An Optimal Clustering Algorithm for the Labeled Stochastic Block Model

Kaito Ariu[♣][◇] **Se-Young Yun**[♠] **Alexandre Proutiere**[◇]

[♣]CyberAgent, Inc.
[◇]KTH Royal Institute of Technology
[♠]Graduate School of AI, KAIST

Abstract

This paper considers the clustering problem in the Labeled Stochastic Block Model (LSBM) from the observations of labels. For this model, we assume that the cluster size increases linearly with the number of nodes n. Our goal is to develop an efficient algorithm to identify the clusters based on the observed labels. We reexamine instance-specific lower bounds on the expected number of misclassified items. These bounds must be satisfied by any clustering algorithm. We propose Instance-Adaptive Clustering (IAC), the first algorithm that matches the lower bounds in expectation. IAC combines a one-time spectral clustering method with an iterative likelihood-based cluster assignment refinement procedure. This technique relies on the instance-specific lower bound and does not necessitate any model parameters, including the number of clusters. IAC retains an overall computational complexity of $\mathcal{O}(n\text{polylog}(n))$. We demonstrate the efficacy of our approach through numerical experiments.

1 Introduction

Community detection or clustering refers to the task of gathering similar items into a few groups from the data that, most often, correspond to observations of pair-wise interactions between items Newman and Girvan [2004]. A benchmark commonly used to assess the performance of clustering algorithms is the celebrated Stochastic Block Model (SBM) Holland et al. [1983], where pair-wise interactions are represented by a random graph. In this graph, the vertices correspond to items, and the presence of an edge between two items indicates their interaction. The SBM has been extensively studied over the last two decades; for a recent survey, see Abbe [2018]. However, it provides a relatively simplistic view of how items may interact. In real applications, interactions can be of different types (e.g., represented by ratings in recommender systems or a level of proximity between users in a social network). To capture this richer information about item interactions, the Labeled Stochastic Block Model (LSBM), proposed and analyzed in Heimlicher et al. [2012], Lelarge et al. [2013], Yun and Proutiere [2016], describes interactions by labels drawn from an arbitrary collection. The objective of this paper is to devise a clustering algorithm that, based on the observation of these labels, reconstructs the clusters of items while minimizing the expected number of misclassified items. In the following, we formally introduce LSBMs and outline our results.

The Labeled Stochastic Block Model. In the LSBM, the set $\mathcal{I}$ consisting of n items or nodes is randomly partitioned into K unknown disjoint clusters $\mathcal{I}_1, \dots, \mathcal{I}_K$. The cluster index of the item i is denoted by $\sigma(i)$. Let $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_K)$ represent the probabilities of items belonging to each cluster, i.e., for all $k \in [K]$ and $i \in \mathcal{I}$, $\mathbb{P}(i \in \mathcal{I}_k) = \alpha_k$. We assume that $\alpha_1, \dots, \alpha_K$ are strictly positive constants and that K and α are fixed as n grows large. Without loss of generality, we also assume that $\alpha_1 \leq \dots \leq \alpha_K$. Let $\mathcal{L} = \{0, 1, \dots, L\}$ be the finite set of labels. For each

ICML 2023 Workshop: Sampling and Optimization in Discrete Space (SODS 2023).

edge $(v, w) \in \mathcal{I}_i \times \mathcal{I}_j$, the learner observes the label ℓ with probability $p(i,j,\ell)$, independently of the labels observed in other edges. The number of clusters K is initially unknown. We have $\forall i,j \in [K]^2,\ \sum_{\ell\in[L]} p(i,j,\ell) = 1$. Without loss of generality, 0 is the most frequent label: $0 = \arg\max_{\ell} \sum_{i=1}^K \sum_{j=1}^K \alpha_i \alpha_j p(i,j,\ell)$. Let $\bar{p} = \max_{i,j,\ell\geq 1} p(i,j,\ell)$ be the maximum probability of observing a label different from 0. We will mostly consider the challenging sparse regime where $\bar{p} = \mathcal{O}((\log n)/n)$ and $\bar{p}n \to \infty$ as $n \to \infty$, but we will precise the assumptions made on n and $\bar{p}$ for each of our results. We further assume for all $i, j, k \in [K]$:

$$\text{(A1)} \quad \forall \ell \in \mathcal{L},\ \frac{p(i,j,\ell)}{p(i,k,\ell)} \leq \eta \quad \text{and} \quad \text{(A2)} \quad \frac{\sum_{k=1}^K (\sum_{\ell=1}^L (p(i,k,\ell) - p(j,k,\ell)))^2}{\bar{p}^2} \geq \varepsilon,$$

where η and ε are positive constants independent of n. (A1) imposes some homogeneity on the edge existence probability, and (A2) implies a certain separation among the clusters. In summary, the LSBM is parametrized by α and $p := (p(i,j,\ell))_{1\leq i,j\leq K, 0\leq \ell\leq L}$. We denote $p(i)$ as the $K \times (L+1)$ matrix whose element on j-th row and $(\ell+1)$-th column is $p(i,j,\ell)$ and denote $p(i,j) \in [0,1]^{L+1}$ the vector describing the probability of the label of a pair of items in $\mathcal{I}_i$ and $\mathcal{I}_j$.

Main results. We design a computationally efficient algorithm that recovers the clusters in the LSBM with a minimal error rate. By *minimal*, we mean that for any given LSBM, the algorithm achieves the best possible error rate for this specific LSBM. This contrasts with the minimax guarantees and demonstrates that the algorithm adapts to the hardness of the LSBM it is applied. We first present an instance-specific lower bound on the expected number of misclassified items satisfied by any algorithm. Let $\mathcal{P}^{K\times(L+1)}$ denote the set of all $K \times (L+1)$ matrices such that each row represents a probability distribution and define the divergence $D(\alpha, p)$ of the parameter (α, p) as follows:

$$D(\alpha, p) = \min_{i,j\in[K]: i\neq j} D_{L+}(\alpha, p(i), p(j)) \tag{1}$$

$$\text{with} \quad D_{L+}(\alpha, p(i), p(j)) = \min_{y\in\mathcal{P}^{K\times(L+1)}} \max\left\{ \sum_{k=1}^K \alpha_k \,\mathrm{kl}(y(k), p(i,k)), \sum_{k=1}^K \alpha_k \,\mathrm{kl}(y(k), p(j,k)) \right\},$$

where kl is the Kullback-Leibler divergence between two label distributions, i.e., $\mathrm{kl}(y(k), p(i,k)) = \sum_{\ell=0}^L y(k,\ell) \log \frac{y(k,\ell)}{p(i,k,\ell)}$. $D_{L+}(\alpha, p(i), p(j))$ can be interpreted as the hardness in distinguishing whether an item belongs to cluster i or cluster j based on the data. Consider a clustering algorithm π. Let $\varepsilon^\pi(n)$ denote the number of misclassified items for a given clustering algorithm π, with $\mathbb{E}[\varepsilon^\pi(n)]$ representing its expected value. This quantity is defined up to a permutation. Specifically, if π returns $(\hat{\mathcal{I}}_k)_k$, then $\varepsilon^\pi(n)$ is calculated as $\min_\theta |\cup_k \hat{\mathcal{I}}_k \backslash \mathcal{I}_{\theta(k)}|$, where θ denotes a permutation of $[K]$. To simplify the notation throughout the remainder of the paper, we assume that the permutation achieving the minimum is given by $\theta(k) = k$ for all $k \in [K]$. We present the following theorem that provides a lower bound on $\mathbb{E}[\varepsilon^\pi(n)]$:

Theorem 1.1. *Let $s = o(n)$. Under the assumptions of (A1), (A2), and $\bar{p}n = \omega(1)$, for any clustering algorithm π that satisfies $\limsup_{n\to\infty} \frac{\mathbb{E}[\varepsilon^\pi(n)]}{s} \leq 1$,*

$$\liminf_{n\to\infty} \frac{nD(\alpha, p)}{\log(n/s)} \geq 1. \tag{2}$$

The proof of Theorem 1.1 is based on the change-of-measure argument frequently used in online stochastic optimization and multi-armed bandit problems Lai and Robbins [1985], Kaufmann et al. [2016]. It is presented in Yun and Proutiere [2016] and in Appendix C for completeness. The main contribution of this paper is an algorithm with performance guarantees that match those of the above lower bound and with computational complexity scaling as $n$polylog(n). This algorithm, referred to as Instance-Adaptive Clustering (IAC) and presented in Section 3, first applies a spectral clustering algorithm to initially guess the clusters and then runs a *likelihood-based local improvement* algorithm to refine the estimated clusters. To analyze the performance of the algorithm, we make the following assumption.

$$\text{(A3)} \quad np(j,i,\ell) \geq (n\bar{p})^\kappa \text{ for all } i, j \text{ and } \ell \geq 1, \text{ for some constant } \kappa > 0.$$

Assumption (A3) excludes the existence of labels that are too sparse compared to $\bar{p}$. The following theorem establishes the optimality of IAC:

Theorem 1.2. *Assume that (A1), (A2), and (A3) hold, and that* $\bar{p} = O(\log n/n)$, $\bar{p}n = \omega(1)$. *Let* $s = o(n)$. *If the parameters* (α, p) *of the LSBM satisfy* (2)*, then IAC misclassifies at most* s *items in high probability and in expectation, i.e.,* $\lim_{n\to\infty} \mathbb{P}[\varepsilon^{\mathrm{IAC}}(n) \leq s] = 1$ *and* $\limsup_{n\to\infty} \frac{\mathbb{E}[\varepsilon^{\mathrm{IAC}}(n)]}{s} \leq 1$. *IAC requires* $\mathcal{O}(n(\log n)^3)$ *floating-point operations.*

As far as we know, IAC is the first algorithm achieving performance that matches the lower bound presented in Theorem 1.1. It improves the previous results on the LSBM [Yun and Proutiere, 2016], moving beyond high-probability performance guarantees. More precisely, the algorithm presented by Yun and Proutiere [2016] misclassifies less than s items with a probability that tends to 1 as n grows large, provided that s satisfies (2). However, the probability of the *failure* event (when the algorithm misclassifies more than s items) is not quantified in their work. It is necessary to quantify this probability for guarantees in expectation. To achieve this goal, we have to significantly revisit the analysis presented in Yun and Proutiere [2016]. (i) We need to re-design some of the components of the algorithm. (ii) Moreover, in every step of the performance analysis, it is necessary to provide a small enough upper bound for the probability of the failure event. The analysis of the error rate after the first step of the algorithm (essentially a spectral clustering algorithm) requires establishing an upper bound on the spectral norm of the noise matrix associated with the observations. To accomplish this, we leverage arguments from the spectral analysis of sparse random graphs, as demonstrated in, for example, Feige and Ofek [2005]. Unfortunately, these arguments hold with a high probability that does not suffice to establish guarantees in expectation. We extend the arguments so that they hold with probability at least $1 - 1/n^c$ for any $c > 0$, which is enough to obtain guarantees in expectation. Such an extension was also proposed in Gao et al. [2017], Le et al. [2017] for the SBM (we compare our results to those of Gao et al. [2017] in Section 2), but our spectral clustering algorithm is different, and our results apply to the general LSBM. The analysis of the likelihood-based improvement step has to be significantly modified to prove that all intermediate statements (e.g., the lower bound of the rate at which the error rate decreases) hold with a sufficiently high probability, typically again $1 - 1/n^c$ for any $c > 0$. Obtaining such a guarantee is challenging due to the correlations created by the initial clustering, which affect the likelihood-based local improvement. However, we have made it possible by using a set of items with desirable properties in the LSBM (set H in Section 3.2.2) and then conducting deterministic proofs on that set.

2 Related Work

2.1 Community Detection in the SBM

Community detection in the SBM and its extensions have received a lot of attention over the last decade. We first briefly outline existing results below and then zoom in on a few papers that are the most relevant for our analysis. The results of the SBM can be categorized depending on the targeted performance guarantees. We distinguish three types of guarantees: (a) detectability, (b) asymptotically accurate recovery, and (c) exact recovery. Most results are concerned with the simple SBM, which is obtained as LSBM characterized by $L = 1$ and the intra- and inter-cluster probabilities $p(i, i, 1)$ and $p(i, j, 1)$ for $i \neq j \in [K]$.

(a) Detectability refers to the requirement of returning estimated clusters that are positively correlated with the true clusters. It is typically studied in the sparse binary SBM where $K = 2$, $\alpha_1 = \alpha_2$, $p(1, 1, 1) = p(2, 2, 1) = a/n$ and $p(1, 2, 1) = p(2, 1, 1) = b/n$, for some constants $a > b$ independent of n. For such SBM, detectability can be achieved if and only if $(a - b) > \sqrt{2(a + b)}$ [Decelle et al., 2011, Mossel et al., 2015a, Massoulié, 2013]. Detectability conditions in more general sparse SBMs have been investigated in Krzakala et al. [2013], Bordenave et al. [2015]. In the sparse SBM, when the edge probabilities scale as $\mathcal{O}(1/n)$, there is a positive fraction of isolated items, and we cannot do much better than merely detecting the clusters.

(b) In this paper, we are interested in scenarios where the edge probabilities are $\omega(1/n)$, allowing us to achieve an asymptotically accurate recovery of the clusters. This means that the proportion of misclassified items tends to 0 as n grows large. A necessary and sufficient condition for asymptotically accurate recovery in the SBM (with any number of clusters of different but linearly increasing sizes) has been derived in Yun and Proutiere [2014b] and Mossel et al. [2015b]. In our work, we conduct more precise analysis and derive the minimal expected proportion of misclassified items. This minimal proportion is characterized by our divergence $D(\alpha, p)$ and is, therefore, *instance-specific*. Our analysis thus provides more accurate results than those derived in a minimax framework [Gao et al., 2017]. An extensive comparison with Gao et al. [2017] is provided below.

(c) An algorithm achieves an asymptotically exact recovery if it only misclassifies $o(1)$ items. Conditions for such exact recovery have also been recently studied in the binary symmetric SBM [Yun and Proutiere, 2014a, Abbe et al., 2016, Mossel et al., 2015b, Hajek et al., 2016] and in more general SBM [Abbe and Sandon, 2015a,b, Wang et al., 2021]. In Yun and Proutiere [2016], these conditions were further extended to the even more general LSBM.

2.2 Optimal Recovery Rate

Next, we discuss two papers Zhang and Zhou [2016], Gao et al. [2017] that are directly related to our analysis. These papers study the standard SBM and present the minimal expected number of misclassified items but in a minimax setting, in the regime where an asymptotically accurate recovery is possible. To simplify the exposition here, we assume that all clusters are of equal size (refer to Zhang and Zhou [2016], Gao et al. [2017] for more details). The authors of Zhang and Zhou [2016], Gao et al. [2017] characterize the minimal expected number of misclassified items in the *worst* possible SBM within the class $\Theta(n, a, b)$ of SBMs satisfying, using our notation, $p(i, i, 1) \geq \frac{a}{n}$ and $p(i, j, 1) \leq \frac{b}{n}$ for all $i \neq j \in [K]$, for some positive constants a, b depending on n[1]. The minimal expected number of misclassified items is defined through the Rényi divergence of order $\frac{1}{2}$ between the Bernoulli random variables of respective means $\frac{a}{n}$ and $\frac{b}{n}$, given by $I^*(n, a, b) = -2\log(\sqrt{\frac{a}{n}}\sqrt{\frac{b}{n}} + \sqrt{1-\frac{a}{n}}\sqrt{1-\frac{b}{n}})$. When $nI^*(n, a, b) = \omega(1)$, it is equal to $n\exp(-(1+o(1))\frac{nI^*(n,a,b)}{K})$. Zhang and Zhou [2016] established that the so-called penalized Maximum Likelihood Estimator (MLE) achieves this minimax optimal recovery rate but does not provide any algorithm to compute it. The authors of Gao et al. [2017] present an algorithm that achieves this minimax lower bound in the following sense (see Theorem 4 in Gao et al. [2017]):

$$\sup_{(\alpha,p)\in\Theta(n,a,b)} \mathbb{P}_{(\alpha,p)}\left(\varepsilon^\pi(n) \geq n\exp\left(-(1+o(1))\frac{nI^*(n,a,b)}{K}\right)\right) \to 0,$$

where $\mathbb{P}_{(\alpha,p)}$ denotes the distribution of the observations generated under the SBM (α, p). One could argue that the above guarantee does not match the minimax lower bound valid for the *expected* number of misclassified items. However, by carefully inspecting the proof of Theorem 4 in Gao et al. [2017], it is easy to see that the guarantee also holds in expectation:

Corollary 2.1. *Assume that $a/b = \Theta(1)$, $(a-b)^2/a = \omega(1)$, and $a = \mathcal{O}(\log(n))$. Let A_{uv} be the observation for the pair of items (u, v). Under Algorithm 1 in* Gao et al. [2017] initialized with USC(τ) in Gao et al. [2017] for $\tau = C\frac{1}{n}\sum_{u\in[n]}\sum_{v\in[n]} A_{uv}$ with some large enough constant $C > 0$,

$$\sup_{(\alpha,p)\in\Theta(n,a,b)} \mathbb{E}_{(\alpha,p)}[\varepsilon^\pi(n)] \leq n\exp\left(-(1+o(1))\frac{nI^*(n,a,b)}{K}\right).$$

The proof is presented in Appendix F.10. The assumptions of Corollary 2.1 are satisfied when our assumptions (A1) and (A2) hold.

The algorithm presented in Gao et al. [2017], which has established performance guarantees, comes with a high computational cost. It requires applying spectral clustering n times, where for each item u, the algorithm builds a modified adjacency matrix by removing the u-th column and the u-th row and then computes a spectral clustering of this matrix. In contrast, our algorithm performs spectral clustering only once. Gao et al. [2017] also proposed an algorithm with reduced complexity (running in $\Omega(n^2)$), but without performance guarantees. Our algorithm not only performs the spectral clustering once but also requires just $\mathcal{O}(n(\log n)^3)$ operations. Additionally, our algorithm empirically exhibits better classification accuracy than the penalized local maximum likelihood estimation algorithm Gao et al. [2017] in several scenarios.

[1] Refer to Gao et al. [2017] for a precise definition of the class of SBMs considered. Compared to our assumptions (A1)-(A2)-(A3) specialized to SBMs, this class of SBMs is slightly more general.

To conclude, compared to Gao et al. [2017], our analysis provides an instance-specific lower bound for the classification error probability (rather than minimax) and introduces a low-complexity algorithm that matches this lower bound. Additionally, our analysis is applicable to the generic Labeled SBMs. It is worth noting, however, that Gao et al. [2017] derives upper bounds for classification error probability under slightly more general assumptions than ours for the SBMs.

3 The Instance-Adaptive Clustering Algorithm and its Optimality

3.1 Algorithms

The Instance-Adaptive Clustering (IAC), whose pseudo-code is presented in Algorithm 1, consists of two phases: a spectral clustering initialization phase and a likelihood-based improvement phase.

(i) Spectral clustering initialization. The algorithm relies on simple spectral techniques to obtain rough but global estimates of the clusters. For details, refer to lines 1-4 in Algorithm 1. The algorithm first constructs an observation matrix $A^\ell = (A^\ell_{uv})_{u,v}$ for each label ℓ (where $A^\ell_{uv} = 1$ iff label ℓ is observed on edge (u, v)), and sums these matrices to create the aggregated matrix A. After trimming (to eliminate rows and columns corresponding to items with too many observed labels – as these would perturb the spectral properties of A), we apply spectral clustering to A, as shown in Algorithm 2. Specifically, we use the iterative power method (instead of using a direct SVD, which has high complexity) combined with singular value thresholding [Chatterjee, 2015]. This approach allows us to control the algorithm's computational complexity and accurately estimate the number of clusters. Notable differences compared to the spectral clustering in Yun and Proutiere [2016] include modifications to the number of matrix multiplications in the iterative power method (we require approximately $(\log n)^2$ multiplications) and an enlargement of the set of centroid candidates in the k-means algorithm (this set now comprises $(\log n)^2$ randomly selected items) for tighter control of the failure event probability, leading to guarantees in expectation.

(ii) Likelihood-based improvements. Using the initial cluster estimates S_i, we can also estimate p from the data. For any i, j, ℓ, we calculate $\hat{p}(i, j, \ell) = \frac{\sum_{u \in S_i} \sum_{v \in S_j} A^\ell_{uv}}{|S_i||S_j|}$. Based on $\hat{p}$, the log-likelihood of item v belonging to cluster S_k is computed as $\sum_{i \in [\hat{K}]} \sum_{w \in S_i} \sum_{\ell=0}^{L} A^\ell_{vw} \log \hat{p}(k, i, \ell)$. Subsequently, v is assigned to the cluster that maximizes this log-likelihood over $[\hat{K}]$. This process is applied to all items and iterated for $\log n$ times.

3.2 Performance analysis

We sketch below the proof of Theorem 1.2. The complete proof is postponed to the appendix.

3.2.1 Spectral clustering initialization

The following theorem establishes performance guarantees for the cluster estimates returned by the spectral clustering algorithm (Algorithm 2). Specifically, we show that the number of clusters is correctly predicted as $\hat{K} = K$, and the number of misclassified items is $\mathcal{O}(1/\bar{p})$.

Theorem 3.1. *Assume that (A1) and (A2) hold. After Algorithm 2, for any $c > 0$, there exists a constant $C > 0$ such that*

$$\left(\hat{K} = K \quad \text{and} \quad \min_\theta \left| \bigcup_{k=1}^{K} S_k \setminus \mathcal{I}_{\theta(k)} \right| \le \frac{C}{\bar{p}} \right) \quad \textit{with probability at least} \quad 1 - \exp(-cn\bar{p}),$$

where the minimization is performed over the permutation θ of $[K]$.

Sketch of proof of Theorem 3.1. Let M^ℓ denote the expectation of the matrix A^ℓ: $M^\ell_{uv} = p(i, j, \ell)$ when $u \in \mathcal{V}_i$ and $v \in \mathcal{V}_j$. Let $M = \sum_{\ell=1}^{L} M^\ell$, and $M_\Gamma \in [0, 1]^{n \times n}$ be the corresponding trimmed matrix: $(M_\Gamma)_{wv} = M_{wv} \mathbb{1}_{\{w,v \in \Gamma\}}$.

(a) The main ingredient of the proof is an upper bound on the norm of the noise matrix $X^\ell_\Gamma = A^\ell_\Gamma - M^\ell_\Gamma$ that holds with a sufficiently high probability, as stated in the following lemma.

Algorithm 1: Instance-Adaptive Clustering

```
Input: Observed adjacency matrices $A^\ell$ for each label $\ell$ ($A^\ell_{uv} = 1$ if $\ell$ is observed between $u$ and $v$)
1. Estimated average degree. $\widetilde{p} \leftarrow \frac{\sum_{\ell=1}^{L} \sum_{v,w \in \mathcal{I}: v>w} A^\ell_{vw}}{n(n-1)}$
2. Aggregated Matrix. $A \leftarrow \sum_{\ell=1}^{L} A^\ell$.
3. Trimming.
Compute $A_\Gamma$, where $\Gamma$ is the set of items with the top-$\lfloor n \exp(-n\widetilde{p}) \rfloor$ largest values of $\sum_{\ell=1}^{L} \sum_{w \in \mathcal{I}} A^\ell_{vw}$.
$(A_\Gamma)_{wv} = A_{wv}$ when $w, v \in \Gamma$ and
$(A_\Gamma)_{wv} = 0$ when $w, v \in \Gamma^c$ .
4. Spectral Clustering.
Run Algorithm 2 with input $A_\Gamma, \widetilde{p}$ and output $\{S_k\}_{k=1,\dots,\hat{K}}$.
5. Estimation of the Statistical Parameters.
$\hat{p}(i,j,\ell) \leftarrow \frac{\sum_{u \in S_i} \sum_{v \in S_j} A^\ell_{uv}}{|S_i||S_j|}$ for all $1 \le i, j \le \hat{K}$ and $0 \le \ell \le L$.
6. Likelihood-based local improvements.
$S_k^{(0)} \leftarrow S_k$ for all $k \in [\hat{K}]$
for $t = 1$ to $\log n$ do
    $S_k^{(t)} \leftarrow \emptyset$ for all $k \in [\hat{K}]$
    for $v \in \mathcal{I}$ do
        $k^* \leftarrow \arg\max_{1 \le k \le \hat{K}} \left\{ \sum_{i \in [\hat{K}]} \sum_{w \in S_i^{(t-1)}} \sum_{\ell=0}^{L} A^\ell_{vw} \log \hat{p}(k,i,\ell) \right\}$ (tie broken uniformly at random)
        $S_{k^*}^{(t)} \leftarrow S_{k^*}^{(t)} \cup \{v\}$
    end
end
$\hat{\mathcal{I}}_k \leftarrow S_k^{(\log n)}$ for all $k \in [\hat{K}]$
Output: $(\hat{\mathcal{I}}_k)_{k=1,\dots,\hat{K}}$.
```

Lemma 3.2. *For any $\ell \in [L]$, for any $C > 0$, there exists $C' > 0$ such that: $||X^\ell_\Gamma||_2 \le C'\sqrt{n\bar{p}}$, with probability at least $1 - \exp(-Cn\bar{p})$.*

The proof, detailed in Appendix F.5, leverages and extends techniques developed for the spectral analysis of random graphs Feige and Ofek [2005], Coja-Oghlan [2010]. Based on the above lemma, we deduce that for any $C > 0$, there exists $C' > 0$ such that: $\|X_\Gamma\| \le \sum_{\ell=1}^{L} \|X^\ell_\Gamma\| \le C'\sqrt{n\bar{p}}$, with probability at least $1 - \exp(-Cn\bar{p})$.

(b) The second ingredient of the proof is the following lemma, whose proof is provided in Appendix F.6. The lemma provides a lower bound on the distance between two columns of M_Γ corresponding to two items in distinct clusters.

Lemma 3.3. *There exists a constant $C' > 0$ such that with probability at least $1 - \exp(-\omega(n))$, $||M_{\Gamma,v} - M_{\Gamma,w}||_2^2 \ge C'n\bar{p}^2$, uniformly over all $v, w \in \Gamma$ with $\sigma(v) \neq \sigma(w)$.*

(c) The final proof ingredient concerns the performance of the iterative power method with singular value thresholding and is proved in Appendix F.7.

Lemma 3.4. *For any $c > 0$, there exists a constant $C > 0$ such that with probability at least $1 - 1/n^c$, $\|A_\Gamma - \hat{A}\|_2 \le C\sigma_{K+1}$, where σ_{K+1} is the $(K+1)$-th singular value of the matrix A_Γ.*

The first two lemmas may resemble those presented in Yun and Proutiere [2016]; however, we needed to extend the analysis so that these results hold with a higher probability. We can now proceed with proving the theorem. We first explain why the number of clusters is accurately estimated. It is straightforward to verify that there exist two strictly positive constants, C_1 and C_2, such that with probability at least $1 - \exp(-\omega(n))$, $C_1\bar{p} \le \widetilde{p} \le C_2\bar{p}$ (refer to Lemma D.5). Consequently, from Lemma 3.2, we deduce that for any $C > 0$, with probability at least $1 - \exp(-Cn\bar{p})$, the

Algorithm 2: Spectral Clustering

```
Input: $A_\Gamma, \widetilde{p}$
1. Iterative Power Method with Singular Value Thresholding
$\chi \leftarrow n$, $k \leftarrow 0$, and $\hat{U} \leftarrow 0^{n\times 1}$
while $\chi \geq \sqrt{n\widetilde{p}}\log(n\widetilde{p})$ do
    $k \leftarrow k+1$, $U_0 \leftarrow n \times 1$ Gaussian random vector
    (Iterative power method) $U_t \leftarrow (A_\Gamma)^{2\lceil(\log n)^2\rceil+1}U_0$
    (Orthonormalizing) $\hat{U}_k \leftarrow \frac{U_t - \hat{U}_{1:k-1}(\hat{U}_{1:k-1}^\top U_t)}{\|U_t - \hat{U}_{1:k-1}(\hat{U}_{1:k-1}^\top U_t)\|_2}$
    (The estimated $k$-th singular value) $\chi \leftarrow \|A_\Gamma \hat{U}_k\|_2$
end
$\hat{K} \leftarrow k-1, \qquad \hat{A} \leftarrow \hat{U}_{1:\hat{K}}^\top A_\Gamma$
2. k-means Clustering
$\mathcal{I}_R \leftarrow$ a subset of $\Gamma$ obtained by randomly selecting $\lceil(\log n)^2\rceil$ items.
for $t = 1$ to $\lceil \log n \rceil$ do
    $Q_v^{(t)} \leftarrow \left\{ w \in \mathcal{I} : \|\hat{A}_w - \hat{A}_v\|_2^2 \leq t\frac{\widetilde{p}}{100} \right\}$ for all $v \in \mathcal{I}_R$
    $T_k^{(t)} \leftarrow \emptyset$ for all $k \in [\hat{K}]$
    for $k = 1$ to $\hat{K}$ do
        $v_k^* \leftarrow \arg\max_{v \in \mathcal{I}_R} \left| Q_v^{(t)} \setminus \cup_{i=1}^{k-1} T_i^{(t)} \right|$
        $T_k^{(t)} \leftarrow Q_{v_k^*}^{(t)} \setminus \cup_{i=1}^{k-1} T_i^{(t)}$
        $\xi_k^{(t)} \leftarrow \sum_{v \in T_k^{(t)}} \frac{\hat{A}_v}{|T_k^{(t)}|}$
    end
    for $v \in \mathcal{I} \setminus \cup_{k=1}^{\hat{K}} T_k^{(t)}$ do
        $k^* \leftarrow \arg\min_{1\leq k \leq \hat{K}} \|\hat{A}_v - \xi_k^{(t)}\|_2^2$
        $T_{k^*}^{(t)} \leftarrow T_{k^*}^{(t)} \cup \{v\}$
    end
    $r_t \leftarrow \sum_{k=1}^{\hat{K}} \sum_{v \in T_k^{(t)}} \|\hat{A}_v - \xi_k^{(t)}\|_2^2$
end
$t^* \leftarrow \arg\min_{1 \leq t \leq \lceil \log n \rceil} r_t$
$S_k \leftarrow T_k^{(t^*)}$ for all $k \in [\hat{K}]$
Output: $\{S_k\}_{k=1,\dots,\hat{K}}$.
```

$(K+1)$-th singular value of A_Γ is significantly smaller than $\sqrt{\widetilde{p}n}\log(n\widetilde{p})$. In conjunction with Lemma 3.4, this indicates that $K = \hat{K}$ with probability at least $1 - \exp(-Cn\bar{p})$. Therefore, we can assume in the remainder of the proof that $K = \hat{K}$.

Without loss of generality, let us denote γ as the permutaion of $[K]$ such that the set of misclassified items is $\bigcup_{k=1}^K S_k \setminus \mathcal{I}_{\gamma(k)}$. Based on Lemma 3.3, we can prove that: with probability at least $1 - \exp(\omega(n))$,

$$\left| \bigcup_{k=1}^K S_k \setminus \mathcal{I}_{\gamma(k)} \right| C' n\bar{p}^2 \leq \sum_{k=1}^K \sum_{v \in S_k \setminus \mathcal{I}_{\gamma(k)}} ||M_{\Gamma,v} - M_{\Gamma,\gamma(k)}||_2^2 \leq 8||M_\Gamma - \hat{A}||_F^2 + 8r_{t^*},$$

where $M_{\Gamma,\gamma(k)} = M_{\Gamma,w}$ for $w \in \mathcal{I}_{\gamma(k)}$, and where r_{t^*} is defined in Algorithm 2. Furthermore, for any $C > 0$, using Lemmas 3.2 and 3.4, we can establish that there exists a constant $C_0 > 0$ such that $||M_\Gamma - \hat{A}||_F^2 \leq C_0 n\bar{p}$ with probability at least $1 - \exp(-Cn\bar{p})$. Through a refined analysis of the k-means algorithm, we can also prove the existence of a constant $C_1 > 0$ such that $r_{t^*} \leq C_1 n\bar{p}$. For details, please refer to Appendix E.

3.2.2 Likelihood-based improvements

To complete the proof of Theorem 1.2, we analyze the likelihood-based improvement phase of the IAC algorithm. For this purpose, we define a set of well-behaved items H as the largest set of items $v \in \mathcal{I}$ that meet the following three conditions for some constant $C_{\mathrm{H1}} > 0$:

(H1) $e(v,\mathcal{I}) \leq C_{\mathrm{H1}} n\bar{p}$;

(H2) when $v \in \mathcal{I}_k$, $\sum_{i=1}^{K} \sum_{\ell=0}^{L} e(v,\mathcal{I}_i,\ell) \log \frac{p(k,i,\ell)}{p(j,i,\ell)} \geq \frac{n\bar{p}}{\log^4 n\bar{p}}$ for all $j \neq k$;

(H3) $e(v,\mathcal{I} \setminus H) \leq \frac{2n\bar{p}}{\log^5(n\bar{p})}$.

In these conditions, we use the following notation: for any $v \in \mathcal{I}$, $S \subset \mathcal{I}$, and $\ell \in [L]$, $e(v,S,\ell) = \sum_{w\in S} A^{\ell}_{vw}$, and $e(v,S) = \sum_{\ell=1}^{L} e(v,S,\ell)$. We will show that all items in H are correctly clustered with high probability, and the expected number of items not in H matches the lower bound on the expected number of misclassified items. Each condition in the definition of H can be interpreted as follows: (H1) imposes some regularity in the degree of the item, (H2) implies that $v \in H$ is correctly classified when using the likelihood, and the last condition (H3) implies that the item does not have too many labels pointing outside of the set H.

First, we show that the expected number of items not in H can be upper bounded by a number s that is of the same order as $n\exp(-nD(\alpha,p))$.

Proposition 3.5. *When* $s \geq n \exp\left(-nD(\alpha,p) + \frac{n\bar{p}}{\log^3 n\bar{p}}\right)$,

$$\frac{\mathbb{E}[|\mathcal{I} \setminus H|]}{s} \leq 1 + \exp\left(-\frac{3n\bar{p}}{4\log^3 n\bar{p}}\right) + \exp(-\omega(n\bar{p})).$$

Moreover, $\lim_{n\to\infty} \mathbb{P}(|\mathcal{I} \setminus H| \leq s) = 1$.

The proof of Proposition 3.5 can be found in Appendix D.3. This proof shows that the probability of an item satisfying (H2) is dominant compared to the probabilities of the other two conditions and is of the order of $\exp(-nD(\alpha,p))$.

Next, we examine the performance of the likelihood-based improvement step (Line 6 in the IAC algorithm) for items in H. In the following proposition, we quantify the improvement achieved with one iteration of this step.

Proposition 3.6. *Assume that there exists a constant* $C > 0$, *such that* $|\bigcup_{k=1}^{K}(S_k^{(0)} \setminus \mathcal{I}_k) \cap H| + |\mathcal{I} \setminus H| \leq C\frac{1}{\bar{p}}$. *Then, for any constant* $C' > 0$, *with probability at least* $1 - \exp\left(-C'n\bar{p}\right)$, *the following statement holds*

$$\frac{|\bigcup_{k=1}^{K}(S_k^{(t+1)} \setminus \mathcal{I}_k) \cap H|}{|\bigcup_{k=1}^{K}(S_k^{(t)} \setminus \mathcal{I}_k) \cap H|} \leq \frac{1}{\sqrt{n\bar{p}}} \quad \textit{for all} \quad t \geq 0.$$

The proof of Proposition 3.6 can be found in Appendix D.4 and takes advantage of the fact that a likelihood-based test using the estimator $\hat{p}(j,i,\ell)$ matches the test that would use the true likelihood, with high probability.

Proof of Theorem 1.2. we can now complete the proof by observing that from Proposition 3.6, after the $\lceil \log n \rceil$ iterations of the likelihood-based improvement step, $|\cup_{k=1}^{K} (S_k^{(\lceil \log n \rceil)} \setminus \mathcal{I}_k) \cap H| = 0$, with probability at least $1 - \exp(-Cn\bar{p})$ for any constant $C > 0$. Combining this result with Proposition 3.5, when $nD(\alpha,p) - \frac{n\bar{p}}{\log(n\bar{p})^3} \geq \log(n/s)$,

$$\begin{aligned}\mathbb{E}[\varepsilon^{\mathrm{IAC}}(n)] &\leq \frac{1}{1-o(1)}\,\mathbb{E}[|\mathcal{I} \setminus H|] + o(1)\\ &\leq \frac{s}{1-o(1)}\left(1 + \exp\left(-\frac{3n\bar{p}}{4\log^3 n\bar{p}}\right) + \exp(-\omega(n\bar{p}))\right) + o(1)\end{aligned}$$

and $\varepsilon^{\mathrm{IAC}}(n) \leq s + o(1)$, with high probability.

4 Numerical Experiments

In this section, we evaluate the proposed algorithm through empirical analysis. Our experiments are based on the code of Wang et al. [2021], and we consider three scenarios from Gao et al. [2017] as well as one additional scenario. The focus of our comparison is on the IAC algorithm (Algorithm 1) and the computationally light version of the penalized local maximum likelihood estimation (PLMLE) algorithm (Algorithm 3 in Gao et al. [2017]). While PLMLE has no analytical performance guarantees, it requires $\Omega(n^2)$ floating-point operations. We consider simple SBMs with $L = 1$.

Model 1: Balanced Symmetric. First, consider the SBM corresponding to the "Balanced case" in Gao et al. [2017]. Assume that $n = 2500$, $K = 10$, and $L = 1$. We fix the community size to be equal as $\forall k \in [10], |\mathcal{I}_k| = 250$. We set the observation probability as $p(k, k, 1) = 0.48$ for all k and $p(i, k, 1) = 0.32$ for all $i \neq k$.

Model 2: Imbalanced. The next SBM corresponds to the "Imbalanced case" in Gao et al. [2017]. We set $n = 2000$, $K = 4$, and $L = 1$. The sizes of the clusters are heterogenous: $|\mathcal{I}_1| = 200$, $|\mathcal{I}_2| = 400$, $|\mathcal{I}_3| = 600$, and $|\mathcal{I}_4| = 800$.

Model 3: Sparse Symmetric. The last experimental setting from Gao et al. [2017] is the sparse and symmetric case. We generate networks with $n = 4000$, $K = 10$, and $L = 1$. Clusters are of equal sizes: $\forall k \in [10]$, $|\mathcal{I}_k| = 400$. We set the statistical parameter as $p(k, k, 1) = 0.032$ for all k and $p(i, k, 1) = 0.005$ for all $i \neq k$.

Model 4: Sparse Asymmetric. Lastly, we consider the cluster recovery problem with a sparse and asymmetric statistical parameter. We set $n = 1200$, $K = 4$, and $L = 1$. Clusters are of equal sizes: $\forall k \in [4], |\mathcal{I}_k| = 300$. We fix the statistical parameter $(p(i, k, 1))_{i,k}$ as

$$(p(i,k,1))_{i,k} = \begin{pmatrix} 0.032 & 0.005 & 0.008 & 0.005 \\ 0.005 & 0.028 & 0.005 & 0.008 \\ 0.008 & 0.005 & 0.032 & 0.005 \\ 0.005 & 0.008 & 0.005 & 0.028 \end{pmatrix}. \tag{3}$$

The results of our experiments are presented in Table 1. The IAC algorithm consistently performs slightly better than Algorithm 3 in Gao et al. [2017]. For additional figures and details, please refer to Appendix G.

Table 1: Number of misclassified items. IAC and PLMLE indicate Algorithm 1 and Algorithm 3 in Gao et al. [2017], respectively. Means and standard deviations are calculated from the results of 100 experiment instances.

	Model 1		Model 2		Model 3		Model 4	
Algorithm	Mean	Std	Mean	Std	Mean	Std	Mean	Std
IAC	2.8800	1.5909	0.0000	0.0000	29.4100	4.9789	45.5600	9.2489
PLMLE	2.9700	1.6542	0.1850	0.4262	31.0400	5.1775	54.7400	10.5329

5 Conclusion

In this paper, we have investigated the problem of recovering hidden communities in the Labeled Stochastic Block Model (LSBM) with a finite number of clusters. We revisited instance-specific lower bounds on the expected number of misclassified items. We proposed IAC, an algorithm whose performance matches these lower bounds both in expectation and with high probability. IAC consists of a one-time spectral clustering algorithm followed by an iterative likelihood-based cluster assignment improvement. This approach is based on the instance-specific lower bound and does not require any model parameters, including the number of clusters. By performing a spectral clustering only once, IAC maintains an overall computational complexity of $\mathcal{O}(n\text{polylog}(n))$.

References

Emmanuel Abbe. Community detection and stochastic block models. *Foundations and Trends in Communications and Information Theory*, 14(1–2):1–162, 2018.

Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. In *FOCS*, 2015a.

Emmanuel Abbe and Colin Sandon. Recovering communities in the general stochastic block model without knowing the parameters. In *NeurIPS*, 2015b.

Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016.

Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: Community detection and non-regular ramanujan graphs. In *FOCS*, 2015.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(2):227–284, 2010.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Inference and phase transitions in the detection of modules in sparse networks. *Physical Review Letters*, 107, 2011.

Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Achieving optimal misclassification proportion in stochastic block models. *Journal of Machine Learning Research*, 18(1): 1980–2024, 2017.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5):2788–2797, 2016.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53 (2):217–288, 2011.

Simon Heimlicher, Marc Lelarge, and Laurent Massoulié. Community detection in the labelled stochastic block model. *arXiv preprint arXiv:1209.2910*, 2012.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17:1–42, 2016.

Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Can M. Le, Elizaveta Levina, and Roman Vershynin. Concentration and regularization of random graphs. *Random Structures & Algorithms*, 51(3):538–561, 2017.

Marc Lelarge, Laurent Massoulié, and Jiaming Xu. Reconstruction in the labeled stochastic block model. In *2013 IEEE Information Theory Workshop*, 2013.

Torgny Lindvall. *Lectures on the Coupling Method*. Dover Books on Mathematics Series. Courier Corporation, 2002.

Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *STOC*, 2013.

Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3):431–461, 2015a.

Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. In *STOC*, 2015b.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.

Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Society, 2012.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, NY, 2008.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Peng Wang, Huikang Liu, Zirui Zhou, and Anthony Man-Cho So. Optimal non-convex exact recovery in stochastic block model via projected power method. In *ICML*, 2021.

Se-Young Yun and Alexandre Proutiere. Accurate community detection in the stochastic block model via spectral algorithms. *arXiv preprint arXiv:1412.7335*, 2014a.

Se-Young Yun and Alexandre Proutiere. Community detection via random and adaptive sampling. In *COLT*, 2014b.

Se-Young Yun and Alexandre Proutiere. Optimal cluster recovery in the labeled stochastic block model. In *NeurIPS*, 2016.

Anderson Y. Zhang and Harrison H. Zhou. Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252 – 2280, 2016.