

Sampling Motif-Constrained Ensembles of Networks

Rico Fischer,¹ Jorge C. Leitão,¹ Tiago P. Peixoto,² and Eduardo G. Altmann¹

¹Max Planck Institute for the Physics of Complex Systems, 01187 Dresden, Germany

²Institut für Theoretische Physik, Universität Bremen, Hochschulring 18, 28359 Bremen, Germany

(Received 30 July 2015; published 29 October 2015)

The statistical significance of network properties is conditioned on null models which satisfy specified properties but that are otherwise random. Exponential random graph models are a principled theoretical framework to generate such constrained ensembles, but which often fail in practice, either due to model inconsistency or due to the impossibility to sample networks from them. These problems affect the important case of networks with prescribed clustering coefficient or number of small connected subgraphs (motifs). In this Letter we use the Wang-Landau method to obtain a multicanonical sampling that overcomes both these problems. We sample, in polynomial time, networks with arbitrary degree sequences from ensembles with imposed motifs counts. Applying this method to social networks, we investigate the relation between transitivity and homophily, and we quantify the correlation between different types of motifs, finding that single motifs can explain up to 60% of the variation of motif profiles.

DOI: 10.1103/PhysRevLett.115.188701

PACS numbers: 89.75.Hc, 05.10.Ln

Networks form the basis of an ample class of complex systems. The observed topological patterns of such systems often yield the only available evidence for the underlying principles behind their formation. However, the significance of any observed property can only be assessed in comparison to a properly defined network ensemble that acts as a “null” model [1–3]. For instance, clustering (i.e., high density of triangles), skewed degree distributions, and community structure are considered significant in real networks because they are absent in Erdős-Renyi networks. To perform such comparisons, it is essential not only to properly define such null models, but also to correctly sample network realizations from them. This is relatively straightforward when the ensemble generates networks where the edges are sampled independently (e.g., Erdős-Renyi and configuration models [4,5], the stochastic block model [6,7]) and it remains feasible when strict edge independence is violated due to hard constraints [8–10]. However, for ensembles with more generic constraints the sampling is significantly more challenging. A particularly important example is ensembles with a prescribed density of connected subgraphs (“motifs”) [11–13]. For this class of models, one often finds abrupt phase transitions, where sampled networks possess either very high or very low motif density [12,13], excluding intermediary values often encountered in real systems. Furthermore, they often show strong nonergodic behavior, with very slow relaxation that forbids unbiased sampling in practical computational time [13]. Since the edge placement is not independent, the densities of different motifs are correlated with each other and also with large-scale network structures [14,15]. Without addressing the issue of correct sampling, these correlations cannot be properly identified, which makes the occurrence of these patterns in real systems difficult to interpret. In particular, it is not possible to

conclude whether a particular motif density profile indicates a topology optimized towards robustness [16,17] or whether it is merely a by-product of a specific large-scale structure [14,18], of combinatorial constraints [19], or of correlations between motifs.

In this Letter we show how to sample from ensembles with prescribed motif densities in polynomial time. We employ a multicanonical Monte Carlo method [20] that allows the entire range of the order parameter to be explored. In this manner, not only is the nonergodicity problem explicitly avoided, but it also becomes possible to sample networks with arbitrary motif densities, even those at intermediate values that are unattainable via traditional importance sampling. This allows us to quantitatively investigate two fundamental problems in social networks: the homophily-transitivity relationship and the interdependence of different motif types.

We are interested in network ensembles that possess one particular observable s of interest, but that are otherwise maximally random. The last requirement is essential to ensure that the ensemble is representative of the networks with a given s and is not subject to additional (hidden) constraints. Both features are achieved by sampling the network from an exponential random graph model (ERGM) [3,10,21–24] \mathcal{G} , where each graph $g \in \mathcal{G}$ occurs with probability

$$\Pi_{\beta}(g) = \frac{e^{\beta s(g)}}{Z_{\beta}}, \quad \text{where } Z_{\beta} = \sum_{g \in \mathcal{G}} e^{\beta s(g)}, \quad (1)$$

where $s(g)$ is the observable associated with network g , and β is an inverse-temperature parameter, in analogy to the canonical ensemble in statistical physics. The distribution of s is $\rho_{\beta}(s) = \sum_{g \in \mathcal{G}} \delta[s(g) - s] \Pi_{\beta}(g) = \rho_0(s) e^{\beta s} / Z_{\beta}$,

where $\rho_0(s) \equiv \rho_{\beta=0}(s)$ is called the state density (the fraction of networks g in the ensemble that have observable equal to s). The ensemble that acts as a null model for an empirical network with $s = s^*$ is usually constructed fixing β in such a way that $\langle s \rangle_\beta \equiv \sum_s s \rho_\beta(s)$ equals s^* . The number of networks in this ensemble typically grows exponentially with the number of nodes, and, thus, besides a small set of observables s that can be treated analytically, investigation of ERGMs requires sampling networks g from \mathcal{G} using Monte Carlo methods [23].

The usual Markov chain Monte Carlo (MCMC) method works as follows: starting from one network $g \in \mathcal{G}$, a new network $g' \in \mathcal{G}_n$ is proposed by choosing two links at random and exchanging one of the nodes of each link, thus preserving the degree sequence of the network [25]. The proposed network is accepted with the Metropolis-Hastings probability $A(g \rightarrow g') = \min\{1, e^{\beta[s(g') - s(g)]}\}$ and the process is repeated from g' (g) if the proposal is accepted (rejected) [26]. Since the moves fulfill ergodicity and detailed balance, for sufficiently long times the values of s in the sampled networks g are distributed as $\rho_\beta(s)$. However, despite this asymptotic guarantee, in practice this method often fails because the time to approximate $\rho_\beta(s)$ grows exponentially with the number of nodes N . This happens whenever ρ_β possesses more than one local maximum (minimum of the free energy) and the barriers between them grow with N . As we show below, this generically happens when the observables s are related to motifs.

As an alternative to the *canonical* (simple Metropolis) sampling method described above, we propose a *multicanonical* sampling to overcome the aforementioned problem. This method aims to sample networks uniformly on a predefined observable range $[s_{\min}, s_{\max}]$, thus overcoming the minima of $\rho_\beta(s)$ that are responsible for the weak performance of the canonical method. This is done by sampling the states according to auxiliary ensemble with probabilities $\Pi(g) \propto 1/\rho_0(s(g))$, achieved by simply changing the acceptance to $A(g \rightarrow g') = \min\{1, \rho_0[s(g')]/\rho_0[s(g)]\}$ [20]. However, in order to perform this sampling we need to know the state density $\rho_0(s)$. In order to estimate it, we use the Wang-Landau algorithm [20,27], which, in short, constructs an adaptive histogram to approximate $\rho_0[s(g)]$ [28]. After convergence, $\rho_\beta(s)$ is estimated for all β 's reweighting $\rho_0(s)$ through $\rho_\beta(s) = \rho_0(s) \exp(-\beta s)/Z_\beta$ [20]. Hence, the auxiliary ensemble allows us to explore the original canonical ensembles without being restricted to the most probable regions. More importantly, we can impose the desired value of the observable as a hard constraint *a posteriori*, i.e., only sample networks with $s(g) = s^*$. The multicanonical approach has recently been applied to investigate the spectral gap of networks [29], and related approaches have been used to investigate percolation [30] and resilience properties of networks [31].

In Fig. 1 we show how the application of multicanonical sampling solves the limitations of canonical sampling in the classical problem of introducing clustering in a k -regular network [11,13]. Here, nodes are forced to have the same degree k and the observable of interest is the number of triangles, $s(g) = n_\Delta$. Fixing n_Δ is the same as fixing the clustering coefficient $c = 3n_\Delta/n_\wedge$, where n_\wedge is the number of connected triples (a constant for all networks with the same degree sequence) [3]. This model exhibits a transition at a specific value of $\beta = \beta_{PT} (\approx 3.54$ for $k = 4$), separating low and high-clustering phases [13]. The canonical sampling is unable to compute $\langle c \rangle$ close to the phase transition because it yields different estimations of $\langle c \rangle$, depending whether β is slowly increased ($\beta \uparrow$, lower branch) or decreased ($\beta \downarrow$, upper branch). This hysteresis is typical around first-order phase transitions (coexisting phases) and indicates that the canonical sampling is in a metastable state. Indeed, $\rho_{\beta_{PT}}(c)$ has two local maxima in which the canonical sampling becomes trapped (inset in Fig. 1). On the other hand, the multicanonical sampling is immune to these problems: it correctly characterizes $\langle c \rangle$ at $\beta = \beta_{PT}$ and reveals the full distribution $\rho_{\beta=\beta_{PT}}$. Hence, the method is not only capable of computing the correct ensemble

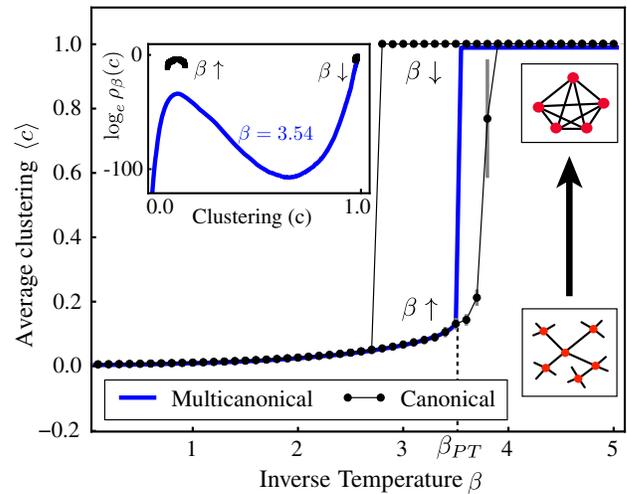


FIG. 1 (color online). Multicanonical sampling of exponential random graphs with imposed clustering avoids the limitations of canonical sampling. The ensemble \mathcal{G} corresponds to k -regular undirected networks with $N = 640$ and degree $k = 4$. The observable is the clustering coefficient $s = c$ (proportional to the number of triangles, n_Δ). The main plot shows $\langle c \rangle$ (and standard deviation) as a function of the inverse temperature β obtained by canonical (symbols) and multicanonical (continuous thick line) sampling. Inset: the distribution $\rho_\beta(c)$ for $\beta = 3.54 \approx \beta_{PT}$ obtained by the two methods. Canonical samplings used 5×10^5 MCMC steps for equilibration, before another 5×10^5 steps were used for estimation. After these steps, the value of β was slowly increased ($\beta \uparrow$) or decreased ($\beta \downarrow$) and the process repeated. The multicanonical sampling used 20 Wang-Landau steps to estimate $\rho_0(c)$ (each step used 5 tunneling steps) [20,32].

average for any β , it yields typical networks with any value of c , including the significant gap $c \in [0.2, 1]$ which is unattainable with the canonical sampling. In Fig. 2 we confirm that the computational cost of the multicanonical method scales polynomially with system size, a dramatic improvement over the exponential scaling of the canonical method.

Next we use the multicanonical method to investigate two important problems of social networks. The first problem we consider is to distinguish between homophily (the tendency of “similar” nodes to connect to each other) and transitivity (the tendency of nodes that already share a common neighbor to connect to each other) in social networks [2, 14, 34–37]. We use the (undirected) network of Email exchange within a university [38]. It consists of $N = 1133$ users, and $M = 5451$ Email exchanges, and a roughly exponential degree distribution. As observables we consider the clustering coefficient c and the degree assortativity r [39], for which we obtain $c^* = 0.166(12)$ and $r^* = 0.08(3)$ (uncertainties in the last digit estimated using the order-10 Jackknife method). We assess the significance of these values by comparing them to those obtained in the following three network ensembles with the same degree sequence as in the original network: (i) Same weight to all networks g (i.e., the configuration model). Canonical

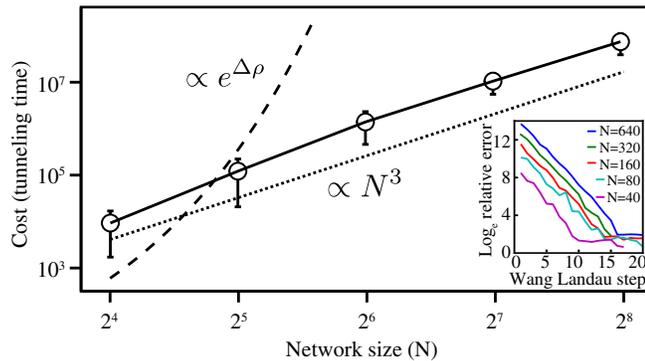


FIG. 2 (color online). Efficiency of the multicanonical method to sample networks with constraints. The computational cost (in number of MCMC steps) to generate an independent realization of a network in the k -regular ensemble with $k = 3$ is plotted as a function of N . In the canonical method close to the critical β , this requires passing the minimum of $\rho_\beta(c)$ (inset of Fig. 1). We measured that the height of this barrier increases as $\Delta\rho \approx 0.4N$, which leads to an exponential increase in the cost (dashed line). Sampling independent realizations in the multicanonical method requires, at most, a tunneling (the number of MCMC steps to do $c = 0 \rightarrow c = 1 \rightarrow c = 0$) [32]. The measured tunneling time (circles and full line) scales polynomially. Inset: convergence of the relative error in the logarithm of the density of states (entropy) during convergence of the Wang-Landau algorithm, estimated comparing the measured value with the exact value on $c = 1$. The saturation of the error observed for a large number of steps does not hinder the sampling of any c (see Ref. [33] for methods to overcome the saturation).

sampling with $\beta = 0$ yields $\langle c \rangle_{\beta=0} = 0.028(1)$ and $\langle r_{\beta=0} \rangle = -0.017(13)$, much smaller than c^* and r^* as typically found in social networks. (ii) ERGMs with $\langle c \rangle = c^*$. In order to determine whether the assortativity is a consequence of high clustering [14] we would like to measure $\langle r \rangle$ from the null model with $\langle c \rangle = c^*$. This canonical sampling fails because $\langle c \rangle_\beta$ vs β shows an hysteresis around $s = c^*$ (inset of Fig. 2, in agreement with our previous discussion). (iii) Hard constraints with $c(g) = c^*$, obtained using multicanonical sampling. As mentioned before, this type of hard constraint is unfeasible with canonical sampling, even if the desired observable value is realizable. With the multicanonical method we sample points after a number of Monte Carlo steps proportional to the tunneling time, which guarantees that the sampled points are independent and unbiased [32]. We performed multicanonical sampling for a desired c and measured the assortativity r . The results are shown in Fig. 3 and reveal that random networks with the same clustering of the Email network $c = c^*$ typically show a much larger assortativity $\langle r \rangle > r^*$. Therefore, although both c^* and r^* are larger than one would expect for a fully random network, the actual value of r^* is significantly less than one would expect by knowing only c^* . From this we conclude that the degree homophily is not explained alone by transitivity.

The second problem we address is the extent to which the occurrence of different motifs (connected subgraphs) are related to each other and the impact of such correlations on the so-called motif profiles [17]. Here we focus on directed networks, and the observable of interest is the

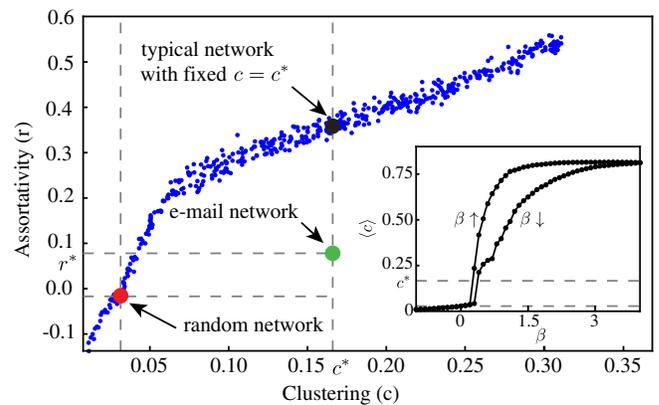


FIG. 3 (color online). Relationship between clustering c and assortativity r in the Email network of Ref. [38]. The assortativity r^* is higher than in a random network with the same degree sequence, but lower than in a typical network with a fixed clustering c^* . The plot shows r and c of different networks: the Email network (green), a typical fully random network (red), a typical random network with $c = c^*$ (black), and networks sampled using the multicanonical method from an ensemble with equal probability for networks with the same c (blue dots). Inset: $\langle c \rangle$ obtained using the canonical method.

number n_i of occurrences of a specific motif i . Again, traditional sampling methods are not suited to address this problem because of the existence of (potentially multiple [13]) discontinuous phase transitions. Instead, using the multicanonical method, we reliably sample networks with a prescribed count of one particular motif. By measuring the counts of all other motifs, we obtain the correlations between them and the constrained motif. In this manner, we obtain [40] the interdependence between all 13 different three-node motifs in a directed acquaintance network between physicians [41] (with $N = 241$ nodes and $M = 1098$ edges). The results in Fig. 4 reveal strong positive and negative correlations between pairs of motifs. Two blocks of motifs can be identified [1,2,3,7,8 and 4,5,6, Fig. 4(a)]. Motifs show positive correlations within their blocks and are anticorrelated with motifs in the other blocks (the motifs 9 to 13 show a mixed behavior). Given that one motif is over (under) represented, one should expect also an over representation in motifs positively (negatively) correlated with it. As a consequence of this correlation, we find that single motifs explain up to 60% of the variance of the motif profile across the other 12 motifs [Fig. 4(b), upper and middle panel]. Furthermore, if the constrained ensembles are used to compute alternative z scores, we find that the resulting motif profiles vary dramatically depending on the constraint, with some motifs j showing variations from $z_j \gg 0$ to $z_j \ll 0$ [Fig. 4(b), lower panel. This sensitivity of the motif profile z_j shows that such profiles bring limited

insights on the over or under representation of individual motifs in a network. In particular, since such nontrivial profiles as those seen in Fig. 4(b) can be obtained by imposing the occurrence of a single motif, it is questionable whether conclusions regarding the underlying formation mechanisms can be reliably reached from them [17,18]. Nevertheless, the null models considered here represent a principled approach of assessing the relative significance of motif occurrences that is more meaningful than the usual comparison to fully random networks.

In summary, we have shown that multicanonical sampling allows for an improved network generation and for the investigation of problems which were otherwise intractable. In particular, we characterize ERGMs in cases where the usual canonical sampling fails and we sample networks imposing hard constraints, an alternative to a direct sampling of ERGMs even when the usual algorithms are feasible. Our analysis of empirical networks demonstrates that using the multicanonical sampling allows the investigation of the interdependence between network properties. In particular, we quantified the correlation between clustering and assortativity, and between different motifs, as well as the extent to which their significance profiles can be explained by single motifs. This opens the possibility of investigating the correlation between motifs as well as other local-scale properties and the large-scale structure of networks [14], such as communities, core peripheries, and many others. The systematic disentangling of these diverse

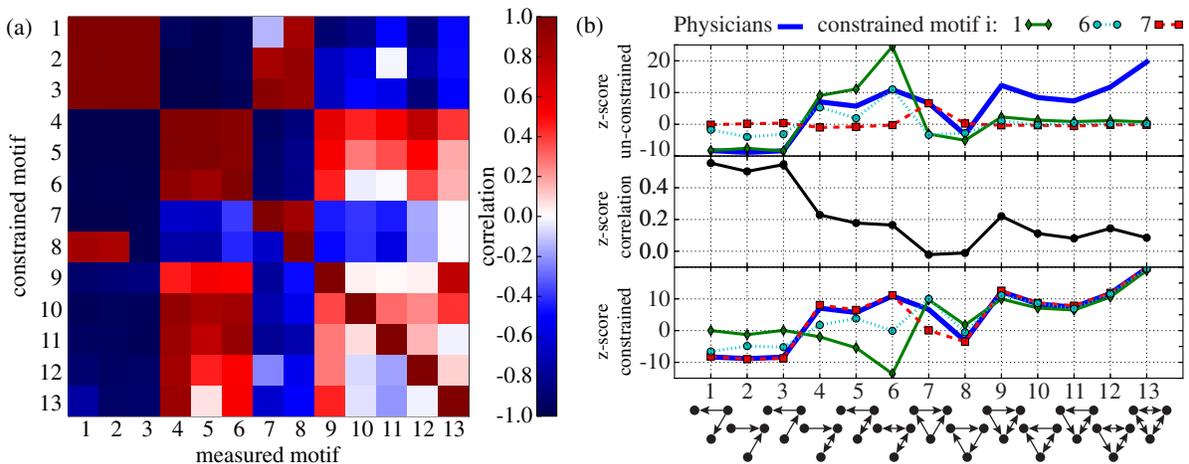


FIG. 4 (color online). Motifs are correlated to each other in blocks. (a) The Pearson correlation coefficient $[R_{ij} = (\langle n_i n_j \rangle - \langle n_i \rangle \langle n_j \rangle) / \sigma_{n_i} \sigma_{n_j}]$ between motifs i and j , computed by varying the constrained motif (in a range which includes the values of the real and random networks [40]). (b) Upper panel: Motif profile [17] built from the z score z_j vs j [$z_j = (n_j - \langle n_j \rangle) / \sigma_{n_j}$, where n_j is the number of motifs j and $\langle \dots \rangle$ and σ_j are the average and standard deviation in the $\beta = 0$ ensemble]. Different lines correspond to the z_j of the real network (z_j^* , blue line) and the expected z_j s in the constrained ensemble in which n_i is equal to the n_i^* of the real network, where i is the constrained motif shown in the legend ($z_j = z_j^*$ for $j = i$). Middle panel: the correlation between the profiles shown in the upper panel, i.e., between the profile z of the real network and the profile z' of the motif- i constrained network (as a function of i), computed as $R_{zz'} = (\langle zz' \rangle - \langle z \rangle \langle z' \rangle) / \sigma_z^2$, where $\langle \dots \rangle$ and σ_z were computed over $j \neq i$. Lower panel: comparison of the z score shown in the upper panel (blue line) and the alternative z score obtained computing $\langle \dots \rangle$ and σ_j in the ensemble constrained by $n_i = n_i^*$, where i is indicated in the legend ($z_j \equiv 0$ for $j = i$).

features is a crucial and open problem in the identification of fundamental models of network formation.

We thank J. M. V. P. Lopes for insightful discussions. This work was partially funded by the University of Bremen, under the program ZF04, and FCT (Portugal), Grant No. SFRH/BD/90050/2012.

-
- [1] L. A. Nunes Amaral and R. Guimera, *Nat. Phys.* **2**, 75 (2006).
- [2] P. Holme and J. Zhao, *Phys. Rev. E* **75**, 046111 (2007).
- [3] M. Newman, *Networks: An Introduction* (Oxford University Press, New York, 2010).
- [4] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, *Phys. Rev. E* **64**, 026118 (2001).
- [5] F. Chung and L. Lu, *Ann. Combinator.* **6**, 125 (2002).
- [6] P. W. Holland, K. B. Laskey, and S. Leinhardt, *Soc. Networks* **5**, 109 (1983).
- [7] B. Karrer and M. E. J. Newman, *Phys. Rev. E* **83**, 016107 (2011).
- [8] J. Blitzstein and P. Diaconis, *Internet Math.* **6**, 489 (2011).
- [9] H. Kim, C. I. D. Genio, K. E. Bassler, and Z. Toroczkai, *New J. Phys.* **14**, 023012 (2012).
- [10] C. Orsini, M. M. Dankulov, A. Jamakovic, P. Mahadevan, P. Colomer-de Simón, A. Vahdat, K. E. Bassler, Z. Toroczkai, M. Boguñá, G. Caldarelli, S. Fortunato, and D. Krioukov, *arXiv:1505.07503* [Nat. Commun. (to be published)].
- [11] D. J. Strauss, *Biometrika* **62**, 467 (1975).
- [12] J. Park and M. E. J. Newman, *Phys. Rev. E* **70**, 066146 (2004).
- [13] D. Foster, J. Foster, M. Paczuski, and P. Grassberger, *Phys. Rev. E* **81**, 046115 (2010).
- [14] D. V. Foster, J. G. Foster, P. Grassberger, and M. Paczuski, *Phys. Rev. E* **84**, 066117 (2011).
- [15] M. E. Beber, C. Fretter, S. Jain, N. Sonnenschein, M. Müller-Hannemann, and M.-T. Hütt, *J. R. Soc. Interface* **9**, 3426 (2012).
- [16] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, *Science* **298**, 824 (2002).
- [17] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, *Science* **303**, 1538 (2004).
- [18] Y. Artzy-Randrup, S. J. Fleishman, N. Ben-Tal, and L. Stone, *Science* **305**, 1107c (2004).
- [19] J. Ugander, L. Backstrom, and J. Kleinberg, in *Proceedings of the 22nd International Conference on World Wide Web, WWW'13, 2013* (ACM Digital Library, Rio de Janeiro, Brazil, 2013).
- [20] D. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, Cambridge, England, 2013).
- [21] D. Strauss, *SIAM Rev.* **28**, 513 (1986).
- [22] T. A. B. Snijders, *J. Soc. Struct.* **3**, 40 (2002).
- [23] J. Park and M. E. J. Newman, *Phys. Rev. E* **70**, 066117 (2004).
- [24] S. Horvát, E. Czabarka, and Z. Toroczkai, *Phys. Rev. Lett.* **114**, 158701 (2015).
- [25] S. Maslov and K. Sneppen, *Science* **296**, 910 (2002).
- [26] M. E. J. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics* (Oxford Univ. Press, New York, 2002).
- [27] F. Wang and D. P. Landau, *Phys. Rev. Lett.* **86**, 2050 (2001).
- [28] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevLett.115.188701> for a description of the multicanonical sampling method proposed in this Letter; we provide an implementation at <https://dx.doi.org/10.5281/zenodo.30626>.
- [29] N. Saito and Y. Iba, *Comput. Phys. Commun.* **182**, 223 (2011).
- [30] A. K. Hartmann, *Eur. Phys. J. B* **84**, 627 (2011).
- [31] T. Dewenter and A. K. Hartmann, *New J. Phys.* **17**, 015005 (2015).
- [32] P. Dayal, S. Trebst, S. Wessel, D. Wurtz, M. Troyer, S. Sabhapandit, and S. N. Coppersmith, *Phys. Rev. Lett.* **92**, 097201 (2004).
- [33] R. E. Belardinelli, S. Manzi, and V. D. Pereyra, *Phys. Rev. E* **78**, 067701 (2008).
- [34] A. Rapoport, *Bull. Math. Biophys.* **15**, 523 (1953).
- [35] M. S. Granovetter, *Am. J. Sociology* **78**, 1360 (1973).
- [36] G. Kossinets and D. J. Watts, *Am. J. Sociology* **115**, 405 (2009).
- [37] G. Bianconi, R. K. Darst, J. Iacovacci, and S. Fortunato, *Phys. Rev. E* **90**, 042806 (2014).
- [38] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, *Phys. Rev. E* **68**, 065103 (2003).
- [39] M. E. J. Newman, *Phys. Rev. Lett.* **89**, 208701 (2002).
- [40] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevLett.115.188701>, Fig. 1.
- [41] J. Coleman, E. Katz, and H. Menzel, *Sociometry* **20**, 253 (1957).