J Stat Phys (2015) 159:1154–1174
DOI 10.1007/s10955-015-1220-2

Statistical Mechanics of the Minimum Dominating Set Problem

Jin-Hua Zhao · Yusupjan Habibulla · Hai-Jun Zhou

Received: 11 November 2014 / Accepted: 12 February 2015 / Published online: 21 February 2015

Abstract The minimum dominating set (MDS) problem has wide applications in network science and related fields. It aims at constructing a node set of smallest size such that any node of the network is either in this set or is adjacent to at least one node of this set. Although this optimization problem is generally very difficult, we show it can be exactly solved by a generalized leaf-removal (GLR) process if the network contains no core. We present a percolation theory to describe the GLR process on random networks, and solve a spin glass model by mean field method to estimate the MDS size. We also implement a message-passing algorithm and a local heuristic algorithm that combines GLR with greedy node-removal to obtain near-optimal solutions for single random networks. Our algorithms also perform well on real-world network instances.

Keywords Dominating set · Spin glass · Core percolation · Leaf removal · Network coarse-graining · Belief propagation

1 Introduction

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? As an example we show in Fig. 1b a minimum dominating set (containing only three nodes) of a small network. A more stringent

J.-H. Zhao · Y. Habibulla · H.-J. Zhou (✉)
State Key Laboratory of Theoretical Physics, Institute of Theoretical Physics, Chinese Academy of Sciences, Zhong-Guan-Cun East Road 55, Beijing 100190, China
e-mail: zhouhj@itp.ac.cn

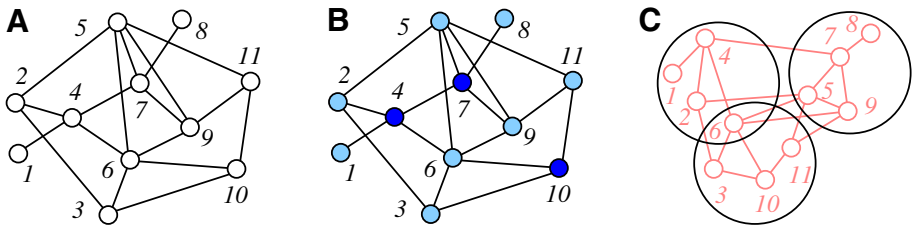

Fig. 1 An example of minimum dominating set. **a** a small network with $N = 11$ nodes and $M = 18$ links. **b** *Blue* (*dark gray*) indicates a node being occupied, while *cyan* (*light gray*) indicates a node being empty but observed. The three occupied nodes form a MDS $\Gamma_0 = \{4, 7, 10\}$ for this network. **c** A coarse-grained representation of the network based on the MDS of (**b**) (Color figure online)

constraint, which is adopted in lattice glass models [3], is to require an empty node i to be surrounded by at least l_i occupied nodes, with l_i being node-dependent. The MDS problem corresponds to the case of $l_i \equiv 1$, 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 node i's degree (i.e., number of adjacent nodes).

The MDS problem has 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 (e.g., resource allocation) [11]. Another application is to build a coarse-grained representation for a complex network starting from a MDS. Such an idea has already been applied to multi-document summarization in the field of information extraction [12]. Each node i of the MDS can be regarded as a representative node for a local domain of the network. We can take the subnetwork induced by node i and all its adjacent nodes (except those in the MDS) as a coarse-grained node, and set up an coarse-grained link between two coarse-grained nodes if the two corresponding subnetworks share at least one node or are connected by at least one link in the original network (see Fig. 1c for an example). If such a coarse-graining process is iterated we will then obtain a hierarchical representation for the original network, which may be very useful for understanding the organization of a complex system and for searching and information transmission in such a system.

Exactly solving the MDS problem, however, 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 $\ln N$ times of the minimum size [13,14]. Many local-search algorithms have been proposed to solve the MDS problem heuristically (see review [1] and [2,6,7,9,15,16]), but theoretical results on the MDS sizes of random network ensembles are still very rare.

In this work we bring several new theoretical and algorithmic contributions. We show in Sect. 2 that a generalized leaf-removal (GLR) process may cause a core percolation transition, and 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 in Sect. 3 to get an upper bound to the MDS size. We then introduce a spin glass model in Sect. 4 and estimate the MDS size by a replica-symmetric (RS) mean field theory, and implement a message-passing algorithm in Sect. 5 to get near-optimal dominating sets for single random network instances. Our algorithms also perform well on real-world network instances. This work shall be useful both for network scientists who are interested in

applying the MDS concept to practical problems, and for applied mathematicians who seek better theoretical understanding on the random MDS problem.

2 Generalized Leaf-Removal and Core Percolation

Consider a simple network W formed by N nodes and M undirected links, each link connecting between two different nodes. Each node with index $i \in \{1, 2, \ldots, N\}$ 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, otherwise it is regarded as unobserved. We need to occupy a set Γ of nodes to make all the N nodes be observed, and the objective is to make the dominating set Γ as small as possible, i.e., to construct a minimum dominating set. It is easy to verify that the three occupied nodes of Fig. 1b form a MDS for that small network. Notice a network may have more than one MDS.

2.1 The GLR Process

Here we extend the leaf-removal idea of [17] (see also more recent work [6,18,19]) and consider a generalized leaf-removal process. This dynamics is based on the following two considerations: first, as pointed out in [6,17], 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 an empty but observed node and at most one of its adjacent nodes is unobserved, then it must be an optimal strategy *not* to occupy i. This second point was not considered in the conventional leaf-removal process [6].

The GLR process simplifies the input network W at discrete evolution steps $t = 0, 1, 2, \ldots$. 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 step $t = 0$ is identical to the original network W, and all the nodes of W_0 are unobserved. We 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. For this latter purpose, let us denote by Γ_0 a MDS of the input network W (there must be at least one such set). The essential idea is to demonstrate that during GLR, we can modify Γ_0 in such a way that its size does not change but all the nodes i that are fixed to be occupied ($c_i = 1$) are in Γ_0 while all the nodes j that are fixed to be unoccupied ($c_j = 0$) are not in Γ_0. Starting from evolution step $t = 0$, let us perform GLR and modify Γ_0 in the following sequential order:

(0) As long as there is an isolated node i in network W_t, fix its occupation state to $c_i = 1$ and delete it from W_t. All such fixed nodes i must also belong to Γ_0.

(1) As long as there is a leaf node i in network W_t which is not yet observed, fix the occupation state of its unique neighbor j to $c_j = 1$ and fix that of i to $c_i = 0$ so that j and all its adjacent nodes (including i) are now observed, see Fig. 2 (left panel). We then delete node j and all its connected links from W_t. If j belongs to Γ_0 then node i must not belong to it, because otherwise Γ_0 could not have been a MDS. On the other hand, if node j does not belong to Γ_0 then node i must belong to it, and in this latter case we modify Γ_0 by adding j to it and deleting i from it.

(2) Then as long as there is a node i which is itself observed and which has only a single unobserved neighbor j, delete the link (i, j) from network W_t, see Fig. 2 (right panel). We do not modify Γ_0 if node i does not belong to it. If node i does belong to Γ_0 then node j must not belong to it, and in this latter case we add j to Γ_0 and delete i from it.

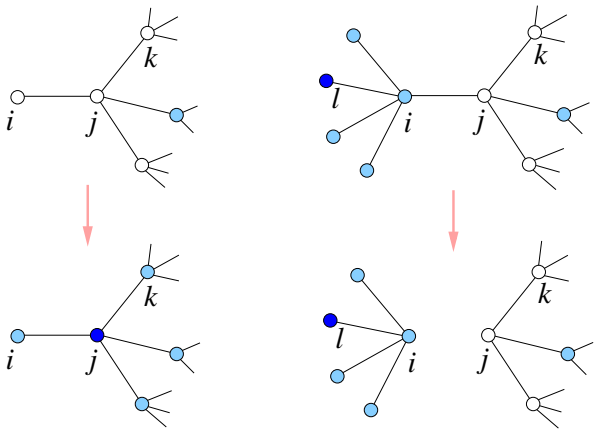

Fig. 2 The two basic operations of the generalized leaf-removal process. *White circles* denote empty and unobserved nodes, *cyan* (*light gray*) *circles* denote empty but observed nodes, and *blue* (*dark gray*) *circles* denote occupied nodes. *Left panel* the unique adjacent node j of an unobserved leaf node i is occupied, and all the neighbors of j are observed. *Right panel* an empty observed node i has only a single unobserved neighbor j, then the link between i and j is deleted (Color figure online)

(3) Then as long as there is an observed node i which is not connected to any unobserved node, 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 Γ_0, for otherwise Γ_0 could not have been a MDS.
(4) If the resulting network W_t is empty or it contains no isolated node nor leaf node, the GLR process stops. If W_t still contains at least one isolated or leaf node, then we increase the evolution step from t to $(t + 1)$ and initialize the network W_{t+1} as identical to W_t. A node i of W_{t+1} is regarded as observed if and only if it is observed in network W_t. We then repeat the above-mentioned operations (1)–(3).

If the final simplified network is non-empty, then there must be some nodes that are still unobserved after the GLR process. 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_{core} the fraction of nodes in this core and by w the fraction of occupied nodes.

If the original network W has no core, then the set Γ of occupied nodes by the GLR process must be identical to the final Γ_0, which is a MDS modified from the original MDS. We have therefore proven that GLR constructs a MDS for a network W if this network contains no core. (All the above-mentioned modification operations on Γ_0 are ignored in the actual implementation of the GLR process. They are introduced here solely for proving that GLR is able to construct a MDS if there is no core.) Furthermore, we notice that if the GLR process finishes with some nodes remaining to be unobserved, the set of nodes occupied during this process must be a MDS for the subnetwork of W induced by all the observed nodes. This is because all these occupied nodes also belong to the modified MDS Γ_0, while all those nodes fixed to be unoccupied are outside of Γ_0.

We generate many large instances of Erdös–Rényi (ER) and scale-free (SF) random networks and run the GLR process on them (details of the network sampling method are given in Sects. 2.3, 2.4, and 2.5). Some representative results are shown in Fig. 3 for ER networks [20,21], in Fig. 4 for SF networks generated through the static model [22], and in Fig. 5 for pure SF networks [20,21]. A major observation is that there is no core in pure SF random networks with minimum node degree $d_{min} = 1$, therefore a MDS for such a network can be easily constructed by the GLR process. Another major observation is that there is a continuous core percolation transition in ER networks and in SF networks generated through the static model. This core percolation transition occurs at certain threshold value of the mean node degree. For example, for ER networks with $N = 10^6$ nodes and $M = (c/2)N$ links,

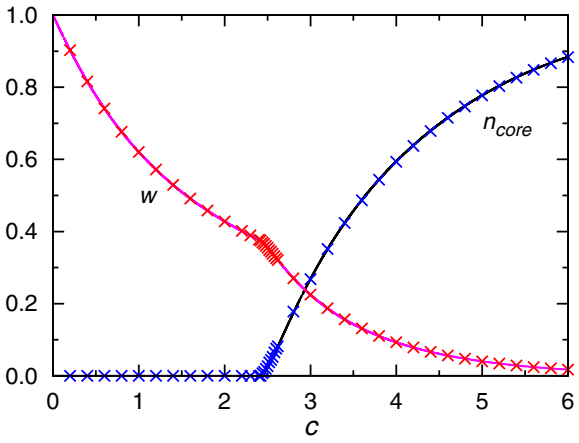

Fig. 3 Generalized leaf-removal on Erdös–Rényi random networks. w and n_{core} are the fractions of occupied and unobserved nodes, respectively. Cross symbols are results obtained by running GLR on a single ER network of size $N = 10^6$ and mean degree c; *solid lines* are the predictions of the percolation theory for $N = \infty$

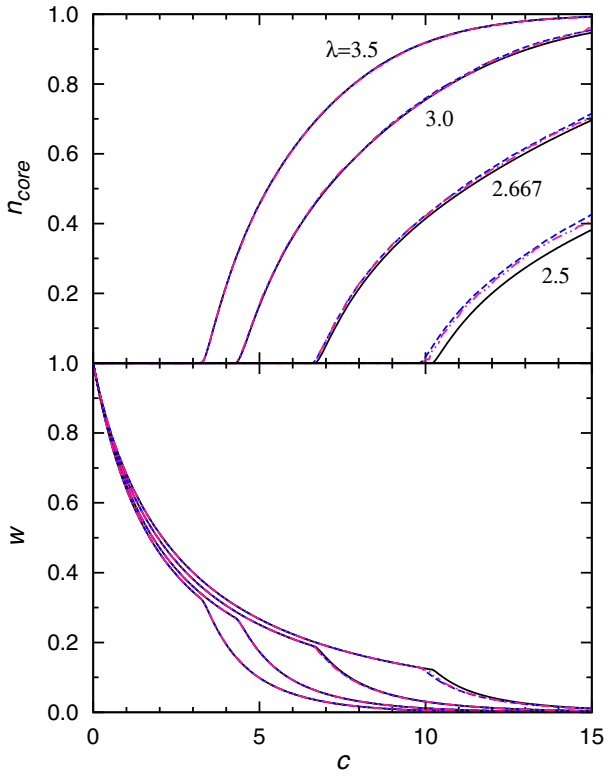

Fig. 4 Generalized leaf-removal on scale-free random networks of decay exponent $\lambda = 3.5, 3.0, 2.667, 2.5$ (from *left to right*) generated through the static model [22]. w and n_{core} are the fractions of occupied and unobserved nodes, respectively. *Red dash-dotted lines* are results obtained by running GLR on a single network instance of size $N = 10^6$ and mean degree c, while *blue dashed lines* are results obtained by the core percolation theory using the degree profile of this network instance as input. *Black solid lines* are the predictions of the percolation theory for $N = \infty$ (Color figure online)

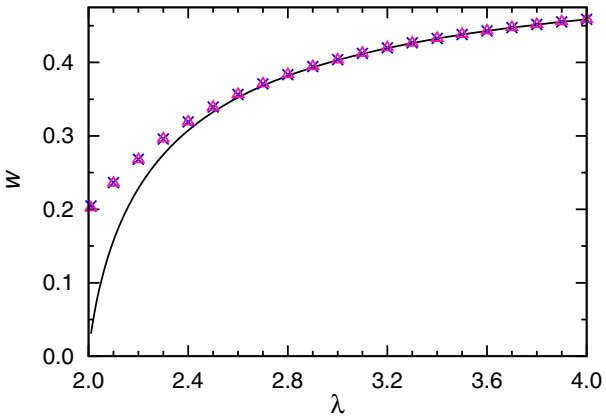

Fig. 5 Generalized leaf-removal on pure scale-free random networks with minimum degree $d = 1$. w is the fraction of occupied nodes. The fraction n_{core} of unobserved nodes is simply $n_{core} = 0$. *Red triangle symbols* are results obtained by running GLR on a single pure SF network of size $N = 10^6$ and decay exponent λ, while blue cross symbols are results obtained by the percolation theory using the degree profile of this network instance as input. The *black solid line* is obtained by the percolation theory at $N = \infty$ (Color figure online)

when the mean node degree $c < 2.41$ there is no core ($n_{core} = 0$), and GLR reaches a MDS for the whole network (Fig. 3). The core emerges at $c \approx 2.41$ and its relative size n_{core} then increases with c continuously from zero. For $c > 2.41$, GLR constructs a MDS only for part of the ER network and it leaves an extensive core of $n_{core}N$ unobserved nodes.

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 resulting from an initial set of randomly chosen occupied nodes. We now develop a percolation theory to thoroughly understand the GLR dynamics on random networks.

2.2 The Core Percolation Theory

A random network is characterized by a degree distribution $P(d)$, which gives the fraction of nodes with degree $d \geq 0$ [20]. We assume that there is no correlation between the degrees of adjacent nodes, therefore the degree d of a node reached by following a randomly chosen link obeys the distribution $Q(d)$ of the form

$$Q(d) \equiv \frac{P(d)d}{c}\ , \tag{1}$$

where $c \equiv \sum_{d\geq 0} P(d)d$ is the mean node degree of the network. Consider a link (i, j) of the network W. Let us neglect for the moment the constraint of node i to node j but only consider the other adjacent nodes of j. If the constraint of node i is neglected, then what is the probability α_t that node j becomes an unobserved leaf node (i.e., it has no other adjacent node except i) at the start of the t-th GLR evolution step? What is the probability β_t that j becomes newly occupied ($c_j = 1$) at the tth GLR step? What is the probability γ_t that j is observed but not occupied at the end of the t-th GLR step? And what is the probability η_t that at the end of the t-th GLR step, node j is an observed and unoccupied node and it has no unobserved adjacent node except i? For an uncorrelated random network these four sets of probability parameters $\{\alpha_0, \alpha_1, \ldots\}$, $\{\beta_0, \beta_1, \ldots\}$, $\{\gamma_0, \gamma_1, \ldots\}$, and $\{\eta_0, \eta_1, \ldots\}$ can be computed by a set of iterative equations.

The expressions of α_0, β_0, γ_0, η_0 for the initial evolution step $t = 0$ are

$$\alpha_0 = Q(1) , \tag{2}$$

$$\beta_0 = 1 - \sum_{d\geq 1} Q(d)(1-\alpha_0)^{d-1} , \tag{3}$$

$$\gamma_0 = \sum_{d\geq 1} Q(d)\left[(1-\alpha_0)^{d-1} - (1-\alpha_0-\beta_0)^{d-1}\right] , \tag{4}$$

$$\eta_0 = \sum_{d\geq 1} Q(d)\left[(\beta_0+\gamma_0)^{d-1} - \gamma_0^{d-1}\right]. \tag{5}$$

Equation (2) is trivial, it simply describes the situation that node j has only a single neighbor (i.e., node i). Equation (3) describes the situation that node j is adjacent to at least one leaf node (except i), which will guarantee j to be occupied at the $t = 0$ GLR step. A random network has only very few short loops and therefore the local network structure around node j is a tree. In the core percolation theory we therefore assume that the adjacent nodes of j are completely independent of each other when j is still unobserved (such an assumption was also exploited in our earlier percolation studies [23–25]). Based on this assumption, the probability of all the adjacent nodes (except i) of j not being unobserved leaves is then written in Eq. (3) as the product of the individual probability $(1-\alpha_0)$ of an adjacent node not being an unobserved leaf. Equation (4) expresses the fact that for node j to be an unoccupied but observed node at the end of the $t = 0$ evolution step, it should not be adjacent to any unobserved leaf node but at least one of its adjacent nodes (except i) should be occupied.

If node j is adjacent to one or more nodes that are occupied at the $t = 0$ evolution step and all its other adjacent nodes (except i) are observed at this evolution step, then at the end of this evolution step j is unoccupied but observed and it is not adjacent to any unobserved node (except i). This then leads to the expression (5) for η_0. Notice such an observed but unoccupied node j will be deleted at the end of the $t = 0$ evolution step. After all such nodes are deleted, some unobserved nodes in the remaining network may become isolated or be connected to only a single node. If this is the case, these isolated or leaf nodes will trigger the next ($t = 1$) evolution step.

Following the same line of theoretical considerations, we obtain the expressions of α_t, β_t, γ_t, and η_t for the tth GLR evolution step ($t \geq 1$) as

$$\alpha_t = \begin{cases} \sum_{d\geq 1} Q(d)(\eta_0)^{d-1} - Q(1) , & (t=1) \\ \sum_{d\geq 1} Q(d)\left[\left(\sum_{l=0}^{t-1}\eta_l\right)^{d-1} - \left(\sum_{l=0}^{t-2}\eta_l\right)^{d-1}\right] , & (t\geq 2) \end{cases} \tag{6}$$

$$\beta_t = \sum_{d\geq 1} Q(d)\left[\left(1-\sum_{l=0}^{t-1}\alpha_l\right)^{d-1} - \left(1-\sum_{l=0}^{t}\alpha_l\right)^{d-1}\right] , \tag{7}$$

$$\gamma_t = \sum_{d\geq 1} Q(d)\left[\left(1-\sum_{l=0}^{t}\alpha_l\right)^{d-1} - \left(1-\sum_{l=0}^{t}(\alpha_l+\beta_l)\right)^{d-1}\right] , \tag{8}$$

$$\eta_t = \sum_{d\geq 1} Q(d)\left[\left(\sum_{l=0}^{t}\beta_l+\gamma_t\right)^{d-1} - (\gamma_t)^{d-1}\right] - \sum_{l=0}^{t-1}\eta_l. \tag{9}$$

Let us denote by γ_{lim} the value of γ_t at the last evolution step $t = t_{lim}$ of the GLR process (notice that the maximal evolution step t_{lim} may approach infinity for a network with $N = \infty$

nodes). Furthermore, we define the accumulated values of α_t, β_t, and η_t as

$$\alpha_{cum} \equiv \sum_{t\geq 0} \alpha_t \ , \quad \beta_{cum} \equiv \sum_{t\geq 0} \beta_t \ , \quad \eta_{cum} \equiv \sum_{t\geq 0} \eta_t .$$

There are the following relationships among α_{cum}, β_{cum}, η_{cum} and γ_{lim}:

$$\alpha_{cum} = \sum_{d\geq 1} Q(d)(\eta_{cum})^{d-1} \ , \tag{10}$$

$$\beta_{cum} = 1 - \sum_{d\geq 1} Q(d)(1-\alpha_{cum})^{d-1} \ , \tag{11}$$

$$\gamma_{lim} = \sum_{d\geq 1} Q(d)\Big[(1-\alpha_{cum})^{d-1} - (1-\alpha_{cum}-\beta_{cum})^{d-1}\Big] , \tag{12}$$

$$\eta_{cum} = \sum_{d\geq 1} Q(d)\Big[(\beta_{cum}+\gamma_{lim})^{d-1} - (\gamma_{lim})^{d-1}\Big]. \tag{13}$$

After all the probability parameters α_t, β_t, γ_t, η_t (for $t = 0, 1, \ldots$) for a node j at the end of a link (i, j) are determined by neglecting the constraint associated with node i, we now ask the following two questions: If the constraint of node i to all its adjacent nodes are considered, then what is the probability n_{core} of i to be unobserved after the whole GLR process? And what is the probability I_t of node i to be occupied at the t-th GLR evolution step? If node i remains to be unobserved during the whole GLR process, it must not be adjacent to any unobserved leaf node nor to any occupied node, and it must have at least two adjacent nodes after the whole GLR process. Therefore we obtain that

$$\begin{aligned} n_{core} &= \sum_{d\geq 2} P(d) \sum_{s=0}^{d-2} C_d^s (\eta_{cum})^s (1-\alpha_{cum}-\beta_{cum}-\eta_{cum})^{d-s} \\ &= \sum_{d\geq 1} P(d)\big[(1-\alpha_{cum}-\beta_{cum})^d - (\eta_{cum})^d \\ &\quad -d(\eta_{cum})^{d-1}(1-\alpha_{cum}-\beta_{cum}-\eta_{cum})\big] , \end{aligned} \tag{14}$$

where $C_d^s \equiv d!/[s!(d-s)!]$ is the binomial coefficient. Notice that if $(\alpha_{cum}+\beta_{cum}+\eta_{cum}) = 1$ then we have $n_{core} = 0$.

It is easy to see that the probability I_0 of a randomly chosen node i to be occupied at the $t = 0$ GLR evolution step is

$$I_0 = 1 - P(1)(1-\frac{\alpha_0}{2}) - \sum_{d\geq 2} P(d)(1-\alpha_0)^d . \tag{15}$$

The coefficient $1/2$ in the second term of the above expression reflects the fact that if node i has only one neighbor j, then i has one-half probability to be occupied if j also has only one neighbor (namely i).

If a randomly chosen node i is not occupied at the $t = 0$ evolution step, then the probability I_1 of it being occupied at the $t = 1$ evolution step is

$$\begin{aligned} I_1 &= \sum_{d\geq 2} P(d)(\eta_0)^d + \sum_{d\geq 2} P(d)\Big[(1-\alpha_0)^d - (1-\alpha_0-\alpha_1)^d \\ &\quad -d\alpha_1\big((\beta_0+\gamma_0)^{d-1} - (\gamma_0)^{d-1}\big)\Big] - \frac{1}{2}\sum_{d\geq 2} P(d)d\alpha_1(\eta_0)^{d-1}. \end{aligned} \tag{16}$$

All the adjacent nodes of i might haven been deleted at the end of the $t = 0$ evolution step. If this is the case node i becomes isolated at the start of the $t = 1$ evolution step, which leads to the first summation of Eq. (16). The second summation of Eq. (16) corresponds to the other situation of node i not being occupied nor being deleted at the $t = 0$ evolution step but it is adjacent to at least one node that becomes an unobserved leaf at the start of the $t = 1$ evolution step. Notice if node i becomes an unobserved leaf node at the start of the $t = 1$ evolution step with its unique neighbor also being such a leaf node, then i has only one-half probability to be occupied at this evolution step. This last situation leads to the third summation term of Eq. (16), which corrects the over-counted probability of occupation in the second summation term.

Following the same line of theoretical considerations, we obtain the probability I_t of a randomly chosen node i changing from being unoccupied to being occupied at the tth GLR evolution step ($t \geq 2$):

$$\begin{aligned} I_t = & \sum_{d\geq 2} P(d) \left[\left(\sum_{l=0}^{t-1} \eta_l \right)^d - \left(\sum_{l=0}^{t-2} \eta_l \right)^d - d\eta_{t-1} \left(\sum_{l=0}^{t-2} \eta_l \right)^{d-1} \right] \\ & + \sum_{d\geq 2} P(d) \left[\left(1 - \sum_{l=0}^{t-1} \alpha_l \right)^d - \left(1 - \sum_{l=0}^{t} \alpha_l \right)^d \right] \\ & - \sum_{d\geq 2} P(d) d\alpha_t \left[\left(\sum_{l=0}^{t-1} \beta_l + \gamma_{t-1} \right)^{d-1} - (\gamma_{t-1})^{d-1} + \left(\sum_{l=0}^{t-2} \eta_l \right)^{d-1} \right] \\ & - \frac{1}{2} \sum_{d\geq 2} P(d) d\alpha_t \left[\left(\sum_{l=0}^{t-1} \eta_l \right)^{d-1} - \left(\sum_{l=0}^{t-2} \eta_l \right)^{d-1} \right]. \end{aligned} \tag{17}$$

The probability w of a randomly chosen node i to be occupied during the GLR process is then

$$w = \sum_{t\geq 0} I_t \tag{18}$$

$$\begin{aligned} = & 1 - P(1)(1 - \alpha_0/2) - \sum_{d\geq 2} P(k) \left[(1 - \alpha_{cum})^d - (\eta_{cum})^d \right] \\ & - \sum_{t\geq 1} \sum_{d\geq 2} P(d) d \left[\eta_t \left(\sum_{l=0}^{t-1} \eta_l \right)^{d-1} + \alpha_t \left(\sum_{l=0}^{t-1} \beta_l + \gamma_{t-1} \right)^{d-1} - \alpha_t (\gamma_{t-1})^{d-1} \right] \\ & - \frac{1}{2} \sum_{t\geq 2} \sum_{d\geq 2} P(d) d\alpha_t \left[\left(\sum_{l=0}^{t-1} \eta_l \right)^{d-1} + \left(\sum_{l=0}^{t-2} \eta_l \right)^{d-1} \right] \\ & - \frac{1}{2} \sum_{d\geq 2} P(d) d\alpha_1 (\eta_0)^{d-1}. \end{aligned} \tag{19}$$

Our core percolation theory can be applied both to single finite random network instances and to random network ensembles at the thermodynamic limit $N \to \infty$. For each t (starting from $t = 0$), we first compute α_t, then use α_t as input to compute β_t, then use α_t and β_t as inputs to compute γ_t, and finally use α_t, β_t and γ_t as inputs to compute η_t. For a finite random network of N nodes, the iteration stops if the evolution step t increases to a value

t_{lim} such that $I_{t_{lim}} < 1/N$. This is because if $NI_t < 1$ the number of newly occupied 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 a sufficiently large evolution step $t = t_{lim}$ until $\alpha_{t_{lim}} \approx 0$.

2.3 Results on Erdös–Rényi Random Networks

We generate an ER random network W of N nodes and $M = (c/2)N$ links by adding links sequentially to an initial network of N isolated nodes. To add a link, we choose two different nodes i and j uniformly at random from the whole node set and set up a link (i, j) between them if this link has not been created before. The mean node degree of the resulting network W is equal to c. When the number N of nodes is sufficiently large the degree distribution $P(d)$ of such a ER network obeys the Poisson distribution [20,21]

$$P(d) = \frac{c^d e^{-c}}{d!} \qquad (d \geq 0). \tag{20}$$

For this network ensemble, the predicted results of n_{core} and w by our core percolation theory are in perfect agreement with simulation results (see Fig. 3). Especially, at the thermodynamic limit $N \to \infty$, the theory predicts a continuous core percolation phase transition at $c \approx 2.4102$, which is slightly lower than the core percolation phase transition point of $c \approx 2.7183$ caused by the conventional leaf-removal process [17]. Before the GLR-induced core percolation transition occurs, the occupation fraction w obtained by Eq. (19) is equal to the ensemble-averaged MDS size (relative to N), but it is only a lower bound to this size when an extensive core emerges in the random network ($n_{core} > 0$).

2.4 Results on Scale-Free Random Networks Generated Through the Static Model

Now let us consider GLR-induced core percolation on more heterogeneous random networks. We generate a scale-free network W of N nodes and $M = (c/2)N$ links according to the static model [22]. Each node $i \in \{1, 2, \ldots, N\}$ is first assigned a fitness value $\theta_i = i^{-\xi}/(\sum_{j=1}^{N} j^{-\xi})$, where $0 \leq \xi < 1$ is a control parameter. Then we add links between pairs of these N nodes in a sequential manner. To create a link, two nodes i and j are chosen independently from the set of N nodes, and the probability that i and j being chosen is equal to $\theta_i \theta_j$; if nodes i and j are different and the link (i, j) has not been created before, this link is added to network. The final network W has a power-law degree distribution $P(d) \propto d^{-\lambda}$ for $d \gg 1$, with degree decay exponent $\lambda = 1 + 1/\xi$. In the thermodynamic limit $N \to \infty$, an explicit expression for $P(d)$ is obtained as [26]

$$P(d) = \frac{[c(1-\xi)]^d}{d!\xi} \int_1^\infty \mathrm{d}x e^{-c(1-\xi)x} x^{k-1-1/\xi} \qquad (d \geq 0). \tag{21}$$

For $N = \infty$, a continuous core percolation phase transition is observed in such a SF network, and this transition occurs at more and more larger value of the mean node degree c as the decay exponent λ decreases (see Fig. 4 for $\lambda = 3.5, 3.0, 8/3 \approx 2.667$, and 2.5 and Fig. 6 for $2 < \lambda \leq 6$). When the decay exponent λ is less then 3.0, theoretical predictions obtained at $N = \infty$ are quantitatively different from theoretical and simulation results obtained on finite (e.g., $N = 10^6$) network instances, with the deviations become more pronounced as λ is closer to 2.0. Such a finite-size effect is mainly caused by the natural cutoff of maximum node degree in finite networks (it was also observed in our earlier work [23] on another type of percolation transitions). We emphasize that for a give finite value of N, the results of the

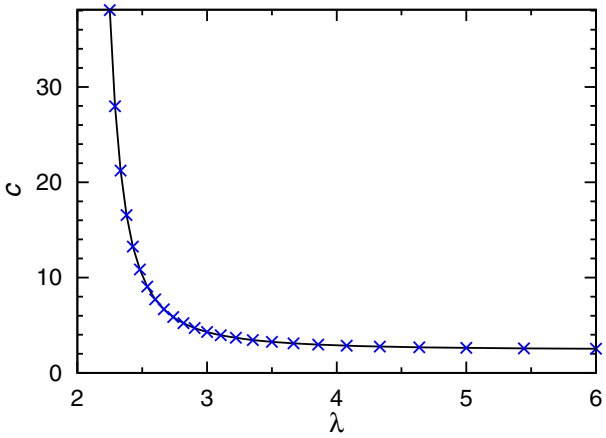

Fig. 6 Core percolation phase transition in infinite ($N = \infty$) random scale-free networks generated through the static model [22]. *Horizontal axis* is the degree decay exponent λ, while vertical axis is the value of the mean node degree c at the phase transition point. *Cross symbols* are predictions of the core-percolation theory, while the *solid line* is just a guide for the eye

core percolation theory agree with the simulation results of the actual GLR process very well, especially if we average the theoretical and simulation results over many network instances to reduce fluctuations.

For random SF networks generated through the static model with $N = \infty$ nodes, the core percolation transition value of mean node degree c is very sensitive to the decay exponent λ in the region of $2 < \lambda < 2.5$, and it diverges as λ approaches 2.0 from above (Fig. 6). At the other limit of $\lambda \to \infty$, the mean node degree at the phase transition approaches the value of $c \approx 2.4102$, which is just the core percolation phase transition point of an infinite ER random network.

2.5 Results on Pure Scale-Free Random Networks

When $N = \infty$, a pure scale-free random network has the following degree distribution

$$P(d) = \frac{1}{\sum_{k=1}^{\infty} k^{-\lambda}} d^{-\lambda} \qquad (d \geq 1) \ , \tag{22}$$

with $\lambda > 2$ to ensure a finite value for the mean node degree c. For such a random SF network our core percolation theory predicts that $n_{core} = 0$, namely there is no core percolation transition and the GLR process will construct a MDS for the whole network. The fraction w of occupied nodes (i.e., the size of a MDS relative to the node number N) decreases with the decreasing of the degree decay exponent λ (see Fig. 5), and it approaches zero as λ approaches 2.0 from above.

We also generate a set of pure SF random networks of finite size N following the same procedure as mentioned in [27] (see also the supplementary information of [23]). The minimum node degree of such a SF network is $d_{min} = 1$, while the maximum node degree is $d_{max} \approx N^{1/(\lambda-1)}$ [27]. When we apply both the GLR process and the core percolation theory on these finite network instances, we find the simulation results on the fraction n_{core} of nodes in the core and the fraction w of occupied nodes are in perfect agreement with the corresponding theoretical results (see Fig. 5). All these finite SF networks contain no core ($n_{core} = 0$), and the MDS relative size w is an increasing function of λ.

Figure 5 also demonstrates strong finite-size effect for pure SF random networks of $\lambda < 3.0$. This finite-size effect is again mainly caused by the cutoff of the maximum node degree of finite networks, which makes the mean node degree of a finite network be smaller than that of an infinite network. For example, at $\lambda = 2.1$ the mean node degree of an infinite network is $c \approx 61.49$, while that of a finite network of size $N = 10^6$ is reduced to $c \approx 5.134$.

3 Hybrid Local Algorithm

There is a very simple greedy algorithm in the literature to solve the MDS problem approximately, which is based on the concept of node impact [1,7,16]). 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 already adjacent to one or more occupied nodes. Starting from an input network W with all the nodes unobserved, the 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 unoccupied nodes is updated and the greedy occupying process is repeated until all the nodes are observed. This pure greedy algorithm is very easy to implement and very fast, but we find that it usually fails to reach a true MDS even when the input network contains no core.

Here we introduce an improved local algorithm by combining the GLR process with the impact-based greedy process. We call this new algorithm the `GLR-Impact` hybrid algorithm. Given an input network W with all the nodes unobserved, we first carry out the GLR process to simplify W as far as possible. If all the nodes are observed during this initial GLR process, a MDS of network W is then constructed. For the nontrivial case of some nodes being left unobserved after this GLR process, we first occupy a randomly chosen node from the subset of highest-impact nodes and then perform the GLR process again to further simplify the network as far as possible. We keep repeating such a occupying-followed-by-GLR process until there is no unobserved node left in the network.

The `GLR-Impact` hybrid algorithm is also very easy to implement and very fast. Its performance is demonstrated in Fig. 7 for single ER networks and regular random (RR) networks. (All the nodes of a RR network have the same integer degree c but the network is otherwise completely random [23]). This hybrid local algorithm outperforms the pure greedy algorithm considerably for $c \leq 10$, but it is still inferior to the belief-propagation-guided decimation (BPD) algorithm of Sect. 5.

We also test the performance of the hybrid algorithm on single SF random networks generated through the static model [22] (see Fig. 8). The `GLR-Impact` algorithm still outperforms the pure greedy algorithm on these heterogeneous networks, and its performance approaches that of the BPD algorithm as the network becomes more and more heterogeneous (i.e., as the decay exponent λ approaches 2.0 from above).

Real-world networks are often very heterogenous, with a small fraction of highly connected nodes [21]. As a test of the algorithms introduced in this work, we apply the GLR process, the pure greedy algorithm, the hybrid algorithm, and the BPD algorithm to a set of twelve real-world networks. Among these network instances, five are infrastructure networks: European express road network (RoadEU [28]), road network of Texas (RoadTX [29]), power grid of western US states (Grid [30]), and two Internet networks at the autonomous systems level (IntNet1 and IntNet2 [31]); three are information networks: Google webpage network (WebPage [29]), European email network (Email [32]), and research citation network (Cita-

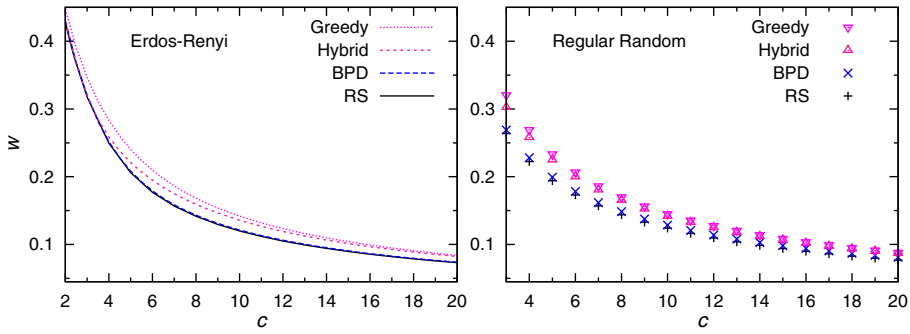

Fig. 7 Constructing dominating sets for Erdös–Rényi networks (*left panel*) and regular random networks (*right panel*). The relative sizes w of dominating sets obtained by a single running of the pure greedy, the hybrid, and the BPD algorithm with $x = 10$ on 96 ER or RR network instances of $N = 10^5$ 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 replica-symmetric mean field theory are also shown (RS)

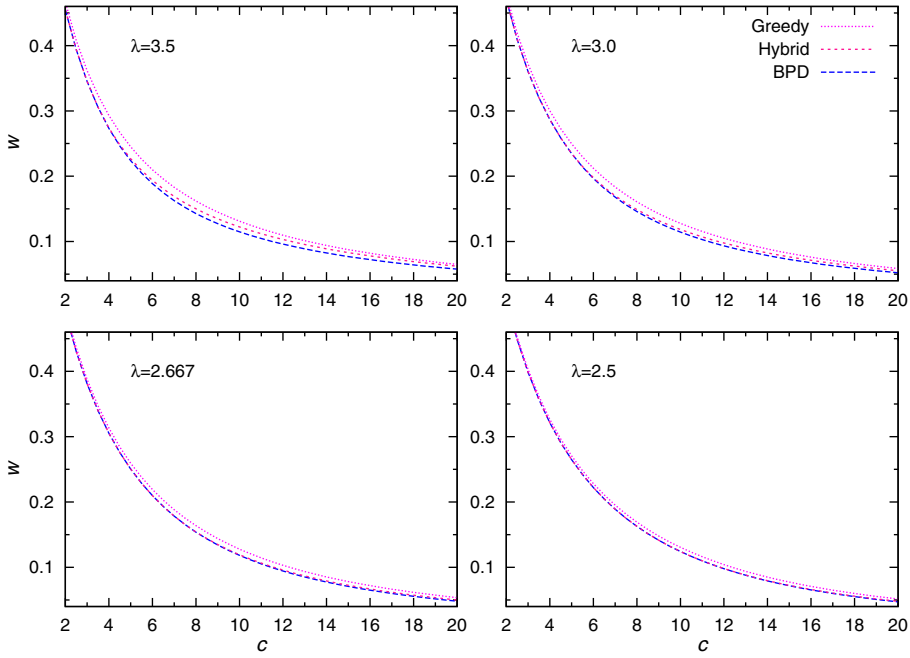

Fig. 8 Constructing dominating sets for scale-free random networks generated through the static model [22]. The relative sizes w of dominating sets obtained by a single running of the pure greedy, the hybrid, and the BPD algorithm with $x = 10$ on 96 SF network instances of $N = 10^5$ and (mean) degree c are compared (fluctuations are of order 10^{-4} and are not shown). The degree decay exponent is $\lambda = 3.5$, 3.0, 2.667, and 2.5, respectively

tion [31]); three are social contact networks: collaboration network of condensed-matter authors (Author [32]), peer-to-peer interaction network (P2P [33]),and on-line friendship network (Friend [34]); the remaining one is the biological network of protein-protein interactions (PPI [35]).

Table 1 Results on twelve real-world network instances

Network	N	M	d_{max}	Core	Greedy	Hybrid	BPD
RoadEU	1177	1417	10	306	428	389	387
PPI	2361	6646	64	17	550	539	539
Grid	4941	6594	19	603	1564	1485	1481
IntNet1	6474	12,572	1458	8	660	656	656
Author	23,133	93,439	279	9052	3686	3612	3604
Citation	34,546	420,877	846	11,178	3335	3168	3095
P2P	62,586	147,892	95	35	12,710	12,582	12,582
Friend	196,591	950,327	14,730	6097	42,536	41,633	41,672
Email	265,214	364,481	7636	470	18,183	18,181	18,181
WebPage	875,713	4,322,051	6332	162,439	81,288	79,928	80,769
RoadTX	1,379,917	1,921,660	12	560,582	477,729	437,503	425,774
IntNet2	1,696,415	11,095,298	35,455	211,244	187,592	183,516	183,248

N and M are, respectively, the total number of nodes and links in the network; d_{max} is the maximum node degree of the network; the column marked by 'Core' records the number of nodes that are left unobserved after the GLR process; the columns marked by 'Greedy', 'Hybrid', and 'BPD' record the sizes of the dominating sets constructed by a single running of the pure greedy, the hybrid, and the BPD algorithm, respectively

The numerical results are summarized in Table 1. The GLR process is able to simplify these networks considerably. After GLR, the remaining number of unobserved nodes is often much smaller than the total number N of nodes in the original network. The BPD algorithm performs slightly better than the `GLR-Impact` hybrid algorithm, and both BPD and the hybrid algorithm outperform the pure greedy algorithm in all the twelve network instances.

4 Spin Glass Model and Replica-Symmetric Mean Field Theory

If a given network instance W contains an extensive core, the GLR process can only give a lower bound to the MDS size. We now discuss the issue of estimating the MDS size by way of a mean field theory. We introduce a partition function Z as

$$Z = \sum_{\underline{c}} \prod_{i \in W} \left\{ e^{-x c_i} \left[1 - (1 - c_i) \prod_{j \in \partial i} (1 - c_j) \right] \right\} , \tag{23}$$

where $\underline{c} \equiv (c_1, c_2, \ldots, c_N)$ denotes one of the 2^N possible occupation configurations, $x > 0$ is a re-weighting parameter, and ∂i denotes node i's set of adjacent nodes. The constraint of each node i leads to a multiplication term $[1 - (1 - c_i) \prod_{j \in \partial i} (1 - c_j)]$, 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.

4.1 Replica-Symmetric Mean Field Theory

We solve the spin glass model (23) by a RS mean field theory, which can be understood from the angle of Bethe-Peierls approximation [36] or derived alternatively through partition function expansion [37,38]. The marginal probability q_i^c of node i's occupation state being c ($\in \{0, 1\}$) is expressed as

$$q_i^c = \frac{e^{-xc}\prod_{j\in\partial i}\sum_{c_j} q_{j\to i}^{(c_j,c)} - \delta_0^c \prod_{j\in\partial i} q_{j\to i}^{(0,0)}}{\sum_{c_i} e^{-xc_i}\prod_{j\in\partial i}\sum_{c_j} q_{j\to i}^{(c_j,c_i)} - \prod_{j\in\partial i} q_{j\to i}^{(0,0)}} , \tag{24}$$

where the Kronecker symbol $\delta_m^n = 1$ if $m = n$ and $\delta_m^n = 0$ if otherwise. The quantity $q_{j\to i}^{(c_j,c_i)}$ 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 (BP) equation:

$$q_{j\to i}^{(c_j,c_i)} = \frac{e^{-xc_j}\prod_{k\in\partial j\backslash i}\sum_{c_k} q_{k\to j}^{(c_k,c_j)} - \delta_0^{c_i+c_j}\prod_{k\in\partial j\backslash i} q_{k\to j}^{(0,0)}}{\sum_{c_i',c_j'} e^{-xc_j'}\prod_{k\in\partial j\backslash i}\sum_{c_k'} q_{k\to j}^{(c_k',c_j')} - \prod_{k\in\partial j\backslash i} q_{k\to j}^{(0,0)}} , \tag{25}$$

where $\partial j\backslash i$ denotes the subset obtained by deleting node i from set ∂j.

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

$$F = \sum_{i\in W} f_i - \sum_{(i,j)\in W} f_{(i,j)} , \tag{26}$$

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:

$$f_i = -\frac{1}{x}\ln\Big[\sum_{c_i} e^{-xc_i}\prod_{j\in\partial i}\sum_{c_j} q_{j\to i}^{(c_j,c_i)} - \prod_{j\in\partial i} q_{j\to i}^{(0,0)}\Big] , \tag{27}$$

$$f_{(i,j)} = -\frac{1}{x}\ln\Big[\sum_{c_i,c_j} q_{i\to j}^{(c_i,c_j)} q_{j\to i}^{(c_j,c_i)}\Big]. \tag{28}$$

From Eqs. (26) and (24) we can compute the free energy density $f \equiv F/N$ 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 $s = (w - f)x$.

4.2 Belief-Propagation Iterations

According to Eq. (25) each probability distribution $q_{j\to i}^{(c_j,c_i)}$ has the property that $q_{j\to i}^{(1,1)} = q_{j\to i}^{(1,0)}$. Therefore in the numerical computations $q_{j\to i}^{(c_j,c_i)}$ can be represented by three non-negative real numbers $q_{j\to i}^{(0,0)}$, $q_{j\to i}^{(0,1)}$, and $q_{j\to i}^{(1,0)}$, which satisfy in addition the normalization condition

$$q_{j\to i}^{(0,0)} + q_{j\to i}^{(0,1)} + 2q_{j\to i}^{(1,0)} = 1. \tag{29}$$

We initialize $q_{j\to i}^{(c_j,c_i)}$ and $q_{i\to j}^{(c_i,c_j)}$ for each link (i, j) of the network between two nodes i and j, for example setting $q_{j\to i}^{(0,0)} = q_{j\to i}^{(0,1)} = q_{j\to i}^{(1,0)} = 1/4$. 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. (25) 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 $q_{j\to i}^{(c_j,c_i)}$ to all its adjacent nodes $i \in \partial j$ are updated according to Eq. (25). The difference $\Delta_{j\to i}(t)$ between an updated message $q_{j\to i}(t)$ at the t-th BP step and the old message $q_{j\to i}(t-1)$ at the $(t-1)$-th BP step is defined as

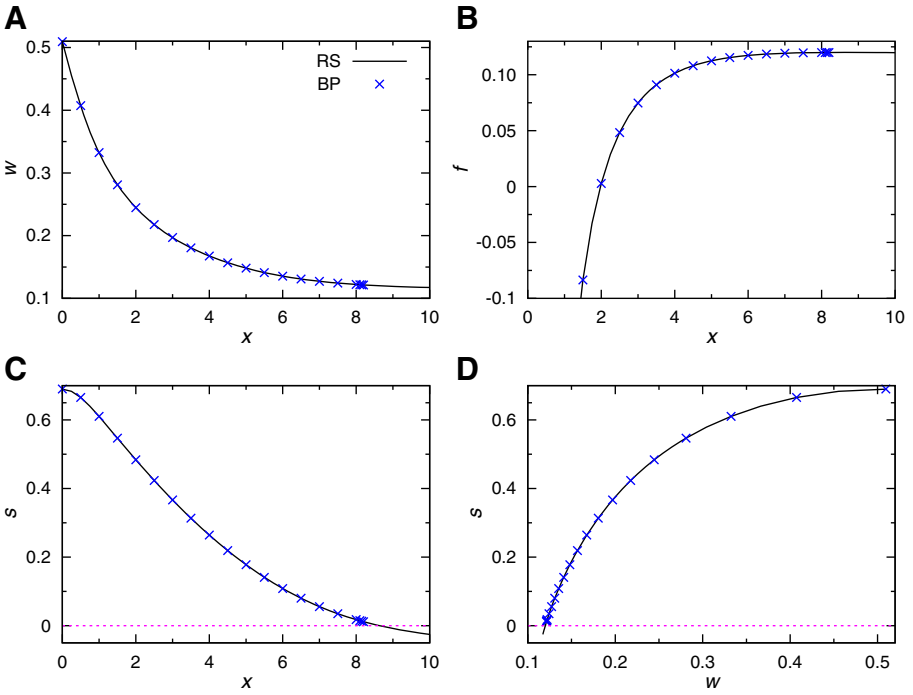

Fig. 9 Replica-symmetric (RS) mean field theory and belief-propagation (BP) results for ER random networks of mean degree $c = 10$. The RS results are obtained by population dynamics simulations, while the BP results are obtained on a single ER network instance of $N = 10^6$ nodes. The BP iteration converges to a fixed point only for $x < 8.22$. **a** Occupation fraction w. **b** Free energy density f. **c** Entropy density s. **d** Entropy density s as a function of occupation fraction w

$$\Delta_{j\to i}(t) \equiv |q^{(0,0)}_{j\to i}(t) - q^{(0,0)}_{j\to i}(t-1)| + |q^{(0,1)}_{j\to i}(t) - q^{(0,1)}_{j\to i}(t-1)| + 2|q^{(1,0)}_{j\to i}(t) - q^{(1,0)}_{j\to i}(t-1)|. \tag{30}$$

If the maximal value among the set of $2M$ difference values $\{\Delta_{j\to i}(t)\}$ 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. (25) 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. 9 the results obtained on a single ER random network of size $N = 10^6$ and mean degree $c = 10$.

For ER networks with mean degree $c > 2.41$ and regular random 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. Systematic theoretical investigations on this spin glass phase will be reported in another publication.

4.3 Ensemble-Averaged Properties

A random network ensemble is characterized by a degree distribution $P(d)$. We perform population dynamics simulations using Eqs. (25), (24) and (26) to obtain ensemble-averaged

results. First, we create a long array A of $\mathcal{N}$ (e.g., 10^5) elements to store a set of messages, each of which represents a probability distribution $q_{j\to i}^{(c_j,c_i)}$ in the form of a three-dimensional vector satisfying Eq. (29): $q_{j\to i} \equiv (q_{j\to i}^{(0,0)}, q_{j\to i}^{(0,1)}, q_{j\to i}^{(1,0)})$. 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 $q_{k\to j}$ from array A uniformly at random, and then use these $(d-1)$ elements as input messages to Eq. (25) to compute a new message $q_{j\to i}$; (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

$$f = \overline{f_i} - \frac{c}{2}\overline{f_{(i,j)}}\,, \tag{31}$$

where $\overline{f_i}$ is the average of the free energy node contribution f_i over all the nodes, and $\overline{f_{(i,j)}}$ is 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 $\overline{f_i}$ and $\overline{f_{(i,j)}}$. The procedure of obtaining a sample of f_i is the same as that of updating an element of the message A, 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. (28) by picking two messages $q_{j\to i}$ and $q_{i\to j}$ uniformly at random from the message array A.

For ER random networks with mean degree $c = 10$, we compare in Fig. 9 the results obtained by this RS population dynamics with the results obtained by BP iteration on a single network instance. The ensemble-averaged results are 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. 9d). 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 set with relative size below w_0. We therefore take the value w_0 as the ensemble-averaged MDS relative size. For ER networks of $c = 10$, we obtain from Fig. 9 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_0 from above. For these latter cases, we simply take w_0 as the ensemble-averaged MDS relative size.

The ensemble-averaged results on the MDS sizes of ER and RR networks are shown in Fig. 7. For ER networks with mean node degree $c < 2.41$ (before the core-percolation transition), the RS mean field results coincide with the results predicted by the core percolation theory. When the random network contains an extensive core, the results obtained by the pure greedy algorithm and the `GLR-Impact` algorithm are higher than the RS mean field predictions, but the results obtained by the BPD algorithm of the next section are very close to the RS mean field predictions.

5 Belief-Propagation-Guided Decimation Algorithm

For a given network W, the RS mean field theory gives an estimate for the occupation probability q_i^{+1} of each node i, see Eq. (24). Such information is exploited in a BPD algorithm

to construct a near-optimal dominating set. (Such an algorithm and its extensions have already been successfully applied to many other combinatorial optimization problems, e.g., the K-satisfiability problem [39,40] and the vertex-cover problem [5]). 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 $q_{j\to i}^{(c_j,c_i)}$ on the link (j,i) between j and node i is updated according to Eq. (25). 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 neighbors. For such a node j, the output message $q_{j\to i}^{(c_j,c_i)}$ on the link (j,i) is then updated according to the following equation:

$$q_{j\to i}^{(c_j,c_i)} = \frac{e^{-xc_j}\prod_{k\in\partial j\backslash i}\sum_{c_k} q_{k\to j}^{(c_k,c_j)}}{\sum_{c'_j,c'_i} e^{-xc'_j}\prod_{k\in\partial j\backslash i}\sum_{c'_k} q_{k\to j}^{(c'_k,c'_j)}}. \tag{32}$$

Similar to Eq. (32), the marginal probability distribution $q_i^{c_i}$ for an observed empty node i is evaluated according to

$$q_i^{c_i} = \frac{e^{-xc_i}\prod_{j\in\partial i}\sum_{c_j} q_{j\to i}^{(c_j,c_i)}}{\sum_{c'_i} e^{-xc'_i}\prod_{j\in\partial i}\sum_{c'_j} q_{j\to i}^{(c'_j,c'_i)}}. \tag{33}$$

It is easy to verify from Eq. (32) that $q_{j\to i}^{(0,0)} = q_{j\to i}^{(0,1)}$ and $q_{j\to i}^{(1,0)} = q_{j\to i}^{(1,1)}$. Notice that if all the nodes in the set $\partial j\backslash i$ are observed, then we derive from Eq. (32) that $q_{j\to i}^{(0,0)} = q_{j\to i}^{(1,0)} = q_{j\to i}^{(0,1)} = q_{j\to i}^{(1,1)} = 1/4$. Because of this property, we need only to consider the links between unobserved nodes and the links between unobserved and observed nodes. All the other links (which are between observed nodes) do not need to be considered in the BP iteration Eqs. (25) and (32).

We implement the BPD algorithm as follows:

(0) Input the network W, set all the nodes to be empty and unobserved and set all the probability distributions $q_{j\to i}^{(c_j,c_i)}$ 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 compute the occupation probability q_i^{+1} of each node i using Eq. (24).

(1) Then occupy a small fraction r (e.g., $r = 0.01$) of the unoccupied nodes that having the highest estimated occupation probabilities.

(2) Then simplify network W by first deleting all the links between observed nodes, and then deleting all the isolated 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 node i is updated either according to Eq. (24) or according to Eq. (33), depending on whether i is unobserved or observed. We then repeat operations (1)–(3) 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.

The results of the BPD algorithm for random networks and for real-world networks are compared with the results obtained by the local heuristic algorithms in Figs. 7, 8, and Table 1. For ER and RR random networks, the BPD algorithm considerably beats both the pure greedy algorithm and the `GLR-Impact` hybrid algorithm; for very heterogeneous (e.g., scale-free) networks, the BPD algorithm only slightly outperforms the `GLR-Impact` algorithm.

6 Discussions

In this work, we proposed two heuristic algorithms (a `GLR-Impact` local algorithm and a BPD message-passing algorithm) and presented a core percolation theory and a replica-symmetric mean field theory for solving the network dominating set problem algorithmically and theoretically. We found that the GLR process may lead to a core percolation transition in the network (see Figs. 3 and 4). Our numerical results shown in Figs. 7, 8 and Table 1 suggested that the `GLR-Impact` algorithm and the BPD algorithm can construct near-optimal dominating sets for random networks and real-world networks.

There are many theoretical issues remaining to be investigated. An easy extension of the core percolation theory is to consider GLR with a subset of initially occupied nodes. By optimizing this initial subset (e.g., following the methods of [41–43]), we may reach an improved lower-bound to the MDS size. Core percolation on degree-correlated random networks [44] and in the more general lattice glass problem [3] are also very interesting. When the random network has an extensive core, we observed that the belief-propagation Eq. (25) fails to converge at large values of the re-weighting parameter x (see Fig. 9), 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 [40,45,46], which may in addition 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 [24,25].

The methods of this work can be readily extended to the MDS problem of directed networks. Our theoretical and algorithmic results on the directed MDS problem will soon be reported in an accompanying paper [47]. A more challenging problem is the connected dominating set problem [48] which has the additional constraint that the nodes in the dominating set should induce a connected subnetwork. Our present work may stimulate further theoretical studies on this hard problem.

Acknowledgments Part of this work was done when H.-J. Zhou was participating in the "Collective Dynamics in Information Systems 2014" Program of the Kavli Institute for Theoretical Physics China (KITPC). H.-J. Zhou thanks Chuang Wang for a helpful discussion, and Alfredo Braunstein, Yang-Yu Liu, Federico Ricci-Tersenghi, and Yi-Fan Sun for helpful comments on the manuscript; J.-H. Zhao and H.-J. Zhou thank Prof. Zhong-Can Ou-Yang for support. Research partially supported by the National Basic Research Program of China (grant number 2013CB932804) and by the National Natural Science Foundation of China (grand numbers 11121403 and 11225526).

References

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 **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. Echenique, P., Gómez-Gardeñes, J., Moreno, Y., Vázquez, A.: Distance-*d* covering problems in scale-free networks with degree correlations. Phys. Rev. E **71**, 035102(R) (2005)
7. Takaguchi, T., Hasegawa, T., Yoshida, Y.: Suppressing epidemics on networks by exploiting observer nodes. Phys. Rev. E **90**, 012807 (2014)
8. Liu, Y.-Y., Slotine, J.-J., Barabási, A.-L.: Controllability of complex networks. Nature **473**, 167–173 (2011)
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. Telecomm. Syst. **18**, 13–36 (2001)
11. Bramoulle, Y., Kranton, R.: Public goods in networks. J. Econom. Theor. **135**, 478–494 (2007)
12. Shen, C., Li, T.: Multi-document summarization via the minimum dominating set. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010, Beijing), pp. 984–992 (Association for Computational Linguistics, 2010)
13. Lund, C., Yannakakis, M.: On the hardness of approximating minimization problems. J. ACM **41**, 960–981 (1994)
14. Raz, R., Safra, S.: A sub-constant error-probability low-degree test, and a sub-constant error-probability pcp characterization of np. In: Proceedings of the 29th Annual ACM Symposium on Theory of Computing, pp. 475–484 (ACM, New York, 1997)
15. Hedar, A.-R., Ismail, R.: Simulated annealing with stochastic local search for minimum dominating set problem. Int. J. Mach. Learn. Cybernet. **3**, 97–109 (2012)
16. Molnár Jr, F., Sreenivasan, S., Szymanski, B.K., Korniss, K.: Minimum dominating sets in scale-free network ensembles. Sci. Rep. **3**, 1736 (2013)
17. Bauer, M., Golinelli, O.: Core percolation in random graphs: a critical phenomena analysis. Eur. Phys. J. B **24**, 339–352 (2001)
18. Lucibello, C., Ricci-Tersenghi, F.: The statistical mechanics of random set packing and a generalization of the Karp–Sipser algorithm. Int. J. Stat. Mech. **2014**, 136829 (2014)
19. Takabe, S., Hukushima, K.: Minimum vertex cover problems on random hypergraphs: replica symmetric solution and a leaf removal algorithm. Phys. Rev. E **89**, 062139 (2014)
20. He, D.-R., Liu, Z.-H., Wang, B.-H.: Complex Systems and Complex Networks. Higher Education Press, Beijing (2009)
21. Albert, R., Barabási, A.-L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**, 47–97 (2002)
22. Goh, K.-I., Kahng, B., Kim, D.: Universal behavior of load distribution in scale-free networks. Phys. Rev. Lett. **87**, 278701 (2001)
23. Zhao, J.-H., Zhou, H.-J., Liu, Y.-Y.: Inducing effect on the percolation transition in complex networks. Nat. Commun. **4**, 2412 (2013)
24. Zhou, H.J.: Long-range frustration in a spin-glass model of the vertex-cover problem. Phys. Rev. Lett. **94**, 217203 (2005)
25. Zhou, H.-J.: Erratum: long-range frustration in a spin-glass model of the vertex-cover problem [phys. rev. lett. 94, 217203 (2005)]. Phys. Rev. Lett. **109**, 199901 (2012)
26. Catanzaro, M., Pastor-Satorras, R.: Analytic solution of a static scale-free network model. Eur. Phys. J. B **44**, 241–248 (2005)
27. Zhou, H.J., Lipowsky, R.: Dynamic pattern evolution on scale-free networks. Proc. Natl. Acad. Sci. USA **102**, 10052–10057 (2005)
28. Šubelj, L., Bajec, M.: Robust network community detection using balanced propagation. Eur. Phys. J. B **81**, 353–362 (2011)
29. 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)
30. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' netowrks. Nature **393**, 440–442 (1998)
31. 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, pp. 177–187 (ACM, New York, 2005)
32. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graph evolution: densification and shrinking diameters. ACM Trans. Knowl. Discov. Data **1**, 2 (2007)
33. 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)

34. Cho, E., Myers, S. A., Leskovec, J.: Friendship and mobility: user movement in localation-based social networks. In: ACM SIGKDD International Conference o Knowledge Discovery and Data Mining, pp. 1082–1090 (San Diego, CA, USA, 2011)
35. Bu, D., et al.: Topological structure analysis of the protein–protein interaction network in budding yeast. Nucleic Acids Res. **31**, 2443–2450 (2003)
36. Mézard, M., Montanari, A.: Information, Physics, and Computation. Oxford Univ. Press, New York (2009)
37. 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 **44**, 425001 (2011)
38. 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)
39. Mézard, M., Parisi, G., Zecchina, R.: Analytic and algorithmic solution of random satisfiability problems. Science **297**, 812–815 (2002)
40. 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)
41. Altarelli, F., Braunstein, A., Dall'Asta, L., Zecchina, R.: Large deviations of cascade processes on graphs. Phys. Rev. E **87**, 062115 (2013)
42. Altarelli, F., Braunstein, A., Dall'Asta, L., Zecchina, R.: Optimizing spread dynamics on graphs by message passing. J. Stat. Mech. (2013). doi:10.1088/1742-5468/2013/09/P09011
43. Guggiola, A., Semerjian, G.: Minimal contagious sets in random regular graphs. J. Stat. Phys. **158**, 300–358 (2015)
44. Hasegawa, T., Takaguchi, T., Masuda, N.: Observability transitions in correlated networks. Phys. Rev. E **88**, 042809 (2013)
45. Mézard, M., Parisi, G.: The bethe lattice spin glass revisited. Eur. Phys. J. B **20**, 217–233 (2001)
46. Mézard, M., Montanari, A.: Reconstruction on trees and spin glass transition. J. Stat. Phys. **124**, 1317–1350 (2006)
47. Habibulla, Y., Zhao, J.-H., Zhou, H.-J.: The directed dominating set problem: generalized leaf removal and belief propagation. (2015, in preparation)
48. Du, D.-Z., Wan, P.-J.: Connected Dominating Set: Theory and Applications. Springer, New York (2013)