Statistical mechanics of the minimum vertex cover problem in stochastic block models

Masato Suzuki and Yoshiyuki Kabashima

Department of Mathematical and Computing Science, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

(Received 20 August 2019; published 2 December 2019)

The minimum vertex cover (Min-VC) problem is a well-known NP-hard problem. Earlier studies illustrate that the problem defined over the Erdös-Rényi random graph with a mean degree c exhibits computational difficulty in searching the Min-VC set above a critical point $c = e = 2.718\ldots$. Here, we address how this difficulty is influenced by the mesoscopic structures of graphs. For this, we evaluate the critical condition of difficulty for the stochastic block model. We perform a detailed examination of the specific cases of two equal-size communities characterized by in and out degrees, which are denoted by $c_{\rm in}$ and $c_{\rm out}$, respectively. Our analysis based on the cavity method indicates that the solution search once becomes difficult when $c_{\rm in} + c_{\rm out}$ exceeds e from below, but becomes easy again when $c_{\rm out}$ is sufficiently larger than $c_{\rm in}$ in the region $c_{\rm out} > e$. Experiments based on various search algorithms support the theoretical prediction.

DOI: 10.1103/PhysRevE.100.062101

I. INTRODUCTION

The minimum vertex cover problem (Min-VCP) is a well-known combinatorial optimization problem. Given a graph, the task of Min-VCP is to obtain a vertex set of minimum size, such that every edge in the graph is connected to at least one vertex in the set. Despite the simplicity of its expression, this problem is related to many real-world problems including monitoring Internet traffic [1], preventing denial-of-service attacks [2], creating immunization strategies in networks [3], solving the placement problem [4], and dealing with data aggregation [5]. In addition, Min-VCP is considered a representative computationally difficult problem that belongs to the class NP-hard. This means that, in the worst case, all known algorithms require an exponentially long time with respect to the number of vertices not only to find a solution but also to verify its validity.

Conventional theory in computer science determines the computational difficulty mainly in the worst case. However, characterizing the difficulty for a typical case is also essential for various practical situations. Recent studies in statistical mechanics carried out the typical case analysis of Min-VCP for the Erdös-Rényi (ER) random graph ensemble [6–8]. Based on the analysis using the replica and cavity methods, these studies reported that searching for the correct solution typically becomes computationally difficult when the mean degree c exceeds $e = 2.718\ldots$, while the solution can be obtained easily for $c < e$. The ER model (ERM) is simple and suitable as the initial step to analyze a typical case. Nonetheless, extending the result to more general ensembles is required since the graphs in real world exhibit several mesoscopic structures, which are absent in ERM.

Based on the above perspective, this study addresses how the critical condition of computational difficulty is influenced when nontrivial structures are introduced into graphs. As a simple but nontrivial example, we examine the Min-VCP defined over the stochastic block model (SBM), which is a generalization of ERM often employed for the community detection problem [9]. Particularly focusing on the equal size two community cases, which can be characterized by two parameters, the mean in and out degrees, denoted by $c_{\rm in}$ and $c_{\rm out}$, respectively, we will show that the solution search once becomes difficult when $c_{\rm in} + c_{\rm out}$ exceeds e from below, but becomes easy again when $c_{\rm out}$ is sufficiently larger than $c_{\rm in}$ in the region $c_{\rm out} > e$. This indicates that mesoscopic structures such as "communities" in graphs strongly influence the computational difficulty.

The remainder of this paper is organized as follows. In the next section, we formulate the problem. In section III, we analyze the Min-VCP defined over SBM using the cavity method, which yields the critical condition of the computational difficulty as a function of $c_{\rm in}$ and $c_{\rm out}$ in typical cases. The theoretical prediction is then tested by numerical experiments in section IV. The final section is devoted to the summary of our results.

II. DEFINITIONS

A. Minimum vertex cover problem

We consider a graph G with a set of N vertices $V = \{1, 2, ..., N\}$ and a set of edges E. A vertex cover (VC) $V_{\rm VC}$ of the graph G is a subset of the vertices V of the graph G which includes at least one vertex of every edge in E. Then, Min-VCP is the problem of finding $V_{\rm VC}$ such that the number of elements in the VC is minimized.

Min-VCP is a well-known NP-hard problem. However, if the graph is bipartite, this problem is equivalent to the maximum matching problem, which belongs to the class P, according to König's theorem [10]. This indicates that the computational difficulty can vary depending on the graph structure of the problem. Our main purpose is to clarify how this change is characterized for the typical case.

B. Random graph models

ERM is a random graph model constructed by connecting every pair of vertices $(i, j) \in V \times V$ independently with

probability $p \in [0, 1]$. The degree distribution of ERM asymptotically converges to the Poisson distribution $P(d) = e^{-c}c^k/k!$, where $c := Np$ is the mean degree and p is set such that c is $O(1)$. Therefore, ERM is also called a Poisson-distributed graph.

On the other hand, SBM is a random graph model generalized from ERM to describe a community (or cluster) structure. For generating a sample of SBM, a hidden label $z_i \in \{1, \ldots, K\}$ $(i = 1, 2, \ldots, N)$ is first assigned to each of the N vertices with a categorical distribution $\pi = \{\pi_1, \pi_2 ..., \pi_K\}$. Given a set of the community labels $\{z_i\}$, every pair of vertices $(i, j) \in V \times V$ is connected independently with probability Θ_{z_i,z_j}, where $\Theta \in [0, 1]^{K\times K}$ is an affinity matrix with the property $\Theta_{z,z'} = \Theta_{z',z}$. Then, in the large system limit $N \to \infty$, which we will focus on below, the probability that a node of community z has d connections to community z' converges to

$$P_{z,z'}(d) := e^{-c_{z,z'}} \frac{c_{z,z'}^d}{d!}, \tag{1}$$

$$c_{z,z'} := \lim_{N\to\infty} N\pi_{z'}\Theta_{z,z'}, \tag{2}$$

where $c_{z,z'}$ is the mean degree from community z to community z', and is assumed to be $O(1)$.

By controlling π and Θ, one can implement various community structures in SBM. As a particular case, we consider the SBM with $\pi_z = 1/K$ and $\Theta_{z,z'} = \{^{Kc_{\rm in}/N \quad z=z'}_{Kc_{\rm out}/N \quad z\neq z'}$ with $c_{\rm in}, c_{\rm out} \sim O(1)$, namely, every vertex is assigned to each community uniformly. In this case, the mean degrees of intracommunity connections and intercommunity connections converge to $c_{\rm in}$ and $c_{\rm out}$, respectively, as N tends to infinity. This is called the symmetric stochastic block model (SSBM) [9].

III. ANALYTICAL FRAMEWORK

This section introduces the analytical framework used in our study.

A. Statistical mechanical formulation

Our analysis is based on the description of Min-VCP in terms of statistical mechanics [6]. First, for a vertex cover $V_{\rm VC} \subset V$ of a graph $G = (V, E)$, we assign binary variables $x = \{x_i\}_{i\in V}$ to each vertex as follows.

$$x_i := \begin{cases} 1 & v_i \in V_{\rm VC} \\ 0 & v_i \notin V_{\rm VC}. \end{cases} \tag{3}$$

This enables us to express the size of $V_{\rm VC}$ simply as

$$H(x) := \sum_{i\in V} x_i. \tag{4}$$

On the other hand, the constraint that every edge must be connected to at least one element of $V_{\rm VC}$ is expressed by

$$\Delta(x) := \prod_{(i,j)\in E} (1 - (1 - x_i)(1 - x_j)), \tag{5}$$

which returns unity if the constraint is satisfied and vanishes, otherwise. Combining these leads to the Boltzmann distribution,

$$P(x) = \frac{1}{Z(\beta)} e^{-\beta H(x)}\Delta(x), \tag{6}$$

$$Z(\beta) = \sum_x e^{-\beta H(x)}\Delta(x), \tag{7}$$

which yields the uniform distribution of the Min-VC sets in the limit of $\beta \to \infty$.

Based on this formulation, the internal energy per variable can be evaluated as

$$\nu(\beta) = \frac{1}{N}\sum_x P(x)H(x) = -\frac{\partial}{\partial\beta}\frac{1}{N}\ln Z(\beta). \tag{8}$$

This is reduced to the ratio of the covered vertices of the Min-VC set as

$$x_c = \frac{1}{N}\min_{V_{\rm vc}}\{|V_{\rm vc}|\} = \lim_{\beta\to\infty}\nu(\beta) \tag{9}$$

in the limit of $\beta \to \infty$.

B. Cavity method

We denote by ∂i and $x_{\partial i}$ the set of nearest neighbors of a vertex i and that of variables indexed by the neighbors, respectively. In addition, we introduce a distribution $p_{\partial i\to i}(x_{\partial i})$ as

$$p_{\partial i\to i}(x_{\partial i}) = \frac{1}{Z_{\setminus i}}\sum_{x\setminus x_{\partial i}} e^{-\beta\sum_{k\neq i}x_k}\prod_{k,l\neq i}(1 - (1 - x_k)(1 - x_l)), \tag{10}$$

where $A \setminus B$ generally indicates exclusion of subset B from set A, and $Z_{\setminus i}$ is the normalization constant. This distribution stands for the joint distribution of $x_{\partial i}$ for the i-cavity system that is defined by excluding vertex i from the original system. We call $p_{\partial i\to i}(x_{\partial i})$ the joint cavity distribution. For any vertex i of general graphs, the following identity holds between the marginal distribution of variable x_i and the joint cavity distribution $p_{\partial i\to i}(x_{\partial i})$.

$$\begin{aligned} p_i(x_i) &= \sum_{x\setminus\{x_i\}} P(x), \\ &\propto \sum_{x\setminus\{x_i\}} e^{-\beta x_i}\prod_{j\in\partial i}(1 - (1 - x_i)(1 - x_j)) \\ &\quad\times e^{-\beta\sum_{k\neq i}x_k}\prod_{k,l\neq i}(1 - (1 - x_k)(1 - x_l)) \\ &\propto \sum_{x\setminus\{x_i\}} e^{-\beta x_i}\prod_{j\in\partial i}(1 - (1 - x_i)(1 - x_j))p_{\partial i\to i}(x_{\partial i}). \end{aligned} \tag{11}$$

Evaluating $p_{\partial i\to i}(x_{\partial i})$ is nontrivial for general graphs. However, when G does not contain any cycles, it is decomposed as

$$p_{\partial i\to i}(x_{\partial i}) = \prod_{j\in\partial i} p_{j\to i}(x_j), \tag{12}$$

where $p_{j\to i}(x_j)$ represents the marginal distribution of x_j in the i-cavity system, since the exclusion of vertex i makes the components of $x_{\partial i}$ statistically independent of one another. This offers simplified expressions for the distributions as

$$p_i(x_i=1)=\frac{e^{-\beta}}{e^{-\beta}+\prod_{j\in\partial i}p_{j\to i}(x_j=1)}, \tag{13}$$

$$p_i(x_i=0)=1-p_i(x_i=1). \tag{14}$$

Further, by considering the process of adding vertex j to the j-cavity system and excluding vertex $i\in\partial j$, we can efficiently compute the marginal cavity distribution $p_{j\to i}(x_j)$ using a message passing algorithm as

$$p_{j\to i}(x_j=1)=\frac{e^{-\beta}}{e^{-\beta}+\prod_{k\in\partial j\setminus i}p_{k\to j}(x_k=1)}, \tag{15}$$

$$p_{j\to i}(x_j=0)=1-p_{j\to i}(x_j=1). \tag{16}$$

This procedure for evaluating the marginal distributions $p_i(x_i)$ by efficiently computing the cavity distributions $p_{j\to i}(x_j)$ for cycle-free graphs is often termed *belief propagation* (BP) [11].

Two issues should be noted here. The first one is about the treatment of graphs containing cycles. When graphs contain cycles, Eq. (12) does not hold, and therefore, BP does not yield an exact assessment of the marginal distributions. However, it can still be employed as an approximate algorithm since Eq. (15) is performable even if there are cycles. In particular, such treatment is expected to offer a good approximation of an accurate solution for large sparse random graphs such as ERM and SBM since the typical length of cycles for these graphs diverges as $O(\ln N)$, and therefore, their influence becomes negligible as N tends to infinity. This motivates us to employ BP for searching the Min-VC set, which corresponds to the Bethe-Peierls approximation known in physics.

The second point to note is that, in the limit of $\beta\to\infty$, the marginal and cavity distributions can be expressed in simpler forms as

$$p_i(x_i=1)=\begin{cases}0 & \sum_{j\in\partial i}u_{j\to i}=0\\ \frac{1}{2} & \sum_{j\in\partial i}u_{j\to i}=1\\ 1 & \sum_{j\in\partial i}u_{j\to i}>1\end{cases}, \tag{17}$$

$$p_{j\to i}(x_j=1)=\begin{cases}0 & \sum_{k\in\partial j\setminus i}u_{k\to j}=0\\ \frac{1}{2} & \sum_{k\in\partial j\setminus i}u_{k\to j}=1\\ 1 & \sum_{k\in\partial j\setminus i}u_{k\to j}>1\end{cases}, \tag{18}$$

using binary messages $\{u_{j\to i}\}$ that are determined by a reduced expression of BP,

$$u_{j\to i}=\begin{cases}1 & \sum_{k\in\partial j\setminus i}u_{k\to j}=0\\ 0 & \sum_{k\in\partial j\setminus i}u_{k\to j}\geqslant 1\end{cases}. \tag{19}$$

This procedure is referred to as *warning propagation* (WP). A comprehensive graphical explanation for WP is given in [7]. After determining the messages by WP, the cover ratio is evaluated as

$$x_{\rm c}=\frac{1}{N}\sum_{i=1}^{N}p_i(x_i=1). \tag{20}$$

C. Density evolution

So far, we have investigated how to obtain the Min-VC set for a given single sample of graphs. However, ignoring the influence of cycles in graphs, which is expected to be valid for $N\to\infty$ in ERM and SBM, makes it possible to characterize the typical properties of graph ensembles using the densities of messages. In addition, in the limit of $\beta\to\infty$, the densities are expressed in particularly simple forms as

$$R_{z,z'}(u):=(1-\rho_{z,z'})\delta(u)+\rho_{z,z'}\delta(u-1), \tag{21}$$

where z and z' denote the labels of communities and $\rho_{z,z'}\in[0,1]$ is the probability that the binary message $u_{j\to i}$ sent from node j of community z to i of z' takes the value of unity.

For ERM, which corresponds to the single community case ($K=1$), Ref. [8] showed that $\rho_{z,z}=W(c)/c$, where $W(\cdot)$ is the Lambert-W function. On the other hand, in the case of SBM, handling WP as the elementary process provides a set of self-consistent equations to determine $P_z(u)$ under the locally treelike assumption, which is often termed the density evolution (DE), as

$$\begin{aligned}\rho_{z,z'}&=\sum_{d_1,\ldots,d_K}\left(\prod_{z''\neq z'}P_{z,z''}(d_{z''})\right)Q_{z,z'}(d_{z'})\\&\quad\times\prod_{z''=1}^{K}(1-\rho_{z'',z})^{d_{z''}}\\&=\exp\left(-\sum_{z''=1}^{K}c_{z,z''}\rho_{z'',z}\right),\end{aligned} \tag{22}$$

where $Q_{z,z'}(d)=(d+1)P_{z,z'}(d+1)/\sum_{d'=0}^{\infty}d'P_{z,z'}(d')$ is the probability that when an edge is randomly chosen, a terminal node i of z has remaining degree d. Note that Eq. (22) is independent of z'. Thus for the simplicity of notation, we hereafter denote $\rho_{z,z'}$ as ρ_z.

Similarly, we express the densities of marginal probabilities as

$$R_z(p)=\nu_z^{(0)}\delta(p)+\nu_z^{(\frac{1}{2})}\delta\left(p-\tfrac{1}{2}\right)+\nu_z^{(1)}\delta(p-1), \tag{23}$$

where $\nu_z^{(p)}\in[0,1]$ for $p\in\{0,\frac{1}{2},1\}$ is the probability that $p_i(x_i=1)=p$ holds for a randomly chosen node i of community z.

After determining ρ_z from Eq. (22), $\nu_z^{(p)}$ in Eq. (23) is evaluated as

$$\nu_z^{(0)}=\sum_{d_1,\ldots,d_K}\left(\prod_{z'=1}^{K}P_{z,z'}(d_{z'})\right)\prod_{z'=1}^{K}(1-\rho_{z'})^{d_{z'}}=\rho_z, \tag{24}$$

$$\begin{aligned}\nu_z^{(\frac{1}{2})}&=\sum_{d_1,\ldots,d_K}\left(\prod_{z'=1}^{K}P_{z,z'}(d_{z'})\right)\sum_{z'=1}^{K}d_{z'}(1-\rho_{z'})^{d_{z'}-1}\rho_{z'}\\&\quad\times\prod_{z''\neq z}(1-\rho_{z''})^{d_{z''}}\\&=\sum_{z'=1}^{K}c_{z,z'}\rho_z\rho_{z'},\end{aligned} \tag{25}$$

$$\nu_z^{(1)}=1-\nu_z^{(0)}-\nu_z^{(\frac{1}{2})}. \tag{26}$$

This yields the typical cover ratio for SBM as

$$x_{\mathrm{c}} = \mathbb{E}\left[\frac{1}{N}\min_{V_{VC}}\{|V_{VC}|\}\right] = \sum_{z=1}^{K}\pi_z\left(\nu_z^{(1)} + \frac{1}{2}\nu_z^{(\frac{1}{2})}\right), \quad (27)$$

where $\mathbb{E}[\cdot]$ stands for the average with respect to the distribution of graphs of SBM. Equations (22) and (27) constitute the first contribution of this study.

D. Stability analysis

In this subsection, we examine the critical points at which the stability of the fixed point solution of WP is broken using two analytical methods. In the following, we assume that the update scheme for WP is a random sequential update, where one directed edge $j \to i$ is picked up uniformly in a given graph at every time step and the binary message on it is updated according to Eq. (19).

1. Linear stability analysis for DE

We denote ρ_z at the tth step in DE by $\rho_z^{(t)}$. We now consider the process of picking up directed edge $j \to i$ at random from SBM at step t of WP. In this process, the probability that the binary message $u_{j\to i}$ sent from node j of community z to i of z' takes the value of unity at the next step $t+1$ is evaluated as

$$\begin{aligned} q_{z,z'} &= \sum_{d_1,\ldots,d_K}\left(\prod_{z''\neq z'}P_{z,z''}(d_{z''})\right)Q_{z,z'}(d_{z'})\prod_{z''=1}^{K}\left(1-\rho_{z''}^{(t)}\right)^{d_{z''}} \\ &= \prod_{z''=1}^{K}\exp\left(-c_{z,z''}\rho_{z''}^{(t)}\right). \end{aligned} \quad (28)$$

Since Eq. (28) is independent of z', we hereafter denote $q_{z,z'}$ as q_z.

In SBM, the probability q_z becomes independent of the value of message $u_{j\to i}$ at step t as the size of graphs tends to infinity. This indicates that the difference between $\rho_z^{(t+1)}$ and $\rho_z^{(t)}$ is expressed as

$$\begin{aligned} \rho_z^{(t+1)} - \rho_z^{(t)} &\simeq \frac{1}{M_z}\left(-\frac{M_z\rho_z^{(t)}}{M}(1-q_z) + \frac{M_z(1-\rho_z^{(t)})}{M}q_z\right) \\ &= \frac{1}{M}\left(-\rho_z^{(t)}(1-q_z) + \left(1-\rho_z^{(t)}\right)q_z\right), \end{aligned} \quad (29)$$

where $M_z := N\pi_z\sum_{z'=1}^{K}c_{z,z'}$ is the number of messages that are sent from the nodes of community z and $M := \sum_{z=1}^{K}M_z$ is the number of all messages. As $M\to\infty$, Eq. (29) can be treated as a set of ordinary differential equations with the rescaled time $\tau = t/M (d\tau = dt/M = 1/M)$ as

$$\begin{aligned} \frac{d\rho_z}{d\tau} &= -\rho_z(1-q_z) + (1-\rho_z)q_z \\ &= -\rho_z + q_z. \end{aligned} \quad (30)$$

This makes it possible to examine the linear stability of DE of WP for SBM numerically. This is the second contribution of this study.

Specifically for the case of SSBM with two communities, Eq. (30) can be expanded as

$$\frac{d\rho}{d\tau} \simeq \begin{pmatrix} -\rho_1 + \exp(-c_{\mathrm{in}}\rho_1 - c_{\mathrm{out}}\rho_2) \\ -\rho_2 + \exp(-c_{\mathrm{in}}\rho_2 - c_{\mathrm{out}}\rho_1) \end{pmatrix}, \quad (31)$$

where $\rho = (\rho_1, \rho_2)^\top$ and $\top$ denotes the operation of matrix/vector transpose. The Jacobian J is then given as

$$J = -I - A, \quad (32)$$

$$A = \begin{pmatrix} c_{\mathrm{in}}\exp(-c_{\mathrm{in}}\rho_1 - c_{\mathrm{out}}\rho_2) & c_{\mathrm{out}}\exp(-c_{\mathrm{in}}\rho_2 - c_{\mathrm{out}}\rho_1) \\ c_{\mathrm{out}}\exp(-c_{\mathrm{in}}\rho_1 - c_{\mathrm{out}}\rho_2) & c_{\mathrm{in}}\exp(-c_{\mathrm{in}}\rho_2 - c_{\mathrm{out}}\rho_1) \end{pmatrix}, \quad (33)$$

where I is the identity matrix. Therefore, the linear stability around a fixed point solution $\hat{\rho}$ of Eq. (22) can be assessed by examining J at $\hat{\rho}$. If all eigenvalues of J at $\hat{\rho}$ are smaller than zero, the solution is linearly stable with respect to DE.

2. Bug proliferation analysis for WP

DE macroscopically characterizes the solution of WP. However, even if DE converges to a stationary density, it does not necessarily imply the convergence of messages to a fixed point since the messages can continue to move while keeping the density stationary. In ERM, earlier studies [6–8], reported that the cover ratio evaluated by WP coincides with that by other methods for $c < e$, but does not for $c > e$. This transition is accompanied by the instability of the fixed point of WP, which is regarded as a consequence of the replica symmetry breaking [6–8]. To examine the possibility of such a transition for SBM, we generalize a method termed the *bug proliferation analysis* that was introduced for ERM in [8].

The analysis starts with the assumption that messages are at a fixed point of WP. At the fixed point, we flip one message $u_{j\to i}$ to the other value, called a bug, and examine whether the bugs proliferate or die out during iterations of WP.

Let us suppose that an incoming message $u_{j\to i}$ sent from node j of community z'' to i of z' is flipped. Then, the outgoing messages $u_{i\to k}$ from i to $k \in \partial i\backslash\{j\}$ change to the other values if and only if all incoming messages from $h \in \partial i\backslash\{j,k\}$ to i are zeros (Fig. 1).

Since node i is selected randomly from z', the expected number of the messages incoming to z that are flipped by Eq. (19) due to a bug incoming from z'' to z' is evaluated as

$$\begin{aligned} s_{z,z',z''} &= \sum_{d_1,\ldots,d_K}\left(\prod_{z^*\neq z'}P_{z,z^*}(d_{z^*})\right)Q_{z,z'}(d_{z'}) \\ &\quad\times d_{z'}\left(\prod_{z^*\neq z'}(1-\rho_{z^*})^{d_{z^*}}\right)(1-\rho_{z'})^{d_{z'}-1} \\ &= c_{z,z'}\prod_{z^*=1}^{K}\exp(-c_{z,z^*}\rho_{z^*}). \end{aligned} \quad (34)$$

Since Eq. (34) is independent of z'', we hereafter denote $s_{z,z',z''}$ as $s_{z,z'}$.

Next, we denote as $p_z^{(t)}(b_z)$ the probability that there are b_z bugs incoming to community z at the tth step of the random

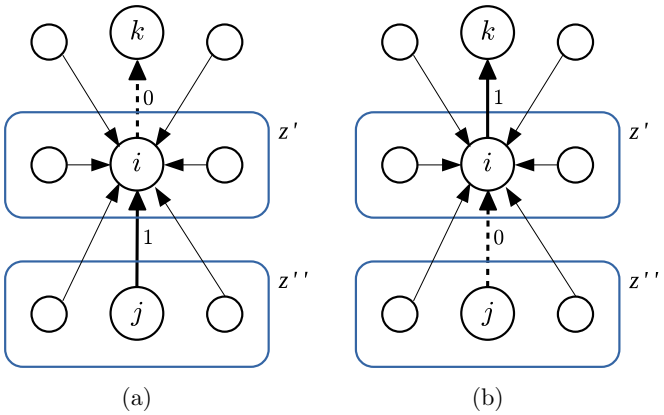

FIG. 1. When $u_{j\to i}=1$ is sent from node j of community z'' to i of z', every outgoing message $u_{i\to k}$ from i to $k\in\partial i\setminus\{j\}$ is set to zero. Therefore, if an incoming message $u_{j\to i}$ is flipped from one to zero [(a) → (b)], each outgoing message $u_{i\to k}$ becomes a new bug if its value is changed to one by Eq. (19), that is, all incoming messages $u_{h\to i}$ from $h\in\partial i\setminus\{j,k\}$ to i are zeros. On the other hand, if an incoming message $u_{j\to i}$ is flipped from zero to one [(b) → (a)], $u_{i\to k}$ becomes a new bug if its original value is one. This condition is equivalent to the previous one.

sequential updates of WP. This probability is updated as

$$\begin{aligned}p_z^{(t+1)}(b_z)=p_z^{(t)}(b_z)&-\frac{b_z}{M}p_z^{(t)}(b_z)+\frac{b_z+1}{M}p_z^{(t)}(b_z+1)\\&-\frac{s_{z,z}b_z}{M}p_z^{(t)}(b_z)+\frac{s_{z,z}(b_z-1)}{M}p_z^{(t)}(b_z-1)\\&-\sum_{z'\neq z}\sum_{b_{z'}=0}^{\infty}\frac{s_{z,z'}b_{z'}}{M}p_{z'}^{(t)}(b_{z'})p_z^{(t)}(b_z)\\&+\sum_{z'\neq z}\sum_{b_{z'}=0}^{\infty}\frac{s_{z,z'}b_{z'}}{M}p_{z'}^{(t)}(b_{z'})p_z^{(t)}(b_z-1).\end{aligned}\tag{35}$$

The second and third terms mean that a bug in the community z is selected for each step, and is fixed. The second term indicates that when a bug in the community z with b_z bugs is selected with the probability b_z/M and is fixed, the number of bugs in community z is no longer b_z. On the other hand, the third term corresponds to the case that the number of bug s in community z varies from b_z+1 to b_z. Similarly, the fourth to seventh terms correspond to the case when a child of bugs which creates a new bug in the community z is selected.

Similar to Eq. (30), in the limit $M\to\infty$, this can be rewritten as ordinal differential equation with the rescaled time $\tau=t/M$ as

$$\begin{aligned}\frac{d}{d\tau}p_z(b_z)=&-b_zp_z(b_z)+(b_z+1)p_z(b_z+1)\\&-s_{z,z}b_zp_z(b_z)+s_{z,z}(b_z-1)p_z(b_z-1)\\&-\sum_{z'\neq z}\sum_{b_{z'}=0}^{\infty}s_{z,z'}b_{z'}p_{z'}(b_{z'})p_z(b_z)\\&+\sum_{z'\neq z}\sum_{b_{z'}=0}^{\infty}s_{z,z'}b_{z'}p_{z'}(b_{z'})p_z(b_z-1).\end{aligned}\tag{36}$$

Let us denote the average of b_z at time τ as $\overline{b_z}:=\sum_{b_z=0}^{\infty}b_zp_z(b_z)$. Equation (36) indicates that $\overline{b_z}$ obeys an ordinary differential equation,

$$\begin{aligned}\frac{d}{d\tau}\overline{b_z}&=-\overline{b_z^2}+\overline{b_z(b_z-1)}-s_{z,z}\overline{b_z^2}+s_{z,z}\overline{b_z(b_z+1)}\\&\quad-\sum_{z'\neq z}s_{z,z'}\overline{b_{z'}b_z}+\sum_{z'\neq z}s_{z,z'}\overline{b_{z'}(b_z+1)}\\&=-\overline{b_z}+\sum_{z'=1}^{K}s_{z,z'}\overline{b_{z'}},\end{aligned}\tag{37}$$

which is the generalization of the bug proliferation analysis proposed for ERM in [8] to SBM. This is the third contribution of this study.

In the case of SSBM with two communities, Eq. (37) is expressed in a simple form as

$$\frac{d}{d\tau}\overline{b}=(-I+A)\cdot\overline{b},\tag{38}$$

where $\overline{b}=(\overline{b}_1,\overline{b}_2)^\top$ and A is introduced at Eq. (33). The corresponding Jacobian is then given by $-I+A$. Unless every eigenvalue of $-I+A$ at $\hat{\rho}$ has values smaller than zero, the fixed point of WP is unstable.

IV. NUMERICAL RESULTS

In this section, we compare the analysis in Sec. III with the results of numerical experiments for SSBM with two communities.

A. Stability analysis for SSBM with two communities

Figures 2 and 3 plot the phase planes of Eq. (31) for SSBM with two communities, where fixed points are denoted by markers. The shape is determined by its stabilities, which are evaluated from the eigenvalues of $-I-A$ [Eq. (32)] and $-I+A$ [Eq. (38)]. The former corresponds to the stability of the macroscopic variables $\rho=(\rho_1,\rho_2)^\top$ while the latter

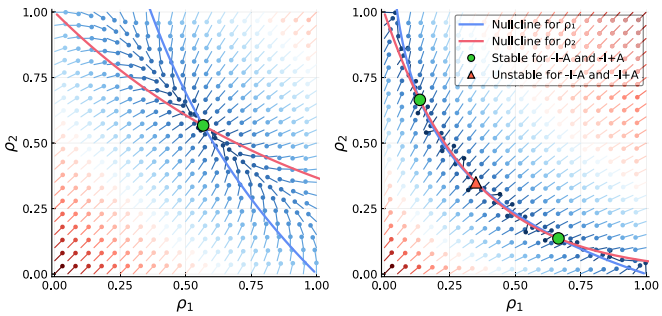

FIG. 2. The phase planes of Eq. (31) for SSBM with two communities. The mean degree of intracommunity is fixed as $c_{\rm in}=0$. Left and right panels correspond to the cases of $c_{\rm out}=1$ and 3, respectively. Solid curves are nullclines of the system, and each marker represents its fixed points. The color of these points indicate the stability evaluated by $-I-A$ [Eq. (32)], that is, the linear stability of $\rho=(\rho_1,\rho_2)^\top$. The difference of the markers indicates the stability assessed by $-I+A$ [Eq. (32)], that is, the microscopic stability with respect to WP.

corresponds to the stability of the microscopic variables $\{u_{j\to i}\}$, respectively.

Due to the underlying symmetry, DE of the two-community SSBM always has a symmetric fixed point of $\rho_1 = \rho_2 = \rho$, which is the unique and stable solution when $c_{\rm in} + c_{\rm out}$ is sufficiently small. In the following, the stability analysis of this solution can be carried out analytically while that for other solutions is performed numerically.

In the case of $c_{\rm in} = 0$ (Fig. 2), at first the symmetric fixed point is the unique and stable solution of Eq. (22). However, as $c_{\rm out}$ increases, a bifurcation occurs, which yields two stable and one unstable fixed points. The stable points assign a value of one to most of the binary messages sent across the two communities. This indicates that either of the two communities is covered more when the bipartite structure becomes sufficiently strong by increasing $c_{\rm out}$.

The behavior for the case of $c_{\rm in} = 1$ (Fig. 3) is somewhat different from that of $c_{\rm in} = 0$. Although the symmetric fixed point is the unique and stable solution when $c_{\rm in} + c_{\rm out}$ is sufficiently small (Fig. 3, top left), its stability is lost microscopically (Fig. 3, top right) at $c_{\rm out} = e - c_{\rm in} = e - 1$. As $c_{\rm out}$ grows further, this fixed point bifurcates two macroscopically stable and one macroscopically unstable fixed points (Fig. 3, bottom left) when the largest eigenvalue of $-I - A$ vanishes. However, the resulting two macroscopically stable fixed points are still unstable microscopically, which indicates that the messages continue to move while keeping the macroscopic distribution stationary. Interestingly, as $c_{\rm out}$ grows further, the macroscopically stable fixed points become microscopically stable as well (Fig. 3, bottom right) at a certain critical point which is larger than e. This behavior can be interpreted to be a consequence of the emergence of strong bipartite structure

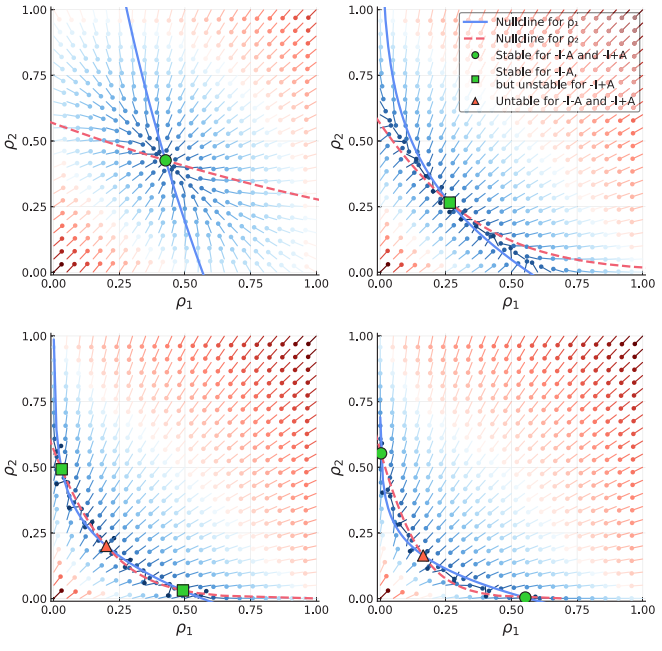

FIG. 3. The phase planes of Eq. (31) for SSBM with two communities. The mean degree of intracommunity is fixed as $c_{\rm in} = 1$. Top left, top right, bottom left, and bottom right panels correspond to the cases of $c_{\rm out} = 1, 4, 7$, and 10, respectively. The description of this figure is the same as in Fig. 2.

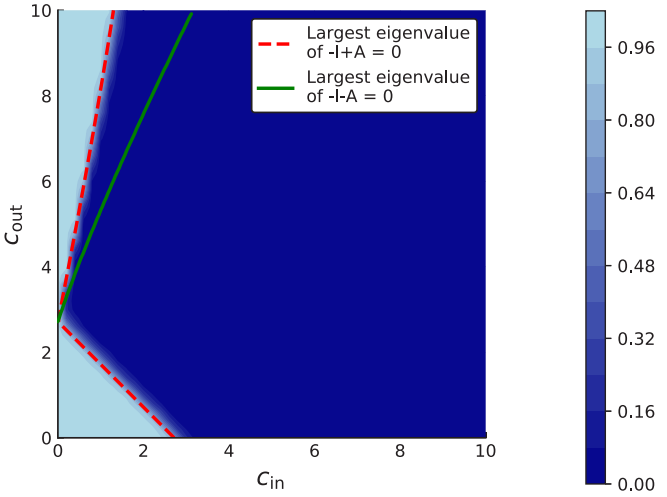

FIG. 4. The convergence probability of WP for SSBM with two communities as a function of $c_{\rm in}$ and $c_{\rm out}$. The red dashed and green solid lines are obtained by conditions that the largest eigenvalues of $-I + A$ of Eq. (38) at the macroscopically stable fixed points and $-I - A$ of Eq. (32) at the symmetric fixed point of $\rho_1 = \rho_2$ vanish, respectively.

created in SBM. As mentioned in Sec II A, Min-VCP belongs to the class P for bipartite graphs.

To confirm these theoretical predictions, we assessed the probability that the sequential random update of WP converges. This probability is calculated according to a method of [8]. Namely, we assessed the fraction of 100 randomly generated graphs with 5000 vertices, for which at least 99% of the messages are converged after 100 sequential updates. The result, in conjunction with the boundary at which the largest eigenvalue of $-I + A$ of Eq. (38) vanishes (red dashed line), is plotted in Fig. 4, which is quite consistent with the above mentioned scenario.

In ERM, it is reported that the instability of WP is associated with the core percolation [12–14]. To examine how this association is relevant in SBM, we plot the ratio of the

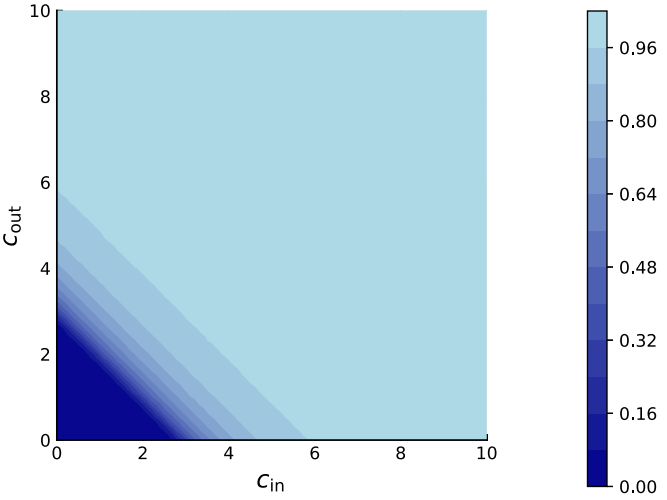

FIG. 5. Core percolation. The ratio of the vertices that belong to the core for SSBM with two communities. The values are evaluated from 100 sample graphs with $N = 2500$.

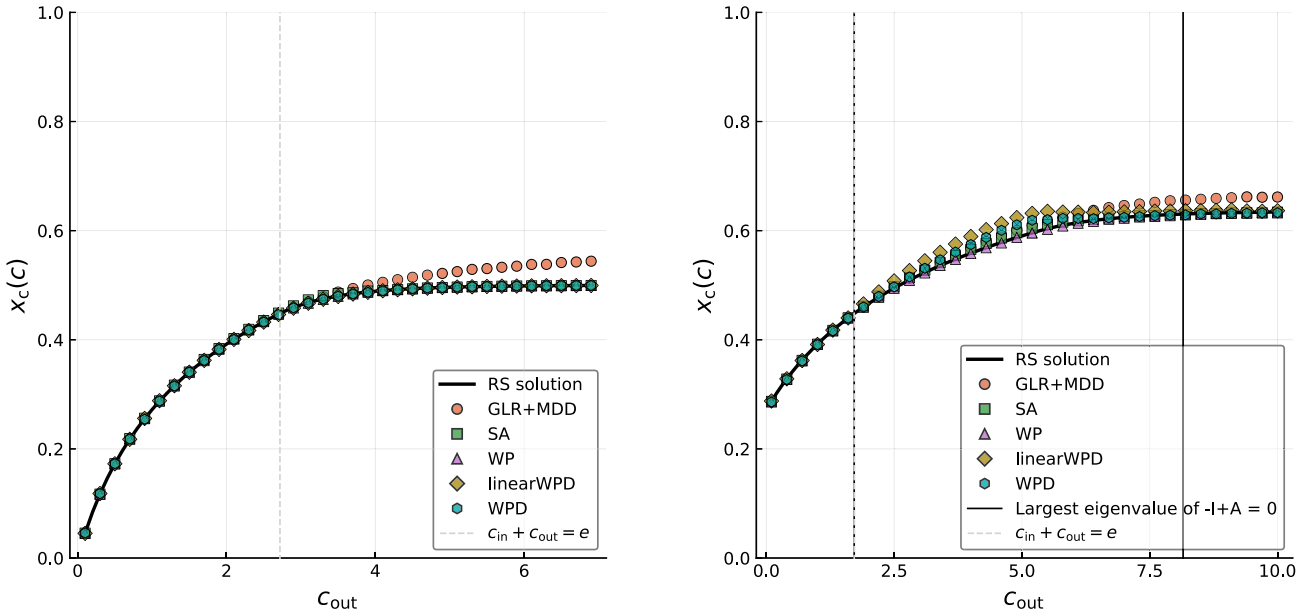

FIG. 6. The cover ratios of Min-VC for SSBM with two communities as a function of c_{out}. Left and right panels correspond to the cases of $c_{in} = 0$ and 1, respectively. Each point indicates the sample average obtained by the corresponding algorithm. Black solid curves represent the replica-symmetric solution obtained by Eq. (27). Gray dotted vertical lines are placed at the point of $c_{in} + c_{out} = e$. Black vertical lines are placed at the point where the largest eigenvalue of $-I + A$ of Eq. (38) vanishes.

vertices that belong to the core in Fig. 5. This figure shows that although the critical condition of the core percolation is in accordance with that of the first transition at $c_{in} + c_{out} = e$, it is not relevant to the second (reentrant) transition that appears in the region of $c_{out} > e$. This indicates that the core percolation is not the sole cause for the emergence of the computational difficulty, and the mesoscopic structures of graphs strongly influence it.

B. Experimental validation

We performed WP on 100 instances of SSBM with $N = 5000$ vertices. In the experiments, we set the maximum number of updates for WP to $100 \times 2M$ times, where M is the number of edges. After WP converged or the number of updates reached the maximum, we calculated the cover ratio x_c using Eq. (20). This is plotted in Fig. 6 along with the analytical result (solid curve).

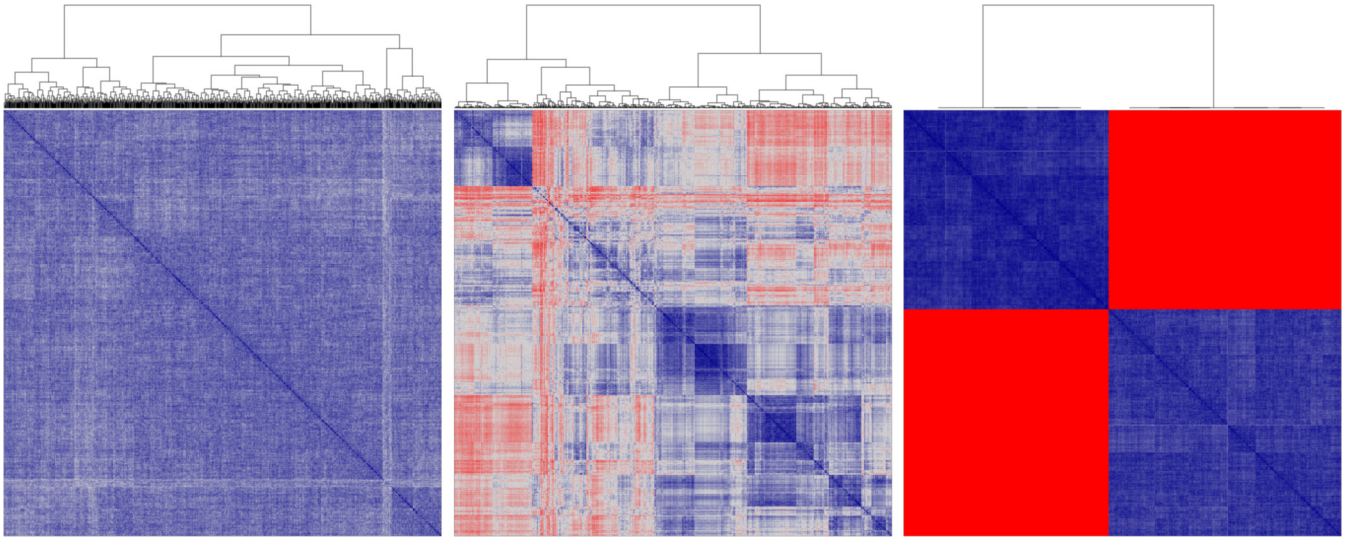

FIG. 7. The heat map of Hamming distance matrix for 1024 Min-VC samples given by SA for SSBM with two communities of $N = 256$. Intradegree c_{in} is fixed to 1. Left, center, and right panels correspond to the cases of $c_{out} = 1, 4,$ and 8, respectively. The arrangement is determined by Ward's method and its hierarchical structure is indicated by the dendrogram.

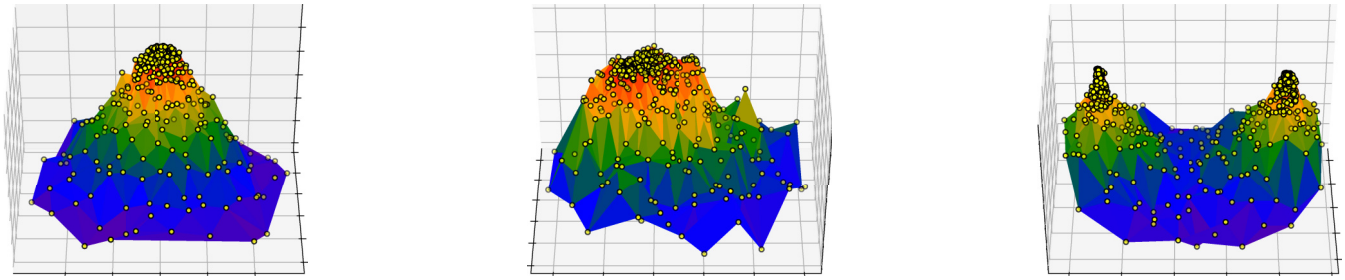

FIG. 8. The landscape of 512 Min-VC samples given by SA for SSBM with two communities of $N = 256$. Intra-degree $c_{\rm in}$ is fixed to 1. Left, center, and right panels correspond to the cases of $c_{\rm out} = 1, 4,$ and 8, respectively. Each solution is mapped from $\{0, 1\}^N$ to $\mathbb{R}^2$ by curvilinear component analysis.

We also tested other algorithms to obtain $x_{\rm c}$ namely, the hybrid algorithm with greedy leaf removal and maximum degree decimation (GLR + MDD) [6], simulated annealing (SA) [15,16], WP, WP decimation (WPD) [6], and WP-based linear-time-and-space algorithm (linearWPD) [17].

In the case of $c_{\rm in} = 0$, the results obtained from each algorithm are in agreement for $c_{\rm out} < e$. However, for $c_{\rm out} > e$, the result obtained by the GLR + MDD algorithm starts to deviate from the results of other algorithms. This can be understood as a consequence of the core percolation which occurs at $c_{\rm out} = e$.

On the other hand, in the case of $c_{\rm in} = 1$, when $1 + c_{\rm out} > e$, the results from the GLR + MDD algorithm as well as those from the WP-based algorithms start to deviate from the result of the SA algorithm, which is regarded as the ground truth. It might be worthwhile to note that the result of WP, which is lower than the "ground truth" in a certain region, does not necessarily mean the value of $x_{\rm c}$ achieved by actual VCs because of degenerate beliefs $p_i(x_i = 1) = p_i(x_i = 0) = 1/2$. For obtaining a VC based on the result of WP, a certain procedure of decimation is necessary for resolving the degeneracy. The results of the two WP-based algorithms (WPD and linear WPD), which are obtained by performing decimation after beliefs are evaluated by WP, consistently yield higher values of $x_{\rm c}$ than that of SA. However, when $c_{\rm out}$ grows further, the results of the WP-based algorithms agree with the result of SA again, supporting the existence of the reentrant transition predicted above.

C. Visualization

Two visualization methods are used to explain the above results graphically. First, we use the Ward's method [18], which is one of the hierarchical clustering methods. This method was applied to the solutions for Min-VCP in earlier studies [7]. We sampled 1024 solutions by SA and clustered them by Hamming distance. Figure 7 plots the dendrogram and heat map of the distance matrix. As $c_{\rm out}$ increases, a hierarchical structure comes out in the solution space, but the solutions are eventually aggregated into two points.

We also use the curvilinear component analysis (CCA) [19], which is one of the nonlinear dimensionality reduction algorithms. The papers [20,21] applied CCA to the solutions of modularity maximization problem. In the current study, we applied CCA to the solutions of Min-VCP. Figure 8 plots the landscape, where we can more visually confirm how the characteristic features of the solution space vary as $c_{\rm out}$ increases.

V. SUMMARY

This study analyzed the typical property of the Min-VCP on SBM using the cavity method. In particular, we examined the critical condition of the computational difficulty for searching for Min-VC sets by the linear stability and bug proliferation analyses for WP. We also performed numerical experiments for SSBM with two communities by various algorithms, which supported predictions obtained by the theoretical analysis.

Min-VC sets are easily found by WP-based algorithms when the total mean degree $c_{\rm in} + c_{\rm out}$ is relatively small. However, it becomes difficult when $c_{\rm in} + c_{\rm out}$ exceeds $e = 2.718...$. This transition is regarded as a consequence of the replica symmetry breaking caused by the core-percolation transition, which is also observed in ERM. However, when $c_{\rm out}$ becomes sufficiently larger than $c_{\rm in}$ in the region of $c_{\rm out} > e$, the solution search by WP-based algorithms becomes easy again. The Min-VCP on a bipartite graph is known to be equivalent to the maximum matching problem that belongs to class P by the König's theorem. The reentrant behavior from the computationally difficult to easy phases presumably reflects this fact. We confirmed this by evaluating the convergence probability of WP and various numerical experiments. These indicate that mesoscopic structures such as "communities" strongly influence the computational difficulty for finding solutions of problems defined on graphs.

Future research scope includes extension of the current analysis to the hypergraphs, which corresponds to *hitting set problem* [22–24], further investigation of RSB using survey propagation [8,25], and validation of the bug analysis [26].

ACKNOWLEDGMENTS

This work was partially supported by JSPS KAKENHI Grant No. 17H00764 and a collaboration with Fujitsu Laboratories Ltd.

[1] Y. Breitbart, C.-Y. Chan, M. Garofalakis, R. Rastogi, and A. Silberschatz, Efficiently monitoring bandwidth and latency in ip networks, in *Proceedings IEEE INFOCOM 2001. Conference on Computer Communications. Twentieth Annual Joint Conference of the IEEE Computer and Communications Society (Cat. No. 01CH37213)*, Vol. 2 (IEEE, Piscataway, 2001), pp. 933–942.

[2] K. Park and H. Lee, On the effectiveness of route-based packet filtering for distributed DOS attack prevention in power-law internets, in *ACM SIGCOMM Computer Communication Review*, Vol. 31 (ACM, New York, 2001), pp. 15–26.

[3] J. Gómez-Gardenes, P. Echenique, and Y. Moreno, Immunization of real complex communication networks, Eur. Phys. J. B **49**, 259 (2006).

[4] M. Weigt and A. K. Hartmann, Number of Guards Needed by a Museum: A Phase Transition in Vertex Covering of Random Graphs, Phys. Rev. Lett. **84**, 6118 (2000).

[5] V. Kavalci, A. Ural, and O. Dagdeviren, Distributed vertex cover algorithms for wireless sensor networks, Int. J. Comp. Netw. Commun. **6**, 95 (2014).

[6] Z. Jin-Hua and Z. Hai-Jun, Statistical physics of hard combinatorial optimization: Vertex cover problem, Chin. Phys. B **23**, 078901 (2014).

[7] A. K. Hartmann and M. Weigt, *Phase Transitions in Combinatorial Optimization Problems: Basics, Algorithms and Statistical Mechanics* (John Wiley & Sons, New York, 2006).

[8] M. Weigt and H. Zhou, Message passing for vertex covers, Phys. Rev. E **74**, 046110 (2006).

[9] E. Abbe, Community detection and stochastic block models: Recent developments, J. Mach. Learn. Res. **18**, 6446 (2017).

[10] J. A. Bondy, U. S. R. Murty *et al.*, *Graph Theory with Applications*, Vol. 290 (Macmillan, London, 1976).

[11] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, Burlington, 1988).

[12] Y.-Y. Liu, E. Csóka, H. Zhou, and M. Pósfai, Core Percolation on Complex Networks, Phys. Rev. Lett. **109**, 205703 (2012).

[13] N. Azimi-Tafreshi, S. Osat, and S. N. Dorogovtsev, Generalization of core percolation on complex networks, Phys. Rev. E **99**, 022312 (2019).

[14] J.-H. Zhao and H.-J. Zhou, Two faces of greedy leaf removal procedure on graphs, J. Stat. Mech.: Theory Exp. (2019), 083401.

[15] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Optimization by simulated annealing, Science **220**, 671 (1983).

[16] A. Lucas, Ising formulations of many NP problems, Frontiers Phys. **2**, 5 (2014).

[17] H. Xu, K. Sun, S. Koenig, and T. K. S. Kumar, A warning propagation-based linear-time-and-space algorithm for the minimum vertex cover problem on giant graphs, in *International Conference on the Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, Lecture Notes in Computer Science Vol 10848 (Springer, Cham, 2018).

[18] J. H. Ward Jr., Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. **58**, 236 (1963).

[19] P. Demartines and J. Hérault, Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets, IEEE Trans. Neural Networks **8**, 148 (1997).

[20] B. H. Good, Y.-A. de Montjoye, and A. Clauset, Performance of modularity maximization in practical contexts, Phys. Rev. E **81**, 046106 (2010).

[21] T. Kawamoto and Y. Kabashima, Counting the number of metastable states in the modularity landscape: Algorithmic detectability limit of greedy algorithms in community detection, Phys. Rev. E **99**, 010301(R) (2019).

[22] M. Mézard and M. Tarzia, Statistical mechanics of the hitting set problem, Phys. Rev. E **76**, 041124 (2007).

[23] S. Takabe and K. Hukushima, Minimum vertex cover problems on random hypergraphs: Replica symmetric solution and a leaf removal algorithm, Phys. Rev. E **89**, 062139 (2014).

[24] B. C. Coutinho, H.-J. Zhou, and Y.-Y. Liu, Percolations on hypergraphs, arXiv:1605.0897.

[25] P. Zhang, Y. Zeng, and H. Zhou, Stability analysis on the finite-temperature replica-symmetric and first-step replica-symmetry-broken cavity solutions of the random vertex cover problem, Phys. Rev. E **80**, 021122 (2009).

[26] T. Castellani, F. Krzakala, and F. Ricci-Tersenghi, Spin glass models with ferromagnetically biased couplings on the Bethe lattice: Analytic solutions and numerical simulations, Eur. Phys. J. B **47**, 99 (2005).