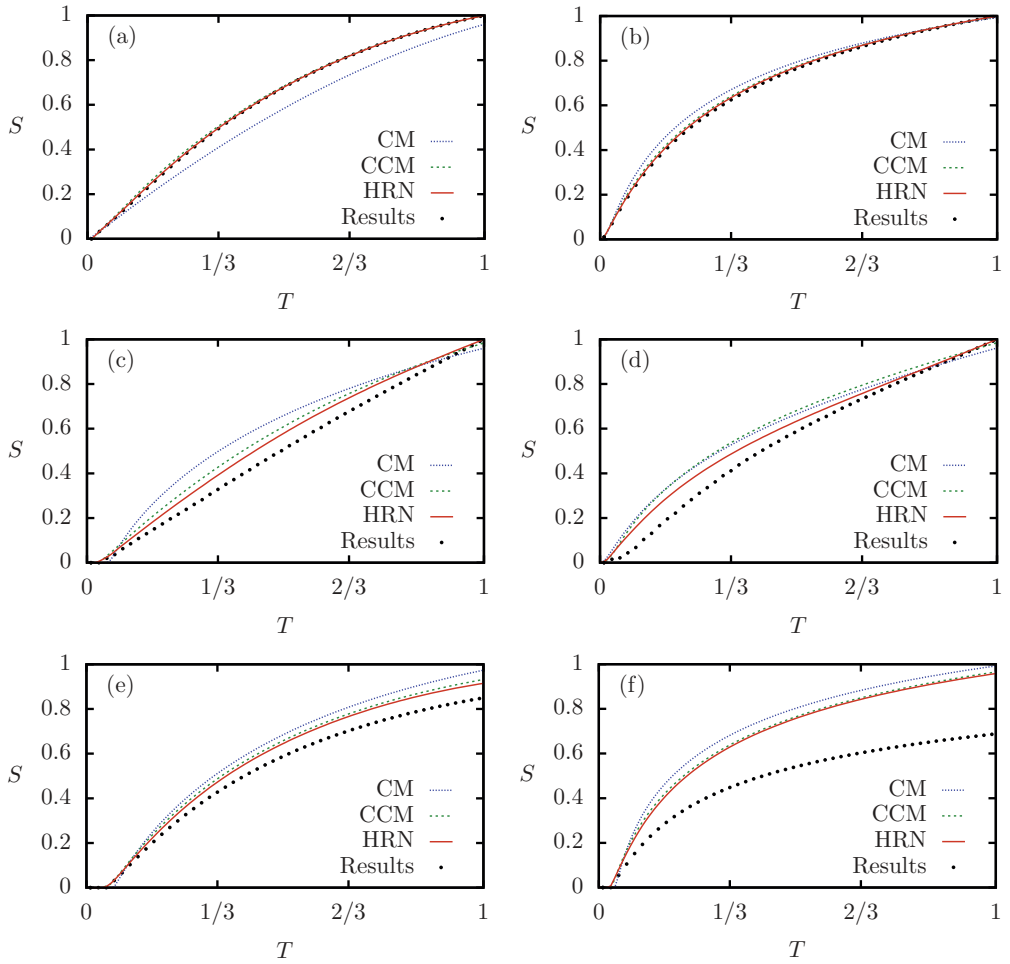


FIG. 3. (Color online) Results of bond percolation on real networks (black dots) compared with analytical predictions obtained with the CM (short dashed blue), CCM (long dashed green), and HRN (full red). The networks are: (a) Internet at the level of autonomous systems, (b) a snapshot of the Gowalla location-based social network, (c) the Pretty-Good-Privacy trust network, (d) a subset of the World Wide Web, (e) the co-authorship network of MathSciNet before 2008, and (f) a large subset of the Facebook social network. See Table I for further details.

In this case, the probability over which $[(1-T)+T\sum_{c',j} R(c',j|c,i)x_{c'j}]^{k_i}$ must be averaged is weighted by the number of blue stubs that the node has (the denominator $\langle k\rangle_c - c$ is simply the normalization constant). For instance, since stubs are paired randomly, a randomly chosen blue stub is ten times more likely to belong to a node that has ten blue stubs than a node that has one.

These pgfs in hand, the size of the giant component can be expressed as [35]

$$S = 1 - \sum_c w_c g_c(\boldsymbol{a}), \tag{9}$$

where $\boldsymbol{a} \equiv \{a_{ci}\}_{c\in\mathbb{N},i\in\{r,b\}}$ is the probability that a node of coreness c reached by one of its stubs of color i does not belong to the giant component. More precisely, a node of coreness c belongs to the giant component if at least one of its neighbors belongs to it, which happens with probability $1-g_c(\boldsymbol{a})$. The size of the giant component is then obtained by averaging this probability over the fraction of nodes that are of coreness c. The probabilities $\boldsymbol{a} \equiv \{a_{ci}\}_{c\in\mathbb{N},i\in\{r,b\}}$ are obtained through a self-consistency argument: If a node of coreness c reached via one of its stubs of color i does not belong to the giant component, then neither should the nodes that can be reached from it. Hence these probabilities correspond to the stable fixed point of the system of equations

$$a_{ci} = f_{ci}(\boldsymbol{a}), \tag{10}$$

with $c \in \mathbb{N}$ and $i \in \{r,b\}$. As the distributions generated by $f_{ci}(\boldsymbol{x})$ are normalized, $\boldsymbol{a} = \boldsymbol{1}$ is always a solution of Eq. (10) and corresponds to the subcritical regime $S = 0$. At $T = T_c$, this fixed point undergoes a transcritical bifurcation and loses its stability to another solution in $[0,1)^{2c_{\max}}$. This supercritical regime corresponds to the existence of a giant component ($S > 0$); the critical point T_c is obtained from a stability analysis of Eq. (10) around $\boldsymbol{a} = \boldsymbol{1}$.

C. Numerical HRN networks

To generate networks with a given core structure, we start with $N \gg 1$ nodes whose number of stubs is drawn from the degree distribution $\{P(k)\}_{k\in\mathbb{N}} = \{\sum_c K_{ck}\}_{k\in\mathbb{N}}$, and randomly match stubs to create edges (as done for the CM [13]). Next, for each node, we assign a coreness c with probability $Q_k(c) = K_{ck}/P(k)$; c of its stubs are then randomly selected

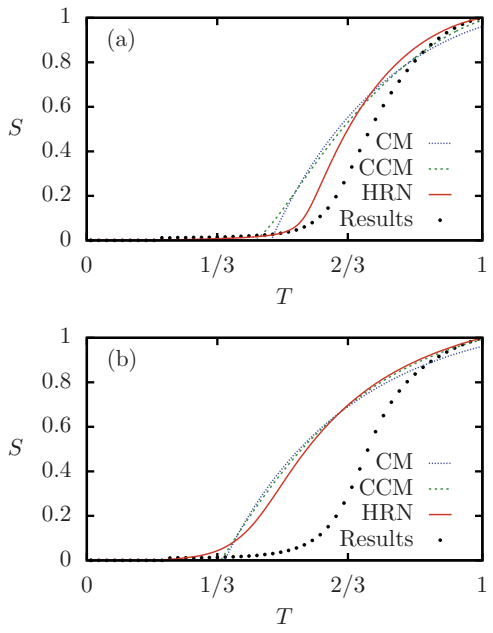


FIG. 4. (Color online) Results of bond percolation on real networks (black dots) compared with analytical predictions obtained with the CM (short dashed blue), CCM (long dashed green), and HRN (full red). The networks are: (a) a subset of the power grid of Poland, and (b) the Western States Power Grid of the United States. See Table I for further details.

as red and the $k - c$ others are identified as blue. Finally, we apply the following Metropolis-Hastings rewiring algorithm (similar to the one proposed in Ref. [14]). At each step, two edges are randomly selected: edge 1 joins nodes of coreness c_1 and c_1' via their respective stubs of color i_1 and j_1 (c_2, i_2, c_2', and j_2 for edge 2). We replace these two edges by edge 3 (c_1, i_1, c_2, and i_2) and edge 4 (c_1', j_1, c_2', and j_2) with probability

$$\min\left\{1, \frac{\Gamma(c_1,i_1;c_2,i_2)\Gamma(c_1',j_1;c_2',j_2)}{\Gamma(c_1,i_1;c_1',j_1)\Gamma(c_2,i_2;c_2',j_2)}\right\},$$

where $\Gamma(c,i;c',j)$ is the wanted fraction of edges that join nodes of coreness c and c' via their respective stubs of color i and j. These fractions are readily obtained from the matrix $\mathbf{C}$ [joint probabilities of Eqs. (5)]

$$\begin{aligned}
&\Gamma(c,r;c',b) = \Gamma(c',b;c,r) = C_{cc'},\\
&\Gamma(c,r;c,b) = \Gamma(c,b;c,r) = w_c(\langle k\rangle_c - c)/\langle k\rangle - \sum_{c''<c} C_{cc''},\\
&\Gamma(c,r;c,r) = 2w_c c/\langle k\rangle - C_{cc} - 2\sum_{c''>c} C_{cc''},
\end{aligned} \tag{11}$$

where $c < c'$, and $\Gamma(c,i;c',j)$ is zero for all other combinations. This procedure preserves the degree distribution, and up to finite-size constraints, has the wanted core structure as its fixed point and is ergodic over the ensemble of networks defined by the HRN model. Figure 2 compares the predictions of Eqs. (9) and (10) with the size of the giant component found in networks generated through this algorithm and shows a perfect agreement.

D. Results

Figures 3 and 4 display the predictions of Eqs. (9) and (10) with the size of the giant component found in real networks (see caption and Table I for a complete description), and with the predictions of the CM and the CCM. These particular networks were chosen to highlight some important results.

TABLE I. Description and properties of the real networks used in Figs. 2–5.

Description	N	$\langle k\rangle$	$k_{\max}$	$c_{\max}$	Fig.	Ref.
Web of trust of the Pretty Good Privacy (PGP) encryption algorithm	10 680	4.55	205	31	3(c), 5	[36]
Structure of the Internet at the level of autonomous systems	22 963	4.22	2390	25	3(a),5	[47]
Large subset of the Facebook social network	63 891	5.74	223	16	3(f), 5	[38]
Snapshot of the Gowalla location-based social network	196 591	9.67	14 730	51	3(b), 5	[39]
Email exchange network from an undisclosed European institution	300 069	2.80	7 631	31	2, 5	[40]
Subset of the World Wide Web	325 729	6.69	10 721	155	3(d), 5	[41]
Co-authorship network of MathSciNet before 2008	391 529	4.46	496	24	2, 3(e), 5	[42]
Subset of the power grid of Poland	3 374	2.41	11	5	2, 4(a), 5	[43]
Western States Power Grid of the United States	4 941	2.67	19	5	4(b), 5	[44]
Email communication within the University Rovira i Virgili	1 134	9.07	1 080	8	5	[45]
Protein-protein interactions in *S. cerevisiae*	2 640	5.00	111	8	5	[45]
Word association graph from the South Florida Free Association norms	7 207	8.82	218	7	5	[45]
Network of hyperlinks between Google's webpages	15 763	18.96	11 401	102	5	[46]
Reply network of the social news website Digg	30 398	5.60	283	9	5	[48]
The cond-mat arXiv co-authorship network circa 2005	30 561	8.24	191	15	5	[45]
Snapshot of the Gnutella peer-to-peer network	36 682	4.82	55	7	5	[37]
Email interchanges between different Enron email addresses	36 692	10.02	1 383	43	5	[49]
Brightkite location-based online social network	58 228	7.35	1 134	52	5	[39]
Network of tagged relationships on the Slashdot news website	77 360	12.13	2 539	54	5	[50]
Friendships between 100 000 Myspace accounts	100 000	16.82	59 108	78	5	[51]
Network of interactions between the users of the English Wikipedia	138 592	10.33	10 715	55	5	[53]
Co-acting network in movies released after December 31st 1999	716 463	21.40	4 625	192	5	[47]

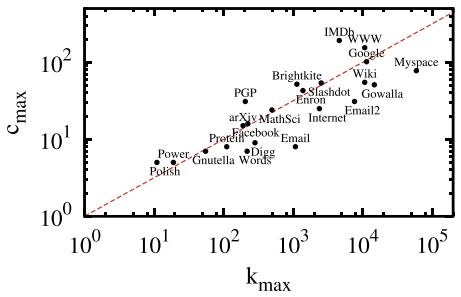


FIG. 5. (Color online) Relation between the highest coreness c_{max} and the highest degree k_{max} for different real networks. The dashed line corresponds to $c_{max} \propto \sqrt{k_{max}}$.

We find that the HRN model performs *at least as well* as the CCM in all investigated cases. First, this observation is interesting as the HRN model requires less input information than the CCM. Indeed the required information scales roughly as $k_{max}c_{max} + c_{max}^2$. As shown in Fig. 5, c_{max} scales approximately as $k_{max}^{1/2}$ in many real networks, hence the input information in the HRN model scales roughly as $k_{max}^{3/2}$. Considering the fact that k_{max} in real networks is often well above 10^2 (see Table I), this difference results in a much faster computation and a major memory gain.

Second, although the HRN model does not account explicitly for the degree-degree correlations, they are effectively captured by the matrices **K** (degree-coreness correlations) and **C** (coreness-coreness correlations). This is confirmed in all available real networks as seen in Figs. 3 and 4 and in Table II (see the correlation coefficient r). The difficulty observed in replicating the Polish power grid correlation coefficient is most likely a coarse-graining effect since power grids a have particularly low number of c shells ($c_{max} = 5$). Table II further investigates the efficiency of the HRN model to reproduce the structural properties of real networks by comparing the clustering coefficients and the mean shortest paths found in both the real networks and their equivalent HRN networks. As expected for any random networks, the HRN model has vanishing clustering. Its relatively good performance for the Internet is a mere consequence of the small size of some of the c shells. Furthermore, the HRN model once more performs at least as well as the CCM in reproducing the mean shortest path. In fact, the extent by which HRN outperforms the CCM in replicating the mean shortest path also appears to be a good indicator of the accuracy gain in predicting percolation.

Moreover, and perhaps surprisingly, we see in Fig. 4(a) that the "S" shape obtained from the Polish power grid, typically due to finite size, is well reproduced by the HRN model, which is formally infinite in size. More precisely, this shape is usually attributed to the finite size of the network ($N = 3374$ for the Polish power grid) as the small components—whose average size formally diverges at $T = T_c$—are misinterpreted as an emerging giant component. Interestingly, the results from the HRN model suggest that this shape is not a numerical artifact of the percolation algorithm, but that it is rather a signature of its geographically embedded nature due to strong *coreness-related* correlations. This unexpected property of the HRN model is confirmed on another, more clustered, Western States power grid in Fig. 4(b). In this case, adding clustering to the HRN is expected to shift its prediction towards higher values of T, i.e., closer to the results from the real network. In fact, the HRN model is more accurate in predicting percolation on the Polish power grid (clustering coefficient $C = 0.02$) than for the Western States power grid ($C = 0.08$). A clustered version of the HRN model seems to offer a promising avenue for the modeling of geographically embedded networks such as power grids.

In this regard, the results of Figs. 3(e) and 3(f) add even more emphasis on the importance of including the effect of clustering in a subsequent version of the HRN model. Indeed, co-authorship networks 3(e) are notoriously clustered networks as authors of a same paper are all connected via a fully connected clique. Similarly, in Facebook 3(f), people belonging to a same social group (e.g., classmates, colleagues, teammates) tend all to be connected to one another, yielding almost fully connected cliques. Again, we expect in this situation that clustering would reduce the size of the giant component (due to redundant connections in cliques), hence bringing the predictions of a clustered HRN model closer to the behaviors observed with the real networks.

TABLE II. Comparison of the clustering coefficient C [44], the degree correlation coefficient r [14], and the mean shortest path ℓ [44] found in three real networks considered in this paper and in the equivalent HRN networks of the same size generated using the algorithm presented in Sec. III C. The three networks correspond to a representative sample of the different behaviors observed with the real networks presented in Table I. The mean shortest paths are expressed as a ratio with the mean shortest paths obtained in the equivalent CCM networks ($\ell^{CCM} = 3.68$, 5.40, and 9.44 for the Internet, PGP, and Polish power grid networks, respectively).

Network	C	r	ℓ/ℓ^{CCM}
Internet	0.230	−0.099	1.04
Internet HRN	0.116	−0.083	1.00
PGP	0.266	0.490	1.39
PGP HRN	0.011	0.508	1.14
Polish grid	0.019	0.677	1.55
Polish grid HRN	0.002	0.528	1.25

IV. CONCLUSION

We have shown that the core structure can be useful beyond the characterization and visualization of networks. It serves modeling efforts well and is efficient in reproducing the structural properties of real networks. Moreover, a few simple connection rules can enforce a core structure in random networks for which the outcome of bond percolation can be predicted with the well-established pgf approach [52]. We feel that this work sets the stage for further improvements (specifically the inclusion of clustering) and paves the way towards a more complete analytical description of percolation on real networks.

ACKNOWLEDGMENTS

The authors would like to acknowledge the financial support of the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada, and the Fonds de recherche du Québec–Nature et technologies.

APPENDIX A: CONFIGURATION MODEL

The most influential quantity with regard to bond percolation on networks is the degree distribution: The distribution of the number of connections (degree) that nodes have. The simplest analytical model that incorporates an arbitrary degree distribution is the CM [12,13]. It defines a maximally random network ensemble that is random in all respects other than the degree distribution $\{P(k)\}_{k\in\mathbb{N}}$: The probability for a randomly chosen node to have a degree equal to k. Networks of this ensemble are generated by creating a set of N nodes, each with a number of stubs drawn from the degree distribution, and then by pairing randomly stubs to form edges.

To compute the size S^{CM} of the giant component and the value T_c^{CM} of the percolation threshold, we define the probability generating function [13]

$$g(x) = \sum_{k=0}^{\infty} P(k)[(1-T)+Tx]^k \tag{A1}$$

that generates the degree distribution. The first derivative of $g(x)$ evaluated at $x = 1$ corresponds to the average degree of the nodes $g'(1) = \langle k\rangle$. We also define

$$f(x) = \frac{g'(x)}{g'(1)} = \frac{1}{\langle k\rangle}\sum_{k'=1}^{\infty} k'P(k')[(1-T)+Tx]^{k'-1} \tag{A2}$$

that generates the number of *other* neighbors of a node that has been reached by following a randomly chosen edge (i.e., the *excess* degree distribution). The size of the giant component is directly obtained via

$$S^{\mathrm{CM}} = 1 - g(a^{\mathrm{CM}}), \tag{A3}$$

where a^{CM} is the probability that a randomly chosen edge does not lead to the giant component. It is the stable fixed point of

$$a^{\mathrm{CM}} = f(a^{\mathrm{CM}}) \tag{A4}$$

in $[0,1]$. The solution $a^{\mathrm{CM}} = 1$ corresponds to the absence of a giant component ($S^{\mathrm{CM}} = 0$). The percolation threshold is the point at which this solution becomes unstable.

To model bond percolation on a given network with the CM, one simply has to extract the degree distribution; the required information therefore scales as $k_{\max}$, the highest degree of the network. The original network is then found within the network ensemble generated by the CM, the ensemble composed of all possible networks one could design with the exact same degree distribution.

APPENDIX B: CORRELATED CONFIGURATION MODEL

Apart from the degree distribution, real networks are typically characterized by strong correlations regarding *who is connected with whom*. One way to include such correlations into a random network model is through the *joint degree distribution* $\{P(k,k')\}_{k,k'\in\mathbb{N}}$ giving the probability that a randomly chosen edge has nodes of degree k and k' at its ends. This yields a *correlated configuration model* (CCM) that defines a maximally random network ensemble having arbitrary degree-degree correlations with a corresponding degree distribution [14,15]. The degree distribution is encoded in $\{P(k,k')\}_{k,k'\in\mathbb{N}}$ through the identity

$$\sum_{k'} P(k,k') = \frac{kP(k)}{\langle k\rangle}. \tag{B1}$$

Generating networks from this ensemble proceeds as for the CM: N nodes, whose degrees are drawn from $\{P(k)\}_{k\in\mathbb{N}}$, are connected via the stub pairing scheme. A Metropolis-Hastings rewiring algorithm [14] is then applied whose fixed point is the network ensemble defined by $\{P(k,k')\}_{k,k'\in\mathbb{N}}$. At each step, two edges are randomly chosen: edge 1 joins nodes m_1 and n_1 with respective degree i_1 and j_1 (m_2, n_2, i_2, and j_2 for edge 2). These two edges are replaced by edge 3 (m_1, m_2, i_1, and i_2) and edge 4 (n_1, n_2, j_1, and j_2) with probability

$$\min\left\{1, \frac{P(i_1,i_2)P(j_1,j_2)}{P(i_1,j_1)P(i_2,j_2)}\right\}. \tag{B2}$$

The size S^{CCM} of the giant component is computed as in the CM [14]

$$S^{\mathrm{CCM}} = 1 - \sum_{k=0}^{\infty} P(k)[(1-T)+Ta_k]^k = 1 - g(\boldsymbol{a}), \tag{B3}$$

where $\boldsymbol{a} = \{a_k\}_{k\in\mathbb{N}}$ is the set of probabilities that an edge leading toward a node with a degree k is not attached to the giant component. They correspond to the stable fixed point in $[0,1]^{k_{\max}}$ of the system of equations

$$a_k = \frac{\sum_{k'} P(k,k')[(1-T)+Ta_{k'}]^{k'-1}}{\sum_{k'} P(k,k')}, \tag{B4}$$

with $k \in \mathbb{N}$. The value T_c of the percolation threshold is the value for which the fixed point $\boldsymbol{a} = \mathbf{1}$ of Eqs. (B4) becomes unstable.

To model bond percolation on a given network with the CCM, one simply has to extract the joint degree distribution. This is achieved by scanning the degree of the two nodes at the end of each edge of the network; the required information therefore scales as $k_{\max}^2$. The original network is then found within the random network ensemble of all networks with the same degree distribution and degree-degree correlations. Note that this ensemble is a subset of the ensemble generated by the CM with the same degree distribution.

APPENDIX C: CONSISTENCY CONDITIONS ON K AND C

The consistency conditions on the matrices $\mathbf{K}$ and $\mathbf{C}$ can be summarized as follows: They must encode an ensemble of *closed* networks. In other words, *all stubs must be paired*, and this must be done in accordance with the stubs matching rules (e.g., two blue stubs cannot be paired). Consequently, there is no k-core structure that the HRN model cannot model as long as it is realistic. This will always be the case when $\mathbf{K}$ and $\mathbf{C}$ are extracted from real networks.

First, there must be as many edges leaving nodes of coreness c toward nodes of coreness c' as there are in the opposite

direction. This requires that $C_{cc'} = C_{c'c}$, a condition that is always fulfilled since $\mathbf{C}$ is defined as a symmetric matrix

$$\mathbf{C} = \mathbf{C}^{\mathrm{T}}. \tag{C1}$$

Second, the degree of each node is bounded from below by its coreness, hence

$$K_{ck} = 0 \quad \text{for} \quad k < c. \tag{C2}$$

Third, both $\mathbf{K}$ and $\mathbf{C}$ must prescribe the same number of stubs stemming from nodes of coreness c,

$$\sum_k k K_{ck} = \langle k \rangle \sum_{c'} C_{cc'}, \tag{C3}$$

where the extra factor $\langle k \rangle$ accounts for the fact that $\mathbf{K}$ "counts" nodes, whereas $\mathbf{C}$ counts stubs (i.e., multiplying both sides by the number of nodes N yields absolute numbers instead of *per capita* averages). Finally, as the coreness of the nodes defines their number of red stubs, the matrix $\mathbf{C}$ is subjected to the following additional constraints for every c:

$$\langle k \rangle \sum_{c'>c} C_{cc'} \leqslant w_c c \leqslant \langle k \rangle \sum_{c' \geqslant c} C_{cc'}. \tag{C4}$$

The first inequality states that there must be at least as many red stubs stemming from nodes of coreness c as there are edges leaving the c shell toward nodes of higher coreness. Equality then means that all red stubs lead to nodes of higher coreness. The second inequality states that all red stubs must lead to nodes of coreness c or higher. Equality occurs when all blue stubs are directed toward nodes of coreness $c' < c$. A similar expression to (C4) can be derived for blue stubs

$$\langle k \rangle \sum_{c'<c} C_{cc'} \leqslant (\langle k \rangle_c - c) w_c \leqslant \langle k \rangle \sum_{c' \leqslant c} C_{cc'}, \tag{C5}$$

and can be interpreted analogously.

[1] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).
[2] L. A. Meyers, Bull. Am. Math. Soc. **44**, 63 (2007).
[3] A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno, and Z. Changsong, Phys. Rep. **469**, 93 (2008).
[4] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Rev. Mod. Phys. **80**, 1275 (2008).
[5] R. Cohen and S. Havlin, *Complex Networks: Structure, Robustness and Function* (Cambridge University Press, Cambridge, 2010).
[6] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2010).
[7] L. Hébert-Dufresne, O. Patterson-Lomba, G. M. Goerg, and B. M. Althouse, Phys. Rev. Lett. **110**, 108103 (2013).
[8] L. Hébert-Dufresne, A. Allard, J.-G. Young, and L. J. Dubé, Sci. Rep. **3**, 2171 (2013).
[9] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[10] S. Melnik, A. Hackett, M. A. Porter, P. J. Mucha, and J. P. Gleeson, Phys. Rev. E **83**, 036112 (2011).
[11] K. Christensen and N. R. Moloney, in *Complexity and Criticality* (Imperial College Press, London, 2005), p. 392.
[12] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).
[13] M. E. J. Newman, Phys. Rev. E **66**, 016128 (2002).
[14] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[15] A. Vázquez and Y. Moreno, Phys. Rev. E **67**, 015101(R) (2003).
[16] M. E. J. Newman, Phys. Rev. E **67**, 026126 (2003).
[17] A. Vazquez, Phys. Rev. E **74**, 066114 (2006).
[18] A. Allard, P.-A. Noël, L. J. Dubé, and B. Pourbohloul, Phys. Rev. E **79**, 036113 (2009).
[19] A. Allard, L. Hébert-Dufresne, P.-A. Noël, V. Marceau, and L. J. Dubé, J. Phys. A **45**, 405005 (2012).
[20] M. E. J. Newman, Phys. Rev. E **68**, 026121 (2003).
[21] M. A. Serrano and M. Boguñá, Phys. Rev. Lett. **97**, 088701 (2006).
[22] M. A. Serrano and M. Boguñá, Phys. Rev. E **74**, 056115 (2006).
[23] X. Shi, L. A. Adamic, and M. J. Strauss, Physica A **378**, 33 (2007).
[24] Y. Berchenko, Y. Artzy-Randrup, M. Teicher, and L. Stone, Phys. Rev. Lett. **102**, 138701 (2009).
[25] J. C. Miller, Phys. Rev. E **80**, 020901(R) (2009).
[26] M. E. J. Newman, Phys. Rev. Lett. **103**, 058701 (2009).
[27] J. P. Gleeson, Phys. Rev. E **80**, 036107 (2009).
[28] B. Karrer and M. E. J. Newman, Phys. Rev. E **82**, 066118 (2010).
[29] V. Zlatić, D. Garlaschelli, and G. Caldarelli, Europhys. Lett. **97**, 28005 (2012).
[30] Recent advances in understanding the global organization of clustering in real networks [54] offers further ideas to incorporate clustering in our model and will be the subject of a subsequent study.
[31] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. Eugene Stanley, and H. A. Makse, Nat. Phys. **6**, 888 (2010).
[32] V. Batagelj and M. Zaveršnik, arXiv:cs/0310049v1[6S-DS].
[33] S. B. Seidman, Social Networks **5**, 269 (1983).
[34] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Phys. Rev. Lett. **96**, 040601 (2006).
[35] The present approach is but a special case of a complete and more general theoretical framework being prepared for publication.
[36] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, Phys. Rev. E **70**, 056122 (2004).
[37] M. Ripeanu and I. Foster, in *Peer-to-Peer Systems*, edited by P. Druschel, F. Kaashoek, and A. Rowstron (Springer, Berlin, 2002), pp. 85–93.
[38] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, in *Proceedings of the 2nd ACM Workshop on Online Social Networks—WOSN '09* (ACM, New York, NY, 2009), pp. 37–42.
[39] E. Cho, S. A. Myers, and J. Leskovec, in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, 2011), pp. 1082–1090.
[40] J. Leskovec, J. Kleinberg, and C. Faloutsos, ACM Trans. Knowl. Discov. Data **1**, 2 (2007).
[41] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[42] G. Palla, I. J. Farkas, P. Pollner, I. Derényi, and T. Vicsek, New J. Phys. **10**, 123026 (2008).

[43] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, IEEE Trans. Power Syst. **26**, 12 (2011).
[44] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[45] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2005).
[46] I. J. Farkas, D. Ábel, G. Palla, and T. Vicsek, New J. Phys. **9**, 186 (2007).
[47] L. Hébert-Dufresne, A. Allard, V. Marceau, P.-A. Noël, and L. J. Dubé, Phys. Rev. Lett. **107**, 158702 (2011).
[48] M. D. Choudhury, H. Sundaram, A. John, and D. D. Seligmann, in *Proceedings of the International Conference on Computational Science and Engineering* (IEEE Computer Society, Washington, DC, 2009), pp. 151–158.
[49] B. Klimt and Y. Yang, in *First Conference on Email and Anti-Spam (CEAS)* (IEEE Computer Society, Washington, DC, 2004).
[50] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, Internet Math. **6**, 29 (2009).
[51] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, in *Proceedings of the 16th International Conference on World Wide Web* (ACM, New York, NY, 2007), pp. 835–844.
[52] Codes solving the theoretical model and generating the networks are available at http://dynamica.phy.ulaval.ca
[53] S. Maniu, T. Abdessalem, and B. Cautis, in *Proceedings of the International Conference on World Wide Web Posters* (ACM, New York, NY, 2011), pp. 87–88.
[54] P. Colomer-de-Simón, M. A. Serrano, M. G. Beiró, J. I. Alvarez-Hamelin, and M. Boguñá, Sci. Rep. **3**, 2517 (2013).