VEC-SBM: Optimal Community Detection with Vectorial Edges Covariates

Guillaume Braun
RIKEN AIP

Masashi Sugiyama
RIKEN AIP
University of Tokyo

Abstract

Social networks are often associated with rich side information, such as texts and images. While numerous methods have been developed to identify communities from pairwise interactions, they usually ignore such side information. In this work, we study an extension of the Stochastic Block Model (SBM), a widely used statistical framework for community detection, that integrates vectorial edges covariates: the Vectorial Edges Covariates Stochastic Block Model (VEC-SBM). We propose a novel algorithm based on iterative refinement techniques and show that it optimally recovers the latent communities under the VEC-SBM. Furthermore, we rigorously assess the added value of leveraging edge's side information in the community detection process. We complement our theoretical results with numerical experiments on synthetic and semi-synthetic data.

1 Introduction

Networks are a powerful tool for representing relational data, where each entity is represented by a node and pairwise connections between these entities are encoded by edges. Over the past decades, numerous clustering methods have been devised to extract meaningful insights from graph-structured datasets. These methods are often evaluated under the Stochastic Block Model (Holland et al., 1983), a random graph model where each edge is sampled independently with a probability depending solely on the latent communities of the corresponding nodes. However, a notable limitation of the SBM is its exclusive consideration of binary interactions.

In recent years, there has been growing interest in developing extensions of the SBM that can incorporate more information. This includes the Weighted SBM (Aicher et al., 2014), which assigns scalar weights to edges, the Labeled SBM (Yun and Proutiere, 2016) where labels correlated to the nodes' community are available, the Multi-Layer SBM (Vallès-Català et al., 2016), accommodating multimodal interactions in distinct layers, the Contextual SBM (Deshpande et al., 2018), linking each node to a covariate vector, or the recently introduced Embedded Topic SBM (Boutin et al., 2023a), where textual information is associated with each edge.

In this work, we consider a variant of the Embedded Topic SBM: the Vectorial Edges Covariate SBM (VEC-SBM). Under the VEC-SBM, each observed edge is associated with a vector, distinguishing it from the Weighted SBM, which solely permits scalar weights, and the Multi-Layer SBM, where edge presence is sampled independently on each layer. Consequently, methodologies and theoretical guarantees established for these existing models cannot be directly applied to the VEC-SBM. Moreover, in contrast to prior work such as that by Boutin et al. (2023a), which primarily focuses on practical applications, our study centers on the statistical analysis of the VEC-SBM. In particular, our analysis quantifies the added value of the side information provided by the edges and shows that it has a multiplicative effect on the signal-to-noise ratio (SNR). This is in contrast with the node's side information that has an additive impact on the SNR (Abbe et al., 2022), further motivating the incorporation of edges' side information in clustering algorithms.

Our contributions. In this work, we make the following contributions.

- We introduce a novel algorithm for graph clustering that incorporates edge vectorial covariates.

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

Our algorithm is computationally efficient and applicable to various settings.

- We rigorously analyze our algorithm under the VEC-SBM and demonstrate that it achieves a statistically optimal convergence rate. We also provide valuable insights by quantifying the information gain resulting from the inclusion of edge covariates: under the VEC-SBM, the SNR will depend on the difference between the means of the covariates of different classes multiplied by the average degree of the nodes in the graph. Even if this difference of means is small this can lead to a considerable increase in the SNR depending on the sparsity level of the graph.
- We conduct comprehensive numerical experiments on synthetic data to further substantiate our findings. These experiments highlight the importance of leveraging both the network structure and covariate information for effectively recovering all communities. Moreover, we apply our algorithm to a real-world dataset, with synthetic edge covariates, demonstrating its practical applicability.

Related Work. Various extensions of the SBM have been developed to incorporate side information. In this paragraph, we describe the main existing algorithmic approaches used for this purpose.

- *Discretization of edge weights.* Xu et al. (2017) introduced an algorithm that achieves the optimal rate of convergence under the Weighted SBM by discretizing edge weights. However, this strategy becomes computationally inefficient when dealing with high-dimensional covariates, and selecting the appropriate level of discretization can be challenging.
- *Tensor and matrix factorization methods.* Tensor methods, such as those presented in Jing et al. (2021) or Paul and Chen (2020), offer an alternative approach to incorporate side information. However, these methods are typically applied to multi-layer graphs, which exhibit a distinct noise structure compared to our setting.
- *Model-based approaches.* Since the Maximum Likelihood Estimator (MLE) is intractable for models based on the SBM, several alternative approaches have been used. Cerqueira and Levina (2023) proposed a pseudo-likelihood approach that avoids the need for discretization, but it relies on the assumption that weights are sampled from a univariate Gaussian mixture. Their method attains a convergence rate matching the lower bound established by Xu et al. Xu et al. (2017) under the Weighted SBM, albeit up to a constant factor. Another prevalent model-based approach involves the use of variational EM algorithms (Léger, 2016; Bouveyron et al., 2016). While effective, these methods are often computationally demanding and lack theoretical guarantees. The recent work of Boutin et al. (2023a,b) combined the variational approach with a deep learning architecture to simultaneously extract topics from edges' textual side information and cluster the network. In this work, we consider a simpler setting where the edge covariates can be directly exploited. Our main purpose is to statistically analyze the model to get insight into the added value of edge-side information.
- *Iterative refinement approaches/ alternating optimization.* Instead of using the pseudolikelihood or the variational method, our algorithm relies on an iterative refinement procedure based on a simplified version of the Maximum A Posteriori (MAP) estimator. The global proof strategy is based on the seminal work of Gao and Zhang (2022) and inspired by the analysis of the Contextual SBM (CSBM) (Braun et al., 2022). However, contrary to the CSBM, there is a dependency between the two available sources of information under the VEC-SBM, since we only access to edge covariates for the observed edges. This poses new challenges and requires the development of new analysis techniques. Due to the particular noise structure in our setting, the error decomposition is different and requires new concentration inequalities and techniques to be controlled. Using alternating optimization to solve non-convex optimization problems is a popular approach that has also been used for other problems including matrix sensing (Stöger and Soltanolkotabi, 2021), dictionary learning (Liang et al., 2022), heterogeneous matrix factorization (Shi et al., 2023), and multi-task regression (Thekumparampil et al., 2021) to mention but a few.

Notations. We will use Landau standard notations $o(.)$ and $O(.)$. For sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, if there is a constant $C > 0$ such that $a_n \leq Cb_n$ (resp. $a_n \geq Cb_n$) for all n we will write $a_n \lesssim b_n$ (resp. $a_n \gtrsim b_n$). If $a_n \lesssim b_n$ and $a_n \gtrsim b_n$, then we write $a_n \asymp b_n$. Matrices will be denoted by uppercase letters. The i-th row of a matrix A will be denoted as $A_{i:}$ and depending on the context can be interpreted as a column vector. The column j of A will be denoted by $A_{:j}$, and the (i, j)th entry by A_{ij}. The transpose of A is denoted by $A^\top$. I_k denotes the $k \times k$ identity matrix. We use $\|\cdot\|$ and $\|\cdot\|_F$ to respectively denote the spectral norm

(or Euclidean norm in the case of vectors) and the Frobenius norm. The number of non-zero entries of a matrix A is denoted $\mathrm{nnz}(A)$. $A^\dagger$ denote the pseudo-inverse of A. The maximum between a and b will be denoted by $a \vee b$. The indicator function of a set C is denoted by $\mathbf{1}_C$.

2 Model and algorithm description

In Section 2.1 we describe the generative model and discuss the assumptions we made for the analysis. Then, in Section 2.2 we introduce our algorithm.

2.1 The Vectorial Edge Covariates Stochastic Block Model (VEC-SBM)

The VEC-SBM is an extension of the SBM where a vector is associated with each observed edge of a graph sampled from an SBM. The distribution of these edge vector covariates only depends on the community the endpoints forming the edge belong to. More formally, the VEC-SBM is defined by the following parameters.

- A set of nodes $\mathcal{N} = [n]$ and a partition of $\mathcal{N}$ into K communities $\mathcal{C}_1, \dots, \mathcal{C}_K$.
- A membership matrix $Z \in \{0,1\}^{n \times K}$ such that there is exaclty one 1 in every row. Each membership matrix Z can be associated bijectively with a partition function $z : [n] \to [K]$ such that $z(i) = z_i = k$ where k is the unique column index satisfying $Z_{ik} = 1$.
- A symmetric connectivity matrix of probabilities between communities
$$\Pi = (\Pi_{kk'})_{k,k' \in [K]} \in [0,1]^{K \times K}.$$
- A family of centroids $\mu = (\mu_{kk'})_{k,k' \in [K]}$ such that $\mu_{kk'} \in \mathbb{R}^d$, where $d = O(1)$ and $\mu_{kk'} = \mu_{k'k}$ for all $k, k' \in [K]$.
- A family of covariance matrices $\Sigma = (\Sigma_{kk'})_{k,k' \in [K]}$ such that $\Sigma_{kk'} \in \mathbb{R}^{d \times d}$ and $\Sigma_{kk'} = \Sigma_{k'k}$ for all $k, k' \in [K]$.

A graph with n nodes and edge covariates of dimension d can be represented by a tensor $R \in \mathbb{R}^{n \times n \times d}$. It is distributed according to the VEC-SBM(Z, Π, μ, Σ) if the entries of R are generated as follows. First, the presence or absence of an edge is encoded by a matrix $A \in \{0,1\}^{n \times n}$ with entries samples independently by
$$A_{ij} \overset{\text{ind.}}{\sim} \mathcal{B}(P_{ij}), \quad i, j \in [n],\ i \leq j$$
where $\mathcal{B}(p)$ denotes a Bernoulli distribution with parameter p, and $P = \mathbb{E}(A) = Z \Pi Z^\top$. The noise is denoted by $E = A - \mathbb{E}(A)$. The sparsity level of the graph is denoted by $p_{max} = \max_{i,j} p_{ij}$. We will focus on the regime where $np_{max} = \Omega(\log n)$ and $np_{max} = o(n)$.

Then, for each couple (i, j), we sample independently (conditionally on Z) an edge covariate $G_{ij} \in \mathbb{R}^d$ such that
$$G_{ij} = \mu_{z(i)z(j)} + \epsilon_{ij},$$
where ϵ_{ij} is a centered Gaussian[1] variable with covariance $\Sigma_{z(i)z(j)}$. Since multiplying all the edges of the graph by a constant will not modify the information contained in the covariates, we can assume that $\|\mu_{kk'}\|_\infty \leq 1$ for all $k, k' \in [K]$. Finally, the observed tensor R is given by $R_{ijd} = A_{ij}(G_{ij})_d$. For the analysis, we will make some additional assumptions.

Assumption A1 (Approximately balanced communities)**.** *The communities $\mathcal{C}_1, \dots, \mathcal{C}_K$ are approximately balanced, i.e., there exists a constant $\alpha \geq 1$ such that for all $k \in [K]$ we have*
$$\frac{n}{\alpha K} \leq |\mathcal{C}_k| \leq \frac{\alpha n}{K}.$$

Assumption A2 (Isotropic variance)**.** *For all $k, k' \in [K]$, $\Sigma_{kk'} = I_d$.*

Assumption A3 (Symmetric SBM)**.** *The connectivity matrix Π is of the form $q\mathbf{1}\mathbf{1}^\top + (p-q)I_K$ with $1 > p > q > 0$.*

Assumption A4 (Limited graph information)**.** *We have $p = o(1)$, $p/q = O(1)$, $np = \Omega(\log n)$ and $n(\sqrt{p} - \sqrt{q})^2/K < \log n$.*

Assumption A1 is standard in clustering literature. When the communities are highly unbalanced, the problem becomes noticeably more difficult, see for example Mukherjee et al. (2023). Assumption A2 is used to simplify the exposition. We believe that our results can be extended to the general case (e.g. Chen and Zhang (2021) for Gaussian mixture models) without changing the proof strategy but at the price of additional technicalities. Moreover, our experiments in Section 4 show that the algorithm we analyzed performs well even if this assumption is not satisfied. Assumption A3 is also used for convenience, but could be removed at the cost of additional technicalities. Finally, Assumption A4 implies that there is not enough information in the graph to recover the clusters, motivating the use of side information. In particular, $n(\sqrt{p} - \sqrt{q})^2/K < \log n$ corresponds to the regime where exact recovery is impossible (Zhang and Zhou, 2016).

The quality of the clustering is evaluated through the

[1] We made this assumption to simplify the exposition, but we believe that the proof can be extended to Sub-Gaussian r.v.

misclustering rate r defined by

$$r(\hat{z}, z) = \frac{1}{n} \min_{\pi \in \mathfrak{S}} \sum_{i \in [n]} \mathbf{1}_{\{\hat{z}(i) \neq \pi(z(i))\}},$$

where $\mathfrak{S}$ denotes the set of permutations on $[K]$. An estimator $\hat{z}$ achieves *exact recovery* if $r(\hat{z}, z) = 0$ with probability $1 - o(1)$ as n tends to infinity. It achieves *weak consistency* (or almost full recovery) if $\mathbb{P}(r(\hat{z}, z) = o(1)) = 1 - o(1)$ as n tends to infinity. A more complete overview of the different types of consistency and the sparsity regimes where they occur can be found in Abbe (2018).

2.2 Algorithm description

Let us denote

$$MAP_i(\mathcal{C}, \Pi, \mu, \Sigma) = \sum_{l \in [K]} \sum_{j \in \mathcal{C}_l} A_{ij} \log(\Pi_{kl}) + (1 - A_{ij}) \log(1 - \Pi_{kl}) - \sum_{l \in [K]} \sum_{j \in \mathcal{C}_l} A_{ij} (G_{ij} - \mu_{kl})^\top \Sigma_{kl} (G_{ij} - \mu_{kl}) - \frac{1}{2} \det(\Sigma_{kl})$$

the logarithm of the MAP of a node i such that $z(i) = k$, given Π, μ, Σ and a partition $\mathcal{C}$ of $[n]$. At each step t, `IR-VEC` (cf. Algorithm 1) estimates the model parameters and then updates the partition encoded by $Z^{(t)}$ based on MAP_i. We will denote by $\mathcal{C}_k^{(t)}$ the set of nodes i such that $Z_{ik}^{(t)} = 1$, i.e. the nodes that are associated with community k at time t.

For the analysis, we will consider a simplified version of `IR-VEC` where at each step $\Sigma_{kk'} = I_d$ for all $k, k' \in [K]$. This version of the algorithm will be referred to as `sIR-VEC`. Despite ignoring the covariance structure, Section 4 shows that `sIR-VEC` performs as well as `IR-VEC` even if the edge covariates are non-isotropic.

Computational complexity. Estimating the model parameters at each step requires at most $O(\mathrm{nnz}(A)d^2)$ elementary operations. Given the estimate of the model parameters, updating the partition requires $O(\mathrm{nnz}(A)Kd)$ operations where $\mathrm{nnz}(A)$ corresponds to the number of non zero entries of A. The global complexity of the algorithm is hence $O(\mathrm{nnz}(A) \max(d^2, Kd))$. Under the VEC-SBM, $\mathrm{nnz}(A) \asymp n^2 p_{max}$.

Initialization. We use the vanilla spectral method on A for initialization. While the accuracy provided by this method can be very poor in challenging scenarios, we show experimentally in Section 4 that it doesn't affect the performances of `sIR-VEC`. We leave as an open problem the analysis of random initialization.

Algorithm 1 Iterative refinement for the VEC-SBM (`IR-VEC`)

Input: The number of communities K, A, G, an initial estimate of the partition $Z^{(0)}$ of the nodes, a number of iteration T.

1: **for** $0 \leq t \leq T - 1$ **do**
2: Estimate the model parameters

$$W^{(t)} = (Z^{(t)})^\dagger, \ \Pi^{(t)} = {W^{(t)}}^\top A W^{(t)}$$
$$\mu_{kk'}^{(t)} = \sum_{\substack{i \in \mathcal{C}_k^{(t)} \\ j \in \mathcal{C}_{k'}^{(t)}}} A_{ij} G_{ij} \Big/ \sum_{\substack{i \in \mathcal{C}_k^{(t)} \\ j \in \mathcal{C}_{k'}^{(t)}}} A_{ij}$$
$$\Sigma_{kk'}^{(t)} = \sum_{\substack{i \in \mathcal{C}_k^{(t)} \\ j \in \mathcal{C}_{k'}^{(t)}}} A_{ij} (G_{ij} - \mu_{kk'}^{(t)})(G_{ij} - \mu_{kk'}^{(t)})^\top \Big/ \sum_{\substack{i \in \mathcal{C}_k^{(t)} \\ j \in \mathcal{C}_{k'}^{(t)}}} A_{ij}.$$

3: Update the partition

$$z_i^{(t+1)} = \arg\max_{k \in [K]} MAP_i(\mathcal{C}^{(t)}, \Pi^{(t)}, \mu^{(t)}, \Sigma^{(t)}).$$

4: **end for**

Output: A partition of the nodes $Z^{(T)}$.

3 Analysis method and main results

In this section, we first introduce some notations associated with the error decomposition, present our main results, and then outline the proof strategy. The details of the proofs can be found in the supplementary material.

3.1 Error decomposition

Our analysis is based on the framework developed by Gao and Zhang (2022). This framework has been used to analyze other clustering models with similar flavors, such as the CSBM (Braun et al., 2022) or the Tensor Block Model (Han et al., 2022). However, we emphasize that previous results cannot be adapted straightforwardly to our setting due to the specific noise structure induced by the VEC-SBM. In particular, in the VEC-SBM, there is a dependence between the two sources of noise: the graph and the covariates. This dependence requires new techniques and concentration inequalities to control the noise.

To understand how `sIR-VEC` can lead to an improvement of the partition, one needs to analyze in which situation a node i is misclassified after one refinement step. It corresponds to the condition

$$a \neq \arg\max_{k \in [K]} MAP_i(\mathcal{C}^{(t)}, \Pi^{(t)}, \mu^{(t)}, I_d).$$

By some elementary algebra, one can show that the previous condition is equivalent to the existence of $b \in$

$[K] \setminus \{a\}$ such that

$$C_i(a,b) < -\Delta^2(a,b) + F_{ib}^{(t)} + G_{ib}^{(t)},$$

where $F_{ib}^{(t)}$ and $G_{ib}^{(t)}$ are error terms specified in the appendix (Section A.1), and the signal term $\Delta^2(a,b)$ and stochastic term $C_i(a,b)$ are given by

$$\Delta^2(a,b) = \log(p/q)(|\mathcal{C}_a|p - |\mathcal{C}_b|q) + \sum_{l\in[K]} |\mathcal{C}_l|\Pi_{al}\,\|\mu_{al} - \mu_{bl}\|^2 \quad (3.1)$$

$$C_i(a,b) = \log(\frac{p}{q})\left(\sum_{j\in\mathcal{C}_a} E_{ij} - \sum_{j\in\mathcal{C}_b} E_{ij}\right) + \sum_{l\in[K]}\sum_{j\in\mathcal{C}_{l,-i}} \left(E_{ij}\,\|\mu_{al} - \mu_{bl}\|^2 + 2A_{ij}\langle \epsilon_{ij}, \mu_{al} - \mu_{bl}\rangle\right).$$

The first part of the signal only depends on the graph while the second part depends on the covariates and the graph connectivity parameters. For instance, in the case where the communities are exactly balanced and $p = q$, the second part corresponds to $(np/K)\sum_l \|\mu_{al} - \mu_{bl}\|^2$, i.e. the sparsity level of the graph np is multiplied by the average distance between the edge covariates means $\sum_l \|\mu_{al} - \mu_{bl}\|^2 / K$, hence the multiplicative effect.

3.2 Convergence guarantee

The following theorem shows that if the initialization $z^{(0)}$ is good enough, then `sIR-VEC` converges in $O(\log n)$ iterations and achieves a misclustering rate that decreases exponentially in the SNR formally defined as $\Delta^2_{min} = \min_{a\neq b} \Delta^2(a,b)$.

Theorem 1. *Assume that $\Delta^2_{min} \asymp \log n$. Under assumptions A1, A2,A3 and A4, if $z^{(0)}$ is such that*

$$r(z, z^{(0)}) \leq \frac{\epsilon}{K^2}$$

for a constant ϵ small enough, then with probability at least $1 - n^{-\Omega(1)}$ we have for all $t \gtrsim \log n$

$$r(z^{(t)}, z) \leq e^{-(c+o(1))\Delta^2_{min}}$$

where $c > 0$ is the constant appearing in Lemma 1.

Remark 1. *The condition on initialization implies having $r(z^{(0)}, z) = O(1/K^2)$. This is a more stringent requirement compared to the condition in Braun et al. (2022), which only necessitates $r(z^{(0)}, z) = O(1/K)$. This dependency on K is likely to be an artifact of the proof. If we could replace the factor K^2 in Lemma 2 with K, we could relax the initial condition to $r(z^{(0)}, z) = O(1/K)$. In Section 4, we experimentally demonstrate that `sIR-VEC` performs well even when initialized with an almost non-informative $z^{(0)}$.*

Remark 2. *The misclustering rate of the SBM is on the order of $\exp(-n(\sqrt{p} - \sqrt{q})^2/K)$. However, when $p \approx q$ accurate recovery of communities becomes challenging. In a similar context under the VEC-SBM, if $\min_{a\neq b\in[K]} \sum_l \|\mu_{al} - \mu_{bl}\|^2 \geq v > 0$, then Δ^2_{min} is on the order $vnp/K = \Omega(\log n) \gg n(\sqrt{p} - \sqrt{q})^2/K$. When $l = 1$, it reduces to a Weighted SBM with Gaussian weights. Utilizing the closed-form expression for the Hellinger distance between Gaussian r.v., the result from Xu et al. (2017) shows that when $p = q$ the misclustering rate is $\exp(-2\log n(1-\exp(-\min_{a\neq b\in[K]} |\mu_a - \mu_b|^2)))$. Through a first-order Taylor approximation, we obtain $1-\exp(-\min_{a\neq b\in[K]} |\mu_a - \mu_b|^2) \approx |\mu_a - \mu_b|^2$, matching our convergence rate up to a constant factor. Notice that under the CSBM (Braun et al., 2022), the SNR is of order $n(p-q) + \min_{a\neq b} \|\mu_a - \mu_b\|^2$. By consequence, the information added by nodes covariate is independent of the sparsity level of the graph. In our setting, it is multiplied by the sparsity level of the graph np. Due to the multiplicative effect, integrating edge-side information could have a stronger impact on the SNR than node-side information.*

Sketch of the proof of Theorem 1. The result is obtained by using the framework developped by Gao and Zhang (2022). The oracle error is controlled by using a conditioning argument, see Lemma 1. The main challenge, as discussed in Section 3.5, is to control the noise. Since the calculations involved are long, we relegated them to the appendix, Section A. □

3.3 Minimax lower-bound

We are going to show that the convergence rate of Theorem 1 is optimal; up to a constant factor. Assume that the covariates are Gaussian, i.e. $\epsilon_{ij} \sim \mathcal{N}(0, I_d)$ and consider the following space of parameters

$$\Theta = \{p = q \in [0,1], \mu_{kk'} \in [-1,1]^d, \forall k, k' \in [K]\}.$$

Theorem 2. *If $\Delta_{min} \to \infty$, there exists a constant $c' > 0$ such that*

$$\inf_{\hat{z}} \sup_{\theta\in\Theta} \mathbb{E}(r(\hat{z}, z)) \geq \exp(-c'\Delta^2_{min}).$$

If $\Delta^2_{min} = O(1)$, then $\inf_{\hat{z}} \sup_{\theta\in\Theta} \mathbb{E}(r(\hat{z}, z)) \geq c$ for some positive constant c.

Remark 3. *We exclusively examined the extreme scenario where the graph provides no information about the community structure. As the minimax lower-bound is only tight up to a constant factor, extending this result to the case where $p > q$ is straightforward by lower bounding the signal using either its first or second part, see equation (3.1).*

Sketch of the proof. First, we lower bound the minimax risk by the error associated with a two-hypothesis testing problem by using the argument of Gao et al.

(2018). Since the optimal test is achieved by the likelihood ratio (Neyman Pearson Lemma), it is sufficient to lower the probability of failure of this optimal test. This can be done by first conditioning on A and using the well-known properties of Gaussian's r.v., and then integrating over A, cf. Section B in the appendix for details. □

3.4 Oracle error

If we ignore the error terms $F_{ib}^{(t)}$ and $G_{ib}^{(t)}$, a node i is misclassified when $C_i(a,b) < -\Delta^2(a,b)$. This is an unavoidable source of error since it corresponds to the error made by the algorithm after one iteration initialized with the ground-truth partition and with the true model parameters. The error occurring in this way can be quantified by the **oracle error** defined for all $\delta \in (0, 1/2]$ by

$$\xi(\delta) = \sum_{i=1}^{n} \sum_{b \in [K] \setminus z_i} \Delta^2(z_i, b) \mathbf{1}_{\{C_i(z_i,b) \leq -(1-\delta)\Delta^2(z_i,b)\}}.$$

Lemma 1. *Let $\delta \in (0, 1/2]$ be a constant, and let us denote for any given $i \in [n]$ and $b \in [K] \setminus z_i$ the event*

$$\Omega_1(z_i, b) = \left\{ C_i(z_i, b) \leq -(1-\delta)\Delta^2(z_i, b) \right\}.$$

Under the assumptions of Theorem 1, there exists a constant $c > 0$ such that for all $z_i \neq b$

$$\mathbb{P}(\Omega_1(z_i, b)) \leq e^{-c\Delta_{min}^2}.$$

Proof. Assume that $z_i = a$. First, let us decompose $C_i(a,b) + \Delta^2(a,b)$ as $Y_1 + Y_2$ where

$$Y_1 = \log(\frac{p}{q}) \left(\sum_{j \in \mathcal{C}_a} A_{ij} - \sum_{j \in \mathcal{C}_b} A_{ij} \right)$$

and

$$Y_2 = \sum_{l \in [K], j \in \mathcal{C}_l} A_{ij} \left\| \mu_{al} - \mu_{bl} \right\|^2 + 2A_{ij} \langle \epsilon_{ij}, \mu_{al} - \mu_{bl} \rangle.$$

Conditionally on $A_{i:}$, we have for all $t \in \mathbb{R}$

$$\begin{aligned} \mathbb{E}(e^{tY_2} | A_{i:}) &= e^{t \sum_{l \in [K], j \in \mathcal{C}_l} A_{ij} \|\mu_{al} - \mu_{bl}\|^2} \\ &\times \mathbb{E}(e^{2t \sum_{l \in [K]} \sum_{j \in \mathcal{C}_l} A_{ij} \langle \epsilon_{ij}, \mu_{al} - \mu_{bl} \rangle} | A_{i:}) \\ &\leq e^{(t+2t^2) \sum_{l \in [K]} \sum_{j \in \mathcal{C}_l} A_{ij} \|\mu_{al} - \mu_{bl}\|^2}. \end{aligned}$$

(since ϵ_{ij} are ind. Sub-Gaussian r.v.)

Let us denote

$$\Delta_A^2 = \Delta_A^2(a,b) = \sum_{l \in [K], j \in \mathcal{C}_l} A_{ij} \left\| \mu_{al} - \mu_{bl} \right\|^2.$$

We have shown that for all t

$$\mathbb{E}(e^{t(Y_1+Y_2)}) \leq \mathbb{E}(e^{tY_1 + (t+2t^2)\Delta_A^2}).$$

We can rewrite $tY_1 + (t+2t^2)\Delta_A^2$ as a weighted sum of independent Bernoulli trials $\sum_j w_j(t) A_{ij}$ where $w_j(t) = t\log(p/q) + (t+2t^2)\left\|\mu_{aa} - \mu_{ba}\right\|^2$ for $j \in \mathcal{C}_a$, $w_j(t) = -t\log(p/q) + (t+2t^2)\left\|\mu_{ab} - \mu_{bb}\right\|^2$ for $j \in \mathcal{C}_b$ and $w_j(t) = (t+2t^2)\left\|\mu_{al} - \mu_{bl}\right\|^2$ when $j \in \mathcal{C}_l$ for $l \neq a, b$. Hence, we obtain

$$\begin{aligned} \log\mathbb{E}(e^{tY_1 + (t+2t^2)\Delta_A^2}) &\leq \sum_j p_{ij}(e^{w_j(t)} - 1) \\ &\leq \sum_j p_{ij} w_j(t) + 0.5 e^{\sup_j |w_j(t)|} \sum_j p_{ij} w_j(t)^2. \end{aligned}$$

(by Taylor-Lagrange formula)

Since $p/q = O(1)$ and $\left\|\mu_{al} - \mu_{bl}\right\|^2 \leq 4$ for all $l \in [K]$, one can choose $t^* < 0$ close to 0 such that

$$e^{\sup_j |w_j(t^*)|} \sum_j p_{ij} w_j(t^*)^2 \leq |\sum_j p_{ij} w_j(t^*)|$$

and

$$\sum_j p_{ij} w_j(t^*) \leq -c' \Delta^2(a,b)$$

for some positive constant c'. By consequence

$$\begin{aligned} \mathbb{P}(\Omega_1(z_i, b)) &= \mathbb{P}(Y_1 + Y_2 \leq \delta\Delta^2(a,b)) \\ &\leq \mathbb{E}(e^{t^*(Y_1+Y_2)}) e^{-t^*\delta\Delta^2(a,b)} \\ &\leq e^{-0.5c'\Delta^2(a,b) - t^*\delta\Delta^2(a,b)} \\ &\leq e^{-0.25c'\Delta^2(a,b)} \end{aligned}$$

for all $\delta > 0$ smaller than $\min\{c'|4t^*|^{-1}, 1/2\} = 1/2$. □

Corollary 1. *Under the assumptions of Theorem 1, with probability at least $1 - e^{-\Delta_{min}^2}$ we have for some constant $c > 0$*

$$\xi(\delta) \leq n e^{-c\Delta_{min}^2}.$$

Proof. By Lemma 1, we have

$$\mathbb{E}(\xi(\delta)) \leq n e^{-(c - o(1))\Delta_{min}^2}$$

since K is constant. Hence, by Markov inequality

$$\mathbb{P}(\xi(\delta) \geq e^{\Delta_{min}} \mathbb{E}(\xi(\delta))) \leq e^{-\Delta_{min}}.$$

Since $e^{\Delta_{min}} \mathbb{E}(\xi(\delta))) \leq n e^{-(c+o(1))\Delta_{min}^2}$ we obtain the result of the lemma. □

3.5 Control of the noise

To apply Theorem 3.1 in Gao and Zhang (2022) to show that the error contracts at each step until reaching the oracle error, one needs to prove that the noise terms $F_i^{(t)}$ and $G_i^{(t)}$ satisfy the following conditions. Let $\tau = \epsilon n \Delta_{min}^2 / K^2$ where $\epsilon > 0$ and let $\delta \in (0, 1/2)$ be a constant. Let us define the weighted Hamming loss

$$l(z, z') = \sum_{i=1}^{n} \Delta^2(z_i, z_i') \mathbf{1}_{\{z_i \neq z_i'\}}.$$

Condition C1 (F-error type)**.** *Assume that*

$$\max_{\{z^{(t)}: l(z, z^{(t)}) \leq \tau\}} \sum_{i=1}^{n} \max_{b \in [K] \setminus z_i} \frac{(F_{ib}^{(t)})^2}{\Delta^2(z_i, b) l(z, z^{(t)})} \leq \frac{\delta^2}{256}$$

for all $t \geq 0$ holds with probability at least $1 - n^{-\Omega(1)}$.

Condition C2 (G-error type)**.** *Assume that*

$$\max_{i \in [n]} \max_{b \in [K] \setminus z_i} \frac{|G_{ib}^{(t)}|}{\Delta^2(z_i, b)} \leq \frac{\delta}{4}$$

holds uniformly on the event $\{z^{(t)} : l(z, z^{(t)}) \leq \tau\}$ for all $t \geq 0$ with probability at least $1 - n^{-\Omega(1)}$.

Condition C1 necessitates a uniform control of the noise induced by the $F^{(t)}$ error term in an ℓ_2 norm sense. Additionally, Condition C2 requires a uniform control of the l_∞ norm of the $G^{(t)}$ error term. Typically, the $F^{(t)}$-error term depends on the estimation error of the partition $\left\|Z^{(t)} - Z\right\|$ while the $G^{(t)}$-error term depends on the parameter estimation error $\left\|\Pi^{(t)} - \Pi\right\|$.

The main technical challenge to prove the consistency or `sIR-VEC` is to show that the previous conditions hold. In particular, one needs to control the estimation error of the model parameters uniformly. For the SBM part, it can be done as in Braun et al. (2022), but bounding uniformly the error $\left\|\mu_{kk'} - \mu_{kk'}^{(t)}\right\|$ requires a new approach: contrary to the CSBM setting, the edge centroids are estimated on random samples that depend on A and the current estimate of the partition $\mathcal{C}^{(t)}$. This is the object of the following lemma.

Lemma 2. *Under the assumption of Theorem 1 we have with probability at least $1 - n^{-\Omega(1)}$, for all $z^{(t)}$ such that $l(z^{(t)}, z) \leq \tau$*

$$\max_{b,l \in [K]} \left\|\mu_{bl} - \mu_{bl}^{(t)}\right\| \lesssim K^{1.5} \left(\sqrt{\frac{l(z^{(t)}, z)}{n \Delta_{min}^2}} \vee \frac{\sqrt{\log K}}{n p_{max}} \right).$$

Sketch of the proof. To obtain an uniform bound over $z^{(t)}$, we first need to control uniformly over all set S the quantities $\left\|\sum_{i \in S} \epsilon_i\right\|$ where ϵ_i are i.i.d. sub-Gaussian r.v. This can be done by using Lemma A.1 in Lu and Zhou (2016). Secondly, we need to control uniformly over $T_1, T_1 \subset [n]$ the sums $\sum_{i \in T_1, j \in T_2} A_{ij}$. This can be done by showing that A satisfies the discrepancy property, e.g. Lei and Rinaldo (2015). The details can be found in the appendix (Section A.1.1, Lemma 4). □

We will also need control of the following term to establish Condition C1.

Lemma 3. *Under the assumption of Theorem 1 we have with probability at least $1 - n^{-\Omega(1)}$*

$$\sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_{k'}, j' \in \mathcal{C}_{k''}} E_{ij} E_{ij'} \lesssim (n/K)^2 p_{max},$$

for all $k, k', k'' \in [K]$.

Proof. Let's fix $k, k', k'' \in [K]$. To simplify the exposition, we will assume that each class is of size n/K. We want to bound

$$S = \sum_{j,j'} \sum_{i \in \mathcal{C}_k} E_{ij} E_{ij'} = \sum_{j,j'} \langle E_{:j}, E_{:j'} \rangle_{\mathcal{C}_k},$$

where the scalar product is restricted to the entries of $E_{:j}$ in $\mathcal{C}_k$. In the following, we will drop the subscript when clear from the context.

Case where k, k', and k'' are all distinct. S is a sum of $(n/K)^3$ bounded and centered independent r.v. with variance of order p_{max}^2 that can be handled with standard concentration inequalities.

Case where $k \neq k' = k''$. We have $S = \sum_{j \neq j'} \langle E_{j:}, E_{:j'} \rangle + \|E\|_F^2$. It is easy to show that w.h.p. $\|E\|_F^2 \lesssim (n/K)^2 p_{max}$. So we can focus on the first summand. Observe that $\sum_{j \neq j'} \mathbb{E}(\langle E_{:j}, E_{:j'} \rangle) = 0$. To remove the dependencies in the sum, we will use a decoupling argument similar to the one used in Rudelson and Vershynin (2013) to prove Hanson-Wright inequality. This strategy has also been used by Braun (2023) in the setting of bipartite graphs.

Let $(\delta_j)_{j \in [n]}$ be independent Bernoulli r.v. with parameter $1/2$ and let us define the set of indices $\Lambda_\delta = \{j \in \mathcal{C}_{k'} : \delta_j = 1\}$ and the random variable

$$S_\delta = \sum_{j,j' \in \mathcal{C}_{k'}} \delta_j (1 - \delta_{j'}) \langle E_{j:}, E_{j':} \rangle = \sum_{j \in \Lambda_\delta} \langle E_{j:}, \sum_{j' \in \Lambda_\delta^c} E_{j':} \rangle.$$

Let us denote by $\mathbb{E}_{\Lambda^c}(.)$ the conditional expectation on

δ and $(E_{j':})_{j'\in\Lambda_\delta^c}$. For all $t>0$, we have

$$\begin{aligned}
\log \mathbb{E}_{\Lambda^c}\left(e^{tS_\delta}\right) &= \sum_{i,j\in\Lambda_\delta} \log(\mathbb{E}_{\Lambda_\delta^c}(e^{E_{ij}t\sum_{j'\in\Lambda_\delta^c}E_{ij'}}) \\
&= \sum_{i,j\in\Lambda_\delta}\left(\log(\mathbb{E}e^{A_{ij}t\sum_{j'\in\Lambda_\delta^c}E_{ij'}}) - p_{ij}t\sum_{j'\in\Lambda_\delta^c}E_{ij'}\right) \\
&\le \sum_{i,j\in\Lambda_\delta}\left(p_{ij}(e^{t\sum_{j'\in\Lambda_\delta^c}E_{ij'}}-1) - p_{ij}t\sum_{j'\in\Lambda_\delta^c}E_{ij'}\right) \\
&\qquad (\log(1+x)\ge x, \text{ for all } x>-1) \\
&\le e^{t\max_i\sum_{j'\in\Lambda_\delta^c}E_{ij'}}0.5t^2p_{max}\sum_{i\in\mathcal{C}_k,j\in\mathcal{C}_{k'}}(\sum_{j'\in\Lambda_\delta^c}E_{ij'})^2. \\
&\qquad (\text{by Taylor-Lagrange formula})
\end{aligned}$$

Let $C_1>1$ be an appropriately large constant and let us denote the events

$$\begin{aligned}
\mathcal{E}(\Lambda_\delta^c) = \{&\sum_{i\in\mathcal{C}_k,j\in\mathcal{C}_{k'}}(\sum_{j'\in\Lambda_\delta^c}E_{ij'})^2 \le C_1(n/K)^3p_{max}\} \\
&\cap\{\max_i\sum_{j'\in\Lambda_\delta^c}E_{ij'}\le C_1np_{max}/K\}
\end{aligned}$$

and

$$\mathcal{E} = \{\max_{\Lambda_\delta}\sum_{i\in\mathcal{C}_k,j\in\mathcal{C}_{k'}}(\sum_{j'\in\Lambda_\delta^c}E_{ij'})^2 \le C_1(n/K)^3p_{max}\}\cap\mathcal{D}$$

where $\mathcal{D}=\{\max_i\sum_j A_{ij}\le C_1np_{max}/K\}$. By Bernstein inequality (cf. appendix Section C) $\mathcal{E}$ occurs with probability at least $1-n^{-5}$ for C_1 large enough. By choosing $t=(C_1np_{max}/K)^{-1}$, we obtain for $C_2>1$ large enough

$$\begin{aligned}
&\mathbb{P}\left(S_\delta \ge C_2(n/K)^2p_{max}\cap\mathcal{E}\right) \\
&\le \mathbb{E}\left(\mathbf{1}_{\mathcal{E}(\Lambda_\delta^c)}\mathbb{P}(S_\delta\ge C_2(n/K)^2p_{max}|\Lambda^c)\right) \\
&\le \mathbb{E}\left(\mathbf{1}_{\mathcal{E}(\Lambda_\delta^c)}\mathbb{P}(e^{tS_\delta}\ge e^{C_2t(n/K)^2p_{max}}|\Lambda^c)\right) \\
&\le e^{-tC_2(n/K)^2p_{max}}\mathbb{E}\left(\mathbf{1}_{\mathcal{E}(\Lambda_\delta^c)}\mathbb{E}_{\Lambda^c}\left(e^{tS_\delta}\right)\right) \\
&\lesssim e^{-C_2C_1^{-1}(n/K)}e^{2n/(C_1K)} \le e^{-5n/K}.
\end{aligned}$$

By an union bound argument,

$$\mathbb{P}(\underbrace{\exists\delta, S_\delta\gtrsim(n/K)^2p_{max}\cap\mathcal{E}}_{\mathcal{E}_1}) \le 2^{n/K}e^{-5n/K}\le e^{-4n/K}.$$

Hence, $\mathbb{P}(\exists\delta, S_\delta\gtrsim(n/K)^2p_{max}) \le e^{-4n/K}+n^{-5}$. Since

$$\sum_{j\neq j'}\langle E_{:j},E_{:j'}\rangle_{\mathcal{C}_k} = 4\mathbb{E}_\delta(S_\delta),$$

we obtain that $S\lesssim(n/K)^2p_{max}$ with probability at least $1-n^{-5}$.

Case $k=k'=k''$. See the appendix, Section C. □

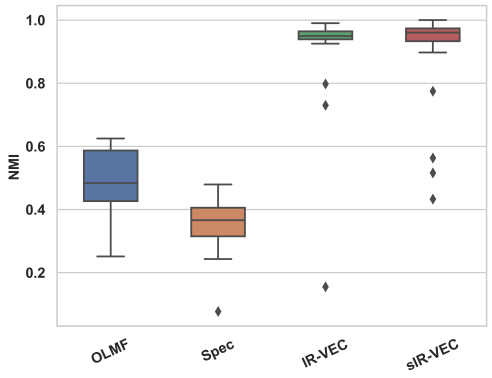

Figure 1: Average performance over 20 runs under Scenario 1.

4 Numerical experiments

In this section, we evaluate our proposed algorithms, `sIR-VEC` and `IR-VEC` (with $T=3$), on synthetic data and the email EU core dataset (Leskovec et al., 2007) with synthetic edge covariates[2].

We compare our methods with `OLMF` (Paul and Chen, 2020), a general matrix factorization approach applicable beyond the multi-layer graph setting, and the vanilla spectral method `Spec` which doesn't incorporate edge side information. We assess the accuracy of clustering using the Normalized Mutual Information (NMI) criterion, where an NMI of zero indicates no significant correlation between the clusters, and an NMI of one signifies a perfect match.

4.1 Network with indistinguishable communities (Scenario 1)

We consider a VEC-SBM with $K=3$, $n=600$, and such that the graph is generated by an SBM with parameters $p=3.5\log n$ and $q=\log n$ where the communities 1 and 2 are indistinguishable. The covariates are such that $\mu_{11}=(1,1,1)$ and $\mu_{22}=-\mu_{11}$ and all the other centroids are zero. Thus, the covariates only separate communities 1 and 2. As shown in Figure 1, `IR-VEC` and `sIR-VEC` outperform `OLMF` and effectively combine both sources of information to recover the clusters. However, we observed that `IR-VEC` is more sensitive to initialization than `sIR-VEC`. This is why we initialized it with `sIR-VEC`.

4.2 Non-isotropic covariance (Scenario 2)

In this scenario, we sample a VEC-SBM with the same parameters as the previous experiment, except for the

[2]The experiments were conducted using R on a CPU Intel Core i7-1255U. The code implementation can be found on `https://github.com/glmbraun/VECSBM/`.

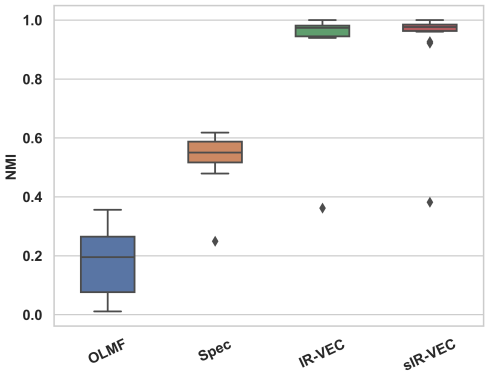

Figure 2: Average performance over 20 runs under Scenario 2.

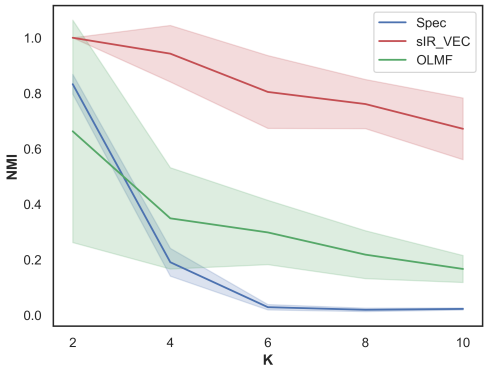

Figure 3: Average performance over 20 runs with varying K (Scenario 3).

edge covariates. Here, $\mu_{kk'}$ is generated uniformly over $[-1, 1]$, and $\Sigma_{kk'}$ are positive definite matrices randomly generated using the `clusterGeneration` package, with the maximal singular value set to 1. As shown in Figure 2, the performance of `OLMF` significantly decreases under this scenario, while `IR-VEC` and surprisingly `sIR-VEC` recover accurately the clusters.

4.3 Influence of the number of communities (Scenario 3)

We evaluate the performance of our method as the number of communities increases. We fix the parameters: $n = 1000$, $p = 8 \log n/n$, $q = p/2$, and generate isotropic edge covariates with centroids sampled uniformly over $[-2, 2]$ for $K \in \{2, 4, 6, 8, 10\}$. As shown in Figure 3, while the spectral method's performance decreases with increasing K, `sIR-VEC` is less sensitive. This is because the edge distribution is dissymmetric, allowing the SNR to remain higher when K increases. Additionally, we observe that `sIR-VEC` performs well when initialized with an almost uninformative $z^{(0)}$ provided by `Spec`.

4.4 Email EU core dataset

The dataset (Leskovec et al., 2007) comprises email communications between members of different European research institutions. We restrict the dataset to six institutions with at least 50 members and consider the institution as the ground-truth partition. Isolated vertices are removed, and for each edge, we simulate a textual distribution across 6 topics depending on the communities of its endpoints. The proportion of topics for each $k, k' \in [K]$ is generated uniformly over $[0, 1]$. We obtained an NMI of 0.49 with the spectral method, while `IR-VEC` is more accurate and provides a clustering with an NMI of 0.81 after 15 iterations. The number of iterations required is higher than when the graph is generated from an SBM, but `IR-VEC` appears robust to variations in graph topology.

5 Conclusion and perspectives

We have quantified the added value of edge-side information within the VEC-SBM framework. Our findings reveal that incorporating edge covariates can significantly improve the SNR, particularly when communities exhibit similar connectivity profiles or when dealing with a large number of communities. Furthermore, we have introduced an efficient iterative algorithm, `sIR-VEC`, which has been proven to achieve the optimal misclustering rate.

However, our work leaves several questions open for future research. These include the challenging task of estimating the number of communities in the presence of covariates, as well as the analysis of random initialization techniques for improved community detection. Furthermore, a promising research direction is to extend this framework to more complex and realistic models where the covariates are high-dimensional, and the network possesses intricate structures beyond the scope of the traditional SBM.

Acknowledgements

G.B. would like to express gratitude to the anonymous reviewers for their constructive feedback, contributing to the overall clarity and quality of the paper. Special thanks to Okan Koc for his valuable assistance concerning the Python implementation. M.S. was supported by JST CREST Grant Number JPMJCR18A2 and a grant from Apple, Inc. Any views, opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and should not be interpreted as reflecting the views, policies or position, either expressed or implied, of Apple Inc.

Bibliography

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

E. Abbe, J. Fan, and K. Wang. An ℓ_p theory of PCA and spectral clustering. *The Annals of Statistics*, 50 (4):2359 – 2385, 2022.

C. Aicher, A. Z. Jacobs, and A. Clauset. Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2):221–248, 2014.

R. Boutin, C. Bouveyron, and P. Latouche. Embedded topics in the stochastic block model. *Statistics and Computing*, 33(5):95, 2023a.

R. Boutin, P. Latouche, and C. Bouveyron. The deep latent position topic model for clustering and representation of networks with textual edges, 2023b.

C. Bouveyron, P. Latouche, and R. Zreik. The stochastic topic block model for the clustering of vertices in networks with textual edges. *Statistics and Computing*, 28(1):11–31, Oct 2016. ISSN 1573-1375.

G. Braun. Strong consistency guarantees for clustering high-dimensional bipartite graphs with the spectral method, 2023.

G. Braun, H. Tyagi, and C. Biernacki. An iterative clustering algorithm for the contextual stochastic block model with optimality guarantees. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 2257–2291, 2022.

A. Cerqueira and E. Levina. A pseudo-likelihood approach to community detection in weighted networks, 2023.

X. Chen and A. Zhang. Optimal clustering in anisotropic gaussian mixture models. *ArXiv*, 2021.

Y. Deshpande, S. Sen, A. Montanari, and E. Mossel. Contextual stochastic block models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

C. Gao and A. Y. Zhang. Iterative algorithm for discrete structure recovery. *The Annals of Statistics*, 50(2):1066 – 1094, 2022.

C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153 – 2185, 2018.

R. Han, Y. Luo, M. Wang, and A. R. Zhang. Exact Clustering in Tensor Block Model: Statistical Optimality and Computational Limit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1666–1698, 2022.

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

B.-Y. Jing, T. Li, and Z. Lyu. Community detection on mixture multilayer networks via regularized tensor decomposition. *The Annals of Statistics*, 49, 12 2021.

J.-B. Léger. Blockmodels: A r-package for estimating in latent block model and stochastic block model, with various probability functions, with or without covariates. *arXiv: Computation*, 2016.

J. Lei and A. Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1), Feb 2015.

J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

G. Liang, G. Zhang, S. Fattahi, and R. Y. Zhang. Simple alternating minimization provably solves complete dictionary learning, 2022.

Y. Lu and H. H. Zhou. Statistical and computational guarantees of lloyd's algorithm and its variants. *ArXiv*, 2016.

C. S. Mukherjee, P. Peng, and J. Zhang. Recovering unbalanced communities in the stochastic block model with application to clustering with a faulty oracle. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

S. Paul and Y. Chen. Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *Ann. Statist.*, 48(1):230–250, 2020.

M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(none):1 – 9, 2013.

N. Shi, R. A. Kontar, and S. Fattahi. Heterogeneous matrix factorization: When features differ by datasets, 2023.

D. Stöger and M. Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. In *Neural Information Processing Systems*, 2021.

K. K. Thekumparampil, P. Netrapalli, P. Jain, and S. Oh, editors. *Sample Efficient Linear Meta-Learning*, 2021.

T. Vallès-Català, F. A. Massucci, R. Guimerà, and M. Sales-Pardo. Multilayer stochastic block models reveal the multilayer structure of complex networks. *Phys. Rev. X*, 6:011036, Mar 2016.

M. Xu, V. Jog, and P.-L. Loh. Optimal rates for community estimation in the weighted stochastic block model. *The Annals of Statistics*, 2017.

S.-Y. Yun and A. Proutiere. Optimal cluster recovery in the labeled stochastic block model. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

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

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material

The proof of Theorem 1 is presented in Section A. More precisely, Section A.1 gives the exact expression of the error decomposition discussed in Section 3.1. Section A.1.1 shows that the F-error term satisfies Condition C1 and Section A.1.2 shows that the G-error term satisfies Condition C2. The proof of the remaining case of Lemma 3 is given in Section C.

A Proof of Theorem 1

We will use the following notations. Let us define the Hamming loss h as

$$h(z_i, z_i') = \sum_i \mathbf{1}_{z_i \neq z_i'}$$

and recall that l was defined as the weighted Hamming loss

$$l(z, z') = \sum_i \Delta^2(z_i, z_i') \mathbf{1}_{z_i \neq z_i'}.$$

We will denote by $h^{(t)}$ the corresponding function applied to z and $z^{(t)}$.

A.1 Error decomposition

Let us define for all $k, k' \in [K]$ the oracle estimators

$$\tilde{p} = \sum_{k \in [K]} \frac{\sum_{i,j \in \mathcal{C}_k} A_{ij}}{K|\mathcal{C}_k|^2}, \tilde{q} = \sum_{k \neq k' \in [K]} \frac{\sum_{i \in \mathcal{C}_k, j \in \mathcal{C}_{k'}} A_{ij}}{K(K-1)|\mathcal{C}_k||\mathcal{C}_k'|}, \tilde{\mu}_{kk'} = \frac{\sum_{i \in \mathcal{C}_k, j \in \mathcal{C}_{k'}} A_{ij} G_{ij}}{|\mathcal{C}_k||\mathcal{C}_k'|}.$$

The node i such that $z(i) = a$ is incorrectly classified at step t iff

$$a \neq \arg\max_k MAP_i(\mathcal{C}^{(t)}, \Pi^{(t)}, \mu^{(t)}, I_d).$$

It implies that there exists a $b \neq a \in [K]$ such that

$$\sum_{l \in [K]} \sum_{j \in \mathcal{C}_{l,-i}^{(t)}} A_{ij} \left(\log(\Pi_{al}^{(t)}) - \log(\Pi_{bl}^{(t)}) - \left\| G_{ij} - \mu_{al}^{(t)} \right\|^2 + \left\| G_{ij} - \mu_{bl}^{(t)} \right\|^2 \right) < 0.$$

This last condition can be further decomposed as

$$\underbrace{\log(\frac{p}{q}) \left(\sum_{j \in \mathcal{C}_a} E_{ij} - \sum_{j \in \mathcal{C}_b} E_{ij} \right) + \sum_{l \in [K]} \sum_{j \in \mathcal{C}_{l,-i}} \left(E_{ij} \left\| \mu_{al} - \mu_{bl} \right\|^2 + 2A_{ij} \langle \epsilon_{ij}, \mu_{al} - \mu_{bl} \rangle \right)}_{C_i(a,b)} < -\Delta^2(a,b) + F_i^{(t)} + G_i^{(t)}$$

where

$$\begin{aligned}
\Delta^2(a,b) &= \log(p/q)(|\mathcal{C}_a| p - |\mathcal{C}_b| q) + \sum_{l \in [K]} |\mathcal{C}_l| \Pi_{al} \left\| \mu_{al} - \mu_{bl} \right\|^2, \\
F_i^{(t)} &= \langle E_{i:}(Z^{(t)} - Z), \log \Pi_{a:} - \log \Pi_{b:} \rangle + \langle E_{i:} Z^{(t)}, \log \Pi_{a:}^{(t)} - \log \tilde{\Pi}_{a:} - \log \Pi_{b:}^{(t)} + \log \tilde{\Pi}_{b:} \rangle \\
&\quad + \sum_{l \in [K]} \sum_{j \in \mathcal{C}_l} \left(2A_{ij} \langle \epsilon_{ij}, \mu_{al}^{(t)} - \tilde{\mu}_{al} + \tilde{\mu}_{bl} - \mu_{bl}^{(t)} \rangle + E_{ij} \left\| \tilde{\mu}_{bl} - \mu_{bl}^{(t)} \right\|^2 - 2E_{ij} \langle \mu_{al} - \mu_{bl}, \tilde{\mu}_{bl} - \mu_{bl}^{(t)} \rangle \right), \\
G_i^{(t)} &= \langle P_{i:}(Z^{(t)} - Z), \log \Pi_{a:} - \log \Pi_{b:} \rangle + \langle P_{i:} Z^{(t)}, \log \Pi_{a:}^{(t)} - \log \Pi_{a:} + \log \Pi_{b:}^{(t)} + \log \Pi_{b:} \rangle \\
&\quad + \sum_{l \in [K]} \sum_{j \in \mathcal{C}_l} P_{ij} \left(\left\| \mu_{bl} - \mu_{bl}^{(t)} \right\|^2 - 2P_{ij} \langle \mu_{al} - \mu_{bl}, \mu_{bl} - \mu_{bl}^{(t)} \rangle \right)
\end{aligned}$$

$$+ \langle E_{i:} Z^{(t)}, \log \tilde{\Pi}_{a:} - \log \Pi_{a:} + \log \Pi_{b:} - \log \tilde{\Pi}_{b:} \rangle - 2 \sum_{l \in [K]} \sum_{j \in \mathcal{C}_l} E_{ij} \langle \mu_{al} - \mu_{bl}, \mu_{bl} - \tilde{\mu}_{bl} \rangle$$
$$+ \sum_{l \in [K]} \sum_{j \in \mathcal{C}_l} E_{ij} \left(\left\| \mu_{bl} - \mu_{bl}^{(t)} \right\|^2 - \left\| \tilde{\mu}_{bl} - \mu_{bl}^{(t)} \right\|^2 \right) + 2A_{ij} \langle \epsilon_{ij}, \tilde{\mu}_{al} - \mu_{al} - \tilde{\mu}_{bl} + \mu_{bl} \rangle.$$

The term $\Delta^2(a,b)$ is deterministic and corresponds to the signal. Under Assumptions A1 and A3, it is easy to see that

$$\Delta^2(a,b) \asymp \frac{n(p-q)}{K} + \frac{np}{K} \sum_l \|\mu_{al} - \mu_{bl}\|^2 .$$

So if the difference between the centroids $\sum_l \|\mu_{al} - \mu_{bl}\|^2 = \Omega(1)$, edges covariate have a multiplicative effect on the signal. The error terms $(F_i(t))_{i \in [n]}$ depend linearly on $\epsilon_{i:}$ and $E_{i:}$ and these errors can be controlled in average. On the other hand, the error terms $(G_i(t))_{i \in [n]}$ need to be controlled uniformly.

A.1.1 F-error term

We need an upper-bound of

$$F = \max_{\{z^{(t)} : l(z, z^{(t)}) \leq \tau\}} \sum_{i=1}^n \max_{b \in [K] \setminus z_i} \frac{(F_{ib}^{(t)})^2}{\Delta^2(z_i, b) l(z, z^{(t)})}.$$

Toward this end, notice that $(F_{ib}^{(t)})^2 \lesssim F_1 + F_2 + F_3 + F_4$ where

$$F_1 = \sum_{i=1}^n \max_{b \in [K] \setminus z_i} \langle E_{i:}(Z^{(t)} - Z), \log \Pi_{a:} - \log \Pi_{b:} \rangle^2$$
$$F_2 = \sum_{i=1}^n \max_{b \in [K] \setminus z_i} \langle E_{i:} Z^{(t)}, \log \Pi_{a:}^{(t)} - \log \tilde{\Pi}_{a:} - \log \Pi_{b:}^{(t)} + \log \tilde{\Pi}_{b:} \rangle^2$$
$$F_3 = \sum_{i=1}^n \max_{b \in [K] \setminus z_i} \left(\sum_{l \in [K]} \sum_{j \in \mathcal{C}_l} A_{ij} \langle \epsilon_{ij}, \mu_{al}^{(t)} - \tilde{\mu}_{al} + \tilde{\mu}_{bl} - \mu_{bl}^{(t)} \rangle \right)^2$$
$$F_4 = \sum_{i=1}^n \max_{b \in [K] \setminus z_i} \left(\sum_{l \in [K]} \sum_{j \in \mathcal{C}_l} E_{ij} \left(\left\| \tilde{\mu}_{bl} - \mu_{bl}^{(t)} \right\|^2 - 2 \langle \mu_{al} - \mu_{bl}, \tilde{\mu}_{bl} - \mu_{bl}^{(t)} \rangle \right) \right)^2 .$$

First, let us establish some useful inequalities that will be used repeatedly.

Lemma 4. *Under the assumption of Theorem 1 we have with probability at least $1 - n^{-\Omega(1)}$, for all $z^{(t)}$ such that $Kl(z^{(t)}, z) \leq n\Delta_{min}^2 \epsilon$*

1. $\max_{k \in [K]} |n_k^{(t)} - n_k| \leq \frac{l(z^{(t)}, z)}{\Delta_{min}^2}$
2. $\max_{k \in [K]} ||Z_{:k}^{(t)} - Z_{:k}|| \leq \left\| Z^{(t)} - Z \right\| \lesssim \frac{K^{0.5}}{n^{0.5} \Delta_{min}^2} l(z^{(t)}, z)$,
3. $|\log(\frac{p^{(t)}}{q^{(t)}}) - \log(\frac{\tilde{p}}{\tilde{q}})| \lesssim K \frac{l(z^{(t)}, z)}{n\Delta_{min}^2}$
4. $\max_{b,l \in [K]} \left\| \tilde{\mu}_{bl} - \mu_{bl}^{(t)} \right\| \lesssim K^{1.5} \left(\sqrt{\frac{l(z^{(t)}, z)}{n\Delta_{min}^2}} \vee \frac{\sqrt{\log K}}{np_{max}} \right)$,
5. $\max_{b,l \in [K]} \|\tilde{\mu}_{bl} - \mu_{bl}\| \lesssim \frac{K}{\sqrt{n^2 p_{max}}}$.

Proof. The first three items are direct consequences of Lemma 13, 14, and 15 in Braun et al. (2022). However, the fourth point doesn't derive immediately from Lemma 4.1 in Gao and Zhang (2022) since in our setting the estimate of μ_{bl} depends on A. We will need several properties of the adjacency matrix A that holds w.h.p. We recall them below.

Fact 1. The adjacency matrix A satisfies the following version of the discrepancy property (see e.g. Lei and Rinaldo (2015))

$$\forall T_1, T_2 \subset [n] \text{ such that } |T_1| \leq |T_2|,\ T(A) \leq \kappa(|T_1|, |T_2|) p_{max} |T_1||T_2|$$

where $T(A) = \sum_{(i,j)\in T_1\times T_2} A_{ij}$ and $\kappa(|T_1|, |T_2|) = \max(\frac{c^* \log(en/|T_2|)}{p_{max}|T_1|}, C^*)$ for constants $c^*, C^* > 0$.The argument is based on Chernoff's bound and a union bound, see the supplementary material of Lei and Rinaldo (2015) for more details.

Fact 2. Let us denote by $I_A(a,b) = \sum_{i\in\mathcal{C}_a, j\in\mathcal{C}_b} A_{ij}$ the number of observed edges between communities a and $b \in [K]$. By using Chernoff's bound and a union bound over $(a,b) \in [K]^2$ we can show that w.h.p for all $a, b \in [K]$

$$I(a,b) \asymp p_{max} \frac{n^2}{K^2}.$$

We will assume from now on that A satisfies the previous properties. Hence, conditionally on A, a straightforward adaptation of Lemma A.1 in Lu and Zhou (2016) implies that w.h.p.

$$\forall T = T_1 \times T_2 \subset [n] \times [n],\ \left\| \sum_{(i,j)\in T} A_{ij}\epsilon_{ij} \right\| \lesssim \sqrt{|T(A)|n}. \tag{A.1}$$

The price to pay to have a uniform bound is an additional $\sqrt{n}$ factor.

Consider the set $T = (\mathcal{C}_a \times \mathcal{C}_b)\Delta(\mathcal{C}_a^{(t)} \times \mathcal{C}_b^{(t)})$ where Δ denotes the symmetric difference operator. It can be further decomposed as $\cup_{c=1}^3 T^{(c)}$ where $T^{(1)} = \mathcal{C}_a \setminus C_a^{(t)} \times \mathcal{C}_b$, $T^{(2)} = (\mathcal{C}_a \cap \mathcal{C}_a^{(t)}) \times (\mathcal{C}_b^{(t)} \Delta \mathcal{C}_b)$, and $T^{(3)} = \mathcal{C}_a^{(t)} \setminus \mathcal{C}_a \times \mathcal{C}_b^{(t)}$. Note that by definition of $h^{(t)}$ we have

$$\begin{aligned}
\max_b |\mathcal{C}_b^{(t)} \Delta \mathcal{C}_b| &\leq h^{(t)} \\
\max_a |\mathcal{C}_a \setminus C_a^{(t)}| &\leq h^{(t)} \\
\max_a |\mathcal{C}_a^{(t)} \setminus C_a| &\leq h^{(t)}.
\end{aligned}$$

Notice that by definition of $h^{(0)}$, we have $|\mathcal{C}_a \cap C_a^{(t)}| \asymp n/K$ and $|\mathcal{C}_b^{(t)}| \asymp n/K$. By Fact 1 and 2, we have for all $c = 1 \dots 3$

$$|T^{(c)}(A)| \lesssim p_{max} \frac{nh^{(t)}}{K} \vee \log(K) \frac{n}{K}. \tag{A.2}$$

Let us denote by $I_A^{(t)}(a,b) = \sum_{i\in\mathcal{C}_a^{(t)}, j\in\mathcal{C}_b^{(t)}} A_{ij}$ the number of edges between nodes in the estimated communities a and b. Since $h^{(t)} \leq h^{(0)} \leq \epsilon n/K^2$ we have,

$$\left| I_A^{(t)}(a,b) - I_A(a,b) \right| \lesssim \epsilon I_A(a,b)/K.$$

This implies that $I_A^{(t)}(a,b) \asymp p_{max} n^2/K^2$. Now, we can decompose

$$\begin{aligned}
\tilde{\mu}_{bl} - \mu_{bl}^{(t)} &= \frac{\sum_{i\in\mathcal{C}_a, j\in\mathcal{C}_b} A_{ij}G_{ij} - \sum_{i\in\mathcal{C}_a^{(t)}, j\in\mathcal{C}_b^{(t)}} A_{ij}G_{ij}}{I_A(a,b)} + \left(\frac{1}{I_A(a,b)} - \frac{1}{I_A^{(t)}(a,b)} \right) \sum_{i\in\mathcal{C}_a^{(t)}, j\in\mathcal{C}_b^{(t)}} A_{ij}G_{ij} \\
&= \underbrace{\frac{\sum_{i\in\mathcal{C}_a, j\in\mathcal{C}_b} A_{ij}\epsilon_{ij} - \sum_{i\in\mathcal{C}_a^{(t)}, j\in\mathcal{C}_b^{(t)}} A_{ij}\epsilon_{ij}}{I_A(a,b)}}_{L_1} + \underbrace{\left(\frac{1}{I_A(a,b)} - \frac{1}{I_A^{(t)}(a,b)} \right) \sum_{i\in\mathcal{C}_a^{(t)}, j\in\mathcal{C}_b^{(t)}} A_{ij}\epsilon_{ij}}_{L_2} \\
&+ \underbrace{\frac{\sum_{i\in\mathcal{C}_a, j\in\mathcal{C}_b} A_{ij}\mu_{ab} - \sum_{i\in\mathcal{C}_a^{(t)}, j\in\mathcal{C}_b^{(t)}} A_{ij}\mu_{z(i)z(j)}}{I_A(a,b)}}_{L_3}
\end{aligned}$$

$$+ \underbrace{\left(\frac{1}{I_A(a,b)} - \frac{1}{I_A^{(t)}(a,b)}\right) \sum_{i\in\mathcal{C}_a^{(t)}, j\in\mathcal{C}_b^{(t)}} A_{ij}\mu_{z(i)z(j)}}_{L_4}.$$

Control of L_1. We have

$$\begin{aligned}\|L_1\| &= \left\| \frac{\sum_{c=1}^3 \sum_{(i,j)\in T^{(c)}} A_{ij}\epsilon_{ij}}{I_A(a,b)} \right\| \\ &\lesssim K^{1.5}\left(\sqrt{\frac{h^{(t)}}{n}} \vee \frac{\sqrt{\log K}}{np_{max}}\right) \leq K^{1.5}\left(\sqrt{\frac{l(z^{(t)},z)}{n\Delta_{min}^2}} \vee \frac{\sqrt{\log K}}{np_{max}}\right). \qquad \text{(by equations (A.2) and (A.1))}\end{aligned}$$

Control of L_3. By a similar argument we obtain

$$\begin{aligned}\|L_3\| &\lesssim \frac{\sum_c \sum_{(i,j)\in T^{(c)}} A_{ij}}{I_A(a,b)} \\ &\lesssim K\frac{h^{(t)}}{n} \vee K\frac{\log K}{np_{max}}.\end{aligned}$$

Control of L_2. By equation (A.1) and the discrepancy property, we obtain

$$\left\| \sum_{i\in\mathcal{C}_a^{(t)}, j\in\mathcal{C}_b^{(t)}} A_{ij}\epsilon_{ij} \right\| \lesssim \sqrt{n}\sqrt{I(a,b)}.$$

Furthermore, we have

$$\left| \frac{1}{I_A(a,b)} - \frac{1}{I_A^{(t)}(a,b)} \right| \lesssim \frac{|I_A(a,b) - I_A^t(a,b)|}{I(a,b)^2} \lesssim \frac{K^3 h^{(t)}}{n^3 p_{max}} \vee \frac{K^3 \log K}{n^3 p_{max}^2}.$$

By consequence

$$\|L_3\| \lesssim \frac{1}{\sqrt{I(a,b)}}\frac{Kh^{(t)}}{n} \vee \frac{\log K}{\sqrt{I(a,b)}np_{max}}$$

Control of L_4. It can be handled in the same way as L_2.

We can conclude by summing all the error terms. □

Control of F_1. This term can be controlled with a similar argument as in Braun et al. (2022). We have

$$\begin{aligned}\frac{F_1}{\Delta^2(z_i,b)l(z^{(t)},z)} &\leq \sum_i \left\| E_{i:}(Z^{(t)} - Z) \right\|^2 \frac{1}{\Delta_{min}^2 l(z^{(t)},z)} \\ &\leq \left\| E_{i:}(Z^{(t)} - Z) \right\|_F^2 \frac{1}{\Delta_{min}^2 l(z^{(t)},z)} \\ &\lesssim K\|E\|^2 \frac{K}{n\Delta_{min}^6} l(z^{(t)},z) \qquad \text{(by Lemma 4)} \\ &\lesssim K\frac{np_{max}}{\Delta_{min}^4}\frac{Kl(z^{(t)},z)}{n\Delta_{min}^2} \to 0\end{aligned}$$

Control of F_2. This term can be handled again by the same techniques as in Braun et al. (2022). We have by triangular inequality

$$\begin{aligned}\frac{F_2}{\Delta^2(z_i,b)l(z^{(t)},z)} &\leq 4\sum_i \left\| E_{i:}Z^{(t)} \right\|^2 \max_k \left\| \log \Pi_{k:}^{(t)} - \log \tilde{\Pi}_{k:} \right\|^2 \frac{1}{\Delta_{min}^2 l(z^{(t)},z)} \\ &\lesssim K^2 \frac{n^2 p_{max} l(z^{(t)},z)}{n^2\Delta_{min}^6} \to 0. \qquad \text{(by Lemma 4)}\end{aligned}$$

Control of F_3 while $\sqrt{\frac{h^{(t)}}{n}} > \frac{\sqrt{logK}}{np_{max}}$. To control F_3 we will use the following lemma.

Lemma 5. *Under the assumption of Theorem 1 we have with probability at least $1 - n^{-\Omega(1)}$*

1. $\max_{a,b\in[K]} \left\|\sum_{i\in\mathcal{C}_a, j\in\mathcal{C}_b} A_{ij}\epsilon_{ij}\epsilon_{ij}^\top\right\| \lesssim (n/K)^2 p_{max}$,
2. $\max_{l,l',a\in[K]} \left\|\sum_{i\in\mathcal{C}_a, j\in\mathcal{C}_l, j'\in\mathcal{C}_{l'}} A_{ij}A_{ij'}\epsilon_{ij}\epsilon_{ij'}^\top\right\| \lesssim (n/K)^2 p_{max}$.

Proof. The first point can be obtained by conditioning on A and using the Lemma A.2 in Lu and Zhou (2016). Since $\sum_{i\in\mathcal{C}_a, j\in\mathcal{C}_b} A_{ij} \lesssim (n/K)^2 p_{max}$ w.h.p. we obtain the stated result. The second result can be obtained by a similar argument and by noticing that by Lemma 3 w.h.p

$$\max_{a,l,l'} \sum_{i\in\mathcal{C}_a, j\in\mathcal{C}_l, j'\in\mathcal{C}_{l'}} A_{ij}A_{ij'} \lesssim (n/K)^2 p_{max}$$

□

By developing the square in F_3 we obtain

$$\begin{aligned}
\frac{F_3}{\Delta^2(z_i,b)l(z^{(t)},z)} &\lesssim \frac{K^4}{\Delta_{min}^2 l(z^{(t)},z)} \left(\left\|\sum_{i,j} A_{ij}\epsilon_{ij}\epsilon_{ij}^\top\right\| + \max_{l,l',a\in[K]} \left\|\sum_{i\in\mathcal{C}_a, j\in\mathcal{C}_l, j'\in\mathcal{C}_{l'}} A_{ij}A_{ij'}\epsilon_{ij}\epsilon_{ij'}^\top\right\|\right) \\
&\quad \times \max_{b\neq a,l} \left\|\mu_{al}^{(t)} - \tilde{\mu}_{al} + \tilde{\mu}_{bl} - \mu_{bl}^{(t)}\right\|^2 \\
&\lesssim K^6 \frac{np_{max}}{\Delta_{min}^4} \to 0. \qquad \text{(by Lemma 4)}
\end{aligned}$$

Control of F_4 while $\sqrt{\frac{h^{(t)}}{n}} > \frac{\sqrt{logK}}{np_{max}}$. Let us write $c_{abl} = \left\|\tilde{\mu}_{bl} - \mu_{bl}^{(t)}\right\|^2 - 2\langle \mu_{al} - \mu_{bl}, \tilde{\mu}_{bl} - \mu_{bl}^{(t)}\rangle$. We have $\max_{a,b,l} |c_{abl}| \lesssim K^2\sqrt{\frac{l(z^{(t)},z)}{n\Delta_{min}^2}}$ By developing the square in F_4 we obtain

$$\begin{aligned}
\frac{F_4}{\Delta^2(z_i,b)l(z^{(t)},z)} &\lesssim \sum_i \sum_{b\neq z(i),l,l'} \sum_{j\in\mathcal{C}_l, j'\in\mathcal{C}_{l'}} E_{ij}E_{ij'}c_{z(i)bl}c_{z(i)bl'} \frac{1}{\Delta_{min}^2 l(z^{(t)},z)} \\
&\lesssim \sum_{b\neq z(i),l,l',l''} c_{l''bl}c_{l''bl'} \sum_{j\in\mathcal{C}_l, j'\in\mathcal{C}_{l'}} \sum_{i\in\mathcal{C}_{l''}} E_{ij}E_{ij'} \\
&\lesssim K^3 \frac{n^2 p_{max}}{n\Delta_{min}^4} \to 0. \qquad \text{(by Lemma 3)}
\end{aligned}$$

Case where $\sqrt{\frac{h^{(t)}}{n}} < \frac{\sqrt{logK}}{np_{max}}$. In this case, one should consider F_3 and F_4 as G-error terms. More precisely, we should consider the following term appearing in the F-error decomposition

$$\sum_{l\in[K]} \sum_{j\in\mathcal{C}_l} \left(2A_{ij}\langle \epsilon_{ij}, \mu_{al}^{(t)} - \tilde{\mu}_{al} + \tilde{\mu}_{bl} - \mu_{bl}^{(t)}\rangle + E_{ij}\left\|\tilde{\mu}_{bl} - \mu_{bl}^{(t)}\right\|^2 - 2E_{ij}\langle \mu_{al} - \mu_{bl}, \tilde{\mu}_{bl} - \mu_{bl}^{(t)}\rangle\right)$$

as a G-error term. By using the fact that

$$\left\|\tilde{\mu}_{bl} - \mu_{bl}^{(t)}\right\| \lesssim \frac{K^{1.5}\log K}{np_{max}}$$

it is easy to show that all the terms are $o(1)$. See also the proof of the bound of G_7 and G_8 in the next subsection.

A.1.2 G-error term

We can upper bound $G_i^{(t)}$ by $G_1 + G_2 + G_3 + G_4 + G_5 + G_6 + G_7 + G_8$ where

$$
\begin{aligned}
G_1 &= \left| \langle P_{i:}(Z^{(t)} - Z), \log \Pi_{a:} - \log \Pi_{b:} \rangle \right| \\
G_2 &= \left| \langle P_{i:} Z^{(t)}, \log \Pi^{(t)}_{a:} - \log \Pi_{a:} + \log \Pi^{(t)}_{b:} + \log \Pi_{b:} \rangle \right| \\
G_3 &= \sum_{l \in [K]} \sum_{j \in \mathcal{C}_l} P_{ij} \left\| \mu_{bl} - \mu^{(t)}_{bl} \right\|^2 \\
G_4 &= \sum_{l \in [K]} \sum_{j \in \mathcal{C}_l} 2P_{ij} \left| \langle \mu_{al} - \mu_{bl}, \mu_{bl} - \mu^{(t)}_{bl} \rangle \right| \\
G_5 &= \left| \langle E_{i:} Z^{(t)}, \log \tilde{\Pi}_{a:} - \log \Pi_{a:} + \log \Pi_{b:} - \log \tilde{\Pi}_{b:} \rangle \right| \\
G_6 &= 2 \left| \sum_{l \in [K]} \sum_{j \in \mathcal{C}_l} E_{ij} \langle \mu_{al} - \mu_{bl}, \mu_{bl} - \tilde{\mu}_{bl} \rangle \right| \\
G_7 &= \left| \sum_{l \in [K]} \sum_{j \in \mathcal{C}_l} E_{ij} \left(\left\| \mu_{bl} - \mu^{(t)}_{bl} \right\|^2 - \left\| \tilde{\mu}_{bl} - \mu^{(t)}_{bl} \right\|^2 \right) \right| \\
G_8 &= 2 \left| \sum_{l \in [K]} \sum_{j \in \mathcal{C}_l} A_{ij} \langle \epsilon_{ij}, \tilde{\mu}_{al} - \mu_{al} - \tilde{\mu}_{bl} + \mu_{bl} \rangle \right|.
\end{aligned}
$$

Control of G_1. Since $\left\| P_{i:}(Z^{(t)} - Z) \right\| \leq \sqrt{K} p_{max} h^{(t)}$ and $\| \log \Pi_{a:} - \log \Pi_{b:} \| = O(1)$ we get

$$
\frac{G_1}{\Delta^2(z_i, b)} \lesssim \frac{\sqrt{K} p_{max} h^{(t)}}{\Delta^2_{min}} \lesssim \frac{K l(z^{(t)}, z)}{n \Delta^2_{min}} \frac{n p_{max}}{\Delta^2_{min} \sqrt{K}} \leq \delta.
$$

Control of G_2. We have $\left\| P_{i:} Z^{(t)} \right\| \lesssim (n/\sqrt{K}) p_{max}$ and by Lemma 4 $\left\| \log \Pi^{(t)}_{a:} - \log \Pi_{a:} \right\| \lesssim K \frac{l(z^{(t)}, z)}{n \Delta^2_{min}}$. By using the triangular inequality, we obtain

$$
\frac{G_2}{\Delta^2(z_i, b)} \lesssim K \frac{l(z^{(t)}, z)}{n \Delta^2_{min}} \frac{n p_{max}}{\Delta^2_{min} \sqrt{K}} \leq \delta.
$$

Control of G_3. By Lemma 4 $\max_{b, l \in [K]} \left\| \mu_{bl} - \mu^{(t)}_{bl} \right\| \lesssim K^2 \sqrt{\frac{l(z^{(t)}, z)}{n \Delta^2_{min}}} + \frac{K}{\sqrt{n}}$. By consequence,

$$
\frac{G_3}{\Delta^2(z_i, b)} \lesssim K^4 \frac{n p_{max}}{\Delta^2_{min}} \left(\frac{l(z^{(t)}, z)}{n \Delta^2_{min}} + \frac{1}{n} \right) \leq \delta.
$$

Control of G_4. The proof is similar to G_3.

Control of G_5. One has $\left\| \log \tilde{\Pi}_{a:} - \log \Pi_{a:} + \log \Pi_{b:} - \log \tilde{\Pi}_{b:} \right\| \lesssim \frac{1}{n^2 p_{max}}$ and $\max_i \left\| E_{i:} Z^{(t)} \right\| \lesssim \sqrt{n p_{max}}$.

Control of G_6. Let $V \in \mathbb{R}^k$ such that $V_l = \langle \mu_{al} - \mu_{bl}, \mu_{bl} - \tilde{\mu}_{bl} \rangle$. By the triangular inequality and Lemma 4, we have $\| V \|_\infty \lesssim \frac{K}{n \sqrt{p}}$. Hence

$$
\begin{aligned}
\frac{G_6}{\Delta^2(z_i, b)} &= \frac{2 \left| \langle E_{i:} Z, V \rangle \right|}{\Delta^2(z_i, b)} \\
&\lesssim \frac{\sqrt{n} p_{max} \| V \|_\infty}{\Delta^2_{min}} = o(1)
\end{aligned}
$$

Control of G_7. By Lemma 4, $\max_{b,l\in[K]} \left\|\mu_{bl} - \mu_{bl}^{(t)}\right\| \lesssim K^2\sqrt{\frac{l(z^{(t)},z)}{n\Delta_{min}^2}} + \frac{K}{\sqrt{n}}$ and $\max_{b,l\in[K]} \left\|\tilde{\mu}_{bl} - \mu_{bl}^{(t)}\right\| \lesssim K\sqrt{\frac{l(z^{(t)},z)}{n\Delta_{min}^2}}$. By consequence,

$$\frac{G_6}{\Delta^2(z_i,b)} \lesssim \frac{\sqrt{np_{max}}}{\Delta_{min}^2}\left(K^4\frac{l(z^{(t)},z)}{n\Delta_{min}^2} + K^2\frac{1}{n}\right) = o(1).$$

Control of G_8. Conditionally on $(A_{ij})_{j\in\mathcal{C}_l}$, $\sum_{j\in\mathcal{C}_l} A_{ij}\langle\epsilon_{ij}, \tilde{\mu}_{al} - \mu_{al} - \tilde{\mu}_{bl} + \mu_{bl}\rangle$ is a centered Gaussian r.v. with variance $\sigma_A^2 = \|\tilde{\mu}_{al} - \mu_{al} - \tilde{\mu}_{bl} + \mu_{bl}\|^2 \sum_{j\in\mathcal{C}_l} A_{ij} \lesssim \frac{K}{\sqrt{n}}\sum_{j\in\mathcal{C}_l} A_{ij}$ by Lemma 4. By consequence,

$$\mathbb{P}_{(A_{ij})_{j\in\mathcal{C}_l}}\left(\left|\sum_{j\in\mathcal{C}_l} A_{ij}\langle\epsilon_{ij}, \tilde{\mu}_{al} - \mu_{al} - \tilde{\mu}_{bl} + \mu_{bl}\rangle\right| \gtrsim \sqrt{\log n}\,\|\tilde{\mu}_{al} - \mu_{al} - \tilde{\mu}_{bl} + \mu_{bl}\|^2 \sum_{j\in\mathcal{C}_l} A_{ij}\right) \leq n^{-\Omega(1)}.$$

Let us denote the event

$$H = \left\{\max_{i,l}\sum_{j\in\mathcal{C}_l} A_{ij} \lesssim \frac{np_{max}}{K}\right\}.$$

This event holds with probability at least $1 - n^{-\Omega(1)}$. We have

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{j\in\mathcal{C}_l} A_{ij}\langle\epsilon_{ij}, \tilde{\mu}_{al} - \mu_{al} - \tilde{\mu}_{bl} + \mu_{bl}\rangle\right| \gtrsim \sqrt{\log n}\frac{np_{max}}{K}\,\|\tilde{\mu}_{al} - \mu_{al} - \tilde{\mu}_{bl} + \mu_{bl}\|^2\right)\\
&\leq \mathbb{P}_H\left(\left|\sum_{j\in\mathcal{C}_l} A_{ij}\langle\epsilon_{ij}, \tilde{\mu}_{al} - \mu_{al} - \tilde{\mu}_{bl} + \mu_{bl}\rangle\right| \gtrsim \sqrt{\log n}\frac{np_{max}}{K}\,\|\tilde{\mu}_{al} - \mu_{al} - \tilde{\mu}_{bl} + \mu_{bl}\|^2\right) + n^{-\Omega(1)}\\
&\lesssim \mathbb{P}_H\left(\mathbb{P}_{(A_{ij})_{j\in\mathcal{C}_l}}\left(\left|\sum_{j\in\mathcal{C}_l} A_{ij}\langle\epsilon_{ij}, \tilde{\mu}_{al} - \mu_{al} - \tilde{\mu}_{bl} + \mu_{bl}\rangle\right| \gtrsim \sqrt{\log n}\,\|\tilde{\mu}_{al} - \mu_{al} - \tilde{\mu}_{bl} + \mu_{bl}\|^2\sum_{j\in\mathcal{C}_l} A_{ij}\right)\right) + n^{-\Omega(1)}\\
&\leq n^{-\Omega(1)}.
\end{aligned}$$

Since $\|\tilde{\mu}_{al} - \mu_{al} - \tilde{\mu}_{bl} + \mu_{bl}\|^2 \lesssim \frac{K^2}{n}$ by Lemma 4, we obtain that

$$\frac{G_8}{\Delta^2(z_i,b)} \lesssim \frac{K^3\sqrt{n}np_{max}}{n\Delta_{min}^2} = o(1).$$

B Proof of Theorem 2

Choose two communities a and $b \in [K]$ such that $\Delta_{\min} = \Delta(a,b)$. For each $k \in [K]$, let T_k a subset of $\mathcal{C}_k$ with cardinality $\frac{3n}{4K}$. Define $T = \cup_{k=1}^K T_k$ and

$$\mathcal{Z} = \{\hat{z} : \hat{z}_i = z_i \text{ for all } i \in T\}.$$

By using the same argument as in the proof of Theorem 2 in Gao et al. (2018) we can reduce the problem to a two-hypothesis testing problem

$$\inf_{\hat{z}}\sup_{\theta\in\Theta}\mathbb{E}(r(\hat{z},z)) \geq \frac{1}{6|T^c|}\sum_{i\in T^c}\frac{1}{2K^2}\inf_{\hat{z}_i}\mathbb{P}_1(\hat{z}_i = 2) + \mathbb{P}_2(\hat{z}_i = 1) \tag{B.1}$$

where $\mathbb{P}_1$ (resp. $\mathbb{P}_2$) denotes the probability distribution of the data when $z_i = a$ (resp. $z_i = b$). By the Neyman Pearson Lemma, the likelihood ratio test achieves the infinimum of the right-hand side of (B.1). Hence we have

$$\inf_{\hat{z}_i}\mathbb{P}_1(\hat{z}_i = 2) + \mathbb{P}_2(\hat{z}_i = 1) = \mathbb{P}\left(\underbrace{\sum_{l\in[K],j\in\mathcal{C}_l} A_{ij}(\|\mu_{al} - \mu_{bl}\|^2 + 2\langle\epsilon_{ij}, \mu_{al} - \mu_{al}\rangle) \leq 0}_{\mathcal{O}}\right).$$

First, assume that $\Delta^2(a,b) \to \infty$. Conditionally on $(A_{ij})_j$, $\mathbb{P}_{(A_{ij})_j}(\mathcal{O}) = \mathbb{P}(X_A \leq -\frac{\sigma_A^2}{2}) = \mathbb{P}(X_A \geq \frac{\sigma_A^2}{2})$ where $X_A \sim \mathcal{N}(0, \sigma_A^2)$ and $\sigma_A^2 = \sum_{l,j} A_{ij} \left\|\mu_{al} - \mu_{bl}\right\|^2$. By using the fact that

$$\int_t^\infty e^{-x^2/2}\,\mathrm{d}x \geq \frac{1}{\sqrt{2\pi}} \frac{t}{t^2+1} e^{-t^2/2},$$

it is easy to show that

$$\mathbb{P}\left(X_A \geq \frac{\sigma_A^2}{2}\right) \gtrsim \frac{e^{-\sigma_A^2/8}}{\sigma_A}.$$

By Chernoff's bound, there exists constants $c_1, c_2 > 1$ such that

$$\mathbb{P}\left(\frac{1}{c_1}\mathbb{E}(\sigma_A^2) \leq \sigma_A^2 \leq c_1 \mathbb{E}(\sigma_A^2)\right) \geq 1 - e^{-c_2 \mathbb{E}(\sigma_A^2)}.$$

Moreover we have

$$\begin{aligned}
\log \mathbb{E}\left(e^{-\sigma_A^2/8}\right) &= \sum_{l,j} \log\left(e^{-\|\mu_{al}-\mu_{bl}\|^2/8} p + 1 - p\right) \\
&\geq (1-o(1)) \sum_{l,j} p\left(e^{-\|\mu_{al}-\mu_{bl}\|^2/8} - 1\right) && \text{(by using } \log(1+x) \geq \tfrac{x}{1+x} \text{ for } x > -1\text{)} \\
&\geq -(1-o(1)) \frac{\Delta^2(a,b)}{8}. && \text{(because } e^{-x} - 1 \geq -x\text{)}
\end{aligned}$$

By consequence,

$$\begin{aligned}
\mathbb{P}(\mathcal{O}) &= \mathbb{E}\left(\mathbb{P}\left(X_A \geq \frac{\sigma_A^2}{2}\right)\right) \\
&\geq \mathbb{E}\left(\frac{e^{-\sigma_A^2/8}}{\sigma_A} \mathbf{1}_{\{\frac{1}{c_1}\mathbb{E}(\sigma_A^2) \leq \sigma_A^2 \leq c_1 \mathbb{E}(\sigma_A^2)\}}\right) \\
&\geq \frac{1}{\sqrt{c_1 \mathbb{E}(\sigma_A^2)}} \mathbb{E}\left(e^{-\sigma_A^2/8} \mathbf{1}_{\{\frac{1}{c_1}\mathbb{E}(\sigma_A^2) \leq \sigma_A^2 \leq c_1 \mathbb{E}(\sigma_A^2)\}}\right) \\
&\geq \frac{1}{\sqrt{c_1}\Delta(a,b)} \mathbb{E}\left(e^{-\sigma_A^2/8}\right) - \frac{e^{-c_2\Delta^2(a,b)}}{\sqrt{c_1}\Delta(a,b)} \\
&\geq e^{-(1-o(1))\Delta^2(a,b)}
\end{aligned}$$

since $c_2 > 1$ and $\Delta(a,b) \to \infty$.

If $\Delta^2(a,b) = O(1)$, then $\sigma_A^2 = O(1)$ with a positive probability and

$$\mathbb{P}(\mathcal{O}) \geq \mathbb{P}\left(\sigma_A^2 = O(1)\right) \mathbb{E}\left(\mathbb{P}\left(X_A \geq \frac{\sigma_A^2}{2}\right)\right) \gtrsim 1.$$

C Proof of Lemma 3

Control of $\mathcal{E}$. The event $\mathcal{D}$ occurs with probability at least $1 - n^{-5}$ by Chernoff's bound. It remains to show that conditionally on $\mathcal{D}$

$$\max_{\Lambda} \sum_{i \in \mathcal{C}_k} \left(\sum_{j \in \Lambda_\delta^c} E_{ij}\right)^2 \lesssim (n/K)^2 p_{max}.$$

Let us fix Λ and define $X_i = (\sum_j E_{ij})^2 \mathbf{1}_{\{|\sum_j E_{ij}| \lesssim np_{max}/K\}}$. By developing the square and using the independence between E_{ij} and $E_{ij'}$ for $j \neq j'$ we obtain $\mathbb{E}(X_i) \lesssim np_{max}$. A similar calculation shows that $\mathrm{Var}(X_i) \lesssim (np)^2$. Hence, Bernstein's inequality gives

$$\mathbb{P}\left(\sum_i X_i \gtrsim (n/K)^2 p_{max}\right) \leq e^{-\Omega(n/K)}.$$

We obtain the result by a union bound and the fact that conditionally on $\mathcal{D}$, it holds that $\max_i \left|\sum_j E_{ij}\right| \lesssim np_{max}/K$.

Case where $k = k' = k''$. One can first decouple the indexes i from j, j' by considering

$$S_\delta = \sum_{i,j\neq j'} \delta_i(1-\delta_j)(1-\delta_{j'})E_{ij}E_{ij'} = \sum_{i\in\Lambda}\sum_{j,j'\in\Lambda^c} E_{ij}E_{ij'}.$$

Then one can use the result from the case $k \neq k' = k''$ to show that conditionally on $\mathcal{E}$, $S_\delta \lesssim (n/K)^2 p_{max}$ with probability at least $1 - e^{-Cn/K}$ for some constant $C > 1$ and conclude as in the previous case.