

Article

Exact Recovery of Stochastic Block Model by Ising Model

Feng Zhao ^{1,†} , Min Ye ^{2,†} and Shao-Lun Huang ^{2,*}

¹ Department of Electronics, Tsinghua University, Beijing 100084, China; zhaof17@mails.tsinghua.edu.cn

² Tsinghua Berkeley Shenzhen Institute, Berkeley, CA 94704, USA; yeemmi@sz.tsinghua.edu.cn

* Correspondence: shaolun.huang@sz.tsinghua.edu.cn

† These authors contributed equally to this work.

Abstract: In this paper, we study the phase transition property of an Ising model defined on a special random graph—the stochastic block model (SBM). Based on the Ising model, we propose a stochastic estimator to achieve the exact recovery for the SBM. The stochastic algorithm can be transformed into an optimization problem, which includes the special case of maximum likelihood and maximum modularity. Additionally, we give an unbiased convergent estimator for the model parameters of the SBM, which can be computed in constant time. Finally, we use metropolis sampling to realize the stochastic estimator and verify the phase transition phenomenon through experiments.

Keywords: stochastic block model; exact recovery; Ising model; maximum likelihood; metropolis sampling

1. Introduction

In network analysis, community detection consists in inferring the group of vertices that are more densely connected in a graph [1]. It has been used in many domains, such as recommendation systems [2], task allocation in distributed computing [3], gene expressions [4], and so on. The stochastic block model (SBM) is one of the most commonly used statistical models for community detection problems [5,6]. It provides a benchmark artificial dataset to evaluate different community detection algorithms and inspires the design of many algorithms for community detection tasks. These algorithms, such as semi-definite relaxation, spectral clustering, and label propagation, not only have theoretical guarantees when applied to the SBM, but also perform well on datasets without the SBM assumption. The study of the theoretical guarantee of the SBM model can be divided between the problem of exact recovery and that of partial recovery. Exact recovery requires that the estimated community should be exactly the same as the underlining community structure of the SBM whereas partial recovery expects the ratio of misclassified nodes to be as small as possible. For both cases, the asymptotic behavior of the detection error is analyzed when the scale of the graph tends to infinity. There are already some well-known results for the exact recovery problem on the SBM. To name but a few, Abbe and Mossel established the exact recovery region for a special sparse SBM with two communities [7,8]. Later on, the result was extended to a general SBM with multiple communities [9].

Parameter inference in the SBM is often considered alongside the exact recovery problem. Previous inference methods require the joint estimation of node labels and model parameters [10], which have high complexity since the recovery and inference tasks are done simultaneously. In this article, we will decouple the inference and recovery problems, and propose an unbiased convergent estimator for SBM parameters when the number of communities is known. Once the estimator is obtained, the recovery condition can be checked to determine whether it is possible to recover the labels exactly. Additionally, the estimated parameter will guide the choice of parameters for our proposed stochastic algorithm.

In this article, the exact recovery of the SBM is analyzed by considering the Ising model, which is a probability distribution of node states [11]. We use the terms node



Citation: Zhao, F.; Ye, M.; Huang, S.-L. Exact Recovery of Stochastic Block Model by Ising Model. *Entropy* **2021**, *23*, 65. <https://doi.org/10.3390/e23010065>

Received: 29 November 2020

Accepted: 30 December 2020

Published: 2 January 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

states and node labels interchangeably throughout this paper, both of which refer to the membership of the underlining community. The Ising model was originally proposed in statistical mechanics to model the ferromagnetism phenomenon but has wide applications in neuroscience, information theory, and social networks. Among different variants of Ising models, the phase transition property is shared. Phase transition can be generally formulated when some information quantity changes sharply in a small neighborhood of parameters. Based on the random graph generated by an SBM with two underlining communities, the connection of the SBM and the Ising model was first studied by [12]. Our work will extend the existing result to the multiple community case, establish the phase transition property, and give the recovery error an upper bound. The error bounds decay in a polynomially fast rate in different phases. Then we will propose an alternative approach to estimate the labels by finding the Ising state with maximal probability. Compared with sampling from the Ising model directly, we will show that the optimization approach has a sharper error upper bound. Solving the optimization problem is a generalization of maximum likelihood and also has a connection with maximum modularity. Additionally, searching the state with maximal probability could also be done within all balanced partitions. We will show that this constrained search is equivalent to the graph minimum cut problem, and the detection error upper bound for the constrained maximization will also be given.

The exact solution to maximize the probability function or exact sampling from the Ising model is NP-hard. Many polynomial time algorithms have been proposed for approximation purposes. Among these algorithms, simulated annealing performs well and produces a solution that is very close to the true maximal value [13]. On the other hand, in the original Ising model, metropolis sequential sampling is used to generate samples for the Ising model [14]. Simulated annealing can be regarded as metropolis sampling with decreasing temperature. In this article, we will use the metropolis sampling technique to sample from the Ising model defined on the SBM. This approximation enables us to verify the phase transition property of our Ising model numerically.

This paper is organized as follows. Firstly, in Section 3 we introduce the SBM and give an estimator for the parameters of the SBM. Then, in Section 4, our specific Ising model is given and its phase transition property is obtained. Derived from the Ising model, in Section 5, the energy minimization method is introduced, and we establish its connection with maximum likelihood and modularity maximization algorithm. Furthermore, in Section 6, we realize the Ising model using the metropolis algorithm to generate samples. Numerical experiments and conclusion are given lastly to finish this paper.

Throughout this paper, the community number is denoted by k ; the random undirected graph G is written as $G(V, E)$ with vertex set V and edge set E ; $V = \{1, \dots, n\} =: [n]$; the label of each node is X_i , which is chosen from $W = \{1, \omega, \dots, \omega^{k-1}\}$, and we further require W to be a cyclic group with order k ; W^n is the n -ary Cartesian power of W ; f is a permutation function on W and is extended to W^n in an element-wise manner; U^c is the complement set of U and $|U|$ is the cardinality of U ; the set S_k is used to represent all permutation functions on W and $S_k(\sigma) := \{f(\sigma) | f \in S_k\}$ for $\sigma \in W^n$; the indicator function $\delta(x, y)$ is defined as $\delta(x, y) = 1$ when $x = y$, and $\delta(x, y) = 0$ when $x \neq y$; $f(n) = O(g(n))$ if there exists a constant $c > 0$ such that $f(n) \leq cg(n)$ for large n ; $f(n) = o(g(n))$ holds if $\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 0$; we define the distance of two vectors as: $\text{dist}(\sigma, \sigma') = |\{i \in [n] : \sigma_i \neq \sigma'_i\}|$ for $\sigma, \sigma' \in W^n$ and the distance of a vector to a space $S \subseteq W^n$ as $\text{dist}(\sigma, S) := \min\{\text{dist}(\sigma, \sigma') | \sigma' \in S\}$. For example, when $n = 2$ and $k = 2$, $\sigma = (1, \omega) \in W^2$; $\omega^0 = 1$; $\omega \cdot \omega = \omega^2 = 1$; let f be a mapping such that $f(1) = \omega$ and $f(\omega) = 1$, then $f \in S_2$ and $f(\sigma) = (\omega, 1)$; $\text{dist}(\sigma, f(\sigma)) = 2$; $S_k(\sigma) = \{\sigma, f(\sigma)\}$; and $S_k^c(\sigma) = \{(1, 1), (\omega, \omega)\}$.

2. Related Works

The classical Ising model is defined on a lattice and confined to two states $\{\pm 1\}$. This definition can be extended to a general graph and multiple-state case [15]. In [16],

Liu considered the Ising model as defined on a graph generated by sparse SBM and his focus was to compute the log partition function, which was averaged over all random graphs. In [17], an Ising model with a repelling interaction was considered on a fixed graph structure, and the phase transition condition was established, which involves both the attracting and repelling parameters. Our Ising model derives from the work of [12], but we extend their results by considering the error upper bound and multiple-community case.

The exact recovery condition for the SBM can be derived as a special case of many generalized models, such as pairwise measurements [18], minimax rates [19], and side information [20]. The Ising model in this paper provides another way to extend the SBM model and derives the recovery condition. Additionally, the error upper bound for exact recovery of the two-community SBM by constrained maximum likelihood has been obtained in [7]. Compared with previous results, we establish a sharper upper bound for the multiple-community case in this paper.

The connection between maximum modularity and maximal likelihood was investigated in [21]. To get an optimal value of maximum modularity approximately, simulated annealing was exploited [22], which proceeds by using the partition approach, while the Metropolis sampling used in this paper is applied to estimate the node membership directly.

3. Stochastic Block Model and Parameter Estimation

In this paper, we consider a special symmetric stochastic block model (SSBM), which is defined as follows:

Definition 1 (SSBM with k communities). Let $0 \leq q < p \leq 1$, $V = [n]$ and $X = (X_1, \dots, X_n) \in W^n$. X satisfies the constraint that $|\{v \in [n] : X_v = u\}| = \frac{n}{k}$ for $u \in W$. The random graph G is generated under $\text{SSBM}(n, k, p, q)$ if the following two conditions are satisfied.

1. There is an edge of G between the vertices i and j with probability p if $X_i = X_j$ and with probability q if $X_i \neq X_j$.
2. The existences of each edge are mutually independent.

To explain SSBM in more detail, we define the random variable $Z_{ij} := \mathbb{1}[\{i, j\} \in E(G)]$, which is the indicator function of the existence of an edge between nodes i and j . Given the node labels X , Z_{ij} follows Bernoulli distribution, whose expectation is given by:

$$\mathbb{E}[Z_{ij}] = \begin{cases} p & \text{if } X_i = X_j \\ q & \text{if } X_i \neq X_j \end{cases} \quad (1)$$

Then the random graph G with n nodes is completely specified by $Z := \{Z_{ij}, 1 \leq i < j \leq n\}$ in which all Z_{ij} are jointly independent. The probability distribution for SSBM can be written as:

$$P_G(G) := P_G(Z = z|X) = p^{\sum_{X_i=X_j} z_{ij}} q^{\sum_{X_i \neq X_j} z_{ij}} \cdot (1-p)^{\sum_{X_i=X_j} (1-z_{ij})} (1-q)^{\sum_{X_i \neq X_j} (1-z_{ij})} \quad (2)$$

We will use the notation $\mathcal{G}_n$ to represent the set containing all graphs with n nodes. By the normalization property, $P_G(\mathcal{G}_n) = \sum_{G \in \mathcal{G}_n} P_G(G) = 1$.

In Definition 1, we have supposed that the node label X is fixed instead of a uniformly distributed random variable. Since the maximum posterior estimator is equivalent to the maximum likelihood when the prior is uniform, these two definitions are equivalent. Although the random variable definition is more commonly used in previous literature [6], fixing X makes our formal analysis more concise.

Given the SBM, the exact recovery problem can be formally defined as follows:

Definition 2 (Exact recovery in SBM). Given X , the random graph G is drawn under $\text{SSBM}(n, k, p, q)$. We say that the exact recovery is solvable for $\text{SSBM}(n, k, p, q)$ if there exists an algorithm that takes G as input and outputs $\hat{X}$ such that:

$$P_a(\hat{X}) := P(\hat{X} \in S_k(X)) \rightarrow 1 \text{ as } n \rightarrow \infty$$

In the above definition, the notation $P_a(\hat{X})$ is called the probability of accuracy for estimator $\hat{X}$. Let $P_e(\hat{X}) = 1 - P_a(\hat{X})$ represent the probability of error. Definition 2 can also be formulated as $P_e(\hat{X}) \rightarrow 0$ as $n \rightarrow \infty$. The notation $\hat{X} \in S_k(X)$ means that we can only expect a recovery up to a global permutation of the ground truth label vector X . This is common in unsupervised learning as no anchor exists to assign labels to different communities. Additionally, given a graph G , the algorithm can be either deterministic or stochastic. Generally speaking, the probability of $\hat{X} \in S_k(X)$ should be understood as $\sum_{G \in \mathcal{G}_n} P_G(G) P_{\hat{X}|G}(\hat{X} \in S_k(X))$, which reduced to $P_G(\hat{X} \in S_k(X))$ for the deterministic algorithm.

For constants p, q , which are irrelevant with the graph size n , we can always find algorithms to recover X such that the detection error decreases exponentially fast as n increases; that is to say, the task with a dense graph is relatively easy to handle. Within this paper, we consider a sparse case when $p = \frac{a \log n}{n}, q = \frac{b \log n}{n}$. This case corresponds to the sparsest graph when exact recovery of the SBM is possible. And under this condition, a well known result [9] states that exact recovery is possible if and only if:

$$\sqrt{a} - \sqrt{b} > \sqrt{k} \quad (3)$$

Before diving into the exact recovery problem, we first consider the inference problem for SBM. Suppose k is known, and we want to estimate a, b from the graph G . We offer a simple method by counting the number of edges T_1 and the number of triangles T_2 of G , and the estimators $\hat{a}, \hat{b}$ are obtained by solving the following equation systems:

$$\frac{x + (k-1)y}{2k} = \frac{T_1}{n \log n} \quad (4)$$

$$\frac{1}{k^2} \left(\frac{x^3}{6} + \frac{k-1}{2} xy^2 + (k-1)(k-2) \frac{y^3}{6} \right) = \frac{T_2}{\log^3 n} \quad (5)$$

The theoretical guarantee for the solution is given by the following theorem:

Theorem 1. When n is large enough, the equation system of Equations (4) and (5) has the unique solution $(\hat{a}, \hat{b})$, which are unbiased consistent estimators of (a, b) . That is, $\mathbb{E}[\hat{a}] = a, \mathbb{E}[\hat{b}] = b$, and $\hat{a}$ and $\hat{b}$ converge to a, b in probability, respectively.

Given a graph generated by the SBM, we can use Theorem 1 to obtain the estimated a, b and determine whether exact recovery of label X is possible by Equation (3). Additionally, Theorem 1 provides good estimation of a, b to initialize their parameters of some recovery algorithm like maximum likelihood or our proposed Metropolis sampling in Section 6.

4. Ising Model for Community Detection

In the previous section, we have defined SBM and its exact recovery problem. While SBM is regarded as obtaining the graph observation G from node label X , the Ising model provides a way to generate estimators of X from G by a stochastic procedure. The definition of such an Ising model is given as follows:

Definition 3 (Ising model with k states). Given a graph G sampled from $\text{SSBM}(n, k, \frac{a \log n}{n}, \frac{b \log n}{n})$, the Ising model with parameters $\gamma, \beta > 0$ is a probability distribution of the state vector $\sigma \in W^n$ whose probability mass function is

$$P_{\sigma|G}(\sigma = \bar{\sigma}) = \frac{\exp(-\beta H(\bar{\sigma}))}{Z_G(\alpha, \beta)} \quad (6)$$

where

$$H(\bar{\sigma}) = \gamma \frac{\log n}{n} \sum_{\{i,j\} \notin E(G)} \delta(\bar{\sigma}_i, \bar{\sigma}_j) - \sum_{\{i,j\} \in E(G)} \delta(\bar{\sigma}_i, \bar{\sigma}_j) \quad (7)$$

The subscript in $P_{\sigma|G}$ indicates that the distribution depends on G , and $Z_G(\alpha, \beta)$ is the normalizing constant for this distribution.

In physics, β refers to the inverse temperature and $Z_G(\gamma, \beta)$ is called the partition function. The Hamiltonian energy $H(\bar{\sigma})$ consists of two terms: the repelling interaction between nodes without edge connection and the attracting interaction between nodes with edge connection. The parameter γ indicates the ratio of the strength of these two interactions. The term $\frac{\log n}{n}$ is added to balance the two interactions because there are only $O(\frac{\log n}{n})$ connecting edges for each node. The probability of each state is proportional to $\exp(-\beta H(\bar{\sigma}))$, and the state with the largest probability corresponds to that with the lowest energy.

The classical definition of the Ising model is specified by $H(\sigma) = -\sum_{(i,j) \in E(G)} \sigma_i \cdot \sigma_j$ for $\sigma_i = \pm 1$. There are two main differences between Definition 3 and the classical one. Firstly, we add a repelling term between nodes without an edge connection. This makes these nodes have a larger probability to take different labels. Secondly, we allow the state at each node to take k values from W instead of the two values ± 1 . When $\gamma = 0$ and $k = 2$, Definition 3 is reduced to the classical definition of the Ising model up to a scaling factor.

Definition 3 gives a stochastic estimator $\hat{X}^*$ for X : $\hat{X}^*$ is one sample generated from the Ising model, which is denoted as $\hat{X}^* \sim \text{Ising}(\gamma, \beta)$. The exact recovery error probability for $\hat{X}^*$ can be written as $P_e(\hat{X}^*) := \sum_{G \in \mathcal{G}_n} P_G(G) P_{\sigma|G}(S_k^c(X))$. From this expression we can see that the error probability is determined by two parameters (γ, β) . When these parameters take proper values, $P_e(\hat{X}^*) \rightarrow 0$, and the exact recovery of the SBM is achievable. On the contrary, $P_e(\hat{X}^*) \rightarrow 1$ if (γ, β) takes other values. These two cases are summarized in the following theorem:

Theorem 2. Define the function $g(\beta), \tilde{g}(\beta)$ as follows:

$$g(\beta) = \frac{be^\beta + ae^{-\beta}}{k} - \frac{a+b}{k} + 1 \quad (8)$$

and:

$$\tilde{g}(\beta) = \begin{cases} g(\beta) & \beta \leq \bar{\beta} = \frac{1}{2} \log \frac{a}{b} \\ g(\bar{\beta}) = 1 - \frac{(\sqrt{a} - \sqrt{b})^2}{k} & \beta > \bar{\beta} \end{cases} \quad (9)$$

where $\bar{\beta} = \arg \min_{\beta > 0} g(\beta)$. Let β^* be defined as:

$$\beta^* = \log \left(\frac{a+b-k - \sqrt{(a+b-k)^2 - 4ab}}{2b} \right) \quad (10)$$

which is the solution to the equation $g(\beta) = 0$ and $\beta^* < \bar{\beta}$. Then depending on how (γ, β) take values, for any given $\epsilon > 0$ and $\hat{X}^* \sim \text{Ising}(\gamma, \beta)$, when n is sufficiently large, we have:

1. If $\gamma > b$ and $\beta > \beta^*$, $P_e(\hat{X}^*) \leq n^{\tilde{g}(\beta)/2 + \epsilon}$;
2. If $\gamma > b$ and $\beta < \beta^*$, $P_e(\hat{X}^*) \leq (1 + o(1)) \max\{n^{\tilde{g}(\beta)}, n^{-g(\beta) + \epsilon}\}$;
3. If $\gamma < b$, $P_e(\hat{X}^*) \leq \exp(-Cn)$ for any $C > 0$.

By simple calculus, $\tilde{g}(\beta) < 0$ for $\beta > \beta^*$ and $g(\beta) > 0$ for $\beta < \beta^*$. $g(\tilde{\beta}) < 0$ follows from Equation (3). The illustration of $g(\beta), \tilde{g}(\beta)$ is shown in Figure 1a. Therefore, for sufficiently small ϵ and as $n \rightarrow \infty$, the upper bounds in Theorem 2 all converge to 0 at least in polynomial speed. Therefore, Theorem 2 establishes the sharp phase transition property of the Ising model, which is illustrated in Figure 1b.

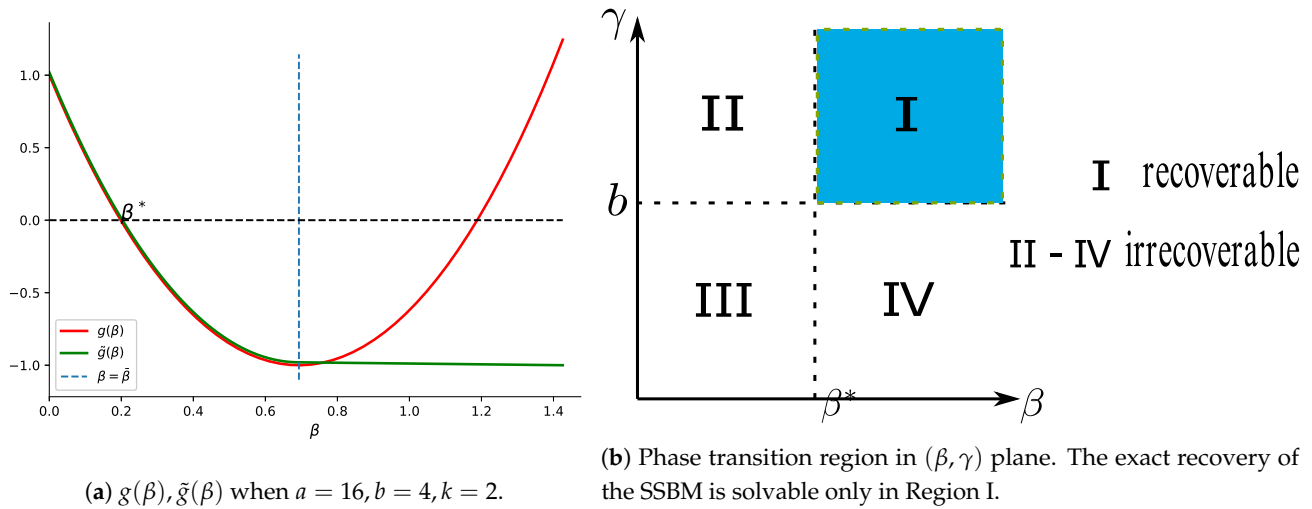


Figure 1. Illustration of Theorem 2.

Theorem 2 can also be understood from the marginal distribution for σ : $P_\sigma(\sigma = \bar{\sigma}) = \sum_{G \in \mathcal{G}_n} P_G(G) P_{\sigma|G}(\sigma = \bar{\sigma})$. Let $D(\sigma, \sigma')$ be the event when σ is closest to σ' among all its permutations. That is,

$$D(\sigma, \sigma') := \{\sigma = \arg \min_{f \in S_k} \text{dist}(f(\sigma), \sigma')\} \quad (11)$$

Then Theorem 2 can be stated with respect to the marginal distribution P_σ :

Corollary 1. Suppose $\gamma > b$, depending on how β takes values:

1. When $\beta > \beta^*$, $P_\sigma(\sigma = X|D(\sigma, X)) = 1 - o(1)$;
2. When $\beta < \beta^*$, $P_\sigma(\sigma = X|D(\sigma, X)) = o(1)$.

Below we outline the proof ideas of Theorem 2. The insight is obtained from the analysis of the one-flip energy difference. This useful result is summarized in the following lemma:

Lemma 1. Suppose $\bar{\sigma}'$ differs from $\bar{\sigma}$ only at position r by $\bar{\sigma}'_r = \omega^s \cdot \bar{\sigma}_r$. Then the change of energy is:

$$\begin{aligned} H(\bar{\sigma}') - H(\bar{\sigma}) &= (1 + \gamma \frac{\log n}{n}) \sum_{i \in N_r(G)} J_s(\bar{\sigma}_r, \bar{\sigma}_i) \\ &\quad + \gamma \frac{\log n}{n} (m(\omega^s \cdot \bar{\sigma}_r) - m(\bar{\sigma}_r) + 1) \end{aligned} \quad (12)$$

where $m(\omega^j) := |\{i \in [n] | \bar{\sigma}_i = \omega^j\}|$, $N_r(G) := \{j | (r, j) \in E(G)\}$ and $J_s(x, y) = \delta(x, y) - \delta(\omega_s \cdot x, y)$.

Lemma 1 gives an explicit way to compare the probability of two neighboring states by the following equality:

$$\frac{P_{\sigma|G}(\sigma = \bar{\sigma}')}{P_{\sigma|G}(\sigma = \bar{\sigma})} = \exp(-\beta(H(\bar{\sigma}') - H(\bar{\sigma}))) \quad (13)$$

Additionally, since the graph is sparse and every node has $O(\log n)$ neighbors, from Equation (12) the computational cost (time complexity) for the energy difference is also $O(\log n)$.

When $H(\bar{\sigma}') > H(\bar{\sigma})$, we can expect $P_{\sigma|G}(\sigma = \bar{\sigma}')$ is far less than $P_{\sigma|G}(\sigma = \bar{\sigma})$. Roughly speaking, if $\sum_{\text{dist}(\sigma', X)=1} \exp(-\beta(H(\bar{\sigma}') - H(X)))$ converges to zero, we can expect the probability of all other states differing from $S_k(X)$ converges to zero. On the contrary, if $\sum_{\text{dist}(\sigma', X)=1} \exp(-\beta(H(\bar{\sigma}') - H(X)))$ tends to infinity, then $P_{\sigma}(S_k(X))$ converges to zero. This illustrates the idea behind the proof of Theorem 2. The rigorous proof can be found in Section 8.

5. Community Detection via Energy Minimization

Since β^* is irrelevant with n , when $\gamma > b$, we can choose a sufficiently large β such that $\beta > \beta^*$, then by Theorem 2, $\sigma \in S_k(X)$ almost surely, which implies that $P_{\sigma|G}(\sigma = X)$ has the largest probability for almost all graphs G sampled from the SBM. Therefore, instead of sampling from the Ising model, we can directly maximize the conditional probability to find the state with the largest probability. Equivalently, we can proceed by minimizing the energy term in Equation (7):

$$\hat{X}' := \arg \min_{\bar{\sigma} \in W^n} H(\bar{\sigma}) \quad (14)$$

In (14), we allow $\bar{\sigma}$ to take values from W^n . Since we know X has equal size $|\{v \in [n] : X_v = u\}| = \frac{n}{k}$ for each label u , another formulation is to restrict the search space to $W^* := \{\sigma \in W^n \mid |\{v \in [n] : \sigma_v = \omega^s\}| = \frac{n}{k}, s = 0, \dots, k-1\}$. When $\sigma \in W^*$, minimizing $H(\sigma)$ is equivalent to:

$$\hat{X}'' := \arg \min_{\sigma \in W^*} \sum_{\{i,j\} \notin E(G)} \delta(\sigma_i, \sigma_j) \quad (15)$$

where the minimal value is the minimum cut between different detected communities.

When $\hat{X}'' \neq X$, we must have $\text{dist}(\hat{X}'', X) \geq 2$ to satisfy the constraint $\hat{X}'' \in W^*$. Additionally, the estimator of $\hat{X}''$ is parameter-free whereas $\hat{X}'$ depends on γ . The extra parameter γ in the expression of $\hat{X}'$ can be regarded as a kind of Lagrange multiplier for this integer programming problem. Thus, the optimization problem for $\hat{X}'$ is the relaxation of that for $\hat{X}''$ by introducing a penalized term and enlarging the searched space from W^* to W^n .

When $\beta > \bar{\beta}$, $\bar{g}(\beta)$ becomes a constant value. Therefore, we can get $n^{g(\bar{\beta})/2}$ as the tightest error upper bound for the Ising estimator $\hat{X}^*$ from Theorem 2. For the estimator $\hat{X}'$ and $\hat{X}''$, we can obtain a sharper error upper bound, which is summarized in the following theorem:

Theorem 3. When $\sqrt{a} - \sqrt{b} > \sqrt{k}$, for sufficiently large n ,

1. If $\gamma > b$, $P_G(\hat{X}' \notin S_k(X)) \leq (k-1 + o(1))n^{g(\bar{\beta})}$;
2. $P_G(\hat{X}'' \notin S_k(X)) \leq ((k-1)^2 + o(1))n^{2g(\bar{\beta})}$.

As $g(\bar{\beta}) < 0$, $n^{2g(\bar{\beta})} < n^{g(\bar{\beta})} < n^{g(\bar{\beta})/2}$, Theorem 3 implies that $P_e(\hat{X}'')$ has the sharpest upper bound among the three estimators. This can be intuitively understood as the result of smaller search space. The proof technique of Theorem 3 is to consider the probability of events $H(X) > H(\bar{\sigma})$ for $\text{dist}(\sigma, X) \geq 1$. Then by union bound, these error probabilities can be summed up. We note that a loose bound $n^{g(\bar{\beta})/4}$ was obtained in [7] for the estimator $\hat{X}''$ when $k = 2$. For a general case, since $\bar{g}(\beta) = 1 - \frac{(\sqrt{a}-\sqrt{b})^2}{k}$, Theorem 3 implies that exact recovery is possible using $\hat{X}'$ as long as $\sqrt{a} - \sqrt{b} > \sqrt{k}$ is satisfied.

Estimator $\hat{X}'$ has one parameter, γ . When γ takes different values, $\hat{X}'$ is equivalent with maximum likelihood or maximum modularity in the asymptotic case. The following analysis shows their relationship intuitively.

The maximum likelihood estimator is obtained by maximizing the log-likelihood function. From (2), this function can be written as:

$$\log P_G(Z|X = \sigma) = -\log \frac{a}{b} \cdot H(\sigma) + C$$

where the parameter γ in $H(\sigma)$ satisfies $\gamma \frac{\log n}{n} = \frac{1}{\log(a/b)} (\log(1 - \frac{a \log n}{n}) - \log(1 - \frac{b \log n}{n}))$ and C is a constant irrelevant with σ . When n is sufficiently large, we have $\gamma \rightarrow \gamma_{ML} := \frac{a-b}{\log(a/b)}$. That is, the maximum likelihood estimator is equivalent to $\hat{X}'$ when $\gamma = \gamma_{ML}$ asymptotically.

The maximum modularity estimator is obtained by maximizing the modularity of a graph [23], which is defined by:

$$Q = \frac{1}{2|E|} \sum_{ij} (A_{ij} - \frac{d_i d_j}{2|E|}) \delta(C_i, C_j) \quad (16)$$

For the i -th node, d_i is its degree and C_i is its community belonging. A is the adjacency matrix. Up to a scaling factor, the modularity Q can be re-written using the label vector σ as:

$$\begin{aligned} Q(\sigma) = & - \sum_{\{i,j\} \notin E(G)} \frac{d_i d_j}{2|E|} \delta(\sigma_i, \sigma_j) \\ & + \sum_{\{i,j\} \in E(G)} (1 - \frac{d_i d_j}{2|E|}) \delta(\sigma_i, \sigma_j) \end{aligned} \quad (17)$$

From (17), we can see that $Q(\sigma) \rightarrow -H(\sigma)$ with $\gamma = \gamma_{MQ} = \frac{a+b}{2}$ as $n \rightarrow \infty$. Indeed, we have $d_i \sim \frac{(a+b) \log n}{2}$, $|E| \sim \frac{1}{2} n d_i$. Therefore, we have $\frac{d_i d_j}{2|E|} \rightarrow \gamma_{MQ} \frac{\log n}{n}$. That is, the maximum modularity estimator is equivalent with $\hat{X}'$ when $\gamma = \gamma_{MQ}$ asymptotically.

Using $a > b$ and the inequality $x - 1 > \log x > 2 \frac{x-1}{x+1}$ for $x > 1$ we can verify that $\gamma_{MQ} > \gamma_{ML} > b$. That is, both the maximum likelihood and the maximum modularity estimator satisfy the exact recovery conditions $\gamma > b$ in Theorem 3.

6. Community Detection Based on Metropolis Sampling

From Theorem 2, if we could sample from the Ising model, then with large probability, the sample is aligned with X . However, exact sampling is difficult when n is very large since the cardinality of the state space increases in the rate of k^n . Therefore, some approximation is necessary, and the most common way to generate an Ising sample is using Metropolis sampling [14]. Empirically speaking, starting from a random state, the Metropolis algorithm updates the state by randomly selecting one position to flip its state at each iteration step. Then after some initial burning time, the generated samples can be regarded as sampling from the Ising model.

The theoretical guarantee of Metropolis sampling is based on the Markov chain. Under some general conditions, Metropolis samples converge to the steady state of the Markov chain, and the steady state follows the probability distribution to be approximated. For the Ising model, there are many previous works which have shown the convergence of Metropolis sampling [24].

For our specific Ising model and energy term in Equation (7), the pseudo code of our algorithm is summarized in Algorithm 1. This algorithm requires that the number of the communities k is known and the strength ratio parameter γ is given. We should choose $\gamma > b$ where b is estimated by $\hat{b}$ in Theorem 1. The iteration time N should also be specified in advance.

Algorithm 1 Metropolis sampling algorithm for SBM.Inputs: the graph G , inverse temperature β , the strength ratio parameter γ Output: $\hat{X} = \bar{\sigma}$

```

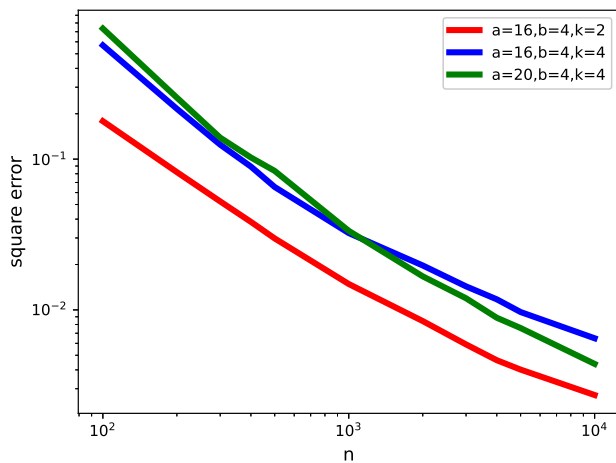
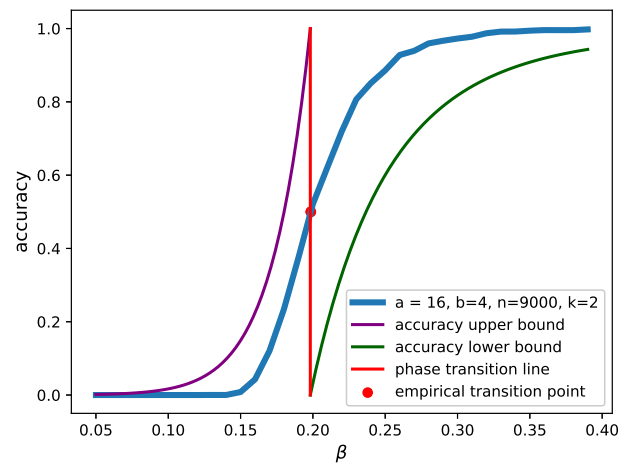
1: random initialize  $\bar{\sigma} \in W^n$ 
2: for  $i = 1, 2, \dots, N$  do
3:   propose a new state  $\bar{\sigma}'$  according to Lemma 1 where  $s, r$  are randomly chosen
4:   compute  $\Delta H(r, s) = H(\bar{\sigma}') - H(\bar{\sigma})$  using (12)
5:   if  $\Delta H(r, s) < 0$  then
6:      $\sigma_r \leftarrow w^s \cdot \sigma_r$ 
7:   else
8:     with probability  $\exp(-\beta \Delta H(r, s))$  such that  $\sigma_r \leftarrow w^s \cdot \sigma_r$ 
9:   end if
10: end for

```

The computation of $\Delta H(r, s)$ needs $O(\log n)$ time from Lemma 1. For some special Ising model, it needs to take $N = O(n \log n)$ to generate the sample for good approximation [25]. For our model, it is unknown whether $O(n \log n)$ is sufficient, and we empirically chose $N = O(n^2)$ in numerical experiments. Then the time complexity of Algorithm 1 is $O(n^2 \log n)$.

In the remaining part of this section, we present experiments conducted to verify our theoretical results. Firstly, we considered several combinations of (a, b, k) and obtained the estimator $(\hat{a}, \hat{b})$ by Theorem 1. Using the empirical mean squared error (MSE) $\frac{1}{m} \sum_{i=1}^m (\hat{a} - a)^2 + (\hat{b} - b)^2$ as the criterion and choosing $m = 1000$, the result is shown in Figure 2a. As we can see, as n increases, the MSE decreases polynomially fast. Therefore, the convergence of $\hat{a} \rightarrow a$ and $\hat{b} \rightarrow b$ was verified.

Secondly, using Metropolis sampling, we conducted a moderate simulation to verify Theorem 2 for the case $\gamma > b$. We chose $n = 9000, k = 2$, and the empirical accuracy was computed by $P_e = \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{1}[\hat{X}^* = \pm X]$. In this formula, m_1 is the number of times the random graph was generated by the SBM, whereas m_2 is the number of times consecutive samples were generated by Algorithm 1 for a given graph. We chose $m_1 = 2100, m_2 = 6000$, which is fairly large and can achieve a good approximation of $P_e(\hat{X}^*)$ by the law of large numbers. The result is shown in Figure 2b.

(a) Estimation error of $\hat{a}, \hat{b}$ with respect to n .(b) The accuracy of exact recovery by $\hat{X}^*$.**Figure 2.** Experimental results.

The vertical red line ($\beta = \beta^* = 0.198$), computed from (10), represents the phase transition threshold. The point $(0.199, \frac{1}{2})$ in the figure can be regarded as the empirical phase transition threshold, whose first coordinate is close to β^* . The green line $(\beta, n^{g(\beta)/2})$ is the theoretical lower bound of accuracy for $\beta > \beta^*$, and the purple line $(\beta, n^{-g(\beta)})$ is the theoretical upper bound of accuracy for $\beta < \beta^*$. It can be expected that as n becomes larger, the empirical accuracy curve (blue line in the figure) will approach the step function, which jumps from 0 to 1 at $\beta = \beta^*$.

7. Conclusions

In this paper, we presented one convergent estimator (in Theorem 1) to infer the parameters of the SBM and analyzed three label estimators to detect communities of the SBM. We gave the exact recovery error upper bound for all label estimators (in Theorems 2 and 3) and studied their relationships. By introducing the Ising model, our work makes a new path to study the exact recovery problem for the SBM. More theoretical and empirical work will be done in the future, such as convergence analyses on modularity (in Equation (17)), the necessary iteration time (in Algorithm 1) for Metropolis sampling, and so on.

8. Proof of Main Theorems

8.1. Proof of Theorem 1

Lemma 2. Consider an Erdős–Rényi random graph G with n nodes, in which edges are placed independently with probability p [26]. Suppose $p = \frac{a \log n}{n}$, the number of edges is denoted by $|E|$ while the number of triangles is T . Then $\frac{|E|}{n \log n} \rightarrow \frac{a}{2}$ and $\frac{T}{\log^3 n} \rightarrow \frac{a^3}{6}$ in probability.

Proof. Let X_{ij} represent a Bernoulli random variable with parameter p . Then $|E| = \sum_{i,j} X_{ij}$, X_{ij} are i.i.d. $\mathbb{E}[T(G)] = \frac{n(n-1)}{2} p = \frac{(n-1) \log n}{2} a$ and $\text{Var}[|E|] = \frac{n(n-1)}{2} p(1-p) < a \frac{(n-1) \log n}{2}$. Then by Chebyshev's inequality,

$$P\left(\left|\frac{|E|}{n \log n} - \frac{a}{2} \frac{n-1}{n}\right| > \epsilon\right) \leq \frac{\text{Var}[|E|/(n \log n)]}{\epsilon^2} < \frac{a(n-1)}{2n^2 \epsilon^2 \log n}$$

For a given ϵ , when n is sufficiently large,

$$P\left(\left|\frac{|E|}{n \log n} - \frac{a}{2}\right| > \epsilon\right) < P\left(\left|\frac{|E|}{n \log n} - \frac{a}{2} \frac{n-1}{n}\right| > 2\epsilon\right) \leq \frac{n-1}{8n^2 \epsilon^2 \log n}$$

Therefore, by the definition of convergence in probability, we have $\frac{|E|}{n \log n} \rightarrow \frac{a}{2}$ as $n \rightarrow \infty$.

Let X_{ijk} represents a Bernoulli random variable with parameter p^3 . Then $T = \sum_{i,j,k} X_{ijk}$. It is easy to compute that $\mathbb{E}[T] = \binom{n}{3} p^3$. Since X_{ijk} are not independent, the variance of T needs careful calculation. From [27] we know that:

$$\begin{aligned} \text{Var}[T] &= \binom{n}{3} p^3 + 12 \binom{n}{4} p^5 + 30 \binom{n}{5} p^6 + 20 \binom{n}{6} p^6 \\ &\quad - \binom{n}{3}^2 p^6 = O(\log^3 n) \end{aligned}$$

Therefore by Chebyshev's inequality,

$$P\left(\left|\frac{T}{\log^3 n} - \frac{a^3}{6} \frac{(n-1)(n-2)}{n^2}\right| > \epsilon\right) \leq \frac{\text{Var}[T/\log^3 n]}{\epsilon^2} = \frac{1}{\epsilon^2} O\left(\frac{1}{\log^3 n}\right)$$

Hence, $\frac{T}{\log^3 n} \rightarrow \frac{a^3}{6}$. $\square$

The convergence of $|E|$ in the Erdős–Rényi graph can be extended directly to the SBM since the existence of each edge is independent. However, for T , it is a little tricky since the existences of each triangle are mutually dependent. The following two lemmas give the formula for the variance of inter-community triangles in the SBM.

Lemma 3. Consider a two-community SBM $(2n, p, q)$ and count the number of triangles T , which has a node in S_1 and an edge in S_2 . Then the variance of T is:

$$\begin{aligned} \text{Var}[T] &= \frac{n^2(n-1)}{2} q^2 p + n^2(n-1)(n-2) p^2 q^3 \\ &\quad + \frac{n^2(n-1)^2}{2} q^4 p - \frac{n^2(n-1)(3n-4)}{2} q^4 p^2 \end{aligned} \quad (18)$$

Lemma 4. Consider a three-community SBM $(3n, p, q)$ and count the number of triangles T , which has a node in S_1 , one node in S_2 , and one node in S_3 . Then the variance of T is:

$$\text{Var}[T] = n^3 q^3 + 3n^3(n-1)q^4 + 3n^3(n-1)^2 q^5 - n^3(3n^2 - 3n + 1)q^6$$

The proof of the above two lemmas uses some counting techniques and is similar to that in [27], and we omit it here.

Lemma 5. For a SBM (n, k, p, q) where $p = \frac{a \log n}{n}$, $q = \frac{b \log n}{n}$. The number of triangles is T . Then $\frac{T}{(\log n)^3}$ converges to $\frac{1}{k^2} \left(\frac{a^3}{6} + \frac{k-1}{2} ab^2 + (k-1)(k-2) \frac{b^3}{6} \right)$ in probability as $n \rightarrow \infty$.

Proof. We split T into three parts: the first is the number of triangles within community i , T_i . There are k terms of T_i . The second is the number of triangles that have one node in community i and one edge in community j , T_{ij} . There are $k(k-1)$ terms of T_{ij} . The third is the number of triangles that have one node in community i , one node in community j and one node in community k .

We only need to show that:

$$\frac{T_i}{\log^3 n} \rightarrow \frac{(a/k)^3}{6} \quad (19)$$

$$\frac{T_{ij}}{\log^3 n} \rightarrow \frac{1}{2} (a/k)(b/k)^2 \quad (20)$$

$$\frac{T_{ijk}}{\log^3 n} \rightarrow (b/k)^3 \quad (21)$$

The convergence of $\frac{T_i}{\log^3 n}$ comes from Lemma 2. For T_{ij} we use the conclusion from Lemma 3. We replace n with n/k , $p = a \frac{\log n}{n}$, and $q = b \frac{\log n}{n}$ in Equation (18).

$\text{Var}[T_{ij}] \sim \frac{ab^2}{2k^3} \log^3 n$. Since the expectation of $\frac{T_{ij}}{\log^3 n}$ is $(n/k) \binom{n/k}{2} pq^2 / (\log^3 n) = \frac{n-1}{2n} \frac{ab^2}{k^3}$, by Chebyshev's inequality we can show that:

$$P\left(\left|\frac{T_{ij}}{\log^3 n} - \frac{n-1}{2n} \frac{ab^2}{k^3}\right| > \epsilon\right) \leq \frac{\text{Var}[T_{ij}/\log^3 n]}{\epsilon^2} = \frac{1}{\epsilon^2} O\left(\frac{1}{\log^3 n}\right)$$

Therefore, $\frac{T_{ij}}{\log^3 n}$ converges to $\frac{1}{2}(a/k)(b/k)^2$.

To prove $\frac{T_{ijk}}{\log^3 n} \rightarrow (b/k)^3$, from Lemma 4 we can get $\text{Var}[T_{ijk}] = O(\log^5 n)$:

$$P\left(\left|\frac{T_{ijk}}{\log^3 n} - \frac{b^3}{k^3}\right| > \epsilon\right) \leq \frac{\text{Var}[T_{ijk}/\log^3 n]}{\epsilon^2} = \frac{1}{\epsilon^2} O\left(\frac{1}{\log n}\right)$$

□

Proof of Theorem 1. Let $e_1^* = \frac{a+(k-1)b}{2k}$, $k^2 e_2^* = \frac{a^3}{6} + \frac{k-1}{2} ab^2 + (k-1)(k-2) \frac{b^3}{6}$ and $e_1 = \frac{T_1}{n \log n}$, $e_2 = \frac{T_2}{\log^3 n}$. From Lemma 2, $e_1 \rightarrow e_1^*$. From Lemma 5, $e_2 \rightarrow e_2^*$ as $n \rightarrow \infty$. Using $x = 2ke_1 - (k-1)y$, we can get:

$$g(y) := (k-1)(y^3 - 6e_1 y^2 + 12e_1^2 y) + 6e_2 - 8ke_1^3 = 0 \quad (22)$$

This equation has a unique real root since $g(y)$ is increasing on $\mathbb{R}$: $g'(y) = 3(k-1)(y - 2e_1)^2 \geq 0$. Next we show that the root lies within $(0, 2e_1)$.

$$\begin{aligned} \lim_{n \rightarrow \infty} g(0) &= 6e_2^* - 8k(e_1^*)^3 = -\frac{3}{k^2} (k-1)(k-2)ab^2 \\ &\quad - \frac{3(k-1)}{k^2} a^2 b - \frac{k-1}{k^2} ((k-2) - (k-1)^2) b^3 < 0 \\ \lim_{n \rightarrow \infty} g(2e_1) &= 6e_2^* - 8(e_1^*)^3 = \frac{(k-1)(a-b)^3}{k^3} > 0 \end{aligned}$$

Therefore, we can get a unique solution y within $(0, 2e_1)$. Since (a, b) is a solution for the equation array, the conclusion follows.

By taking expectation on both sides of Equations (4) and (5) we can show $\mathbb{E}[\hat{a}] = a$, $\mathbb{E}[\hat{b}] = b$. By the continuous property of $g(y)$, $\hat{b} \rightarrow b$ and $\hat{a} \rightarrow a$ follows similarly. □

8.2. Proof of Theorem 2

Proof of Lemma 1. First we rewrite the energy term in (7) as:

$$H(\bar{\sigma}) = \gamma \frac{\log n}{n} \sum_{i < j} \delta(\bar{\sigma}_i, \bar{\sigma}_j) - (1 + \gamma \frac{\log n}{n}) \sum_{\{i,j\} \in E(G)} \delta(\bar{\sigma}_i, \bar{\sigma}_j)$$

Then calculating the energy difference term by:

$$\begin{aligned}
 H(\bar{\sigma}') - H(\bar{\sigma}) &= (1 + \gamma \frac{\log n}{n}) \\
 &\quad \cdot \sum_{i \in N_r(G)} (\delta(\bar{\sigma}_r, \bar{\sigma}_i) - \delta(\omega^s \cdot \bar{\sigma}_r, \bar{\sigma}_i)) \\
 &\quad + \gamma \frac{\log n}{n} \sum_{i \neq r} (\delta(\omega^s \cdot \bar{\sigma}_r, \bar{\sigma}_i) - \delta(\bar{\sigma}_r, \bar{\sigma}_i)) \\
 &= (1 + \gamma \frac{\log n}{n}) \sum_{i \in N_r(G)} J_s(\bar{\sigma}_r, \bar{\sigma}_i) \\
 &\quad + \gamma \frac{\log n}{n} \sum_{i=1}^n (\delta(\omega^s \cdot \bar{\sigma}_r, \bar{\sigma}_i) - \delta(\bar{\sigma}_r, \bar{\sigma}_i) + 1) \\
 &= (1 + \gamma \frac{\log n}{n}) \sum_{i \in N_r(G)} J_s(\bar{\sigma}_r, \bar{\sigma}_i) \\
 &\quad + \gamma \frac{\log n}{n} (m(\omega^s \cdot \bar{\sigma}_r) - m(\bar{\sigma}_r) + 1)
 \end{aligned}$$

□

Before diving into the technical proof of Theorem 2, we need to introduce some extra notations. When $\bar{\sigma}$ differs from X only at position r , taking $\bar{\sigma} = X$ in Lemma 1, we have:

$$H(\bar{\sigma}') - H(\bar{\sigma}) = (1 + \gamma \frac{\log n}{n})(A_r^0 - A_r^s) + \gamma \frac{\log n}{n} \quad (23)$$

where A_r^s is defined as $A_r^s = |\{j \in [n] \setminus \{r\} : \{j, r\} \in E(G), X_j = \omega^s \cdot X_r\}|$. Since the existence of each edge in G is independent, $A_r^s \sim \text{Bernoulli}(\frac{n}{k}, \frac{b \log n}{n})$ for $s \neq 0$ and $A_0^s \sim \text{Bernoulli}(\frac{n}{k} - 1, \frac{a \log n}{n})$.

For the general case, we can write:

$$H(\bar{\sigma}) - H(X) = (1 + \gamma \frac{\log n}{n})[A_{\bar{\sigma}} - B_{\bar{\sigma}}] + \gamma \frac{\log n}{n} N_{\bar{\sigma}} \quad (24)$$

in which we use $A_{\bar{\sigma}}$ or $B_{\bar{\sigma}}$ to represent the binomial random variable with parameter $\frac{a \log n}{n}$ or $\frac{b \log n}{n}$, respectively, and $N_{\bar{\sigma}}$ is a deterministic positive number depending on $\bar{\sigma}$ but irrelevant with the graph structure. The following lemma gives the expression of $A_{\bar{\sigma}}$, $B_{\bar{\sigma}}$ and $N_{\bar{\sigma}}$:

Lemma 6. For $\text{SSBM}(n, k, p, q)$, we assume $\bar{\sigma}$ differs from the ground truth label vector X in the $|\mathcal{I}| := \text{dist}(\bar{\sigma}, X)$ coordinate. Let $I_{ij} = |\{r \in [n] | X_r = w^i, \bar{\sigma}_r = w^j\}|$ for $i \neq j$ and $I_{ii} = 0$. We further denote the row sum as $I_i = \sum_{j=0}^{k-1} I_{ij}$ and the column sum as $I'_i = \sum_{j=0}^{k-1} I_{ji}$. Then:

$$N_{\bar{\sigma}} = \frac{1}{2} \sum_{i=0}^{k-1} (I_i - I'_i)^2 \quad (25)$$

$$B_{\bar{\sigma}} \sim \text{Bernoulli}(\frac{n}{k} |\mathcal{I}| + \frac{1}{2} \sum_{i=0}^{k-1} (-2I'_i I_i + I_i^2 - \sum_{j=0}^{k-1} I_{ji}^2), q) \quad (26)$$

$$A_{\bar{\sigma}} \sim \text{Bernoulli}(\frac{n}{k} |\mathcal{I}| - \frac{1}{2} \sum_{i=0}^{k-1} (I_i^2 + \sum_{j=0}^{k-1} I_{ij}^2), p) \quad (27)$$

The proof of Lemma 6 is mainly composed of careful counting techniques, and we omit it here. When $|\mathcal{I}|$ is small compared to n , we have the following Lemma, which is an extension of Proposition 6 in [12].

Lemma 7. For $t \in [\frac{1}{k}(b-a), 0]$ and $|\mathcal{I}| \leq n/\sqrt{\log n}$

$$\begin{aligned} & P_G(B_{\bar{\sigma}} - A_{\bar{\sigma}} \geq t|\mathcal{I}| \log n) \\ & \leq \exp\left(|\mathcal{I}| \log n \left(f_{\beta}(t) - \beta t - 1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right)\right) \end{aligned} \quad (28)$$

where $f_{\beta}(t) = \min_{s \geq 0}(g(s) - st) + \beta t \leq \tilde{g}(\beta)$.

Corresponding to the three cases of Theorem 2, we use three non-trivial lemmas to establish the properties of the Ising model.

Lemma 8. Let $\gamma > b$. When $\text{dist}(\bar{\sigma}, X) \geq \frac{n}{\sqrt{\log n}}$ and $D(\bar{\sigma}, X)$, the event $P_{\sigma|G}(\sigma = \bar{\sigma}) > \exp(-Cn)P_{\sigma|G}(\sigma = X)$ happens with a probability (with respect to SSBM) less than $\exp(-\tau(\alpha, \beta)n\sqrt{\log n})$, where C is an arbitrary constant and $\tau(\alpha, \beta)$ is a positive number.

Proof. We denote the event $P_{\sigma|G}(\sigma = \bar{\sigma}) > \exp(-Cn)P_{\sigma|G}(\sigma = X)$ as $\tilde{D}(\bar{\sigma}, C)$. By Equation (24), $\tilde{D}(\bar{\sigma}, C)$ is equivalent to:

$$\left(1 + \frac{\gamma \log n}{n}\right)[B_{\bar{\sigma}} - A_{\bar{\sigma}}] > \frac{\gamma \log n}{n}N_{\bar{\sigma}} - \frac{C}{\beta}n \quad (29)$$

We claim that $\bar{\sigma}$ must satisfy at least one of the following two conditions:

1. $\exists i \neq j$ s.t. $\frac{1}{k(k-1)} \frac{n}{\sqrt{\log n}} \leq I_{ij} \leq \frac{n}{k} - \frac{1}{k(k-1)} \frac{n}{\sqrt{\log n}}$
2. $\exists i \neq j$ s.t. $I_{ij} > \frac{n}{k} - \frac{1}{k(k-1)} \frac{n}{\sqrt{\log n}}$ and $I_{ji} < \frac{1}{k(k-1)} \frac{n}{\sqrt{\log n}}$

If neither of the above two condition holds, then from condition 1 we have $I_{ij} < \frac{1}{k(k-1)} \frac{n}{\sqrt{\log n}}$ or $I_{ij} > \frac{n}{k} - \frac{1}{k(k-1)} \frac{n}{\sqrt{\log n}}$ for any $0 \leq i, j \leq k-1$. Since $\sum_{i,j} I_{ij} = |\mathcal{I}| \geq \frac{n}{\sqrt{\log n}}$, there exists i, j such that $I_{ij} > \frac{n}{k} - \frac{1}{k(k-1)} \frac{n}{\sqrt{\log n}}$. Under such conditions, we also assume $I_{ji} > \frac{n}{k} - \frac{1}{k(k-1)} \frac{n}{\sqrt{\log n}}$. Let X' be the vector that exchanges the value of w^i with w^j in X . We consider:

$$\begin{aligned} \text{dist}(\bar{\sigma}, X') - \text{dist}(\bar{\sigma}, X) &= |\{r \in [n] | X_r = w^i, \bar{\sigma}_r \neq w^j\}| \\ &\quad + |\{r \in [n] | X_r = w^j, \bar{\sigma}_r \neq w^i\}| \\ &\quad - |\{r \in [n] | X_r = w^i, \bar{\sigma}_r \neq w^i\}| \\ &\quad - |\{r \in [n] | X_r = w^j, \bar{\sigma}_r \neq w^j\}| \\ &= \frac{n}{k} - I_{ij} + \frac{n}{k} - I_{ji} - I_i - I_j \\ &< \frac{2}{k(k-1)} \frac{n}{\sqrt{\log n}} - I_i - I_j < 0 \end{aligned} \quad (30)$$

which contracts with the fact that $\bar{\sigma}$ is nearest to X . Therefore, we should have $I_{ji} < \frac{1}{k(k-1)} \frac{n}{\sqrt{\log n}}$. Now the (i, j) pair satisfies condition 2, which contracts with the fact that $\bar{\sigma}$ satisfies neither of the two conditions.

Under condition 1, we can get a lower bound on $|A_{\bar{\sigma}}|$ from Equation (27). Let $I'_{ij} = I_{ij}$ for $i \neq j$ and $I'_{ii} = \frac{n}{k} - I_i$. Then we can simplify $|A_{\bar{\sigma}}|$ as:

$$\begin{aligned} |A_{\bar{\sigma}}| &= \frac{n}{k} |\mathcal{I}| - \frac{1}{2} \sum_{i=0}^{k-1} (I_i^2 + \sum_{j=0}^{k-1} I_{ij}^2) \\ &= \frac{n^2}{2k} - \frac{1}{2} \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} I_{ij}^2 \end{aligned}$$

We further have $\sum_{i=0}^{k-1} \sum_{j=0}^{k-1} I_{ij}^2 \leq (k-1) \frac{n^2}{k^2} + (\frac{n}{k} - I_{ij})^2 + I_{ij}^2$ where I_{ij} satisfies condition 1. Therefore, $\sum_{i=0}^{k-1} \sum_{j=0}^{k-1} I_{ij}^2 \leq (k-1) \frac{n^2}{k^2} + (\frac{1}{k(k-1)} \frac{n}{\sqrt{\log n}})^2 + (\frac{n}{k} - \frac{1}{k(k-1)} \frac{n}{\sqrt{\log n}})^2 = \frac{n^2}{k} - \frac{2n^2}{k^2(k-1)\sqrt{\log n}}(1 + o(1))$. As a result,

$$A_{\bar{\sigma}} \geq \frac{n^2}{k^2(k-1)\sqrt{\log n}}(1 + o(1)) \quad (31)$$

Under condition 2, we can get a lower bound on $N_{\bar{\sigma}}$. Since $\text{dist}(\bar{\sigma}, X') - \text{dist}(\bar{\sigma}, X) \geq 0$, from (30) we have $I_{ij} + I_{ji} + I_i + I_j \leq \frac{2n}{k}$. Since $I_i \geq I_{ij} > \frac{n}{k} - \frac{1}{k(k-1)} \frac{n}{\sqrt{\log n}}$, we have $I_j \leq \frac{2}{k(k-1)} \frac{n}{\sqrt{\log n}}$. Now consider $I'_j - I_j \geq \frac{n}{k} - \frac{3}{k(k-1)} \frac{n}{\sqrt{\log n}}$. From (25): $N_{\bar{\sigma}} \geq \frac{1}{2} (\frac{n}{k} - \frac{3}{k(k-1)} \frac{n}{\sqrt{\log n}})^2 = \frac{n^2}{2k^2} (1 + o(1))$.

Now we use the Chernoff inequality to bound Equation (29); we can omit $\frac{\gamma \log n}{n}$ on the left-hand side since it is far smaller than 1. Let $Z \sim \text{Bernoulli}(\frac{a \log n}{n})$, $Z' \sim \text{Bernoulli}(\frac{b \log n}{n})$, then:

$$\begin{aligned} P_G(\tilde{D}(\bar{\sigma}, C)) &\leq (\mathbb{E}[\exp(sZ)])^{|B_{\bar{\sigma}}|} (\mathbb{E}[\exp(-sZ')])^{|A_{\bar{\sigma}}|} \\ &\cdot \exp(-s(\frac{\gamma \log n}{n} N_{\bar{\sigma}} - \frac{C}{\beta} n)) \\ &\leq \exp\left(|B_{\bar{\sigma}}| \frac{b \log n}{n} (e^s - 1) + |A_{\bar{\sigma}}| \frac{a \log n}{n} (e^{-s} - 1) - s(\frac{\gamma \log n}{n} N_{\bar{\sigma}} - \frac{C}{\beta} n)\right) \end{aligned}$$

Using $|B_{\bar{\sigma}}| = N_{\bar{\sigma}} + |A_{\bar{\sigma}}|$ we can further simplify the exponential term as:

$$\frac{\log n}{n} [|A_{\bar{\sigma}}| (b(e^s - 1) + a(e^{-s} - 1)) + N_{\bar{\sigma}} (b(e^s - 1) - \gamma s)] + s \frac{C}{\beta} n$$

Now we investigate the function $g_1(s) = b(e^s - 1) + a(e^{-s} - 1)$ and $g_2(s) = b(e^s - 1) - \gamma s$. Both functions take zero values at $s = 0$ and $g'_1(s) = (be^s - ae^{-s})$, $g'_2(s) = be^s - \gamma$. Therefore, $g'_1(0) = b - a < 0$, $g'_2(0) = b - \gamma < 0$ and we can choose $s^* > 0$ such that $g_1(s^*) < 0$, $g_2(s^*) < 0$. To compensate the influence of the term sCn/β we only need to make sure that the order of $\frac{\log n}{n} \min\{|A_{\bar{\sigma}}|, N_{\bar{\sigma}}\}$ is larger than n . This requirement is satisfied since either $|A_{\bar{\sigma}}| \geq \frac{n^2}{k^2(k-1)\sqrt{\log n}}(1 + o(1))$ or $N_{\bar{\sigma}} \geq \frac{n^2}{2k^2}(1 + o(1))$. $\square$

Lemma 9. If $\gamma > b$, $\beta > \beta^*$, For $1 \leq r \leq \frac{n}{\sqrt{\log n}}$ and $\forall \epsilon > 0$, there is a set $\mathcal{G}^{(r)}$ such that:

$$P_G(\mathcal{G}_n^{(r)}) \geq 1 - n^{r(\bar{\delta}(\beta)/2 + \epsilon)} \quad (32)$$

and for every $G \in \mathcal{G}_n^{(r)}$,

$$\frac{P_{\sigma|G}(\text{dist}(\sigma, X) = r | D(\sigma, X))}{P_{\sigma|G}(\sigma = X | D(\sigma, X))} < n^{r\tilde{g}(\beta)/2} \quad (33)$$

For $r > \frac{n}{\sqrt{\log n}}$, there is a set $\mathcal{G}^{(r)}$ such that:

$$P(G \in \mathcal{G}_n^{(r)}) \geq 1 - e^{-n} \quad (34)$$

and for every $G \in \mathcal{G}_n^{(r)}$,

$$\frac{P_{\sigma|G}(\text{dist}(\sigma, X) = r | D(\sigma, X))}{P_{\sigma|G}(\sigma = X | D(\sigma, X))} < e^{-n} \quad (35)$$

Proof. We distinguish the discussion between two cases: $r \leq \frac{n}{\sqrt{\log n}}$ and $r > \frac{n}{\sqrt{\log n}}$.

When $r \leq \frac{n}{\sqrt{\log n}}$, we can show that $\text{dist}(\sigma, X) = r$ implies $D(\sigma, X)$ by using the triangle inequality of dist . For $f \in S_k \setminus \{\text{id}\}$, where id is the identity mapping, we have:

$$\frac{2n}{k} \leq \text{dist}(f(X), X) \leq \text{dist}(\sigma, f(X)) + \text{dist}(\sigma, X)$$

Therefore, $\text{dist}(\sigma, f(X)) \geq \frac{2n}{k} - \frac{n}{\sqrt{\log n}} \geq \text{dist}(\sigma, X)$ and Equation (33) is equivalent with:

$$\frac{P_{\sigma|G}(\text{dist}(\sigma, X) = r)}{P_{\sigma|G}(\sigma = X)} < n^{r\tilde{g}(\beta)/2} \quad (36)$$

The left-hand side can be written as:

$$\begin{aligned} \frac{P_{\sigma|G}(\text{dist}(\sigma, X) = r)}{P_{\sigma|G}(\sigma = X)} &= \sum_{\text{dist}(\bar{\sigma}, X)=r} \exp(-\beta(H(\bar{\sigma}) - H(X))) \\ &\text{by (24)} \leq \sum_{\text{dist}(\bar{\sigma}, X)=r} \exp(\beta_n(B_{\bar{\sigma}} - A_{\bar{\sigma}})) \end{aligned}$$

where $\beta_n = \beta(1 + \gamma \frac{\log n}{n})$.

Define $\Xi_n(r) := \sum_{\text{dist}(\bar{\sigma}, X)=r} \exp(\beta_n(B_{\bar{\sigma}} - A_{\bar{\sigma}}))$ and we only need to show that:

$$P_G(\Xi_n(r) \geq n^{r\tilde{g}(\beta)/2}) \leq n^{r(\tilde{g}(\beta)/2 + \epsilon)} \quad (37)$$

Define the event $\Lambda_n(G, r) := \{B_{\bar{\sigma}} - A_{\bar{\sigma}} < 0, \forall \bar{\sigma} \text{ s.t. } \text{dist}(\bar{\sigma}, X) = r\}$, and we proceed as follows:

$$\begin{aligned} P_G(\Xi_n(r) \geq n^{r\tilde{g}(\beta)/2}) &\leq P_G(\Lambda_n(G, r)^c) \\ &\quad + P_G(\Xi_n(r) \geq n^{r\tilde{g}(\beta)/2} | \Lambda_n(G, r)) \end{aligned}$$

For the first term, since $|\{\bar{\sigma} | \text{dist}(\bar{\sigma}, X) = r\}| = (k-1)^r n^r$, by Lemma 7, $P_G(\Lambda_n(G, r)^c) \leq (k-1)^r n^{r\tilde{g}(\beta)} \leq n^{r(\tilde{g}(\beta)/2 + \epsilon/2)}$. For the second term, we use Markov inequality:

$$P_G(\Xi_n(r) \geq n^{r\tilde{g}(\beta)/2} | \Lambda_n(G, r)) \leq \mathbb{E}[\Xi_n(r) | \Lambda_n(G, r)] n^{-r\tilde{g}(\beta)/2}$$

The conditional expectation can be estimated as follows:

$$\begin{aligned} \mathbb{E}[\Xi_n(r)|\Lambda_n(G, r)] &= \sum_{\text{dist}(\bar{\sigma}, X)=r} \sum_{tr \log n=-\infty}^{-1} \\ &P_G(B_{\bar{\sigma}} - A_{\bar{\sigma}} = tr \log n) \exp(\beta_n tr \log n) \\ &\leq (k-1)^r n^{r+r\beta_n(b-a)/k} + \sum_{\text{dist}(\bar{\sigma}, X)=r} \sum_{tr \log n=\frac{b-a}{k} \log n}^{-1} \\ &P_G(B_{\bar{\sigma}} - A_{\bar{\sigma}} = tr \log n) \exp(\beta_n tr \log n) \\ r + r\beta_n(b-a)/k &= f_{\beta_n}(\frac{b-a}{k}) < \tilde{g}(\beta_n), \text{ therefore, } (k-1)^r n^{r+r\beta_n(b-a)/k} n^{-r\tilde{g}(\beta)/2} \leq \\ &n^{r(\tilde{g}(\beta)/2+\epsilon/2)}. \text{ Using Lemma 7, we have:} \end{aligned}$$

$$P_G(B_{\bar{\sigma}} - A_{\bar{\sigma}} = tr \log n) \exp(\beta_n tr \log n) \leq n^{r(f_{\beta_n}(t)-1+O(\frac{1}{\sqrt{\log n}}))}$$

Since $\beta_n \rightarrow \beta, \forall \epsilon$, when n is sufficiently large we have $\tilde{g}(\beta_n) \leq \tilde{g}(\beta) + \epsilon/2$. Therefore,

$$\begin{aligned} &\sum_{\text{dist}(\bar{\sigma}, X)=r} \sum_{\substack{tr \log n= \\ r(b-a)/k \log n}}^{-1} P_G(B_{\bar{\sigma}} - A_{\bar{\sigma}} = t \log n) \exp(\beta_n tr \log n) \\ &\leq n^{r(\tilde{g}(\beta_n)-\tilde{g}(\beta)/2)} \\ &\leq n^{r(\tilde{g}(\beta)/2+\epsilon/2)} O(\log n) (k-1)^r \end{aligned}$$

Combining the above equations, we have:

$$\begin{aligned} P_G(\Xi_n(r) \geq n^{r\tilde{g}(\beta)/2}) &\leq n^{r(\tilde{g}(\beta)/2+\epsilon/2)} O(\log n) (k-1)^r \\ &\leq n^{r(\tilde{g}(\beta)/2+\epsilon)} \end{aligned}$$

When $r > \frac{n}{\sqrt{\log n}}$, using Lemma 8, we can choose a sufficiently large constant $C > 1$ such that $k^n \exp(-Cn) < e^{-n}$:

$$\begin{aligned} \frac{P_{\sigma|G}(\text{dist}(\sigma, X) = r | D(\sigma, X))}{P_{\sigma|G}(\sigma = X | D(\sigma, X))} &= \sum_{\substack{D(\sigma, X) \\ \text{dist}(\sigma, X)=r}} \frac{P_{\sigma|G}(\sigma = \bar{\sigma})}{P_{\sigma|X}(\sigma = X)} \\ &> \exp(-n) \end{aligned}$$

happens with probability less than e^{-n} . Therefore, Equation (35) holds. $\square$

If $\gamma > b$ and $\beta < \beta^*$, we have the following lemma:

Lemma 10. *If $\gamma > b$ and $\beta < \beta^*$, there is a set $\mathcal{G}_n^{(1)}$ such that $P_G(\mathcal{G}_n^{(1)}) \geq 1 - n^{g(\tilde{\beta})}$ and:*

$$\mathbb{E}[\sum_{r=1}^n \exp(\beta_n(A_r^s - A_r^0)) | G \in \mathcal{G}_n^{(1)}] = (1 + o(1)) n^{g(\beta_n)} \quad (38)$$

$$\text{Var}[\sum_{r=1}^n \exp(\beta_n(A_r^s - A_r^0)) | G \in \mathcal{G}_n^{(1)}] \leq (1 + o(1)) n^{g(2\beta_n)} \quad (39)$$

Lemma 10 is an extension of Proposition 10 in [12] and can be proved using almost the same analysis. Thus we omit the proof of Lemma 10 here.

Lemma 11. If $\gamma > b$ and $\beta < \beta^*$, there is a set $\mathcal{G}'_n$ such that:

$$P_G(\mathcal{G}'_n) \geq 1 - (1 + o(1)) \max\{n^{g(\bar{\beta})}, n^{\tilde{g}(2\beta_n) - 2g(\beta_n) + \epsilon}\} \quad (40)$$

and for every $G \in \mathcal{G}'_n$,

$$\frac{P_{\sigma|G}(\text{dist}(\sigma, X) = 1 | D(\sigma, X))}{P_{\sigma|G}(\sigma = X | D(\sigma, X))} \geq (1 + o(1)) n^{g(\beta_n)} \quad (41)$$

Proof. The left-hand side of Equation (41) can be rewritten as:

$$\frac{P_{\sigma|G}(\text{dist}(\sigma, X) = 1)}{P_{\sigma|G}(\sigma = X)} = (1 + o(1)) \sum_{s=1}^{k-1} \sum_{r=1}^n \exp(\beta_n(A_r^s - A_r^0)) \quad (42)$$

Let $\mathcal{G}_n^{(1)}$ be defined in Lemma 10 and $\mathcal{G}_n^{(2)} := \{|\sum_{r=1}^n \exp(\beta_n(A_r^s - A_r^0)) - (1 + o(1))n^{g(\beta_n)}| \leq n^{g(\beta_n) - \epsilon/2}\}$.

Using Chebyshev's inequality, we have:

$$P_G(G \notin \mathcal{G}_n^{(2)} | G \in \mathcal{G}_n^{(1)}) \leq n^{\tilde{g}(2\beta_n) - 2g(\beta_n) + \epsilon}$$

Let $\mathcal{G}'_n = \mathcal{G}_n^{(1)} \cap \mathcal{G}_n^{(2)}$:

$$\begin{aligned} P_G(G \in \mathcal{G}'_n) &= P_G(\mathcal{G}_n^{(1)}) P_G(G \in \mathcal{G}_n^{(2)} | G \in \mathcal{G}_n^{(1)}) \\ &\geq (1 - n^{\tilde{g}(2\beta_n) - 2g(\beta_n) + \epsilon}) (1 - n^{g(\bar{\beta})}) \\ &= 1 - (1 + o(1)) \max\{n^{g(\bar{\beta})}, n^{\tilde{g}(2\beta_n) - 2g(\beta_n) + \epsilon}\} \end{aligned}$$

and for every $G \in \mathcal{G}'_n$,

$$\sum_{r=1}^n \exp(\beta_n(A_r^s - A_r^0)) = (1 + o(1)) n^{g(\beta_n)}$$

Therefore, from Equation (42) we have:

$$\frac{P_{\sigma|G}(\text{dist}(\sigma, X) = 1)}{P_{\sigma|G}(\sigma = X)} \geq (1 + o(1)) n^{g(\beta_n)}$$

□

Let $\Lambda := \{\omega^j \cdot \mathbf{1}_n | j = 0, \dots, k-1\}$ where $\mathbf{1}_n$ is the all-ones vector with dimension n , and we have the following lemma:

Lemma 12. Suppose $\gamma < b$ and $\bar{\sigma}$ satisfies $\text{dist}(\bar{\sigma}, \mathbf{1}_n) \geq \frac{n}{\sqrt{\log n}}$ and $D(\bar{\sigma}, \mathbf{1}_n)$. Then the event $P_{\sigma|G}(\sigma = \bar{\sigma}) > \exp(-Cn) P_{\sigma|G}(\sigma = \mathbf{1}_n)$ happens with a probability (with respect to SSBM) less than $\exp(-\tau(\alpha, \beta)n\sqrt{\log n})$ where C is an arbitrary constant, $\tau(\alpha, \beta)$ is a positive number.

Proof. Let $n_r = |\{\bar{\sigma}_i = w^r | i \in [n]\}|$. Then $n_0 \geq n_r$ for $r = 1, \dots, k-1$ since $\arg \min_{\sigma' \in \Lambda} \text{dist}(\bar{\sigma}, \sigma') = \mathbf{1}_n$. Without loss of generality, we suppose $n_0 \geq n_1 \geq \dots \geq n_{k-1}$.

Define $N_{\bar{\sigma}} = \frac{1}{2}(n(n-1) - \sum_{r=0}^{k-1} n_r(n_r-1)) = \frac{1}{2}(n^2 - \sum_{r=0}^{k-1} n_r^2)$. Denote the event $P_{\sigma|G}(\sigma = \bar{\sigma}) > \exp(-Cn)P_{\sigma|G}(\sigma = \mathbf{1}_n)$ as $D'(\bar{\sigma}, C)$, which can be transformed as:

$$\left(1 + \frac{\gamma \log n}{n}\right) \left(\sum_{\bar{\sigma}_i \neq \bar{\sigma}_j, X_i = X_j} Z_{ij} + \sum_{\bar{\sigma}_i \neq \bar{\sigma}_j, X_i \neq X_j} Z_{ij} \right) \leq \frac{\gamma \log n}{n} N_{\bar{\sigma}} + \frac{C}{\beta} n \quad (43)$$

Firstly we estimate the order of $N_{\bar{\sigma}}$, and obviously $N_{\bar{\sigma}} \leq \frac{1}{2}n^2$. Using the conclusion in Appendix A of [18], we have:

$$\sum_{r=0}^{k-1} n_r^2 \leq \begin{cases} nn_0 & n_0 \leq \frac{n}{2} \\ n^2 - 2n_0(n - n_0) & n_0 > \frac{n}{2} \end{cases} \quad (44)$$

By assumption of $\text{dist}(\bar{\sigma}, \mathbf{1}_n) \geq \frac{n}{\sqrt{\log n}}$, we have $n_0 \leq n - \frac{n}{\sqrt{\log n}}$ and $n_0 \geq \frac{n}{k}$ follows from $n_0 \geq n_r$. When $n_0 > \frac{n}{2}$, we have $N_{\bar{\sigma}} \geq n_0(n - n_0) \geq \frac{n^2}{\sqrt{\log n}}(1 + o(1))$. The second inequality is achieved if $n_0 = n - \frac{n}{\sqrt{\log n}}$. When $n_0 < \frac{n}{2}$, $N_{\bar{\sigma}} \geq \frac{n^2 - nn_0}{2} \geq \frac{n^2}{4}$ and the second inequality is achieved when $n_0 = \frac{n}{2}$. Thus generally we have $\frac{n^2}{\sqrt{\log n}}(1 + o(1)) \leq N_{\bar{\sigma}} \leq \frac{n^2}{2}$.

Since $\frac{C}{\beta}n = o(\frac{\log n}{n}N_{\bar{\sigma}})$ we can rewrite Equation (43) as:

$$\left(\sum_{\substack{\bar{\sigma}_i \neq \bar{\sigma}_j \\ X_i = X_j}} -Z_{ij} + \sum_{\substack{\bar{\sigma}_i \neq \bar{\sigma}_j \\ X_i \neq X_j}} -Z_{ij} \right) \geq -\gamma \frac{\log n N_{\bar{\sigma}}}{n} (1 + o(1)) \quad (45)$$

Let $N_1 = \sum_{\bar{\sigma}_i \neq \bar{\sigma}_j, X_i = X_j} 1$ and $N_2 = \sum_{\bar{\sigma}_i \neq \bar{\sigma}_j, X_i \neq X_j} 1 = N_{\bar{\sigma}} - N_1$.

Using the Chernoff inequality we have:

$$\begin{aligned} P_G(D'(\bar{\sigma}, C)) &\leq (\mathbb{E}[\exp(-sZ)])^{N_1} (\mathbb{E}[\exp(-sZ')])^{N_2} \\ &\cdot \exp\left(\gamma \frac{\log n N_{\bar{\sigma}} s}{n} (1 + o(1))\right) \\ &= \exp\left(\frac{\log n}{n} (1 + o(1)) (e^{-s} - 1) (aN_1 + bN_2)\right. \\ &\quad \left.+ \gamma \frac{\log n N_{\bar{\sigma}} s}{n} (1 + o(1))\right) \end{aligned}$$

Since $s > 0$ and $a > b$, we further have:

$$P_G(D'(\bar{\sigma}, C)) \leq \exp\left(\frac{N_{\bar{\sigma}} \log n}{n} (b(e^{-s} - 1) + \gamma s + o(1))\right)$$

Let $h_b(x) = x - b - x \log \frac{x}{b}$, which satisfies $h_b(x) < 0$ for $0 < x < b$, and take $s = -\log \frac{\gamma}{b} > 0$, using $N_{\bar{\sigma}} \geq \frac{n^2}{\sqrt{\log n}}$ we have:

$$\begin{aligned} P_G(D'(\bar{\sigma}, C)) &\leq \exp\left(N_{\bar{\sigma}} \frac{\log n}{n} h_b(\gamma) (1 + o(1))\right) \\ &\leq \exp(h_b(\gamma) n \sqrt{\log n} (1 + o(1))) \end{aligned}$$

□

Proof of Theorem 2. (1) Since $P_\sigma(\sigma \notin S_k(X)) = \sum_{f \in S_k} P_\sigma(\sigma \neq f(X) | D(\sigma, f(X)))$
 $P_\sigma(D(\sigma, f(X)))$, we only need to establish $P_\sigma(\sigma \neq X | D(\sigma, X)) \leq n^{\tilde{g}(\beta)/2+\epsilon}$. From Lemma 9,
we can find $\mathcal{G}_n^{(r)}$ for $r = 1, \dots, n$. Let $\mathcal{G}'_n = \cap_{r=1}^n \mathcal{G}_n^{(r)}$ and choose $\frac{\epsilon}{2}$; from Equations (32) and
(34), we have:

$$\begin{aligned} P_G(\mathcal{G}_n^{lc}) &= P_G(\cup_{r=1}^n (\mathcal{G}_n^{(r)})^c) \\ &\leq \sum_{r=1}^{n/\sqrt{\log n}} n^{r(\tilde{g}(\beta)/2+\epsilon/2)} + ne^{-n} \\ &\leq \frac{1}{2} n^{\tilde{g}(\beta)/2+\epsilon} \end{aligned}$$

where the last equality follows from the estimation of sum of geometric series. On the other
hand, for every $G \in \mathcal{G}'_n$, from Equations (33) and (35), we have:

$$\begin{aligned} \frac{P_{\sigma|G}(\sigma \neq X | D(\sigma, X))}{1 - P_{\sigma|G}(\sigma \neq X | D(\sigma, X))} &= \frac{P_{\sigma|G}(\sigma \neq X | D(\sigma, X))}{P_{\sigma|G}(\sigma = X | D(\sigma, X))} \\ &< \sum_{r=1}^{n/\sqrt{\log n}} n^{r\tilde{g}(\beta)/2} + ne^{-n} \end{aligned}$$

from which we can get the estimation $P_{\sigma|G}(\sigma \neq X | D(\sigma, X)) \leq \frac{1}{2} n^{\tilde{g}(\beta)/2+\epsilon}$. Finally,

$$\begin{aligned} P_\sigma(\sigma \neq X | D(\sigma, X)) &= \sum_{G \in \mathcal{G}'_n} P_G(G) P_{\sigma|G}(\sigma \neq X | D(\sigma, X)) \\ &\quad + P_G(\mathcal{G}_n^{lc}) \leq n^{\tilde{g}(\beta)/2+\epsilon}. \end{aligned}$$

(2) When $\beta < \beta^*$, using Lemma 11, for every $G \in \mathcal{G}'_n$ we can obtain:

$$\frac{1 - P_{\sigma|G}(\sigma = X | D(\sigma, X))}{P_{\sigma|G}(\sigma = X | D(\sigma, X))} \geq (1 + o(1)) n^{g(\beta_n)}$$

We then have:

$$P_{\sigma|G}(\sigma = X | D(\sigma, X)) \leq (1 + o(1)) n^{-g(\beta_n)}$$

Then:

$$\begin{aligned} P_\sigma(\sigma = X | D(\sigma, X)) &\leq P(\mathcal{G}_n^{lc}) \\ &\quad + \sum_{G \in \mathcal{G}'_n} P_G(G) P_{\sigma|G}(\sigma = X | D(\sigma, X)) \\ &\leq (1 + o(1)) n^{-g(\beta_n)} + (1 + o(1)) \max\{n^{g(\tilde{\beta})}, n^{\tilde{g}(2\beta_n)-2g(\beta_n)+\epsilon}\} \\ &\leq (1 + o(1)) \max\{n^{g(\tilde{\beta})}, n^{-g(\beta)+\epsilon}\} \end{aligned}$$

(3) When $\gamma < b$, for any $f \in S_k$, we have $\text{dist}(f(X), \Lambda) = \frac{(k-1)n}{k} > \frac{n}{\sqrt{\log n}}$. Therefore,
using Lemma 12, we can find a graph $\mathcal{G}'_n$ such that $P_C(\mathcal{G}'_n) \leq \exp(-nC)$ and for any
 $G \notin \mathcal{G}'_n$, $P_{\sigma|G}(\sigma = f(X)) \leq \exp(-Cn)$. Therefore,

$$P_a(\hat{X}^*) \leq P_G(\mathcal{G}'_n) + k! \exp(-Cn) = (1 + k!) \exp(-Cn)$$

The conclusion of $P_a(\hat{X}^*) \leq \exp(-Cn)$ follows since C can take any positive value. $\square$

8.3. Proof of Theorem 3

Lemma 13 (Lemma 8 of [7]). Let m be a positive number larger than n . When $Z_1, \dots, Z_m$ are i.i.d. Bernoulli($\frac{b \log n}{n}$) and $W_1, \dots, W_m$ are i.i.d. Bernoulli($\frac{a \log n}{n}$), independent of $Z_1, \dots, Z_m$, then:

$$P\left(\sum_{i=1}^m (Z_i - W_i) \geq 0\right) \leq \exp\left(-\frac{m \log n}{n} (\sqrt{a} - \sqrt{b})^2\right) \quad (46)$$

Proof of Theorem 3. Let $P_F^{(r)}$ denote the probability when there is σ satisfying $\text{dist}(\sigma, X) = r$ and $H(\sigma) < H(X)$.

From Equation (23), when σ differs from X only by one coordinate, from Lemma 13 the probability for $H(\sigma) < H(X)$ is bounded by $P_G(A_r^s - A_r^0 > 0) \leq n^{-\frac{(\sqrt{a}-\sqrt{b})^2}{k}}$. Therefore, $P_F^{(1)} \leq (k-1)n^{g(\tilde{\beta})}$. Using Lemma 7, we can get $P_F^{(r)} \leq (k-1)^r n^{rg(\tilde{\beta})}$ for $r \leq \frac{n}{\sqrt{\log n}}$. For $r \geq \frac{n}{\sqrt{\log n}}$, choosing $C = 0$ in Lemma 8 we can get $\sum_{r \geq \frac{n}{\sqrt{\log n}}} P_F^{(r)} \leq e^{-n}$. That is, the dominant term is $\sum_{r \leq \frac{n}{\sqrt{\log n}}} P_F^{(r)}$ since the other part decreases exponentially fast. Therefore, the upper bound for error rate of $\hat{X}'$ is:

$$\begin{aligned} P_F &= \sum_{r=1}^n P_F^{(r)} \leq (1+o(1)) \sum_{r=1}^{\infty} (k-1)^r n^{rg(\tilde{\beta})} \\ &\leq (1+o(1)) \frac{(k-1)n^{g(\tilde{\beta})}}{1 - (k-1)n^{g(\tilde{\beta})}} = (k-1+o(1))n^{g(\tilde{\beta})} \end{aligned}$$

When $\sigma \in W^*$, since $|\{r \in [n] | \sigma_r = \omega^i\}| = I'_i + \frac{n}{k} - I_i$, we have $I'_i = I_i$. From Lemma 6, we can see $|B_{\tilde{\sigma}}| = |A_{\tilde{\sigma}}|$ and $N_{\tilde{\sigma}} = 0$. Then from Equation (24), $H(\tilde{\sigma}) < H(X)$ is equivalent with $B_{\tilde{\sigma}} > A_{\tilde{\sigma}}$.

Therefore, when $\text{dist}(\tilde{\sigma}, X) \geq \frac{n}{\sqrt{\log n}}$ and $D(\sigma, X)$, from Equation (31), we have $A_{\tilde{\sigma}} \geq \frac{n^2}{k^2(k-1)\sqrt{\log n}}(1+o(1))$. We use Lemma 13 by letting $m = |A_{\tilde{\sigma}}|$ when $m \geq \frac{n}{\sqrt{\log n}}$; the error term is bounded by $\sum_{r \geq \frac{n}{\sqrt{\log n}}} P_F^{(r)} \leq \sum_{r \geq \frac{n}{\sqrt{\log n}}} (k-1)^r \exp\left(-\frac{(\sqrt{a}-\sqrt{b})^2}{k^2(k-1)} n \sqrt{\log n}\right) \leq \exp(-n)$, which decreases exponentially fast. For $m < \frac{n}{\sqrt{\log n}}$, we can use Lemma 7 directly by considering $\sum_{r=2}^{\frac{n}{\sqrt{\log n}}} P_F^{(r)}$. The summation starts from $r = 2$ since $\sigma \in W^*$. Therefore,

$$\begin{aligned} P_F &= \sum_{r=2}^n P_F^{(r)} \leq (1+o(1)) \sum_{r=2}^{\infty} (k-1)^r n^{rg(\tilde{\beta})} \\ &\leq ((k-1)^2 + o(1))n^{2g(\tilde{\beta})}. \end{aligned}$$

□

Author Contributions: The work of this paper was conceptualized by M.Y., who also developed the mathematical analysis methodology for the proof of the main theorem of this paper. All simulation code was written by F.Z., who also extended the idea of M.Y., concreted the proof and wrote this paper. The work is supervised and funded by S.-L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Natural Science Foundation of China under Grant 61807021, in part by the Shenzhen Science and Technology Program under Grant KQTD20170810150821146, and in part by the Innovation and Entrepreneurship Project for Overseas High-Level Talents of Shenzhen under Grant KQJSCX20180327144037831.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SBM	Stochastic Block Model
SSBM	Symmetric Stochastic Block Model
MSE	Mean Squared Error
ML	maximum likelihood
MQ	maximum modularity

References

- Fortunato, S. Community detection in graphs. *Phys. Rep.* **2010**, *486*, 75–174. [\[CrossRef\]](#)
- Feng, H.; Tian, J.; Wang, H.J.; Li, M. Personalized recommendations based on time-weighted overlapping community detection. *Inf. Manag.* **2015**, *52*, 789–800. [\[CrossRef\]](#)
- Hendrickson, B.; Kolda, T.G. Graph partitioning models for parallel computing. *Parallel Comput.* **2000**, *26*, 1519–1534. [\[CrossRef\]](#)
- Cline, M.S.; Smoot, M.; Cerami, E.; Kuchinsky, A.; Landys, N.; Workman, C.; Christmas, R.; Avila-Campilo, I.; Creech, M.; Gross, B.; et al. Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2007**, *2*, 2366. [\[CrossRef\]](#) [\[PubMed\]](#)
- Holland, P.W.; Laskey, K.B.; Leinhardt, S. Stochastic blockmodels: First steps. *Soc. Netw.* **1983**, *5*, 109–137. [\[CrossRef\]](#)
- Abbe, E. Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **2017**, *18*, 6446–6531.
- Abbe, E.; Bandeira, A.S.; Hall, G. Exact recovery in the stochastic block model. *IEEE Trans. Inf. Theory* **2015**, *62*, 471–487. [\[CrossRef\]](#)
- Mossel, E.; Neeman, J.; Sly, A. Consistency thresholds for the planted bisection model. *Electron. J. Probab.* **2016**, *21*, 24. [\[CrossRef\]](#)
- Abbe, E.; Sandon, C. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science, Berkeley, CA, USA, 17–20 October 2015; pp. 670–688.
- Nowicki, K.; Snijders, T.A.B. Estimation and prediction for stochastic blockstructures. *J. Am. Stat. Assoc.* **2001**, *96*, 1077–1087. [\[CrossRef\]](#)
- Ising, E. Beitrag zur theorie des ferromagnetismus. *Z. Für Phys.* **1925**, *31*, 253–258. [\[CrossRef\]](#)
- Ye, M. Exact recovery and sharp thresholds of Stochastic Ising Block Model. *arXiv* **2020**, arXiv:2004.05944.
- Liu, J.; Liu, T. Detecting community structure in complex networks using simulated annealing with k-means algorithms. *Phys. A Stat. Mech. Appl.* **2010**, *389*, 2300–2309. [\[CrossRef\]](#)
- Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092. [\[CrossRef\]](#)
- Potts, R.B. Some generalized order-disorder transformations. In *Mathematical Proceedings of the Cambridge Philosophical Society*; Cambridge University Press: Cambridge, UK 1952; Volume 48, pp. 106–109.
- Liu, L. On the Log Partition Function of Ising Model on Stochastic Block Model. *arXiv* **2017**, arXiv:1710.05287.
- Berthet, Q.; Rigollet, P.; Srivastava, P. Exact recovery in the Ising blockmodel. *Ann. Stat.* **2019**, *47*, 1805–1834. [\[CrossRef\]](#)
- Chen, Y.; Suh, C.; Goldsmith, A.J. Information recovery from pairwise measurements. *IEEE Trans. Inf. Theory* **2016**, *62*, 5881–5905. [\[CrossRef\]](#)
- Hajek, B.; Wu, Y.; Xu, J. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Trans. Inf. Theory* **2016**, *62*, 2788–2797. [\[CrossRef\]](#)
- Saad, H.; Nosratinia, A. Community detection with side information: Exact recovery under the stochastic block model. *IEEE J. Sel. Top. Signal Process.* **2018**, *12*, 944–958. [\[CrossRef\]](#)
- Newman, M.E. Equivalence between modularity optimization and maximum likelihood methods for community detection. *Phys. Rev. E* **2016**, *94*, 052315. [\[CrossRef\]](#)
- He, J.; Chen, D.; Sun, C. A fast simulated annealing strategy for community detection in complex networks. In Proceedings of the 2016 2nd IEEE International Conference on Computer and Communications (ICCC), Chengdu, China, 14–17 October 2016; pp. 2380–2384.
- Clauset, A.; Newman, M.E.; Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **2004**, *70*, 066111. [\[CrossRef\]](#)
- Diaconis, P.; Saloff-Coste, L. What do we know about the Metropolis algorithm? *J. Comput. Syst. Sci.* **1998**, *57*, 20–36. [\[CrossRef\]](#)
- Frigessi, A.; Martinelli, F.; Stander, J. Computational Complexity of Markov Chain Monte Carlo Methods for Finite Markov Random Fields. *Biometrika* **1997**, *84*, 1–18. [\[CrossRef\]](#)
- Erdős, P.; Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci* **1960**, *5*, 17–60.
- Holland, P.W.; Leinhardt, S. A Method for Detecting Structure in Sociometric Data. *Am. J. Sociol.* **1970**, *76*, 492–513. [\[CrossRef\]](#)