The minimum dominating set (MDS) problem [1] has fundamental importance in network science. For example, to ensure the proper functioning of a complex networked system such as a nation-wide power grid, it is often necessary to monitor the system’s microscopic dynamics in real-time by placing sensors on the nodes. A sensor may have the capability of observing the instantaneous states of the residing node and all its adjacent nodes in the network [2], so they may not need to occupy all the nodes. We then have the MDS problem: How to place sensors on as few nodes as possible to minimize costs but still ensure that each node is either occupied or adjacent to at least one occupied node? A more stringent constraint, which is adopted in lattice glass models [3], is to require an empty node to be surrounded by at least \( l_i \) occupied nodes, with \( l_i \) being node-dependent. The MDS problem corresponds to the case of \( l_i = 1 \), while the other limiting case of \( l_i = d_i \) is just the vertex cover (or independent set) problem [4, 5], where \( d_i \) is the degree (i.e., number of adjacent nodes) of node \( i \).

The MDS problem has many wide practical applications, such as monitoring large-scale power grids and other transportation systems [2], controlling the spreading of infectious diseases and other network dynamical processes [6–9], efficient routing in wireless networks [10], and network public goods games [11]. However, exactly solving it is extremely difficult in general, since it is a nondeterministic polynomial-complete (NP-complete) optimization problem [1]. Even the task of approximately solving the MDS problem is very hard. For a general network of \( N \) nodes, so far the best polynomial algorithms can only guarantee to get dominating sets with sizes not exceeding \( N \) times of the minimum size [12, 13]. Many local-search algorithms have been proposed to solve the MDS problem heuristically (see review [1] and [2, 6, 9, 13]), but theoretical results on the MDS sizes of random network ensembles are very sparse.

In this work we bring several new theoretical contributions. We show that a generalized leaf-removal (GLR) process may cause a core percolation transition, and we propose a quantitative theory to describe this percolation. If the network contains no core, GLR reaches an exact MDS; if an extensive core exists, we combine GLR with a local greedy process to get an upper bound to the MDS size. We then introduce a spin glass model and estimate the MDS size by a replica-symmetric (RS) mean field theory, and implement a message-passing algorithm to get near-optimal dominating sets for single network instances. Our algorithms also perform well on a set of real-world large network instances.

Core percolation. Consider a simple network \( W \) of \( N \) nodes and \( M \) links, each link connecting between two different nodes. Each node \( i \in \{1, 2, \ldots, N\} \) is either empty (indicated by the occupation state \( c_i = 0 \) or occupied by sensors \( c_i = 1 \)). A node \( i \) is regarded as observed if it is occupied or it is empty but adjacent to one or more occupied nodes. We need to occupy a set \( \Gamma \) of nodes to make all the \( N \) nodes be observed, and the objective is to make the dominating set \( \Gamma \) as small as possible. Here we extend the leaf-removal idea of [15] and consider a generalized leaf-removal process. This dynamics is based on the following considerations: first, as pointed out in [15], if node \( i \) is an unobserved leaf node (which has only a single neighbor, say \( j \)), then occupying \( j \) but leaving \( i \) empty must be an optimal strategy; second, we notice that if \( i \) is empty but observed and at most one of its adjacent nodes is unobserved, then it must be an optimal strategy not to occupy \( i \).

Starting from the original network \( W \) with all the links present and all the nodes unobserved, GLR at each evolution step \( t = 0, 1, 2, \ldots \) performs the following operations sequentially: (1) Occupy all the isolated nodes; then as long as there is an unobserved leaf node in the network, occupy its unique adjacent node. (2) Simplify the network: first delete all the links between two observed nodes; then as long as there is an observed node with only a single link, delete this node and the attached link; and then delete all the isolated observed nodes. (3) If there is at least one isolated or leaf unobserved node in the simplified network, increase the evolution step \( t \) by one and repeat operations (1) and (2).

If all the nodes of the original network \( W \) become observed after GLR, it is easy to prove that the set \( \Gamma \) of occupied nodes must be a MDS [16]. On the other hand,
If the final simplified network is non-empty, then there must be some nodes that are still unobserved. The subnetwork induced by these unobserved nodes is referred to as the core of the original network $W$. This core is connected only to observed empty nodes but not to occupied nodes. We denote by $n_{\text{core}}$ the fraction of nodes in this core and by $w$ the fraction of occupied nodes. The cross symbols of Fig. 1 show the GLR simulation results on single Erdős-Rényi (ER) random networks [17] with $N = 10^6$ nodes and $M = (c/2)N$ links. When the mean node degree $c < 2.41$ there is no core ($n_{\text{core}} = 0$), and GLR reaches a MDS for the whole network. The core emerges at $c \approx 2.41$ and its relative size $n_{\text{core}}$ then increases with $c$ continuously from zero, suggesting a continuous core percolation. For $c > 2.41$, GLR constructs a MDS only for part of the network and it leaves an extensive core of $n_{\text{core}}N$ nodes unobserved. Qualitatively the same core percolation phenomenon occurs on more heterogeneous (scale-free) [18] random networks. Notice the core percolation transition resulting from the GLR optimization process is qualitatively different from the simpler observability transition discussed in [2], which considers the appearance of a giant connected component of observed nodes.

We now develop a percolation theory to describe the GLR dynamics on random networks characterized by a degree distribution $P(d)$, which gives the fraction of nodes with degree $d \geq 0$ [17]. This theory has four sets of parameters $\alpha_t$, $\beta_t$, $\gamma_t$, and $\eta_t$ ($t = 0, 1, \ldots$). Consider a randomly chosen link $(i,j)$ between nodes $i$ and $j$. When there is no degree correlation in the network, the probability $Q(d)$ of node $j$ to have $d$ attached links is $Q(d) = dP(d)/c$, with $c = \sum_{d \geq 0} dP(d)$ being the mean degree. Let us suppose node $i$ is externally fixed to $c_i = 0$. Conditional on this, then $\alpha_t$ is defined as the probability of the adjacent node $j$ becoming an unobserved leaf node at the start of the $t$-th GLR step, $\beta_t$ is the probability of $j$ being newly occupied at this $t$-th step, $\gamma_t$ is that of $j$ being empty but observed at the end of the $t$-th step, and $\eta_t$ is that of $j$ being an empty and observed leaf node at the end of the $t$-th step.

The probability of a randomly chosen node $k$ to be in the final unobserved core is simply $n_{\text{core}}$. Since a random network is locally tree-like, we can assume independence among the adjacent nodes of $k$ when it is not yet observed. Following the derivation of [19], we obtain that

$$n_{\text{core}} = \sum_{d \geq 2} \frac{d-2}{s=0} P(d) \sum_{s=0}^{d-2} C^s_{d-2} \left( \sum_{l \geq 0} \eta_l \right)^s \left( 1 - \sum_{l \geq 0} (\alpha_l + \beta_l + \eta_l) \right)^{d-s},$$

where $C^s_{d-2} \equiv d!/[s!(d-s)!]$ is the binomial coefficient. If node $k$ does not belong to the core, we denote by $I_t$ the probability of it transiting to be occupied at the $t$-th GLR step. Then the fraction of occupied nodes during the whole GLR process is computed by

$$w = \sum_{t \geq 0} I_t. \quad (2)$$

After some combinatorial considerations similar to those of [19], we obtain the following iterative equations for $\alpha_t$, $\beta_t$, $\gamma_t$, and $\eta_t$:

$$\sum_{t=0}^{t} \alpha_t = Q(1) + \sum_{d \geq 2} Q(d) \left( \sum_{l=0}^{t-1} \eta_l \right)^{d-1} , \quad (3)$$

$$\sum_{t=0}^{t} \beta_t = 1 - Q(1) - \sum_{d \geq 2} Q(d) \left( 1 - \sum_{l=0}^{t} \alpha_l \right)^{d-1} , \quad (4)$$

$$\gamma_t = \sum_{d \geq 2} Q(d) \left[ \left( 1 - \sum_{l=0}^{t} \alpha_l \right)^{d-1} - \left( 1 - \sum_{l=0}^{t} (\alpha_l + \beta_l) \right)^{d-1} \right] , \quad (5)$$

$$\eta_t = \sum_{d \geq 2} Q(d) \left[ \left( \sum_{l=0}^{t} \beta_l + \gamma_l \right)^{d-1} - (\gamma_l)^{d-1} \right] . \quad (6)$$

The expression for the occupation fraction $I_t$ can also be derived, but the resulting iterative equation is more complicated:

$$\sum_{t=0}^{t} I_t = P(0) + P(1) \frac{Q_0}{2} + \sum_{d \geq 2} P(d) \left\{ \left( \sum_{l=0}^{t-1} \eta_l \right)^{d-1} - d \sum_{l=1}^{t-1} \eta_l \left( \sum_{m=0}^{l-1} \eta_m \right)^{d-1} + 1 - \left( 1 - \sum_{l=0}^{t} \alpha_l \right)^{d-1} \right\} - d \sum_{l=1}^{t} \alpha_l \left\{ \left( \sum_{m=0}^{l-1} \beta_m + \gamma_{l-1} \right)^{d-1} - \left( \gamma_{l-1} \right)^{d-1} \right\} + \frac{1}{2} \left( \sum_{m=0}^{l-1} \eta_m \right)^{d-1} + \frac{1}{2} \left( \sum_{m=0}^{l-2} \eta_m \right)^{d-1} \right\} . \quad (7)$$

When $N \to \infty$ the degree distribution of ER networks is $P(d) = c^d e^{-c}/d!$. For this network ensemble, the predicted results of $n_{\text{core}}$ and $w$ by our percolation theory are in perfect agreement with simulation results (Fig. 1). Especially the theory predicts a continuous core percolation transition at $c = 2.4102$. Before this transition, the occupation fraction $w$ obtained by Eq. 6 is equal to the ensemble-averaged MDS size (relative to $N$), but it is only a lower bound to this size when $n_{\text{core}} > 0$. This percolation theory works also excellently for scale-free [18] random networks.

If GLR leaves a core behind, we can combine it with a local impact-based greedy process [11] to construct a sub-optimal dominating set. The impact of an unoccupied node $i$ is defined as the number of nodes that will be observed by occupying $i$. Our hybrid local algorithm...
works by repeating the following operation: if there is an unobserved leaf node, perform GLR; otherwise occupy a randomly chosen node of maximal impact [16]. The performance of this hybrid algorithm is demonstrated in Fig. 2 for single ER networks and regular random (RR) networks. (All the nodes of a RR network have the same integer degree c.) This hybrid algorithm outperforms the pure greedy algorithm considerably for c ≤ 10.

Minimum occupation fraction. We now discuss the issue of estimating the MDS size. For a given network instance W, we introduce a partition function Z as

\[ Z = \sum_{\xi} \prod_{i \in W} \left\{ e^{-x c_i} \left[ 1 - (1 - c_i) \prod_{j \in \partial i} (1 - c_j) \right] \right\}, \quad (8) \]

where \( \xi \equiv (c_1, c_2, \ldots, c_N) \) denotes one of the \( 2^N \) possible occupation configurations, \( x > 0 \) is a re-weighting parameter, and \( \partial i \) denotes node \( i \)'s set of adjacent nodes. The constraint of each node \( i \) leads to a multiplication term \( \left[ 1 - (1 - c_i) \prod_{j \in \partial i} (1 - c_j) \right] \), which equals to 0 if \( i \) and all its adjacent nodes are empty and equals to 1 if otherwise. The partition function therefore only takes into account all the dominating sets, and at \( x \to \infty \) it is contributed exclusively by the MDS configurations.

We can solve the spin glass model [8] by a replica-symmetric mean-field theory, which can be understood from the angle of Bethe-Peierls approximation [20] or derived alternatively through partition function expansion [21, 22]. The marginal probability \( q_i^c \) of node \( i \)'s occupation state being \( c \in \{0, 1\} \) is expressed as

\[ q_i^c = \frac{e^{-xc} \prod_{j \in \partial i} q_j^{(c_j,c_j)} - \delta_0^c \prod_{j \in \partial i} q_j^{(0,0)}}{\sum_{c_i} e^{-xc_i} \prod_{j \in \partial i} \sum_{c_j} q_j^{(c_j,c_j)} - \prod_{j \in \partial i} q_j^{(0,0)}}, \quad (9) \]

where the Kronecker symbol \( \delta_0^m \) is \( \delta_0^n = 1 \) if \( m = n \) and \( \delta_0^m = 0 \) if otherwise. The quantity \( q_j^{(c_j,c_i)} \) is defined as the joint probability that node \( i \) is in occupation state \( c_i \) and its adjacent node \( j \) is in occupation state \( c_j \) when the constraint of node \( i \) is not considered. This probability can be evaluated through the following belief-propagation equation:

\[ q_j^{(c_j,c_i)} = \frac{e^{-xc_j} \prod_{k \in \partial j \setminus i} q_k^{(c_k,c_j)} - \delta_0^{c_j} \prod_{k \in \partial j \setminus i} q_k^{(0,0)}}{\sum_{c_i'} e^{-xc_i'} \prod_{k \in \partial j \setminus i} \sum_{c_k} q_k^{(c_k,c_j)} - \prod_{k \in \partial j \setminus i} q_k^{(0,0)}}, \quad (10) \]

where \( \partial j \setminus i \) denotes the subset obtained by deleting node \( i \) from set \( \partial j \).

The total free energy \( F \) is related to the partition function by \( F = -1/x \ln Z \). According to the RS mean-field theory, its expression is

\[ F = \sum_{i \in W} f_i - \sum_{(i,j) \in W} f_{(i,j)}, \quad (11) \]

where \( f_i \) and \( f_{(i,j)} \) are the free energy contributions of a node \( i \) and a link \((i,j)\) between nodes \( i \) and \( j \):

\[ f_i = -\frac{1}{x} \ln \left[ \sum_{c_i} e^{-xc_i} \prod_{j \in \partial i} \sum_{c_j} q_j^{(c_j,c_i)} - \prod_{j \in \partial i} q_j^{(0,0)} \right], \quad (12) \]

\[ f_{(i,j)} = -\frac{1}{x} \ln \left[ \sum_{c_i, c_j} q_j^{(c_j,c_i)} - q_j^{(c_j,c_i)} \right]. \quad (13) \]

From Eqs. (11) and (9) we can compute the free energy density \( f = F/N \) and the mean occupation fraction \( w = (1/N) \sum_{i \in W} q_i^1 \). The entropy density of the system is then evaluated as \( s = (w - f)x \).

For ER and RR networks, we determine the ensemble-averaged minimum occupation fraction \( w \) by choosing the largest re-weighting value \( x \) under the constraint of non-negative entropy density [16]. We then construct dominating sets for single network instances through a belief-propagation-guided decimation (BPD) process [21, 22], namely we repeatedly fix a tiny fraction \( r \leq 0.01 \) of the unoccupied nodes \( i \) with the largest estimated occupation probability \( q_i^1 \) and then update the occupation probabilities for all the remaining unoccupied nodes by belief-propagation iterations [16]. The results of this BPD algorithm (with fixed \( x = 10 \)) and the predicted minimum occupation fractions are shown in Fig. 2 and compared with the results obtained by the pure greedy and the hybrid local algorithms. When the network contains an extensive core, the sizes of the dominating sets obtained by

FIG. 1: Generalized leaf-removal on Erdős-Rényi random networks. \( w \) and \( n_{\text{core}} \) are the fractions of occupied nodes and unobserved nodes, respectively. Cross symbols are obtained by running GLR on a single ER network of size \( N = 10^6 \) and mean degree \( c \), solid lines are the predictions of the percolation theory. Dashed vertical line marks the core percolation transition point.
FIG. 2: Constructing dominating sets. The relative sizes \( w \) of dominating sets obtained by a single running of the pure greedy, the hybrid, and the BPD algorithm with \( x = 10 \) on 96 ER or RR network instances of \( N = 10^5 \) and (mean) degree \( c \) are compared (fluctuations are of order \( 10^{-4} \) and are not shown). The ensemble-averaged MDS relative sizes obtained by the RS theory are also shown. Dashed vertical line marks the core percolation transition point of ER networks.

the two local algorithms are extensively larger than the MDS size. On the other hand, the sizes of the dominating sets obtained by BPD are very close to the predicted MDS size by the RS theory, indicating that BPD is able to construct near-optimal dominating sets for random networks.

Real-world networks are often very heterogenous, with a small fraction of highly connected nodes [18]. For such networks we find that the local hybrid algorithm and the BPD algorithm have comparable performance, and they both outperform the pure greedy algorithm [16].

Discussion. We have proposed two algorithms, a percolation theory, and a mean-field spin-glass theory for the network dominating set problem. These results may be very useful in many real-world applications. Our research will be followed by more thorough theoretical investigations. An easy extension of the percolation theory is to consider GLR with a subset of initially occupied nodes. By optimizing this initial subset, we may reach an improved lower bound to the MDS size. Core percolation on degree-correlated random networks [23] and in the more general lattice glass problem [3] are also interesting issues. When the random network has an extensive core, we notice the belief-propagation equation (10) fails to converge at large values of the re-weighting parameter \( x \), indicating a spin glass phase transition. A systematic study of the spin glass phase will be carried out using the first-step replica-symmetry-breaking mean-field theory [25–27], which may offer an improved estimate on the ensemble-averaged MDS size. The possible deep connections between core percolation and the complexity of the random MDS problem will also be addressed by adapting the long-range frustration theory [28]. The methods of this work can also be readily extended to the MDS problem of directed networks.

Acknowledgments

Part of this work was done when HJZ was participating in the “Collective Dynamics in Information Systems 2014” Program of Kavli Institute for Theoretical Physics China (KITPC). HJZ thanks Chuang Wang for a helpful discussion, and Yang-Yu Liu and Yi-Fan Sun for helpful comments on the manuscript. HJZ thanks Prof. Zhong-Can Ou-Yang for support. HJZ was supported by the National Basic Research Program of China (2013CB932804) and the National Science Foundation of China (11121403, 11225526).

[1] Haynes, T. W., Hedetniemi, S. T. & Slater, P. J. Fundamentals of Domination in Graphs (Marcel Dekker, New York, 1998).
[2] Yang, Y., Wang, J. & Motter, A. E. Network observability transitions. Phys. Rev. Lett. 109, 258701 (2012).
[3] Biroli, G. & Mézard, M. Lattice glass models. Phys. Rev. Lett. 88, 025501 (2002).
[4] Hartmann, A. K. & Weigt, M. Statistical mechanics of the vertex-cover problem. J. Phys. A: Math. Gen. 36, 11069–11093 (2003).
[5] Zhao, J.-H. & Zhou, H.-J. Statistical physics of hard combinatorial optimization: Vertex cover problem. Chin. Phys. B 23, 078901 (2014).
[6] Takaguchi, T., Hasegawa, T. & Yoshida, Y. Suppressing epidemics on networks by exploiting observer nodes. Phys. Rev. E 90, 012807 (2014).
[7] Liu, Y.-Y., Slotine, J.-J. & Barabási, A.-L. Controllability of complex networks. Nature 473, 167–173 (2011).
[8] Nacher, J. C. & Akutsu, T. Analysis on critical nodes in controlling complex networks using dominating sets. In International Conference on Signal-Image Technology
[9] Wuchty, S. Controllability in protein interaction networks. *Proc. Natl. Acad. Sci. USA* **111**, 7156–7160 (2014).

[10] Wu, J. & Li, H. A dominating-set-based routing scheme in ad hoc wireless networks. *Telecommun. Syst.* **18**, 13–36 (2001).

[11] Bramoulle, Y. & Kranton, R. Public goods in networks. *J. Econom. Theor.* **135**, 478–494 (2007).

[12] Lund, C. & Yannakakis, M. On the hardness of approximating minimization problems. *J. ACM* **41**, 960–981 (1994).

[13] Raz, R. & Safra, S. A sub-constant error-probability low-degree test, and a sub-constant error-probability pcp characterization of np. In *Proc. 29th Annual ACM Symposium on Theory of Computing*, 475–484 (ACM, New York, 1997).

[14] Molnár Jr., F., Sreenivasan, S., Szymanski, B. K. & Korniss, K. Minimum dominating sets in scale-free network ensembles. *Sci. Rep.* **3**, 1736 (2013).

[15] Bauer, M. & Golinelli, O. Core percolation in random graphs: a critical phenomena analysis. *Eur. Phys. J. B* **24**, 339–352 (2001).

[16] See supplementary information for more details.

[17] He, D.-R., Liu, Z.-H. & Wang, B.-H. *Complex Systems and Complex Networks* (Higher Education Press, Beijing, 2009).

[18] Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).

[19] Zhao, J.-H., Zhou, H.-J. & Liu, Y.-Y. Inducing effect on the percolation transition in complex networks. *Nature Communications* **4**, 2412 (2013).

[20] Mézard, M. & Montanari, A. *Information, Physics, and Computation* (Oxford Univ. Press, New York, 2009).

[21] Xiao, J.-Q. & Zhou, H. J. Partition function loop series for a general graphical model: free-energy corrections and message-passing equations. *J. Phys. A: Math. Theor.* **44**, 425001 (2011).

[22] Zhou, H. J. & Wang, C. Region graph partition function expansion and approximate free energy landscapes: Theory and some numerical results. *J. Stat. Phys.* **148**, 513–547 (2012).

[23] Mézard, M., Parisi, G. & Zecchina, R. Analytic and algorithmic solution of random satisfiability problems. *Science* **297**, 812–815 (2002).

[24] Hasegawa, T., Takaguchi, T. & Masuda, N. Observability transitions in correlated networks. *Phys. Rev. E* **88**, 042809 (2013).

[25] Mézard, M. & Parisi, G. The bethe lattice spin glass revisited. *Eur. Phys. J. B* **20**, 217–233 (2001).

[26] Mézard, M. & Montanari, A. Reconstruction on trees and spin glass transition. *J. Stat. Phys.* **124**, 1317–1350 (2006).

[27] Krzakala, F., Montanari, A., Ricci-Tersenghi, F., Semerjian, G. & Zdeborova, L. Gibbs states and the set of solutions of random constraint satisfaction problems. *Proc. Natl. Acad. Sci. USA* **104**, 10318–10323 (2007).

[28] Zhou, H. J. Long-range frustration in a spin-glass model of the vertex-cover problem. *Phys. Rev. Lett.* **94**, 217203 (2005); Erratum, *Phys. Rev. Lett.* **109**, 199901 (2012).
Supplementary Information

**Generalized leaf-removal (GLR)**

Given an input network $W$, let us denote a minimum dominating set (MDS) of this network as $\Gamma_0$ (there must be at least one such set). For the convenience of description, let us denote by $W_t$ the simplified network at the start of the $t$-th evolution step of GLR. $W_0$ at the initial evolution step $t = 0$ is identical to the original network $W$, and all the nodes of $W_0$ are unobserved. We now prove that, if GLR makes the whole input network $W$ be observed, then the set of nodes occupied during this process must be a MDS. The essential idea is to demonstrate that during GLR, we can modify $\Gamma_0$ in such a way that its size does not change but all the nodes $i$ that are fixed to be occupied ($c_i = 1$) by GLR are in $\Gamma_0$ while all the nodes $j$ that are fixed to be unoccupied ($c_j = 0$) by GLR are not in $\Gamma_0$.

Starting from evolution step $t = 0$, let us perform GLR and modify $\Gamma_0$ in the following sequential order:

1. As long as there is a leaf node $i$ in network $W_t$, we fix its occupation state to $c_i = 1$ and delete it from $W_t$. All such fixed nodes $i$ must also belong to $\Gamma_0$, so at the moment there is no need to modify $\Gamma_0$.

2. Then as long as there is a node $i$ which is itself observed and which has only a single unobserved neighbor $j$, we delete the link $(i, j)$ from network $W_t$, see right panel of Fig. 3. We do not modify $\Gamma_0$ if node $i$ does not belong to it. If node $i$ does belong to $\Gamma_0$, then node $j$ must not belong to it, and in this later case we add $j$ to $\Gamma_0$ and delete $i$ from it.

3. Then as long as there is an observed node $i$ which is not connected to any unobserved node, we fix its occupation state to $c_i = 0$ and delete it and all its attached links from $W_t$. Such a node $i$ must not belong to $\Gamma_0$, for otherwise $\Gamma_0$ could not have been a MDS. Therefore we do not modify $\Gamma_0$.

4. If the resulting network $W_t$ contains at least one isolated node or a leaf node, then we increase the evolution step from $t$ to $(t + 1)$ and set the network $W_{t+1}$ as identical to $W_t$. It is easy to verify that all the nodes of $W_{t+1}$ are not occupied, and we regard a node $i$ of $W_{t+1}$ to be observed if it is observed in network $W_t$, otherwise $i$ is regarded as unobserved. We then repeat the above-mentioned operations (1)–(3) sequentially.

If all the nodes of the original network $W$ are observed at the end of the GLR process, then the set $\Gamma$ of occupied nodes by this process must be identical to the final $\Gamma_0$, which is a MDS modified from the original MDS.

It is also easy to verify that, if the GLR process finishes with a core of unobserved nodes, the set of occupied nodes by it must be a MDS for the subnetwork of $W$ induced by all the nodes outside the core.

Solution of the percolation theory

Our percolation theory can be applied both to single finite random network instances and to random network ensembles at the thermodynamic limit $N \to \infty$. This
theory has four sets of parameters $\alpha_t$, $\beta_t$, $\gamma_t$ and $\eta_t$, with the subscript $t$ being the evolution step of GLR ($t = 0, 1, 2, \ldots$). Given a degree probability distribution profile $P(d)$ as input, we first obtain another degree probability distribution profile $Q(d)$ through $Q(d) = dP(d)/c$ ($c$ being the mean node degree), and then compute these four sets of parameters iteratively, starting from $t = 0$. For each $t$, we first compute $\alpha_t$ from Eq. (3), then use $\alpha_t$ as input to compute $\beta_t$ from Eq. (4), then use $\alpha_t$, $\beta_t$, and $\gamma_t$ as inputs to compute $\eta_t$ from Eq. (5), and finally use $\alpha_t$, $\beta_t$, $\gamma_t$, and $\eta_t$ as inputs to compute $Q(d)$ from Eq. (6). After the four sets of parameters $\alpha_t$, $\beta_t$, $\gamma_t$, and $\eta_t$ are numerically determined in this way, we then evaluate the relative size $n_{\text{core}}$ of the unobserved core and the fraction $w$ of occupied nodes using Eqs. (1) and (2).

The initial condition of this numerical iteration is $\alpha_0 = Q(1)$. For a finite random network of $N$ nodes, the iteration stops if the evolution step $t$ increases to a threshold value $t^*$ such that $\alpha_{t^*} < 1/N$. This is because if $N\alpha_{t^*} < 1$, the number of newly generated unobserved leaf nodes has a high probability to be zero and then GLR will be unable to continue. For the case of $N \to \infty$, the numerical iteration process can be carried out to sufficiently large evolution steps $t$ until $\alpha_t \to 0$.

**Local algorithms**

We consider two local algorithms, the pure greedy algorithm and the hybrid algorithm. The pure greedy algorithm is based on the heuristic idea of occupying an empty node that has the maximal impact value in the network. The impact of an unoccupied node $i$ equals to the number of nodes that will be observed by occupying $i$. For example, if node $i$ has 3 unobserved neighbors, its impact is 4 if $i$ is itself unobserved and is 3 if $i$ is observed (i.e., it is adjacent to one or more observed nodes). Starting from an input network $W$ with all the nodes unobserved, this greedy algorithm selects uniformly at random a node $i$ from the subset of nodes with the highest impact and fix its occupation state to $c_i = 1$. All the adjacent nodes of $i$ are then observed. If there are still unobserved nodes in the network, the impact value for each of the unobserved nodes is updated and the greedy occupying process is repeated.

The hybrid local algorithm combines this impact-based greedy process with the GLR process. Given an input network $W$ with all the nodes unobserved, we first carry out the GLR process to simplify the network as far as possible. If some nodes are left unobserved after this GLR process (i.e., there is a core in the original network $W$), we first fix a randomly chosen highest-impact node $i$ to the occupation state $c_i = 1$, which makes $i$ and all its adjacent nodes be observed, and then perform the GLR process to further simplify the network as far as possible. We keep repeating such a combined fixation-plus-GLR process until there is no unobserved node left in the network.

**Belief-propagation (BP) iterations**

According to Eq. (10) each probability distribution $q_j^{(0,0)}$, $q_j^{(1,0)}$ has the property that $q_j^{(1,1)} = q_j^{(0,1)}$. Therefore in the numerical computations $q_j^{(c_i,c_j)}$ can be represented by three non-negative real numbers $q_j^{(0,0)}$, $q_j^{(0,1)}$, and $q_j^{(1,0)}$, which satisfy in addition the normalization condition

$$q_j^{(0,0)} + q_j^{(0,1)} + 2q_j^{(1,0)} = 1.$$  \hfill (14)

We initialize $q_j^{(c_i,c_j)}$ and $q_j^{(c_i,c_j)}$ for each link $(i, j)$ of the network between two nodes $i$ and $j$, for example setting $q_j^{(0,0)} = q_j^{(0,1)} = q_j^{(1,0)} = 1/4$. We then perform BP iteration a number $T$ of times at a given value of the re-weighting parameter $x$, until a fixed-point solution of Eq. (10) is reached or $T$ exceeds a pre-specified number (e.g., 1000). In each BP iteration step we treat all the nodes of the network in a random order. When node $j$ is examined, the output messages $q_j^{(c_i,c_j)}$ to all its adjacent nodes $i \in \partial j$ are updated according to Eq. (10). The difference $\Delta_j^{(t)}$ between an updated message $q_j^{(t)}$ at the $t$-th BP step and the old message $q_j^{(t-1)}$ at the $(t-1)$-th BP step is defined as

$$\Delta_j^{(t)} = |q_j^{(0,0)}(t) - q_j^{(0,0)}(t-1)| + |q_j^{(1,0)}(t) - q_j^{(1,0)}(t-1)| + 2|q_j^{(1,0)}(t) - q_j^{(1,0)}(t-1)|.$$  \hfill (15)

If the maximal value among the set of $2M$ difference values $\{\Delta_j^{(t)}\}$ is less than certain pre-specified threshold value (e.g., $10^{-3}$ or even smaller), then BP iteration is regarded as being converged. At a fixed point of Eq. (10), we then compute the free energy density $f$, the mean occupation fraction $w$, and the entropy density $s$ through the RS mean field theory. As an example, we show in Fig. 4 the results obtained on a single Erdős-Rényi (ER) random network with $N = 10^6$ nodes and mean degree $c = 10$.

For ER networks with mean degree $c > 2.41$ and regular random (RR) networks with integer degree $c \geq 3$, we find that when the re-weighting parameter $x$ is larger than certain threshold value, BP iteration is unable to converge to a fixed point. Such a non-convergence phenomenon indicates that, when the random network system has an extensive core, it will be in a spin glass phase at sufficiently large values of $x$. Our systematical theoretical results on this spin glass phase will be reported in another publication.
Ensemble-averaged properties

A random network ensemble is characterized by a degree distribution $P(d)$. We perform population dynamics simulations using Eqs. (10), (9) and (11) to obtain ensemble-averaged results. First, we create a long array $A$ of $N$ (e.g., $10^5$) elements to store a set of messages, each of which represents a probability distribution $q_j^{(c_j,c_i)}$ in the form of a three-dimensional vector satisfying Eq. (14): $(q_j^{(0,0)}, q_j^{(0,1)}, q_j^{(1,0)})$. We then repeatedly update elements of this array by the following procedure: (1) generate a random integer $d \geq 1$ according to the degree probability distribution $Q(d)$; (2) draw $(d - 1)$ elements $q_k^{(c_k,c_j)}$ from array $A$ uniformly at random, and then use these $(d - 1)$ elements as input messages to Eq. (10) to compute a new message $q_j^{(c_j,c_i)}$; (3) replace a randomly chosen element of array $A$ with this new message. The message array $A$ is expected to reach a steady state after it is updated a sufficient number of times (e.g., after each element of this array is updated 10,000 times on average).

We then keep updating the message array $A$ and at the same time compute the thermodynamic quantities $f$, $w$, and $s$. For example, the free energy density $f$ is obtained by

$$f = \bar{f}_i - \frac{c}{2} \bar{f}_{(i,j)},$$

where $\bar{f}_i$ means the average of the free energy node contribution $f_i$ over all the nodes, and $\bar{f}_{(i,j)}$ means the average of the free energy link contribution $f_{(i,j)}$ over all the links. We generate many samples of $f_i$ and $f_{(i,j)}$ to compute their averages $\bar{f}_i$ and $\bar{f}_{(i,j)}$. The procedure of obtaining a sample of $f_i$ is the same as the procedure of obtaining a sample of the message $q_j^{(c_j,c_i)}$, the only difference being that the degree $d_i$ of node $i$ should be generated according to the distribution $P(d)$ instead of $Q(d)$. A sample of $f_{(i,j)}$ is obtained very easily through Eq. (13) by picking two messages $q_j^{(c_j,c_i)}$ and $q_j^{(c_i,c_j)}$ uniformly at random from the message array $A$.

For ER random networks with mean degree $c = 10$, we compare in Fig. 4 the results obtained by this replica-symmetric (RS) population dynamics with the results obtained by BP iteration on single network instances. The ensemble-averaged results are indeed in perfect agreement with the BP iteration results (provided the BP iteration is able to converge).

The entropy density $s$ as a function of the mean occupation fraction $w$ can be obtained from these RS population dynamics results (see for example Fig. 4D). In some random network systems, the entropy density $s$ become negative if $w$ decreases below certain threshold value $w_0$, indicating that there is no dominating sets with relative size below $w_0$. We therefore take the value $w_0$ as the ensemble-averaged MDS relative size. For the ER networks of $c = 10$, we obtain from Fig. 4 that $w_0 \approx 0.120$ (the corresponding value of $x$ is $x \approx 8.637$). In some other random network systems (e.g., ER random networks with $c < 2.41$, before the core percolation transition), the entropy density $s$ approaches a non-negative limiting value as $w$ approaches a limiting value $w_{\text{bp}}$ from above. For these later cases, we simply take $w_0$ as the ensemble-averaged MDS relative size.

Belief-propagation-guided decimation (BPD)

For a given network $W$, the RS mean field theory gives an estimate for the occupation probability $q_j^{+1}$ of each node $i$, see Eq. (9). Such information is exploited in the BPD algorithm to reduce the size of a dominating set. At each round of the BPD process, unoccupied nodes with the highest estimated occupation probabilities are added to the dominating set, and the occupation probabilities for the remaining unoccupied nodes are then updated.

If a node $j$ is unobserved (it is empty and has no adjacent occupied node), the output message $q_j^{(c_j,c_i)}$ on the link $(j,i)$ between nodes $j$ and $i$ is updated according to Eq. (10). On the other hand, if node $j$ is empty but observed (it has at least one adjacent occupied node), this node then presents no restriction to the occupation states of all its unoccupied adjacent nodes. For such a
node \( j \), the output message \( q_{j \rightarrow i}^{(c_j,c_i)} \) on the link \((j,i)\) is then updated according to the following equation:

\[
q_{j \rightarrow i}^{(c_j,c_i)} = \frac{e^{-xc_j} \prod_{k \in \partial j \setminus i} \sum_{c_k} q_{k \rightarrow j}^{(c_k,c_i)}}{\sum_{c_j',c_i'} e^{-xc_j'} \prod_{k \in \partial j \setminus i} \sum_{c_k} q_{k \rightarrow j}^{(c_k,c_i')}}. \tag{17}
\]

Similar to Eq. (17), the marginal probability distribution \( q_i^{(c_i)} \) for an observed empty node \( i \) is evaluated according to

\[
q_i^{(c_i)} = \frac{e^{-xc_i} \prod_{j \in \partial i} \sum_{c_j} q_{j \rightarrow i}^{(c_j,c_i)}}{\sum_{c_i'} e^{-xc_i'} \prod_{j \in \partial i} \sum_{c_j} q_{j \rightarrow i}^{(c_j,c_i')}}. \tag{18}
\]

It is easy to verify from Eq. (17) that \( q_i^{(0,0)} = q_{j \rightarrow i}^{(0,1)} \) and \( q_i^{(1,0)} = q_{j \rightarrow i}^{(1,1)} \). Notice that if all the nodes in the set \( \partial j \setminus i \) are observed, then we derive from Eq. (17) that \( q_i^{(0,0)} = q_i^{(1,0)} = q_j^{(0,1)} = q_j^{(1,1)} = 1/4 \). Because of this property, we need only to consider the links between two unobserved nodes and the links between an unobserved node and an observed node. All the other links (which are between two observed nodes) do not need to be considered in the BP iteration equations (10) and (17).

We implement the BPD algorithm as follows:

0. Input the network \( W \), set all the nodes to be empty and set all the probability distributions \( q_{j \rightarrow i}^{(c_j,c_i)} \) to be the uniform distribution. Set the re-weighting parameter \( x \) to a sufficiently large value (e.g., \( x = 10 \)).

Then perform the BP iteration a number \( T_0 \) of rounds (e.g., \( T_0 = 200 \)). After these \( T_0 \) iterations, we then compute the occupation probability \( q_i^{+1} \) of each node \( i \) using Eq. (9).

1. We then occupy a small fraction \( r \) (e.g., \( r = 0.01 \)) of the unoccupied nodes that have the highest estimated occupation probabilities.

2. We then simplify network \( W \) by first deleting all the links between observed nodes, and then deleting all the isolated and observed nodes.

3. If the resulting network \( W \) still contains unobserved nodes, we perform BP iteration for a number of \( T_1 \) rounds (e.g., \( T_1 = 10 \)). The output message of an unobserved node \( i \) is updated either according to Eq. (9) or according to Eq. (18), depending on whether \( i \) is unobserved or observed. We then repeat operations (1) and (2) in sequential order until all the nodes are observed.

In addition, we may first carry out the GLR process to simplify the network \( W \) as far as possible before running the BPD process. For real-world networks with some nodes being highly connected, we find that such a GLR simplifying step reduces the BPD running time considerably and also slightly reduces the size of the constructed dominating set.

### Real-world network instances

As an initial test of the algorithms introduced in this work, we apply the GLR process, the pure impact-based greedy algorithm, the hybrid algorithm, and the BPD algorithm to a set of twelve real-world networks. The numerical results are summarized in Table I.

| Network   | \( N \) | \( M \) | \( d_{max} \) | Core | Greedy | Hybrid | BPD |
|-----------|--------|--------|--------------|------|--------|--------|-----|
| EuroRoad  | 1177   | 1417   | 10           | 306  | 428    | 389    | 387 |
| PPI       | 2361   | 6646   | 64           | 17   | 550    | 539    | 539 |
| PowerGrid | 4941   | 6594   | 19           | 603  | 1564   | 1485   | 1481 |
| Internet-1| 4674   | 12752  | 1458         | 8    | 660    | 566    | 656 |
| CoAuthor  | 23133  | 93439  | 279          | 9052 | 3686   | 3612   | 3604 |
| Citation  | 34546  | 420877 | 846          | 1178 | 3335   | 3168   | 3095 |
| PeerToPeer| 62586  | 147892 | 95           | 35   | 12710  | 12582  | 12582 |
| Friendship| 196691 | 950327 | 14730        | 6097 | 42536  | 41633  | 41672 |
| Email     | 265214 | 364481 | 7636         | 470  | 18183  | 18181  | 18181 |
| WebPages  | 875713 | 4322051| 6332         | 162439| 81288  | 79928  | 80769 |
| RoadTX    | 1379917| 1921660| 12           | 560582| 477729 | 437503 | 425774 |
| Internet-2| 1696415| 11095298| 35455       | 211244| 187592 | 183516 | 183248 |

TABLE I: \( N \) and \( M \) are, respectively, the total number of nodes and links in the network; \( d_{max} \) is the maximal node degree of the network; the column marked by ‘Core’ lists the number of nodes that are left unobserved after the GLR process; the columns marked by ‘Greedy’, ‘Hybrid’, and ‘BPD’ list, respectively, the size of the dominating set constructed by a single running of the impact-based pure greedy algorithm, the hybrid local algorithm, and the BPD message-passing algorithm.
than the hybrid local algorithm: In ten out of the twelve examined networks instances, the size of the dominating sets constructed by BPD is equal to or slightly smaller than that of the dominating set constructed by the hybrid algorithm. Both BPD and the hybrid algorithm outperform the pure greedy algorithm in all the twelve network instances.

**Additional references**

29. Šubelj, L. & Bajec, M. Robust network community detection using balanced propagation. Eur. Phys. J. B 81, 353–362 (2011).

30. Leskovec, J., Lang, K. J., Dasgupta, A. & Mahoney, M. W. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. Internet Math. 6, 29–123 (2009).

31. Watts, D. J. & Strogatz, S. H. Collective dynamics of ’small-world’ networks. Nature 393, 440–442 (1998).

32. Leskovec, J., Kleinberg, J. & Faloutsos, C. Graphs over time: densification laws, shrinking diameters and possible explanations. In “Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining”, 177–187 (ACM, New York, 2005).

33. Leskovec, J., Kleinberg, J. & Faloutsos, C. Graph evolution: Densification and shrinking diameters. ACM Transactions on Knowledge Discovery from Data 1, 2 (2007).

34. Ripeanu, M., Foster, I. & Iamnitchi, A. Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. IEEE Internet Comput. 6, 50–57 (2002).

35. Cho, E., Myers, S. A. & Leskovec, J. Friendship and mobility: User movement in location-based social networks. In “ACM SIGKDD International Conference on Knowledge Discovery and Data Mining”, 1082–1090 (San Diego, CA, USA, 2011).

36. Bu, D., et al. Topological structure analysis of the protein-protein interaction network in budding yeast. Nucleic Acids Res. 31, 2443–2450 (2003).