Supplementary Material: Optimal Cluster Recovery in the Labeled Stochastic Block Model

Se-Young Yun
CNLS, Los Alamos National Lab.
Los Alamos, NM 87545
syun@lanl.gov

Alexandre Proutiere
Automatic Control Dept., KTH
Stockholm 100-44, Sweden
alepro@kth.se

Abstract

We consider the problem of community detection or clustering in the labeled Stochastic Block Model (LSBM) with a finite number K of clusters of sizes linearly growing with the global population of items n. Every pair of items is labeled independently at random, and label ℓ appears with probability $p(i, j, \ell)$ between two items in clusters indexed by i and j, respectively. The objective is to reconstruct the clusters from the observation of these random labels.

Clustering under the SBM and their extensions has attracted much attention recently. Most existing work aimed at characterizing the set of parameters such that it is possible to infer clusters either positively correlated with the true clusters, or with a vanishing proportion of misclassified items, or exactly matching the true clusters. We find the set of parameters such that there exists a clustering algorithm with at most s misclassified items in average under the general LSBM and for any $s = o(n)$, which solves one open problem raised in [2]. We further develop an algorithm, based on simple spectral methods, that achieves this fundamental performance limit within $O(n\text{polylog}(n))$ computations and without the a-priori knowledge of the model parameters.

1   Introduction

Community detection consists in extracting (a few) groups of similar items from a large global population, and has applications in a wide spectrum of disciplines including social sciences, biology, computer science, and statistical physics. The communities or clusters of items are inferred from the observed pair-wise similarities between items, which, most often, are represented by a graph whose vertices are items and edges are pairs of items known to share similar features.

The stochastic block model (SBM), introduced three decades ago in [13], constitutes a natural performance benchmark for community detection, and has been, since then, widely studied. In the SBM, the set of items $\mathcal{V} = \{1, \ldots, n\}$ are partitioned into K non-overlapping clusters $\mathcal{V}_1, \ldots, \mathcal{V}_K$, that have to be recovered from an observed realization of a random graph. In the latter, an edge between two items belonging to clusters $\mathcal{V}_i$ and $\mathcal{V}_j$, respectively, is present with probability $p(i, j)$, independently of other edges. The analyses presented in this paper apply to the SBM, but also to the *labeled* stochastic block model (LSBM) [12], a more general model to describe the similarities of items. There, the observation of the similarity between two items comes in the form of a *label* taken from a finite set $\mathcal{L} = \{0, 1, \ldots, L\}$, and label ℓ is observed between two items in clusters $\mathcal{V}_i$ and $\mathcal{V}_j$, respectively, with probability $p(i, j, \ell)$, independently of other labels. The standard SBM can be seen as a particular instance of its labeled counterpart with two possible labels 0 and 1, and where the edges present (resp. absent) in the SBM correspond to item pairs with label 1 (resp. 0). The problem

of cluster recovery under the LSBM consists in inferring the hidden partition $\mathcal{V}_1, \ldots, \mathcal{V}_K$ from the observation of the random labels on each pair of items.

Over the last few years, we have seen remarkable progresses for the problem of cluster recovery under the SBM (see [8] for an exhaustive literature review), highlighting its scientific relevance and richness. Most recent work on the SBM aimed at characterizing the set of parameters (i.e., the probabilities $p(i,j)$ that there exists an edge between nodes in clusters i and j for $1 \leq i, j \leq K$) such that some qualitative recovery objectives can or cannot be met. For sparse scenarios where the average degree of items in the graph is $O(1)$, parameters under which it is possible to extract clusters positively correlated with the true clusters have been identified [5, 20, 18]. When the average degree of the graph is $\omega(1)$, one may predict the set of parameters allowing a cluster recovery with a vanishing (as n grows large) proportion of misclassified items [25, 19], but one may also characterize parameters for which an asymptotically exact cluster reconstruction can be achieved [1, 24, 9, 19, 2, 3, 14].

In this paper, we address the finer and more challenging question of determining, under the general LSBM, the minimal number of misclassified items given the parameters of the model. Specifically, for any given $s = o(n)$, our goal is to identify the set of parameters such that it is possible to devise a clustering algorithm with at most s misclassified items. Of course, if we achieve this goal, we shall recover all the aforementioned results on the SBM.

Main results. We focus on the labeled SBM as described above, and where each item is assigned to cluster $\mathcal{V}_k$ with probability $\alpha_k > 0$, independently of other items. We assume w.l.o.g. that $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_K$. We further assume that $\alpha = (\alpha_1, \ldots, \alpha_K)$ does not depend on the total population of items n. Conditionally on the assignment of items to clusters, the pair or edge $(v, w) \in \mathcal{V}^2$ has label $\ell \in \mathcal{L} = \{0, 1, \ldots, L\}$ with probability $p(i, j, \ell)$, when $v \in \mathcal{V}_i$ and $w \in \mathcal{V}_j$. W.l.o.g., 0 is the most frequent label, i.e., $0 = \arg\max_\ell \sum_{i=1}^{K} \sum_{j=1}^{K} \alpha_i \alpha_j p(i, j, \ell)$. Throughout the paper, we typically assume that $\bar{p} = o(1)$ and $\bar{p}n = \omega(1)$ where $\bar{p} = \max_{i,j,\ell \geq 1} p(i, j, \ell)$ denotes the maximum probability of observing a label different than 0. We shall explicitly state whether these assumption are made when deriving our results. In the standard SBM, the second assumption means that the average degree of the corresponding random graph is $\omega(1)$. This also means that we can hope to recover clusters with a vanishing proportion of misclassified items. We finally make the following assumption: there exist positive constants η and ε such that for every $i, j, k \in [K] = \{1, \ldots, K\}$,

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

(A2) imposes a certain separation between the clusters. For example, in the standard SBM with two communities, $p(1, 1, 1) = p(2, 2, 1) = \xi$, and $p(1, 2, 1) = \zeta$, (A2) is equivalent to $2(\xi - \zeta)^2/\xi^2 \geq \epsilon$. In summary, the LSBM is parametrized by α and $p = (p(i, j, \ell))_{1 \leq i, j \leq K, 0 \leq \ell \leq L}$, and recall that α does not depend on n, whereas p does.

For the above LSBM, we derive, for any arbitrary $s = o(n)$, a necessary condition under which there exists an algorithm inferring clusters with s misclassified items. We further establish that under this condition, a simple extension of spectral algorithms extract communities with less than s misclassified items. To formalize these results, we introduce the *divergence* of (α, p). We denote by $p(i)$ the $K \times (L+1)$ matrix whose element on the j-th row and the $(\ell + 1)$-th column is $p(i, j, \ell)$ and denote by $p(i, j) \in [0, 1]^{L+1}$ the vector describing the probability distribution of the label of a pair of items in $\mathcal{V}_i$ and $\mathcal{V}_j$, respectively. Let $\mathcal{P}^{K \times (L+1)}$ denote the set of $K \times (L+1)$ matrices such that each row represents a probability distribution. The divergence $D(\alpha, p)$ of (α, p) is defined as follows: $D(\alpha, p) = \min_{i,j:i \neq j} D_{L+}(\alpha, p(i), p(j))$ with

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

where KL denotes the Kullback-Leibler divergence between two label distributions, i.e., $KL(y(k), p(i, k)) = \sum_{\ell=0}^{L} y(k, \ell) \log \frac{y(k,\ell)}{p(i,k,\ell)}$. Finally, we denote by $\varepsilon^\pi(n)$ the number of misclassified items under the clustering algorithm π, and by $\mathbb{E}[\varepsilon^\pi(n)]$ its expectation (with respect to the randomness in the LSBM and in the algorithm).

We first derive a tight lower bound on the average number of misclassified items when the latter is $o(n)$. Note that such a bound was unknown even for the SBM [2].

Theorem 1 *Assume that (A1) and (A2) hold, and that $\bar{p}n = \omega(1)$. Let $s = o(n)$. If there exists a clustering algorithm π misclassifying in average less than s items asymptotically, i.e., $\limsup_{n\to\infty} \frac{\mathbb{E}[\varepsilon^\pi(n)]}{s} \leq 1$, then the parameters (α, p) of the LSBM satisfy:*

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

To state the corresponding positive result (i.e., the existence of an algorithm misclassifying only s items), we make an additional assumption to avoid extremely sparse labels: (A3) there exists a constant $\kappa > 0$ such that $np(j, i, \ell) \geq (n\bar{p})^\kappa$ for all i, j and $\ell \geq 1$.

Theorem 2 *Assume that (A1), (A2), and (A3) hold, and that $\bar{p} = o(1)$, $\bar{p}n = \omega(1)$. Let $s = o(n)$. If the parameters (α, p) of the LSBM satisfy (1), then the Spectral Partition (SP) algorithm presented in Section 4 misclassifies at most s items with high probability, i.e., $\lim_{n\to\infty} \mathbb{P}[\varepsilon^{SP}(n) \leq s] = 1$.*

These theorems indicate that under the LSBM with parameters satisfying (A1) and (A2), the number of misclassified items scales at least as $n\exp(-nD(\alpha, p)(1 + o(1)))$ under any clustering algorithm, irrespective of its complexity. They further establish that the Spectral Partition algorithm reaches this fundamental performance limit under the additional condition (A3). We note that the SP algorithm runs in polynomial time, i.e., it requires $O(n^2\bar{p}\log(n))$ floating-point operations.

We further establish a necessary and sufficient condition on the parameters of the LSBM for the existence of a clustering algorithm recovering the clusters exactly with high probability. Deriving such a condition was also open [2].

Theorem 3 *Assume that (A1) and (A2) hold. If there exists a clustering algorithm that does not misclassify any item with high probability, then the parameters (α, p) of the LSBM satisfy: $\liminf_{n\to\infty} \frac{nD(\alpha, p)}{\log(n)} \geq 1$. If this condition holds, then under (A3), the SP algorithm recovers the clusters exactly with high probability.*

The paper is organized as follows. Section 2 presents the related work and example of application of our results. In Section 3, we sketch the proof of Theorem 1, which leverages change-of-measure and coupling arguments. We present in Section 4 the Spectral Partition algorithm, and analyze its performance (we outline the proof of Theorem 2). All results are proved in details in the supplementary material.

2 Related Work and Applications

2.1 Related work

Cluster recovery in the SBM has attracted a lot of attention recently. We summarize below existing results, and compare them to ours. Results are categorized depending on the targeted level of performance. First, we consider the notion of *detectability*, the lowest level of performance requiring that the extracted clusters are just positively correlated with the true clusters. Second, we look at *asymptotically accurate recovery*, stating that the proportion of misclassified items vanishes as n grows large. Third, we present existing results regarding exact cluster recovery, which means that no item is misclassified. Finally, we report recent work whose objective, like ours, is to characterize the optimal cluster recovery rate.

Detectability. Necessary and sufficient conditions for *detectability* have been studied for the binary symmetric SBM (i.e., $L = 1$, $K = 2$, $\alpha_1 = \alpha_2$, $p(1, 1, 1) = p(2, 2, 1) = \xi$, and $p(1, 2, 1) = p(2, 1, 1) = \zeta$). In the sparse regime where $\xi, \zeta = o(1)$, and for the binary symmetric SBM, the main focus has been on identifying the phase transition threshold (a condition on ξ and ζ for *detectability*: It was conjectured in [5] that if $n(\xi - \zeta) < \sqrt{2n(\xi + \zeta)}$ (i.e., under the threshold), no algorithm can perform better than a simple random assignment of items to clusters, and above the threshold, clusters can partially be recovered. The conjecture was recently proved in [20] (necessary condition), and [18] (sufficient condition). The problem of detectability has been also recently studied in [27] for the asymmetric SBM with more than two clusters of possibly different sizes. Interestingly, it is shown that in most cases, the phase transition for detectability disappears.

The present paper is not concerned with conditions for detectability. Indeed detectability means that only a strictly positive proportion of items can be correctly classified, whereas here, we impose that the proportion of misclassified items vanishes as n grows large.

Asymptotically accurate recovery. 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 [25] and [19]. Using our notion of divergence specialized to the SBM, this condition is $nD(\alpha, p) = \omega(1)$. Our results are more precise since the minimal achievable number of misclassified items is characterized, and apply to a broader setting since they are valid for the generic LSBM.

Asymptotically exact recovery. Conditions for exact cluster recovery in the SBM have been also recently studied. [1, 19, 9] provide a necessary and sufficient condition for asymptotically exact recovery in the binary symmetric SBM. For example, it is shown that when $\xi = \frac{a \log(n)}{n}$ and $\zeta = \frac{b \log(n)}{n}$ for $a > b$, clusters can be recovered exactly if and only if $\frac{a+b}{2} - \sqrt{ab} \geq 1$. In [2, 3], the authors consider a more general SBM corresponding to our LSBM with $L = 1$. They define CH-divergence as:

$$D_+(\alpha, p(i), p(j)) = \frac{n}{\log(n)} \max_{\lambda \in [0,1]} \sum_{k=1}^{K} \alpha_k \left((1-\lambda)p(i,k,1) + \lambda p(j,k,1) - p(i,k,1)^{1-\lambda} p(j,k,1)^{\lambda} \right),$$

and show that $\min_{i \neq j} D_+(\alpha, p(i), p(j)) > 1$ is a necessary and sufficient condition for asymptotically exact reconstruction. The following claim, proven in the supplementary material, relates D_+ to D_{L+}.

Claim 4 *When $\bar{p} = o(1)$, we have for all i, j:*

$$D_{L+}(\alpha, p(i), p(j)) \overset{n \to \infty}{\sim} \max_{\lambda \in [0,1]} \sum_{\ell=1}^{L} \sum_{k=1}^{K} \alpha_k \left((1-\lambda)p(i,k,\ell) + \lambda p(j,k,\ell) - p(i,k,\ell)^{1-\lambda} p(j,k,\ell)^{\lambda} \right).$$

Thus, the results in [2, 3] are obtained by applying Theorem 3 and Claim 4.

In [14], the authors consider a symmetric labeled SBM where communities are balanced (i.e., $\alpha_k = \frac{1}{K}$ for all k) and where label probabilities are simply defined as $p(i,i,\ell) = p(\ell)$ for all i and $p(i,j,\ell) = q(\ell)$ for all $i \neq j$. It is shown that $\frac{nI}{\log(n)} > 1$ is necessary and sufficient for asymptotically exact recovery, where $I = -\frac{2}{K} \log \left(\sum_{\ell=0}^{L} \sqrt{p(\ell)q(\ell)} \right)$. We can relate I to $D(\alpha, p)$:

Claim 5 *In the LSBM with K clusters, if $\bar{p} = o(1)$, and for all i, j, ℓ such that $i \neq j$, $\alpha_i = \frac{1}{K}$, $p(i,i,\ell) = p(\ell)$, and $p(j,k,\ell) = q(\ell)$, we have: $D(\alpha, p) \overset{n \to \infty}{\sim} -\frac{2}{K} \log \left(\sum_{\ell=0}^{L} \sqrt{p(\ell)q(\ell)} \right).$*

Again from this claim, the results derived in [14] are obtained by applying Theorem 3 and Claim 5.

Optimal recovery rate. In [6, 21], the authors consider the binary SBM in the sparse regime where the average degree of items in the graph is $O(1)$, and identify the minimal number of misclassified items for very specific intra- and inter-cluster edge probabilities ξ and ζ. Again the sparse regime is out of the scope of the present paper. [26, 8] are concerned with the general SBM corresponding to our LSBM with $L = 1$, and with regimes where asymptotically accurate recovery is possible. The authors first characterize the optimal recovery rate in a minimax framework. More precisely, they consider a (potentially large) set of possible parameters (α, p), and provide a lower bound on the expected number of misclassified items for the worst parameters in this set. Our lower bound (Theorem 1) is more precise as it is model-specific, i.e., we provide the minimal expected number of misclassified items for a given parameter (α, p) (and for a more general class of models). Then the authors propose a clustering algorithm, with time complexity $O(n^3 \log(n))$, and achieving their minimax recovery rate. In comparison, our algorithm yields an optimal recovery rate $O(n^2 \bar{p} \log(n))$ for any given parameter (α, p), exhibits a lower running time, and applies to the generic LSBM.

2.2 Applications

We provide here a few examples of application of our results, illustrating their versatility. In all examples, $f(n)$ is a function such that $f(n) = \omega(1)$, and a, b are fixed real numbers such that $a > b$.

The binary SBM. Consider the binary SBM where the average item degree is $\Theta(f(n))$, and represented by a LSBM with parameters $L = 1$, $K = 2$, $\alpha = (\alpha_1, 1 - \alpha_1)$, $p(1, 1, 1) = p(2, 2, 1) = \frac{af(n)}{n}$, and $p(1, 2, 1) = p(2, 1, 1) = \frac{bf(n)}{n}$. From Theorems 1 and 2, the optimal number of misclassified vertices scales as $n \exp(-g(\alpha_1, a, b) f(n)(1 + o(1)))$ when $\alpha_1 \leq 1/2$ (w.l.o.g.) and where

$$g(\alpha_1, a, b) := \max_{\lambda \in [0,1]} (1 - \alpha_1 - \lambda + 2\alpha_1 \lambda) a + (\alpha_1 + \lambda - 2\alpha \lambda) b - \alpha_1 a^\lambda b^{(1-\lambda)} - (1 - \alpha_1) a^{(1-\lambda)} b^\lambda.$$

It can be easily checked that $g(\alpha_1, a, b) \geq g(1/2, a, b) = \frac{1}{2}(\sqrt{a} - \sqrt{b})^2$ (letting $\lambda = \frac{1}{2}$). The worst case is hence obtained when the two clusters are of equal sizes. When $f(n) = \log(n)$, we also note that the condition for asymptotically exact recovery is $g(\alpha_1, a, b) \geq 1$.

Recovering a single hidden community. As in [10], consider a random graph model with a hidden community consisting of αn vertices, edges between vertices belonging the hidden community are present with probability $\frac{af(n)}{n}$, and edges between other pairs are present with probability $\frac{bf(n)}{n}$. This is modeled by a LSBM with parameters $K = 2$, $L = 1$, $\alpha_1 = \alpha$, $p(1, 1, 1) = \frac{af(n)}{n}$, and $p(1, 2, 1) = p(2, 1, 1) = p(2, 2, 1) = \frac{bf(n)}{n}$. The minimal number of misclassified items when searching for the hidden community scales as $n \exp(-h(\alpha, a, b) f(n)(1 + o(1)))$ where

$$h(\alpha, a, b) := \alpha \left(a - (a - b) \frac{1 + \log(a - b) - \log(a \log(a/b))}{\log(a/b)} \right).$$

When $f(n) = \log(n)$, the condition for asymptotically exact recovery of the hidden community is $h(\alpha, a, b) \geq 1$.

Optimal sampling for community detection under the SBM. Consider a dense binary symmetric SBM with intra- and inter-cluster edge probabilities a and b. In practice, to recover the clusters, one might not be able to observe the entire random graph, but sample its vertex (here item) pairs as considered in [25]. Assume for instance that any pair of vertices is sampled with probability $\frac{\delta f(n)}{n}$ for some fixed $\delta > 0$, independently of other pairs. We can model such scenario using a LSBM with three labels, namely $\times$, 0 and 1, corresponding to the absence of observation (the vertex pair is not sampled), the observation of the absence of an edge and of the presence of an edge, respectively, and with parameters for all $i, j \in \{1, 2\}$, $p(i, j, \times) = 1 - \frac{\delta f(n)}{n}$, $p(1, 1, 1) = p(2, 2, 1) = a\frac{\delta f(n)}{n}$, and $p(1, 2, 1) = p(2, 1, 1) = b\frac{\delta f(n)}{n}$. The minimal number of misclassified vertices scales as $n \exp(-l(\delta, a, b) f(n)(1 + o(1)))$ where $l := \delta(1 - \sqrt{ab} - \sqrt{(1 - a)(1 - b)})$. When $f(n) = \log(n)$, the condition for asymptotically exact recovery is $l(\alpha, a_+, a_-, b_+, b_-) \geq 1$.

Signed networks. Signed networks [16, 23] are used in social sciences to model positive and negative interactions between individuals. These networks can be represented by a LSBM with three possible labels, namely 0, + and -, corresponding to the absence of interaction, positive and negative interaction, respectively. Consider such LSBM with parameters: $K = 2$, $\alpha_1 = \alpha_2$, $p(1, 1, +) = p(2, 2, +) = \frac{a_+ f(n)}{n}$, $p(1, 1, -) = p(2, 2, -) = \frac{a_- f(n)}{n}$, $p(1, 2, +) = p(2, 1, +) = \frac{b_+ f(n)}{n}$, and $p(1, 2, -) = p(2, 1, -) = \frac{b_- f(n)}{n}$, for some fixed a_+, a_-, b_+, b_- such that $a_+ > b_+$ and $a_- < b_-$. The minimal number of misclassified individuals here scales as $n \exp(-m(\alpha, a_+, a_-, b_+, b_-) f(n)(1 + o(1)))$ where

$$m(\alpha, a_+, a_-, b_+, b_-) := \frac{1}{2} \left((\sqrt{a_+} - \sqrt{b_+})^2 + (\sqrt{a_-} - \sqrt{b_-})^2 \right).$$

When $f(n) = \log(n)$, the condition for asymptotically exact recovery is $l(\alpha, a_+, a_-, b_+, b_-) \geq 1$.

3 Fundamental Limits: Change of Measures through Coupling

In this section, we explain the construction of the proof of Theorem 1. The latter relies on an appropriate *change-of-measure* argument, frequently used to identify upper performance bounds in

online stochastic optimization problems [15]. In the following, we refer to Φ, defined by parameters (α, p), as the true stochastic model under which all the observed random labels are generated, and denote by $\mathbb{P}_\Phi = \mathbb{P}$ (resp. $\mathbb{E}_\Phi[\cdot] = \mathbb{E}[\cdot]$) the corresponding probability measure (resp. expectation). In our change-of-measure argument, we construct a second stochastic model Ψ (whose corresponding probability measure and expectation are $\mathbb{P}_\Psi$ and $\mathbb{E}_\Psi[\cdot]$, respectively). Using a change of measures from $\mathbb{P}_\Phi$ to $\mathbb{P}_\Psi$, we relate the expected number of misclassified items $\mathbb{E}_\Phi[\varepsilon^\pi(n)]$ under any clustering algorithm π to the expected (w.r.t. $\mathbb{P}_\Psi$) log-likelihood ratio $\mathcal{Q}$ of the observed labels under $\mathbb{P}_\Phi$ and $\mathbb{P}_\Psi$. Specifically, we show that, roughly, $\log(n/\mathbb{E}_\Phi[\varepsilon^\pi(n)])$ must be smaller than $\mathbb{E}_\Psi[\mathcal{Q}]$ for n large enough. Now to identify the appropriate second stochastic model Ψ, we proceed in two steps:

First, we guess the scaling of the lower bound of expected number of misclassified items under any algorithm, i.e., we identify its connection to the divergence $D(\alpha, p)$.

Second, we find a model Ψ such that $\mathbb{E}_\Psi[\mathcal{Q}]$ approximates $D(\alpha, p)$ as n grows large. To that aim, we build Ψ from Φ using coupling techniques [17]. This (miraculous) coupling is inspired by the reasoning used in the first step to identify the optimal recovery rate, and is the corner-stone of the proof.

Putting the arguments together, we show that for any clustering algorithm π, $\log(n/\mathbb{E}_\Phi[\varepsilon^\pi(n)])$ must be smaller than $D(\alpha, p)$ as n grows large, which gives Theorem 1. Next, we describe the two steps of the proof as well as the analysis of $\mathcal{Q}$ in more details.

3.1 Guessing the optimal recovery rate

Consider a LSBM with parameters (α, p). The optimal recovery rate is obtained using the following heuristic argument. Assume that (α, p) are known and that all items have been already correctly classified except for $v \in \mathcal{V}_i$. Then, applying the maximum a posteriori probability (MAP) estimator constitutes the best way of classifying v. Let $e(v, \mathcal{V}_k, \ell)$ denote the number of item pairs (v, w) such that $w \in \mathcal{V}_k$ and with observed label ℓ. Further introduce $\mu(v, \mathcal{V}_k) = [e(v, \mathcal{V}_k, \ell)/|\mathcal{V}_k|]_{0 \leq \ell \leq L}$ as the empirical probability vector defined by the label densities between v and $\mathcal{V}_k$. Under the MAP estimator, v is misclassified and assigned to $\mathcal{V}_j$ when

$$
\begin{aligned}
0 \quad > \quad & \log\left(\frac{\alpha_i}{\alpha_j}\right) + \sum_{k=1}^K \sum_{\ell=0}^L e(v, \mathcal{V}_k, \ell) \log\left(\frac{p(i, k, \ell)}{p(j, k, \ell)}\right) \\
= \quad & \log\left(\frac{\alpha_i}{\alpha_j}\right) + \sum_{k=1}^K \sum_{\ell=0}^L |\mathcal{V}_k| \frac{e(v, \mathcal{V}_k, \ell)}{|\mathcal{V}_k|} \left(\log\left(\frac{e(v, \mathcal{V}_k, \ell)/|\mathcal{V}_k|}{p(j, k, \ell)}\right) - \log\left(\frac{e(v, \mathcal{V}_k, \ell)/|\mathcal{V}_k|}{p(i, k, \ell)}\right)\right) \\
= \quad & \log\left(\frac{\alpha_i}{\alpha_j}\right) + \sum_{k=1}^K |\mathcal{V}_k| \left(KL\left(\mu(v, \mathcal{V}_k), p(j, k)\right) - KL\left(\mu(v, \mathcal{V}_k), p(i, k)\right)\right). \quad (2)
\end{aligned}
$$

Observe that the term $\log(\alpha_i/\alpha_j)$ can be neglected in the r.h.s. of (2) as n grows large. Hence, from the definition of $D_{L+}(\alpha, p(i), p(j))$, v is misclassified to $\mathcal{V}_j$ (i.e., (2) holds) when the following event occurs:

$$
\sum_{k=1}^K |\mathcal{V}_k| KL\left(\mu(v, \mathcal{V}_k), p(i, k)\right) \geq n D_{L+}(\alpha, p(i), p(j)).
$$

We can evaluate the probability of this event in the particular case when $\bar{p} = o(\frac{1}{\sqrt{n}})$ using the following claim.

Claim 6 *When* $|\mathcal{V}_k| = \Omega(n)$, $\bar{p} = o(\frac{1}{\sqrt{n}})$, $n\bar{p} = \omega(1)$, *and* $D = \Omega(\bar{p})$,

$\log\left(\mathbb{P}\left\{\sum_{k=1}^K |\mathcal{V}_k| KL\left(\mu(v, \mathcal{V}_k), p(i, k)\right) \geq nD\right\}\right) \overset{n \to \infty}{\sim} -nD.$

Now by choosing $(i, j) = (i^\star, j^\star) = \arg\min_{i,j : i < j} D_{L+}(p(i), p(j))$, we can expect, from the previous claim applied to $D = D(\alpha, n)$, that the probability of misclassifying $v \in \mathcal{V}_{i^\star}$ scales as $\exp(-D(\alpha, n))$. Since $|\mathcal{V}_{i^\star}|$ grows linearly with n, this implies that the expected number of misclassified items scales at least as $n \exp(-nD(\alpha, P)(1 + o(1)))$.

3.2 Construction of ψ

Let $(i^\star, j^\star) = \arg\min_{i,j:i<j} D_{L+}(\alpha, p(i), p(j))$, and let $v^\star$ denote the smallest item index that belongs to cluster $i^\star$ or $j^\star$. If both $\mathcal{V}_{i^\star}$ and $\mathcal{V}_{j^\star}$ are empty, we define $v^\star = n$. Let $q \in \mathcal{P}^{K \times (L+1)}$ such that: $D(\alpha, p) = \sum_{k=1}^K \alpha_k KL(q(k), p(i^\star, k)) = \sum_{k=1}^K \alpha_k KL(q(k), p(j^\star, k))$. The existence of such q is proved in Lemma 8 in the supplementary material. Now to define the stochastic model Ψ, we couple the generation of labels under Φ and Ψ as follows.

1. We first generate the random clusters $\mathcal{V}_1, \ldots, \mathcal{V}_K$ under Φ, and extract $i^\star, j^\star$, and $v^\star$. The clusters generated under Ψ are the same as those generated under Φ. For any $v \in \mathcal{V}$, we denote by $\sigma(v)$ the cluster of item v.

2. For all pairs (v, w) such that $v \neq v^\star$ and $w \neq v^\star$, the labels generated under Ψ are the same as those generated under Φ, i.e., the label ℓ is observed on the edge (v, w) with probability $p(\sigma(v), \sigma(w), \ell)$.

3. Under Ψ, for any $v \neq v^\star$, the observed label on the edge $(v, v^\star)$ under Ψ is ℓ with probability $q(\sigma(v), \ell)$.

Let $x_{v,w}$ denote the label observed for the pair (v, w). We introduce $\mathcal{Q}$, the log-likelihood ratio of the observed labels under $\mathbb{P}_\Phi$ and $\mathbb{P}_\Psi$ as:

$$\mathcal{Q} = \sum_{v=1}^{v^\star - 1} \log \frac{q(\sigma(v), x_{v^\star, v})}{p(\sigma(v^\star), \sigma(v), x_{v^\star, v})} + \sum_{v=v^\star+1}^{n} \log \frac{q(\sigma(v), x_{v^\star, v})}{p(\sigma(v^\star), \sigma(v), x_{v^\star, v})}. \tag{3}$$

Let π be a clustering algorithm with output $(\hat{\mathcal{V}}_k)_{1 \leq k \leq K}$, and let $\mathcal{E} = \bigcup_{1 \leq k \leq K} \hat{\mathcal{V}}_k \setminus \mathcal{V}_k$ be the set of misclassified items under π. Note that in general in our analysis, we always assume without loss of generality that $|\bigcup_{1 \leq k \leq K} \hat{\mathcal{V}}_k \setminus \mathcal{V}_k| \leq |\bigcup_{1 \leq k \leq K} \hat{\mathcal{V}}_{\gamma(k)} \setminus \mathcal{V}_k|$ for any permutation γ, so that the set of misclassified items is indeed $\mathcal{E}$. By definition, $\varepsilon^\pi(n) = |\mathcal{E}|$. Since under Φ, items are interchangeable (remember that items are assigned to the various clusters in an i.i.d. manner), we have: $n \mathbb{P}_\Phi \{v \in \mathcal{E}\} = \mathbb{E}_\Phi[\varepsilon^\pi(n)] = \mathbb{E}[\varepsilon^\pi(n)]$.

Next, we establish a relationship between $\mathbb{E}[\varepsilon^\pi(n)]$ and the distribution of $\mathcal{Q}$ under $\mathbb{P}_\Psi$. For any function $f(n)$, we can prove that: $\mathbb{P}_\Psi \{\mathcal{Q} \leq f(n)\} \leq \exp(f(n)) \frac{\mathbb{E}_\Phi[\varepsilon^\pi(n)]}{(\alpha_{i^\star} + \alpha_{j^\star})n} + \frac{\alpha_{j^\star}}{\alpha_{i^\star} + \alpha_{j^\star}}$. Using this result with $f(n) = \log(n/\mathbb{E}_\Phi[\varepsilon^\pi(n)]) - \log(2/\alpha_{i^\star})$, and Chebyshev's inequality, we deduce that: $\log(n/\mathbb{E}_\Phi[\varepsilon^\pi(n)]) - \log(2/\alpha_{i^\star}) \leq \mathbb{E}_\Psi[\mathcal{Q}] + \sqrt{\frac{4}{\alpha_{i^\star}} \mathbb{E}_\Psi[(\mathcal{Q} - \mathbb{E}_\Psi[\mathcal{Q}])^2]}$, and thus, a necessary condition for $\mathbb{E}[\varepsilon^\pi(n)] \leq s$ is:

$$\log(n/s) - \log(2/\alpha_{i^\star}) \leq \mathbb{E}_\Psi[\mathcal{Q}] + \sqrt{\frac{4}{\alpha_{i^\star}} \mathbb{E}_\Psi[(\mathcal{Q} - \mathbb{E}_\Psi[\mathcal{Q}])^2]}. \tag{4}$$

3.3 Analysis of $\mathcal{Q}$

In view of (4), we can obtain a necessary condition for $\mathbb{E}[\varepsilon^\pi(n)] \leq s$ if we evaluate $\mathbb{E}_\Psi[\mathcal{Q}]$ and $\mathbb{E}_\Psi[(\mathcal{Q} - \mathbb{E}_\Psi[\mathcal{Q}])^2]$. To evaluate $\mathbb{E}_\Psi[\mathcal{Q}]$, we can first prove that $v^\star \leq \log(n)^2$ with high probability. From this, we can approximate $\mathbb{E}_\Psi[\mathcal{Q}]$ by $\mathbb{E}_\Psi[\sum_{v=v^\star+1}^{n} \log \frac{q(\sigma(v), x_{v^\star, v})}{p(\sigma(v^\star), \sigma(v), x_{v^\star, v})}]$, which is itself well-approximated by $n D(\alpha, p)$. More formally, we can show that:

$$\mathbb{E}_\Psi[\mathcal{Q}] \leq \left(n + 2\log(\eta)\log(n)^2\right) D(\alpha, p) + \frac{\log \eta}{n^3}. \tag{5}$$

Similarly, we prove that $\mathbb{E}_\Psi[(\mathcal{Q} - \mathbb{E}_\Psi[\mathcal{Q}])^2] = O(n\bar{p})$, which in view of Lemma 9 (refer to the supplementary material) and assumption (A2), implies that: $\mathbb{E}_\Psi[(\mathcal{Q} - \mathbb{E}_\Psi[\mathcal{Q}])^2] = o(n D(\alpha, p))$.

We complete the proof of Theorem 1 by putting the above arguments together: From (4), (5) and the above analysis of $\mathcal{Q}$, when the expected number of misclassified items is less than s (i.e., $\mathbb{E}[\varepsilon^\pi(n)] \leq s$), we must have: $\liminf_{n \to \infty} \frac{n D(\alpha, p)}{\log(n/s)} \geq 1$.

4 The Spectral Partition Algorithm and its Optimality

In this section, we sketch the proof of Theorem 2. To this aim, we present the Spectral Partition (SP) algorithm and analyze its performance. The main pseudo-code of SP is presented in Algorithm 1. The

SP algorithm consists in two parts. In the first part, corresponding to Lines 1-4 in the pseudo-code, we apply a spectral decomposition of the matrix $A = \sum_{\ell=1}^{L} w_\ell A^\ell$ constructed from the observed labels. This matrix is first trimmed, and then treated by applying the spectral decomposition algorithm, whose pseudo-code is presented in Algorithm 2. The second part of the SP algorithm, corresponding to Lines 5 and 6 in Algorithm 1, consists in improving the clusters initially identified in the first step.

Algorithm 1 Spectral Partition

Input: Observation matrices A^ℓ for every label ℓ ($A^\ell_{uv} = 1$ if ℓ is observed between u and v).

1. Estimated average degree. $\tilde{p} \leftarrow \frac{\sum_{\ell=1}^{L} \sum_{u,v} A^\ell_{uv}}{n(n-1)}$

2. Random Weights. $A \leftarrow \sum_{\ell=1}^{L} w_\ell A^\ell$ where the weights w_ℓ's are i.i.d and uniformly distributed on $[0,1]$.

3. Trimming. Construct $A_\Gamma = (A_{vw})_{v,w \in \Gamma}$, where Γ is the set of nodes obtained after removing $\lfloor n \exp(-n\tilde{p}) \rfloor$ nodes having the largest $\sum_\ell \sum_{w \in V} A^\ell_{vw}$.

4. Spectral Decomposition. Run Algorithm 2 with input $A_\Gamma, \tilde{p}$, and output $(S_k)_{k=1,\ldots,\hat{K}}$.

5. Estimated 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 $1 \leq i,j \leq \hat{K}$ and $0 \leq \ell \leq L$.

6. Improvement.
$S_k^{(0)} \leftarrow S_k$, for all k
for $t = 1$ **to** $\log n$ **do**
    $S_k^{(t)} \leftarrow \emptyset$, for all k
    for $v \in \mathcal{V}$ **do**
        Find $k^\star = \arg\max_{1 \leq k \leq \hat{K}} \{\sum_{i=1}^{\hat{K}} \sum_{w \in S_i^{(t-1)}} \sum_{\ell=0}^{L} A^\ell_{vw} \log \hat{p}(k,i,\ell)\}$ (tie broken uniformly at random)
        $S_{k^\star}^{(t)} \leftarrow S_{k^\star}^{(t)} \cup \{v\}$
    end for
end for
$\hat{\mathcal{V}}_k \leftarrow S_k^{(\log n)}$, for all k
Output: $(\hat{\mathcal{V}}_k)_{k=1,\ldots,\hat{K}}$.

The first part of the algorithm can be interpreted as an initialization for its second part, and consists in applying a spectral decomposition of a $n \times n$ random matrix A constructed from the observed labels. More precisely, $A = \sum_{\ell=1}^{L} w_\ell A^\ell$, where A^ℓ is the binary matrix identifying the item pairs with observed label ℓ, i.e., for all $v, w \in \mathcal{V}$, $A^\ell_{vw} = 1$ if and only if (v,w) has label ℓ. The weight w_ℓ for label $\ell \in \{1, \ldots, L\}$ is generated uniformly at random in $[0,1]$, independently of other weights. From the spectral decomposition of A, we estimate the number of communities and provide asymptotically accurate estimates $S_1, \ldots, S_K$ of the hidden clusters asymptotically accurately, i.e., we show that when $n\bar{p} = \omega(1)$, with high probability, $\hat{K} = K$ and there exists a permutation γ of $\{1, \ldots, K\}$ such that $\frac{1}{n} \left| \cup_{k=1}^{K} \mathcal{V}_k \setminus S_{\gamma(k)} \right| = O\left(\frac{\log(n\bar{p})^2}{n\bar{p}} \right)$. This first part of the SP algorithm is adapted from algorithms proposed for the standard SBM in [4, 25] to handle the additional labels in the model without the knowledge of the number K of clusters.

The second part is novel, and is critical to ensure the optimality of the SP algorithm. It consists in first constructing an estimate $\hat{p}$ of the true parameters p of the model from the matrices $(A^\ell)_{1 \leq \ell \leq L}$ and the estimated clusters $S_1, \ldots, S_K$ provided in the first part of SP. We expect p to be well estimated since $S_1, \ldots, S_K$ are asymptotically accurate. Then our cluster estimates are iteratively improved. We run $\lfloor \log(n) \rfloor$ iterations. Let $S_1^{(t)}, \ldots, S_K^{(t)}$ denote the clusters estimated after the t-th iteration, initialized with $(S_1^{(0)}, \ldots, S_K^{(0)}) = (S_1, \ldots, S_K)$. The improved clusters $S_1^{(t+1)}, \ldots, S_K^{(t+1)}$ are obtained by assigning each item $v \in \mathcal{V}$ to the cluster maximizing a log-likelihood formed from $\hat{p}, S_1^{(t)}, \ldots, S_K^{(t)}$, and the observations $(A^\ell)_{1 \leq \ell \leq L}$: v is assigned to $S_{k^\star}^{(t+1)}$ where $k^\star = \arg\max_k \{\sum_{i=1}^{K} \sum_{w \in S_i^{(t-1)}} \sum_{\ell=0}^{L} A^\ell_{vw} \log \hat{p}(k,i,\ell)\}$.

Part 1: Spectral Decomposition. The spectral decomposition is described in Lines 1 to 4 in Algorithm 1. As usual in spectral methods, the matrix A is first trimmed (to remove lines and columns

---

Algorithm 2 Spectral decomposition

Input: $A_\Gamma, \tilde{p}$

1. Iterative Power Method with singular value thresholding

(Initialization) $\chi \leftarrow n$, $k \leftarrow 0$, and $\hat{U} \leftarrow 0^{n \times 1}$

while $\chi \geq \sqrt{n\tilde{p}} \log(n\tilde{p})$ **do**

   $k \leftarrow k + 1$, $U_0 \leftarrow n \times 1$ Gaussian random vector

   (Iterative power method) $U_t \leftarrow (A_\Gamma)^{\lceil 2 \log(n) \rceil} U_0$

   (Orthonormalizing U_t) $\hat{U}_k \leftarrow \dfrac{U_t - \hat{U}_{1:k-1}(\hat{U}_{1:k-1}^\top U_t)}{\|U_t - \hat{U}_{1:k-1}(\hat{U}_{1:k-1}^\top U_t)\|_2}$

   (Estimating the k-th singular value) $\chi \leftarrow \|A_\Gamma \hat{U}_k\|_2$

end while

$\tilde{K} \leftarrow k - 1$, $\hat{V} \leftarrow \hat{U}_{1:\tilde{K}}^\top A_\Gamma$

2. Clustering

$\mathcal{V}_R \leftarrow$ a subset of Γ obtained by randomly selecting $\lceil \log(n) \rceil$ items of Γ

$Q_v \leftarrow \{w \in \Gamma : \|\hat{V}_w - \hat{V}_v\|_2^2 \leq \frac{n\tilde{p}^2}{\log(n\tilde{p})}\}$ for all $v \in \mathcal{V}_R$

(Initialization) $S_0 \leftarrow \emptyset$, $k \leftarrow 0$, and $\rho \leftarrow |\Gamma|$

while $\rho \geq \frac{\log(n\tilde{p})^4}{\tilde{p}}$ **do**

   $k \leftarrow k + 1$, $v_k^\star \leftarrow \arg\max_{v \in \mathcal{V}_R} |Q_v \setminus \bigcup_{l=0}^{k-1} S_l|$, $S_k \leftarrow Q_{v_k^\star} \setminus \bigcup_{l=0}^{k-1} S_l$ and $\rho \leftarrow |S_k|$.

end while

$\hat{K} \leftarrow k - 1$

for $v \in \Gamma \setminus \bigcup_{k=1}^{\hat{K}} S_k$ **do**

   $k_\star \leftarrow \arg\min_k \|\hat{V}_{v_k^\star} - \hat{V}_v\|_2$, $S_{k_\star} \leftarrow S_{k_\star} \cup \{v\}$

end for

Output: $(S_k)_{k=1,\ldots,\hat{K}}$.

---

corresponding to items with too many observed labels – as they would perturb the spectral analysis). To this aim, we estimate the average number of labels per item, and use this estimate, denoted by $\tilde{p}$ in Algorithm 1, as a reference for the trimming process. Γ and A_Γ denote the set of remaining items after trimming, and the corresponding trimmed matrix, respectively.

If the number of clusters K is known and if we do not account for time complexity, the two step algorithm in [4] can extract the clusters from A_Γ: first the optimal rank-K approximation $A^{(K)}$ of A_Γ is derived using the SVD; then, one applies the k-mean algorithm to the columns of $A^{(K)}$ to reconstruct the clusters. The number of misclassified items after this two step algorithm is obtained as follows. Let $M^\ell = \mathbb{E}[A_\Gamma^\ell]$, and $M = \sum_{\ell=1}^L w_\ell M^\ell$ (using the same weights as those defining A). Then, M is of rank K. If v and w are in the same cluster, $M_v = M_w$ and if v and w do not belong to the same cluster, from (A2), we must have with high probability: $\|M_v - M_w\|_2 = \Omega(\bar{p}\sqrt{n})$. Thus, the k-mean algorithm misclassifies v only if $\|A_v^{(K)} - M_v\|_2 = \Omega(\bar{p}\sqrt{n})$. By leveraging elements of random graph and random matrix theories, we can establish that $\sum_v \|A_v^{(k)} - M_v\|_2^2 = \|A^{(k)} - M\|_F^2 = O(n\bar{p})$ with high probability. Hence the algorithm misclassifies $O(1/\bar{p})$ items with high probability.

Here the number of clusters K is not given a-priori. In this scenario, Algorithm 2 estimates the rank of M using a singular value thresholding procedure. To reduce the complexity of the algorithm, the singular values and singular vectors are obtained using the iterative power method instead of a direct SVD. It is known from [11] that with $\Theta(\log(n))$ iterations, the iterative power method find singular values and the rank-K approximation very accurately. Hence, when $n\bar{p} = \omega(1)$, we can easily estimate the rank of M by looking at the number of singular values above the threshold $\sqrt{n\tilde{p}} \log(n\tilde{p})$, since we know from random matrix theory that the $(K + 1)$-th singular value of A_Γ is much less than $\sqrt{n\tilde{p}} \log(n\tilde{p})$ with high probability. In the pseudo-code of Algorithm 2, the estimated rank of M is denoted by $\tilde{K}$.

The rank-$\tilde{K}$ approximation of A_Γ obtained by the iterative power method is $\hat{A} = \hat{U}\hat{V} = \hat{U}\hat{U}^\top A_\Gamma$. From the columns of $\hat{A}$, we can estimate the number of clusters and classify items. Almost every

column of $\hat{A}$ is located around the corresponding column of M within a distance $\frac{1}{2}\sqrt{\frac{n\bar{p}^2}{\log(n\bar{p})}}$, since $\sum_v \|\hat{A}_v - M_v\|_2^2 = \|\hat{A} - M\|_F^2 = O(n\bar{p}\log(n\bar{p})^2)$ with high probability (we rigorously analyze this distance in the supplementary material Section C.2). From this observation, the columns can be categorised into K groups. To find these groups, we randomly pick $\log(n)$ reference columns and for each reference column, search all columns within distance $\sqrt{\frac{n\bar{p}^2}{\log(n\bar{p})}}$. Then, with high probability, each cluster has at least one reference column and each reference column can find most of its cluster members. Finally, the K groups are identified using the reference columns. To this aim, we compute the distance of $n\log(n)$ column pairs $\hat{A}_v, \hat{A}_w$. Observe that $\|\hat{A}_v - \hat{A}_w\|_2 = \|\hat{V}_v - \hat{V}_w\|_2$ for any $u, v \in \Gamma$, since the columns of $\hat{U}$ are orthonormal. Now $\hat{V}_v$ is of dimension $\tilde{K}$, and hence we can identify the groups using $O(n\tilde{K}\log(n))$ operations.

Theorem 7 *Assume that (A1) and (A2) hold, and that $n\bar{p} = \omega(1)$. After Step 4 (spectral decomposition) in the SP algorithm, with high probability, $\hat{K} = K$ and there exists a permutation γ of $\{1, \dots, K\}$ such that:* $\left| \cup_{k=1}^{K} \mathcal{V}_k \setminus S_{\gamma(k)} \right| = O\left(\frac{\log(n\bar{p})^2}{\bar{p}}\right).$

Part 2: Successive clusters improvements. Part 2 of the SP algorithm is described in Lines 5 and 6 in Algorithm 1. To analyze the performance of each improvement iteration, we introduce the set of items H as the largest subset of $\mathcal{V}$ such that for all $v \in H$: (H1) $e(v, \mathcal{V}) \leq 10\eta n\bar{p}L$; (H2) when $v \in \mathcal{V}_k$, $\sum_{i=1}^{K}\sum_{\ell=0}^{L} e(v, \mathcal{V}_i, \ell) \log \frac{p(k,i,\ell)}{p(j,i,\ell)} \geq \frac{n\bar{p}}{\log(n\bar{p})^4}$ for all $j \neq k$; (H3) $e(v, \mathcal{V} \setminus H) \leq 2\log(n\bar{p})^2$, where for any $S \subset \mathcal{V}$ and ℓ, $e(v, S, \ell) = \sum_{w \in S} A_{vw}^{\ell}$, and $e(v, S) = \sum_{\ell=1}^{L} e(v, S, \ell)$. Condition (H1) means that there are not too many observed labels $\ell \geq 1$ on pairs including v, (H2) means that an item $v \in \mathcal{V}_k$ must be classified to $\mathcal{V}_k$ when considering the log-likelihood, and (H3) states that v does not share too many labels with items outside H.

We then prove that $|\mathcal{V} \setminus H| \leq s$ with high probability when $nD(\alpha, p) - \frac{n\bar{p}}{\log(n\bar{p})^3} \geq \log(n/s) + \sqrt{\log(n/s)}$. This is mainly done using concentration arguments to relate the quantity $\sum_{i=1}^{K}\sum_{\ell=0}^{L} e(v, \mathcal{V}_i, \ell) \log \frac{p(k,i,\ell)}{p(j,i,\ell)}$ involved in (H2) to $nD(\alpha, p)$.

Finally, we establish that if the clusters provided after the first part of the SP algorithm are asymptotically accurate, then after $\log(n)$ improvement iterations, there is no misclassified items in H. To that aim, we denote by $\mathcal{E}^{(t)}$ the set of misclassified items after the t-th iteration, and show that with high probability, for all t, $\frac{|\mathcal{E}^{(t+1)} \cap H|}{|\mathcal{E}^{(t)} \cap H|} \leq \frac{1}{\sqrt{n\bar{p}}}$. This completes the proof of Theorem 2, since after $\log(n)$ iterations, the only misclassified items are those in $\mathcal{V} \setminus H$.

Acknowledgments

References

[1] E. Abbe, A. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *CoRR*, abs/1405.3267, 2014.

[2] E. Abbe and C. Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. In *FOCS*, 2015.

[3] E. Abbe and C. Sandon. Recovering communities in the general stochastic block model without knowing the parameters. In *NIPS*, 2015.

[4] A. Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability & Computing*, 19(2):227–284, 2010.

[5] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Inference and phase transitions in the detection of modules in sparse networks. *Phys. Rev. Lett.*, 107, Aug 2011.

[6] Y. Deshpande, E. Abbe, and A. Montanari. Asymptotic mutual information for the two-groups stochastic block model. *CoRR*, abs/1507.08685, 2015.

[7] U. Feige and E. Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

[8] C. Gao, Z. Ma, A. Zhang, and H. Zhou. Achieving optimal misclassification proportion in stochastic block model. *CoRR*, abs/1505.03772, 2015.

[9] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *CoRR*, abs/1412.6156, 2014.

[10] B. Hajek, Y. Wu, and J. Xu. Information limits for recovering a hidden community. *CoRR*, abs/1509.07859, 2015.

[11] N. Halko, P. Martinsson, and J. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[12] S. Heimlicher, M. Lelarge, and L. Massoulié. Community detection in the labelled stochastic block model. In *NIPS Workshop on Algorithmic and Statistical Approaches for Large Social Networks*, 2012.

[13] P. Holland, K. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109 – 137, 1983.

[14] V. Jog and P. Loh. Information-theoretic bounds for exact recovery in weighted stochastic block models using the renyi divergence. *CoRR*, abs/1509.06418, 2015.

[15] T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[16] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *CHI*, 2010.

[17] T. Lindvall. *Lectures on the Coupling Method*. Dover Books on Mathematics Series. Dover Publications, Incorporated, 2002.

[18] L. Massoulié. Community detection thresholds and the weak ramanujan property. In *STOC*, 2014.

[19] E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for binary symmetric block models. In *STOC*, 2015.

[20] E. Mossel, J. Neeman, and A. Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015.

[21] E. Mossel and J. Xu. Density evolution in the degree-correlated stochastic block model. *CoRR*, abs/1509.03281, 2015.

[22] T. Tao. *Topics in random matrix theory*, volume 132. AMS Bookstore, 2012.

[23] V. Traag and J. Bruggeman. Community detection in networks with positive and negative links. *Physical Review E*, 80(3):036115, 2009.

[24] S. Yun and A. Proutiere. Accurate community detection in the stochastic block model via spectral algorithms. *CoRR*, abs/1412.7335, 2014.

[25] S. Yun and A. Proutiere. Community detection via random and adaptive sampling. In *COLT*, 2014.

[26] A. Zhang and H. Zhou. Minimax rates of community detection in stochastic block models. *CoRR*, abs/1507.05313, 2015.

[27] P. Zhang, C. Moore, and M. Newman. Community detection in networks with unequal groups. *CoRR*, abs/1509.00107, 2015.

A   Properties of the divergence $D(\alpha, p)$ and related quantities

In this section, we prove the two claims of Section 2, as well as other results on the divergence $D(\alpha, p)$ that will be instrumental in the proofs of Theorems.

A.1   Proof of Claim 4

$D_{L+}(p(i), p(j))$ is the minimum of the objective function of the following convex optimization problem:

$$\min_{y \in \mathcal{P}^{K \times (L+1)}} \quad \sum_{k=1}^{K} \alpha_k \left(\sum_{\ell=1}^{L} y(k, \ell) \log \left(\frac{y(k, \ell)}{p(i, k, \ell)} \right) + (1 - \sum_{\ell=1}^{L} y(k, \ell)) \log \left(\frac{1 - \sum_{\ell=1}^{L} y(k, \ell)}{1 - \sum_{\ell=1}^{L} p(i, k, \ell)} \right) \right)$$

$$\text{s.t.} \quad \sum_{k=1}^{K} \alpha_k KL(y(k), p(i, k)) \geq \sum_{k=1}^{K} \alpha_k KL(y(k), p(j, k)).$$

$$(6)$$

Note that we define $y(k, 0) = 1 - \sum_{\ell=1}^{L} y(k, \ell)$ for all k. Since $\bar{p} = o(1)$, one can easily check that the solution of (6) has to be $\sum_{\ell=1}^{L} y(k, \ell) = o(1)$ for all k. The objective function converges to infinity when $\sum_{\ell=1}^{L} y(k, \ell) = \Omega(1)$, while it has $o(\bar{p})$ when $y(k, \ell) = p(j, k, \ell)$ for all k and ℓ. Thus, we consider $\sum_{\ell=1}^{L} y(k, \ell) = o(1)$. The associated Lagrangian is:

$$g(y, \lambda) = \sum_{k=1}^{K} \alpha_k \left(\sum_{\ell=1}^{L} y(k, \ell) \log \left(\frac{y(k, \ell)}{p(i, k, \ell)} \right) + (1 - \sum_{\ell=1}^{L} y(k, \ell)) \log \left(\frac{1 - \sum_{\ell=1}^{L} y(k, \ell)}{1 - \sum_{\ell=1}^{L} p(i, k, \ell)} \right) \right) +$$

$$\sum_{k=1}^{K} \alpha_k \lambda \left(\sum_{\ell=1}^{L} y(k, \ell) \log \left(\frac{p(i, k, \ell)}{p(j, k, \ell)} \right) + (1 - \sum_{\ell=1}^{L} y(k, \ell)) \log \left(\frac{1 - \sum_{\ell=1}^{L} p(i, k, \ell)}{1 - \sum_{\ell=1}^{L} p(j, k, \ell)} \right) \right).$$

$$(7)$$

The derivative of $g(y, \lambda)$ w.r.t. $y(k, \ell)$ is computed as follows:

$$\frac{\partial g(y, \lambda)}{\partial y(k, \ell)} = \alpha_k \left(\log \left(\frac{y(k, \ell)}{p(i, k, \ell)} \right) - \log \left(\frac{1 - \sum_{m=1}^{L} y(k, m)}{1 - \sum_{m=1}^{L} p(i, k, m)} \right) \right) +$$

$$\alpha_k \lambda \left(\log \left(\frac{p(i, k, \ell)}{p(j, k, \ell)} \right) - \log \left(\frac{1 - \sum_{m=1}^{L} p(i, k, m)}{1 - \sum_{m=1}^{L} p(j, k, m)} \right) \right).$$

Observe that, since (A1) holds, $\bar{p} = o(1)$ and $\sum_{\ell=1}^{L} y(k, \ell) = o(1)$, as n grows large, $\log \left(\frac{1 - \sum_{m=1}^{L} y(k, m)}{1 - \sum_{m=1}^{L} p(i, k, m)} \right)$ and $\log \left(\frac{1 - \sum_{m=1}^{L} p(i, k, m)}{1 - \sum_{m=1}^{L} p(j, k, m)} \right)$ converges to 0. Thus, (7) is minimized at

$$y(k, \ell) = p(i, k, \ell) \left(\frac{p(j, k, \ell)}{p(i, k, \ell)} \right)^{\lambda} (1 + o(1)).$$

$$(8)$$

When we put (8) onto (7) and use the approximation $\lim_{x \to 0} \log(1 + x) = x$ (again using $\bar{p} = o(1)$),

$$\min_{y \in \mathcal{P}^{K \times \{0,1\}}} g(y, \lambda)$$

$$= \min_{y \in \mathcal{P}^{K \times \{0,1\}}} \sum_{k=1}^{K} \sum_{\ell=1}^{L} \alpha_k \left(o(\bar{p}) + \right.$$

$$(1 - \sum_{\ell=1}^{L} y(k, \ell)) \log \left(\frac{1 - \sum_{\ell=1}^{L} y(k, \ell)}{1 - \sum_{\ell=1}^{L} p(i, k, \ell)} \right) \left(\frac{1 - \sum_{\ell=1}^{L} p(i, k, \ell)}{1 - \sum_{\ell=1}^{L} p(j, k, \ell)} \right)^{\lambda} \right)$$

$$= \min_{y \in \mathcal{P}^{K \times \{0,1\}}} \sum_{k=1}^{K} \sum_{\ell=1}^{L} \alpha_k \left(o(\bar{p}) - \right.$$

$$\sum_{\ell=1}^{L} y(k,\ell)(1+o(1)) + (1-\lambda)\sum_{\ell=1}^{L} p(i,k,\ell)(1+o(1)) + \lambda\sum_{\ell=1}^{L} p(j,k,\ell)(1+o(1))\Bigg).$$

Therefore, the minimum value of (6) is equivalent to

$$\max_{\lambda\in[0,1]} \sum_{k=1}^{K}\sum_{\ell=1}^{L} \alpha_k \left((1-\lambda)p(i,k,\ell) + \lambda p(j,k,\ell) - p(i,k,1)^{1-\lambda}p(j,k,\ell)^{\lambda}\right) + o(\bar{p}).$$

A.2 Proof of Claim 5

When $\bar{p} = o(1)$, for all $i \neq j$, $\alpha_i = \frac{1}{K}$, $p(i,i,\ell) = p(\ell)$, and $p(i,j,\ell) = q(\ell)$, from Claim 4,

$$
\begin{aligned}
D_{L+}(\alpha, p(i), p(j)) &= \max_{\lambda\in[0,1]} \sum_{k=1}^{K}\sum_{\ell=1}^{L} \alpha_k \left((1-\lambda)p(i,k,\ell) + \lambda p(j,k,\ell) - p(i,k,\ell)^{1-\lambda}p(j,k,\ell)^{\lambda}\right) \\
&= \frac{1}{K} \max_{\lambda\in[0,1]} \sum_{\ell=1}^{L} \left(p(\ell) + q(\ell) - p(\ell)^{1-\lambda}q(\ell)^{\lambda} - p(\ell)^{\lambda}q(\ell)^{1-\lambda}\right) \\
&= \frac{1}{K} \sum_{\ell=1}^{L} \left(p(\ell) + q(\ell) - 2\sqrt{p(\ell)q(\ell)}\right). \tag{9}
\end{aligned}
$$

Now, since $\sqrt{1+x} = 1 + \frac{x}{2}(1+o(1))$ and $\log(1+x) = x(1+o(1))$ when $x = o(1)$,

$$
\begin{aligned}
-\frac{2}{K}\log\left(\sum_{\ell=0}^{L}\sqrt{p(\ell)q(\ell)}\right) &= -\frac{2}{K}\log\left(\sqrt{p(0)q(0)} + \sum_{\ell=1}^{L}\sqrt{p(\ell)q(\ell)}\right) \\
&= -\frac{2}{K}\log\left(1 - \frac{\sum_{\ell=1}^{L}p(\ell) + q(\ell)}{2}(1+o(1)) + \sum_{\ell=1}^{L}\sqrt{p(\ell)q(\ell)}\right) \\
&= \frac{2}{K}\left(\frac{\sum_{\ell=1}^{L}p(\ell) + q(\ell)}{2} - \sum_{\ell=1}^{L}\sqrt{p(\ell)q(\ell)}\right)(1+o(1)). \tag{10}
\end{aligned}
$$

The claim follows from (9) and (10).

A.3 Other properties

Lemma 8 *Let $(i^\star, j^\star) = \arg\min_{i,j} D_{L+}(p(i), p(j))$ and $i^\star < j^\star$. Then, there exists $q \in \mathcal{P}^{K\times(L+1)}$ such that*

$$D(\alpha, p) = \sum_{k=1}^{K} \alpha_k KL(q(k), p(i^\star, k)) = \sum_{k=1}^{K} \alpha_k KL(q(k), p(j^\star, k)).$$

Proof. We check by contradiction that such a q exists. Indeed, assume that

$$D(\alpha, p) = \sum_{k=1}^{K} \alpha_k KL(q(k), p(i^\star, k)) > \sum_{k=1}^{K} \alpha_k KL(q(k), p(j^\star, k)).$$

Then there exists k_0 such that $KL(q(k_0), p(i^\star, k_0)) > KL(q(k_0), p(j^\star, k_0))$. Observe that by positivity of the KL divergence, $q(k_0) \neq p(i^\star, k_0)$. Hence by continuity of the KL divergence, we can construct q' such that $q(k) = q'(k)$ for all $k \neq k_0$, and such that: $KL(q(k_0), p(i^\star, k_0)) - \epsilon < KL(q'(k_0), p(i^\star, k_0)) < KL(q(k_0), p(i^\star, k_0))$ and $KL(q'(k_0), p(j^\star, k_0)) < KL(q(k_0), p(j^\star, k_0)) + \epsilon$ for some $0 < \epsilon < (KL(q(k_0), p(i^\star, k_0)) - KL(q(k_0), p(j^\star, k_0)))/2$. With this choice of q', we get:

$$D(\alpha, p) > \sum_{k=1}^{K} \alpha_k KL(q'(k), p(i^\star, k)) > \sum_{k=1}^{K} \alpha_k KL(q'(k), p(j^\star, k)),$$

which contradicts the definition of $D(\alpha, p)$. ∎

Lemma 9 *When $\bar{p} = o(1)$,*

$$\lim_{n \to \infty} \frac{D(\alpha, p)}{\sum_{k=1}^{K} \frac{\alpha_k}{2} \left(\sum_{\ell=1}^{L} (\sqrt{p(i^\star, k, \ell)} - \sqrt{p(j^\star, k, \ell)})^2 \right)} \geq 1.$$

Proof. Let $(i^\star, j^\star) = \arg\min_{i,j} D_{L+}(\alpha, p(i), p(j))$ and $i^\star < j^\star$. From Lemma 8, there exists q satisfying that

$$D(\alpha, p) = \sum_{k=1}^{K} \alpha_k KL(q(k), p(i^\star, k)) = \sum_{k=1}^{K} \alpha_k KL(q(k), p(j^\star, k)).$$

Then,

$$
\begin{aligned}
nD(\alpha, p) &= n \frac{\sum_{k=1}^{K} \left(\alpha_k KL(q(k), p(i^\star, k)) + \alpha_k KL(q(k), p(j^\star, k)) \right)}{2} \\
&= -n \sum_{k=1}^{K} \alpha_k \sum_{\ell=0}^{L} q(k, \ell) \log \left(\frac{\sqrt{p(i^\star, k, \ell) p(j^\star, k, \ell)}}{q(k, \ell)} \right) \\
&\geq n \sum_{k=1}^{K} \alpha_k \sum_{\ell=0}^{L} \left(q(k, \ell) - \sqrt{p(i^\star, k, \ell) p(j^\star, k, \ell)} \right) \\
&= n \sum_{k=1}^{K} \alpha_k \left(\frac{\sum_{\ell=1}^{L} (p(i^\star, k, \ell) + p(j^\star, k, \ell))}{2} - \sum_{\ell=1}^{L} \sqrt{p(i^\star, k, \ell) p(j^\star, k, \ell)} \right) (1 - o(1)) \\
&= n \sum_{k=1}^{K} \frac{\alpha_k}{2} \left(\sum_{\ell=1}^{L} (\sqrt{p(i^\star, k, \ell)} - \sqrt{p(j^\star, k, \ell)})^2 \right) (1 - o(1)).
\end{aligned}
$$

∎

Lemma 10 *Under condition (A1), when $\bar{p} = o(1)$, $\limsup_{n \to \infty} \frac{D(\alpha, p)}{\eta \bar{p} L} \leq 1$.*

Proof. From the definition of $D(\alpha, p)$, for any $i \neq j$,

$$
\begin{aligned}
D(\alpha, p) &\leq \max \left\{ \sum_{k=1}^{K} \alpha_k KL(p(i, k), p(i, k)), \sum_{k=1}^{K} \alpha_k KL(p(i, k), p(j, k)) \right\} \\
&= \sum_{k=1}^{K} \alpha_k KL(p(i, k), p(j, k)) \\
&\leq \sum_{k=1}^{K} \alpha_k \sum_{\ell=1}^{L} \frac{(p(i, k, \ell) - p(j, k, \ell))^2}{p(j, k, \ell)} (1 + o(1)) \\
&\leq \sum_{k=1}^{K} \alpha_k \sum_{\ell=1}^{L} \eta \bar{p} (1 + o(1)) \\
&= \eta \bar{p} L (1 + o(1)),
\end{aligned}
$$

where we use $\log(1 + x) = x(1 + o(1))$ when $x = o(1)$. ∎

B Proof of Theorem 1

The proof consists in an appropriate *change-of-measure* argument. The originality of the proof stems from the fact that the change of measures is obtained by a judicious coupling argument [17]. In the following, we refer to Φ as the true stochastic model under which all the observed random labels are generated, and denote by $\mathbb{P}_\Phi = \mathbb{P}$ (resp. $\mathbb{E}_\Phi[\cdot] = \mathbb{E}[\cdot]$) the corresponding probability measure (resp. expectation). We recall that Φ is defined by the parameters (α, p), and that under Φ, the nodes are first attached to the various clusters according to the distribution α, and the labels between two nodes are then generated using distributions p. The proof consists in constructing a perturbed stochastic model

15

Ψ coupling the labels generated under Φ with those generated under Ψ. We denote by $\mathbb{P}_\Psi$ (resp. $\mathbb{E}_\Psi[\cdot] = \mathbb{E}[\cdot]$) the probability measure (resp. expectation) under the perturbed model Ψ. We then relate the proportion of misclassified nodes under any given clustering algorithm π to the distribution under $\mathbb{P}_\Psi$ of a quantity $\mathcal{Q}$ that resembles the log-likelihood ratio of the observed labels under $\mathbb{P}_\Phi$ and $\mathbb{P}_\Psi$. The analysis of the likelihood ratio finally provides the desired lower bound on the expected misclassified nodes under π. Next, we detail each step of the proof.

Coupling and the perturbed stochastic model Ψ. Let $(i^\star, j^\star) = \arg\min_{i,j:i<j} D_{L+}(p(i), p(j))$, and let $v^\star$ denote the smallest node index that belongs to cluster $i^\star$ or $j^\star$. If both $\mathcal{V}_{i^\star}$ and $\mathcal{V}_{j^\star}$ are empty, we define $v^\star = n$. Let $q \in [0,1]^{K \times (L+1)}$ satisfy:

$$D(\alpha, p) = \sum_{k=1}^{K} \alpha_k KL(q(k), p(i^\star, k)) = \sum_{k=1}^{K} \alpha_k KL(q(k), p(j^\star, k)).$$

There exists such a q from Lemma 8. Now to define the perturbed stochastic model Ψ, we couple the generation of labels under Φ and Ψ as follows.

1. We first generate construct the random clusters $\mathcal{V}_1, \ldots, \mathcal{V}_K$ under Φ, and extract $i^\star$, $j^\star$, and $v^\star$. The clusters generated under Ψ are the same as those generated under Φ. For any $v \in \mathcal{V}$, we denote by $\sigma(v)$ the cluster of node v.

2. For all nodes $v, w \neq v^\star$, the labels generated under Ψ are the same as those generated under Φ, i.e., the label ℓ is observed on the edge (v,w) with probability $p(\sigma(v), \sigma(w), \ell)$.

3. Under Ψ, for any $v \neq v^\star$, the observed label on the edge $(v, v^\star)$ under Ψ is ℓ with probability $q(\sigma(v), \ell)$.

The log-likelihood ratio and its connection to the expected number of misclassified nodes. Let $x_{v,w}$ denote the label observed on the edge (v,w). We introduce $\mathcal{Q}$, referred to as the pseudo-log-likelihood ratio of the observed labels under $\mathbb{P}_\Phi$ and $\mathbb{P}_\Psi$) as:

$$\mathcal{Q} = \sum_{v=1}^{v^\star-1} \log \frac{q(\sigma(v), x_{v^\star,v})}{p(\sigma(v^\star), \sigma(v), x_{v^\star,v})} + \sum_{v=v^\star+1}^{n} \log \frac{q(\sigma(v), x_{v^\star,v})}{p(\sigma(v^\star), \sigma(v), x_{v^\star,v})}. \tag{11}$$

Let π denote a clustering algorithm with output $(\hat{\mathcal{V}}_k)_{1 \leq k \leq K}$, and let $\mathcal{E} = \bigcup_{1 \leq k \leq K} \hat{\mathcal{V}}_k \setminus \mathcal{V}_k$ be the set of misclassified nodes under π. Note that in general in our proofs, we always assume without loss of generality that $|\bigcup_{1 \leq k \leq K} \hat{\mathcal{V}}_k \setminus \mathcal{V}_k| \leq |\bigcup_{1 \leq k \leq K} \hat{\mathcal{V}}_{\gamma(k)} \setminus \mathcal{V}_k|$ for any permutation γ, so that the set of misclassified nodes is really $\mathcal{E}$. We denote by $\varepsilon^\pi(n) = |\mathcal{E}|$. Since under Φ, nodes are interchangeable (remember that nodes are assigned to the various clusters in an i.i.d. manner), we have:

$$n\mathbb{P}_\Phi\{v \in \mathcal{E}\} = \mathbb{E}_\Phi[\varepsilon^\pi(n)] = \mathbb{E}[\varepsilon^\pi(n)].$$

Next, we establish a relationship between $\mathbb{E}[\varepsilon^\pi(n)]$ and the distribution of $\mathcal{Q}$ under $\mathbb{P}_\Psi$. For any function $f(n)$, we have:

$$\mathbb{P}_\Psi\{\mathcal{Q} \leq f(n)\} = \mathbb{P}_\Psi\{\mathcal{Q} \leq f(n), v^\star \in \mathcal{E}\} + \mathbb{P}_\Psi\{\mathcal{Q} \leq f(n), v^\star \notin \mathcal{E}\}. \tag{12}$$

Using $\mathcal{Q}$, we get:

$$\begin{aligned}
\mathbb{P}_\Psi\{\mathcal{Q} \leq f(n), v^\star \in \mathcal{E}\} &= \int_{\{\mathcal{Q} \leq f(n), v^\star \in \mathcal{E}\}} d\mathbb{P}_\Psi \\
&= \int_{\{\mathcal{Q} \leq f(n), v^\star \in \mathcal{E}\}} \exp(\mathcal{Q}) d\mathbb{P}_\Phi \\
&\leq \exp(f(n))\mathbb{P}_\Phi\{\mathcal{Q} \leq f(n), v^\star \in \mathcal{E}\} \\
&\leq \exp(f(n))\mathbb{P}_\Phi\{v^\star \in \mathcal{E}\} \\
&\leq \exp(f(n))\frac{\mathbb{E}_\Phi[\varepsilon^\pi(n)]}{(\alpha_{i^\star} + \alpha_{j^\star})n}, \tag{13}
\end{aligned}$$

where the last inequality is obtained from the fact that we cannot distinguish between $v^\star$ and any other $v \in \mathcal{V}_{\sigma(v^\star)}$. Indeed,

$$\mathbb{P}_\Phi\{v^\star \in \mathcal{E}\} = \mathbb{P}_\Phi\{v \in \mathcal{E} | v \in \mathcal{V}_{i^\star} \cup \mathcal{V}_{j^\star}\}$$

16

$$= \frac{\mathbb{P}_\Phi\{v \in \mathcal{E}, v \in \mathcal{V}_{i^\star} \cup \mathcal{V}_{j^\star}\}}{\mathbb{P}_\Phi\{v \in \mathcal{V}_{i^\star} \cup \mathcal{V}_{j^\star}\}}$$

$$\leq \frac{\mathbb{P}_\Phi\{v \in \mathcal{E}\}}{\mathbb{P}_\Phi\{v \in \mathcal{V}_{i^\star} \cup \mathcal{V}_{j^\star}\}} = \frac{\mathbb{E}_\Phi[\varepsilon^\pi(n)]}{(\alpha_{i^\star} + \alpha_{j^\star})n}.$$

Furthermore, since under the stochastic model Ψ, the observed labels do not depend on whether $v^\star$ belongs to cluster $i^\star$ or $j^\star$, we have:

$$\mathbb{P}_\Psi\{v^\star \in \hat{\mathcal{V}}_{i^\star}|v^\star \in \mathcal{V}_{i^\star}\} = \mathbb{P}_\Psi\{v^\star \in \hat{\mathcal{V}}_{i^\star}|v^\star \in \mathcal{V}_{j^\star}\} \quad \text{and}$$
$$\mathbb{P}_\Psi\{v^\star \in \hat{\mathcal{V}}_{j^\star}|v^\star \in \mathcal{V}_{i^\star}\} = \mathbb{P}_\Psi\{v^\star \in \hat{\mathcal{V}}_{j^\star}|v^\star \in \mathcal{V}_{j^\star}\}.$$

Finally, since $\mathbb{P}_\Psi\{v^\star \in \hat{\mathcal{V}}_{i^\star}|v^\star \in \mathcal{V}_{i^\star}\} + \mathbb{P}_\Psi\{v^\star \in \hat{\mathcal{V}}_{j^\star}|v^\star \in \mathcal{V}_{i^\star}\} \leq 1$, we also have:

$$\begin{aligned}
& \mathbb{P}_\Psi\{\mathcal{Q} \leq f(n), v^\star \notin \mathcal{E}\} \\
\leq \; & \mathbb{P}_\Psi\{v^\star \notin \mathcal{E}\} \\
= \; & \frac{\alpha_{i^\star}}{\alpha_{i^\star} + \alpha_{j^\star}}\mathbb{P}_\Psi\{v^\star \in \hat{\mathcal{V}}_{i^\star}|v^\star \in \mathcal{V}_{i^\star}\} + \frac{\alpha_{j^\star}}{\alpha_{i^\star} + \alpha_{j^\star}}\mathbb{P}_\Psi\{v^\star \in \hat{\mathcal{V}}_{j^\star}|v^\star \in \mathcal{V}_{j^\star}\} \\
= \; & \frac{\alpha_{i^\star}}{\alpha_{i^\star} + \alpha_{j^\star}}\mathbb{P}_\Psi\{v^\star \in \hat{\mathcal{V}}_{i^\star}|v^\star \in \mathcal{V}_{i^\star}\} + \frac{\alpha_{j^\star}}{\alpha_{i^\star} + \alpha_{j^\star}}\mathbb{P}_\Psi\{v^\star \in \hat{\mathcal{V}}_{j^\star}|v^\star \in \mathcal{V}_{i^\star}\} \\
\leq \; & \frac{\alpha_{j^\star}}{\alpha_{i^\star} + \alpha_{j^\star}}.
\end{aligned} \tag{14}$$

Combining (12), (13), and (14), we conclude that:

$$\mathbb{P}_\Psi\{\mathcal{Q} \leq f(n)\} \leq \exp(f(n))\frac{\mathbb{E}_\Phi[\varepsilon^\pi(n)]}{(\alpha_{i^\star} + \alpha_{j^\star})n} + \frac{\alpha_{j^\star}}{\alpha_{i^\star} + \alpha_{j^\star}}. \tag{15}$$

The previous equation provides the desired generic relationship between $\mathbb{E}_\Phi[\varepsilon^\pi(n)]$ and $\mathbb{P}_\Psi\{\mathcal{Q} \leq f(n)\}$ from which can deduce a necessary condition for $\mathbb{E}[\varepsilon^\pi(n)] \leq s$. Applying (15) with $f(n) = \log(n/\mathbb{E}_\Phi[\varepsilon^\pi(n)]) - \log(2/\alpha_{i^\star})$, we have:

$$\mathbb{P}_\Psi\{\mathcal{Q} \leq \log(n/\mathbb{E}_\Phi[\varepsilon^\pi(n)]) - \log(2/\alpha_{i^\star})\} \leq 1 - \frac{\alpha_{i^\star}}{2} < 1 - \frac{\alpha_{i^\star}}{4}. \tag{16}$$

In addition, from Chebyshev's inequality,

$$\mathbb{P}_\Psi\left\{\mathcal{Q} \leq \mathbb{E}_\Psi[\mathcal{Q}] + \sqrt{\frac{4}{\alpha_{i^\star}}\mathbb{E}_\Psi[(\mathcal{Q} - \mathbb{E}_\Psi[\mathcal{Q}])^2]}\right\} \geq 1 - \frac{\alpha_{i^\star}}{4}. \tag{17}$$

From (16) and (17), we deduce that:

$$\log(n/\mathbb{E}_\Phi[\varepsilon^\pi(n)]) - \log(2/\alpha_{i^\star}) \leq \mathbb{E}_\Psi[\mathcal{Q}] + \sqrt{\frac{4}{\alpha_{i^\star}}\mathbb{E}_\Psi[(\mathcal{Q} - \mathbb{E}_\Psi[\mathcal{Q}])^2]},$$

and thus, a necessary condition for $\mathbb{E}[\varepsilon^\pi(n)] \leq s$ is:

$$\log(n/s) - \log(2/\alpha_{i^\star}) \leq \mathbb{E}_\Psi[\mathcal{Q}] + \sqrt{\frac{4}{\alpha_{i^\star}}\mathbb{E}_\Psi[(\mathcal{Q} - \mathbb{E}_\Psi[\mathcal{Q}])^2]}. \tag{18}$$

Analysis of the log-likelihood ratio. In view of (18), we can obtain a necessary condition for $\mathbb{E}[\varepsilon^\pi(n)] \leq s$ if we evaluate $\mathbb{E}_\Psi[\mathcal{Q}]$ and $\mathbb{E}_\Psi[(\mathcal{Q} - \mathbb{E}_\Psi[\mathcal{Q}])^2]$.

(i) We first compute $\mathbb{E}_\Psi[\mathcal{Q}]$. Note that in view of the definition of $v^\star$, a node whose index is smaller than $v^\star$ cannot be in $\mathcal{V}_{i^\star}$ or $\mathcal{V}_{j^\star}$, whereas a node whose index v is larger than $v^\star$ can be in any cluster (and the cluster of such a v is drawn according to the distribution α independently of other nodes). This slightly complicates the computation of the expectation of the two sums defining $\mathcal{Q}$ in (11). To circumvent this problem, we can observe that $v^\star$ is rather small, i.e., less $\log(n)^2$ with high probability, and that hence, we can approximate $\mathbb{E}_\Psi[\mathcal{Q}]$ by $\mathbb{E}_\Psi[\sum_{v=v^\star+1}^n \log \frac{q(\sigma(v),x_{v^\star,v})}{p(\sigma(v^\star),\sigma(v),x_{v^\star,v})}]$, which is itself well-approximated by $nD(\alpha, p)$. More formally, since $\mathbb{P}\{v^\star \leq m\} = 1 - (1 - \alpha_{i^\star} - \alpha_{j^\star})^m$,

$$\mathbb{P}\{v^\star \leq \log(n)^2\} \geq 1 - \frac{1}{n^4}. \tag{19}$$

17

Hence from condition (A1), (19), and the definition of $\mathcal{Q}$,

$$
\begin{aligned}
\mathbb{E}_\Psi[\mathcal{Q}] &= \mathbb{P}\{v^\star > \log(n)^2\}\mathbb{E}_\Psi[\mathcal{Q}|v^\star > \log(n)^2] + \mathbb{P}\{v^\star \le \log(n)^2\}\mathbb{E}_\Psi[\mathcal{Q}|v^\star \le \log(n)^2] \\
&\le \frac{\log\eta}{n^3} + \mathbb{E}_\Psi[\mathcal{Q}|v^\star \le \log(n)^2] \\
&\le \frac{\log\eta}{n^3} + \mathbb{E}_\Psi\left[\sum_{v=1}^{v^\star-1} \log\frac{q(\sigma(v), x_{v^\star,v})}{p(\sigma(v^\star),\sigma(v),x_{v^\star,v})}|v^\star \le \log(n)^2\right] + nD(\alpha,p) \\
&\le \frac{\log\eta}{n^3} + \mathbb{E}_\Psi\left[(v^\star-1)\sum_{k\notin\{i^\star,j^\star\}}\frac{\alpha_k KL(q(k),p(\sigma(v^\star),k)))}{1-\alpha_{i^\star}-\alpha_{j^\star}}|v^\star \le \log(n)^2\right] + nD(\alpha,p) \\
&\le \left(n + 2\log(n)^2\log\eta\right)D(\alpha,p) + \frac{\log\eta}{n^3},
\end{aligned}
\tag{20}
$$

where the last inequality stems from the fact that $2KL(q(i),p(\sigma(v^\star,i)))\log\eta \ge KL(q(j),p(\sigma(v^\star,j)))$ for all i and j from condition (A1).

(ii) To compute $\mathbb{E}_\Psi[(\mathcal{Q}-\mathbb{E}_\Psi[\mathcal{Q}])^2]$, we evaluate $\mathbb{E}_\Psi[(\mathcal{Q}-nD(\alpha,p))^2|\sigma(v^\star)=i^\star]$ and $\mathbb{E}_\Psi[(\mathcal{Q}-nD(\alpha,p))^2|\sigma(v^\star)=j^\star]$. From condition (A1), (19), and the definition of $\mathcal{Q}$,

$$
\begin{aligned}
&\mathbb{E}_\Psi[(\mathcal{Q}-nD(\alpha,p))^2|\sigma(v^\star)=i^\star] \\
&= \mathbb{P}\{v^\star \le \log(n)^2\}\mathbb{E}_\Psi[(\mathcal{Q}-nD(\alpha,p))^2|\sigma(v^\star)=i^\star, v^\star \le \log(n)^2] \\
&\quad + \mathbb{P}\{v^\star > \log(n)^2\}\mathbb{E}_\Psi[(\mathcal{Q}-nD(\alpha,p))^2|\sigma(v^\star)=i^\star, v^\star > \log(n)^2] \\
&\le \mathbb{E}_\Psi[(\mathcal{Q}-nD(\alpha,p))^2|\sigma(v^\star)=i^\star, v^\star \le \log(n)^2] \\
&\quad + \frac{1}{n^4}\mathbb{E}_\Psi[(\mathcal{Q}-nD(\alpha,p))^2|\sigma(v^\star)=i^\star, v^\star > \log(n)^2] \\
&= O(n\bar{p}).
\end{aligned}
$$

To derive the above inequality, we have used:

$$
\begin{aligned}
\mathbb{E}_\Psi&\left[\left(\sum_{v=v^\star+1}^{n}\left(\log\frac{q(\sigma(v),x_{v^\star,v})}{p(\sigma(v^\star),\sigma(v),x_{v^\star,v})}-D(\alpha,p)\right)\right)^2|\sigma(v^\star)=i^\star\right] \\
&= \sum_{v=v^\star+1}^{n}\mathbb{E}_\Psi\left[\left(\log\frac{q(\sigma(v),x_{v^\star,v})}{p(i^\star,\sigma(v),x_{v^\star,v})}-D(\alpha,p)\right)^2|\sigma(v^\star)=i^\star\right] \\
&= O(n\bar{p}) \quad\text{and}
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_\Psi&\left[\left(\sum_{v=1}^{v^\star-1}\left(\log\frac{q(\sigma(v),x_{v^\star,v})}{p(\sigma(v^\star),\sigma(v),x_{v^\star,v})}-D(\alpha,p)\right)\right)^2|\sigma(v^\star)=i^\star\right] \\
&= O(v^\star\bar{p}+(v^\star\bar{p})^2),
\end{aligned}
$$

where we use (A1) and the fact that every label is generated independently. Using the same approach, we can also conclude that $\mathbb{E}_\Psi[(\mathcal{Q}-nD(\alpha,p))^2|\sigma(v^\star)=j^\star]=O(n\bar{p})$. In summary, we have:

$$
\mathbb{E}_\Psi[(\mathcal{Q}-\mathbb{E}_\Psi[\mathcal{Q}])^2] = O(n\bar{p}).
\tag{21}
$$

We are ready to complete the proof of Theorem 1. From (18), (20), (21), and Lemma 9, when the expected number of misclassified nodes is less than s (i.e., $\mathbb{E}[\varepsilon^\pi(n)] \le s$), we must have:

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

∎

C   Performance of the SP Algorithm – Proof of Theorem 2

Notations. We use the standard matrix norm $\|A\| = \sup_{x:\|x\|_2=1}\|Ax\|_2$. We denote by M^ℓ the expectation of the matrix of A^ℓ, i.e., $M^\ell_{u,v} = p(i,j,\ell)$ when $u \in \mathcal{V}_i$ and $v \in \mathcal{V}_j$. Let $M =$

$\sum_{\ell=1}^{L} w_\ell M^\ell$. We define A_Γ to denote the adjacency matrix obtained after trimming (Step 3 in Algorithm 1). For any matrix $R \in \mathbb{R}^{n \times n}$, we define the matrix R_Γ the square matrix formed by the lines and columns of R whose indexes are in Γ. Hence, we can define A_Γ^ℓ, M_Γ^ℓ, and M_Γ where Γ is the set of items obtained after the trimming process (Line 3) in the SP algorithm (when taking the expectation to get for example M_Γ, we condition on Γ). We introduce the noise matrices $X_\Gamma^\ell = A_\Gamma^\ell - M_\Gamma^\ell$ and $X_\Gamma = \sum_{\ell=1}^{L} w_\ell X_\Gamma^\ell$. We also denote by $e(v, S, \ell) = \sum_{w \in S} A_{vw}^\ell$ the total number of item pairs with observed label ℓ including the item v and an item from S and $\mu(v, S, \ell) = \frac{e(v,S,\ell)}{|S|}$ the empirical density of label ℓ. Let $e(v, S) = \sum_{\ell=1}^{L} e(v, S, \ell)$ and $\mu(v, S) = [\mu(v, S, \ell)]_{0 \le \ell \le L}$. In what follows, $e(v, \mathcal{V})$ is referred to as the *degree* of item v (the number of observed labels different than 0 of pairs of items including v).

Outline of the proof. To analyze the performance of the SP algorithm, we first state preliminary lemmas. Lemma 11 is concerned with the concentration of the degree of the various items. Lemma 12 provides an upper bound of the matrix norm of random noise matrix X_Γ^ℓ. From these two lemmas, we analyze the performance of the first part of the SP algorithm, and prove Theorem 7. To analyze the second part of the SP algorithm consisting of $\log(n)$ improvement iterations, we introduce an appropriate set of items H such that that $\mathcal{V} \setminus H$ is of cardinality less than s with high probability under the condition that $nD(\alpha, p) - \frac{n\bar{p}}{\log(n\bar{p})^3} \ge \log(n/s) + \sqrt{\log(n/s)}$. We further bound the rate of improvement of our cluster estimates in each iteration when restricted to the set of items H, and deduce that after $\log(n)$ iterations, no item in H is misclassified.

C.1   Preliminary lemmas

Lemma 11 *For every $v \in \mathcal{V}$ and $c \ge 1$, we have*

$$\mathbb{P}\{e(v, \mathcal{V}) \ge 10cn\bar{p}L\} \le \exp(-10cn\bar{p}L).$$

Proof. From Markov inequality,

$$
\begin{aligned}
\mathbb{P}\{e(v, \mathcal{V}) \ge 10n\bar{p}L\} &\le \inf_{\theta > 0} \frac{\prod_{k=1}^{K} \mathbb{E}\left[\exp(\theta e(v, \mathcal{V}_k))\right]}{\exp(\theta 10 cn\bar{p}L)} \\
&\le \inf_{\theta > 0} \frac{\prod_{k=1}^{K} \left(1 + \bar{p}L(\exp(\theta) - 1)\right)^{\alpha_k n}}{\exp(\theta 10 cn\bar{p}L)} \\
&\le \inf_{\theta > 0} \frac{\prod_{k=1}^{K} \left(\exp(\bar{p}L(\exp(\theta) - 1))\right)^{\alpha_k n}}{\exp(\theta 10 cn\bar{p}L)} \\
&\le \exp(-10cn\bar{p}L),
\end{aligned}
$$

where we derive the last inequality choosing $\theta = 2$. ∎

Lemma 12 (Lemma 8.5 of [4]) *When $e(v, \mathcal{V}, \ell) \le \Delta$ for all $v \in \Gamma$, with high probability,*

$$\|X_\Gamma^\ell\| = O(\sqrt{n\bar{p} + \Delta}).$$

The proof of Lemma 12 relies on arguments used in the spectral analysis of random graphs, see [7] and [4].

Lemma 13 *For all $v \in \mathcal{V}_k$ and $D \ge 0$,*

$$
\mathbb{P}\left\{ \left(\sum_{i=1}^{K} |\mathcal{V}_i| KL(\mu(v, \mathcal{V}_i), p(k, i)) \ge nD\right) \cap \left(e(v, \mathcal{V}) \le 10\eta n\bar{p}L\right) \right\}
$$
$$
\le \exp\left(-nD + KL\log(10\eta Ln\bar{p}) + \frac{100\eta^2 n\bar{p}^2 L^2}{\alpha_1}\right).
$$

Proof. Let $\mathcal{X}$ be a set of $K \times (L+1)$ matrices such that

$$\mathcal{X} = \left\{ \boldsymbol{x} \in \mathbb{Z}^{K \times (L+1)} : \quad \sum_{i=1}^{K} \sum_{\ell=1}^{L} x_{i,\ell} \leq 10\eta n\bar{p}L, \quad \text{and} \quad \sum_{\ell=0}^{L} x_{i,\ell} = |\mathcal{V}_i| \quad \text{for all} \quad 1 \leq i \leq K \right\}.$$

For notational simplicity, we use $[\frac{x_{i,\ell}}{|\mathcal{V}_i|}]$ instead of $[\frac{x_{i,\ell}}{|\mathcal{V}_i|}]_{0 \leq \ell \leq L}$ to represent the probability mass vector on labels defined by x_i. With a slight abuse of notation, we denote by $e(v)$ the $K \times (L+1)$ matrix whose (i,ℓ) element is $e(v, \mathcal{V}_i, \ell)$. Then, for $v \in \mathcal{V}_k$,

$$\mathbb{P}\left\{ \left(\sum_{i=1}^{K} |\mathcal{V}_i| KL(\mu(v, \mathcal{V}_i), p(k,i)) \geq nD \right) \cap \left(e(v, \mathcal{V}) \leq 10n\bar{p}L \right) \right\}$$

$$= \sum_{\boldsymbol{x} \in \mathcal{X}} \mathbb{P}\{e(v) = \boldsymbol{x}\} \mathbb{P}\left\{ \sum_{i=1}^{K} |\mathcal{V}_i| KL(\mu(v, \mathcal{V}_i), p(k,i)) \geq nD \Big| e(v) = \boldsymbol{x} \right\}$$

$$\leq \sum_{\boldsymbol{x} \in \mathcal{X}} \mathbb{P}\{e(v) = \boldsymbol{x}\} \frac{\exp\left(\sum_{i=1}^{K} |\mathcal{V}_i| KL([\frac{x_{i,\ell}}{|\mathcal{V}_i|}], p(k,i)) \right)}{\exp(nD)}$$

$$\leq \sum_{\boldsymbol{x} \in \mathcal{X}} \mathbb{P}\{e(v) = \boldsymbol{x}\} \frac{\prod_{i=1}^{K} \prod_{\ell=0}^{L} \left(\frac{x_{i,\ell}}{|\mathcal{V}_i| p(k,i,\ell)} \right)^{x_{i,\ell}}}{\exp(nD)}$$

$$\overset{(a)}{\leq} \frac{1}{\exp(nD)} \sum_{\boldsymbol{x} \in \mathcal{X}} \prod_{i=1}^{K} \left(\left(1 - \frac{\sum_{\ell=1}^{L} x_{i,\ell}}{|\mathcal{V}_i|} \right)^{x_{i,0}} \exp(\sum_{\ell=1}^{L} x_{i,\ell}) \right)$$

$$= \frac{1}{\exp(nD)} \sum_{\boldsymbol{x} \in \mathcal{X}} \prod_{i=1}^{K} \exp\left((|\mathcal{V}_i| - \sum_{\ell=1}^{L} x_{i,\ell}) \log\left(1 - \frac{\sum_{\ell=1}^{L} x_{i,\ell}}{|\mathcal{V}_i|} \right) + \sum_{\ell=1}^{L} x_{i,\ell} \right)$$

$$\leq \frac{1}{\exp(nD)} \sum_{\boldsymbol{x} \in \mathcal{X}} \prod_{i=1}^{K} \exp\left(\frac{(\sum_{\ell=1}^{L} x_{k,\ell})^2}{|\mathcal{V}_i|} \right)$$

$$\leq \frac{(10\eta n\bar{p}L)^{KL} \exp(100\eta^2 n\bar{p}^2 L^2/\alpha_1)}{\exp(nD)}$$

$$= \exp\left(-nD + KL \log(10\eta Ln\bar{p}) + \frac{100\eta^2 n\bar{p}^2 L^2}{\alpha_1} \right),$$

where (a) stems from the following inequality:

$$\mathbb{P}\{e(v, \mathcal{V}_i, \ell) = x_{i,\ell} \quad \text{for all} \quad i, \ell\}$$

$$\leq \prod_{i=1}^{K} \left(p(k,i,0)^{x_{i,0}} \prod_{\ell=1}^{L} \binom{|\mathcal{V}_i|}{x_{i,\ell}} p(k,i,\ell)^{x_{k,\ell}} \right)$$

$$\leq \prod_{i=1}^{K} \left(p(k,i,0)^{x_{i,0}} \prod_{\ell=1}^{L} \left(\frac{e|\mathcal{V}_i|}{x_{i,\ell}} \right)^{x_{i,\ell}} p(k,i,\ell)^{x_{i,\ell}} \right).$$

∎

C.2 Part 1 of the SP algorithm – Proof of Theorem 7

Recall that $\hat{A} = \hat{U}\hat{V} = \hat{U}\hat{U}^\top A_\Gamma$ and $\|\hat{A}_u - \hat{A}_v\| = \|\hat{V}_u - \hat{V}_v\|$. We can bound the number of misclassified items as follows:

- with high probability, we have

$$\|\hat{A} - M_\Gamma\|_F^2 = \sum_{v \in \Gamma} \|\hat{A}_v - M_{v,\Gamma}\|_2^2 = O(n\bar{p}\log(n\bar{p})^2); \tag{22}$$

20

- with high probability, every item pair u and v satisfies that when $\sigma(v)$ represents the cluster of v and $M_{v,\Gamma}$ denotes the column vector of M_Γ on v,

$$\|M_{u,\Gamma} - M_{v,\Gamma}\|_2^2 = \Omega\left(n\bar{p}^2\right) \quad \text{when} \quad \sigma(u) \neq \sigma(v), \tag{23}$$

  since every w_ℓ is generated uniformly at random in $[0,1]$ and (A2) holds;
- (23) suggests that if v is misclassified by Algorithm 2, then we should have:

$$\|\hat{A}_v - M_{v,\Gamma}\|_2^2 = \Omega\left(n\bar{p}^2\right); \tag{24}$$

- from (22) and (24), with high probability,

$$\left|\bigcup_{k=1}^K (V_k \setminus S_k)\right| = O\left(\frac{\log(n\bar{p})^2}{\bar{p}}\right).$$

Next, we prove (22) and (24).

Proof of (22). First observe that from the definition of Γ,

$$
\begin{aligned}
\mathbb{P}\left\{\max_{v \in \Gamma} e(v, \mathcal{V}) \geq 10n\bar{p}L\right\} &= \mathbb{P}\left\{|\{v : e(v, \mathcal{V}) \geq 10n\bar{p}L\}| > \lfloor n\exp(-n\tilde{p})\rfloor\right\} \\
&\leq \frac{n\exp(-10n\bar{p}L)}{\lfloor n\exp(-n\tilde{p})\rfloor + 1} \\
&\leq \exp(-5n\bar{p}L),
\end{aligned}
$$

where the first inequality stems from Lemma 11 and Markov inequality. Therefore, with high probability,

$$\max_{v \in \Gamma} e(v, \mathcal{V}) \leq 10n\bar{p}L. \tag{25}$$

When the degrees of items are bounded, the standard matrix norm of each noise matrix X_Γ^ℓ can be bounded using Lemma 12. From (25) and Lemma 12,

$$
\begin{aligned}
\|X_\Gamma\| &\leq \sum_{\ell=1}^L w_\ell \|X_\Gamma^\ell\| \\
&= \sum_{\ell=1}^L O(w_\ell \sqrt{n\bar{p} + 10n\bar{p}L}) \\
&= O(\sqrt{n\bar{p}}). \tag{26}
\end{aligned}
$$

Let $\tilde{K}$ be the number of columns of $\hat{U}$. Since $\hat{A}$ is the $\tilde{K}$-rank approximation of A_Γ obtained by the iterative power method with $2\log(n)$ iterations, from Theorem 9.1 and Theorem 9.2 in [11], with high probability,

$$\frac{1}{2}s_k(A_\Gamma) \leq \|A_\Gamma \hat{U}_k\| \leq s_k(A_\Gamma) \quad \text{and} \quad \|A_\Gamma(I - \hat{U}_{1:k}\hat{U}_{1:k}^\top)\| \leq 2s_{k+1}(A_\Gamma). \tag{27}$$

Since $\|A_\Gamma \hat{U}_K\| \leq s_{K+1}(A_\Gamma) \leq \|X_\Gamma\| = O(\sqrt{n\bar{p}})$ from Lemma 12 and (27), $\tilde{K} \leq K$ and thus the rank of $(\hat{A} - M_\Gamma)$ is less than $2K$. Therefore,

$$
\begin{aligned}
\|\hat{A} - M_\Gamma\|_F^2 &\leq 2K\|\hat{A} - M_\Gamma\|^2 \\
&\leq 4K\left(\|\hat{A} - A_\Gamma\|^2 + \|A_\Gamma - M_\Gamma\|^2\right) \\
&\leq O(n\bar{p}\log(n\bar{p})^2), \tag{28}
\end{aligned}
$$

where the last inequality stems from the fact that $\|A_\Gamma - M_\Gamma\| = \|X_\Gamma\| = O(\sqrt{n\bar{p}})$ and $\|\hat{A} - A_\Gamma\| \leq 2s_{\tilde{K}+1}(A_\Gamma) = O(\sqrt{n\bar{p}}\log(n\bar{p}))$ from (27).

Proof of (24). Define the following sets:

$$
\begin{aligned}
\mathcal{I}_k &= \{v \in \mathcal{V}_k \cap \Gamma : \|\hat{A}_v - M_\Gamma^k\|^2 \leq \frac{1}{4}\frac{n\tilde{p}^2}{\log(n\tilde{p})}\} \\
\mathcal{O} &= \{v \in \Gamma : \|\hat{A}_v - M_\Gamma^k\|^2 \geq 4\frac{n\tilde{p}^2}{\log(n\tilde{p})} \quad \text{for all} \quad 1 \leq k \leq K\}.
\end{aligned}
$$

These sets are designed so that

(i) $|(\cup_{k=1}^K \mathcal{I}_k) \cap Q_v| = 0$ for all $v \in \mathcal{O} \cap \mathcal{V}_R$, since $\|\hat{A}_v - \hat{A}_w\|^2 \geq \frac{1}{2}\|\hat{A}_v - M_\Gamma^k\|^2 - \|\hat{A}_w - M_\Gamma^k\|^2 > \frac{n\bar{p}^2}{\log(n\bar{p})}$ for all $w \in \mathcal{I}_k$;

(ii) $|\Gamma \setminus (\cup_{k=1}^K \mathcal{I}_k)| \leq \frac{\|\hat{A} - M_\Gamma\|_F^2}{\min_{v \in \Gamma \setminus (\cup_{k=1}^K I_k)} \|\hat{A}_v - M_\Gamma^k\|^2} = O\left(\frac{\log(n\bar{p})^3}{\bar{p}}\right)$;

(iii) $\mathcal{I}_k \subset Q_v$ for all $v \in \mathcal{I}_k \cap \mathcal{V}_R$, since $\|\hat{A}_v - \hat{A}_w\|^2 \leq 2\|\hat{A}_v - M_\Gamma^k\|^2 + 2\|\hat{A}_w - M_\Gamma^k\|^2 \leq \frac{n\bar{p}^2}{\log(n\bar{p})}$ for all $w \in \mathcal{I}_k$;

(iv) If $|Q_v \cap \mathcal{I}_k| \geq 1$, $|Q_v \cap \mathcal{I}_j| = 0$ for all $j \neq k$, since $\|M_\Gamma^k - M_\Gamma^j\| = \Omega(n\bar{p}^2)$ is much larger than the radius $\frac{n\bar{p}^2}{\log(n\bar{p})} = O(\frac{n\bar{p}^2}{\log(n\bar{p})})$;

From the properties of $\mathcal{I}_k$ and $\mathcal{O}$, we state the following results.

- From (i) and (ii), we deduce that

$$|Q_v| = O\left(\frac{\log(n\bar{p})^3}{\bar{p}}\right) \quad \text{for all} \quad v \in \mathcal{O} \cap V_R, \tag{29}$$

since every $w \in (\cup_{k=1}^K \mathcal{I}_k)$ is outside of Q_v (i.e., $w \in \Gamma \setminus (\cup_{k=1}^K I_k)$ is necessary for $w \in Q_v$);

- since α_k is a constant for all k and $\frac{|\Gamma \setminus (\cup_{k=1}^K \mathcal{I}_k)|}{|\Gamma|} = o(1)$ from (ii), with high probability,

$$|\mathcal{I}_k \cap \mathcal{V}_R| \geq 1 \quad \text{for all} \quad 1 \leq k \leq K; \tag{30}$$

- The properties (ii), (iii), and (iv) and (30) imply that

$$|Q_v \setminus \cup_{l=0}^{k-1} S_l| \geq m_k, \quad \exists v \in (\cup_{m=1}^K \mathcal{I}_k \cap V_R) \setminus (\cup_{l=0}^{k-1} S_l), \tag{31}$$

where m_k is the k-th largest value among $\{|\mathcal{I}_1|, \ldots, |\mathcal{I}_K|\}$;

- since $|\mathcal{I}_k| \geq |V_k \cap (\Gamma \setminus \mathcal{O})| \geq \alpha_k n(1 - o(1))$ from (ii) and (iii),

$$|\mathcal{I}_k| \geq |V_k \cap (\Gamma \setminus \mathcal{O})| \geq \alpha_k n(1 - o(1)). \tag{32}$$

Thus, we can conclude that $\hat{K} = K$ from (31) and (32) and the property (ii); and from (29), there exists a permutation γ such that $\|\hat{A}_{v_k^\star} - M_\Gamma^{\gamma(k)}\|^2 \leq 4\frac{n\bar{p}^2}{\log(n\bar{p})}$ for all k. Hence from (23), $\|\hat{A}_v - M_{v,\Gamma}\|^2 = \Omega\left(n\bar{p}^2\right)$ when v is misclassified. $\blacksquare$

C.3 Proof of Theorem 2

From Chernoff bound, with high probability,

$$||\mathcal{V}_k| - \alpha_k n| \leq \sqrt{n}\log(n) \quad \text{for all} \quad k. \tag{33}$$

In what follows, we hence just prove the theorem assuming that (33) holds.

Let H be the largest set of items $v \in \mathcal{V}$ satisfying:

(H1) $e(v, \mathcal{V}) \leq 10\eta n\bar{p}L$,

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

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

(H1) regularizes degrees, (H2) means that $v \in H$ is correctly classified when using the log-likelihood estimate, and (H3) means that v does not share too many labels with items outside H.

The proof of the theorem follows from the following propositions. The first provides an upper bound of $|\mathcal{V} \setminus H|$, and the second provides the rate at which our estimated clusters improve in each iteration when we restrict our attention to items in H.

Proposition 14 *When $nD(\alpha, p) - \frac{n\bar{p}}{\log(n\bar{p})^3} \geq \log(n/s) + \sqrt{\log(n/s)}$, $|\mathcal{V} \setminus H| \leq s$ with high probability.*

22

Proposition 15 *If* $|\bigcup_{k=1}^{K}(S_k^{(0)} \setminus \mathcal{V}_k) \cap H| + |\mathcal{V} \setminus H| = O(1/\bar{p})$, *with high probability, the following statement holds*

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

From Proposition 15, after $\log(n)$ iterations (remember that $n\bar{p} = \omega(1)$, so when n is large enough $1/\sqrt{n\bar{p}} \leq e^{-2}$), no item in H can be misclassified with high probability. Hence the number of misclassified items cannot exceed $|\mathcal{V} \setminus H| \leq s$, $nD(\alpha, p) - \frac{n\bar{p}}{\log(n\bar{p})^3} \geq \log(n/s) + \sqrt{\log(n/s)}$. The proof is completed by remarking that if the previous condition on $D(\alpha, p)$ holds, then

$$1 \leq \lim_{n \to \infty} \frac{nD(\alpha, p) - \frac{n\bar{p}}{\log(n\bar{p})^3}}{\log(n/s) + \sqrt{\log(n/s)}} = \lim_{n \to \infty} \frac{nD(\alpha, p)}{\log(n/s)},$$

where we used $D(\alpha, p) = \Omega(\bar{p})$ from condition (A2) and Lemma 9. ∎

C.3.1 Proof of Proposition 14 – Size of $\mathcal{V} \setminus H$

We compute the number of items satisfying (H1), (H2), and (H3) in (34), (35), and Lemma 16, respectively.

Number of items satisfying (H1): From Lemma 11, we get:

$$\mathbb{P}\{e(v, \mathcal{V}) \leq 10\eta n\bar{p}L\} \geq 1 - \exp(-10\eta n\bar{p}L). \tag{34}$$

Number of items satisfying (H2): We shall prove that when v satisfies (H1), v satisfies (H2) as well with probability at least

$$1 - \exp\left(-nD(\alpha, p) + \frac{n\bar{p}}{2\log(n\bar{p})^3}\right). \tag{35}$$

To this aim, we first establish that if v satisfies

$$\sum_{i=1}^{K} |\mathcal{V}_i| KL(\mu(v, \mathcal{V}_i), p(k, i)) \leq \left(1 - \frac{\log(n)^2}{\sqrt{n}}\right) nD(\alpha, p) - \frac{n\bar{p}}{\log(n\bar{p})^4}, \tag{36}$$

then v satisfies (H2). Indeed, assume that (36) holds, then

(i) $\sum_{i=1}^{K} \alpha_i n KL(\mu(v, \mathcal{V}_i), p(k, i)) \leq \left(1 + \frac{\log(n)^2}{\sqrt{n}}\right) \sum_{i=1}^{K} |\mathcal{V}_i| KL(\mu(v, \mathcal{V}_i), p(k, i)) < nD(\alpha, p)$, since $||\mathcal{V}_i| - \alpha_i n| \leq \sqrt{n}\log(n)$ and (36) holds;

(ii) $\sum_{i=1}^{K} \alpha_i n KL(\mu(v, \mathcal{V}_i), p(j, i)) \geq nD(\alpha, p)$, since
$\max\left\{\sum_{i=1}^{K} \alpha_i KL(\mu(v, \mathcal{V}_i), p(j, i)), \sum_{i=1}^{K} \alpha_i KL(\mu(v, \mathcal{V}_i), p(k, i))\right\} \geq D(\alpha, p)$ and $\sum_{i=1}^{K} \alpha_i KL(\mu(v, \mathcal{V}_i), p(k, i)) < D(\alpha, p)$;

(iii) $\sum_{i=1}^{K} |\mathcal{V}_i| KL(\mu(v, \mathcal{V}_i), p(j, i)) \geq \left(1 - \frac{\log(n)^2}{\sqrt{n}}\right) nD(\alpha, p)$, from ii) and the fact that $||\mathcal{V}_i| - \alpha_i n| \leq \sqrt{n}\log(n)$;

(iv) from (36) and iii), for all $j \neq i$,

$$\sum_{i=1}^{K} \sum_{\ell=0}^{L} e(v, \mathcal{V}_i, \ell) \log \frac{p(k, i, \ell)}{p(j, i, \ell)} = \sum_{i=1}^{K} |\mathcal{V}_i| \left(KL(\mu(v, \mathcal{V}_i), p(j, i)) - KL(\mu(v, \mathcal{V}_i), p(k, i))\right)$$

$$\geq \frac{n\bar{p}}{\log(n\bar{p})^4}.$$

Hence v satisfies (H2). It remains to evaluate the probability of the event (36), which is done by applying Lemma 13 and proves (35).

23

Number of items satisfying (H3): From (34), (35), and the Markov inequality, we deduce that with probability at least $1 - \exp\left(-\sqrt{\log(n/s)}\right)$, the number of items that do not satisfy either (H1) or (H2) is less than $s/3$ when $nD(\alpha, p) - \frac{n\bar{p}}{\log(n\bar{p})^3} \geq \log(n/s) + \sqrt{\log(n/s)}$, since

$$\frac{\mathbb{E}\{\text{The number of items that do not satisfy either (H1) or (H2)}\}}{s/3}$$

$$\leq \frac{n \exp(-10\eta n\bar{p}L) + n \exp\left(-nD(\alpha, p) + \frac{n\bar{p}}{2\log(n\bar{p})^3}\right)}{s/3}$$

$$\leq \frac{n}{s} \exp\left(-nD(\alpha, p) + \frac{n\bar{p}}{\log(n\bar{p})^3}\right) \leq \exp\left(-\sqrt{\log(n/s)}\right), \tag{37}$$

where we have used Lemma 10 for the last inequality. Lemma 16 allows us to complete the proof of Proposition. ∎

Lemma 16 *When the number of items that do not satisfy either (H1) or (H2) is less than $s/3$, $|\mathcal{V} \setminus H| \leq s$, with high probability.*

Proof. Let $e(S, S) = \sum_{v \in S} e(S, S)$. Next we prove the following intermediate claim: there is no subset $S \subset \mathcal{V}$ such that $e(S, S) \geq s \log(n\bar{p})^2$ and $|S| = s$ with high probability. For any subset $S \in \mathcal{V}$ such that $|S| = s$, by Markov inequality,

$$\begin{aligned}
\mathbb{P}\{e(S, S) \geq s \log(n\bar{p})^2\} &\leq \inf_{t \geq 0} \frac{\mathbb{E}[\exp(e(S, S)t)]}{st \log(n\bar{p})^2} \\
&\leq \inf_{t \geq 0} \frac{\prod_{i=1}^{s^2/2}(1 + L\bar{p}\exp(t))}{st \log(n\bar{p})^2} \\
&\leq \inf_{t \geq 0} \exp\left(\frac{s^2 L\bar{p}}{2}\exp(t) - st \log(n\bar{p})^2\right) \\
&\leq \exp\left(-n\bar{p}s\left(\log n\bar{p} - \frac{sL}{2n}\exp(\frac{n\bar{p}}{\log n\bar{p}})\right)\right) \\
&\leq \exp\left(-\frac{n\bar{p}s \log n\bar{p}}{2}\right), \tag{38}
\end{aligned}$$

where, in the last two inequalities, we have set $t = \frac{n\bar{p}}{\log n\bar{p}}$ and used the fact that: $\frac{n}{s} \geq \exp(\frac{n\bar{p}}{\log n\bar{p}})$, which comes from the assumptions made in the theorem. Since the number of subsets $S \subset \mathcal{V}$ with size s is $\binom{n}{s} \leq (\frac{en}{s})^s$, from (38), we deduce:

$$\begin{aligned}
\mathbb{E}[|\{S : e(S, S) \geq s \log(n\bar{p})^2 \text{ and } |S| = s\}|] &\leq (\frac{en}{s})^s \exp\left(-\frac{n\bar{p}s \log n\bar{p}}{2}\right) \\
&= \exp\left(-s(\frac{n\bar{p}\log n\bar{p}}{2} - \log\frac{en}{s})\right) \\
&\leq \exp\left(-\frac{n\bar{p}s \log n\bar{p}}{4}\right).
\end{aligned}$$

Therefore, by Markov inequality, we can conclude that there is no $S \subset \mathcal{V}$ such that $e(S, S) \geq s \log(n\bar{p})^2$ and $|S| = s$ with high probability.

To conclude the proof of the lemma, we build the following sequence of sets. Let Z_1 denote the set of items that do not satisfy at least one of (H1) and (H2). Let $\{Z(t) \subset \mathcal{V}\}_{1 \leq t \leq t^*}$ be generated as follows:

- $Z(0) = Z_1$.
- For $t \geq 1$, $Z(t) = Z(t-1) \cup \{v_t\}$ if there exists $v_t \in \mathcal{V}$ such that $e(v_t, Z(t-1)) > 2\log(n\bar{p})^2$ and $v_t \notin Z(t-1)$. If such an item does not exist, the sequence ends.

The sequence ends after the construction of $Z(t^\star)$. We show that if we assume that the cardinality of items that do not satisfy (H3) is strictly larger than $s/2$, then one the set of the sequence $\{Z(t) \subset \mathcal{V}\}_{1 \leq t \leq t^\star}$ contradicts the claim we just proved.

Assume that the number of items do not satisfy (H3) is strictly larger than $s/2$, then these items will be at some point added to the sets $Z(t)$, and by definition, each of these node contributes with more than $2\log(n\bar{p})^2$ in $e(Z(t), Z(t))$. Hence if starting from Z_1, we add $s/2$ items not satisfying (H3), we get a set $Z(t)$ of cardinality less than $s/3 + s/2$ and such that $e(Z(t), Z(t)) > s\log(n\bar{p})^2$. We can further add arbitrary items to $Z(t)$ so that it becomes of cardinality s, and the obtained set contradicts the claim. ∎

C.3.2 Proof of Proposition 15

Recall that $\{S_j^{(t)}\}_{1 \leq j \leq K}$ is the partition after the t-th improvement iteration. Also recall that with loss of generality, we assume that the set of misclassified items in H after the t-th step is $\mathcal{E}^{(t)} = \left(\cup_k (S_k^{(t)} \setminus \mathcal{V}_k)\right) \cap H$ (it should be defined through an appropriate permutation γ of $\{1, \ldots, K\}$ by $\mathcal{E}^{(t)} = (\cup_k (S_k^{(t)} \setminus \mathcal{V}_{\gamma(k)})) \cap H$, but we omit γ). With this notational convention, we can define $\mathcal{E}_{jk}^{(t)} = (S_j^{(t)} \cap \mathcal{V}_k) \cap H$ and $\mathcal{E}^{(t)} = \bigcup_{j,k:j \neq k} \mathcal{E}_{jk}^{(t)}$. At each improvement step, items move to the most likely cluster (according to the log-likelihood defined in the SP algorithm). Thus, for all i,

$$0 \leq \sum_{j,k:j \neq k} \sum_{v \in \mathcal{E}_{jk}^{(t+1)}} \sum_{i=1}^{K} \sum_{\ell=0}^{L} e(v, S_i^{(t)}, \ell) \log \frac{\hat{p}(j,i,\ell)}{\hat{p}(k,i,\ell)}$$

$$\leq \sum_{j,k:j \neq k} \sum_{v \in \mathcal{E}_{jk}^{(t+1)}} \sum_{i=1}^{K} \sum_{\ell=0}^{L} e(v, S_i^{(t)}, \ell) \log \frac{p(j,i,\ell)}{p(k,i,\ell)} + |\mathcal{E}^{(t+1)}|(n\bar{p})^{1-\kappa} \log(n\bar{p})^3 \qquad (39)$$

$$\leq \sum_{j,k:j \neq k} \sum_{v \in \mathcal{E}_{jk}^{(t+1)}} \sum_{i=1}^{K} \sum_{\ell=0}^{L} e(v, \mathcal{V}_i, \ell) \log \frac{p(j,i,\ell)}{p(k,i,\ell)}$$

$$+ \sum_{w \in \mathcal{E}^{(t+1)}} e(w, \mathcal{E}^{(t)}) \log(2\eta) + 2|\mathcal{E}^{(t+1)}|(n\bar{p})^{1-\kappa} \log(n\bar{p})^3 \qquad (40)$$

$$\leq -\frac{n\bar{p}}{\log(n\bar{p})^4}|\mathcal{E}^{(t+1)}| + \sum_{w \in \mathcal{E}^{(t+1)}} e(w, \mathcal{E}^{(t)}, \ell) \log(2\eta) + 2|\mathcal{E}^{(t+1)}|(n\bar{p})^{1-\kappa} \log(n\bar{p})^3 \quad (41)$$

$$\leq -\frac{n\bar{p}}{\log(n\bar{p})^4}|\mathcal{E}^{(t+1)}| + \sqrt{|\mathcal{E}^{(t)}||\mathcal{E}^{(t+1)}|n\bar{p}\log n\bar{p}} + 3|\mathcal{E}^{(t+1)}|(n\bar{p})^{1-\kappa} \log(n\bar{p})^3. \qquad (42)$$

Therefore, from the above inequalities, we conclude that

$$\frac{|\mathcal{E}^{(t+1)}|}{|\mathcal{E}^{(t)}|} \leq \frac{\log(n\bar{p})^{10}}{n\bar{p}} \leq \frac{1}{\sqrt{n\bar{p}}}.$$

Next we prove all the steps of the previous analysis.

Proof of (39)*:* From $\log(1+x) \leq x$, when $p(j,i,\ell) - |\hat{p}(j,i,\ell) - p(j,i,\ell)| > 0$,

$$\left| \log \frac{\hat{p}(j,i,\ell)}{p(j,i,\ell)} \right| \leq \frac{|\hat{p}(j,i,\ell) - p(j,i,\ell)|}{p(j,i,\ell) - |\hat{p}(j,i,\ell) - p(j,i,\ell)|}.$$

Thus, we just provide an upper bound of $|\hat{p}(j,i,\ell) - p(j,i,\ell)|$ to show (39). From the triangle inequality,

$$|\hat{p}(j,i,\ell) - p(j,i,\ell)|$$
$$= \frac{\left| e(S_i^{(0)}, S_j^{(0)}, \ell) - p(j,i,\ell)|S_i^{(0)}||S_j^{(0)}| \right|}{|S_i^{(0)}||S_j^{(0)}|}$$
$$\leq \frac{\left| e(S_i^{(0)}, S_j^{(0)}, \ell) - \mathbb{E}[e(S_i^{(0)}, S_j^{(0)}, \ell)] \right| + \left| \mathbb{E}[e(S_i^{(0)}, S_j^{(0)}, \ell)] - p(j,i,\ell)|S_i^{(0)}||S_j^{(0)}| \right|}{|S_i^{(0)}||S_j^{(0)}|} (43)$$

25

We first find an upper bound of $\left| e(S_i^{(0)}, S_j^{(0)}, \ell) - \mathbb{E}[e(S_i^{(0)}, S_j^{(0)}, \ell)] \right|$. Let $\mathcal{S}$ be the of partitions such that

$$\left| \cup_{k=1}^K \mathcal{V}_k \setminus S_k \right| \le \xi = O\left(\frac{\log(n\bar{p})^2}{\bar{p}} \right) \quad \text{for all} \quad \{S_k\}_{1 \le k \le K} \in \mathcal{S}.$$

Then,

$$
\begin{aligned}
|\mathcal{S}| &\le \binom{n}{\xi} K^\xi \\
&\le \left(\frac{ken}{\xi} \right)^\xi \\
&= \exp\left(O\left(\frac{\log(n\bar{p})^3}{\bar{p}} \right) \right).
\end{aligned}
\tag{44}
$$

For all $\{S_k\}_{1 \le k \le K} \in \mathcal{S}$ and for all $\ell \ge 1$ and $1 \le i, j \le K$, $e(S_i, S_j, \ell)$ is the sum of $|S_i||S_j|$ (or $\frac{|S_i|^2}{2}$ when $i = j$) independent Bernoulli random variables. Since the variance of $e(S_i, S_j, \ell)$ is always less than $n^2 \bar{p}$, by Chernoff inequality (e.g., Theorem 2.1.3 in [22]), with probability at least $1 - \exp\left(-\Theta\left(\frac{\log(n\bar{p})^4}{\bar{p}} \right) \right)$,

$$|e(S_i, S_j, \ell) - \mathbb{E}[e(S_i, S_j, \ell)]| \le n\log(n\bar{p})^2 \quad \text{for all} \quad i, j, \ell. \tag{45}$$

From (44) and (45), with high probability,

$$|e(S_i, S_j, \ell) - \mathbb{E}[e(S_i, S_j, \ell)]| \le n\log(n\bar{p})^2 \quad \text{for all} \quad i, j, \ell \quad \text{and} \quad \{S_k\}_{1 \le k \le K} \in \mathcal{S}.$$

Since $\{S_k^{(0)}\}_{1 \le k \le K} \in \mathcal{S}$, from the above inequality,

$$\left| e(S_i^{(0)}, S_j^{(0)}, \ell) - \mathbb{E}[e(S_i^{(0)}, S_j^{(0)}, \ell)] \right| \le n\log(n\bar{p})^2 \quad \text{for all} \quad i, j, \ell. \tag{46}$$

We now devote to the remaining part of (43). Since $|\mathcal{E}^{(0)}| = O\left(\frac{\log(n\bar{p})^2}{\bar{p}} \right)$ from Theorem 7,

$$\left| \mathbb{E}[e(S_i^{(0)}, S_j^{(0)}, \ell)] - |S_i^{(0)}||S_j^{(0)}|p(i, j, \ell) \right| \le \eta |\mathcal{E}^{(0)}| np(i, j, \ell) = O(n\log(n\bar{p})^2). \tag{47}$$

From (43), (46) and (47), with high probability,

$$|\hat{p}(j, i, \ell) - p(j, i, \ell)| = O(\log(n\bar{p})^2/n) \quad \text{for all} \quad i, j, \ell,$$

which implies that:

$$\left| \log \frac{\hat{p}(j, i, \ell)}{p(j, i, \ell)} \right| \le \frac{|\hat{p}(j, i, \ell) - p(j, i, \ell)|}{p(j, i, \ell) - |\hat{p}(j, i, \ell) - p(j, i, \ell)|} = O\left(\frac{\log(n\bar{p})^2}{np(j, i, \ell)} \right) \quad \text{for all} \quad i, j, \ell.$$

Since $e(v, S_i^{(t)}, \ell) \le e(v, \mathcal{V}) \le 10\eta n\bar{p}L$ from (H1) and $np(j, i, \ell) \ge (n\bar{p})^\kappa$ from (A3), we deduce that, for all $v \in \Gamma$ and i, j, k,

$$\sum_{\ell=0}^L e(v, S_i^{(t)}, \ell) \left| \log \frac{\hat{p}(j, i, \ell)}{\hat{p}(k, i, \ell)} - \log \frac{p(j, i, \ell)}{p(k, i, \ell)} \right| = O\left(\log(n\bar{p})^2 (n\bar{p})^{1-\kappa} \right).$$

26

Proof of (40)*:* Since $\log \frac{p(j,i,0)}{p(k,i,0)} = O(\bar{p})$ for all i, j, k and $|\mathcal{E}^{(t)}| = O(\log(n\bar{p})^2/\bar{p})$,

$$\sum_{i=1}^{K}\sum_{\ell=0}^{L} e(v, S_i^{(t)}, \ell) \log \frac{p(j,i,\ell)}{p(k,i,\ell)}$$

$$= \sum_{i=1}^{K}\left(|S_i^{(t)}|\log\frac{p(j,i,0)}{p(k,i,0)} + \sum_{\ell=1}^{L} e(v, S_i^{(t)}, \ell)\log\frac{p(j,i,\ell)p(k,i,0)}{p(k,i,\ell)p(j,i,0)}\right)$$

$$\leq \sum_{i=1}^{K}\left(|\mathcal{V}_i|\log\frac{p(j,i,0)}{p(k,i,0)} + \sum_{\ell=1}^{L} e(v, S_i^{(t)}, \ell)\log\frac{p(j,i,\ell)p(k,i,0)}{p(k,i,\ell)p(j,i,0)}\right) + \log(n\bar{p})^3$$

$$\leq \sum_{i=1}^{K}\sum_{\ell=0}^{L} e(v, \mathcal{V}_i, \ell)\log\frac{p(j,i,\ell)}{p(k,i,\ell)} + \sum_{i=1}^{K}\sum_{\ell=1}^{L} e(v, \mathcal{V}_i \setminus S_i^{(t)}, \ell)\log(2\eta) + \log(n\bar{p})^3$$

$$= \sum_{i=1}^{K}\sum_{\ell=0}^{L} e(v, \mathcal{V}_i, \ell)\log\frac{p(j,i,\ell)}{p(k,i,\ell)} + \left(e(v, \mathcal{E}^{(t)}) + e(v, \mathcal{V} \setminus H)\right)\log(2\eta) + \log(n\bar{p})^3$$

$$\leq \sum_{i=1}^{K}\sum_{\ell=0}^{L} e(v, \mathcal{V}_i, \ell)\log\frac{p(j,i,\ell)}{p(k,i,\ell)} + \log(2\eta)e(v, \mathcal{E}^{(t)}) + 2\log(n\bar{p})^3,$$

where the last inequality stems from (H3), i.e., from $e(v, \mathcal{V} \setminus H) \leq 2\log(n\bar{p})^2$ when $v \in H$.

Proof of (41)*:* Since $\mathcal{E}^{(t+1)} \subset H$ and every $v \in H$ satisfies (H2), every $v \in \mathcal{E}_{jk}^{(i+1)}$ satisfies:

$$\sum_{i=1}^{K}\sum_{\ell=0}^{L} e(v, \mathcal{V}_i, \ell)\log\frac{p(j,i,\ell)}{p(k,i,\ell)} \leq -\frac{n\bar{p}}{\log(n\bar{p})^4}.$$

Proof of (42)*:* Let $\bar{\Gamma} = \{v : e(v, \mathcal{V}) \leq 10\eta n\bar{p}L\}$ and $A_{\bar{\Gamma}}^{\ell}$ be the trimmed matrix of A^{ℓ} whose elements in rows and columns corresponding to $w \notin \bar{\Gamma}$ are set to 0. $\bar{\Gamma}$ is the set of all items that satisfy (H1) and $H \subset \bar{\Gamma}$. Let $X_{\bar{\Gamma}} = \sum_{\ell=1}^{L}(A_{\bar{\Gamma}}^{\ell} - M_{\bar{\Gamma}}^{\ell})$. We have:

$$\sum_{v\in\mathcal{E}^{(t+1)}} (e(v, \mathcal{E}^{(t)}) - \mathbb{E}[e(v, \mathcal{E}^{(t)})]) \leq 1_{\mathcal{E}^{(t)}}^{T} \cdot X_{\bar{\Gamma}} \cdot 1_{\mathcal{E}^{(t+1)}},$$

where 1_S is the vector whose v-th component is equal to 1 if $v \in S$ and to 0 otherwise. Since $\mathbb{E}[e(v, \mathcal{E}^{(t)})] \leq \bar{p}L|\mathcal{E}^{(t)}|$ and $\|X_{\bar{\Gamma}}\|_2 \leq \sqrt{n\bar{p}\log n\bar{p}}$ with high probability from Lemma 12,

$$\sum_{v\in\mathcal{E}^{(t+1)}} e(v, \mathcal{E}^{(t)}) = \sum_{v\in\mathcal{E}^{(t+1)}} \left(e(v, \mathcal{E}^{(t)}) - \mathbb{E}[e(v, \mathcal{E}^{(t)})]\right) + \bar{p}L|\mathcal{E}^{(t)}||\mathcal{E}^{(t+1)}|$$

$$\leq \|1_{\mathcal{E}^{(t)}}^{T} \cdot X_{\bar{\Gamma}} \cdot 1_{\mathcal{E}^{(t+1)}}\|_2 + |\mathcal{E}^{(t+1)}|\log(n\bar{p})$$

$$\leq \|1_{\mathcal{E}^{(t)}}^{T}\|_2\|X_{\bar{\Gamma}}\|_2\|1_{\mathcal{E}^{(t+1)}}\|_2 + |\mathcal{E}^{(t+1)}|\log(n\bar{p})$$

$$\leq \sqrt{|\mathcal{E}^{(t)}||\mathcal{E}^{(t+1)}|n\bar{p}\log(n\bar{p})} + |\mathcal{E}^{(t+1)}|\log(n\bar{p}).$$

∎

D   Proof of Theorem 3

The positive result is obtained by applying Theorem 2 to $s = \frac{1}{2}$. When $\liminf_{n\to\infty} \frac{nD(\alpha,p)}{\log(n)} \geq 1$, SP algorithm find clusters exactly with high probability. Thus, it suffices to show the negative result.

We prove the negative part by contradiction. Consider a maximum a posteriori (MAP) estimation with full parameter information. When we observe a labeld information A, the MAP estimates the clusters as follows:

$$(\hat{S}_k)_{k=1,\dots,k} = \arg\max_{(S_k)_{k=1,\dots,K}} \mathbb{P}\{(S_k)_{k=1,\dots,K}|\alpha, p, K, A\}. \tag{48}$$

Let ε^{MAP} denote the number of misclassified nodes by the MAP estimation. From the definition of the MAP estimation, for any clustring algorithm π, we have

$$\mathbb{P}\left\{\varepsilon^{\pi} \geq 1\right\} \geq \mathbb{P}\left\{\varepsilon^{\text{MAP}} \geq 1\right\}. \tag{49}$$

Thus, in what follows, we show that when $\liminf_{n\to\infty} \frac{nD(\alpha,p)}{\log(n)} < 1$, the MAP estimation is failed to find the exact clusters with high probability.

We start by Lemma 17 which finds a large deviation inequality for edge connections.

Lemma 17 *Let $\boldsymbol{x} \in \mathbb{Z}^{K\times(L+1)}$ whose $(k, \ell+1)$ element is $x_{k,\ell}$, and such that $\sum_{\ell=0}^{L} x_{k,\ell} = |\mathcal{V}_k|$ for all $1 \leq k \leq K$, $\sum_{\ell=1}^{L} x_{k,\ell} = \Theta(n\bar{p})$ for all k, and*

$$\sum_{k=1}^{K} |\mathcal{V}_k| KL(\mu(v, \mathcal{V}_k), p(i, k)) = nD \quad \text{when} \quad e(v) = \boldsymbol{x},$$

where we denote by $e(v)$ the $K \times (L+1)$ matrix whose $(k, \ell+1)$ element is $e(v, \mathcal{V}_k, \ell)$. Then,

$$\log\left(\mathbb{P}\left\{e(v) = \boldsymbol{x}\right\}\right) \geq -nD(1 + o(1)) \quad \text{when} \quad v \in \mathcal{V}_i \quad \text{and} \quad D = \Omega(\bar{p}).$$

Proof. When using the convention $\sum_{\ell=a}^{b}$ as 0 when $a > b$, we have

$\log\left(\mathbb{P}\left\{e(v) = \boldsymbol{x}\right\}\right)$

$$= \sum_{k=1}^{K}\left(\left(|\mathcal{V}_k| - \sum_{\ell=1}^{L} x_{k,\ell}\right) \log\left(p(i,k,0)\right) + \sum_{\ell=1}^{L} \log\left(p(i,k,\ell)^{x_{k,\ell}}\binom{|\mathcal{V}_k| - \sum_{m=1}^{\ell-1} x_{k,m}}{x_{k,\ell}}\right)\right)$$

$$\geq \sum_{k=1}^{K}\left(\left(|\mathcal{V}_k| - \sum_{\ell=1}^{L} x_{k,\ell}\right) \log\left(p(i,k,0)\right) + \sum_{\ell=1}^{L} \log\left(p(i,k,\ell)^{x_{k,\ell}} \frac{\left(|\mathcal{V}_k| - \sum_{m=1}^{L} x_{k,m}\right)^{x_{k,\ell}}}{x_{k,\ell}!}\right)\right)$$

$$\overset{(a)}{\geq} \sum_{k=1}^{K}\left(\left(|\mathcal{V}_k| - \sum_{\ell=1}^{L} x_{k,\ell}\right) \log\left(p(i,k,0)\right) + \sum_{\ell=1}^{L} \log\left(\left(\frac{p(i,k,\ell)e}{\frac{x_{k,\ell}}{|\mathcal{V}_k| - \sum_{m=1}^{L} x_{k,m}}}\right)^{x_{k,\ell}} \frac{1}{e\sqrt{x_{k,\ell}}}\right)\right)$$

$$\overset{(b)}{=} \sum_{k=1}^{K}\left(\left(|\mathcal{V}_k| - \sum_{\ell=1}^{L} x_{k,\ell}\right) \log\left(p(i,k,0)\right) + \sum_{\ell=1}^{L} \log\left(\frac{p(i,k,\ell)e}{\frac{x_{k,\ell}}{|\mathcal{V}_k| - \sum_{m=1}^{L} x_{k,m}}}\right)^{x_{k,\ell}}\right) - o\left(\sum_{k=1}^{K} \sum_{\ell=1}^{L} x_{k,\ell}\right)$$

$$\overset{(c)}{\geq} \sum_{k=1}^{K}\left(|\mathcal{V}_k| - \sum_{\ell=1}^{L} x_{k,\ell}\right) \log\left(p(i,k,0)\left(1 + \frac{\sum_{\ell=1}^{L} x_{k,\ell}}{|\mathcal{V}_k| - \sum_{\ell=1}^{L} x_{k,\ell}}\right)\right)$$

$$+ \sum_{k=1}^{K}\left(\sum_{\ell=1}^{L} x_{k,\ell} \log\left(\frac{p(i,k,\ell)}{\frac{x_{k,\ell}}{|\mathcal{V}_k| - \sum_{m=1}^{L} x_{k,m}}}\right)\right) - o\left(\sum_{k=1}^{K} \sum_{\ell=1}^{L} x_{k,\ell}\right)$$

$$= \sum_{k=1}^{K}\left(|\mathcal{V}_k| - \sum_{\ell=1}^{L} x_{k,\ell}\right) \log\left(\frac{p(i,k,0)}{(|\mathcal{V}_k| - \sum_{\ell=1}^{L} x_{k,\ell})/|\mathcal{V}_k|}\right) + \sum_{k=1}^{K}\left(\sum_{\ell=1}^{L} x_{k,\ell} \log\left(\frac{p(i,k,\ell)}{x_{k,\ell}/|\mathcal{V}_k|}\right)\right)$$

$$+ \sum_{k=1}^{K}\left(\sum_{\ell=1}^{L} x_{k,\ell} \log\left(\frac{|\mathcal{V}_k| - \sum_{m=1}^{L} x_{k,m}}{|\mathcal{V}_k|}\right)\right) - o\left(\sum_{k=1}^{K} \sum_{\ell=1}^{L} x_{k,\ell}\right)$$

$$\overset{(d)}{\geq} -nD - o\left(\sum_{k=1}^{K} \sum_{\ell=1}^{L} x_{k,\ell}\right)$$

$$\overset{(e)}{\geq} -nD(1 + o(1)),$$

where (a) is obtained from $n! \leq e\sqrt{n}\left(\frac{n}{e}\right)^n$; (b) stems from $\sum_{k=1}^{K} \sum_{\ell=1}^{L} x_{k,\ell} = \omega(1)$; to derive (c), we use $e^{\sum_{\ell=1}^{L} x_{k,\ell}} \geq \left(1 + \frac{\sum_{\ell=1}^{L} x_{k,\ell}}{|\mathcal{V}_k| - \sum_{\ell=1}^{L} x_{k,\ell}}\right)^{|\mathcal{V}_k| - \sum_{\ell=1}^{L} x_{k,\ell}}$ since $e \geq (1 + 1/x)^x$ for all $x > 0$; to

28

prove (d), we use the definition of $\boldsymbol{x}$ and the following inequality:

$$\sum_{\ell=1}^{L} x_{k,\ell} \log\left(\frac{|\mathcal{V}_k|}{|\mathcal{V}_k| - \sum_{m=1}^{L} x_{k,m}}\right) = \frac{\left(\sum_{\ell=1}^{L} x_{k,\ell}\right)^2}{|\mathcal{V}_k| - \sum_{\ell=1}^{L} x_{k,\ell}}(1 + o(1)) = o(\sum_{\ell=1}^{L} x_{k,\ell});$$

and (e) is obtained from the definition of $\boldsymbol{x}$ that $\sum_{\ell=1}^{L} x_{k,\ell} = \Theta(n\bar{p})$ for all k. ∎

Assume that there exists a constant $\eta > 0$ such that $\frac{nD(\alpha,p)}{\log(n)} < 1 - \eta$.

Let $(i^\star, j^\star) = \arg\min_{i,j:i<j} D_{L+}(p(i), p(j))$ (i.e., it is the hardest case to discriminate cluster $i^\star$ and cluster $j^\star$). When $n \to \infty$, one can easily check using the continuity of the KL divergence that there exists $\boldsymbol{x}^\star$ such that when $e(v) = \boldsymbol{x}^\star$,

$$\frac{\eta}{2}\log n + \sum_{k=1}^{K} |\mathcal{V}_k| KL(\mu(v, \mathcal{V}_k), p(j^\star, k)) < \sum_{k=1}^{K} |\mathcal{V}_k| KL(\mu(v, \mathcal{V}_k), p(i^\star, k)) \quad \text{and} \tag{50}$$

$$\sum_{k=1}^{K} |\mathcal{V}_k| KL(\mu(v, \mathcal{V}_k), p(i^\star, k)) \le (1 - \eta/2)\log(n). \tag{51}$$

Let $\mathcal{V}_e = \{v \in \mathcal{V}_{i^\star} : e(v) = \boldsymbol{x}^\star\}$. From (51) and Lemma 17, $\mathbb{E}[|\mathcal{V}_e|] \ge n^{\eta/4}$. Thus, from Markov inequality, with probability at least $1 - n^{-\eta/4}$, $\mathcal{V}_e$ is not empty (i.e., $|\mathcal{V}_e| \ge 1$).

Let $v^\star \in \mathcal{V}_e$ be a node in $\mathcal{V}_e$. We denote by Φ the original partition and define a slightly modified partition Ψ as follows:

$$\hat{\mathcal{V}}_{i^\star} = \mathcal{V}_{i^\star} \setminus \{i^\star\}, \quad \hat{\mathcal{V}}_{j^\star} = \mathcal{V}_{j^\star} \cup \{i^\star\}, \quad \text{and} \quad \hat{\mathcal{V}}_k = \mathcal{V}_k \quad \text{otherwise.}$$

Then, Ψ is a more likely partition than Φ from (50), i.e.,

$$\mathbb{P}\{\Phi|\alpha, p, K, A\} \ge \mathbb{P}\{\Psi|\alpha, p, K, A\} \tag{52}$$

which means that the MAP estimator does not select the exact partition when $\mathcal{V}_e$ is not empty. Therefore, from (49), every clustering algorithm π has the error probability that

$$\mathbb{E}\{\varepsilon^\pi \ge 1\} \ge 1 - n^{-\eta/4}$$

when there exists a constant $\eta > 0$ such that $\frac{nD(\alpha,p)}{\log(n)} < 1 - \eta$. ∎

E   Proof of Claim 6

When $\bar{p} = o(\frac{1}{\sqrt{n}})$, we have from Lemmas 11 and 13:

$$\lim_{n\to\infty} \frac{\log\left(\mathbb{P}\left\{\sum_{k=1}^{K} |\mathcal{V}_k| KL(\mu(v, \mathcal{V}_k), p(i, k)) \ge nD\right\}\right)}{nD}$$

$$\le \lim_{n\to\infty} \frac{\log\left(\mathbb{P}\left\{\sum_{k=1}^{K} |\mathcal{V}_k| KL(\mu(v, \mathcal{V}_k), p(i, k)) \ge nD, e(v, \mathcal{V}) \le 10\eta n\bar{p}L\right\} + \mathbb{P}\{e(v, \mathcal{V}) \ge 10\eta n\bar{p}L\}\right)}{nD}$$

$$\le \lim_{n\to\infty} \frac{\log\left(\exp\left(-nD(1 - o(1))\right) + \exp(-10\eta n\bar{p}L)\right)}{nD}$$

$$\le -1.$$

Thus, to prove the result, it suffices to show that:

$$\lim_{n\to\infty} \frac{\log\left(\mathbb{P}\left\{\sum_{k=1}^{K} |\mathcal{V}_k| KL(\mu(v, \mathcal{V}_k), p(i, k)) \ge nD\right\}\right)}{nD} \ge -1.$$

Next, we denote by $e(v)$ the $K \times (L + 1)$ matrix whose $(k, \ell + 1)$ element is $e(v, \mathcal{V}_k, \ell)$. Let $\boldsymbol{x} \in \mathbb{Z}^{K \times (L+1)}$ whose $(k, \ell + 1)$ element is $x_{k,\ell}$, and such that $\sum_{\ell=0}^{L} x_{k,\ell} = |\mathcal{V}_k|$ for all $1 \le k \le K$, $\sum_{\ell=1}^{L} x_{k,\ell} = \Theta(n\bar{p})$ for all k, and

$$\sum_{k=1}^{K} |\mathcal{V}_k| KL(\mu(v, \mathcal{V}_i), p(i, k)) = nD(1 + o(1)) \quad \text{when} \quad e(v) = \boldsymbol{x}.$$

We can easily check using the continuity of the KL divergence that such a choice for x is possible. Then, from Lemma 17,

$$\log\left(\mathbb{P}\left\{\sum_{k=1}^{K}|\mathcal{V}_k|KL\left(\mu(v,\mathcal{V}_k),p(i,k)\right)\geq nD\right\}\right)\geq\log\left(\mathbb{P}\left\{e(v)=x\right\}\right)$$
$$\geq -nD(1+o(1)).$$

∎