Université de Lille

Inria

ED MADIS 631 - Université de Lille - INRIA

EFFICIENT ITERATIVE METHODS FOR CLUSTERING AND MATCHING PROBLEMS ON GRAPHS

MÉTHODES ITÉRATIVES EFFICIENTES POUR DES PROBLÈMES DE CLASSIFICATION ET D'APPARIEMENT DE GRAPHES

Thèse de Doctorat soutenue le 06/12/2022, par GUILLAUME BRAUN pour obtenir le grade de Docteur en **Mathématiques et leurs interactions** devant un jury composé de :

Marc LELARGE Directeur de recherche, CNRS	Rapporteur, Président du jury
Catherine MATIAS Directrice de Recherche, CNRS	Rapporteure
Alexandre d'ASPREMONT Directeur de recherche, CNRS	Examinateur
Olga KLOPP Professeur, ESSEC Business School	Examinatrice
Christophe BIERNACKI Professeur, Université de Lille	Directeur de thèse
Hemant TYAGI Chargé de recherche, INRIA	Co-encadrant de thèse

École doctorale MADIS 631
Spécialité : Mathématiques Appliquées
Unité de recherche : INRIA, Université de Lille, CNRS,
Laboratoire de mathématiques Painlevé, 59650 Villeneuve d'Ascq, France

Guillaume Braun: *Efficient iterative methods for clustering and matching problems on graphs,* © December 2022

SUPERVISORS:
Christophe Biernacki
Hemant Tyagi

LOCATION:
Villeneuve d'Ascq

TIME FRAME:
December 2022

智に働けば角が立つ。情に棹させば流される。意地を通せば窮屈だ。とかくに人の世は住みにくい。

— 夏目漱石、『草枕』

ABSTRACT

Graph-structured datasets arise naturally in many fields including biology with protein-to-protein interaction networks, ecology with predator-prey networks, and economy with financial market networks. In order to extract relevant information from these networks, one often uses clustering methods to gather nodes having similar connectivity patterns into communities. Numerous clustering algorithms have been proposed in the past decade and have been analyzed under the Stochastic Block Model (SBM), a popular random graph model. However, in practice, one often has access to side information, and it is typically unclear how this information can be incorporated into existing methods and to what extent it can help to improve the clustering performance.

We will first address this question under the Contextual Stochastic Block Model (CSBM) – a simple extension of the SBM with independent Gaussian covariates associated with each node – and propose an iterative refinement method that is fast and achieves the information-theoretic threshold for exact recovery. Our method is inspired by the Generalized Power Method (GPM) and principles of Expectation Maximization (EM) type algorithms.

Next, we extend the method to be applied to networks with heterogeneous degrees or mixed-membership, but also with different kinds of covariates like multilayer graphs with possibly missing values, or high dimensional bipartite graphs, hence showing the flexibility of the approach.

Lastly, we consider the graph matching problem where the additional graphs can be considered as correlated side information. We show that we can also use the GPM for this problem to significantly improve the matching obtained by state-of-the-art methods, and provide consistency guarantees for the GPM under the Correlated Wigner Model (CoWM).

野を横に馬牽むけよほととぎす

— 松尾芭蕉

REMERCIEMENTS

Par ordre chronologique, il me faut d'abord remercier Matsuo Basho, le célèbre poète de la grenouille qui plonge, sans lequel je n'aurais sans doute pas entrepris cette thèse. C'est en déchiffrant l'un de ses haïkus gravé sur une pierre recouverte de mousses évoquant un cavalier excité par le cri d'un coucou que j'ai réalisé que la poésie me serait toujours plus étrangère que les bornes minimax.

Il me faut ensuite remercier les personnes qui m'ont mené vers l'équipe Modal grâce à leurs conseils, encouragements et contacts : Philippe, David et Pascal.

Bien évidemment, ce travail n'aurait pas vu le jour si Christophe ne m'avait pas chaleureusement accueilli et proposé un sujet de recherche en collaboration avec Hemant. Je leur suis particulièrement redevable du temps et de l'énergie qu'ils ont consacré à me former. En particulier à Hemant qui a fait preuve d'une rare patience à relire et corriger mes preuves.

Je tiens également à exprimer ma gratitude aux rapporteurs, Catherine Matias et Marc Lelarge qui m'ont fait l'intimidant honneur de se pencher sur ma thèse. Je suis également reconnaissant à Olga Klopp et Alexandre d'Aspremont d'avoir très gentiment accepté de participer au jury.

Si j'ai pu mener ce travail à son terme c'est également grâce au personnel administratif de l'Inria (Anne et Charlotte en particulier) ainsi que quelques anciens collègues de l'Insee qui m'ont aidé dans les procédures administratives. Enfin, je dois également saluer tous les membres de l'équipe Modal qui ont supporté mes jérémiades dans la bonne humeur pendant trois ans. En particulier à Cristian qui a accepté le rôle de souffre douleur aux échecs et m'a été d'un soutien précieux durant ma thèse.

Il me reste encore quelques lignes pour remercier, même s'ils ne liront vraissemblablement jamais ce manuscrit, ma famille, mes amis plus ou moins lointains, et Agathe, l'araignée vivant derrière le réfrigérateur, pour leurs présences réconfortantes.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACRONYMS

GMM Gaussian Mixture Model

SBM Stochastic Block Model

SSBM Symmetric SBM

CSBM Contextual Stochastic Block Model

CSSBM Contextual Symmetric SBM

MLSBM Multi-Layer Stochastic Block Model

DCSBM Degree Corrected SBM

MMSBM Mixed Membership SBM

RDPG Random Dot Product Graph

BiSBM Bipartite Stochastic Block Model

SBiSBM Symmetric BiSBM

CoWM Correlated Wigner Model

CoERM Correlated Erdös-Renyi Model

OSBM Overlapping SBM

CoSBM Correlated Stochastic Block Model

OCCAM Overlapping Continuous Community Assignment Model

SBMO SBM with Overlap

PCA Principal Component Analysis

MLE Maximum Likelihood Estimator

LAP Linear Assignment Problem

QAP Quadratic Assignment Problem

VEM Variational Expectation Maximization

ELBO Evidence Lower Bound

GPM Generalized Power Method

SDP Semi-Definite Programming

EM Expectation Maximization

OLMF Orthogonal Linked Matrix Factorization

GOE Gaussian Orthogonal Ensemble

SNR Signal-to-Noise Ratio

MAP Maximum A Posteriori

NMI Normalized Mutual Information

GPM Generalized Power Method

NOTATIONS

- The cardinality of a set $\mathcal{C}$ is denoted by $|\mathcal{C}|$.
- The all-ones vector of size n is denoted by $\mathbf{1}_n$.
- The indicator function of an event A is denoted by $\mathbb{1}_A$.
- The set of permutation over n elements is denoted by $\mathfrak{S}_n$ and the corresponding set of matrices is denoted by $\mathcal{P}_n$
- The empty set is denoted by $\emptyset$.
- The column j of A will be denoted by $A_{:j}$, and the (i,j)th entry by A_{ij}. 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 transpose of A is denoted by $A^\top$ and $A_{:j}^\top$ corresponds to the jth row of $A^\top$ by convention.
- I_k denotes the $k \times k$ identity matrix.
- The number of non zero entries of a matrix A is denoted by $nnz(A)$.
- For matrices, we use $\|.\|$ and $\|.\|_F$ to respectively denote the spectral norm (or Euclidean norm in case of vectors) and Frobenius norm.
- The trace of a matrix A is denoted by $\mathrm{Tr}(A)$.
- We use lowercase letters ($\epsilon, a, b, \ldots$) to denote scalars and vectors, except for universal constants that will be denoted by $c_1, c_2, \ldots$ for lower bounds and $C_1, C_2, \ldots$ for upper bounds. We use uppercase letters to denote matrices and random variables (with possible local exceptions).
- For any sequences $(a_n)_{n\geqslant 1}$ and $(b_n)_{n\geqslant 1}$ if there is a constant $C > 0$ such that $a_n \leqslant Cb_n$ (resp. $a_n \geqslant Cb_n$), we will use the notation $a_n \lesssim b_n$ (resp. $a_n \gtrsim b_n$). If the inequalities only hold

for n large enough, we will use the notation $a_n = O(b_n)$ (resp. $a_n = \Omega(b_n)$). If $a_n \lesssim b_n$ (resp. $a_n = O(b_n)$) and $a_n \gtrsim b_n$ (resp. $a_n = \Omega(b_n)$), then we write $a_n \asymp b_n$ (resp. $a_n = \Theta(b_n)$).

1 INTRODUCTION

Identifying and recognizing patterns is a natural task for the human brain. It helps to extract relevant information from one's environment and discover general laws in nature. Hence, *clustering* – the act of gathering objects by similarities – is a fundamental activity that lies at the root of knowledge elaboration. This is probably why one spends so much time trying to organize one's library by sorting one's books by topic, color, shape, etc.

However, even if the ability of human brain to categorize and organize information is remarkable, it is not capable of processing large amounts of data and can fail to detect some patterns. This is a reason why one has tried to automate this procedure and started to develop clustering algorithms. The first systematic numerical approaches were developed for applications in social sciences (Czekanowski, 1909; Zubin, 1938). More historical references can be found in the introduction of the book of Bouveyron et al., 2019. Most of the initially proposed methods were based on the definition of a distance between the objects of interest. Then, for a given distance, one can use optimization methods to obtain a partition of the objects such that the objects belonging to the same class have a small distance between them and a large distance with objects from another class. It is exemplified by the `k-means` algorithm, one of the most popular clustering methods.

The previous heuristic approach has several drawbacks. The choice of the distance is arbitrary and it is impossible to assess the quality of the resulting clustering. Also, methods that usually work in practice can fail in particular instances. This is for example the case of the spectral method applied on the cockroach graph (Luxburg, 2004).

A way to overcome these issues is to introduce a probabilistic model for the data and formulate the clustering problem as a statistical inference problem.

1.1 A STATISTICAL APPROACH OF CLUSTERING

The first generative model proposed for clustering numerical vector data $(X_i)_{i=1}^n \subset \mathbb{R}^d$ of dimension d is the Gaussian Mixture Model (GMM). Each vector X_i is assigned to a community $k \in [K]$ and then generated from a Gaussian probability distribution than only depends on k, independently of the other vectors. More formally, for each i, assign to i a label $z_i \in [K]$ and generate X_i by

$$X_i = \mu_{z_i} + \epsilon_i, \text{ where } \epsilon_i \overset{\text{ind.}}{\sim} \mathcal{N}(0, \Sigma_{z_i})$$

where for all $k \leqslant K$, $\mu_k \in \mathbb{R}^d$ and $\Sigma_k \in \mathbb{R}^{d\times d}$ is a positive definite matrix. Clustering $(X_i)_{i=1}^n$ is now equivalent to estimate the latent partition given by $z : [n] \to [K]$ that associates i to z_i.

A natural approach is to maximize the likelihood associated with the problem. For example, under the isotropic case where all $\Sigma_k = \mathrm{Id}$, one can easily check that the (complete) Maximum Likelihood Estimator (MLE) associated to the GMM is related to the one provided by the `k-means` optimization problems

$$\arg \min_{\mu_1,\dots,\mu_K\in\mathbb{R}^d,\mathcal{C}} \sum_{k=1}^{K} \sum_{i\in\mathcal{C}_k} \|X_i - \mu_k\|^2 \tag{1.1.1}$$

where $\mathcal{C}$ is a partition of $[n]$ into K sets $\mathcal{C}_1, \dots, \mathcal{C}_K$.

Unfortunately, solving (1.1.1) requires doing an exhaustive search over all the possible K^n partitions and cannot be done in a reasonable time (i.e. polynomial in n). Even if k is fixed to 2, `k-means` clustering is a NP-hard problem as shown in Dasgupta, 2008. So in practice, one has to use other strategies. For example, there are approximate `k-means` algorithms, see for example Kumar et al., 2004, that run in polynomial time in n and produce an output with an estimation error proportional to the error associated with the optimal solution. In the following section, we will describe an alternative strategy, particularly useful when one can obtain a first rough estimate of the latent partition.

1.2 ALTERNATING OPTIMIZATION

A popular approach to solve (1.1.1) is to use Lloyd's algorithm: starting from a first estimate $\mathcal{C}^{(0)} = \{\mathcal{C}_1^{(0)}, \dots, \mathcal{C}_K^{(0)}\}$ of the latent partition one can estimate the centroids of the partition by

$$\mu_k^{(0)} = \frac{1}{|\mathcal{C}_k^{(0)}|} \sum_{i\in\mathcal{C}_k^{(0)}} X_i$$

and then update the partition to obtain $\mathcal{C}^{(1)}$ by assigning each X_i to the closest centroid. One can check that this is equivalent to solve (1.1.1) by alternatively fixing $\mathcal{C}$ or μ.

This approach for solving nonconvex optimization problems is referred to as *alternating optimization*. It has been used with success in discrete optimization problems including approximate ranking, group synchronization, and multireference alignment (see Gao et al., 2019) as well as for continuous optimization problems, e.g. matrix completion, phase retrieval (Chi et al., 2019) or Principal Component Analysis (PCA) with missing data (Zhu et al., 2019). This principle is also at the core of model-based approaches: EM-like algorithms (Dempster et al., 1977) use alternating optimization to maximize a proxy of the likelihood, the Evidence Lower Bound (ELBO).

One of the drawbacks of this approach is that it requires an initial estimate of the partition. But one can often use spectral methods to obtain a cheap initial estimate. One can also rely on random sampling strategies. This is a very common approach in the model-based literature and it can be proved to work under some assumptions, see for example Wu et al., 2022.

1.3 NETWORK STRUCTURED DATA

The past two decades have witnessed the increase of network structured data with the rise of social networks. Contrary to the previously discussed vector data setting where each measurement X_i is independent of the other, networks encode pairwise interactions between the agents of interest represented by nodes, and hence there are structural dependencies among the observations. This is a fundamental difference. But before discussing its algorithmic implications, let us first introduce a generative model to give a statistical meaning to the clustering task.

The most popular random graph model that incorporates a community structure is the SBM. It was introduced by (Holland et al., 1983) and is often used as a benchmark to evaluate clustering algorithms on graphs. It is defined by the following parameters.

- The set of nodes $\mathcal{N} = [n]$.
- Communities $\mathcal{C}_1, \dots, \mathcal{C}_K$, of respective sizes $n_1, \dots, n_K$, forming a partition of $\mathcal{N}$.
- A membership matrix $Z \in \mathcal{M}_{n,K}$ where $\mathcal{M}_{n,K}$ denotes the class of membership matrices. Here, $Z_{ik} = 1$ if node i belongs to $\mathcal{C}_k$, and is 0 otherwise. Each membership matrix Z can be associated bijectively with a 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
$$\Pi = (\Pi_{kk'})_{k,k' \in [K]} \in [0,1]^{K \times K}.$$

Denoting by $P = (p_{ij})_{i,j \in [n]} := Z \Pi Z^\intercal$, a graph $\mathcal{G}$ is distributed according to $\mathrm{SBM}(Z, \Pi)$ if the entries of the corresponding symmetric adjacency matrix A are generated by

$$A_{ij} \overset{\text{ind.}}{\sim} \mathcal{B}(p_{ij}), \quad 1 \leqslant i \leqslant j \leqslant n,$$

where $\mathcal{B}(p)$ denotes a Bernoulli distribution with parameter p. Hence the probability that two nodes are connected depends only on their respective community membership. Here we allowed self-loop for analysis convenience, but usually the diagonal of the matrix A is removed.

Usually observed networks are sparse – most of the entries are zeros – and the sparsity level of a SBM is measured by $p_{max} = \max_{i,j} p_{ij}$.

A direct application of Lloyd algorithm to this setting will require to estimate the centroids $\mu_k := \Pi_{k:} Z^\top$ by averaging the rows of the adjacency matrix A that belong to the same community. By doing that we completely ignore the fact that the columns of A have a similar block structure. Typically, we would expect that such an approach won't be robust to noise since it doesn't take into account the low-rank structure of the problem.

An alternative is to solve the following optimization problem that takes into account the row and columns block structure:

$$\arg \min_{Z \in \mathcal{M}_{n,K}, \Pi \in [0,1]^{K \times K}} \left\| A - Z \Pi Z^\top \right\|^2. \tag{1.3.1}$$

Unfortunately, this is again an hard combinatorial optimization problem. But we can try an alternating optimization strategy. If the partition Z were known, the solution $\hat{\Pi}$ of (1.3.1) will be given by $Z^\top A Z$. But for a fixed Π, (1.3.1) is still difficult to solve due to the quadratic nature of the objective function in Z. We will show in Chapter 3 that it can be done by using a low-rank embedding of A computed from an initial estimate of the latent partition. Another possible approach would be to relax the condition $Z \in \mathcal{M}_{n,K}$ to $Z \in \mathcal{O}_{n,K}$ where

$$\mathcal{O}_{n,K} = \{M \in \mathbb{R}^{n \times K} \text{ such that } M^\top M = I_K\}$$

is the set of matrices with orthonormal columns and then use continuous optimization tools such as gradient descent. This approach is a special case of the Orthogonal Linked Matrix Factorization (OLMF) method developed for multilayer graphs or multiview clustering, see e.g. (Paul et al., 2020; Dong et al., 2014). We will discuss in more detail the existing method for clustering graphs in Chapter 2.

One can also wonder why one should use a least square criterion instead of the likelihood as an objective function. We will show in Chapter 3 that using a least square proxy of the likelihood doesn't induce any loss of accuracy – under some parameter regimes – compared to the solution obtained by solving the MLE using a similar alternating optimization strategy. Moreover, it leads in practice to a significant gain in term of computational time. Popular model-based approaches rely on variational EM algorithm (Celisse et al., 2011) and stochastic variant (Brault et al., 2014).

1.4 SIDE INFORMATION

In practice, real-world networks often come with side information in the form of nodes or edges covariates. For example, when analyzing interactions among people on a social network, we have access to additional features such as gender, age, or ethnicity that can be relevant

for the clustering task. Whereas several algorithms have been proposed to incorporate nodes or edges side information, little is known about the value of this side information to improve the clustering performance. The question has not only a theoretical interest but also practical implications on the design of algorithms that should incorporate optimally the available side information.

In this dissertation we will consider several kinds of side information

- nodes side information provided by a GMM;
- edges side information that can be represented by a multilayer graph;
- edges side information that comes from correlated but unlabeled graphs.

We will further discuss the last setting in the next section.

1.5 GRAPH MATCHING

In many applications, one observes several correlated realizations of a graph. For example, in computer vision, one can observe different images of an object – than can be modeled by a graph – in different positions. In order to connect these observations it is necessary to align the graphs, that is to say, find the permutations that maximize the similarity between each pair of graphs. When there are only two graphs, this optimization problem can be formulated as follows. Let A and B two adjacency matrices associated with graphs G and H. The permutation that best aligns G and H is given by the solution of

$$\max_{\pi \in \mathfrak{S}_n} \sum_{i,j} A_{ij} B_{\pi(i)\pi(j)} \tag{1.5.1}$$

where $\mathfrak{S}_n$ denotes the set of permutations over $[n]$ or equivalently

$$\min_{P \in \mathfrak{S}_n} \left\| A - PBP^\top \right\|_F^2. \tag{1.5.2}$$

This is a hard combinatorial optimization problem known as the Quadratic Assignment Problem (QAP). Indeed the problem is known to be NP-hard (Sahni et al., 1976). Fortunately, under some generative models it can be solved exactly. The most studied models for correlated random graphs are the following.

CORRELATED WIGNER MODEL $CoWM(n, \sigma, \pi^*)$. The following Gaussian model has been proposed in Ding et al., 2021.

$$A_{ij} \sim \begin{cases} \mathcal{N}(0, \frac{1}{n}) \text{ if } i \neq j, \\ \mathcal{N}(0, \frac{2}{n}) \text{ if } i = j, \end{cases}$$

and $B_{x^*(i)x^*(j)} = \sqrt{1-\sigma^2}A_{ij} + \sigma Z_{ij}$, where $Z \overset{d}{=} A$. Both A and B are distributed as the Gaussian Orthogonal Ensemble (GOE). Here the parameter $\sigma > 0$ should be interpreted as the noise parameter and in that sense, B can be regarded as a "noisy perturbation" of A. Moreover, $x^* \in \mathfrak{S}_n$ is the ground-truth (or latent) permutation that we seek to recover. It is not difficult to verify that the problem (1.5.1) is in fact the MLE of x^* under the CoWM.

CORRELATED ERDÖS-RENYI MODEL $\mathrm{CoERM}(n, q, s, x^*)$. For $q, s \in [0, 1]$, the correlated Erdos-Renyi model with latent permutation $x^* \in \mathfrak{S}_n$ can be described in two steps.

1. A is generated according to the Erdös-Renyi model $\mathrm{ERM}(n, q)$: for all $i, j \in [n]$ such that $i < j$, $A_{ij} \overset{\text{ind.}}{\sim} \mathcal{B}(q)$.
2. Conditional on A, the entries of B are i.i.d according to the law

$$B_{x^*(i),x^*(j)} \sim \begin{cases} \mathcal{B}(s) & \text{if} \quad A_{ij} = 1, \\ \mathcal{B}\left(\frac{q}{1-q}(1-s)\right) & \text{if } A_{ij} = 0. \end{cases} \tag{1.5.3}$$

The graph matching problem (1.5.2) share a lot of similarities with the clustering problem (1.3.1). The main difference is that the latter involves a latent low-dimension structure since usually $K \ll n$. But if $K = n$ and the matrix Π were given, the two problems would be equivalent. So it is not surprising that there are strong links between methods used for addressing each of these problems. However, graph matching is generally harder. It is usually difficult to find a good enough initial matching. This can be explained intuitively by the absence of a latent low-dimension structure. We will discuss the information-theoretic limits and techniques used for graph matching in more detail in Chapter 6.

1.6 THESIS CONTRIBUTION AND OUTLINE

In this dissertation, we investigate the value of side information in clustering and matching problems on graphs from an information-theoretic perspective but also from a practical aspect by developing efficient algorithms. The first part of this thesis is devoted to clustering problems where *independent* side information is available in the form of nodes or edges covariates.

In Chapter 2, we provide an overview of the relevant literature on graph clustering– with a focus on theoretical results – and present the principal methods: spectral and Semi-Definite Programming (SDP) relaxations, variational methods, and approximate message passing algorithms.

In Chapter 3 we study a simple model – the CSBM – where each node of a graph generated from a SBM is associated with Gaussian

covariates generated from an GMM, independent of the graph but sharing the same latent partition. We propose an algorithm that iteratively estimates the model parameters and refines the partition based on a least square criterion. We show that our algorithm achieves the threshold for exact recovery under a symmetric version of the CSBM, and that up to a constant factor, its rate of convergence is minimax optimal in the general setting. The technical tools introduced in this chapter will be further used in the following chapters.

In Chapter 4, we consider multilayer graphs where each layer describes a different modality of the pairwise interactions between the agents of interest. Multilayer graphs can be thought of as a way to represent networks with edges side information. We further assume that each node doesn't appear on each layer and propose a generative model for this setting based on the Multi-Layer Stochastic Block Model (MLSBM). We study an extension of classical methods used for clustering multilayer graphs and derive consistency guarantees. We also explain how the method developed in Chapter 3 to this setting can be applied to the multilayer setting and provide some numerical experiments.

In Chapter 5, we study an extension of our iterative refinement method to graphs that are not generated by a SBM. First, we consider high-dimensional bipartite graphs – generated from the Bipartite Stochastic Block Model (BiSBM) – and show that our method can achieve the minimax optimal rate of convergence, up to a constant factor. Secondly, we show how to adapt the algorithm to graphs with heterogeneous degrees and overlapping communities.

The second part of the thesis is devoted to graph matching problems. In this setting, we observe two, or more, graphs that are generated from the same parent graphs but where the node labels are not known and consequently the graph cannot be directly aligned. Finding the permutation that best aligns these graphs is the goal of matching algorithms. If this permutation could be found, then we would obtain a multilayer graph. But contrary to the setting developed in Chapter 4, the layers are no longer *independent* but *correlated*. Hence, graph matching can be used as a subroutine for clustering tasks. For example, it has recently been shown (Racz et al., 2021; Gaudio et al., 2022) that the side information provided by correlated graphs can help to improve the threshold for exact recovery under the Correlated Stochastic Block Model (CoSBM). Unfortunately, the proposed algorithm to achieve this task is not implementable in polynomial time.

In Chapter 6 we further discuss the literature on graph matching and present the main methods and associated theoretical guarantees.

In Chapter 7 we analyze an iterative refinement matching method under the CoWM. Provided that one can have access to an initial estimate of the ground-truth permutation that correctly matches a suffi-

ciently large proportion of nodes, one can show that this refinement method – based on the GPM – significantly improves the matching.

Finally, the appendix gathers some general results that are used in different chapters.

1.7 PUBLICATIONS

This dissertation is in large parts based on the following publications and technical reports. The publications relevant to each chapter will be precised at the appropriate places.

Araya, Ernesto, Guillaume Braun, and Hemant Tyagi (2022). *Seeded graph matching for the correlated Wigner model via the projected power method*. arXiv: 2204.04099 [math.ST].

Braun, Guillaume and Hemant Tyagi (2022a). *Minimax Optimal Clustering of Bipartite Graphs with a Generalized Power Method*. arXiv: 2205.12104 [math.ST].

Braun, Guillaume, Hemant Tyagi, and Christophe Biernacki (2021). "Clustering multilayer graphs with missing nodes." In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Vol. 130, pp. 2260–2268.

Braun, Guillaume, Hemant Tyagi, and Christophe Biernacki (2022b). "An iterative clustering algorithm for the Contextual Stochastic Block Model with optimality guarantees." In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162, pp. 2257–2291.

Part I

CLUSTERING GRAPHS WITH SIDE INFORMATION

In the first part of this dissertation, we will investigate the adding value of side information for clustering graphs. First we will study a simple setting where each node is associated to vector covariates. Then we will show that the proposed optimal strategy to tackle this problem can be extended to other settings involving edges covariates (multilayer graphs) and also adapted to more general random graphs models incorporating degree heterogeneity and overlapping communities.

2

INTRODUCTION TO GRAPH CLUSTERING

In this chapter, we will first present the information-theoretic limits of clustering under the SBM in Section 2.1. Then, we will discuss the principal methods used for clustering in Section 2.2. Finally, we will give an overview of the possible extensions of the SBM that incorporate properties often observed in real-life networks but not modeled by the SBM. Interestingly, some of the clustering strategies used for SBM can be extended to some of these more complex random graph models.

2.1 INFORMATION THEORETIC LIMITS OF CLUSTERING

In this section, to simplify the exposition, we will only present the information-theoretic thresholds associated with the Symmetric SBM (SSBM), a particular instance of the SBM satisfying the following constraints.

- The communities have the same size : $|\mathcal{C}_k| = \frac{n}{K}$ for all $k \in [K]$.
- The connectivity matrix $\Pi = (p-q)I_K + q\mathbb{1}_K\mathbb{1}_K^\top$ where $1 \geqslant p > q \geqslant 0$.

This model will be referred to as $SSBM(n, K, p, q)$. It is often used in theoretical analysis since the simplicity of the model allows some simplification in the calculations due to symmetries. Also, the different information-theoretic thresholds are easier to interpret for this model since they only depend on the difference $p-q$, than for the general SBM that involves the Chernoff-Hellinger divergence. A more complete presentation can be found in Abbe, 2018.

The usual way to quantify (for theoretical purposes) the accuracy of an estimate $\hat{z}$ of the latent partition is the following quantity referred to as the **misclustering rate**

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

One can have different recovery requirements.

- **Exact recovery:** $r(\hat{z}, z) = 0$ with probability $1 - o(1)$ when $n \to \infty$, that is to say one recovers exactly the partition w.h.p. This condition is also referred to in the literature as **strong consistency**.

- **Almost exact recovery:** $r(\hat{z}, z) = o(1)$ with probability $1 - o(1)$ when $n \to \infty$, that is to say the proportion of misclustered nodes goes to zero when $n \to \infty$. This condition is also referred to in the literature as **weak consistency**.
- **Partial recovery:** $r(\hat{z}, z) \geqslant \alpha$ with probability $1 - o(1)$ when $n \to \infty$, that is to say one can correctly classify at least a proportion α of the nodes w.h.p. This regime is interesting only if $\alpha > 1/K$, i.e. we can do better than a random guess of the communities. This last regime is referred to as the **detection** regime in the literature.

For exact recovery, we need to be in a regime where $p_{max} = \Theta(\frac{\log n}{n})$. The following theorem gives a more precise statement.

Theorem 1 (Abbe et al., 2015a)**.** *Under the SSBM*$(n, K, \frac{a \log n}{n}, \frac{b \log n}{n})$*, exact recovery is possible only when*

$$\frac{(\sqrt{a} - \sqrt{b})^2}{K} > 1.$$

Moreover, in this parameter regime there are polynomial time algorithms that efficiently recover the latent partition.

For almost exact recovery, the necessary condition is relaxed to $np_{max} \to \infty$.

Theorem 2 (Zhang et al., 2016a)**.** *Assume that* K *is a constant. Under the SSBM*(n, K, p, q)*, a sufficient and necessary condition for almost exact recovery is*

$$n(p - q) \to \infty.$$

On the other hand, partial recovery only requires $p_{max} = \Omega(\frac{1}{n})$.

Theorem 3 (Abbe et al., 2015b)**.** *Under the SSBM*$(n, K, \frac{a}{n}, \frac{b}{n})$*, it is possible to detect communities in polynomial time if*

$$\mathrm{SNR} := \frac{(a - b)^2}{K(a + (K - 1)b)} > 1.$$

When $K > 4$ *and* $\mathrm{SNR} < 1$*, it is still possible to detect communities with a non-efficient algorithm.*

Remark 1. *There are several algorithms, including spectral methods and SDP relaxation that have been shown to achieve the threshold for exact recovery or lead to almost exact recovery. However, the partial recovery regime is usually more challenging. We will discuss these methods in Section 2.2.*

2.2 CLUSTERING METHODS

2.2.1 *Spectral methods*

Spectral methods rely on the computation of the eigenvectors of a wisely chosen matrix that contains information about the community structure. Their popularity comes from the fact that they are usually fast (one can compute quickly the top K eigenvalues and associated eigenvectors of a matrix), easy to implement, and often accurate. It is also common to use an additional refinement step to the estimate provided by a spectral method.

Let us explain the method when directly applied to the adjacency matrix. In this case, the method will be referred to as the *vanilla spectral method*. Assume that $A \sim SSBM(n, 2, p, q)$ where $p - q = \Omega(p)$ and $np \to \infty$. In expectation, A has a block structure[1]

$$\mathbb{E}(A) = \begin{pmatrix} p\mathbb{1}_{n/2}\mathbb{1}_{n/2}^\top & q\mathbb{1}_{n/2}\mathbb{1}_{n/2}^\top \\ q\mathbb{1}_{n/2}\mathbb{1}_{n/2}^\top & p\mathbb{1}_{n/2}\mathbb{1}_{n/2}^\top \end{pmatrix}.$$

It is easy to check that $\mathbb{E}(A)$ has two non zero eigenvalues $\lambda_1 = n(p+q)/2$ and $\lambda_2 = n(p-q)/2$ associated with eigenvectors $u_1 = \mathbb{1}_n$ and

$$u_2 = (\underbrace{1, \dots, 1}_{n/2 \text{ times}}, \underbrace{-1, \dots, -1}_{n/2 \text{ times}})^\top.$$

Hence if there were no noise, communities could be identified by the sign of the entries of u_2. One can show that this argument also works when there is noise. Let $\hat{u}_2$ the eigenvector associated with the second largest eigenvalue of A denoted by $\hat{\lambda}_2$. By using Weyl's inequality (Weyl, 1912) we have for $k = 1, \dots, n$

$$|\lambda_k - \hat{\lambda}_k| \leqslant \|A - \mathbb{E}(A)\|$$

where $\hat{\lambda}_k$ is the kth largest eigenvalue of A. By concentration inequalities, when $np_{max} \gtrsim \log n$, $\|A - \mathbb{E}(A)\| \lesssim \sqrt{np}$ w.h.p. This shows that the spectrum of A is close to the spectrum of $\mathbb{E}(A)$ because the largest eigenvalues are of magnitude $np \gg \sqrt{np}$ in the almost exact recovery regime. Moreover Davis-Kahan theorem (Yu et al., 2014) implies that the angles θ_i between the eigenvectors corresponding to the ith eigenvalue have a sinus bounded by

$$\sin\theta_i \lesssim \frac{\|A - \mathbb{E}(A)\|}{\min_{i \neq j} |\hat{\lambda}_j - \hat{\lambda}_i|} \lesssim \frac{1}{\sqrt{np}}.$$

Consequently $\|\hat{u}_2 \pm u_2\| = o(1)$ and one can consistently use the sign of the entries of $\hat{u}_2$ to estimate the community membership. This idea can be generalized to general SBM, see Lei et al., 2015.

1 We can assume w.l.o.g that the first $n/2$ entries are in community 1.

This proof technique only leads to weak consistency guarantees. But one can use recent results (Abbe et al., 2020c; Lei, 2019) on entrywise perturbation of eigenvectors to show that the vanilla spectral method achieves the threshold for exact recovery.

A popular alternative is to use the Laplacian $L = D - A$ or the symmetric[2] Laplacian $L_{sym} = I - D^{-1/2}AD^{-1/2}$ where D is a diagonal matrix defined by $D_{ii} = A_{i:}\mathbb{1}_n$. Spectral clustering on the Laplacian can be interpreted as a relaxation of the (normalized or not) Ratio Cut optimization problem as explained in Luxburg, 2004. Strong consistency guarantees for this algorithm have been derived in the recent work of (Deng et al., 2021).

However, in the detection regime, the vanilla spectral method will fail because nodes with high degrees prevent the eigenvectors from concentrating on the communities. A simple way to fix this problem is to use regularization by reducing the degree of the problematic nodes of A (Le et al., 2017) (or adding $\epsilon \mathbf{1}_n \mathbf{1}_n^\top$ for an appropriate choice of ϵ to the Laplacian). Unfortunately, this approach doesn't achieve the information-theoretic threshold for detection. But instead of using the adjacency matrix or a regularized version, on can use the non-backtracking matrix (Bordenave et al., 2018) – a matrix indexed by oriented edges where cycles of order two are removed – or the powered adjacency (Abbe et al., 2020a). The main idea is to incorporate more information about the neighborhood of a node at a certain depth and use the fact that Erdös-Renyi like graphs have a local tree structure.

2.2.2 *Semi-definite programming (SDP)*

A semi-definite program is an optimization program of the form

$$\begin{aligned} &\min_{X \in \mathbb{R}^{n\times n}} \langle A, X\rangle \\ &\text{s.t. } X \succeq 0 \\ &\text{and } \langle B_i, X\rangle \leqslant b_i \text{ for } i = 1, \dots, m \end{aligned}$$

where $A, B_1, \dots, B_m \in \mathbb{R}^{n\times n}$ and $b_i \in \mathbb{R}$ for all $i \in [m]$. It is a convex optimization problem since the set of positive semi-definite matrices is a cone. Consequently, such optimization problems can be solved in polynomial time by using for example interior point methods (Vandenberghe et al., 1996; Huang et al., 2021). But the computational cost of SDP methods is usually expensive. Recent works proposed different strategies to solve SDP optimization problem more efficiently. In particular, the Burer-Monteiro method (Burer et al., 2001) has surged in popularity during the past decade.

A common strategy to solve NP-hard combinatorial problems is to relax them into a SDP optimization problem. This approach has

2 One can also use the random-walk Laplacian $L_w = I - D^{-1}A$.

been used for example to solve phase synchronization problem (Gao et al., 2022), clustering sub-Gaussian mixtures (Giraud et al., 2019) or community detection (Li et al., 2021). Let us illustrate the principle of the method for a $SSBM(n, 2, p, q)$ where the diagonal elements are removed.

Let us define the binary variable X_{ij} equal to 1 when i and j are in the same cluster and 0 otherwise and $X = (X_{ij})_{i,j\in[n]} \in \mathbb{R}^{n\times n}$ be the corresponding partition matrix. The set of all partition matrices is denoted by $\mathcal{X}_n$. It is easy to show (Li et al., 2021) that maximizing the likelihood of the model is equivalent to

$$\max_{X\in\mathcal{X}_n} \sum_{i<j} (A_{ij} - \lambda) X_{ij} \tag{2.2.1}$$

where $\lambda = \frac{\log \frac{1-p}{1-q}}{\log \frac{p(1-q)}{(1-p)q}}$ and the maximum is taken over all partition matrices X. We can now relax the constraints on X and only use the following properties of X:

1. X is positive semi-definite;
2. $X_{ij} \geqslant 0$ for all i, j;
3. $X_{ii} = 1$ for all i.

Then the optimization problem (2.2.1) can be relaxed to

$$\begin{aligned} &\max_{X\in\mathbb{R}^{n\times n}} \langle A - \lambda J_n, X\rangle \\ &\text{s.t. } X \succeq 0, X_{ii} = 1 \text{ for all } i \text{ and } X \geqslant 0 \end{aligned} \tag{2.2.2}$$

where $J_n = \mathbb{1}_n \mathbb{1}_n^\top$.

Under some noise regime, one can show that the solution $\hat{X}$ of (2.2.2) is equal to X^*, the ground truth partition matrix. The proof strategy is as follows. First one can show by using KKT conditions and constructing a dual certificate (it is generally difficult) that X^* is indeed a solution of (2.2.2). It remains then to show that the solution is unique. One can also have weaker requirement and show that $\hat{X}$ is close to X^*.

2.2.3 *Variational Expectation Maximization (VEM)*

EM algorithms are a family of iterative methods for finding a local maximum of the log-likelihood when the statistical model depends on unobservable latent variables. VEM algorithms involve an additional approximation to approximate a posterior distribution which computation is usually intractable. Let us detail the algorithm for the SBM. It was first proposed by Daudin et al., 2008.

Let assume that $A \sim \mathrm{SBM}(Z, \Pi)$ and that $Z_i \overset{\text{i.i.d.}}{\sim} \mathrm{Multi}(\alpha_1, \dots, \alpha_K)$ where $\mathrm{Multi}(\alpha_1, \dots, \alpha_K)$ is a multinomial distribution with parameters

$$(\alpha_k)_{k \in [K]} \in [0,1]^K \text{ such that } \sum_k \alpha_k = 1.$$

In other words, we added a prior distribution over the partition. The coefficients α_K can be interpreted as the proportion of each community. In the approximately equal size case, we will choose $\alpha_1 = \ldots = \alpha_K = 1/K$. Let us denote $\theta = (\Pi, \alpha_1, \dots, \alpha_K)$.

The complete log-likelihood of the model $\log \mathcal{L}_\theta(A, Z)$ can be written as

$$\log \mathcal{L}_\theta(A, Z) = \sum_{i,k} Z_{ik} \log \alpha_k + \sum_{\substack{i,k,k' \\ j>i}} Z_{ik} Z_{jk'} \log b(A_{ij}, \Pi_{kk'}) \quad (2.2.3)$$

where $b(A_{ij}, \Pi_{kk'}) = \Pi_{kk'}^{A_{ij}} (1 - \Pi_{kk'})^{1 - A_{ij}}$ is the density of a Bernoulli with parameter $\Pi_{kk'}$.

One can show that for every distribution Q on Z

$$\log \mathcal{L}_\theta(A) = \log \mathbb{E}_Z \mathcal{L}_\theta(A, Z) \geqslant \underbrace{\mathbb{E}_{Z \sim Q}(\log \mathcal{L}_\theta(A, Z)) - \mathbb{E}_{Z \sim Q}(\log Z)}_{J_\theta(A, Q)}.$$

This lower bound of the observed likelihood is often called the evidence lower bound (ELBO). By using the identity

$$J_\theta(A, Q) = \log \mathcal{L}_\theta(A) - \mathrm{KL}(Q || \mathcal{L}_\theta(Z|A))$$

one can see that the ELBO is maximised when Q correspond to the posterior distribution $\mathcal{L}_\theta(Z|A)$. Unfortunately, due to dependencies in the model one cannot compute this posterior distribution. This is why we use a *variational approximation*: we will restrict the optimization problem to the probability distributions Q that can be factorized

$$Q(Z) = \prod_i Q_i(Z_i).$$

This approximation is also often referred to as the *mean field* approximation. Now we can use alternative optimization to find the couple (θ, Q) that maximizes $J_\theta(A, Q)$.

VE-STEP. Assume that the model parameter θ is fixed. The distribution Q is by definition fully characterised by the the variational parameters $\tau_{ik} = \mathbb{P}_Q(Z_{ik} = 1)$ for all $i \in [n], k \in [K]$ that satisfies the constraint $\sum_k \tau_{ik} = 1$ for all i. It is easy to check that

$$J_\theta(A, Q) = \sum_{i,k} \tau_{ik} (\log \alpha_k - \log \tau_{ik}) + \sum_{\substack{i,k,k' \\ j>i}} \tau_{ik} \tau_{jk'} \log b(A_{ij}, \Pi_{kk'}).$$

By the KKT conditions, the solution $\hat{\tau} = (\hat{\tau}_{ik})_{i,k}$ that maximizes $J_\theta(Q, A)$ satisfies for all $i \in [n]$ and $k \in [K]$

$$\hat{\tau}_{ik} \propto \alpha_k \prod_{j \neq i} \prod_{k'} b(A_{ij}, \Pi_{kk'})^{\hat{\tau}_{jk'}}.$$

Consequently, $\hat{\tau}$ can be estimated by solving a fixed point problem and using the constraint $\sum_k \tau_{ik} = 1$.

M-STEP. When the distribution Q is fixed, it is easy to check that the parameter θ that maximizes $J_\theta(Q, A)$ is given by

$$\hat{\alpha}_k = \frac{\sum_i \tau_{ik}}{n} \quad \hat{\Pi}_{kk'} = \frac{\sum_{i<j} \tau_{ik}\tau_{jk'}A_{ij}}{\sum_{i<j} \tau_{ik}\tau_{jk'}} \quad \forall k, k' \in [K].$$

The VE-step and the M-steps are repeated until convergence of the algorithm.

The consistency of the estimator provided by this VEM-algorithm has been analyzed in Celisse et al., 2011 then extended by Bickel et al., 2013. The analysis has been extended to more general models including SBM with missing observations (Mariadassou et al., 2020; Gaucher et al., 2021), the Latent Block Model (Brault et al., 2020) and dynamic SBM (Longepierre et al., 2019). Whereas the previously mentioned work focus on estimating the model parameters, the accuracy of the method for estimating the partition has been investigated in Zhang et al., 2020a. An alternative to the mean field approximation has been proposed by Yin et al., 2020. The effect of random initialization has been studied in Sarkar et al., 2021.

In practice, the VEM-algorithm can handle graphs with tens of thousands of nodes, but the running time becomes prohibitive for very large networks.

2.2.4 *Modularity based clustering*

A popular way to measure the quality of the partition of a graph into communities is to rely on the quality function Q (Newman et al., 2004) defined by

$$Q = \frac{1}{|E|} \sum_{i,j} (A_{ij} - P_{ij}) \mathbb{1}_{z_i = z_j}$$

where $|E|$ is the number of edges in A, $z_i \in [K]$ denotes the community i belongs to, and P_{ij} correspond to a choice of a null model (a common choice is $\frac{deg(i)deg(j)}{2|E|}$). In general, it is NP-hard to find the partition that minimizes Q. However, there is a popular heuristic to solve this problem: the Louvain algorithm (Blondel et al., 2008). It consists of several steps that are repeated iteratively.

1. Start with an initial partition $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$.

2. For each vertex i and $k \in [K]$ define a partition $\mathcal{C}^{k,i}$ where i is removed from the cluster it was previously associated with and is added to the cluster k.

3. Define $Q^{k,i}$ the modularity function associated with $\mathcal{C}^{k,i}$ minus Q and let be $k_i^* = \arg\max_k Q^{k,i}$ (breaking ties arbitrarily).

4. For all vertices i such that $\mathcal{C}_{z_i} \neq \mathcal{C}_{k_i^*}$ pick a pair (i,j) at random such that $\mathcal{C}_{z_i} = \mathcal{C}_{k*_j}$ and $\mathcal{C}_{z_j} = \mathcal{C}_{k*_i}$, and define a new partition $\mathcal{C}'$ by moving i to $\mathcal{C}_{k*_i}$ and j to $\mathcal{C}_{k_j^*}$.

5. Repeat the previous steps until convergence.

This is a local search algorithm that aims at optimizing the modularity Q. In practice, the algorithm is fast and can be parallelized. Recently, the work of Cohen-Addad et al., 2020 established consistency guarantees of this algorithm under the SBM.

2.2.5 *Methods for partial recovery*

The previously described will usually fail in the very sparse regime where $p_{max} = O(1/n)$. Even if SDP approaches can achieve detection in some regimes, they don't reach the threshold for partial recovery (Ricci-Tersenghi et al., 2016). One of the invoked reason of this failure is that these methods were developed to approximate the Maximum A Posteriori (MAP) (or minimum bisection when $K = 2$), a global criterion that is not suitable for this regime as explained in Abbe, 2018, Section 2.6. Instead of maximizing a global criterion, one should try to maximize the posterior distribution of each singular vertex $\max_{\hat{z_i}} \mathbb{P}(\hat{z_i} = z_i|A)$. One can approximate this posterior distribution by using a belief propagation algorithm. A linear approximation of this algorithm is related to the nonbacktracking operator studied in Bordenave et al., 2018. See also Abbe et al., 2016. Instead of counting nonbacktracking walks, Abbe et al., 2020a showed that one can use spectral method on the new adjacency matrix where two nodes are connected if there is a path of a certain length connecting the two nodes in the original graph. Another simpler matrix related to the nonbacktracking operator is the Bethe-Hessian matrix Saade et al., 2014.

2.3 BEYOND SBM: OTHER RANDOM GRAPHS MODELS

2.3.1 *Degree Corrected SBM (DCSBM)*

Under the SBM assumptions, all the nodes within the same community are exchangeable. But in real-life networks, nodes often have heterogeneous degrees. For example, in social networks, there are often hubs corresponding to very popular people. A way to integrate this

property is to add a degree correction parameters $\theta_i > 0$ for each node $i \in [n]$. A graph $\mathcal{G}$ is distributed according to $\mathrm{DCSBM}(Z, \Theta, \Pi)$ if the entries of the corresponding symmetric adjacency matrix A are generated by

$$A_{ij} \overset{\text{ind.}}{\sim} \mathcal{B}(\theta_i \theta_j \Pi_{z(i)z(j)}), \quad 1 \leqslant i \leqslant j \leqslant n.$$

So in expectation we have $\mathbb{E}(A) = \Theta Z \Pi Z^\top \Theta$ where Θ is the diagonal matrix with $\Theta_{ii} = \theta_i$. For identifiability, we can further assume $\sum_{i \in \mathcal{C}_k} \theta_i = 1$ for all k (if it were different from 1, then we can absorb this factor in the connectivity matrix Π). A particular case of this model is analyzed in Gao et al., 2018. They proposed a feasible two-stage algorithm that achieves the threshold for exact recovery.

2.3.2 *Mixed Membership SBM (MMSBM)*

The assumption that each node belongs to only one community can be too restrictive. For example, in a citation network, it would be more realistic to assume that some articles can cover several topics. To overcome the expressive limitation of the SBM, Airoldi et al., 2008 introduced the MMSBM. Here, each node is associated with a probability distribution over the different communities. This can be interpreted geometrically. Pure nodes – the ones that are associated with only one community – correspond to the extremal points of a polytope, and the positions of the other nodes on this polytope are given by their associated probability distributions that can be interpreted as barycentric coordinates.

More formally, instead of requiring $Z \in \mathcal{M}_K$ we will consider matrices $Z \in \mathcal{M}_K^c$ where $\mathcal{M}_K^c$ is the set of mixed membership matrix defined as follows: $Z \in \mathcal{M}_K^c$ if and only if $Z \in [0,1]^{n \times K}$ and:

- $\|Z_{i:}\|_1 = 1$ for all $1 \leqslant i \leqslant n$;
- for all $k \in [K]$, there is at least one index i such that $Z_{ik} = 1$ (these indices correspond to **pure nodes**).

Other models like the Overlapping SBM (OSBM) (Latouche et al., 2011), the Overlapping Continuous Community Assignment Model (OCCAM) (Zhang et al., 2020b), or SBM with Overlap (SBMO) (Kaufmann et al., 2018) have also been proposed, but we will focus on the MMSBM because it is easier to interpret than OCCAM and more general than the OSBM or SBMO. The MMSBM can also be interpreted in certain cases as a RDPG (Athreya et al., 2018).

2.3.3 *Random Dot Product Graph (RDPG)*

RDPGs are a particular case of latent position graphs where each node is associated with an unknown latent position in some space and

the connectivity probability between two nodes is determined by the distance (or a measure of similarity) between the corresponding latent positions. Formally, let us define $X = (X_1, \dots, X_n)^\top \in \mathbb{R}^{n \times d}$ where X_i is the latent position associated with the node i. Further assume that for each $i, j \in [n], X_i^\top X_j \in [0, 1]$. Then an adjacency matrix A is generated from a $\mathrm{RDPG}(n, X)$ if

$$A_{ij} \overset{\text{ind.}}{\sim} \mathcal{B}(X_i^\top X_j), \quad 1 \leqslant i \leqslant j \leqslant n.$$

RDPG encompasses the DCSBM and MMSBM as special cases. More information can be found in the survey of Athreya et al., 2018.

3

CLUSTERING UNDER THE CONTEXTUAL STOCHASTIC BLOCK MODEL

This chapter is based on Braun et al., 2021a where we introduce a new clustering method for graphs that incorporate nodes side information. In contrast to previously proposed methods, our algorithm comes with strong consistency guarantees under the CSBM and is minimax optimal under the Contextual Symmetric SBM (CSSBM). The theoretical results are confirmed by experiments on synthetic data that show that our algorithm significantly outperforms concurrent methods. We also show that it can be applied to weighted graph and demonstrate the practical interest of our method on real data.

The chapter is organized as follows. In Section 3.1 we introduce the problem and discuss the related work. The statistical framework is presented in Section 3.2 and the algorithms in Section 3.3. We then state our main results and outline the proofs in Section 3.4. Experimental results are presented in Section 3.5. Possible extensions of our framework is discussed in Section 3.6. The additional proofs details are gathered in Section 3.7.

3.1 INTRODUCTION

Real-world networks often come with side information in the form of nodes covariates which can be used to improve clustering performance. For example, when analyzing interactions among people on a social network, we have access to additional features such as gender, age, or ethnicity that can be relevant for the clustering task. Other examples, including biological networks and predator-prey interaction networks are discussed in Newman et al., 2015. The Contextual Stochastic Block Model (CSBM) is a simple extension of the SBM that incorporates such side information: each node is associated with a Gaussian vector of parameters depending only on the community to which the node belongs; see Section 3.2 for details.

Several variants of this model and clustering algorithms have been proposed in the literature to incorporate side information. These methods include model-based approaches (Yang et al., 2013; Weng et al., 2022; Hric et al., 2016; Emmons et al., 2019; Stanley et al., 2019; Contisciani et al., 2020; Fajardo-Fontiveros et al., 2022), spectral methods (Binkiewicz et al., 2017; Mele et al., 2019; Abbe et al., 2020b), modularity based optimization methods (Zhang et al., 2016b), belief propagation Deshpande et al., 2018 and SDP based approaches (Yan et al., 2020). Even if some of these algorithms come with certain theoretical

guarantees, the added value of side information is in general not well understood. The recent works of Abbe et al., 2020b, Lu et al., 2020 and Ma et al., 2021 clarify the situation by establishing information theoretic thresholds for exact recovery and detection in a special case with two communities. However, the algorithm presented in Abbe et al., 2020b is not likely to be extended to a general CSBM with more than two (possibly unequal-sized) communities, while the latter two results focus on detection rather than consistency.

3.1.1 *Our contributions*

We make the following contributions in this paper.

- We propose a new iterative algorithm for clustering networks that is fast and is applicable to various settings including the general CSBM and also signed weighted graphs as shown in experiments.
- The proposed algorithm is theoretically analyzed under the CSBM, and we show that its rate of convergence is statistically optimal under the CSSBM. As a byproduct, we derive the threshold for exact recovery with K communities under the CSSBM, thus extending the recent result of Abbe et al., 2020b which was obtained for $K = 2$.
- We confirm the theoretical properties of our algorithm through experiments on simulated data showing that our method outperforms existing algorithms, not only under the CSBM but also under the Signed SBM the latter of which models community structure in signed networks (Cucuringu et al., 2019). Finally, we provided a real data application of our algorithm.

3.1.2 *Related work.*

As outlined earlier, covariate-assisted clustering methods have been studied from various perspectives and it is outside the scope of this work to provide an exhaustive survey. Here, we will discuss the literature that is most relevant to our work.

Our iterative method can be thought of as a Classification-EM algorithm (Celeux et al., 1992), hereafter referred to as C-EM, where instead of using the likelihood we use a least squares criterion. Such ideas were first applied and analyzed under various models including associative SBM by Lu et al., 2016 and then extended to a general framework by Gao et al., 2019. Recently, such ideas were also successfully applied to the Gaussian Tensor Block Model (Han et al., 2020) and a general GMM (Chen et al., 2021). However, the above results can not be directly applied to the CSBM due to the heterogeneity of the data.

Iterative refinement methods can also be derived naturally from the Power Method (Wang et al., 2021; Ndaoud et al., 2022) or alternative optimization methods (Chi et al., 2019). They have been successfully deployed in other settings as well, e.g., SBM (Wang et al., 2021), group synchronization (Boumal, 2016), joint alignment from pairwise differences (Chen et al., 2018), graph matching (Onaran et al., 2017) and low-rank matrix recovery (Chi et al., 2019).

Despite the huge amount of work on covariate-assisted clustering, there are only limited consistency results. Binkiewicz et al., 2017 and Yan et al., 2020 obtained some weak consistency guarantees that depend on both sources of information, but as noted in Abbe et al., 2020b, those bounds are not tight. Abbe et al., 2020b were the first to propose a method that achieves the threshold for exact recovery, but their algorithm only works when $K = 2$ and it seems difficult to extend it to a more general setting.

3.2 THE STATISTICAL FRAMEWORK

The CSBM consists of a graph encoded in an adjacency matrix $A \in \{0,1\}^{n\times n}$ and nodes covariates forming a matrix $X = [X_1\ X_2 \cdots X_n]^\top \in \mathbb{R}^{n\times d}$ where d is the dimension of the covariate space. The graph and the covariates are generated as follows.

- The **graph** part of the data is generated from a $\mathrm{SBM}(Z,\Pi)$ with $Z \in \mathcal{M}_{n,K}$ as defined in Section 1.3.

We further assume in this chapter that the communities are approximately well balanced and that the sparsity level of the graph p_{max} is of order $\log n/n$.

Assumption A1. *There is a constant α such that for all $k \in [K]$*

$$\frac{n}{\alpha K} \leqslant n_k \leqslant \frac{\alpha n}{K}.$$

This is a common assumption in the literature. The problem becomes harder when the communities have very unbalanced sizes.

Assumption A2. *We will assume throughout that $p_{max} \asymp \log n/n$.*

If the graph is denser, we are in the exact recovery regime and the problem is easy. If we are in a sparser regime, we would need to regularize the adjacency matrix to enforce concentration, or use completely different methods (see e.g. Lu et al., 2020).

For the analysis, we will also consider a special case of the SBM where the communities are equal sized, i.e., $n_k = n/K$ for all $k \in [K]$, and the connectivity matrix is given by

$$\Pi = \begin{pmatrix} p & q & \dots & q \\ q & p & \dots & q \\ \vdots & \vdots & \ddots & \vdots \\ q & q & \dots & p \end{pmatrix} \in [0,1]^{K \times K}.$$

We will further assume that $p = p' \frac{\log n}{n}$ and $q = q' \frac{\log n}{n}$ for constants p', q' such that $p' > q' > 0$. This model will be referred to as the Symmetric SBM (SSBM) and denoted by $SSBM(n, K, p, q)$.

- The **nodes covariates** are generated by a GMM, independent of A conditionally on the partition Z. More formally, for each i,

$$X_i = \mu_{z_i} + \epsilon_i, \text{ where } \epsilon_i \overset{\text{ind.}}{\sim} \mathcal{N}(0, \sigma^2 I_d)$$

  with $\mu_k \in \mathbb{R}^d$ for all $k \leqslant K$ and $\sigma > 0$. We assume that $d = O(n)$.

Remark 2. *For ease of exposition we further assume that σ is known but our method can be extended to anisotropic GMM with unknown variance as in Chen et al., 2021. We also assume K to be known – this assumption is common in the clustering literature. Estimating K is a non-trivial task which is outside the scope of this work, see Jin et al., 2021 for a procedure for SBM. We also leave as further work the incorporation of other forms of covariates, e.g., discrete covariates as in Ahn et al., 2018.*

3.3 HOW TO INTEGRATE HETEROGENEOUS SOURCES OF INFORMATION?

The use of side information should intuitively help to recover clusters that are not well separated on each individual source of information. However, it is not well understood how to integrate two heterogeneous sources of information in the clustering process. Previous attempts (Binkiewicz et al., 2017; Yan et al., 2020) proceed by directly aggregating the adjacency matrix and a Gram matrix (or Kernel matrix) formed by the covariates, but a lot of information can be lost in the aggregation process. Moreover, it is not clear what is the best linear combination of the two matrices. Here, we propose a different approach based on a two step algorithm (see Algorithm 1) that fully exploits all information. In the first step, we obtain a rough estimate of the model parameters from the previous estimate of the partition; the initialization methods that can be used are discussed in Section 3.3.2. Then, in the second step, we iteratively refine the partition, as further explained in Section 3.3.1. In Section 3.5.1 we illustrate via experiments that Algorithm 1 outperforms existing methods for cluster

recovery in the setting where the clusters are insufficiently separated on a single source of information.

Algorithm 1 Iterative Refinement with Least Squares

Input: $A \in \mathbb{R}^{n\times n}$, $X \in \mathbb{R}^{n\times d}$,$K \in \mathbb{N}^*$, $\sigma > 0$, $Z^{(0)} \in \{0,1\}^{n\times K}$ a membership matrix and $T \geqslant 1$.

1: **for** $0 \leqslant t \leqslant T-1$ **do**
2: Given $Z^{(t)}$, estimate the model parameters: $n_k^{(t)} = |\mathcal{C}_k^{(t)}|$, $W^{(t)} = Z^{(t)}(D^{(t)})^{-1}$ where $D^{(t)} = \mathrm{diag}(n_k^{(t)})_{k\in[K]}$, $\Pi^{(t)} = {W^{(t)}}^\top A W^{(t)}$, and $\mu_k^{(t)} = {W_{k:}^{(t)}}^\top X$, for all $k \leqslant K$.
3: Refine the partition by solving for each $i \in [n]$

$$z_i^{(t+1)} = \arg\min_k ||(A_{i:}W^{(t)} - \Pi_{k:}^{(t)})\sqrt{\Sigma_k^{(t)}}||^2 + \frac{||X_i - \mu_k^{(t)}||^2}{\sigma^2}$$

where

$$\Sigma_k^{(t)} = \begin{cases} \mathrm{diag}(\frac{n_{k'}^{(t)}}{\Pi_{kk'}^{(t)}})_{k'\in[K]} & \texttt{(IR-LS)} \\ \frac{\min_{k'} n_{k'}^{(t)}}{\max_{k',k''} \Pi_{k'k''}^{(t)}} I_K & \texttt{(sIR-LS)} \\ \frac{n}{K(p^{(t)}-q^{(t)})} \log(\frac{p^{(t)}(1-q^{(t)})}{q^{(t)}(1-p^{(t)})}) I_K & \texttt{(IR-LSS)} \end{cases}$$

with $p^{(t)} = K^{-1}\sum_{k\in[K]} \Pi_{kk}^{(t)}$ and $q^{(t)} = (K^2-K)^{-1}\sum_{k\neq k'\in[K]} \Pi_{kk'}^{(t)}$.
4: Form the matrices $Z^{(t+1)}$ from $z^{(t+1)}$.
5: **end for**

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

3.3.1 *The refinement mechanism*

At each step t, Algorithm 1 estimates the model parameters given a current estimate of the partition ($W^{(t)}$), then updates the partition by reassigning each node to its closest community. Here, the proximity of a node i to a community k is measured by the distance between its estimated (graph) connectivity profile ($A_{i:}W^{(t)}$) and its covariates (X_i) to the current estimate of the community parameters ($\Pi_{k:}^{(t)}, \mu_k^{(t)}$). Instead of using the MAP estimator as in C-EM algorithms, we use a least-square criterion. In a model-based perspective, this can be interpreted as a Gaussian approximation of the connectivity profile of each node. We will see later in the experiments that this doesn't lead to a loss of accuracy (see Section 3.5), and is also faster (see Table 1).

Different variants of our algorithm are possible depending on the way the variance of each community is estimated and integrated in

the criterion used for the partition refinement. The general method will be referred as `IR-LS`, the simplified spherical version is denoted by `sIR-LS` (it assumes implicitly that each community has the same variance) and the version of the algorithm used for CSSBM is denoted by `IR-LSS`(the optimal choice in this setting to balance the weight attributed to each source of information).

Computational cost. In each iteration, the complexity of estimating the parameters is $O(nnz(A) + nd)$ while that of estimating the partition is $O(nK(K + d))$. So the total cost of `IR-LS` is $O(T(nnz(A) + nK(\max(K, d)))$. In our setting A is sparse, hence $nnz(A) \asymp n \log n$.

Remark 3. *Algorithm 1 can also be used for clustering weighted signed graphs, as shown later in the experiments. Moreover, it is interesting to note that when there are no covariates, the algorithm can be applied to graphs generated from a general SBM. This is in contrast to the iterative algorithm proposed by Lu et al., 2016 that can only be applied to assortative SBMs (see appendix).*

3.3.2 *Initialization*

Different strategies can be adopted for initialization. If we assume that the communities are separated on each source of information and that the Signal-to-Noise Ratio (SNR) is large enough to recover a sufficient proportion in each cluster, we can use a spectral method on one source of information (the graph for example). However, it is in general better to combine both sources of information. While one could use the methods proposed in Yan et al., 2020 or Binkiewicz et al., 2017 that also come with some theoretical guarantees, we instead use Algorithm 2 to initialize the partition. This algorithm will be referred to as `EM-Emb`. In our experiments, we used the package `clusterR` (Mouselimis, 2022) for estimating the Gaussian mixture with an EM algorithm. This algorithm is fast, provides a sufficiently accurate estimate of the partition, and avoids hyperparameter tuning. As shown in the experiments (see Section 3.5.4.2), under some parameters regime, we can also use a random initialization strategy.

3.4 MAIN RESULTS AND ANALYSIS PRINCIPLE

In this section, we analyze the variants `sIR-LS` and `IR-LSS` of Algorithm 1. While it is possible to extend the analysis to `IR-LS`, it would be considerably more technical and tedious due to its non-spherical structure. Hence, we will assume here that the covariance matrix $\Sigma_k^{(t)}$ in Algorithm 1 has the form $\lambda^{(t)} I_k$ where $\lambda^{(t)}$ is an appropriate scalar depending on whether we use `sIR-LS` or `IR-LSS`.

In Section 3.4.1 we will present the general principle for the analysis. Then we will specialize it for analyzing `IR-LSS` (under the CSSBM)

Algorithm 2 EM on graph embedding and covariates (`EM-Emb`)

Input: The number of communities K, the adjacency matrix A, covariates X.

1: Compute $U_K \in \mathbb{R}^{n \times K}$ the matrix formed by the eigenvectors associated with the top-K eigenvalues (in absolute order) of A.
2: Merge the columns of U_K with the columns of X to obtain a matrix X'.
3: Cluster the rows of X' by using an EM algorithm for GMM.

Output: A partition of the nodes $Z^{(0)}$.

in Section 3.4.2, and prove that the convergence rate obtained is optimal in Section 3.4.3. Finally we show that the same framework can be used to bound the convergence rate of `sIR-LS` (under the CSBM) in Section 3.4.4. The details of the proofs are outlined in the appendix.

3.4.1 *Analysis principle*

Our analysis is motivated by the general framework recently developed by Gao et al., 2019, and also borrows some decomposition techniques used for analyzing Gaussian tensors from Han et al., 2020. However, these results are not directly applicable to our setting due to dependencies arising from symmetry in the SBM. Moreover, we need tighter control of the error terms then provided by these works.

We will assume w.l.o.g. that $\sigma = 1$ (since σ is assumed to be known in our framework) and that the permutation π that minimizes the distance between $z^{(0)}$ and z is the identity (if not, then replace z by $\pi^{-1}(z)$). Hence there is no label switching ambiguity in the community labels of $z^{(t)}$ because they are determined from $z^{(0)}$.

The first step is to analyze the event "after one refinement step, the node i will be incorrectly clustered given the current estimation of the partition $z^{(t)}$ at time t". This corresponds to the condition

$$a \neq \arg\min_k \|X_i - \mu_k^{(t)}\|^2 + \hat{\lambda}^{(t)} \|A_{i:} W^{(t)} - \Pi_{k:}^{(t)}\|^2$$

for a node i such that $z_i = a$. One can see that this condition is equivalent to the existence of $b \in [K] \setminus a$ such that

$$\underbrace{\langle \epsilon_i, \tilde{\mu}_a - \tilde{\mu}_b \rangle + \lambda \langle E_{i:} W, \tilde{\Pi}_{a:} - \tilde{\Pi}_{b:} \rangle}_{C_i(a,b)} \leqslant \frac{-\Delta^2(a,b)}{2} + \mathrm{Err}_{ib}^{(t)}.$$

Here,

$$\Delta^2(a,b) = \|\mu_a - \mu_b\|^2 + \lambda \|\Pi_{a:} - \Pi_{b:}\|^2,$$
$$\tilde{\mu}_k = X^\top W_{:a}, \; \tilde{\Pi}_{k:} = W_{k:}^\top A W,$$

$$\text{and } \lambda = \frac{n_{min}}{p_{max}} \text{ or } \frac{n}{K(p-q)} \log\left(\frac{p(1-q)}{q(1-p)}\right)$$

depending whether we are analyzing `sIR-LS` or `IR-LSS`. Moreover, $Err_{ib}^{(t)}$ is an error term that can be further decomposed as a sum $F_{ib}^{(t)} + G_{ib}^{(t)} + H_{ib}^{(t)}$ of different kinds of error terms which will be controlled in different ways. If we ignore the error term, we obtain the condition corresponding to having an incorrect result after one iteration starting from the ground truth partition. The errors occurring in this way will be quantified by the **ideal oracle error**

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

Let us denote

$$\Delta_{min} = \min_{a \neq b \in [K]} \Delta(a, b)$$

to quantify the separation of the parameters associated with the different communities. If $\Delta_{min} = 0$, it would imply that at least two communities are indistinguishable and the model would not be identifiable. For $t \geqslant 1$ and $\delta \in [0, 1)$, let

$$\delta^{(t)} = \max\left(\frac{7}{8}\frac{\tau^{(t-1)}}{\tau^{(0)}}, \delta\right), \ \tau^{(t)} = \tau^{(0)}\delta^{(t)}$$

be sequences where $\tau^{(0)} = \epsilon n \Delta_{min}^2 / K$ for a small enough constant $\epsilon > 0$.

In general the rate of decay of $\xi(\delta)$ leads to the convergence rate of iterative refinement algorithms, hence it is important to control this quantity.

Condition C1 (ideal error)**.** *Assume that*

$$\xi(\delta^{(t)}) \leqslant \frac{3}{4}\tau^{(t-1)}, \text{ for all } t \geqslant 1$$

holds with probability at least $1 - \eta_1$.

We now have to analyze the error terms and prove that their contribution is negligible compared to the ideal oracle error rate. Let

$$l(z, z') = \sum_{i \in [n]} \Delta^2(z_i, z'_i) \mathbb{1}_{\{z_i \neq z'_i\}}$$

be a measure of distance between two partitions $z, z' \in [K]^n$. We will control the error terms by showing that the following conditions are satisfied.

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

$$\max_{\{z^{(t)} : l(z, z^{(t)}) \leqslant \tau^{(0)}\}} \sum_{i=1}^{n} \max_{b \in [K] \setminus z_i} \frac{(F_{ib}^{(t)})^2}{\Delta^2(z_i, b) l(z, z^{(t)})} \leqslant \frac{\delta^2}{256}$$

for all $t \geqslant 0$ *holds with probability at least* $1 - \eta_2$.

Condition C3 (GH-error type). *Assume that*

$$\max_{i\in[n]}\max_{b\in[K]\setminus z_i}\frac{|H_{ib}^{(t)}|+|G_{ib}^{(t)}|}{\Delta(z_i,b)^2}\leqslant\frac{\delta^{(t+1)}}{4}$$

holds uniformly on the event $\{z^{(t)} : l(z,z^{(t)})\leqslant\tau^{(t)}\}$ *for all* $t\geqslant 0$ *with probability at least* $1-\eta_3$.

We can now show under these conditions that there is a contraction of the error if the initial estimate of the partition is close enough to the ground truth partition.

Theorem 4. *Assume that* $l(z^{(0)},z)\leqslant\tau^{(0)}$ *and* $\delta<1$. *Additionally assume that Conditions C1, C2, and C3 hold. Then with probability at least* $1-\sum_{i=1}^{3}\eta_i$

$$l(z^{(t)},z)\leqslant\xi(\delta^{(t)})+\frac{1}{8}l(z^{(t-1)},z),\ \forall t\geqslant 1. \tag{3.4.1}$$

Remark 4. *This is an adaptation of Theorem 3.1 in Gao et al., 2019 where we allow at each step to choose a different* $\delta^{(t)}$. *It allows us to obtain a weaker condition for initialization than the one used in Theorem 4.1 in Gao et al., 2019. Indeed, they require* $l(z^{(0)},z)=o(\frac{n\Delta_{\min}^2}{K})$ *in order to have* $\delta=o(1)$, *but we only need* $l(z^{(0)},z)=O(\frac{n\Delta_{\min}^2}{K})$.

Proof of Theorem 4. By definition $\delta^{(t)}<1$ for all $t\geqslant 1$. Let $i\in[n]$ such that $z_i=a$ and assume that $l(z^{(t-1)},z)\leqslant\tau^{(t-1)}$ for some given $t\geqslant 1$. Denoting $I_i^{(t)}(a,b):=\mathbb{1}_{\{C_i(a,b)\leqslant\frac{-(1-\delta^{(t)})\Delta^2(a,b)}{2}\}}$, observe that

$$\begin{aligned}
&\mathbb{1}_{\{z_i^{(t)}=b\}}\\
&\overset{(1)}{\leqslant}\mathbb{1}_{\{C_i(a,b)\leqslant\frac{-\Delta^2(a,b)}{2}+F_{ib}^{(t-1)}+G_{ib}^{(t-1)}+H_{ib}^{(t-1)}\}}\\
&\overset{(2)}{\leqslant}I_i^{(t)}(a,b)+\mathbb{1}_{\{\frac{\delta^{(t)}}{2}\Delta^2(a,b)\leqslant F_{ib}^{(t-1)}+G_{ib}^{(t-1)}+H_{ib}^{(t-1)}\}}\\
&\overset{(3)}{\leqslant}I_i^{(t)}(a,b)+\mathbb{1}_{\{\frac{\delta^{(t)}}{4}\Delta^2(a,b)\leqslant F_{ib}^{(t-1)}\}}\\
&\leqslant I_i^{(t)}(a,b)+\mathbb{1}_{\{\frac{\delta}{4}\Delta^2(a,b)\leqslant F_{ib}^{(t-1)}\}}\\
&\overset{(4)}{\leqslant}I_i^{(t)}(a,b)+\frac{32(F_{ib}^{(t-1)})^2}{\delta^2\Delta^4(a,b)}.
\end{aligned}$$

The inequality (1) follows from the definition of $z_i^{(t)}$ and the error decomposition. Inequality (2) comes from a union bound while (3) uses Condition C3. Finally, (4) follows from Markov inequality. Hence,

$$\begin{aligned}
l(z^{(t)},z)&=\sum_{i\in[n]}\sum_{b\in[K]\setminus\{z_i\}}\Delta^2(z_i,b)\mathbb{1}_{\{z_i^{(t)}=b\}}\\
&\leqslant\sum_{i\in[n]}\sum_{b\in[K]\setminus\{z_i\}}\Delta^2(z_i,b)\mathbb{1}_{\{C_i(a,b)\leqslant\frac{-(1-\delta^{(t)})\Delta^2(z_i,b)}{2}\}}
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{i\in[n]} \sum_{b\in[K]\setminus\{z_i\}} \Delta^2(z_i,b)\mathbb{1}_{\{z_i^{(t)}=b\}} \frac{32(F_{ib}^{(t-1)})^2}{\delta^2\Delta^4(z_i,b)} \\
&\leqslant \xi(\delta^{(t)}) + \sum_{i\in[n]} \max_{b\in[K]\setminus\{z_i\}} \frac{32(F_{ib}^{(t-1)})^2}{\delta^2\Delta^2(z_i,b)} \\
&\leqslant \xi(\delta^{(t)}) + \frac{1}{8} l(z^{(t-1)},z).
\end{aligned}
$$

Using Condition C1, we hence obtain

$$
l(z^{(t)},z) \leqslant \xi(\delta^{(t)}) + \frac{1}{8}\tau^{(t-1)} \leqslant \frac{7}{8}\tau^{(t-1)}.
$$

Thus $\tau^{(t)}$ is an upper bound for $l(z^{(t)},z)$ and the theorem is proved by induction. □

By iteratively unwrapping (3.4.1) we obtain for t large enough the following bound on $l(z^{(t)},z)$.

Corollary 1. *Assume that the assumptions of Theorem 1 hold. Denoting* $t^* = \lceil \frac{\log(1/\delta)}{\log(8/7)} \rceil$, *we have for all* $t \geqslant t^*$ *that* $\tau^{(t)} = \tau$, $\delta^{(t)} = \delta$ *and*

$$
l(z^{(t)},z) \leqslant \frac{8}{7}\xi(\delta) + \frac{3}{28}\tau^{(0)} \left(\frac{1}{8}\right)^{t-t^*}.
$$

This bound shows that if t is suitably large, then the estimation error is of the order of the oracle error.

Proof of Corollary 1. For convenience, denote $\tau = \tau^{(0)}\delta$. Let t^* be the smallest integer such that $\tau \geqslant (\frac{7}{8})^t\tau^{(0)}$; clearly, $t^* = \lceil \frac{\log(1/\delta)}{\log(8/7)} \rceil$. Then for $t \geqslant t^*$, we have $\tau^{(t)} = \tau = \tau^{(0)}\delta$, and hence (from the definition of $\delta^{(t)}$), $\delta^{(t)} = \delta$. Therefore Theorem 4 implies that for $t \geqslant t^*$,

$$
\begin{aligned}
l(z^{(t)},z) &\leqslant \xi(\delta^{(t)}) + \frac{1}{8}\xi(\delta^{(t-1)}) + \ldots + \left(\frac{1}{8}\right)^{t-t^*} \xi(\delta^{(t^*)}) \\
&\quad + \left[\left(\frac{1}{8}\right)^{t-t^*+1} \xi(\delta^{(t^*-1)}) + \ldots + \left(\frac{1}{8}\right)^{t-1} \xi(\delta^{(1)})\right] \\
&\leqslant \frac{8}{7}\xi(\delta) + \left[\left(\frac{1}{8}\right)^{t-t^*+1} \xi(\delta^{(t^*-1)}) + \ldots + \left(\frac{1}{8}\right)^{t-1} \xi(\delta^{(1)})\right] \\
&\qquad\qquad \text{(since } \delta^{(t)} = \delta \text{ for } t \geqslant t^*\text{)} \\
&\leqslant \frac{8}{7}\xi(\delta) + \frac{3}{4}\tau^{(0)} \left[\left(\frac{1}{8}\right)^{t-1} + \ldots + \left(\frac{1}{8}\right)^{t-t^*+1}\right] \\
&\qquad\qquad \text{(since for all } t,\ \xi(\delta^{(t)}) \leqslant \tfrac{3}{4}\tau^{(0)} \text{ by Condition C1)} \\
&\leqslant \frac{8}{7}\xi(\delta) + \frac{3}{28}\tau^{(0)} \left(\frac{1}{8}\right)^{t-t^*}.
\end{aligned}
$$

□

3.4.2 *Convergence guarantees for `IR-LSS` under the CSSBM*

Let us define the SNR

$$\tilde{\Delta}^2 = \frac{1}{8} \min_{k \neq k'} \|\mu_k - \mu_{k'}\|^2 + \frac{\log n}{K} (\sqrt{p'} - \sqrt{q'})^2.$$

It is easy to see that $\tilde{\Delta} \asymp \Delta_{min}$. The following lemma shows that $\xi(\delta)$ decreases exponentially fast in $\tilde{\Delta}$ provided Δ_{min} is suitably large.

Lemma 1. *Assume that* $K^{1.5}/\Delta_{min} \to 0$ *and* $\delta = \delta(n) \to 0$ *at a suitably slow rate. Then with probability at least* $1 - \exp(-\tilde{\Delta})$*, we have*

$$\xi(\delta) \leqslant n \exp(-(1 + o(1))\tilde{\Delta}^2).$$

The following theorem shows that if $z^{(0)}$ is close enough to z, then the misclustering rate decreases exponentially fast with the SNR $\tilde{\Delta}$ after $O(\log n)$ iterations.

Theorem 5. *Assume that* $K^{1.5}/\Delta_{min} \to 0$ *and* $\tilde{\Delta}^2 \asymp \log n/K$*. Under the CSSBM*$(n, K, p, q)$ *assumption, if* $z^{(0)}$ *is such that*

$$l(z, z^{(0)}) \leqslant \frac{\epsilon n \Delta_{min}^2}{K}$$

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

$$r(z^{(t)}, z) \leqslant \exp(-(1 + o(1))\tilde{\Delta}^2).$$

Sketch of proof. We first show that Conditions C2 and C3 are satisfied. Then we show that Condition C1 is satisfied for the sequences $\delta^{(t)}$ and $\tau^{(t)}$, hence Theorem 4 can be applied to obtain a contraction of the error at each step. □

Remark 5. *By assumption* $\tilde{\Delta}^2 \gtrsim \log n/K$*, and so the condition* $\tilde{\Delta}^2 \asymp \log n/K$ *is not very restrictive. Indeed, if the information provided by the GMM part was not of the same order as the graph part, it would not be useful to aggregate information. If* $\tilde{\Delta}^2 \gg \log n$ *then we would be in the exact recovery setting and the problem becomes easy.*

Remark 6. *The initial condition implies that* $h(z^{(t)}, z)/n \leqslant \epsilon/K$ *where* h *denotes Hamming distance, see appendix for details. This is a detection condition. It is outside the scope of this work to analyze an algorithm that achieves this condition under CSBM. The numerical experiment done in Section 3.7 suggests that the algorithm can still work with a random initialization, at least in particular cases.*

3.4.3 *Minimax lower-bound for CSSBM*

We are going to establish that the convergence rate established in Theorem 5 is optimal. Let

$$\Theta = \{(\mu_k)_{k\in[K]} \in \mathbb{R}^K, p, q \in [0,1] \text{ such that } p > q\}$$

be the admissible parameter space.

Theorem 6. *Under the assumption* $\tilde{\Delta}/\log K \to \infty$*, we have*

$$\inf_{\hat{z}} \sup_{\theta\in\Theta} \mathbb{E}(r(\hat{z}, z)) \geqslant \exp(-(1+o(1))\tilde{\Delta}^2).$$

If $\tilde{\Delta} + \log K = O(1)$*, then* $\inf_{\hat{z}} \sup_{\theta\in\Theta} \mathbb{E}(\frac{r(\hat{z},z)}{n}) \geqslant c$ *for some positive constant* c*.*

Remark 7. *This lower-bound shows that if* $\tilde{\Delta}^2 < \log n$ *then every estimator fails to achieve exact recovery with a probability bounded below from zero because* $\sup_{\theta\in\Theta} \mathbb{E}(r(\hat{z}, z)) > n^{\epsilon}$ *for some* $\epsilon > 0$*. On the other hand, Theorem 5 shows that when* $\tilde{\Delta}^2 > \log n$ *then* `IR-LSS` *achieves exact recovery. Hence the threshold for exact recovery is* $\tilde{\Delta}^2/\log n$*. When* $K = 2$ *and* $\mu_1 = -\mu_2 = \mu$ *this matches the result obtained by Abbe et al., 2020b.*

Sketch of proof. We can use the same argument as in Theorem 3.3 Lu et al., 2016 to reduce the problem to a hypothesis testing problem. The solution of the latter is given by the maximum likelihood test according to the Neyman-Pearson lemma. Then the probability of error can be controlled by using concentration inequalities. □

3.4.4 *Convergence guarantees for* `sIR-LS` *under the CSBM*

The proof techniques used in the previous section can be extended in a straightforward way to obtain consistency results for `sIR-LS` under the CSBM. The main difference is that the specialized concentration inequality used to prove Lemma 1 can no longer be applied to this setting.

Theorem 7. *Assume that*

$$\frac{K^{1.5}}{\Delta_{min}} \to 0,\ \Delta^2_{min} \asymp \log n/K \text{ and } \max_{a,b\in[K]} \Delta^2(a,b) \lesssim \Delta^2_{min}.$$

Under the CSBM with approximately balanced communities, if for some small enough constant $\epsilon > 0$*,* $l(z, z^{(0)}) \leqslant \frac{\epsilon n \Delta^2_{min}}{K}$*, then with probability at least* $1 - n^{-\Omega(1)}$ *we have*

$$r(z^{(t)}, z) \leqslant \exp\left(-\frac{1}{8}\Delta^2_{min}\right) \text{ for all } t \gtrsim \log n.$$

3.4.5 *Comparison with existing theoretical results*

Abbe et al., 2020b obtain the same bound as us when $K = 2$. Their method – which is an aggregated spectral method – requires computing the largest eigenvector of $XX^\top$ (with diagonal set to zero) which has a complexity $O(n^2)$. In contrast, our method has complexity $O(n \log n + nd)$ and is thus faster when $d = o(n)$. Binkiewicz et al., 2017 consider a spectral method applied on a regularized Laplacian. When $d \asymp \text{polylog}(n)$ they show that the misclustering rate is $O(1/\text{polylog}(n))$. Yan et al., 2020 used a similar regularization idea but with a SDP and obtain an error bound (in Frobenius norm) for estimating a clustering matrix. Their bound depends on the two sources of information, but as also noted by Abbe et al., 2020b, it is unclear how the bound improves with side information. Moreover the bounds in Binkiewicz et al., 2017; Yan et al., 2020 are not optimal.

3.5 NUMERICAL EXPERIMENTS

We now empirically evaluate our method on both synthetic and real data[1]. Section 3.5.1 contains simulations for the CSBM and Section 3.5.2 contains results for clustering signed networks under a Signed SBM. In Section 3.5.3, we test our method on a dataset consisting of a (weighted) signed graph along with covariate information for the nodes.

3.5.1 *CSBM with not well separated communities*

In this experiment the graph is generated from a SBM with parameters $n = 1000$, $K = 3$, $Z_i \overset{\text{i.i.d}}{\sim} \text{Multinomial}(1; 1/3, 1/3, 1/3)$, and

$$\Pi = 0.02 * \begin{pmatrix} 1.6 & 1.2 & 0.05 \\ 1.2 & 1.6 & 0.05 \\ 0.05 & 0.05 & 1.2 \end{pmatrix}.$$

The covariates are generated from a GMM with variance $\sigma^2 = 0.2$ and class centers $\mu_1 = (0, 0, 1), \mu_2 = (-1, 1, 0), \mu_3 = (0, 0, 1)$. Note that $\mathcal{C}_1, \mathcal{C}_3$ cannot be separated by the covariate information, while $\mathcal{C}_1, \mathcal{C}_2$ are not well separated in the graph information (as seen from Π). Hence, one would expect in this example that using only a single source of information should not yield good clustering results. To demonstrate this, we use the Normalized Mutual Information (NMI) criterion to measure the quality of the resulting clusters. It is an information theoretic measure of similarity taking values in $[0, 1]$, with 1 denoting a perfect match, and 0 denoting the absence of correlation between partitions.

1 The source code is available on https://github.com/glmbraun/CSBM

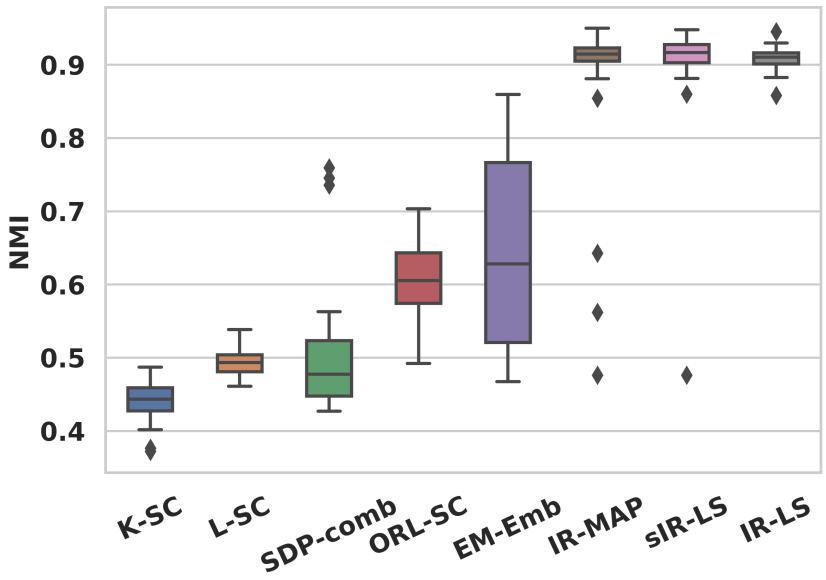

Figure 1: Average performance over 40 runs of different algorithms under CSBM.

3.5.1.1 *Performance comparison*

We will use `K-SC` and `L-SC` to denote respectively the results obtained by applying spectral clustering on the Gaussian kernel matrix K formed from the covariates, and spectral clustering (SC) applied on the Laplacian of the graph. Additionally, `SDP-Comb` refers to the method proposed by Yan et al., 2020; `IR-MAP` is similar to `IR-LS` but with the least-square criterion replaced by the MAP to update the partition; `ORL-SC` (Oracle Regularized Laplacian SC) corresponds to SC applied on $A + \lambda K$ where λ is chosen to maximize the NMI between the (oracle known) true partition and the one obtained by using SC on $A + \lambda K$. For the implementation of `SDP-Comb`, we used the Matlab code provided by Yan et al., 2020 with λ given by `ORL-SC`. We limited our comparison to the aforementioned methods for concreteness; a comparison with all existing methods from the literature would need a separate study and is outside the scope of the paper.

Figure 1 shows that the three iterative methods considered (`IR-MAP`, `IR-LS`, `sIR-LS`), initialized with `EM-Emb`, provide significantly better clustering performance compared to the other methods. The variance of `sIR-LS` is a bit larger than `IR-LS`. On the other hand, other methods based on aggregating the two sources of information (`SDP-Comb` and `ORL-SC`) lead to a limited improvement in clustering performance. Additional experiment results in the appendix suggest that the iterative methods considered also work with random initialization.

3.5.1.2 *Computational cost*

We took the average of CPU time (in seconds) over 20 repetitions. There is an important gain in speed obtained by replacing the MAP objective by a least square criterion. Moreover, the initialization obtained with `EM-Emb` is very fast. The results are gathered in Table 1.

	`L-SC`	`ORL-SC`	`EM-Emb`	`IR-LS`	`IR-MAP`
Time	1.4	7.9	0.5	1.2	37.1
Ratio	2.7	15	1	2.3	70

Table 1: Comparison between computation times (averaged over 20 runs).

3.5.2 *Signed SBM*

A graph is generated from the Signed SBM as follows. First we generate an Erdös-Renyi graph where each edge appears with probability p and each edge takes the value 1 if both extremities are in the same community and -1 otherwise. Then we flip the sign of each edge independently with probability $\eta \in [0, 1/2)$. Our method `sIR-LS` can be directly applied to this setting, but we can also use the fact that the connectivity matrix Π is assortative to design a more specialised algorithm `IR-SSBM` (see Algorithm 3) that assigns a node to the community which maximizes its intra-connectivity estimated probability. For initialization, we use `Sponge-sym` (Cucuringu et al., 2019) for clustering signed graphs. Figure 2 shows that 20 refinement steps improve the clustering.

Algorithm 3 `IR-SSBM`

Input: The number of communities K, initial partition $z^{(0)}$, $T \geqslant 1$.

1: **for** $0 \leqslant t \leqslant T-1$ **do**
2: Compute $W^{(t)} = Z^{(t)}(D^{(t)})^{-1}$ where $D^{(t)} = \mathrm{diag}(n_k^{(t)})_{k\in[K]}$, and $C^{(t)} = AW^{(t)}$.
3: Update the partition for each $i \leqslant n$

$$z_i^{(t+1)} = \arg\max_k C_{ik}^{(t)}$$

4: **end for**

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

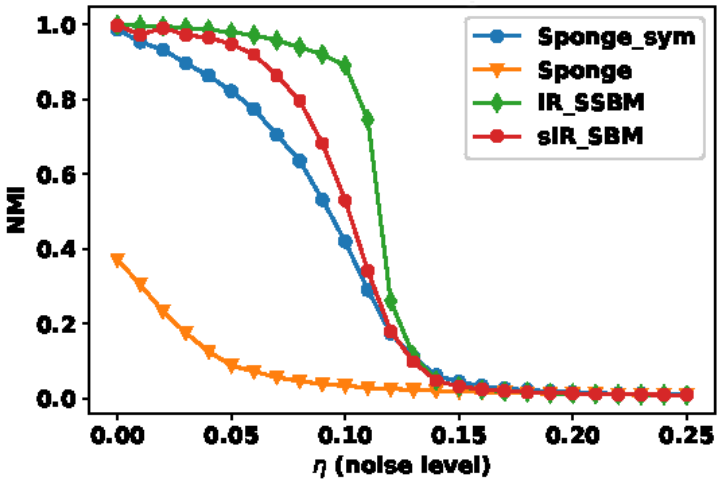

Figure 2: NMI versus η (noise) under signed SBM, $K = 20$, $n = 10000$, $p = 0.01$.

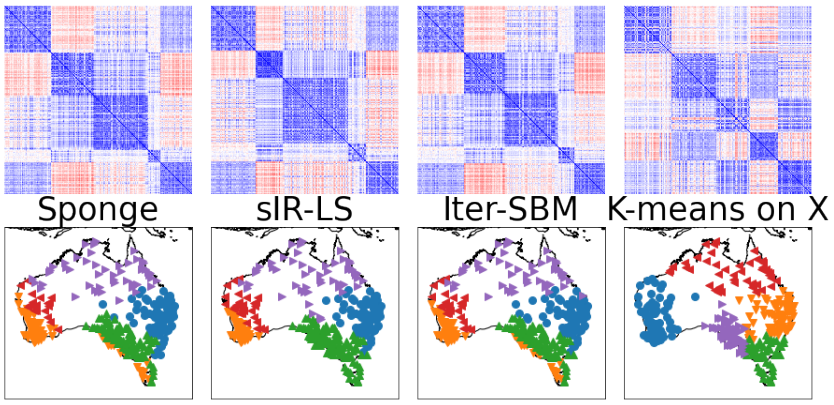

Figure 3: Sorted adjacency matrices and maps for Australian rainfall dataset ($K = 5$).

3.5.3 *Australia Rainfall Dataset*

We consider the time series data of historical rainfalls in locations throughout Australia, this was also studied in Cucuringu et al., 2019. Edge weights are obtained from the pairwise Pearson correlation, leading to a complete signed graph on $n = 306$ nodes. We use the longitude and latitude as covariates X, and `Sponge` (Cucuringu et al., 2019) to obtain an initial partition for `sIR-LS` and `Iter-SBM` (the version of `sIR-LS` without covariates). We exclude `IR-LS` here due to its relative instability on this dataset (see appendix). This shows that in some situations it can be better to use `sIR-LS` rather than `IR-LS`. Figure 3 illustrates the clustering obtained with `Sponge` (using only the graph), `sIR-LS` (integrating the covariates), `Iter-SBM` (refinement without covariates), and `K-means` applied on the covariates. The use

of covariates in the refinement steps reinforces the geographical structure (orange points in the bottom right part of the map disappeared), increases the size of the smallest cluster (the violet cluster on the three first maps), and strengthens the original clustering as seen in the sorted adjacency matrix, whereas `Iter-SBM` ignores the geography and `K-means` ignores the graph structure. We reproduced the experiment for different choices of $K \in \{3, 7, 10\}$ in Figure 4, 5 and 6

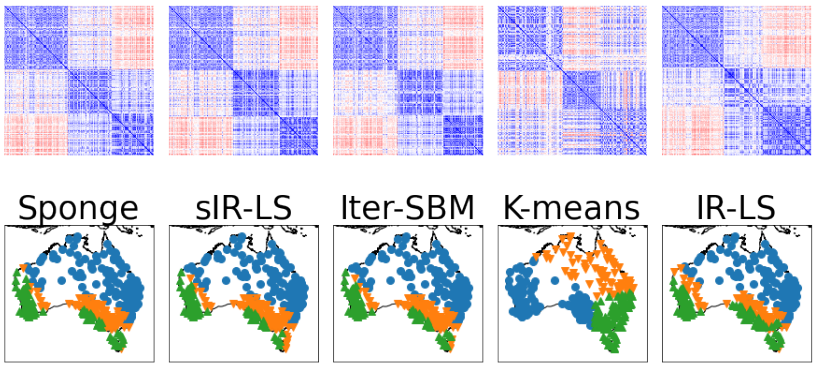

Figure 4: Sorted adjacency matrices of the Australian rainfall data set and corresponding maps for $K = 3$.

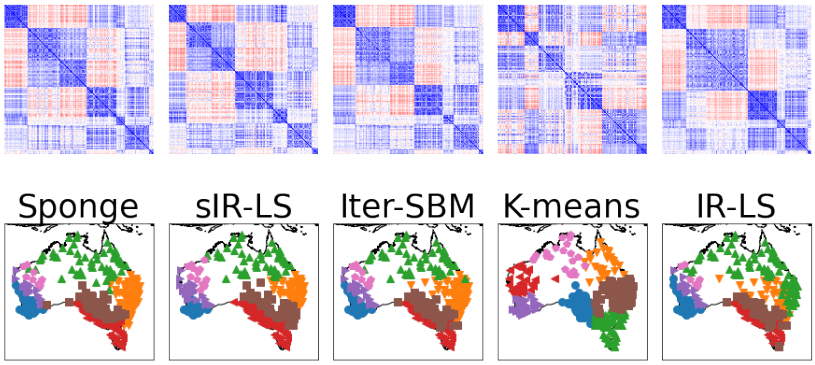

Figure 5: Sorted adjacency matrices of the Australian rainfall data set and corresponding maps for $K = 7$.

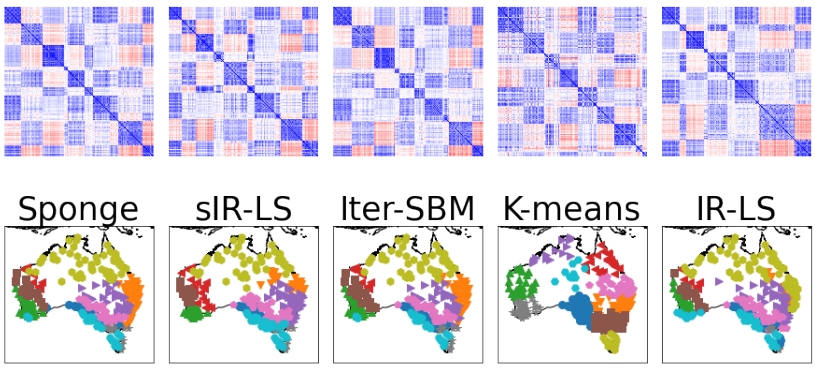

Figure 6: Sorted adjacency matrices of the Australian rainfall data set and corresponding maps for $K = 10$.

3.5.4 *Random initialization*

In this subsection, we show that under some parameters regime, our method can work in practice with a random initialization. But in general it seems better to use `EM-Emb`.

3.5.4.1 *Comparison between `IR-MAP` and `IR-LS`*

In this experiment we fix $K = 2$ and $n = 1000$. Let $c > 0$ and define $p = 4c \log n/n$ and $q = c \log n/n$. Also define a Gaussian mixture in $\mathbb{R}$ with centers 1 (for community 1) and 2 (for the second community). The variance parameter σ is chosen such that the information provided by the Gaussian part of the model is equal to the one provided by the graph. More precisely, we set $\sigma^2 = \frac{1}{2c \log n}$. The value $c = 0.5$ corresponds to the threshold for exact recovery.

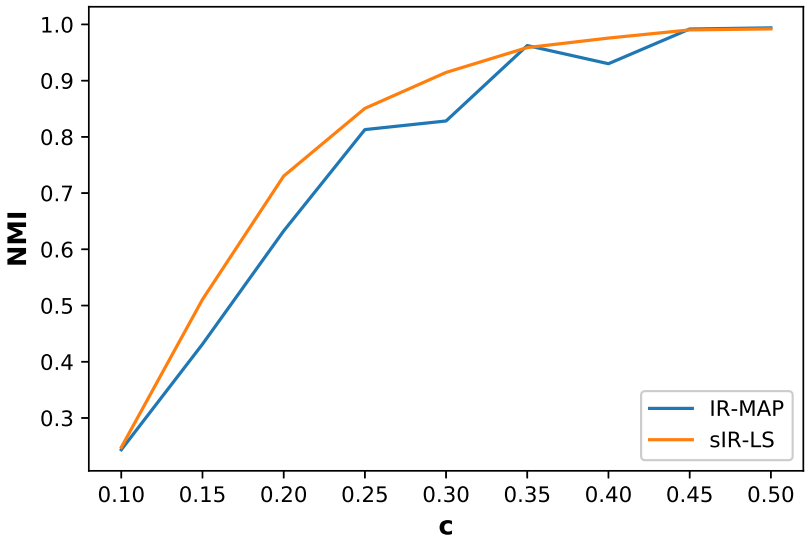

Figure 7: Average performance measured by NMI obtained with random initialization over 20 runs.

Figure 7 shows that when initialized with a partition $Z^{(0)}$ where each entry is generated independently from a Multinomial$(1; 1/2, 1/2)$, `sIR-LS` performs slightly better than `IR-MAP`, has less variability, and is robust to random initialization, hence justifying the Gaussian approximation.

3.5.4.2 *Comparison between random initialization and initialization with `EM-Emb`*

We use a similar experimental setting as the one described in Section 5.1. We only slightly change the connectivity matrix

$$\Pi = 0.02 * \begin{pmatrix} 1.6 & 1.2 & 0.5 \\ 1.2 & 1.6 & 0.5 \\ 0.5 & 0.5 & 1.2 \end{pmatrix}.$$

Figure 8 shows that when randomly initialized ours algorithms `IR-LS` and `sIR-LS` can suffer from numerical instability. That's why we recommend to use `EM-Emb` instead. But it could interesting to develop a strategy based on random initialization by identifying and disregarding the random initialization that lead to atypical results.

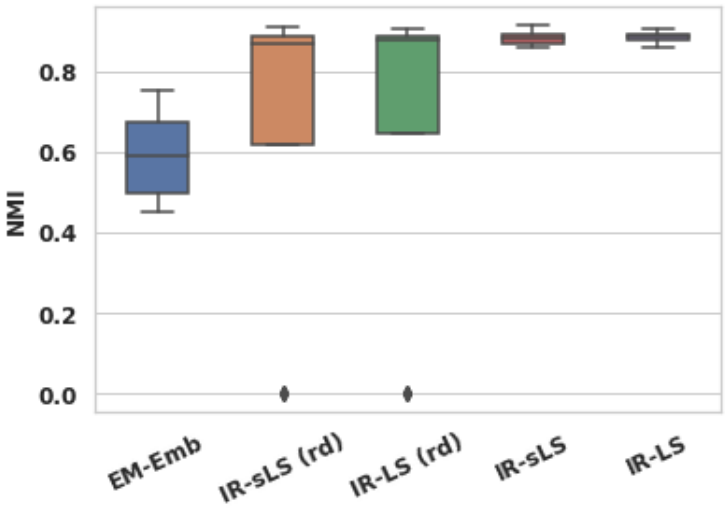

Figure 8: Average performance over 20 runs of our algorithms under the experimental setting of Section 5.1. Here (rd) indicates random initialization.

3.5.5 *Heterophilic SBM*

If we disregard the covariates, our algorithm can be used for inference under a general SBM, in contrast to the method proposed by Lu et al., 2016 which was restricted to the assortative setting. In particular,

our algorithm also works for networks with heterophilic communities. The following experiment illustrates the gain in term of accuracy for `IR-LS` initialized with `A-SC` (spectral clustering on the adjacency matrix). It also shows the interest of using more than one iteration in the refinement step with the MAP (this corresponds to `IR-MAP(1)`).

We consider $n = 1000, K = 3, Z_i \overset{i.i.d}{\sim} \text{Multinomial}(1; 1/3, 1/3, 1/3)$ and

$$\Pi = \begin{pmatrix} 0.2 & 0.05 & 0.1 \\ 0.05 & 0.15 & 0.05 \\ 0.1 & 0.05 & 0.03 \end{pmatrix}.$$

The NMI is averaged over 40 repetitions; the results are shown in Figure 9. We also considered the Variational Expectation Maximization (VEM) algorithm implemented in the R package `blockmodels` (Léger, 2016), but the running time was prohibitive (approximately one hour for a single Monte Carlo run, whereas our algorithm takes a few seconds). It nevertheless returned the exact partition as `IR-LS`.

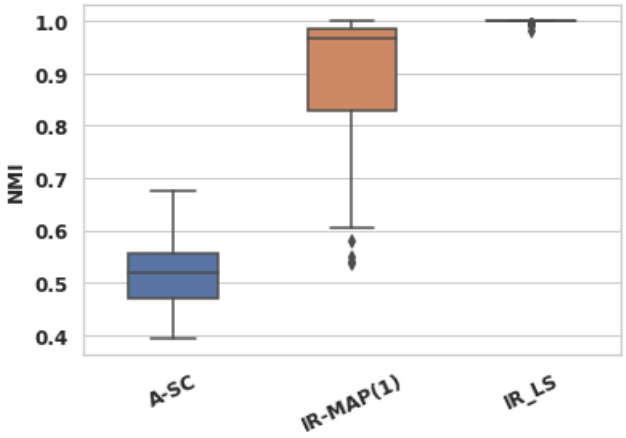

Figure 9: Average performance of different algorithms on a heterophilic SBM, over 40 Monte Carlo runs.

3.6 POSSIBLE EXTENSIONS AND FUTURE WORK

Although commonly used as a benchmark for clustering on graphs, the SBM doesn't encode several properties often observed in real-life networks such as degree heterogeneity, mixed-membership, power-law, and so on. In this section, we will discuss how the method we proposed can be extended to other models.

3.6.1 *Degree-corrected SBM*

The DCSBM has been defined in Section 2.3.1. Due to the degree heterogeneity, the refinement mechanism used in `IR-LS` won't work. But it can be fixed in the following manner. Due to the property that

$Z^\top \Theta W = I_K$ since $\sum_{i \in \mathcal{C}_k} \theta_i = 1$ for all k, we can still estimate $\Pi^{(t)}$ by $W^{(t)\top} A W^{(t)}$. But now we have to take into account the fact that each node connectivity profile $A_{i:} W^{(t)}$ scale differently. A solution could be to update according to the following rule

$$\arg \min_{k \in [K], \theta \in \mathbb{R}} \left\| A_i W^{(t)} - \theta \Pi^{(t)}_{k:} \right\|^2 .$$

It can be solved by testing all the possibilities for k and computing the corresponding θ (solution of a least square problem).

A similar idea using angles (but that could be rewritten in this form) has been proposed by the independent work of Hu et al., 2022 in the same time as we released the present work. They used this method for clustering tensors with heterogeneous degrees and their results don't apply directly to DCSBM but we believe that one could show that by incorporating in this way nodes heterogeneity in the refinement process, the output of the algorithm is minimax optimal under the model considered in Gao et al., 2018.

3.6.2 *Mixed-Membership SBM*

We can also adapt `IR-LS` to MMSBM (see Section 2.3.2 for the definition of the model) by Algorithm 4. This leads to an algorithm close to the one proposed by Arroyo et al., 2021, but that doesn't involve hyper-parameter tuning.

3.6.2.1 *Algorithm description*

Algorithm 4 Generalized Power Method (`GPM`)

Input: $A \in \mathbb{R}^{n \times n}$, $K \in \mathbb{N}^*$, $Z^{(0)} \in [0,1]^{n \times K}$ a mixed membership matrix, a threshold parameter δ and $T \geqslant 1$, the number of iterations.

1: **for** $0 \leqslant t \leqslant T-1$ **do**
2: Given $Z^{(t)}$, form the matrix $W^{(t)} = Z^{(t)}(Z^{(t)\top} Z^{(t)})^{-1}$ estimate the model parameters: $\Pi^{(t)} = W^{(t)\top} A W^{(t)}$
3: Refine the partition by solving for each $i \in [n]$

$$Z_i^{(t+1)} = \mathcal{P}_{\mathcal{C}}(A_i W^{(t)} {\Pi^{(t)}}^{-1}) \tag{3.6.1}$$

where $\mathcal{C} = \{x \in \mathbb{R}_+^K : \|x\|_1 = 1\}$.
4: Form the matrices $Z^{(t+1)}$ from $(Z_i^{(t+1)})_i$.
5: **end for**

Output: $Z^{(T)}$.

At each step t, Algorithm 4 first estimates the model parameters ($\Pi^{(t)}$) given a current estimate of the partition ($W^{(t)}$), and then updates the

latent position of each node by projecting $A_i W^{(t)} \Pi^{(t)^{-1}}$ onto the simplex $\mathcal{C}$. It can be shown using some properties of the pseudo-inverse that the matrix $\tilde{T}^t$ in the work of Arroyo et al., 2021 corresponds to $AW^{(t)}\Pi^{(t)^{-1}}$. This latter way of writing $\tilde{T}^t$ can be interpreted in a easy way: $A_i W^{(t)}$ corresponds to the estimated connectivity profile of node i. In the noiseless case it would be equal to $P_i W = Z_i \Pi$. So in order to obtain Z_i, it only remains to multiply by right with the inverse of Π.

The main difference with the GPM presented in Arroyo et al., 2021 lies in the projection step. We directly project on the simplex, whereas Arroyo et al., 2021 first applies a thresholding operator and then normalize the rows of the resulting matrix. It is unclear to us if the composition of these operations is a projection. Moreover, due to a difference in the model considered, we use the l_1 norm to impose a constraint on the rows of Z instead of the l_2 norm.

COMPUTATIONAL COMPLEXITY. Computing $Z^{(t)\top} Z^{(t)}$ has a complexity $O(K^2 n)$ and the inverting a $K \times K$ matrix has complexity $O(K^3)$, so this operation is negligible. Hence, computing $W^{(t)}$ has a complexity $O(nK^2)$. The complexity of estimating $\Pi^{(t)}$ is $O(\texttt{Knnz}(A) + K^2 n)$. Under the sparsity assumption $p_{max} \asymp \log^c n / n$, $c > 1$, we have $nnz(A) \asymp n \log^c n$ and since $K \ll \log n$ the dominant term in the complexity is $\texttt{Knnz}(A)$. So computing $AW^{(t)}\Pi^{(t)^{-1}}$ has complexity $O(\texttt{Knnz}(A))$. Finally, since there are T iterations we get a total complexity of $O(T\texttt{Knnz}(A))$. Usually T is of the order $\log n$, so in a relatively sparse regime, the computational cost of the refinement procedure is $O(n\text{polylog}(n))$. As shown in the experiments, there are fast spectral methods to get a rough initial estimate $Z^{(0)}$, but still good enough to lead to accurate estimate after applying `GPM`.

3.6.2.2 *Numerical experiments*

We considered the following methods to compare with: `SPACL` (Mao et al., 2021b) a fast spectral method, `M-SCORE` (Jin et al., 2017) a method based on a vertex hunting algorithm that is shown to outperform previously proposed methods and `sPCA` (Arroyo et al., 2021) an algorithm based on a generalization of the power method.

We didn't add the method proposed by Mao et al., 2018 since the authors only provide a Matlab implementation of their algorithm requiring a licence to run. We also tried the method proposed by Majmudar et al., 2020 with inputs A and the normalized correlation matrix $D^{-1/2}A^2D^{-1/2}$[2], but the algorithm didn't return any estimate of the latent positions.

2 The algorithm proposed by Majmudar et al., 2020 is for weighted adjacency matrix, and the matrix $D^{-1/2}A^2D^{-1/2}$ can be interpreted as a Gram matrix

The accuracy of the clustering is measured by

$$\|\hat{Z} - Z\|_1 := \min_{\sigma \in \mathfrak{S}_K} \sum_i \|\hat{Z}_{i:}\sigma - Z_{i:}\|_1$$

where $\mathfrak{S}_K$ is the set of permutation matrices in $[0,1]^{K\times K}$. The permutation that best aligns $\hat{Z}$ and Z is found by solving a Linear Assignment problem with the cost matrix $C \in \mathbb{R}^{K\times}$ defined by $C_{kk'} = \|Z_{:k} - \hat{Z}_{:k'}\|$.

The latent positions (Z_i) are generated by independent Dirichlet random variables with parameter $\alpha = (1/K, \dots, 1/K)$. We then apply $\mathcal{T}_{1/K}$ so that $\|Z_i\|_1 = 1$. The connectivity matrix Π as the symmetric form $\Pi = (p-q)\mathbf{1}_K\mathbf{1}_K^\top + qI_K$ where $p > q > 0$. In the following, the parameter δ of our algorithm is set to $1/(1.1K)$.

INITIALIZATION We considered two methods for initializing our algorithm. First we used `SCORE+`, a spectral method proposed by Jin et al., 2022 for clustering SBM. When initialized with this method, Algorithm 4 will be referred as `IR-SCORE`. Since `SCORE+` is not designed for MMSBM, we also considered the fast spectral method `SPACL` (Mao et al., 2021b) to initialize Algorithm 4. In this case, the algorithm is referred as `IR-SPACL`.

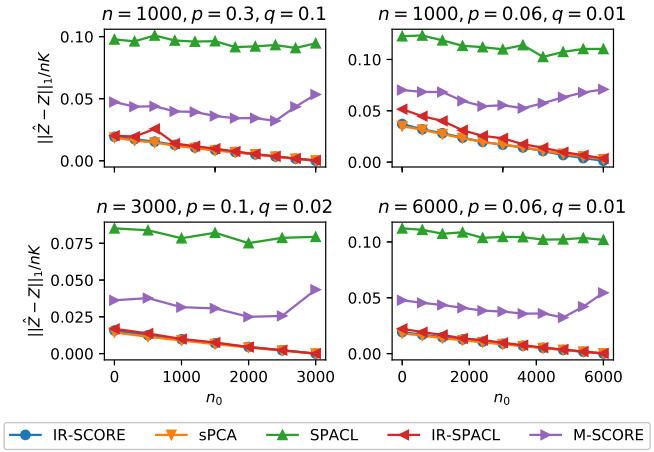

Figure 10: Estimation error associated to a MMSBM with parameters $n, K = 3, p, q$ and n_0 (x-axis) averaged over 20 runs.

INFLUENCE OF THE NUMBER OF ANCHOR NODES For this experiment we fixed $K = 3$. As shown in Figure 10, `SPACL` is not sensitive to the number of pure nodes. While having better performances, `M-SCORE` suffers from the same drawback. In particular, its performances deteriorate when almost all the nodes are pure. On the other hands, the iterative methods `IR-MM`, `IR-SPACL` and `sPCA` have better performances and achieve exact recovery when there are only pure nodes. Even in the case when $n_0 = 0$ our methods perform well

suggesting that the assumption that there exists pure node is not necessary provided that the columns of Z are enough separated.

INFLUENCE OF THE NUMBER OF COMMUNITIES Since the computational complexity of `M-SCORE` is exponential in the number of communities, we didn't integrate this algorithm in this experiment. For this experiment we fix $n = 5000$. Figure 11 shows that `IR-MM` has the best performances when the number of communities increases. When the number of communities is small, whether we initialize our algorithm with `Score+` or `SPACL` we get the same result. On the other hand, if the number of communities is large and the latent positions not enough separated (n_0 small) iterative methods can suffer from numerical instability and does no longer provide accurate estimates.

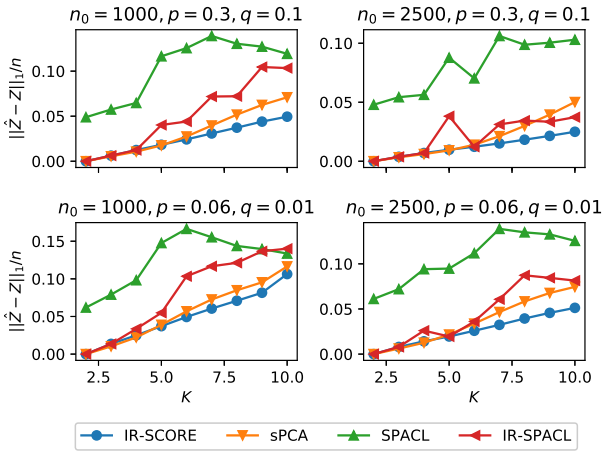

Figure 11: Estimation error associated to a MMSBM with parameters $n = 5000$, K (x-axis), p, q and n_0 averaged over 20 runs.

3.6.3 *Other extensions*

In Chapter 4 we study a setting where partial contextual information is provided by multilayer graphs. The iterative strategy considered in this chapter can be adapted to this setting, see Section 4.7. Instead of considering Gaussian covariates, we can have access to discrete, high-dimensionnal covariates. A first step to extend the method to this setting is presented in Chapter 5.

It would be also interesting to further study when random initialization based strategies could work and develop variant of `IR-LS` robust to missing values and outliers.

3.7 ADDITIONAL PROOFS

3.7.1 *Proof of Lemma 1*

Let $\delta, \bar{\delta} > 0$. The ideal oracle error term can be upper bounded as follows

$$\xi(\delta) \leqslant \underbrace{\sum_{i=1}^{n} \sum_{b \in [K] \setminus z_i} \Delta^2(z_i, b) \mathbb{1}_{\{\langle \epsilon_i, \mu_{z_i} - \mu_b \rangle + \lambda \langle E_{i:} W, \Pi_{z_i:} - \Pi_{b:} \rangle \leqslant \frac{-(1-\delta-\bar{\delta})\Delta^2(z_i,b)}{2}\}}}_{M_1}$$
$$+ \underbrace{\sum_{i=1}^{n} \sum_{b \in [K] \setminus z_i} \Delta^2(z_i, b) \mathbb{1}_{\{\langle \epsilon_i, \tilde{\mu}_{z_i} - \mu_{z_i} \rangle + \lambda \langle E_{i:} W, \tilde{\Pi}_{z_i:} - \Pi_{z_i:} \rangle \leqslant \frac{-\bar{\delta}\Delta^2(z_i,b)}{4}\}}}_{M_2}$$
$$+ \underbrace{\sum_{i=1}^{n} \sum_{b \in [K] \setminus z_i} \Delta^2(z_i, b) \mathbb{1}_{\{\langle -\epsilon_i, \tilde{\mu}_b - \mu_b \rangle - \lambda \langle E_{i:} W, \tilde{\Pi}_{b:} - \Pi_{b:} \rangle \leqslant \frac{-\delta\Delta^2(z_i,b)}{4}\}}}_{M_3}.$$

We will first obtain upper bounds for each $\mathbb{E}(M_i)$, $i = 1 \dots 3$. In particular we will show that the dominant term is $\mathbb{E}(M_1)$. Then, we will use Markov inequality to control $\xi(\delta)$ with high probability.

UPPER BOUND OF $\mathbb{E}(M_1)$. Let us denote for any given $i \in [n]$ and $b \in [K] \setminus z_i$ the event

$$\Omega_1 = \left\{ \langle \epsilon_i, \mu_{z_i} - \mu_b \rangle + \lambda \langle E_{i:} W, \Pi_{z_i:} - \Pi_{b:} \rangle \leqslant \frac{-(1-\delta-\bar{\delta})\Delta^2(z_i,b)}{2} \right\}.$$

By using an analogous argument as the one presented in Lemma 7 we obtain

$$\mathbb{P}(\Omega_1) \leqslant \exp(-(1+o(1))\tilde{\Delta}^2).$$

Thus by taking $\delta = \bar{\delta}$ going to zero as $n \to \infty$, we obtain

$$\begin{aligned}
\mathbb{E}(M_1) &\leqslant \sum_{i=1}^{n} \sum_{b \in [K] \setminus z_i} \Delta^2(z_i, b) \exp(-(1+o(1))\tilde{\Delta}^2) \\
&\leqslant nK \exp(-(1+o(1))\tilde{\Delta}^2) \\
&\leqslant n \exp(-(1+o(1))\tilde{\Delta}^2).
\end{aligned}$$

In the second line we used the fact that $\Delta^2(z_i, b) = \Delta_{min} \lesssim \tilde{\Delta}^2$ for all $z_i \neq b$ and $\tilde{\Delta} \asymp \sqrt{\log n/K} \to \infty$. In the third line we used the assumption $\tilde{\Delta}^2/\log(K) \to \infty$.

UPPER BOUND OF $\mathbb{E}(M_2)$. Let us denote for any given $i \in [n]$ and $b \in [K] \setminus z_i$ the events

$$\Omega_2 = \left\{ \langle \epsilon_i, \tilde{\mu}_{z_i} - \mu_{z_i} \rangle + \lambda \langle E_{i:} W, \tilde{\Pi}_{z_i:} - \Pi_{z_i:} \rangle \leqslant \frac{-\bar{\delta}\Delta^2(z_i,b)}{4} \right\},$$

$$\Omega_2' = \left\{ \langle \epsilon_i, \tilde{\mu}_{z_i} - \mu_{z_i} \rangle \leqslant \frac{-\bar{\delta}\Delta^2(z_i, b)}{8} \right\},$$

and

$$\Omega_2'' = \left\{ \lambda \langle E_{i:} W, \tilde{\Pi}_{z_i:} - \Pi_{z_i:} \rangle \leqslant \frac{-\bar{\delta}\Delta^2(z_i, b)}{8} \right\}.$$

Clearly $\mathbb{P}(\Omega_2) \leqslant \mathbb{P}(\Omega_2') + \mathbb{P}(\Omega_2'')$ by a union bound argument.

Let us first upper bound $\mathbb{P}(\Omega_2')$. Recall that $n_k = n/K$ under the CSSBM by assumption, let us also define $n_{min} := \min_k n_k$. We keep this general notation because it is shorter and indicates how the proof can be generalized to the unbalanced setting. By definition $\tilde{\mu}_{z_i} - \mu_{z_i} = \sum_{j \in \mathcal{C}_{z_i}} \frac{\epsilon_j}{n_{z_i}}$, hence

$$\langle \epsilon_i, \tilde{\mu}_{z_i} - \mu_{z_i} \rangle = \frac{\|\epsilon_i\|^2 + \epsilon_i^\top \sum_{\substack{j \in \mathcal{C}_{z_i} \\ j \neq i}} \epsilon_j}{n_{z_i}}.$$

This last quantity is lower bounded by $\epsilon_i^\top \eta_i$ where $\eta_i = \frac{\sum_{j \in \mathcal{C}_{z_i}, j \neq i} \epsilon_j}{n_{z_i}}$. In particular ϵ_i and η_i are independent and their entries are also independent. Moreover η_i is a centered gaussian random variables with independent entries such that $\mathrm{Var}((\eta_i)_k) \leqslant 1/n_k$. So by Bernstein inequality, it holds for all $x > 0$ that

$$\mathbb{P}\left(\|\eta_i\|^2 \geqslant \frac{1}{n_k}(K + 2\sqrt{Kx} + 2x) \right) \leqslant \exp(-x)$$

which in turn implies

$$\begin{aligned}
\mathbb{P}\left(\epsilon_i^\top \eta_i \leqslant -\frac{\bar{\delta}\Delta_{min}^2}{8} \right) &\leqslant \mathbb{P}\left(\epsilon_i^\top \eta_i \leqslant -\frac{\bar{\delta}\Delta_{min}^2}{8} \;\middle|\; \|\eta_i\|^2 \leqslant \frac{1}{n_k}(K + 2\sqrt{Kx} + 2x) \right) \\
&\quad + \mathbb{P}\left(\|\eta_i\|^2 \geqslant \frac{1}{n_k}(K + 2\sqrt{Kx} + 2x) \right) \\
&\leqslant \exp\left(-c \frac{n_k(\bar{\delta}\Delta_{min}^2)^2}{K + 2\sqrt{Kx} + 2x} \right) + \exp(-x).
\end{aligned}$$

Setting $x = \sqrt{n_k}\bar{\delta}\Delta_{min}^2$ we obtain $\mathbb{P}(\Omega_2') \leqslant 2\exp(-C\bar{\delta}\sqrt{n_k}\Delta_{min}^2)$. Since $\bar{\delta} \to 0$ (as $n \to \infty$) at a suitably slow rate, we have $\bar{\delta}\sqrt{n/K} \to +\infty$. Consequently,

$$\mathbb{P}(\Omega_2') = o(\exp(-(C + o(1))\Delta_{min}^2) = o(\exp(-(1 + o(1))\tilde{\Delta}^2)).$$

Let us now bound $\mathbb{P}(\Omega_2'')$. Since $\tilde{\Pi}_{kk'} - \Pi_{kk'} = \sum_{i \in \mathcal{C}_k, j \in \mathcal{C}_{k'}} \frac{E_{ij}}{n_k n_{k'}} = W_{:k}^\top E W_{:k'}$, we obtain the decomposition

$$\lambda \langle E_{i:} W, \tilde{\Pi}_{z_i:} - \Pi_{z_i:} \rangle = \lambda \langle E_{i:} W, W_{z_i:}^\top E^{(i)} W \rangle + \lambda \langle E_{i:} W, W_{z_i:}^\top E^{(-i)} W \rangle$$

where $E^{(i)}$ is obtained from E by only keeping the ith row and column, and E_{-i} is the matrix obtained from E by replacing the ith row and column by zero. In particular, $E^{(-i)}$ is independent from

$E_{i:}$. The second term can be controlled by using the same techniques as before. Indeed, the entries of $W_{z_i:}^\top E^{(-i)} W$ are independent and $\mathrm{Var}(W_{z_i:}^\top E^{(-i)} W_{:k}) \leqslant C\frac{p_{max}}{n_{min}^2}$ for all k. Denoting $\eta'_{ik} = W_{z_i:}^\top E^{(-i)} W_{:k}$ and $\eta'_i = (\eta'_{ik})_{k\in[K]}$, this implies

$$\begin{aligned}
\mathbb{P}\left(\lambda\langle E_{i:}W, \eta'_i\rangle \leqslant -\frac{\bar\delta\Delta_{min}^2}{16}\right) \leqslant \\
\mathbb{P}\left(\lambda\langle E_{i:}W, \eta'_i\rangle \leqslant -\frac{\bar\delta\Delta_{min}^2}{16} \;\middle|\; \forall k,\, |\eta'_{ik}|^2 \leqslant C\frac{p_{max}}{n_k^2}(K+x)\right) \\
+ \mathbb{P}\left(\exists k,\, |\eta'_{ik}|^2 \geqslant C\frac{p_{max}}{n_{min}^2}(K+x)\right) \\
\leqslant K\exp\left(-C\frac{n_{min}^3(\bar\delta\Delta_{min}^2)^2}{p_{max}\lambda^2(K+x)}\right) + K\exp(-x).
\end{aligned}$$

Here, we used a union bound argument for the first inequality. The second inequality uses Lemma 33 – which provides a concentration bound for binomial random variables – along with the fact $\lambda \asymp \frac{n}{Kp_{max}}$.

Setting $x = C\sqrt{n_k p_{max}}\bar\delta\Delta_{min}^2$ we obtain

$$\begin{aligned}
\mathbb{P}\left(\lambda\langle E_{i:}W, \eta'_i\rangle \leqslant -\frac{\bar\delta\Delta_{min}^2}{16}\right) &\leqslant 2\exp(-c\sqrt{n_k p_{max}}\bar\delta\Delta_{min}^2) \\
&= o(\exp(-\tilde\Delta^2))
\end{aligned}$$

since $\bar\delta$ can be chosen such that $\sqrt{n_k p_{max}}\bar\delta \to +\infty$ (because by assumption $np_{max} \asymp \log n \gg K^4$) and $\Delta_{min} \asymp \tilde\Delta$.

It remains to control $\langle E_{i:}W, W_{z_i:}^\top E^{(i)} W\rangle$. Using the fact

$$(W_{z_i:}^\top E^{(i)} W)_k = \begin{cases} \sum_{j'\in\mathcal{C}_k} \frac{E_{ij'}}{n_{z_i} n_k} & \text{if } k \neq z_i \\ 2\sum_{j'\in\mathcal{C}_k} \frac{E_{ij'}}{n_{z_i}^2} & \text{if } k = z_i \end{cases}$$

we have

$$\begin{aligned}
\langle E_{i:}W, W_{z_i:}^\top E^{(i)} W\rangle &= \sum_{k\neq z_i} \sum_{\substack{j\in\mathcal{C}_k \\ j'\in\mathcal{C}_k}} \frac{E_{ij}}{n_k}\frac{E_{ij'}}{n_{z_i} n_k} + 2n_{z_i}\left(\sum_{j\in\mathcal{C}_{z_i}} \frac{E_{ij'}}{n_{z_i}^2}\right)^2 \\
&= \frac{1}{n_{z_i}}\left(\sum_{j\in\mathcal{C}_k} \frac{E_{ij}}{n_k}\right)^2 + 2n_{z_i}\left(\sum_{j\in\mathcal{C}_{z_i}} \frac{E_{ij'}}{n_{z_i}^2}\right)^2 \\
&\geqslant 0.
\end{aligned}$$

Consequently, $\mathbb{P}\left(\Omega_2''\right)$ can be bounded as

$$\mathbb{P}\left(\Omega_2''\right) \leqslant \mathbb{P}\left(\lambda\langle E_{i:}W, \eta'_i\rangle \leqslant -\frac{\bar\delta\Delta_{min}^2}{16}\right) = o(\exp(-\tilde\Delta^2)).$$

UPPER BOUND OF $\mathbb{E}(M_3)$. Let us denote for any given $i \in [n]$ and $b \in [K] \setminus z_i$ the event

$$\Omega_3 = \left\{ \langle -\epsilon_i, \tilde{\mu}_b - \mu_b \rangle - \lambda \langle E_{i:} W, \tilde{\Pi}_{b:} - \Pi_{b:} \rangle \leqslant \frac{-\bar{\delta}\Delta^2(z_i, b)}{4} \right\}.$$

First observe that

$$\langle \epsilon_i, \tilde{\mu}_b - \mu_b \rangle = \frac{\epsilon_i^\top \sum_{j \in \mathcal{C}_b} \epsilon_j}{n_{z_i}},$$

therefore this term can be handled in the same way as before. Moreover, we have

$$\lambda \langle E_{i:} W, \tilde{\Pi}_{b:} - \Pi_{b:} \rangle = \lambda \langle E_{i:} W, W_{b:}^\top E^{(i)} W \rangle + \lambda \langle E_{i:} W, W_{b:}^\top E^{(-i)} W \rangle.$$

The second term can be handled in the same way as before by using a conditioning argument. Now observe that

$$\langle E_{i:} W, W_{b:}^\top E^{(i)} W \rangle = \frac{1}{n_{z_i}^2 n_b} \left(\sum_{j \in \mathcal{C}_b} E_{ij} \right) \left(\sum_{j' \in \mathcal{C}_{z_i}} E_{ij'} \right)$$

where $\sum_{j \in \mathcal{C}_b} E_{ij}$ and $\sum_{j' \in \mathcal{C}_{z_i}} E_{ij'}$ are independent subgaussian random variables. Thus this term can also be controlled by using the same conditioning argument as before.

CONCLUSION. The previously obtained upper bounds imply

$$\mathbb{E}(\xi(\delta)) \leqslant 3\mathbb{E}(M_1) \leqslant n \exp(-(1 + o(1))\tilde{\Delta}^2).$$

Finally, by Markov inequality, we obtain

$$\mathbb{P}(\xi(\delta) \geqslant \exp(\tilde{\Delta}) \mathbb{E}\xi(\delta)) \leqslant \exp(-\tilde{\Delta}).$$

But since

$$\exp(\tilde{\Delta}) \mathbb{E}\xi(\delta) \leqslant n \exp(-(1 + o(1))\tilde{\Delta}^2)$$

we obtain that with probability at least $1 - \exp(-\tilde{\Delta})$

$$\xi(\delta) \leqslant n \exp(-(1 + o(1))\tilde{\Delta}^2).$$

3.7.2 *Proof of Theorem 5*

The general proof strategy has been presented in Section 3.4.1. In Section 3.7.2.1 we will make the error decomposition explicit. Then, we will control the different error terms in Sections 3.7.2.2, 3.7.2.3 and 3.7.2.4. Finally, we will conclude by applying Theorem 4 in Section 3.7.2.5.

3.7.2.1 *Error decomposition for the one-step analysis of* `IR-LSS`

We will assume without lost of generality that $\sigma = 1$ to simplify the exposition. Let $i \in [n]$ and $a \in [K]$ be such that[3] $z_i = a$, and let

$$\lambda^{(t)} = \frac{n}{K(p^{(t)} - q^{(t)})} \log\left(\frac{p^{(t)}(1-q^{(t)})}{q^{(t)}(1-p^{(t)})}\right)$$

denote the scalar corresponding to the diagonal entry of the inverse covariance matrix $\Sigma_k^{(t)}$. Similarly, let us denote

$$\lambda = \frac{n}{K(p-q)} \log\left(\frac{p(1-q)}{q(1-p)}\right).$$

Given the current estimator of the partition $Z^{(t)}$, node i will be incorrectly estimated after one refinement step if

$$a \neq \arg\min_k \|X_i - \mu_k^{(t)}\|^2 + \hat{\lambda}^{(t)} \|A_{i:} W^{(t)} - \Pi_{k:}^{(t)}\|^2$$

or equivalently, if there exists $b \in [K] \backslash a$ such that

$$\|X_i - \mu_b^{(t)}\|^2 + \hat{\lambda}^{(t)} \|A_{i:} W^{(t)} - \Pi_{b:}^{(t)}\|^2 \leqslant \|X_i - \mu_a^{(t)}\|^2 + \hat{\lambda}^{(t)} \|A_{i:} W^{(t)} - \Pi_{a:}^{(t)}\|^2.$$

The above inequality is equivalent to

$$\langle \epsilon_i, \tilde{\mu}_a - \tilde{\mu}_b \rangle + \lambda \langle E_{i:} W, \tilde{\Pi}_{a:} - \tilde{\Pi}_{b:} \rangle \leqslant \frac{-\Delta^2(a,b)}{2} + F_{ib}^{(t)} + G_{ib}^{(t)} + H_{ib}^{(t)}$$

where

$$\begin{aligned} \Delta^2(a,b) &= \|\mu_a - \mu_b\|^2 + \lambda \|\Pi_{a:} - \Pi_{b:}\|^2, \\ \tilde{\mu}_k &= X^\top W_{:a}, \\ \text{and } \tilde{\Pi}_{k:} &= W_{k:}^\top A W \end{aligned}$$

for all $k \in [K]$. Furthermore, the terms $F_{ib}^{(t)}, G_{ib}^{(t)}$ and $H_{ib}^{(t)}$ are given by

$$\begin{aligned} F_{ib}^{(t)} &= \langle \epsilon_i, (\tilde{\mu}_a - \mu_a^{(t)}) - (\tilde{\mu}_b - \mu_b^{(t)}) \rangle + \\ &\quad \lambda^{(t)} \langle E_{i:} W^{(t)}, (\tilde{\Pi}_{a:} - \Pi_{a:}^{(t)}) - (\tilde{\Pi}_{b:} - \Pi_{b:}^{(t)}) \rangle \\ &\quad + \lambda^{(t)} \langle E_{i:} (W - W^{(t)}), \tilde{\Pi}_{a:} - \tilde{\Pi}_{b:} \rangle \\ &\quad + (\lambda - \lambda^{(t)}) \langle E_{i:} W, \tilde{\Pi}_{a:} - \tilde{\Pi}_{b:} \rangle, \\ 2G_{ib}^{(t)} &= (\|\mu_a - \mu_a^{(t)}\|^2 - \|\mu_a - \tilde{\mu}_a\|^2) \\ &\quad - (\|\mu_a - \mu_b^{(t)}\|^2 - \|\mu_a - \tilde{\mu}_b\|^2) \\ &\quad + \lambda^{(t)} (\|P_{i:} W^{(t)} - \Pi_{a:}^{(t)}\|^2 - \|P_{i:} W^{(t)} - W_{a:}^\top A W^{(t)}\|^2) \end{aligned}$$

3 Depending on the context we will interchangeably use the notation z_i and a.

$$
\begin{aligned}
&\quad - \lambda^{(t)}(\|P_{i:}W^{(t)} - \Pi^{(t)}_{b:}\|^2 - \|P_{i:}W^{(t)} - W^{\top}_{b:}AW^{(t)}\|^2) \\
\text{and} \quad 2H^{(t)}_{ib} &= \|\mu_a - \tilde{\mu}_a\|^2 - \|\mu_a - \tilde{\mu}_b\|^2 + \|\mu_a - \mu_b\|^2 \\
&\quad + \lambda^{(t)}(\|P_{i:}W^{(t)} - W^{\top}_{a:}AW^{(t)}\|^2 \\
&\quad - \|P_{i:}W^{(t)} - W^{\top}_{b:}AW^{(t)}\|^2 + \|\Pi_{a:} - \Pi_{b:}\|^2) \\
&\quad + (\lambda - \lambda^{(t)})\|\Pi_{a:} - \Pi_{b:}\|^2.
\end{aligned}
$$

The main term in this decomposition is

$$
\langle \epsilon_i, \tilde{\mu}_a - \tilde{\mu}_b \rangle + \lambda \langle E_{i:}W, \tilde{\Pi}_{a:} - \tilde{\Pi}_{b:} \rangle \leqslant \frac{-\Delta^2(a,b)}{2}
$$

and corresponds to the error when the current estimation of the partition is the ground truth partition. It is controlled by Lemma 1

The three error terms will be controlled in different ways. The error term $F^{(t)}_{ib}$ depends in a crucial way on i and t, it will be controlled with a l_2-type norm (see Condition C2). The square of the error terms $G^{(t)}_{ib}$ and $|H^{(t)}_{ib}|$ will be controlled uniformly (see Condition C3).

3.7.2.2 *Bounding the error term* $F^{(t)}_{ib}$

In this section we are going to show that Condition C2 is satisfied.

Lemma 2. *Under the assumptions of Theorem 7 (that are also satisfied by Theorem 5) we have w.h.p. that for all* $z^{(t)}$ *such that* $l(z^{(t)}, z) \leqslant \tau^{(0)}$,

$$
\sum_{i=1}^{n} \max_{b \in [K] \setminus z_i} \frac{(F^{(t)}_{ib})^2}{\Delta^2(z_i, b) l(z, z^{(t)})} \leqslant \frac{\delta^2}{256}.
$$

Proof. We need to upper-bound of

$$
F = \max_{\{z^{(t)} : l(z, z^{(t)}) \leqslant \tau^{(0)}\}} \sum_{i=1}^{n} \max_{b \in [K] \setminus z_i} \frac{(F^{(t)}_{ib})^2}{\Delta(z_i, b)^2 l(z, z^{(t)})}.
$$

To this end, we can decompose $F^{(t)}_{ib} = F^{1,(t)}_{ib} + F^{2,(t)}_{ib}$ where

$$
F^{1,(t)}_{ib} = \langle \epsilon_i, (\tilde{\mu}_a - \mu^{(t)}_a) - (\tilde{\mu}_b - \mu^{(t)}_b) \rangle
$$

is the error arising from the GMM part of the model and

$$
\begin{aligned}
F^{2,(t)}_{ib} &= \lambda^{(t)} \langle E_{i:}W^{(t)}, (\tilde{\Pi}_{a:} - \Pi^{(t)}_{a:}) - (\tilde{\Pi}_{b:} - \Pi^{(t)}_{b:}) \rangle \\
&\quad + \lambda^{(t)} \langle E_{i:}(W - W^{(t)}), \tilde{\Pi}_{a:} - \tilde{\Pi}_{b:} \rangle + (\lambda - \lambda^{(t)}) \langle E_{i:}W, \tilde{\Pi}_{a:} - \tilde{\Pi}_{b:} \rangle
\end{aligned}
$$

is the error coming from the SBM part of the model. We have

$$
F \leqslant 2 \max_{\{z^{(t)} : l(z, z^{(t)}) \leqslant \tau^{(0)}\}} \underbrace{\sum_{i=1}^{n} \max_{b \in [K] \setminus z_i} \frac{(F^{1,(t)}_{ib})^2}{\Delta(z_i, b)^2 l(z, z^{(t)})}}_{F^{(t)}_1}
$$

$$+2 \max_{\{z^{(t)}:l(z,z^{(t)})\leqslant\tau^{(0)}\}} \underbrace{\sum_{i=1}^{n} \max_{b\in[K]\setminus z_i} \frac{(F_{ib}^{2,(t)})^2}{\Delta(z_i,b)^2 l(z,z^{(t)})}}_{F_2^{(t)}}$$

and it is sufficient to individually control each term.

CONTROL OF F_1. We follow the same steps as in Gao et al., 2019, the only difference is that we use a different definition for Δ. To begin with,

$$\begin{aligned} F_1^{(t)} &\leqslant \sum_{i=1}^{n} \sum_{b\in[K]\setminus z_i} \frac{\langle \epsilon_i, (\tilde{\mu}_{z_i} - \mu_{z_i}^{(t)}) - (\tilde{\mu}_b - \mu_b^{(t)})\rangle^2}{\Delta(z_i,b)^2 l(z,z^{(t)})} \\ &\leqslant \sum_{b\in[K]} \sum_{a\in[K]\setminus b} \sum_{i\in\mathcal{C}_a} \frac{\langle \epsilon_i, (\tilde{\mu}_a - \mu_a^{(t)}) - (\tilde{\mu}_b - \mu_b^{(t)})\rangle^2}{\Delta(a,b)^2 l(z,z^{(t)})} \\ &\leqslant \sum_{b\in[K]} \sum_{a\in[K]\setminus b} \| \sum_{i\in\mathcal{C}_a} \epsilon_i \epsilon_i^\top \| \frac{\|(\tilde{\mu}_a - \mu_a^{(t)}) - (\tilde{\mu}_b - \mu_b^{(t)})\|^2}{\Delta(a,b)^2 l(z,z^{(t)})}. \end{aligned}$$

We first need to control $\| \sum_{i\in\mathcal{C}_a} \epsilon_i \epsilon_i^\top \|$ which can be done using the following lemma.

Lemma 3. *Let* $\epsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I_d)$. *With probability at least* $1 - \exp(-0.5n)$, *we have*

$$\| \sum_{i\in[n]} \epsilon_i \epsilon_i^\top \| \lesssim n + d.$$

Proof. See Lemma A.2 in Lu et al., 2016. □

Next, we need to control $\|\tilde{\mu}_a - \mu_a^{(t)}\|^2$ for all $a \in [K]$, this can be done with the following lemma.

Lemma 4. *Under the assumptions of Theorem 5, the following holds with probability at least* $1 - n^{-\Omega(1)}$. *If* $z^{(t)}$ *satisfies* $l(z^{(t)}, z) \leqslant \tau^{(0)} = \frac{\epsilon n \Delta_{\min}^2}{K}$ *then it implies*

1. $\max_{k\in[K]} \|\tilde{\mu}_k - \mu_k\| \lesssim \sqrt{\frac{K(d+\log n)}{n}}$,
2. $\max_{k\in[K]} \|\mathbb{E}(X)^\top (W_{:k}^{(t)} - W_{:k})\| \lesssim \frac{K}{n\Delta_{\min}} l(z^{(t)}, z)$,
3. $\max_{k\in[K]} \|(X - \mathbb{E}(X))^\top W_{:k}^{(t)}\| \lesssim \frac{K\sqrt{(d+n) l(z^{(t)},z)}}{n\Delta_{\min}} + \frac{K\sqrt{K(d+\log n)} l(z^{(t)},z)}{n\sqrt{n}\Delta_{\min}^2}$,
4. $\|\tilde{\mu}_k - \mu_k^{(t)}\| \leqslant C_3 \frac{K\sqrt{(d+n) l(z^{(t)},z)}}{n\Delta_{\min}}$.

Proof. Straightforward adaptation the proof of Lemma 4.1 in Gao et al., 2019. □

By combining the different bounds, we can now conclude that with high probability,

$$\max_{\{z^{(t)}:l(z^{(t)},z)\leqslant\tau^{(0)}\}} F_1^{(t)} \lesssim \frac{K^2(Kd/n+1)}{\Delta_{min}^2}\left(1+\frac{K(d/n+1)}{\Delta_{min}^2}\right).$$

This quantity goes to zero when $\Delta_{min}^2/K^3 \to +\infty$.

CONTROL OF $F_2^{(t)}$. Here we can not directly apply the framework developed by Gao et al., 2019. Different changes are necessary and we need to deal with additional dependencies.

Let $b \in [K] \neq z_j$, we then have the bound

$$\begin{aligned}(F_{ib}^{2,(t)})^2 &\leqslant 3(\lambda^{(t)}\langle E_{i:}W^{(t)}, (\tilde{\Pi}_{a:}-\Pi_{a:}^{(t)})-(\tilde{\Pi}_{b:}-\Pi_{b:}^{(t)})\rangle)^2\\ &+3(\lambda^{(t)}\langle E_{i:}(W-W^{(t)}), \tilde{\Pi}_{a:}-\tilde{\Pi}_{b:}\rangle)^2\\ &+3(\lambda-\lambda^{(t)})^2\langle E_{i:}W, \tilde{\Pi}_{a:}-\tilde{\Pi}_{b:}\rangle^2\\ &= F_{21}^2+F_{22}^2+F_{33}^2.\end{aligned}$$

We drop the superscript (t) in the notation for the terms F_{21}, F_{22} and F_{23} for convenience, but clearly they depend on t as well. We will now bound each of the terms F_{2i} for $i = 1\ldots3$ separately. Starting with F_{21}, first note that

$$|\langle E_{i:}W^{(t)}, (\tilde{\Pi}_{a:}-\Pi_{a:}^{(t)})-(\tilde{\Pi}_{b:}-\Pi_{b:}^{(t)})\rangle|^2 \leqslant 4\|E_{i:}W^{(t)}\|^2 \max_k \|\tilde{\Pi}_{k:}-\Pi_{k:}^{(t)}\|^2.$$

With high probability, for all $z^{(t)}$ such that $l(z^{(t)},z) \leqslant \tau^{(0)}$ we have

$$\begin{aligned}\max_{k\in[K]}\|\tilde{\Pi}_{k:}-\Pi_{k:}^{(t)}\|^2 &\leqslant 2\max_{k\in[K]}\|(W_{:k}-W_{:k}^{(t)})^\top AW\|^2\\ &+2\max_{k\in[K]}\|W_{:k}^{(t)\top}A(W-W^{(t)})\|^2\\ &\lesssim \left(\frac{K^2\sqrt{p_{max}}l(z^{(t)},z)}{n^{1.5}\Delta_{min}}\right)^2 \qquad \text{(by Lemma 11)}\\ &\lesssim \frac{K^4 p_{max}l(z^{(t)},z)^2}{n^3\Delta_{min}^2}\end{aligned}$$

and since $\lambda^{(t)} \lesssim \lambda$ (by Lemma 13) it follows that

$$\begin{aligned}\sum_{i=1}^n \max_{b\in[K]\setminus z_i}\frac{F_{21}^2}{\Delta^2(z_i,b)l(z,z^{(t)})}&\\ &\lesssim \lambda^2\sum_i \|E_{i:}W^{(t)}\|^2\frac{K^4p_{max}l(z^{(t)},z)^2}{n^3\Delta_{min}^2}\\ &\lesssim \lambda^2\|EW^{(t)}\|_F^2\frac{K^4p_{max}l(z^{(t)},z)}{n^3\Delta_{min}^4}\\ &\lesssim \lambda^2K\|EW^{(t)}\|^2\frac{K^4p_{max}l(z^{(t)},z)}{n^3\Delta_{min}^4}\end{aligned}$$

$$\begin{aligned}
&\lesssim \lambda^2 K^2 p_{max} \frac{K^4 p_{max} l(z^{(t)}, z)}{n^3 \Delta_{min}^4} \\
&\qquad \text{(by a consequence of Lemma 11, fifth item)} \\
&\lesssim \frac{K^4 l(z^{(t)}, z)}{n \Delta_{min}^4} \qquad \text{(since } \lambda \lesssim \tfrac{n}{K p_{max}}\text{)} \\
&\lesssim \frac{K^3}{\Delta_{min}^2} \to 0
\end{aligned}$$

where we used the fact $\frac{K l(z^{(t)},z)}{n\Delta_{min}^2} \leqslant \epsilon$ for the last line. Indeed, $\frac{K l(z^{(t)},z)}{n\Delta_{min}^2} \leqslant \frac{K\tau^{(0)}}{n\Delta_{min}^2} = \epsilon$.

Let us define $\Delta_2^2(a,b) := \|\Pi_{a:} - \Pi_{b:}\|^2$ for $a, b \in [K]$. Since

$$\begin{aligned}
F_{22}^2 &\lesssim \lambda^2 \|E_{i:}(W - W^{(t)})\|^2 \|\tilde{\Pi}_{z_i:} - \tilde{\Pi}_{b:}\|^2 \\
&\lesssim \lambda^2 \|E_{i:}(W - W^{(t)})\|^2 \|\Pi_{z_i:} - \Pi_{b:}\|^2
\end{aligned}$$

hence we have w.h.p. for all $z^{(t)}$ such that $l(z^{(t)}, z) \leqslant \tau^{(0)}$

$$\begin{aligned}
&\sum_{i=1}^n \max_{b \in [K] \setminus z_i} \frac{F_{22}^2}{\Delta(z_i, b)^2 l(z, z^{(t)})} \\
&\qquad \lesssim \lambda \sum_i \|E_{i:}(W - W^{(t)})\|^2 \frac{1}{l(z^{(t)}, z)} \max_{b \in [K] \setminus z_i} \frac{\lambda \|\Pi_{z_i:} - \Pi_{b:}\|^2}{\Delta(z_i, b)^2} \\
&\qquad \lesssim \|E(W - W^{(t)})\|_F^2 \frac{\lambda}{l(z^{(t)}, z)} \\
&\qquad\qquad \text{(because } \Delta^2(z_i, b) \geqslant \lambda \Delta_2^2(z_i, b)\text{)} \\
&\qquad \lesssim K \|E(W - W^{(t)})\|^2 \frac{\lambda}{l(z^{(t)}, z)} \\
&\qquad \lesssim \lambda K n p_{max} \frac{K^3 l(z^{(t)}, z)}{n^3 \Delta_{min}^4} \qquad \text{(by Lemma 11)} \\
&\qquad \lesssim \frac{K^3}{\Delta_{min}^2} \to 0.
\end{aligned}$$

Using the same proof technique as in Lemma 1

$$(\langle E_{i:} W, \tilde{\Pi}_{a:} - \tilde{\Pi}_{b:} \rangle)^2 \lesssim K p_{max} \Delta_2^2(a, b)$$

holds w.h.p. Since by Lemma 13 we have w.h.p. that for all $z^{(t)}$ such that $l(z^{(t)}, z) \leqslant \tau^{(0)}$

$$|\lambda^{(t)} - \lambda| \lesssim \lambda \frac{K^2 l(z^{(t)}, z)}{\sqrt{n p_{max}} n \Delta_{min}},$$

we obtain

$$\sum_{i=1}^n \max_{b \in [K] \setminus z_i} \frac{F_{23}^2}{\Delta(z_i, b)^2 l(z, z^{(t)})}$$

$$
\begin{aligned}
&\lesssim \lambda^2 \left(\frac{K^2 l(z^{(t)}, z)}{\sqrt{np_{max}} n\Delta_{min}}\right)^2 \sum_i \|E_{i:}W\|^2 \frac{\max_{b\in[K]\setminus z_i} \|\tilde{\Pi}_{z_i:} - \tilde{\Pi}_{b:}\|^2}{\Delta^2(z_i, b) l(z^{(t)}, z)} \\
&\lesssim \lambda \frac{K^4 l(z^{(t)}, z)}{n^3 p_{max} \Delta_{min}^2} \|EW\|_F^2 \max_{b\in[K]\setminus z_i} \frac{\lambda \|\Pi_{z_i:} - \Pi_{b:}\|^2}{\Delta^2(z_i, b)} \\
&\lesssim \frac{n}{Kp_{max}} \frac{K^4 l(z^{(t)}, z)}{n^3 p_{max} \Delta_{min}^2} Kp_{max} \qquad \text{(by Lemma 35)} \\
&\lesssim \frac{K^4 l(z^{(t)}, z)}{n^2 p_{max} \Delta_{min}^2} \\
&\lesssim \frac{K^3}{np_{max}} \\
&\lesssim \frac{K^3}{\Delta_{min}^2} \to 0.
\end{aligned}
$$

Consequently, we have established that Condition C2 holds for all $\delta = o(1)$ such that $\delta^2 = \omega(K^3/\Delta_{min}^2)$. □

3.7.2.3 *Error term* $G_{ib}^{(t)}$

In this section we are going to show that the G - error term satisfied condition C3.

Lemma 5. *Under the assumptions of Theorem 7 (that are also satisfied by Theorem 5) we have w.h.p. for all $z^{(t)}$ such that $l(z^{(t)}, z) \leqslant \tau^{(t)}$*

$$
\max_{i\in[n]} \max_{b\in[K]\setminus z_i} \frac{|G_{ib}^{(t)}|}{\Delta(z_i, b)^2} \leqslant \frac{\delta^{(t+1)}}{8}.
$$

Proof. As for $F_{ib}^{(t)}$ we can split $G_{ib}^{(t)} = G_{ib}^{1,(t)} + G_{ib}^{2,(t)}$ where

$$
\begin{aligned}
G_{ib}^{1,(t)} &= \|\mu_a - \mu_a^{(t)}\|^2 - \|\mu_a - \tilde{\mu}_a\|^2) - (\|\mu_a - \mu_b^{(t)}\|^2 - \|\mu_a - \tilde{\mu}_b\|^2) \\
G_{ib}^{2,(t)} &= \lambda^{(t)}(\|P_{i:}W^{(t)} - \Pi_{a:}^{(t)}\|^2 - \|P_{i:}W^{(t)} - W_{a:}AW^{(t)}\|^2) \\
&\quad - \lambda^{(t)}(\|P_{i:}W^{(t)} - \Pi_{b:}^{(t)}\|^2 - \|P_{i:}W^{(t)} - W_{b:}AW^{(t)}\|^2).
\end{aligned}
$$

By the proof of Lemma 4.1 in Gao et al., 2019 (last inequality of page 46, equations (115) and (118)), we have w.h.p. for all $z^{(t)}$ such that $l(z^{(t)}, z) \leqslant \tau^{(t)}$,

$$
\begin{aligned}
\frac{|G_{ib}^{1,(t)}|}{\Delta^2(a, b)} \lesssim &\left(\frac{Kl(z^{(t)}, z)}{n\Delta_{min}} + K\sqrt{\frac{Kl(z^{(t)}, z)}{n\Delta_{min}}}\right)^2 \Delta_{min}^{-2} \\
&+ \left(\frac{Kl(z^{(t)}, z)}{n\Delta_{min}} + K\sqrt{\frac{Kl(z^{(t)}, z)}{n\Delta_{min}}}\right) \frac{Kl(z^{(t)}, z)}{n\Delta_{min}} \Delta_{min}^{-2} \\
&+ \left(\frac{Kl(z^{(t)}, z)}{n\Delta_{min}} + K\sqrt{\frac{Kl(z^{(t)}, z)}{n\Delta_{min}}}\right) \Delta_{min}^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&\lesssim \frac{Kl(z^{(t)},z)^2}{n^2\Delta_{min}^4} + K\frac{l(z^{(t)},z)}{n\Delta_{min}^2} \\
&\lesssim K\frac{l(z^{(t)},z)}{n\Delta_{min}^2} \qquad \text{(since } K\tfrac{l(z^{(t)},z)}{n\Delta_{min}^2} < 1 \text{ and } \tfrac{Kl(z^{(t)},z)^2}{n^2\Delta_{min}^4} < \tfrac{1}{K}\text{)} \\
&\lesssim \frac{K\tau^{(t)}}{n\Delta_{min}^2} \\
&\lesssim \max\left(\frac{7}{8}\frac{K\tau^{(t-1)}}{n\Delta_{min}^2}, \frac{K\tau}{n\Delta_{min}^2}\right) \qquad \text{(by definition of } \tau^{(t)}\text{)} \\
&\lesssim \max\left(\frac{7}{8}(\frac{8}{7}\epsilon)\delta^{(t)}, \epsilon\delta\right) \\
&\qquad \text{(using } \tau = \tau^{(0)}\delta \text{ and also the definition of } \delta^{(t)}\text{)} \\
&\lesssim \epsilon\delta^{(t)}. \qquad \text{(since } \delta \leqslant \delta^{(t)}\text{)}
\end{aligned}
$$

Now by choosing ϵ to be a suitably small constant (< 1), since $\delta^{(t+1)} \leqslant \frac{7}{8}\delta^{(t)}$ we obtain

$$\frac{|G_{ib}^{1,(t)}|}{\Delta^2(a,b)} \leqslant \frac{\delta^{(t+1)}}{16}.$$

To bound $G_{ib}^{2,(t)}$ we will adapt the method developed in Han et al., 2020. We have by direct calculation

$$
\begin{aligned}
\frac{G_{ib}^{2,(t)}}{\lambda^{(t)}} &= (\|\Pi_{a:}^{(t)} - W_{a:}^\top A W^{(t)}\|^2 - \|\Pi_{b:}^{(t)} - W_{b:}AW^{(t)}\|^2) \\
&\quad + 2\langle P_{i:}W^{(t)} - W_{a:}AW^{(t)}, W_{:a}^\top AW^{(t)} - \Pi_{a:}^{(t)}\rangle \\
&\quad - 2\langle P_{i:}W^{(t)} - W_{b:}AW^{(t)}, W_{:b}^\top AW^{(t)} - \Pi_{b:}^{(t)}\rangle \\
&\leqslant \left|\|\Pi_{a:}^{(t)} - W_{:a}^\top AW^{(t)}\|^2 - \|\Pi_{b:}^{(t)} - W_{:b}^\top AW^{(t)}\|^2\right| \\
&\quad + 4\max_{a\in[K]}\left|\langle W_{:a}^\top EW^{(t)}, (W_{:b} - W_{:b}^{(t)})^\top AW^{(t)}\rangle\right| \\
&\quad + 2\left|\langle(\Pi_{a:} - \Pi_{b:})Z^\top W^{(t)}, (W_{:b} - W_{:b}^{(t)})^\top AW^{(t)}\rangle\right| \\
&=: G_{21} + G_{22} + G_{23}.
\end{aligned}
$$

We drop the superscript (t) in the notation for the terms G_{21}, G_{22} and G_{23} for convenience, but clearly they depend on t as well. First observe that w.h.p, it holds for all $z^{(t)}$ such that $l(z^{(t)},z) \leqslant \tau^{(t)}$ that

$$
\begin{aligned}
G_{21} &\leqslant \max_{a\in[K]}\|\Pi_{a:}^{(t)} - W_{:a}^\top AW^{(t)}\|^2 = \max_{a\in[K]}\|(W_{a:}^{(t)} - W_{:a})^\top AW^{(t)}\|^2 \\
&\lesssim \frac{K^3 p_{max} l(z^{(t)},z)^2}{n^3\Delta_{min}^2}
\end{aligned}
$$

where we used Lemma 11 for the last inequality. This implies by Lemma 12 that

$$\max_{b\in[K]\setminus z_i}\frac{\lambda^{(t)}G_{21}}{\Delta^2(z_i,b)} \lesssim \frac{K^2 l(z^{(t)},z)^2}{n^2\Delta_{min}^4} \leqslant \frac{K^2(\tau^{(0)}\delta^{(t)})^2}{n^2\Delta_{min}^4}$$

which implies $\max_{b\in[K]\setminus z_i} \frac{\lambda^{(t)} G_{21}}{\Delta^2(z_i,b)} \leqslant (\frac{\delta^{(t+1)}}{16})^2$ using the same argument used earlier (for suitably small constant $\epsilon < 1$). Next, in order to bound G_{22}, note that w.h.p, it holds for all $z^{(t)}$ such that $l(z^{(t)}, z) \leqslant \tau^{(t)}$ that

$$\begin{aligned} G_{22} &\leqslant 4 \max_{a\in[K]} \|W_{:a}^\top E W^{(t)}\| \max_{a\in[K]} \|(W_{:b} - W_{:b}^{(t)})^\top A W^{(t)}\| \\ &\lesssim \frac{K\sqrt{p_{max}}}{\sqrt{n}} \frac{K^{1.5}\sqrt{p_{max}} l(z^{(t)}, z)}{n^{1.5}\Delta_{min}} \qquad \text{(by Lemma 11)} \\ &\lesssim K^{2.5} \frac{p_{max}}{n} \frac{l(z^{(t)}, z)}{n\Delta_{min}} \end{aligned}$$

which in turn implies

$$\begin{aligned} \max_{b\in[K]\setminus z_i} \frac{\lambda^{(t)} G_{22}}{\Delta(z_i,b)^2} &\lesssim \frac{\sqrt{K}}{\Delta_{min}} \frac{K l(z^{(t)}, z)}{n\Delta_{min}^2} \qquad \text{(since } \lambda^{(t)} \lesssim n/(Kp_{max})\text{)} \\ &= o(\delta) \end{aligned}$$

as $\sqrt{K}/\Delta_{min} \to 0$ and $K\frac{l(z^{(t)},z)}{n\Delta_{min}^2} < 1$.

Finally, in order to bound G_{23}, note that w.h.p, it holds for all $z^{(t)}$ such that $l(z^{(t)}, z) \leqslant \tau^{(t)}$,

$$\begin{aligned} G_{23} &\lesssim \|\Pi_{a:} - \Pi_{b:})Z^\top W^{(t)}\| \max_{b\in[K]} \|(W_{:b} - W_{:b}^{(t)})^\top A W^{(t)}\| \\ &\lesssim \Delta_2(a,b) \frac{K^{1.5}\sqrt{p_{max}} l(z^{(t)}, z)}{n^{1.5}\Delta_{min}} \end{aligned}$$

which implies

$$\begin{aligned} \max_{b\in[K]\setminus z_i} \frac{\lambda^{(t)} G_{23}}{\Delta(z_i,b)^2} &\lesssim \lambda \frac{\Delta_2(z_i,b)}{\Delta(z_i,b)} \frac{K^{1.5}\sqrt{p_{max}} l(z^{(t)}, z)}{n^{1.5}\Delta_{min}^2} \\ &\lesssim \sqrt{\lambda} \frac{\sqrt{Kp_{max}}}{\sqrt{n}} \frac{K l(z^{(t)}, z)}{n\Delta_{min}^2} \\ &\lesssim \frac{K l(z^{(t)}, z)}{n\Delta_{min}^2}. \qquad \text{(since } \lambda \lesssim \tfrac{n}{Kp_{max}}\text{)} \end{aligned}$$

For a suitably constant $\epsilon < 1$, this then implies

$$\max_{b\in[K]\setminus z_i} \frac{\lambda^{(t)} G_{23}}{\Delta(z_i,b)^2} \leqslant \frac{\delta^{(t+1)}}{16}.$$

□

3.7.2.4 *Error term* $H_{ib}^{(t)}$

In this section we are going to show that the H - error term satisfied condition C3.

Lemma 6. *Under the assumptions of Theorem 7 (that are also satisfied by Theorem 5) we have w.h.p. that for all $z^{(t)}$ such that $l(z^{(t)}, z) \leqslant \tau^{(t)}$,*

$$\max_{i\in[n]} \max_{b\in[K]\setminus z_i} \frac{|H_{ib}^{(t)}|}{\Delta(z_i, b)^2} \leqslant \frac{\delta^{(t+1)}}{8}.$$

Proof. As before, we can split $H_{ib}^{(t)} = H_{ib}^{1,(t)} + H_{ib}^{2,(t)}$ where

$$\begin{aligned}
2H_{ib}^{1,(t)} &= \|\mu_a - \tilde{\mu}_a\|^2 - \|\mu_a - \tilde{\mu}_b\|^2 + \|\mu_a - \mu_b\|^2 \\
2H_{ib}^{2,(t)} &= \lambda^{(t)}(\|P_{i:}W^{(t)} - W_{a:}^\top A W^{(t)}\|^2 - \|P_{i:}W^{(t)} - W_{b:}^\top A W^{(t)}\|^2 \\
&\quad + \|\Pi_{a:} - \Pi_{b:}\|^2) + (\lambda - \lambda^{(t)})\|\Pi_{a:} - \Pi_{b:}\|^2.
\end{aligned}$$

By an immediate adaptation of Lemma 4.1 in Gao et al., 2019 it holds w.h.p. that for all $z^{(t)}$ such that $l(z^{(t)}, z) \leqslant \tau^{(t)}$,

$$\frac{|H_{ib}^{1,(t)}|}{\Delta(z_i, b)^2} \lesssim \frac{K(d + \log n)}{n\Delta_{min}^2} + \sqrt{\frac{K(d + \log n)}{n\Delta_{min}^2}} \to 0$$

as long as $K/\Delta_{min}^2 \to 0$.

It remains to uniformly control $H_{ib}^{2,(t)}$, let us split it[4] as $H_{ib}^{2,(t)} = \lambda^{(t)} H_1 + H_2$. First note that by Lemma 13 it holds w.h.p. that for all $z^{(t)}$ such that $l(z^{(t)}, z) \leqslant \tau^{(t)}$,

$$H_2 := (\lambda - \lambda^{(t)})\|\Pi_{a:} - \Pi_{b:}\|^2 \lesssim \lambda \frac{K l(z^{(t)}, z)}{n\Delta_{min}^2} \Delta_2^2(a, b)$$

which implies

$$\frac{|H_2|}{\Delta(z_i, b)^2} \lesssim \frac{K l(z^{(t)}, z)}{n\Delta_{min}^2}.$$

Then, we obtain $\frac{|H_2|}{\Delta(z_i,b)^2} \leqslant \frac{\delta^{(t+1)}}{16}$ for a small enough constant $\epsilon < 1$.

Now observe that

$$\begin{aligned}
H_1 :&= \|P_{i:}W^{(t)} - W_{a:}^\top A W^{(t)}\|^2 - \|P_{i:}W^{(t)} - W_{b:}^\top A W^{(t)}\|^2 \\
&\quad + \|\Pi_{a:} - \Pi_{b:}\|^2 \\
&= \|W_{a:}^\top E W^{(t)}\|^2 \\
&\quad + (\|\Pi_{a:} - \Pi_{b:}\|^2 - \|P_{i:}W^{(t)} - W_{b:}^\top P W^{(t)}\|^2) \\
&\quad - (\|P_{i:}W^{(t)} - W_{b:}^\top A W^{(t)}\|^2 - \|P_{i:}W^{(t)} - W_{b:}^\top P W^{(t)}\|^2) \\
&= (\|\Pi_{a:} - \Pi_{b:}\|^2 - \|P_{i:}W^{(t)} - W_{b:}^\top P W^{(t)}\|^2) \\
&\quad + (\|W_{a:}^\top E W^{(t)}\|^2 - \|W_{b:}^\top E W^{(t)}\|^2) \\
&\quad + 2\langle P_{i:}W^{(t)} - W_{b:}^\top P W^{(t)}, W_{:b}^\top E W^{(t)}\rangle \\
&= H_{11}^{(t)} + H_{12}^{(t)} + H_{13}^{(t)}.
\end{aligned}$$

4 We drop the superscript (t) in the notation for H_1, H_2 for convenience, but clearly they both depend on t as well.

By writing

$$P_{i:}W^{(t)} - W_{b:}^{\top}PW^{(t)} = (\Pi_{a:} - \Pi_{b:})Z^{\top}W^{(t)},$$

we obtain w.h.p. that for all $z^{(t)}$ such that $l(z^{(t)}, z) \leqslant \tau^{(t)}$,

$$\begin{aligned}
|H_{13}^{(t)}| &\lesssim \|P_{i:}W^{(t)} - W_{b:}^{\top}PW^{(t)}\|\|W_{:b}^{\top}EW^{(t)}\| \\
&\lesssim \|\Pi_{a:} - \Pi_{b:}\|\frac{K\sqrt{p_{max}}}{\sqrt{n}} \\
&\lesssim \|\Pi_{a:} - \Pi_{b:}\|^2\frac{K}{\sqrt{np_{max}}}.
\end{aligned}$$

In particular,

$$\frac{\lambda^{(t)}|H_{13}^{(t)}|}{\Delta^2(z_i, b)} \lesssim \frac{\lambda\Delta_2^2(z_i, b)K}{\Delta^2(z_i, b)\sqrt{np_{max}}} \lesssim \frac{K}{\sqrt{np_{max}}} \to 0.$$

Next observe that w.h.p., it holds for all $z^{(t)}$ such that $l(z^{(t)}, z) \leqslant \tau^{(t)}$ that

$$\left|\|W_{a:}^{\top}EW^{(t)}\|^2 - \|W_{b:}^{\top}EW^{(t)}\|^2\right| \leqslant \max_{k\in[K]}\|W_{k:}^{\top}EW^{(t)}\|^2 \lesssim \frac{K^2p_{max}}{n}$$

where the last inequality uses Lemma 11. This implies

$$\frac{\lambda^{(t)}|H_{12}^{(t)}|}{\Delta^2(z_i, b)} \lesssim \frac{K}{\Delta_{min}^2} \to 0.$$

Finally, it remains to bound $|H_{11}^{(t)}|$. To begin with,

$$\begin{aligned}
|H_{11}^{(t)}| &:= \left|\|\Pi_{a:} - \Pi_{b:}\|^2 - \|P_{i:}W^{(t)} - W_{b:}^{\top}PW^{(t)}\|^2\right| \\
&= \left|\|\Pi_{a:} - \Pi_{b:}\|^2 - \|(\Pi_{a:} - \Pi_{b:})Z^{\top}W^{(t)}\|^2\right|.
\end{aligned}$$

Using the fact

$$\begin{aligned}
&\left|\|\Pi_{a:} - \Pi_{b:}\|^2 - \|(\Pi_{a:} - \Pi_{b:})Z^{\top}W^{(t)}\|^2\right| \\
&\qquad \leqslant \left(\left\|Z^{\top}W^{(t)} - I\right\|^2 + 2\left\|Z^{\top}W^{(t)} - I\right\|\right)\|\Pi_{a:} - \Pi_{b:}\|^2,
\end{aligned}$$

we obtain by the proof of part 1 of Lemma 11 that w.h.p., it holds for all $z^{(t)}$ such that $l(z^{(t)}, z) \leqslant \tau^{(t)}$

$$|H_{11}^{(t)}| \lesssim \|\Pi_{a:} - \Pi_{b:}\|^2\frac{Kl(z^{(t)}, z)}{n\Delta_{min}^2}$$

which implies

$$\frac{\lambda^{(t)}|H_{11}^{(t)}|}{\Delta^2(z_i, b)} \lesssim \frac{Kl(z^{(t)}, z)}{n\Delta_{min}^2}.$$

Then as before, for a suitably small constant $\epsilon < 1$, this implies

$$\frac{\lambda^{(t)}|H_{11}^{(t)}|}{\Delta^2(z_i, b)} \leqslant \frac{\delta^{(t+1)}}{16}.$$

By summing all these inequalities we see that $G_{ib}^{(t)}$ and $H_{ib}^{(t)}$ satisfy Condition C3. □

3.7.2.5 *Conclusion*

By Lemma 2, 5 and 6 the Conditions C2 and C3 are satisfied. In order to apply Theorem 4, we also need to show that Condition C1 is satisfied. To this end, let us define the events

$$\Omega(\delta^{(t)}) := \{C_i(a,b) \leqslant -\left(\frac{1-\delta^{(t)}}{2}\right)\Delta^2(a,b)\}, \quad t \geqslant 1,$$

where we recall that

$$C_i(a,b) = \langle e_i, \tilde{\mu}_a - \tilde{\mu}_b\rangle + \lambda\langle E_{i:}W, \tilde{\Pi}_{a:} - \tilde{\Pi}_{b:}\rangle.$$

Note that by definition of $\delta^{(t)}$, we have for all $t \geqslant 1$

$$\Omega(\delta^{(t)}) \subseteq \Omega(\delta^{(1)}) \implies \mathbb{1}_{\Omega(\delta^{(t)})} \leqslant \mathbb{1}_{\Omega(\delta^{(1)})} \implies \xi(\delta^{(t)}) \leqslant \xi(\delta^{(1)}). \tag{3.7.1}$$

Hence it suffices to bound $\xi(\delta^{(1)})$. To achieve this, it is crucial to bound $\mathbb{P}(\Omega(\delta^{(1)}))$ since the bound on $\xi(\delta^{(1)})$ then follows via an easy adaptation of Lemma 1.

We can bound $\mathbb{P}(\Omega(\delta^{(1)}))$ by an immediate adaptation of Lemma 7 – we just need to replace $\Delta^2/4$ by $(1-\delta^{(1)})\Delta^2/4$ in the last step of the lemma, leading to

$$\mathbb{P}(\Omega(\delta^{(1)})) \leqslant \exp(-(1-\delta^{(1)} + o(1))\tilde{\Delta}^2).$$

Then as mentioned earlier, an easy adaptation of Lemma 1 along with (3.7.1) implies that w.h.p., it holds for all $t \geqslant 1$ that

$$\xi(\delta^{(t)}) \leqslant n\exp(-(1-\delta^{(1)} + o(1))\tilde{\Delta}^2).$$

But we know that

$$n\exp(-(1-\delta^{(1)} + o(1))\tilde{\Delta}^2) \leqslant \frac{3}{4}\tau = \frac{3}{4}\tau^{(0)}\delta$$

for n large enough, by a suitable choice of $\delta = o(1)$, and the fact that $\tilde{\Delta}^2 \asymp \log n/K \to \infty$. Hence the assumptions of Theorem 4 are satisfied. Moreover, for all $t \gtrsim \log(1/\delta)$, Corollary 1 yields the bound

$$\begin{aligned} l(z^{(t)}, z) &\lesssim \xi(\delta) + \tau^{(0)}(1/8)^{t-\Theta(\log(1/\delta))} \\ &\lesssim ne^{-(1+o(1))\tilde{\Delta}^2} + \tau^{(0)}(1/8)^{t-\Theta(\log(1/\delta))}. \end{aligned}$$

Since $\tau^{(0)} \lesssim n\log n/K$ and δ can be chosen such that $\log(1/\delta) \lesssim \log n$, we obtain that for $t \gtrsim \log n$

$$l(z^{(t)}, z) \leqslant ne^{-(1+o(1))\tilde{\Delta}^2}.$$

3.7.3 *Proof of Theorem 6*

Proof. The proof follows the same steps as in the proof of Theorem 3.3 in Lu et al., 2016, only the last part needs to be changed. For the sake of completeness, we reproduce the arguments below. Let us denote

$$h(z', z'') = \sum_{i \in [n]} \mathbb{1}_{\{z'(i) \neq z''(i)\}} \tag{3.7.2}$$

to be the unnormalized Hamming distance between $z', z'' \in [K]^n$. Without lost of generality we can assume that

$$\min_{k,k'} \|\mu_k - \mu_{k'}\| = \|\mu_1 - \mu_2\|.$$

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

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

For all $\hat{z} \neq \tilde{z} \in \mathcal{Z}$ we have

$$\frac{h(\hat{z}, \tilde{z})}{n} \leqslant \frac{1}{4}$$

and for all permutations $\sigma \in \mathfrak{S}_K, \sigma \neq Id$ (where Id denotes the identity permutation) we have

$$\frac{h(\sigma(\hat{z}), \tilde{z})}{n} \geqslant \frac{1}{2}.$$

Thus $r(\hat{z}, \tilde{z}) = \frac{h(\hat{z}, \tilde{z})}{n}$. Then following the same arguments as in the proof of Theorem 2 in Gao et al., 2018 we can obtain

$$\inf_{\hat{z}} \sup_{\theta \in \Theta} \mathbb{E}(r(\hat{z}, z)) \geqslant \frac{1}{6|T^c|} \sum_{i \in T^c} \frac{1}{2K^2} \inf_{\hat{z}_i} (\mathbb{P}_1(\hat{z}_i = 2) + \mathbb{P}_2(\hat{z}_i = 1)) \tag{3.7.3}$$

where $\mathbb{P}_k$ denotes the probability distribution of the data when $z_i = k$. By the Neyman Pearson Lemma, the infinimum of the right hand side of (3.7.3) is achieved by the likelihood ratio test. From Section 3.1 in Zhang et al., 2016a, the log-likelihood of the SBM part can be rewritten as

$$\log\left(\frac{p(1-q)}{q(1-p)}\right) \sum_{i<j} A_{ij} \mathbb{1}_{\{z_i = z_j\}} + f(A)$$

where $f(A)$ doesn't depend on z. Consequently,

$$\frac{1}{2} \inf_{\hat{z}_i} (\mathbb{P}_1(\hat{z}_i = 2) + \mathbb{P}_2(\hat{z}_i = 1)) =$$
$$\mathbb{P}(-0.5\|\epsilon_i\|^2 + \log\left(\frac{p(1-q)}{q(1-p)}\right) \sum_{j \in \mathcal{C}_1} A_{ij} \leqslant -0.5\|\mu_1 + \epsilon_i - \mu_2\|^2$$

$$+\log\left(\frac{p(1-q)}{q(1-p)}\right)\sum_{j\in\mathcal{C}_2}A_{ij}) \quad (3.7.4)$$

Let us denote $Z_i = \log(\frac{p(1-q)}{q(1-p)})(\sum_{j\in\mathcal{C}_2}A_{ij} - \sum_{j\in\mathcal{C}_1}A_{ij})$, this is a random variable independent of ϵ_i. So we get

$$\begin{aligned}
(3.7.4) &= \mathbb{P}(0.5\|\mu_1-\mu_2\|^2 - Z_i \leqslant -\langle\epsilon_i, \mu_1-\mu_2\rangle)\\
&\geqslant \mathbb{P}(0.5\|\mu_1-\mu_2\|^2 - Z_i \leqslant -\langle\epsilon_i, \mu_1-\mu_2\rangle \mid Z_i > 0)\mathbb{P}(Z_i > 0)\\
&\geqslant \mathbb{P}(\|\mu_1-\mu_2\|^2 \leqslant -2\langle\epsilon_i, \mu_1-\mu_2\rangle)\mathbb{P}(Z > 0)\\
&\geqslant \exp\left(-\frac{\Delta_1^2}{8}\right)\exp\left(-n\frac{(1+o(1))(\sqrt{p}-\sqrt{q})^2}{K}\right)\\
&\geqslant \exp(-(1+o(1))\tilde{\Delta}^2).
\end{aligned}$$

Here we used for the penultimate inequality a result from the proof of Theorem 3.3 in Lu et al., 2016 and also use Lemma 5.2 in Zhang et al., 2016a. □

3.7.4 *Proof of technical lemmas*

Lemma 7 (Concentration rate for the ideal oracle error under CSSBM)**.** *Recall that*

$$\Omega_1 = \left\{\langle\epsilon_i, \mu_{z_i}-\mu_b\rangle + \lambda\langle E_{i:}W, \Pi_{z_i:}-\Pi_{b:}\rangle \leqslant \frac{-(1-\delta-\bar{\delta})\Delta^2(z_i,b)}{2}\right\}$$

and suppose that $\delta, \bar{\delta} = o(1)$. Then under the assumptions of Theorem 5, we have

$$\mathbb{P}(\Omega_1) \leqslant \exp(-(1+o(1))\tilde{\Delta}^2)$$

where

$$\tilde{\Delta}^2 = \frac{1}{8}\min_{k,k'}\|\mu_k-\mu_{k'}\|^2 + \frac{\log n}{K}(\sqrt{p'}-\sqrt{q'})^2.$$

Proof. We are going to bound the m.g.f of

$$Z = \langle\epsilon_i, \mu_{z_i}-\mu_b\rangle + \lambda\langle E_{i:}W, \Pi_{z_i:}-\Pi_{b:}\rangle$$

and use Chernoff method. We have for all $t < 0$

$$\begin{aligned}
\log\mathbb{E}e^{tZ} &\leqslant \log\mathbb{E}e^{t\langle\epsilon_i,\mu_{z_i}-\mu_b\rangle} + \log\mathbb{E}e^{t\lambda\langle E_{i:}W,\Pi_{z_i:}-\Pi_{b:}\rangle}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(by independence)}\\
&\leqslant \|\mu_{z_i}-\mu_b\|^2\frac{t^2}{2} + \frac{n}{K}\log(pe^{t\lambda(p-q)K/n}+1-p)\\
&+ \frac{n}{K}\log(qe^{-t\lambda(p-q)K/n}+1-q) - t\lambda(p-q)^2.
\end{aligned}$$

By choosing $t = -1/2$ we get

$$e^{t\lambda(p-q)K/n} = \sqrt{\frac{q(1-p)}{p(1-q)}}$$

and thus

$$\log(pe^{t\lambda(p-q)K/n}+1-p)(qe^{-t\lambda(p-q)K/n}+1-q) \\ = \log(pq+(1-p)(1-q)+2\sqrt{pq}\sqrt{(1-p)(1-q)}).$$

$$\log(pe^{t\lambda(p-q)K/n}+1-p)(qe^{-t\lambda(p-q)K/n}+1-q) \\ = \log(pq+(1-p)(1-q)+2\sqrt{pq}\sqrt{(1-p)(1-q)}).$$

This last quantity is equal to $-(1+o(1))(\sqrt{p}-\sqrt{q})^2$. We can now conclude by remarking that

$$\mathbb{P}(\Omega_1) \leqslant \mathbb{P}\left(-\frac{1}{2}\langle \epsilon_i, \mu_{z_i}-\mu_b\rangle - \frac{1}{2}\lambda\langle E_{i:}W, \Pi_{z_i:}-\Pi_{b:}\rangle \geqslant (1+o(1))\frac{\Delta_{min}^2}{4}\right)$$

hence

$$\begin{aligned}\mathbb{P}(\Omega_1) &\leqslant \mathbb{E}e^{-\frac{Z}{2}-\frac{\Delta_{min}^2}{4}} \\ &\leqslant \exp(\frac{\|\mu_{z_i}-\mu_b\|^2}{8} \\ &- \frac{n}{K}(1+o(1))(\sqrt{p}-\sqrt{q})^2+\frac{\lambda}{2}(p-q)^2-(1+o(1))\frac{\Delta_{min}^2}{4}) \\ &\leqslant \exp(-(1+o(1))\tilde{\Delta}^2),\end{aligned}$$

since $\Delta_{min}^2 = \|\mu_{z_i}-\mu_b\|^2+2\lambda(p-q)^2$. □

Lemma 8. *Recall the definition of Ω_1 from the previous lemma. Under the assumptions of Theorem 7 we have*

$$\mathbb{P}(\Omega_1) \leqslant \exp\left(-\frac{1}{8}\Delta_{min}^2\right).$$

Proof. First observe that

$$\begin{aligned}t\lambda\langle E_{i:}W, \Pi_{z_i:}-\Pi_{b:}\rangle &= t\lambda\sum_{k\in[K]}(\Pi_{z_ik}-\Pi_{bk})\frac{\sum_{j\in\mathcal{C}_k}E_{ij}}{n_k} \\ &= \sum_{k\in[K]}t\lambda(\Pi_{z_ik}-\Pi_{bk})\frac{\sum_{j\in\mathcal{C}_k}A_{ij}-\Pi_{z_ik}}{n_k}.\end{aligned}$$

The sum over k involves independent random variables so in order to bound the m.g.f. of $\langle E_{i:}W, \Pi_{z_i:}-\Pi_{b:}\rangle$ it is sufficient to control the m.g.f. of $\sum_{j\in\mathcal{C}_k}A_{ij}-\Pi_{z_ik}$ for each k. Setting $t' = \lambda t\frac{|\Pi_{z_ik}-\Pi_{bk}|}{n_k}$, we have

$$\begin{aligned}\log\mathbb{E}(e^{t'\sum_{j\in\mathcal{C}_k}(A_{ij}-\Pi_{z_ik})}) &= \log(\Pi_{z_ik}e^{t'}+1-\Pi_{z_ik})-n_kt'\Pi_{z_ik} \\ &\leqslant n_k\Pi_{z_ik}(e^{t'}-t'-1)\end{aligned}$$

$$\begin{aligned}
&\leqslant n_k \Pi_{z_i k} \frac{e(t')^2}{2} \\
&\qquad\qquad (\text{ by Taylor-Lagrange inequality}) \\
&\leqslant 1.5 n_k p_{max} (\lambda t \frac{\Pi_{z_i k} - \Pi_{bk}}{n_k})^2 \\
&\leqslant 1.5\lambda |\Pi_{z_i k} - \Pi_{bk}|^2 t^2.
\end{aligned}$$

For the second inequality we used the fact that for $0 < x < 1$, $\log(1 - x) \leqslant -x$.

Consequently,

$$\log \mathbb{E} e^{tZ} \leqslant ||\mu_{z_i} - \mu_b||^2 \frac{t^2}{2} + 1.5\lambda ||\Pi_{z_i:} - \Pi_{b:}||^2 t^2$$

and

$$\mathbb{P}(\Omega_1) \leqslant e^{||\mu_{z_i} - \mu_b||^2 \frac{t^2}{2} + 1.5\lambda ||\Pi_{z_i:} - \Pi_{b:}||^2 t^2 - \frac{\Delta^2(z_i, b)}{4}}.$$

For $t = 1/2$ we then obtain

$$\mathbb{P}(\Omega_1) \leqslant e^{-\frac{||\mu_{z_i} - \mu_b||^2}{8} - \frac{\lambda}{8} ||\Pi_{z_i:} - \Pi_{b:}||^2} \leqslant e^{-\frac{1}{8}\Delta^2_{min}}.$$

□

Recall the Hamming distance h defined in (3.7.2).

Lemma 9. *For all $z, z' \in [K]^n$ we have*

$$h(z, z') \leqslant \frac{l(z, z')}{\Delta^2_{min}}.$$

Proof.

$$\sum_{i \in [n]} \mathbb{1}_{z_i \neq z'_i} \leqslant \sum_{i \in [n]} \frac{\Delta(z_i, z'_i)^2}{\Delta^2_{min}} \mathbb{1}_{z_i \neq z'_i} = \frac{l(z, z')}{\Delta^2_{min}}.$$

□

Lemma 10. *Assume that for some $\alpha > 1$*

$$\frac{n}{\alpha K} \leqslant n_k \leqslant \frac{\alpha n}{K}.$$

If $l(z, z^{(t)}) \leqslant n\Delta^2_{min}/(2\alpha K)$ then for all $k \in [K]$

$$\frac{n}{2\alpha K} \leqslant n_k^{(t)} \leqslant \frac{2\alpha n}{K}.$$

Proof. Since for all $k \in [K]$ we have $n/(\alpha K) \leqslant n_k \leqslant \alpha n/K$,

$$\begin{aligned}
\sum_{i \in \mathcal{C}_k^{(t)}} 1 &\geqslant \sum_{i \in \mathcal{C}_k \cap \mathcal{C}_k^{(t)}} 1 \geqslant \sum_{i \in \mathcal{C}_k} 1 - \sum_{i \in [n]} \mathbb{1}_{z_i \neq z_i^{(t)}} \\
&\geqslant \frac{n}{\alpha K} - h(z, z^{(t)}) \overset{\text{Lemma 9}}{\geqslant} \alpha \frac{n}{K} - \frac{l(z, z^{(t)})}{\Delta^2_{min}} \\
&\geqslant \frac{\alpha n}{2K}
\end{aligned}$$

by assumption. The other inequality is proved in a similar way. □

Lemma 11. *Assume that* $A \sim \mathrm{SBM}(Z, \Pi)$ *with equal size communities and suppose that the conditions of Theorem 7 are satisfied*[5]*. Then with probability at least* $1 - n^{-\Omega(1)}$ *the following holds for all* $z^{(t)}$ *such that* $l(z^{(t)}, z) \leqslant \tau^{(0)}$.

1. $\max_{k\in[K]} \|W^{(t)}_{:k} - W_{:k}\| \lesssim \frac{K^{1.5}}{n^{1.5}\Delta^2_{min}} l(z^{(t)}, z)$,
2. $\max_{k\in[K]} \|(W^{(t)}_{:k} - W_{:k})^\top AW\| \lesssim \frac{K^{1.5}\sqrt{p_{max}}}{n^{1.5}\Delta_{min}} l(z, z^{(t)})$,
3. $\max_{k\in[K]} \|W^{(t)\top}_{:k} A(W - W^{(t)})\| \lesssim \frac{K^2\sqrt{p_{max}}\, l(z^{(t)}, z)}{n^{1.5}\Delta_{min}}$,
4. $\max_{k\in[K]} \|(W^{(t)}_{:k} - W_{:k})^\top AW^{(t)}\| \lesssim \frac{K^{1.5}\sqrt{p_{max}}\, l(z^{(t)}, z)}{n^{1.5}\Delta_{min}}$,
5. $\|Z^\top W^{(t)}\| \lesssim 1$.

Proof. This is a rather straightforward adaptation of Lemma 4 in Han et al., 2020, but for completeness we include a proof adapted to our setting with our notations.

PROOF OF 1. First observe that Z is rank K and $\lambda_K(Z) = \sqrt{n_{min}}$ so that

$$\|W^{(t)}_{:k} - W_{:k}\| \leqslant n_{min}^{-1/2} \|I - Z^\top W^{(t)}\|.$$

For any $k \in [K]$, denote $\delta_k = 1 - (Z^\top W^{(t)})_{kk}$. Since for all $k, k' \in [K]$

$$(Z^\top W^{(t)})_{kk'} = \frac{\sum_{i\in\mathcal{C}_k} \mathbb{1}_{z_i^{(t)}=k'}}{n^{(t)}_{k'}},$$

we have

$$0 \leqslant \delta_k \leqslant 1, \qquad \sum_{k'\in[K]\setminus k} (Z^\top W^{(t)})_{k'k} = \delta_k.$$

Therefore,

$$\begin{aligned}
\|Z^\top W^{(t)} - I\| &= \sqrt{\sum_{k\in[K]} \left(\delta_k^2 + \sum_{k'\in[K]\setminus k} (Z^\top W^{(t)})^2_{k'k} \right)} \\
&\leqslant \sqrt{\sum_{k\in[K]} \left(\delta_k^2 + \left(\sum_{k'\in[K]\setminus k} (Z^\top W^{(t)})_{k'k} \right)^2 \right)} \\
&\leqslant \sqrt{2\sum_{k\in[K]} \delta_k^2} \leqslant \sqrt{2} \sum_{k\in[K]} \delta_k \\
&= \sqrt{2} \sum_{k\in[K]} \frac{\sum_{i\in\mathcal{C}^{(t)}_k} \mathbb{1}_{z_i\neq k}}{n^{(t)}_k} \\
&\leqslant \sqrt{2} \max_k (n^{(t)}_k)^{-1} \sum_{i\in[n]} \mathbb{1}_{z_i \neq z_i^{(t)}}
\end{aligned}$$

5 These assumptions are clearly satisfied by Theorem 5 as well.

$$\overset{\text{Lemma 10}}{\lesssim} \frac{K}{n} h(z, z^{(t)}) \overset{\text{Lemma 9}}{\lesssim} K \frac{l(z, z^{(t)})}{n\Delta_{min}^2}. \tag{3.7.5}$$

PROOF OF 2. Observe that with probability at least $1 - n^{-\Omega(1)}$ we have

$$\begin{aligned}
\max_{k\in[K]} \|(W_{:k}^{(t)} - W_{:k})^\top AW\| &\leqslant \max_{k\in[K]} \|(W_{:k}^{(t)} - W_{:k})^\top PW\| \\
&\quad + \max_{k\in[K]} \|(W_{:k}^{(t)} - W_{:k})^\top EW\| \\
&\leqslant \max_{k\in[K]} \|(W_{:k}^{(t)} - W_{:k})^\top Z\Pi\| \\
&\quad + \|EW\| \max_{k\in[K]} \|(W_{:k}^{(t)} - W_{:k})\| \\
&\leqslant \|\Pi_{b:} - \sum_{j\in\mathcal{C}_b^{(t)}} \frac{\Pi_{z_j:}}{n_b^{(t)}}\| \\
&\quad + C\sqrt{Kp_{max}} \max_{k\in[K]} \|(W_{:k}^{(t)} - W_{:k})\| \\
&\lesssim \|\Pi_{b:} - \sum_{j\in\mathcal{C}_b^{(t)}} \frac{\Pi_{z_j:}}{n_b^{(t)}}\| + \frac{K^2\sqrt{p_{max}}}{n^{1.5}\Delta_{min}^2} l(z^{(t)}, z).
\end{aligned}$$

Recall that $\Delta_2^2(a, b) := \|\Pi_{a:} - \Pi_{b:}\|^2$. Then we have

$$\begin{aligned}
\left\|\Pi_{b:} - \sum_{j\in\mathcal{C}_b^{(t)}} \frac{\Pi_{z_j:}}{n_b^{(t)}}\right\| &= \left\| \sum_{\substack{j\in\mathcal{C}_b^{(t)} \\ b'\in[K]\setminus b}} \frac{\mathbb{1}_{\{z_j=b'\}}}{n_b^{(t)}} (\Pi_{b:} - \Pi_{b':}) \right\| \\
&\leqslant C\frac{K}{n} \sum_{\substack{j\in\mathcal{C}_b^{(t)} \\ b'\in[K]\setminus b}} \max_{b,b'} \Delta_2(b, b') \mathbb{1}_{\{z_j=b'\}} \\
&\leqslant C\frac{K}{n} \max_{b,b'} \Delta_2(b, b') h(t, t^{(t)}) \\
&\leqslant C\frac{K\Delta_{min}}{\sqrt{\lambda} n\Delta_{min}^2} l(z, z^{(t)}) \\
&\quad (\text{ since } \max_{b,b'} \Delta_2(b, b') \lesssim \tfrac{\Delta_{min}}{\sqrt{\lambda}} \text{ for SBM}) \\
&\leqslant C\frac{K^{1.5}\sqrt{p_{max}}}{n^{1.5}\Delta_{min}} l(z, z^{(t)}).
\end{aligned}$$

Consequently, by summing the previous bounds and using the first inequality of the Lemma we get

$$\max_{k\in[K]} \|(W_{:k}^{(t)} - W_{:k})^\top AW\| \lesssim \frac{K^{1.5}\sqrt{p_{max}}}{n^{1.5}\Delta_{min}} l(z, z^{(t)}) + \frac{K^2\sqrt{p_{max}}}{n^{1.5}\Delta_{min}^2} l(z^{(t)}, z).$$

In our setting $\Delta_{min}^2 \asymp \log n$ so the first term is dominant.

PROOF OF 3. First let's bound $\max_{k\in[K]} \|W_{:k}^{(t)\top} P(W - W^{(t)})\|$. By Lemma 10 we have $\|W_k^{(t)}\| \lesssim \sqrt{K/n}$, so

$$\begin{aligned}\max_{k\in[K]} \|W_{:k}^{(t)\top} P(W - W^{(t)})\| &\leqslant \max_{k\in[K]} \|W_{:k}^{(t)\top} Z\|\|\Pi Z^\top (W - W^{(t)})\| \\ &\lesssim \|\Pi Z^\top (W - W^{(t)})\|_F \\ &\lesssim \sqrt{K} \max_{k\in[K]} \|(W_{:k}^{(t)} - W_{:k})^\top Z\Pi\| \\ &\lesssim \frac{K^2\sqrt{p_{max}}}{n^{1.5}\Delta_{min}} l(z, z^{(t)}). \\ &\qquad\qquad \text{(by the proof of part 2)}\end{aligned}$$

We now give an upper bound for $\max_{k\in[K]} \|W_{:k}^{(t)\top} E(W - W^{(t)})\|$. By triangle inequality,

$$\|W_{:k}^{(t)\top} E(W - W^{(t)})\| \leqslant \|W_{:k}^\top E(W - W^{(t)})\| + \|(W_{:k}^{(t)} - W_{k:})^\top E(W - W^{(t)})\|.$$

First we have

$$\begin{aligned}\|W_{:k}^\top E(W - W^{(t)})\| &\leqslant \|W_{:k}\|\|E\|\|(W - W^{(t)})\| \\ &\lesssim \frac{K^2\sqrt{p_{max}}}{n^{1.5}\Delta_{min}^2} l(z^{(t)}, z).\end{aligned}$$

On the other hand, we also have

$$\begin{aligned}\|(W_{:k}^{(t)} - W_{k:})^\top E(W - W^{(t)})\| &\leqslant \|W_{:k} - W_{:k}^{(t)}\|\|E(W - W^{(t)})\| \\ &\leqslant \|E\|\sqrt{K} \max_k \|W_{:k} - W_{:k}^{(t)}\|^2 \\ &\lesssim \frac{K^{3.5}\sqrt{np_{max}} l(z^{(t)}, z)}{n^2\Delta_{min}^2} \frac{l(z^{(t)}, z)}{n\Delta_{min}^2} \\ &\lesssim \frac{K^{3.5}\sqrt{p_{max}} l(z^{(t)}, z)}{n^{1.5}\Delta_{min}^2}\end{aligned}$$

where the last inequality comes from the fact that by assumption $l(z, z^{(t)}) \leqslant \tau \leqslant \epsilon \frac{n\Delta_{min}^2}{K}$.

Thus it follows that

$$\max_{k\in[K]} \|W_{:k}^{(t)\top} P(W - W^{(t)})\| \lesssim \frac{K^2\sqrt{p_{max}} l(z^{(t)}, z)}{n^{1.5}\Delta_{min}}.$$

PROOF OF 4. First note that

$$\begin{aligned}\|(W_{:k}^{(t)} - W_{:k})^\top P W^{(t)}\| &\leqslant \|(W_{:k}^{(t)} - W_{:k})^\top Z\Pi\|\|Z^\top W^{(t)}\| \\ &\lesssim \frac{K^{1.5}\sqrt{p_{max}} l(z^{(t)}, z)}{n^{1.5}\Delta_{min}}.\end{aligned}$$

Furthermore, by the same argument as before,

$$\|(W_{:k}^{(t)} - W_{:k})^\top E W^{(t)}\| \leqslant \|(W_{:k}^{(t)} - W_{:k})^\top E(W^{(t)} - W)\|$$

$$
\begin{aligned}
&+ \|(W_{:k}^{(t)} - W_{:k})^\top E W\| \\
&\lesssim K\|E\| \max_k \|W_{:k} - W_{:k}^{(t)}\|^2 \\
&+ \frac{K^2\sqrt{p_{max}}}{n^{1.5}\Delta_{min}^2} l(z^{(t)}, z) \\
&\lesssim \frac{K^2\sqrt{p_{max}}}{n^{1.5}\Delta_{min}^2} l(z^{(t)}, z).
\end{aligned}
$$

We obtain the result by triangle inequality.

PROOF OF 5. Since $Z^\top W = I_K$ we have

$$
\begin{aligned}
\|Z^\top W^{(t)}\| &\leqslant 1 + \|Z^\top (W^{(t)} - W)\| \\
&\lesssim 1 + \|I - Z^\top W^{(t)}\| \\
&\lesssim 1 + K\frac{l(z^{(t)}, z)}{n\Delta_{min}^2} && \text{(by Equation (3.7.5))} \\
&\lesssim 1. && \text{(by assumption on } \tau^{(0)})
\end{aligned}
$$

□

Lemma 12. *For* `sIR-LS` *we have with probability at least* $1 - n^{-\Omega(1)}$ *that for all* $z^{(t)}$ *such that* $l(z^{(t)}, z) \leqslant \tau^{(0)}$

$$
\max_{k\in[K]} |n_k^{(t)} - n_k| \leqslant \frac{l(z^{(t)}, z)}{\Delta_{min}^2}, \quad |\lambda^{(t)} - \lambda| \lesssim \lambda\frac{Kl(z^{(t)}, z)}{n\Delta_{min}^2}.
$$

Proof. First observe that

$$
\max_{k\in[K]} |n_k^{(t)} - n_k| = \max_{k\in[K]} |\sum_i \mathbb{1}_{\{z_i^{(t)}=k\}} - \mathbb{1}_{\{z_i=k\}}| \leqslant h(z^{(t)}, z).
$$

Moreover by Lemma 9 we have $h(z^{(t)}, z) \leqslant \frac{l(z^{(t)},z)}{\Delta_{min}^2}$.

Then note that we have

$$
\begin{aligned}
|p_{max}^{(t)} - p_{max}| &\leqslant \max_{k,k'} \|(W_{:k}^{(t)})^\top A W_{:k'}^{(t)} - W_{:k}^\top P W_{:k'}\| \\
&\qquad \text{(the max is 1-Lipschitz)} \\
&\leqslant \max_{k,k'} (\|(W_{:k}^{(t)})^\top E W_{:k'}^{(t)}\| + \|(W_{:k}^{(t)} - W_{:k})^\top P W_{:k'}^{(t)}\| \\
&\quad + \|W_{:k}^\top P (W_{:k'}^{(t)} - W_{:k'})\|) \\
&\lesssim \max_k \|W_{:k}^{(t)}\|^2 \|E\| + \frac{K^{1.5}\sqrt{p_{max}} l(z^{(t)}, z)}{n^{1.5}\Delta_{min}} \\
&\qquad \text{(by the proof of Lemma 11)} \\
&\lesssim \frac{K\sqrt{p_{max}}}{\sqrt{n}} + \frac{K^{1.5}\sqrt{p_{max}} l(z^{(t)}, z)}{n^{1.5}\Delta_{min}} \\
&\lesssim \frac{K^{1.5}\sqrt{p_{max}} l(z^{(t)}, z)}{n^{1.5}\Delta_{min}}
\end{aligned}
$$

$$\lesssim \sqrt{\frac{K}{np_{max}}}\frac{Kl(z^{(t)},z)}{n\Delta_{min}}p_{max}$$
$$\lesssim \frac{Kl(z^{(t)},z)}{n\Delta_{min}^2}p_{max}$$

since $\Delta_{min} \asymp \sqrt{\log n/K} \asymp \sqrt{np_{max}/K}$. Consequently,

$$\begin{aligned}\left|\frac{\lambda^{(t)}}{\lambda}-1\right| &\leqslant \left|\frac{n_{min}^{(t)}}{n_{min}}\frac{p_{max}}{p_{max}^{(t)}}-1\right| \\ &\leqslant \left|\frac{n_{min}^{(t)}-n_{min}}{n_{min}}\frac{p_{max}}{p_{max}^{(t)}}\right| + \left|\frac{p_{max}-p_{max}^{(t)}}{p_{max}^{(t)}}\right| \\ &\lesssim \frac{Kl(z^{(t)},z)}{n\Delta_{min}^2}+\frac{Kl(z^{(t)},z)}{n\Delta_{min}^2} \\ &\lesssim \frac{Kl(z^{(t)},z)}{n\Delta_{min}^2}.\end{aligned}$$

□

Lemma 13. *For IR-LSS, we have with probability at least $1-n^{-\Omega(1)}$ that for all $z^{(t)}$ such that $l(z^{(t)},z) \leqslant \tau^{(0)}$*

$$|\lambda^{(t)}-\lambda| \lesssim \lambda\frac{Kl(z^{(t)},z)}{n\Delta_{min}^2}$$

Proof. By a similar argument as the one used in Lemma 12 we have with probability at least $1-n^{-\Omega(1)}$ that

$$|p^{(t)}-p| \lesssim \frac{Kl(z^{(t)},z)}{n\Delta_{min}^2}p, \quad |q^{(t)}-q| \lesssim \frac{Kl(z^{(t)},z)}{n\Delta_{min}^2}q.$$

This implies

$$\frac{p^{(t)}-q^{(t)}}{p-q} = 1+O\left(\frac{(p+q)Kl(z^{(t)},z)}{(p-q)n\Delta_{min}^2}\right) = 1+O\left(\frac{Kl(z^{(t)},z)}{n\Delta_{min}^2}\right)$$

because $p-q \gtrsim p$. Thus,

$$\begin{aligned}\left|\log\left(\frac{p^{(t)}}{q^{(t)}}\right)-\log\left(\frac{p}{q}\right)\right| &= \left|\log\left(\frac{p^{(t)}}{p}\frac{q}{q^{(t)}}\right)\right| \\ &= 2\log\left(1+O\left(\frac{Kl(z^{(t)},z)}{n\Delta_{min}^2}\right)\right) \\ &= O\left(\frac{Kl(z^{(t)},z)}{n\Delta_{min}^2}\right).\end{aligned}$$

Hence

$$\left|\frac{\log(\frac{p^{(t)}}{q^{(t)}})}{\log(\frac{p}{q})}-1\right| = O\left(\frac{K^{1.5}l(z^{(t)},z)}{n\Delta_{min}^2}\right)$$

since $\log(p/q)$ is bounded above by assumption. Consequently,

$$\frac{\lambda^{(t)}}{\lambda} - 1 = O\left(\frac{Kl(z^{(t)}, z)}{n\Delta_{min}^2}\right).$$

□

4

CLUSTERING MULTILAYER GRAPHS WITH MISSING NODES

In this chapter we consider partial side information given by graphs instead of Gaussian covariates as in the previous chapter. This work is based on Braun et al., 2021b where we propose different extensions of existing clustering methods for multilayer graphs to the setting were each node doesn't appear in all the layers. We then complete this work by showing how the method studied in Chapter 1 can be extended to this setting.

The chapter is organized as follows. Section 4.1 motivates the problem and discusses the related work. The generative model is introduced in Section 4.2. In Section 4.3 we propose a final aggregation method for the missing nodes setting. In Section 4.4 we study an early aggregation method based on the sum of adjacency matrices and in Section 4.5 we propose an extension of an intermediate fusion method to the missing nodes setting. We then explain how the method of Chapter 3 can be applied to this problem in Section 4.7. Finally we perform numerical experiments in Section 4.6. Technical results are gathered in Section 4.8

4.1 INTRODUCTION

Graphs are a powerful tool to represent relationships between agents. Due to applications in a wide array of fields including biology, sociology, ecology and economics (see for e.g., Braun et al., 2015; Han et al., 2015; Kivelä et al., 2014; Kim et al., 2015), the analysis of networks has received significant interest over the last two decades. One fundamental problem of network analysis is *clustering* which involves detecting communities by regrouping nodes having similar connectivity properties. Numerous clustering algorithms have been developed over the years based on different approaches such as modularity maximization, maximum likelihood, random walks, semi-definite programming and spectral clustering (see for instance the survey articles by Fortunato, 2010 and Abbe, 2018).

Often, relationships are better understood through different modalities. These multiple aspects of relationships can be represented by a multilayer graph where each layer is a graph representing the interactions between agents for one modality. For e.g., social interaction between a set of people can be recorded via email exchanges, phone calls, professional links, and so on. Each level of interaction can be encoded into a simple graph and the collection of these graphs leads

to a multilayer representation. Another important example of a multilayer graph is given by a time-varying network where each view of the network at a given time corresponds to a different layer.

Over the last decade, many methods have been proposed for clustering multilayer graphs such as those based on matrix factorization, spectral methods, maximisation of a modularity function or probability model-based approaches; see Kim et al., 2015 for a survey. Consistency results for the recovery of the partition under a stochastic generative model have also been shown for some algorithms, see for example Paul et al., 2020, Pensky et al., 2019, Lei et al., 2022 and Bhattacharyya et al., 2018.

Most existing approaches assume that all the layers share the same set of nodes. In practice, however, data are often incomplete; in particular, the set of observed nodes can clearly vary across layers. For example, in social networks evolving over time, the set of nodes can change due to people leaving/joining the network. This is the setting considered in the present paper.

4.1.1 *Related work*

CLUSTERING ON MULTI-LAYER GRAPHS. As noted by Paul et al., 2020, clustering strategies for multilayer graphs can be roughly categorized into three groups: **early fusion** methods where all views are aggregated and then clustering is performed, **intermediate fusion** methods where the algorithm finds a factor common to all the views, and **final aggregation** methods where each individual view is processed separately and a consensus partition is formed. In the complete setting, different algorithms have been proven to be consistent under a multilayer stochastic block model assumption (see Section 4.2.1). Among them are spectral clustering on the sum of adjacency matrices (e.g., Bhattacharyya et al., 2018; Paul et al., 2020) or on the sum of squared adjacency matrices with bias correction (e.g., Lei et al., 2022; Bhattacharyya et al., 2020), orthogonal linked matrix factorization (e.g., Paul et al., 2020), and co-regularized spectral clustering (e.g., Paul et al., 2020). Existing misclustering bounds for these methods are gathered in the supplementary material.

INCOMPLETE MULTI-VIEW CLUSTERING (IMVC). Recently a similar problem has been addressed in the context of IMVC, see for example Liu et al., 2021, Hu et al., 2019 and references therein. To the best of our knowledge, no consistency results for the recovery of the ground truth clustering structure are shown in this line of work. Algorithms designed for the IMVC framework cannot be directly applied to our setting since they apply to a collection of feature vectors. However they could possibly be adapted, in a non trivial manner, to our framework. For example, in the complete setting, the OMVC method

proposed by Hu et al., 2019 can be considered as a variant of the OLMF estimator proposed by Paul et al., 2020 where the optimization problem is modified in order to take into account the symmetry of the inputs. Similarly, if there were no missing views, the algorithm proposed by Liu et al., 2021 resembles a variant of the co-regularized spectral clustering method of Paul et al., 2020 for clustering multilayer graphs. We leave the adaptation of the algorithm proposed by Liu et al., 2021 to our setting for future work.

4.1.2 *Contributions*

We consider the problem of clustering multilayer graphs with missing nodes under a Multi-Layer Stochastic Block Model (MLSBM) described in Section 4.2. Our contributions are as follows.

- In Section 4.3.1 we propose a final aggregation method based on a variant of k-means for incomplete data (Algorithm 5), and derive a bound for the misclustering rate.
- Section 4.4 extends a popular early fusion method – based on spectral clustering applied to the sum of adjacency matrices – to the missing nodes setting. Section 4.4.1 studies this by imputing the missing entries with zeros (Algorithm 6), and contains an upper bound for the misclutering rate. Section 4.4.2 proposes an alternative method (Algorithm 7) wherein the missing entries are imputed iteratively. This method is shown to perform well in our experiments.
- Section 4.5.2 proposes an extension of an intermediate fusion method – namely the Orthogonal Linked Matrix Factorization (OLMF) method studied by Paul et al., 2020 – to the missing nodes setting.
- In Section 4.6 we empirically evaluate our algorithms on synthetic data, and also on real datasets.

4.2 PROBLEM SETUP

A multilayer graph is a sequence of graphs $\mathcal{G} = (\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(L)})$. If all the graphs are defined on the same set of nodes $\mathcal{N}$ indexed by $[n]$, then $\mathcal{G}$ is said to be pillar. Throughout, we will assume that for all $l \leqslant L$ each graph $\mathcal{G}^{(l)}$ is undirected and has no self-loop. This implies that its associated adjacency matrix $A^{(l)} \in \{0,1\}^{n\times n}$ is symmetric with $A^{(l)}_{ii} = 0$ for all i.

Given $\mathcal{G}$ as input, our goal is to recover a partition of $\mathcal{N}$ into K disjoint sets (or communities), so that nodes belonging to the same community share a similar connectivity profile. To make the setup more precise, we will study this problem in the setting where $\mathcal{G}$ is

generated via an underlying (unknown) stochastic model, with a latent community structure. This model is a common extension of the well-studied stochastic block model (SBM) for the unilayer case which we now describe.

4.2.1 *Multilayer Stochastic Block Model (MLSBM)*

We now describe the multilayer stochastic block model (MLSBM), which is a common extension of the SBM to the setting of multilayer graphs (see for e.g., Paul et al., 2020; Bhattacharyya et al., 2018; Lei et al., 2020). The MLSBM is parametrized by the number of layers L, a common block membership matrix $Z \in \mathcal{M}_{n,K}$, and connectivity matrices $\Pi^{(1)}, \ldots, \Pi^{(L)} \in [0,1]^{K \times K}$.

Similar to the unilayer case, let us denote $P^{(l)} = Z\Pi^{(l)}Z^T$ for $l = 1, \ldots, L$. A multilayer graph $\mathcal{G}$ is distributed according to the model $\text{MLSBM}(Z, \Pi^{(1)}, \ldots, \Pi^{(L)})$ if the adjacency matrix $A^{(l)}$ of each layer is distributed according to a $\text{SBM}(Z, \Pi^{(l)})$ for $l = 1, \ldots, L$. Hence, while the probability that two nodes are connected can vary across layers, the block membership of each node remains unchanged. As in the unilayer case we can define the quantities $p^{(l)}_{max} = \max_{i,j} p^{(l)}_{ij}$, $p_{max} = \max_l p^{(l)}_{max}$.

4.2.2 *Missing nodes*

The assumption that all the layers share the same set of nodes is quite restrictive since real world multilayer networks are often 'non-pillar'. We propose to deal with such networks by considering nodes present in some layers but not in others as missing. Let $w^{(l)}_i$ be a binary variable that records the presence of node i in the layer l where $w^{(l)}_i = 1$ if node i is observed in layer l and 0 otherwise. Denoting $w^{(l)} = (w^{(l)}_1, \ldots, w^{(l)}_n)^T$, let $\Omega^{(l)} = w^{(l)}(w^{(l)})^T$ be the mask matrices and $\tilde{A}^{(l)} = A^{(l)} \odot \Omega^{(l)}$ for $l \leqslant L$ where $\odot$ is the usual Hadamard product. Let J_l denote the set of non-missing nodes in layer l with $n_{J_l} = |J_l|$. By a slight abuse of notation we will denote by A_{J_l} the matrix $A^{(l)}_{J_l}$. The number of observed nodes in $\mathcal{C}_k$ will also be denoted by $n_{J_l,k}$. Throughout, we assume that the missing nodes are generated as $w^{(l)}_i \overset{\text{ind.}}{\sim} \mathcal{B}(\rho)$ for $i = 1, \ldots, n$.

4.3 FINAL AGGREGATION METHODS

A natural way to extend unilayer graph clustering to the multilayer setting is to analyze each layer separately and then find a consensus partition – such approaches are referred to as final aggregation methods. For example, one can apply any clustering method on each

individual layer, take one layer's labels as a reference, find for each remaining layer the permutation of its labels that maximizes the agreement with the reference layer, and then define a consensus community by majority voting as discussed in Han et al., 2015. There exist alternative ways to avoid the cumbersome issue of label switching ambiguity such as the 'aggregate spectral kernel' considered in Paul et al., 2020. Such methods rely on the quality of each individual layer and are often empirically outperformed by other methods as shown in Paul et al., 2020; Han et al., 2015.

Final aggregation methods are still relevant in the missing nodes context. Indeed, if we have exact recovery for each layer, and if for all k there is at least one common node between two layers belonging to $\mathcal{C}_k$, then we can easily reconstruct the whole partition even when the set of common nodes is very small. Hence such methods can be considered as baseline methods.

4.3.1 *A method based on a variant of k-means for incomplete data*

We now propose a final aggregation method for clustering multilayer graphs in the incomplete setting; it avoids the aforementioned label switching problem.

For each layer l, we can compute the matrix $\hat{U}_{J_l}$ of size $|J_l| \times K$ corresponding to the eigenvectors associated with the top K eigenvalues (in absolute value) of $A_{J_l} \in \mathbb{R}^{|J_l| \times |J_l|}$. The matrix $\hat{U}_{J_l}$ can be transformed to a matrix $\hat{U}^{(l)}$ of size $n \times K$ by completing with zeros the rows of the nodes that haven't been observed[1]. Let $\hat{U}$ be the $n \times KL$ matrix obtained by stacking $\hat{U}^{(l)}$.

Analogously, let U_{J_l} be the matrix formed by the K eigenvectors corresponding to non-zero eigenvalues of $Z_{J_l} \Pi^{(l)} Z_{J_l}^T$, $U^{(l)}$ be the $n \times K$ matrix obtained from U_{J_l} by filling the rows corresponding to unobserved nodes with the row corresponding to an observed node (belonging to the same community), and U be the matrix obtained by stacking all the matrices $U^{(l)}$. For each l, let O_l be a $K \times K$ orthogonal matrix such that

$$O_l \in \underset{O^T O = I_k}{\operatorname{argmin}} \|\hat{U}_{J_l} - U_{J_l} O\|_F.$$

As in the unilayer setting, k-means could be applied on the rows of $\hat{U}^{(l)}$ in order to recover the community structure for each l. But in order to avoid the label switching problem we propose to apply on the rows of $\hat{U}$ a variant of k-means described in Chi et al., 2015 that can handle missing values, see Algorithm 5.

1 It is easy to verify that $\hat{U}^{(l)}$ is also the eigenvector matrix corresponding to the top K eigenvalues (in absolute value) of $A^{(l)} \odot \Omega^{(l)}$.

Let us describe the principle behind this algorithm. The classical k-means problem seeks a partition Z and centroid values (encoded in the matrix C) that solves

$$\min_{\substack{Z\in\mathcal{M}_{n,K}\\ C\in\mathbb{R}^{K\times KL}}} \|\hat{U} - ZM\|_F^2.$$

When there are missing values one can instead solve

$$\min_{\substack{Z\in\mathcal{M}_{n,K}\\ C\in\mathbb{R}^{K\times KL}}} \|(\hat{U} - ZM)\odot\Omega_U\|_F^2 \tag{4.3.1}$$

where $\Omega_U = (w^{(1)}\otimes \mathbf{1}_K \ \cdots \ w^{(L)}\otimes \mathbf{1}_K)$ is the $n\times KL$ mask matrix with $\mathbf{1}_K \in \mathbb{R}^{1\times K}$ denoting the all ones vector. It is a matrix composed of L blocks where the rows of each block are 1 if the corresponding node is observed and 0 otherwise.

Algorithm 5 k-pod clustering

Input: The number of communities K, the sets J_l and the adjacency matrices A_{J_l}.

1: Form $\hat{U}^{(l)}$ from A_{J_l} as explained at the beginning of Section 3.1.
2: Form the matrix $\hat{U}$ by stacking the matrices $\hat{U}^{(l)}$.
3: Initialize the partition $\hat{Z}$ and the centroid matrix $\hat{M}$.
4: **repeat**
5: Replace $\hat{U}$ by $\hat{U}\odot\Omega_U + (\hat{Z}\hat{M})\odot(\mathbf{1}\mathbf{1}^T - \Omega_U)$.
6: Apply k-means on the complete matrix $\hat{U}$ and update $\hat{M}$ and $\hat{Z}$.
7: **until** convergence.

Output: A partition of the nodes $\mathcal{N} = \cup_{i=1}^K \mathcal{C}_i$ based on $\hat{Z}$.

In the worst case, the complexity of the algorithm is $O((L+K)n^2)$. But in practice the layers are often sparse and so the complexity will be much less[2].

Theorem 8. *Consider the missing nodes MLSBM in Section 4.2.2, and suppose that* $\rho L \geqslant 1$, $KL \leqslant C_0 n$, $\rho n_{min} \geqslant C_1 K^2 \max(\log^2 n, \sqrt{np_{max}})$ *and* $np_{max}^{(l)} \geqslant C_2\rho^{-1}\log n$. *Let* $\lambda_K^{(l)}$ *be the* K*-th largest singular value of* $\Pi^{(l)}$ *and recall that* $\beta = n_{max}/n_{min}$. *If*

$$\frac{1}{\rho Ln}\sum_l \frac{p_{max}^{(l)}}{(\lambda_K^{(l)})^2} < (30C_3\beta^4K^3)^{-1}$$

then with probability at least $1 - O(n^{-1})$, *it holds that the solution* $\hat{Z}\in\mathcal{M}_{n,K}$ *of* (4.3.1) *satisfies*

$$r(\hat{Z}, Z) \leqslant C_4\exp(-c'\rho L) + \frac{C_5\beta^3K^2}{\rho Ln}\sum_l \frac{p_{max}^{(l)}}{(\lambda_K^{(l)})^2}.$$

2 This remark regarding the complexity applies to our other methods as well.

Remark 8. *The assumption* $\rho L \geqslant 1$ *is natural since* ρL *corresponds to the expected total number of times a node is observed, and a node needs to be observed at least once in order to be classified. The condition* $\rho n_{min} \geqslant C_1 K^2 \log^2 n$ *ensures that* ρ *and* n_{min} *are not too small. If the communities are well-balanced and the parameters* ρ *and* K *are fixed independently of* n*, then the previous condition is satisfied for* n *large enough.*

Remark 9. *Our analysis assumes that each layer is sufficiently informative, and doesn't use the fact that there is more information contained in the whole set of layers than in individual layers. This is why the bound does not improve when* L *increases. The obtained upper-bound is unlikely to be optimal since as shown in the experiments, the clustering performance does seem to improve a bit when* L *increases.*

Proof. Let $\hat{Z}$ and $\hat{C}$ be solutions of the optimization problem (4.3.1) and write $\bar{U} := \hat{Z}\hat{C}$. Define U' as the block matrix obtained by stacking the matrices $U^{(l)}O_l$, $L_i = \{l \in [L] : i \in J_l\}$ be the indices of layers where the node i appears, and $\mathcal{N}_u = \{i : |L_i| \geqslant \rho L/c\}$ where $c > 1$ is a constant that will be fixed later. Let $\mathcal{S}_k$ be the set of 'bad nodes' defined as

$$\mathcal{S}_k := \{i \in \mathcal{C}_k \cap \mathcal{N}_u : \forall l \in L_i, \|U^{(l)}_{i*}O_l - \bar{U}^{(l)}_{i*}\| \geqslant \delta^{(l)}_k/2\}$$

where

$$\delta^{(l)}_k := \min_{\substack{i\in\mathcal{C}_k \\ i'\in\mathcal{C}_{k'} \\ k'\neq k}} \|U^{(l)}_{i'*} - U^{(l)}_{i*}\| = \min_{\substack{i\in\mathcal{C}_k \\ i'\in\mathcal{C}_{k'} \\ k'\neq k}} \|U^{(l)}_{i'*}O_l - U^{(l)}_{i*}O_l\|$$

is the smallest distance between two rows of $U^{(l)}$ corresponding to different communities. Let $\mathcal{T}_k := (\mathcal{C}_k \setminus \mathcal{S}_k) \cap \mathcal{N}_u$ be the complement of $\mathcal{S}_k$ in $\mathcal{N}_u \cap \mathcal{C}_k$ and $\mathcal{T} = \cup_k \mathcal{T}_k$.

Step 1. First let us show by contradiction that if for all k, $|\mathcal{T}_k| > n_k/30$ and n_k satisfies the assumptions of the theorem, then all the nodes in $\mathcal{T}$ are well classified with probability at least $1 - O(n^{-1})$. Assume that there exist $i \in \mathcal{T}_k$ and $j \in \mathcal{T}_{k'}$ such that $\bar{U}_i = \bar{U}_j$. If $L_i \cap L_j \neq \varnothing$, every $l \in L_i \cap L_j$ satisfies

$$\begin{aligned}
\max(\delta^{(l)}_k, \delta^{(l)}_{k'}) &\leqslant \|U^{(l)}_{i*} - U^{(l)}_{j*}\| \\
&\leqslant \|U^{(l)}_{i*} - \bar{U}^{(l)}_{i*}\| + \|U^{(l)}_{j*} - \bar{U}^{(l)}_{j*}\| \\
&< \frac{\delta^{(l)}_k}{2} + \frac{\delta^{(l)}_{k'}}{2}
\end{aligned}$$

contradicting the fact that $i \in \mathcal{T}_k$ and $j \in \mathcal{T}_{k'}$. It remains to treat the case $L_i \cap L_j = \emptyset$. Let C_1 be a cluster induced by $\bar{U}$ containing the nodes i and j. If there were other nodes belonging to $\mathcal{C}_k$ and $\mathcal{C}_{k'}$ but appearing in a common layer, the previous argument can be used to obtain a contradiction. So we can assume that all the nodes of community $\mathcal{C}_k$ in C_1 and all nodes of community $\mathcal{C}_{k'}$ in C_1 appear

on distinct layers. We are going to show this property implies that for all k the size of $\mathcal{C}_k \cap C_1$, and thus the size of C_1, is small with high probability. Let l_1 be a layer where a node in $\mathcal{C}_{k'} \cap C_1$ appears. The probability that none of the nodes in $\mathcal{C}_k \cap C_1$ appears in l_1 is $(1-\rho)^{|\mathcal{C}_k \cap C_1|}$ and this probability is $O(1/n^2)$ if $|\mathcal{C}_k \cap C_1| \geqslant 2\rho^{-1} \log n$ (we used the fact that $-\log(1-\rho) \geqslant \rho$). By symmetry, the result holds for every k such that $|\mathcal{C}_k \cap C_1| > 0$. Therefore we can assume that $|C_1 \cap \mathcal{C}_k| \leqslant 2\rho^{-1} \log n$. Since for all k, $|\mathcal{T}_k| \geqslant n_k/30 \geqslant 3K^2\rho^{-1} \log n$ by assumption, there are nodes in $\mathcal{T}_k$ and $\mathcal{T}_{k'}$ that are not in C_1. Hence there is another cluster C_2 induced by $\hat{U}$ containing nodes from two different communities. The same argument can be applied to C_2 and iteratively to $C_3, \dots, C_K$. At the end, since the C_k form a partition of the set of nodes, we obtain

$$|\mathcal{T}_{k'}| = \sum_k |C_k \cap \mathcal{T}_{k'}| \leqslant 2K^2\rho^{-1} \log n$$

contradicting the fact that $|\mathcal{T}_k| \geqslant 3K^2\rho^{-1} \log n$.

We are now going to show that under the assumptions of the theorem, for all k, $\mathcal{T}_k$ satisfies $|\mathcal{T}_k| > n_k/30$ with probability at least $1 - O(n^{-1})$. In order to prove this result we will first show that $|\mathcal{S}_k|$ is small (Step 2) and then show that $\mathcal{N}_u \cap \mathcal{C}_k$ is large (Step 3).

Step 2. Observe that if $i \in \mathcal{S}_k$ then $\forall l \in L_i$, $4(\delta_k^{(l)})^{-2}\|(U^{(l)}O_l)_{i*} - \bar{U}_{i*}^{(l)}\|^2 \geqslant 1$. So for all k,

$$\begin{aligned} |\mathcal{S}_k|\delta_k^2 &\leqslant 4 \sum_{i \in \mathcal{C}_k \cap \mathcal{N}_u} \min_{l \in L_i} \|(U^{(l)}O_l)_{i*} - \bar{U}_{i*}^{(l)}\|^2 \\ &\leqslant 4 \sum_{i \in \mathcal{C}_k \cap \mathcal{N}_u} \frac{\sum_{l \in L_i} \|(U^{(l)}O_l)_{i*} - \bar{U}_{i*}^{(l)}\|^2}{|L_i|} \end{aligned} \tag{4.3.2}$$

where we used the fact $\delta_k^{(l)} \geqslant \delta_k$ for the first inequality, and the fact that the minimum is always bounded by the mean for the second inequality.

By summing over k, and using the fact that $|L_i| \geqslant \rho L/c$ for $i \in \mathcal{N}_u$, we get

$$\begin{aligned} \sum_k |\mathcal{S}_k|\delta_k^2 &\leqslant \frac{4c}{\rho L} \sum_{i \in \mathcal{N}_u} \sum_{l \in L_i} \|(U^{(l)}O_l)_{i*} - \bar{U}_{i*}^{(l)}\|^2 \\ &\leqslant \frac{C}{\rho L} \|(U' - \bar{U}) \odot \Omega_U\|_F^2. \end{aligned} \tag{4.3.3}$$

Using triangular inequality we get

$$\begin{aligned} \|(U' - \bar{U}) \odot \Omega_U\|_F^2 &\leqslant \|(U' - \hat{U}) \odot \Omega_U\|_F^2 + \|(\hat{U} - \bar{U}) \odot \Omega_U\|_F^2 \\ &\leqslant 2\|(\hat{U} - U') \odot \Omega_U\|_F^2 \end{aligned} \tag{4.3.4}$$

where the second inequality follows from the fact that U' is feasible for (4.3.1), i.e., it can be written as a product of a membership matrix Z and a centroid matrix $C \in \mathbb{R}^{K \times KL}$.

Notice that

$$\|(\hat{U} - U') \odot \Omega_U\|_F^2 = \sum_l \|\hat{U}_{J_l} - U_{J_l} O_l\|_F^2.$$

Let λ_{K,J_l} be the Kth largest singular value of $Z_{J_l} \Pi^{(l)} Z_{J_l}^T$. This last quantity depends on the missing patterns, but the concentration results established in Lemma 14 shows that for all l, $n_{J_l} \leqslant 1.5\rho n$ with probability at least $1 - O(n^{-1})$ and Lemma 17 applied with Z_{J_l} instead of Z and $n_{J_l,min}$ instead of n_{min} shows that $\lambda_{K,J_l} \geqslant n_{J_l,min}\lambda_K^{(l)} \geqslant 0.5\rho n_{min}\lambda_K^{(l)}$ with probability at least $1 - O(n^{-1})$. The concentration inequality used in Lemma 16 and Lemma 14 shows that with probability at least $1 - O(n^{-1})$, $\|A_{J_l} - \mathbb{E}(A_{J_l})\| \leqslant C\sqrt{n_{J_l} p_{max}^{(l)}} \leqslant C\sqrt{\rho n p_{max}^{(l)}}$. But $\rho n_{min}\lambda_K^{(l)} \geqslant 4C\sqrt{\rho n p_{max}^{(l)}}$ for all l due to our assumptions. Moreover, since with high probability, $n_{J_l} p_{max}^{(l)} \geqslant c \log n$ for each l (using the fact that w.h.p, $n_{J_l} \geqslant c'\rho n$ for each l, the condition in the theorem statement suffices), hence Lemma 16 applies and we get that for for each l that with probability $1 - O(n^{-2})$

$$\|\hat{U}_{J_l} - U_{J_l} O_l\|_F^2 \leqslant \frac{C\|A_{J_l} - \mathbb{E}(A_{J_l})\|_F^2}{\lambda_{K,J_l}^2} \leqslant CK\frac{n_{J_l} p_{max}^{(l)}}{\lambda_{K,J_l}^2}. \tag{4.3.5}$$

So by Lemma 17 and Lemma 14 there exists $C > 0$ such that with probability at least $1 - O(Ln^{-2})$ (via union bound), we have for all $l \leqslant L$ that

$$\frac{n_{J_l} p_{max}^{(l)}}{\lambda_{K,J_l}^2} \leqslant C\frac{np_{max}^{(l)}}{\rho(n_{min}\lambda_K^{(l)})^2}. \tag{4.3.6}$$

Plugging equations (4.3.3), (4.3.4), (4.3.6) and (4.3.5) into (4.3.2) we obtain with probability at least $1 - O(n^{-1})$

$$\sum_k |\mathcal{S}_k|\delta_k^2 \leqslant CK \sum_l \frac{np_{max}^{(l)}}{\rho^2 L(n_{min}\lambda_K^{(l)})^2}.$$

We have $\delta_k = \min_l \delta_k^{(l)} = \min_l \sqrt{\frac{1}{n_{k,J_l}}}$ by Lemma 2.1 in Lei et al., 2015. Moreover $\min_l \sqrt{\frac{1}{n_{k,J_l}}} \geqslant \frac{c}{\sqrt{\rho n_k}}$ with probability at least $1 - O(n^{-1})$ by Lemma 14 since $\rho n_k \geqslant C \log^2 n$ by assumption. Thus we obtain

$$\sum_k |\mathcal{S}_k| \leqslant \sum_k |\mathcal{S}_k|(c^{-1}\sqrt{\rho n_k})^2(\delta_k)^2 \leqslant CKn_{max} \sum_l \frac{np_{max}^{(l)}}{\rho L(n_{min}\lambda_K^{(l)})^2}.$$

Observe that $\frac{n_{max}}{n} \leqslant \frac{\beta}{K}$. If

$$\sum_l \frac{np_{max}^{(l)}}{\rho L(n_{min}\lambda_K^{(l)})^2} < (30C\beta^2 K)^{-1},$$

then $|\mathcal{S}_K| < n_k/30$ for all k By using $n_{min} \geqslant \frac{n}{\beta K}$ this last condition can be simplified as

$$\frac{1}{\rho L n} \sum_l \frac{p_{max}^{(l)}}{(\lambda_K^{(l)})^2} < (30C\beta^4K^3)^{-1}.$$

Step 3. We are now going to show that $|\mathcal{N}_u \cap \mathcal{C}_k|$ is large. Let $p(\rho, L) = \mathbb{P}(|L_i| < \rho L/c)$. For the choice $c = 25$, we always have $p < 8/10$ since $\rho L \geqslant 1$ by assumption. Chernoff bound (Lemma 32) shows that $p(\rho, L) \leqslant e^{-\rho L(1-c^{-1})/3}$. If $\rho L > 12 \log n$ then with probability at least $1 - O(n^{-2})$, $\mathcal{N}_u = \mathcal{N}$ and $|\mathcal{N}_u^c| = 0$. Let us assume that $\rho L < 12 \log n$. The number of nodes in $\mathcal{N}_u^c \cap \mathcal{C}_k$ can be written as a sum n_k independent Bernoulli variables with parameter $p = p(\rho, L)$ (we will omit the dependence on ρ and L in the following for notation convenience):

$$|\mathcal{N}_u^c \cap \mathcal{C}_k| = \sum_{i \leqslant n_k} b_i.$$

In expectation $\mathbb{E}(|\mathcal{N}_u^c \cap \mathcal{C}_k|) = pn_k$ and Hoeffding's bound implies that $\mathbb{P}(||\mathcal{N}_u^c \cap \mathcal{C}_k| - pn_k| \geqslant t) \leqslant 2e^{-t^2/n_k}$ for any choice of $t > 0$. So we can take $t = C\sqrt{n_k \log n} = o(n_k)$ and obtain that with probability at least $1 - O(Kn^{-2})$ for all k

$$|\mathcal{N}_u^c \cap \mathcal{C}_k| \leqslant n_k p + C\sqrt{n_k \log n}.$$

Thus $|\mathcal{C}_k \cap \mathcal{N}_u| \geqslant n_k(1 - p - \sqrt{\frac{C \log n}{n_k}})$. If n is large enough, then $\sqrt{\frac{C \log n}{n_k}} < 1/30$.

Since the sets $\mathcal{S}_k$ have cardinalities at most $\frac{n_k}{30}$ we obtain that $|\mathcal{T}_k| \geqslant \frac{5n_k}{30}$.

Conclusion. Steps 1,2 and 3 show that all nodes that belong to $\mathcal{T}_k$ are well classified with probability at least $1 - O(n^{-1})$. Hence the number of misclustered nodes is bounded by the sum of the cardinalities of $\mathcal{S}_k$ plus $|\mathcal{N}_u^c|$. So with probability at least $1 - O(n^{-1})$ we get

$$\begin{aligned} r(\hat{Z}, Z) &\leqslant \frac{1}{n}(|\mathcal{N}_u^c| + \sum_k |\mathcal{S}_k|) \\ &\leqslant \frac{31}{30} p(\rho, L) + C\beta \sum_l \frac{n p_{max}^{(l)}}{\rho L (n_{min} \lambda_K^{(l)})^2} \\ &\leqslant C \exp(-c' \rho L) + \frac{C\beta^3 K^2}{\rho L n} \sum_l \frac{p_{max}^{(l)}}{(\lambda_K^{(l)})^2}. \end{aligned}$$

□

4.4 EARLY FUSION METHODS: SPECTRAL CLUSTERING ON SUM OF ADJACENCY MATRICES

Late fusion methods rely heavily on the quality of each layer. However, by simultaneously using all the information contained in all lay-

ers, the clustering performance can be improved in some settings (see the numerical experiments in Paul et al., 2020 or Han et al., 2015). One way to do this is to aggregate the information across layers and then apply a suitable clustering method. This approach will be referred to as an early fusion method. One simple but popular way to do this is to take the mean of the adjacency matrices (see for e.g., Bhattacharyya et al., 2018; Paul et al., 2020). Then, the k-means algorithm can be applied to the rows of the $n \times K$ eigenvector matrix associated with the top K eigenvalues (in absolute value) of $A = L^{-1} \sum_l A^{(l)}$.

4.4.1 *Imputing missing entries with zeros*

A natural way to extend the aforementioned approach to the setting of missing nodes is to fill the missing entries with zeros, thus leading to Algorithm 6. The worst-case complexity of the algorithm is $O((L + K)n^2)$.

Algorithm 6 Sum of adjacency matrices with missing entries filled with zeros

Input: The number of communities K, the matrices $A^{(l)}$ and $\Omega^{(l)}$.

1: Compute $A = L^{-1} \sum_l A^{(l)} \odot \Omega^{(l)}$.
2: Compute the eigenvectors $u_1, \ldots, u_K$ associated with the K largest eigenvalues of A (ordered in absolute values) and form $U_K = [u_1\ u_2\ \cdots\ u_K]$.
3: Apply `k-means` on the rows of U_K to obtain a partition of $\mathcal{N}$ into K communities.

Output: A partition of the nodes $\mathcal{N} = \cup_{i=1}^{K} \mathcal{C}_i$.

Let us denote $\tilde{A} = \rho^{-2} L^{-1} \sum_l A^{(l)} \odot \Omega^{(l)}$ (clustering on A or $\tilde{A}$ is equivalent since the two matrices are proportional, but for the analysis it is more convenient to work with $\tilde{A}$). Since the diagonal entries of $A^{(l)}$ are zero, $\mathbb{E}(\tilde{A}) = L^{-1} \sum_l \mathbb{E}(A^{(l)})$. Denote by $\mathbb{E}(X|\Omega)$ to be the expectation of X conditionally on $\Omega = (\Omega^{(1)}, \ldots, \Omega^{(L)})$ and let λ_K denote the Kth largest singular value of $\mathbb{E}(\tilde{A})$. We have $\mathbb{E}(\tilde{A}|\Omega) = \rho^{-2} L^{-1} \sum_l \mathbb{E}(A^{(l)}) \odot \Omega^{(l)}$. Using the same kind of perturbation arguments and concentration inequalities as in Lei et al., 2015, we can relate $\tilde{A}$ to $\mathbb{E}(\tilde{A}|\Omega)$ and then use Bernstein inequality to relate $\mathbb{E}(\tilde{A}|\Omega)$ with $\mathbb{E}(\tilde{A})$. This leads to the following bound on the misclustering rate.

Theorem 9. *Under the missing nodes MLSBM in Section 4.2.2, there exist constants $C_0, C_1 > 0$ such that with probability at least $1 - O(n^{-1})$, the solution $\hat{Z} \in \mathcal{M}_{n,K}$ obtained from Algorithm 6 satisfies*

$$r(\hat{Z}, Z) \leqslant \underbrace{\frac{C_0 K}{\rho^4 \lambda_K^2}\left(\frac{np_{max}}{L} + \frac{\log n}{L}\right)}_{\text{noise error}} + \underbrace{C_1 K \frac{(\rho^{-2}-1)^2}{\lambda_K^2}\left((np_{max})^2 \frac{\log(n)}{L} + \left(\frac{np_{max}\log n}{L}\right)^2\right)}_{\text{missing data error}}.$$

If L is small then the missing data error could be larger than one making the upper bound trivial. In the best case scenario, we expect that λ_K scales as np_{max}. So we need at least $C \log n$ layers to get a non trivial upper bound. In order to obtain asymptotic consistency, it is necessary that $L \gg \log n$. However, experiments show that even when L is small, Algorithm 6 gives good results as long as the layers are dense enough and the number of missing nodes is not too large.

When $\rho = 1$ and $np_{max} \geqslant \log n$, $r(\hat{Z}, Z) = O(\frac{1}{Lnp_{max}})$. That improves the one obtained by Bhattacharyya et al., 2018 under a more general setting. See the supplementary material for other comparisons.

In order to prove Theorem 9, we are going to show that $\tilde{A}$ is close to $\mathbb{E}(\tilde{A}|\Omega)$ with high probability for every realization of Ω and that $\mathbb{E}(\tilde{A}|\Omega)$ concentrates around $\mathbb{E}(\tilde{A})$ if L is large enough. These results are summarized in the following proposition.

Proposition 1. *There exist constants c_1 and c_2 such that the following holds.*

1. $\mathbb{P}\left(\|\tilde{A} - \mathbb{E}(\tilde{A}|\Omega)\| \geqslant c_1 \rho^{-2}\left(\sqrt{\frac{np_{max}}{L}} + \sqrt{\frac{\log n}{L}}\right)|\Omega\right) \leqslant n^{-1}$;
2. $\|\mathbb{E}(\tilde{A}|\Omega) - \mathbb{E}(\tilde{A})\| \leqslant c_2(\rho^{-2}-1)\left[np_{max}\left(\sqrt{\frac{\log n}{L}} + \frac{\log n}{L}\right)\right]$ *with probability at least* $1 - o(n^{-1})$.

Proof. The proof of the first statement is the same as the proof of the corresponding inequality if there are no missing values. Since we reason conditionally to the missingness mechanism, the zero entries of $\tilde{A}$ can also be considered as the realization of independent Bernoulli variables with parameter zero.

Let $E = \rho^2(\tilde{A} - \mathbb{E}(\tilde{A}|\Omega))$ and E' be an independent copy of E. Define $E^s = E - E'$ as the symmetrized version of E. Jensen's inequality implies that $\|E\| = \|\mathbb{E}(E - E'|E)\| \leqslant \mathbb{E}(\|E^s\| \mid E)$, so it is enough to control $\|E^s\|$.

The ψ_2 norm (see for example Vershynin, 2016, Proposition 1.2.1) of each entry of E^s is bounded by $K_L := C\sqrt{L^{-1}}\mathcal{K}$ where

$$\mathcal{K} = \max_{i,j,l} \|A_{ij}^{(l)}\|_{\psi_2}$$

and $A_{ij}^{(l)}$ are centered Bernoulli random variables with parameters $p_{ij}^{(l)}$. By definition of the ψ_2 norm there exists a constant c_0 such that for each $i, j \leqslant n$

$$\mathbb{P}(|E_{ij}^s| \geqslant c_0 K_L \sqrt{\log n}) \leqslant n^{-4}.$$

Define $T_{ij} = E_{ij}^s \mathbb{1}_{|E_{ij}^s| \leqslant c_0 K_L \sqrt{\log n}}$ and let $T = (T_{ij}) \in \mathbb{R}^{n \times n}$. By a union bound argument the matrix $E^s - T$ has entries that are not zero with probability at most n^{-2}, thus $\|E^s\| = \|T\|$ with probability at least $1 - O(n^{-2})$. Since the entries of E^s are symmetric, the matrix T is centered and has entries bounded by $c_0 K_L \sqrt{\log n}$ by construction. So we can apply the bound from Lemma 15 to T and obtain

$$\|T\| \leqslant C\sqrt{\frac{np_{max}}{L}} + K_L \log n$$

with probability at least $1 - O(n^{-1})$. We can use the following theorem to get a sharp bound for K_L.

Theorem 10 (Buldygin et al., 2013, Theorem 2.1, Lemma 2.1 (K6))**.** *Let Y be a centered Bernoulli random variable with parameter p, i.e., $Y = 1 - p$ with probability p, and $Y = -p$ with probability $1 - p$. Then,*

$$\|Y\|_{\psi_2}^2 = \begin{cases} 0\,; & p \in \{0, 1\}, \\ 1/4\,; & p = 1/2, \\ \frac{1-2p}{2\log(\frac{1-p}{p})}\,; & p \in (0,1) \setminus \left\{\frac{1}{2}\right\}. \end{cases}$$

In particular, it holds that $\|Y\|_{\psi_2} \leqslant \frac{1}{\sqrt{2|\log(\min\{2p, 2(1-p)\})|}}$.

If $np_{max} \leqslant \log^2 n$, then $K_L \leqslant C(L \log n)^{-1/2}$ and we obtain the first part of the proposition by dividing by ρ^2. If $np_{max} \geqslant \log^2 n$ then we can bound use the trivial bound $K_L \leqslant CL^{-1/2}$ to see that $K_L \log n \leqslant C\sqrt{\frac{np_{max}}{L}}$. Hence

$$\|\tilde{A} - \mathbb{E}(\tilde{A}|\Omega)\| \leqslant C\rho^{-2}\left(\sqrt{\frac{np_{max}}{L}} + \sqrt{\frac{\log n}{L}}\right)$$

with probability at least $1 - O(1/n)$ for all Ω.

It remains to bound the difference between $\mathbb{E}(\tilde{A}|\Omega)$ and $\mathbb{E}(\tilde{A})$. We do so using the matrix Bernstein inequality (Lemma 34). Let $X_l := \rho^{-2}\mathbb{E}(A^{(l)}) \odot \Omega^{(l)} - \mathbb{E}(A^{(l)})$; clearly each X_l is centered. Moreover $\|X_l\| \leqslant \|X_l\|_F \leqslant p_{max} n(\rho^{-2} - 1)$.

For notation convenience, we will write X instead of X_l. We have $\mathbb{E}(X^2)_{ij} = \sum_{k \leqslant n} X_{ik}X_{jk}$ because X is symmetric. Recall that $X_{ik} = a_{ik}(\rho^{-2}\omega_i\omega_k - 1)$ where a_{ik} corresponds to $A_{ik}^{(l)}$. A simple calculation shows that

$$\mathbb{E}(X^2)_{ij} = \sum_k \mathbb{E}(a_{ik}a_{jk}(\rho^{-2}\omega_i\omega_k - 1)(\rho^{-2}\omega_j\omega_k - 1)))$$

$$= \sum_k a_{ik}a_{jk}\mathbb{E}((\rho^{-2}\omega_i\omega_k - 1)(\rho^{-2}\omega_j\omega_k - 1))).$$

If $i = j$, $\mathbb{E}((\rho^{-2}\omega_i\omega_k - 1)^2) = \rho^{-2} - 1$ and if $i \neq j$, $\mathbb{E}((\rho^{-2}\omega_i\omega_k - 1)(\rho^{-2}\omega_j\omega_k - 1))) = \rho^{-1} - 1$. So in both cases, $|\mathbb{E}(X^2)_{ij}| \leqslant np_{max}^2(\rho^{-2} - 1)$. We can now bound $\|\mathbb{E}(X_l^2)\|$ by $\|\mathbb{E}(X_l^2)\|_F \leqslant [np_{max}(\rho^{-2} - 1))]^2$ and $\sigma^2 := \|\sum_l \mathbb{E}(X_l^2)\|$ by $L[np_{max}(\rho^{-2} - 1)]^2$.

Therefore matrix Bernstein inequality implies that

$$\|\sum_l X_l\| \leqslant C(\rho^{-2} - 1)(np_{max}\sqrt{L\log n} + np_{max}\log n)$$

with probability at least $1 - O(n^{-1})$ for a constant C chosen appropriately. □

Proof of Theorem 9. Triangle inequality gives

$$\|\tilde{A} - \mathbb{E}(\tilde{A})\| \leqslant \|\tilde{A} - \mathbb{E}(\tilde{A}|\Omega)\| + \|\mathbb{E}(\tilde{A}|\Omega) - \mathbb{E}(\tilde{A})\|$$

and we can use Proposition 1 to bound with high probability each term. So with probability at least $1 - O(n^{-1})$

$$\begin{aligned}\|\tilde{A} - \mathbb{E}(\tilde{A})\| \leqslant & \frac{C}{\rho^2}\left(\sqrt{\frac{np_{max}}{L}} + \sqrt{\frac{\log n}{L}}\right) \\ & + C(\rho^{-2} - 1)\left(np_{max}\sqrt{\frac{\log n}{L}} + \frac{np_{max}\log n}{L}\right).\end{aligned}$$

We can now use the relation established in Lei et al., 2015, Lemma 2.1, and an immediate adaptation of Lemma 16 to conclude as in Theorem 8. □

4.4.2 *Iteratively imputing the missing entries*

When the number of missing nodes is important, filling missing entries with zero can lead to a huge bias and hence poor clustering performances. In order to reduce the bias we propose an alternative way of imputing the missing values (outlined as Algorithm 7) based on the fact that each adjacency matrix is a noisy realization of a structured matrix.

At iteration t, given an initial estimate $\hat{U}_K^t \in \mathbb{R}^{n\times K}$ of the common subspace we can estimate the membership matrix $\hat{Z}^t$ by applying k-means on $\hat{U}_K^t$. Then, we can estimate the connectivity matrix $\hat{\Pi}^{(l),t}$ for each l as

$$\hat{\Pi}^{(l),t} = ((\hat{Z}^t)^T\hat{Z}^t)^{-1}(\hat{Z}^t)^T A^{(l),t}\hat{Z}^t((\hat{Z}^t)^T\hat{Z}^t)^{-1}. \tag{4.4.1}$$

Given $\hat{Z}^t$ and $\hat{\Pi}^{(l),t}$ we estimate the rows and columns corresponding to missing nodes. Indeed, the connectivity profile of a node i in layer l is given by the ith row of $\hat{Z}^t\hat{\Pi}^{(l),t}(\hat{Z}^t)^T$. By replacing the rows and

columns of missing nodes by their estimated profiles, and leaving the value of observed nodes unchanged, we obtain the updated imputed matrix $A^{(l),t+1}$. Applying spectral clustering on $L^{-1}\sum_l A^{(l),t+1}$ then leads to an updated estimate $\hat{U}_K^{t+1}$ of the common subspace. The procedure can be repeated using $\hat{U}_K^{t+1}$ and $A^{(l),t+1}$, thus iteratively imputing the missing values in order to obtain "completed" adjacency matrices that share the same K rank structure across layers. In the worst case, the complexity of the algorithm run with T iterations is $O((K+L)n^2T + LKnT)$.

Similar iterative imputation methods have been studied in the context of principal component analysis, see for e.g., Zhang et al., 2022; Zhu et al., 2019. In our experiments, Algorithm 7 is seen to perform

Algorithm 7 Sum of adjacency matrices with missing entries filled iteratively

Input: Number of communities K; J_l and $A_{J_l} \in \mathbb{R}^{n\times n}$ for each l; initial estimate of the common subspace $\hat{U}_K^0 \in \mathbb{R}^{n\times K}$ (with orthonormal columns) obtained from Algorithm 6; number of iterations T.

1: Initialize $t = 0$ and $A^{(l),0} = A_{J_l}$ for all l.
2: **repeat**
3: Given $\hat{U}_K^t$, estimate the membership matrix $\hat{Z}^t$ and the connectivity parameters $\hat{\Pi}^{(l),t}$ for all l by using (4.4.1).
4: For each l, replace rows (and corresponding columns) of $A^{(l)}$ corresponding to a missing node i by the ith row of $\hat{Z}^t\hat{\Pi}^{(l),t}\hat{Z}^{t^T}$ to form $A^{(l),t+1}$.
5: Compute the eigenvector matrix $\hat{U}_K^{t+1} = [u_1^{t+1}\ u_2^{t+1}\ \cdots\ u_K^{t+1}]$ associated with the K largest (in absolute order) eigenvalues of $L^{-1}\sum_l A^{(l),t+1}$. Update $t \leftarrow t+1$.
6: **until** $t \geqslant T$
7: Apply K-means on $\hat{U}_K^T$ to get a partition of $\mathcal{N}$.

Output: A partition of the nodes $\mathcal{N} = \cup_{i=1}^K \mathcal{C}_i$.

significantly better than other methods when ρ decreases. While we do not currently have any statistical performance guarantee for Algorithm 7, establishing this is an interesting direction for future work.

4.5 INTERMEDIATE FUSION METHODS: OLMF ESTIMATOR

Orthogonal Linked Matrix Factorization (OLMF) is a clustering method for multilayer graphs that originated in the work of Tang et al., 2009 in the complete data setup, and was later analysed in Paul et al., 2020. It shows good performance in various settings and outperforms spectral clustering when the multilayer network contains homophilic and

heterophilic communities (see the numerical experiments in Paul et al., 2020).

4.5.1 *The complete data setting*

In the complete data setting, the OLMF estimator is a solution of the following optimization problem

$$(\hat{Q}, \hat{B}^{(1)}, \dots, \hat{B}^{(L)}) \in \underset{\substack{Q^T Q = I_k \\ B^{(1)}, \dots, B^{(L)}}}{\operatorname{argmin}} \sum_{l} \|A^{(l)} - Q B^{(l)} Q^T\|_F^2, \tag{4.5.1}$$

where $Q \in \mathbb{R}^{n \times K}$, $B^{(l)} \in \mathbb{R}^{K \times K}$. Note that there is no constraint on the values taken by the entries of $B^{(l)}$.

A little algebra (see Paul et al., 2020) shows that the optimization problem (4.5.1) is equivalent to

$$\hat{Q} \in \underset{Q^T Q = I_k}{\operatorname{argmax}} \sum_{l} \|Q^T A^{(l)} Q\|_F^2, \quad \hat{B}^{(l)} = \hat{Q}^T A^{(l)} \hat{Q} \tag{4.5.2}$$

for $l = 1, \dots, L$. The OLMF estimator can be computed with a gradient descent on the Stiefel manifold (see Paul et al., 2020 and supplementary material therein). The community estimation is then obtained by applying K-means on the rows of $\hat{Q}$.

4.5.2 *Extension to the missing nodes setting*

We now present an extension of the OLMF estimator to the setting of missing nodes. By replacing the matrices $A^{(l)}$, Q in the objective function in (4.5.1) with $A_{J_l} \in \mathbb{R}^{n \times n}$, $Q_{J_l} \in \mathbb{R}^{n \times K}$, we end up with the following modification for the incomplete setting

$$(\hat{Q}, \hat{B}^{(1)}, \dots, \hat{B}^{(L)}) \in \underset{\substack{Q^T Q = I_k \\ B^{(1)}, \dots, B^{(L)}}}{\operatorname{argmin}} \sum_{l} \|A_{J_l} - Q_{J_l} B^{(l)} Q_{J_l}^T\|_F^2. \tag{4.5.3}$$

In our experiments, we employ a BFGS algorithm for solving (4.5.3). The worst-case complexity of the algorithm is $O(LK(n^2 + Kn))$. Denoting the objective function in (4.5.2) by F, its gradients are given by

$$\begin{aligned} \frac{\partial F}{\partial Q} &= -2 \sum_{l} (A_{J_l} - Q_{J_l} B^{(l)} Q_{J_l}^T) Q_{J_l} B^{(l)}, \\ \frac{\partial F}{\partial B^{(l)}} &= -Q_{J_l}^T (A_{J_l} - Q_{J_l} B^{(l)} Q_{J_l}^T) Q_{J_l}. \end{aligned}$$

We relax the constraint that the gradient remains on the Stiefel manifold of $n \times k$ matrices, and initialize the parameters using Algorithm 6.

The optimization problem in (4.5.2) can be motivated via the missing nodes MLSBM as follows. If we replace the noisy realization A_{J_l} with $(Z\Pi^{(l)}Z^T) \odot \Omega^{(l)}$ then one can show (under some conditions) that the solution $\hat{Q}$ of (4.5.3) has the same column span as the ground truth $Z \in \mathcal{M}_{n,K}$. This is shown formally in the following proposition.

Proposition 2. *Assume that $\Pi^{(l)}$ is full rank for each l, and that for each l, l' the sets $J_l \cap J_{l'}$ intersect all communities. Then if $A_{J_l} = (Z\Pi^{(l)}Z^T) \odot \Omega^{(l)}$, it holds that the solution of (4.5.3) is given by $\hat{Q} = Z(Z^TZ)^{-1/2}$ and $\hat{B}^{(l)} = (Z^TZ)^{1/2}\Pi^{(l)}(Z^TZ)^{1/2}$ and is unique up to an orthogonal transformation. Moreover if i, j belong to the same community, then $\hat{Q}_{i*} = \hat{Q}_{j*}$.*

The matrix $\mathbb{E}(A^{(l)})$ can be considered as a slight perturbation of $Z\Pi^{(l)}Z^T$ since the former has zeros on the diagonal. Thus the proposition shows that when there is no noise, the column-span of $\hat{Q}$ (the solution of (4.5.3)) is the same as the ground truth partition Z.

4.6 NUMERICAL EXPERIMENTS

4.6.1 *Synthetic data*

We now describe simulation results when the multilayer graph is generated from the missing nodes MLSBM. The Normalized Mutual Information (NMI) criterion is used to compare the estimated community to the ground truth partition. It is an information theoretic measure of similarity taking values in $[0, 1]$, with 1 denoting a perfect match, and 0 denoting completely independent partitions. Nodes that are not observed at least once are removed. The diagonal (resp. off-diagonal) entries of the connectivity matrices are generated uniformly at random over $[0.18, 0.19]$ (resp. $0.7 * [0.18, 0.19]$). The ground truth partition is generated from a multinomial law with parameters $1/K$. While $K = 3$ is fixed throughout, the parameters n, ρ and L are varied suitably. The average NMI is reported over 20 Monte Carlo trials. As shorthand, we denote Alg. 5 by `k-pod`, Alg. 6 by `sumAdj0`, Alg. 7 by `sumAdjIter`, and (4.5.3) by `OLMFm`.

Figure 12 shows that `sumAdj0` gives good results unless ρ is too small. Then, the performance of this method decreases quickly. This suggests that there is a threshold involving ρ and the difference between intra and inter connectivity parameters. Figure 14 supports this claim. When ρ is small, the performance of `sumAdj0` doesn't improve when n increases. So even if the separation between communities improves, the intra and inter connectivity parameters remain the same suggesting a link between these parameters and ρ.

When L increases (see Figs. 12 and 13), the performance of all methods improves. However, performance of `k-pod` improves less quickly than other methods. This is expected since contrary to other methods,

`k-pod` relies more on the quality of each individual layer. `OLMFm` and `sumAdjIter` exhibit better performance than others in the challenging situation when ρ is small, and perform as well as the others when $\rho \approx 1$. They perform significantly better than `k-pod`, especially when L is large.

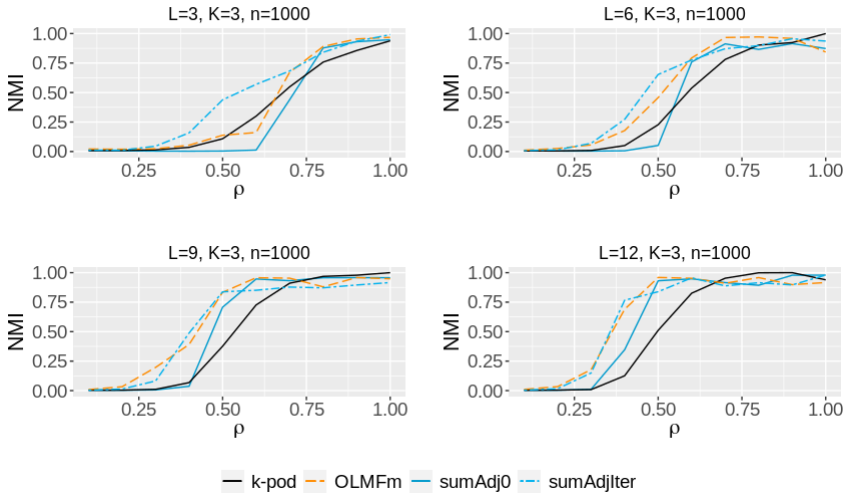

Figure 12: NMI vs ρ for different values of L.

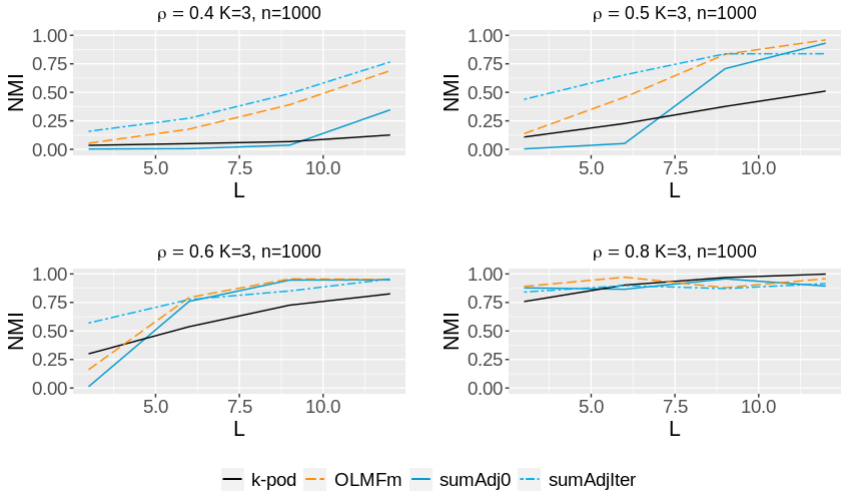

Figure 13: NMI vs L for different values of ρ.

4.6.2 *MIT Reality Mining dataset*

This dataset records interactions (measured by cell phones activities) between 96 students and staff at MIT in the 2004-05 school year (see Eagle et al., 2006). We used the dataset as provided by the R package 'GreedySTBM'. As in Han et al., 2015 we removed the first and last layers, then discretized the time into one week intervals. The number

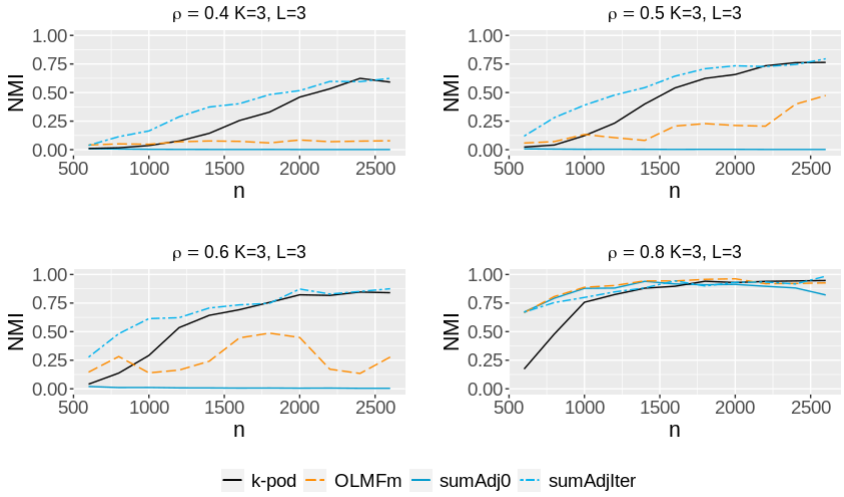

Figure 14: NMI vs n for different values of ρ.

of times two persons had an interaction during the week is not conserved in order to have a simple undirected graph corresponding to each layer. In total we obtained 32 layers. For different values of ρ, we randomly removed nodes in each layer of the multilayer network. The average NMI over 50 Monte Carlo trials is reported in Table 2 for our methods. The ground truth partition here is taken to be that obtained from `sumAdj0` when $\rho = 1$. We disregarded `k-pod` because even

ρ	`sumAdj0`	`OLMFm`	`sumAdjIter`
1	1.00	1.00	1.00
0.9	0.99	0.96	0.99
0.8	0.97	0.86	0.97
0.7	0.96	0.93	0.96
0.6	0.94	0.79	0.94
0.5	0.89	0.91	0.90
0.4	0.76	0.73	0.78
0.3	0.56	0.57	0.62
0.2	0.26	0.41	0.36
0.1	0.09	0.10	0.11

Table 2: NMI vs ρ for MIT Reality Mining dataset.

when $\rho = 1$, its performance was disappointing and very sensitive to the initialization. This is not very surprising since this method works only if each layer is informative enough while we have a multilayer network where individual layers can be very sparse.

The performance of the other three methods studied are quite similar when ρ is not too small ($\rho \geqslant 0.4$). However, the performance

of `OLMFm` seems to be quite sensitive to initialization since for $\rho \in \{0.6, 0.8\}$ its performance is worse than `sumAdj0` and `sumAdjIter`. Even if we remove half of the nodes in each layer we can still approximately recover the partition.

4.6.3 *Malaria parasite genes network*

The dataset was constituted by Larremore et al., 2013 to study the var genes parasite Plasmodium falciparum involved in Malaria. The nodes of the dataset correspond to 307 different amino acid sequences and each of the 9 layers corresponds to a highly variable region (HVR). Two nodes are linked in a given layer if there is a common block sequence between the corresponding amino acid sequences within the HVR associated to the layer. The analysis in Larremore et al., 2013 and Jing et al., 2021 shows that the first six layers share the same community structure with $K = 4$. Hence we restrict our study to the first six layers with $K = 4$. We use the same procedure as before to delete nodes and to select the ground truth partition. `k-pod` was disregarded for the same reason as the previous experiment. As ρ de-

ρ	`sumAdj0`	`OLMFm`	`sumAdjIter`
1	1.00	0.99	1.00
0.9	0.75	0.75	0.72
0.8	0.63	0.62	0.58
0.7	0.47	0.49	0.47
0.6	0.32	0.37	0.34
0.5	0.22	0.20	0.26
0.4	0.13	0.07	0.16

Table 3: NMI vs ρ for Malaria parasite genes network.

creases, the clustering performance decreases rapidly due to a weak separation between the clusters.

4.7 POSSIBLE EXTENSIONS AND CONCLUDING REMARKS

4.7.1 *Extension of `IR-LS` to the multilayer setting*

In this section, we will describe how to extend the algorithm designed for clustering the CSBM to the multilayer graph with missing nodes setting.

The principle of the method is similar to the one exposed in Section 3.3. Instead of having Gaussian covariates, we have graph side information given by the other layers. Consequently, we need to change the criterion used to refine the partition at each step. But this is quite

straightforward, it is sufficient to take the sum of the least square criterion used for individual SBM in Algorithm 1. Since a node i may not appear in each layer, we should disregard the layer were i is missing since these layers doesn't provide any information about the community i belong to.

Algorithm 8 Iterative Refinement for MLSBM (IR-MLSBM)

Input: Adjacency matrices $A^{(l)} \in \mathbb{R}^{n\times n}$ for $l \leqslant L$ and a initial estimate of the latent partition $Z^{(0)} \in \{0,1\}^{n\times K}$.

1: **for** $0 \leqslant t \leqslant T-1$ **do**
2: Given $Z^{(t)}$, estimate the model parameters: $n_k^{(t)} = |\mathcal{C}_k^{(t)}|$, $W^{(t)} = Z^{(t)}(D^{(t)})^{-1}$ where $D^{(t)} = \operatorname{diag}(n_k^{(t)})_{k\in[K]}$, $\Pi^{(l,t)} = {W^{(t)}}^\top A^{(l)} W^{(t)}$ for all $l \leqslant L$.
3: Refine the partition by solving for each $i \in [n]$

$$z_i^{(t+1)} = \arg\min_k \sum_{l\leqslant L} \|(A^{(l)}_{i:}W^{(t)} - \Pi^{(l,t)}_{k:})\sqrt{\Sigma_k^{(l,t)}}\|^2 \mathbb{1}_{i\in J_l}$$

where

$$\Sigma_k^{(l,t)} = \begin{cases} \operatorname{diag}\left(\frac{n^{(t)}_{k'}}{\Pi^{(l,t)}_{kk'}}\right)_{k'\in[K]} & \texttt{(IR-LS')} \\ \frac{\min_{k'} n^{(t)}_{k'}}{\max_{k',k''}\Pi^{(l,t)}_{k'k''}} I_K & \texttt{(sIR-LS')} \\ \frac{n}{K(p^{(l,t)}-q^{(l,t)})}\log\left(\frac{p^{(l,t)}(1-q^{(l,t)})}{q^{(l,t)}(1-p^{(l,t)})}\right) I_K & \texttt{(IR-LSS')} \end{cases}$$

with $p^{(l,t)} = K^{-1}\sum_{k\in[K]}\Pi^{(l,t)}_{kk}$ and $q^{(l,t)} = (K^2-K)^{-1}\sum_{k\neq k'\in[K]}\Pi^{(l,t)}_{kk'}$.
4: Form the matrices $Z^{(t+1)}$ from $z^{(t+1)}$.
5: **end for**

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

This algorithm can be thought as an intermediate fusion method. But contrary to OLMF that optimizes on the set of orthogonal matrices of rank K, Algorithm 8 optimize directly on the discrete set $\mathcal{M}_{n,K}$.

We have shown in Chapter 3 that under the SBM, the misclustering rate associated with the output of Algorithm 1 decreases exponentially in np_{max} (up to a constant factor). So it is natural to conjecture that under the MLSBM the signal contained in each layer will add up and we would obtain a convergence rate of order $e^{-\Theta(Lnp_{max})}$. When we delete some node with probability ρ, in average each node will appear in ρL layers and will have an average degree of order ρnp_{max} instead of np_{max}. This leads to the following conjecture.

Conjecture 1. *Under the MLSBM with missing nodes, if* $\rho^2 Ln p_{max} \to \infty$ *and* L *is large enough, the output* $Z^{(T)}$ *of Algorithm 8 initialized with a estimate* $Z^{(0)}$ *such that* $r(Z^{(0)}, Z) = O(1/K)$ *satisfies*

$$r(Z^{(T)}, Z) \leqslant e^{-\Theta(\rho^2 Ln p_{max})}.$$

To prove this conjecture one could use the same proof techniques as used in Chapter 3. We believe that the adaptation is quite straightforward, but requires some effort to be presented with all the details. We leave that for further work.

However, the preliminary numerical experiments results presented in Figure 15 are not very encouraging. We considered a similar setting as in Section 4.6.1: $L = 3, K = 3, \rho = 0.9$ and the connectivity matrices are the diagonal (resp. off-diagonal) entries of the connectivity matrices are generated uniformly at random over $[0.18, 0.19]$ (resp. $0.7 * [0.18, 0.19]$). We initialized Algorithm 8 with `sumAdj0`. The performances are averaged over 20 Monte-Carlo simulations. Contrary to our expectation, the performance of the clustering slightly decreases after using `IR-MLSBM`.

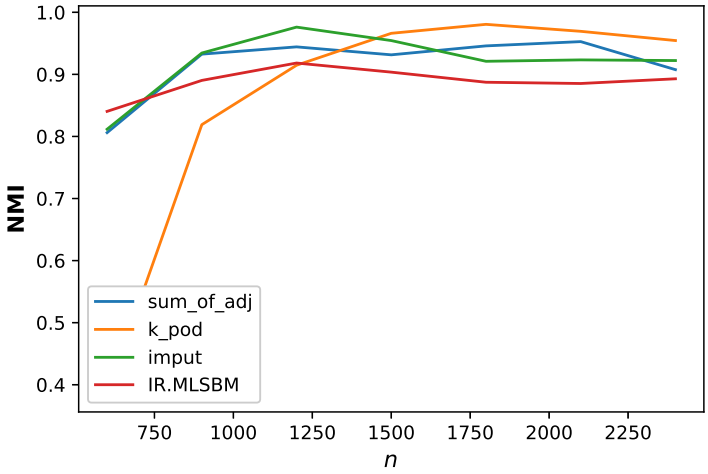

Figure 15: NMI for varying n with $\rho = 0.9$.

4.7.2 *Other research direction*

We assumed for simplicity that the nodes are missing under a Bernoulli sampling scheme, but other missing patterns could be considered. Another important direction would be to relax the strong condition imposed by MLSBM that all layers share the same common partition. For example, it would be more realistic to assume that the partition of networks evolving over time also evolves slowly. It would also be interesting to study the case where missing values are missing not at random (MNAR). In this case, the positions where there is a missing

value could contain information about the underlying communities and hence be considered as side information.

4.8 ADDITIONAL PROOFS

4.8.1 *Auxiliary Lemmas*

We first recall the standard Chernoff bound for sum of independent Bernoulli random variables.

Lemma 14. *Under the assumptions of Theorem 8, with probability at least* $1 - O(KL/n^2)$*, it holds for each* $k = 1, \dots, K$ *and* $l = 1, \dots, \leqslant L$ *that*

$$\frac{\rho}{2} n_k \leqslant n_{k,J_l} \leqslant 2\rho n_k.$$

Proof. Recall that $n_{k,J_l} = \sum_{i \in C_k} \mathbb{1}_{i \in J_l}$ is a sum of n_k independent Bernoulli random variables with parameter ρ. By applying Lemma 32 with $\delta = 1/2$ we get

$$n_{k,J_l} \geqslant \frac{\rho}{2} n_k$$

and

$$n_{k,J_l} \leqslant 2 n_k$$

with probability at least $1 - O(1/n^2)$, provided that $n_k \rho \geqslant C \log n$ for a constant C large enough as assumed in Theorem 8. The lemma follows from a union bound. □

Lemma 15. *Let* X *be an* $n \times n$ *symmetric matrix whose entries* X_{ij} *are independent centered random variables. Then there exists for any* $0 < \epsilon \leqslant 1/2$ *a universal constant* c_ϵ *such that for every* $t \geqslant 0$

$$\mathbb{P}(\|X\| \geqslant 2(1+\epsilon)\tilde{\sigma} + t) \leqslant \exp\left(-\frac{t^2}{\tilde{c}_\epsilon \tilde{\sigma}_*}\right)$$

where $\tilde{\sigma} = \max_i \sqrt{\sum_j \mathbb{E}(X_{ij}^2)}$ *and* $\tilde{\sigma}_* = \max_{i,j} \mathbb{E}\|X_{ij}\|_\infty$.

Proof. See Bandeira et al., 2016, Corollary 3.12 and Remark 3.13. □

Lemma 16. *Let* $A \in \mathbb{R}^{n \times n}$ *be an adjacency matrix generated by a* $\mathrm{SBM}(Z, \Pi)$*. Denote* λ_K *to be the* K*th largest singular value of* $P = Z \Pi Z^T$*. If* $\lambda_K > 2\|A - \mathbb{E}(A)\|$*, then with probability at least* $1 - O(n^{-2})$

$$\|\hat{U} - UO\|_F \leqslant C\sqrt{K} \frac{\sqrt{\max(np_{max}, \log n)}}{\lambda_K}$$

where $\hat{U}$ *is the matrix formed by the first* K *left singular vectors of* A*,* $U = Z(Z^T Z)^{1/2}$ *and* O *is the orthogonal matrix that aligns* $\hat{U}$ *and* U.

Proof. By Remark 3.13 in Bandeira et al., 2016 we get that

$$\|A - \mathbb{E}(A)\| \leqslant C\sqrt{\max(np_{max}, \log n)}$$

with probability at least $1 - O(n^{-2})$. Moreover, since $\hat{U}$ and UO are at most rank K matrices we have

$$\|\hat{U} - UO\|_F \leqslant \sqrt{2K}\|\hat{U} - UO\|.$$

Wedin's theorem (see Wedin, 1972) implies that

$$\|\hat{U} - UO\| \leqslant \frac{\|A - \mathbb{E}(A)\|}{\delta} \tag{4.8.1}$$

where $\delta := |\lambda_K(A) - \lambda_{K+1}(\mathbb{E}(A))|$ represents the spectral gap. By Weyl's inequality,

$$|\lambda_{K+1}(\mathbb{E}(A)) - \lambda_{K+1}(Z\Pi Z^T)| \leqslant \|\mathbb{E}(A) - Z\Pi Z^T\|.$$

Since $\mathbb{E}(A) - Z\Pi Z^T$ is a diagonal matrix, its spectral norm is bounded by its largest coefficient that is bounded by p_{max}. Moreover since $\lambda_{K+1}(Z\Pi Z^T) = 0$ we get $\lambda_{K+1}(\mathbb{E}(A)) \leqslant p_{max}$. The same argument can be used to show that $\lambda_K(\mathbb{E}(A)) \geqslant \lambda_K(Z\Pi Z^T) - p_{max}$.

Weyl's inequality also implies that

$$|\lambda_K(\mathbb{E}(A)) - \lambda_K(A)| \leqslant \|A - \mathbb{E}(A)\|.$$

Thus

$$\begin{aligned}\lambda_K(A) &\geqslant \lambda_K(\mathbb{E}(A)) - \|A - \mathbb{E}(A)\| \\ &\geqslant \lambda_K(Z\Pi Z^T) - p_{max} - \|A - \mathbb{E}(A)\| \\ &\geqslant \frac{1}{2}\lambda_K(Z\Pi Z^T) - p_{max}.\end{aligned}$$

The last inequality follows from the assumption that $\lambda_K(Z\Pi Z^T) \geqslant 2\|A - \mathbb{E}(A)\|$. Since $p_{max} \leqslant \epsilon(n)\lambda_K(Z\Pi Z^T)$ where $\epsilon(n) \to 0$ when $n \to \infty$, $\lambda_K(A) \geqslant c\lambda_K(Z\Pi Z^T)$ and then $\delta \geqslant c\lambda_K(Z\Pi Z^T)$. Therefore the concentration bound stated at the beginning of the proof and (4.8.1) implies

$$\|\hat{U} - UO\|_F \leqslant \sqrt{2K}\frac{\|A - \mathbb{E}(A)\|}{\delta} \leqslant C\sqrt{K}\frac{\sqrt{\max(np_{max}, \log n)}}{\lambda_K}$$

with probability at least $1 - O(n^{-2})$. □

Lemma 17. *We have* $\lambda_K(Z\Pi Z^T) \geqslant n_{min}\lambda_K(\Pi)$.

Proof. Let $\mu_1 \geqslant \mu_2 \geqslant \cdots \geqslant \mu_K$ be the K *non-zero* eigenvalues of $Z\Pi Z^T$. By the variational characterization of eigenvalues we have for all $k = 1, \ldots, K$

$$\mu_k(Z\Pi Z^T) = \min_{V \subset G_{n-k+1}} \max_{\substack{x \in V \\ \|x\|=1}} x^T Z\Pi Z^T x \tag{4.8.2}$$

where G_{n-k+1} denotes the set of $n-k+1$ dimensional subset of $\mathbb{R}^n$. Observe that $\ker(Z^T) = \mathrm{Im}(Z)^\perp$. An element $x \in \ker(Z^T)$ cannot be a solution because $Z\Pi Z^T$ is a rank K matrix and thus $\mu_k(Z\Pi Z^T) \neq 0$, so for $k \leqslant K$ the optimization problem (4.8.2) is equivalent to

$$\mu_k(Z\Pi Z^T) = \min_{V \subset G_{n-k+1}} \max_{\substack{x \in V \cap \mathrm{Im}(Z) \\ ||x||=1}} x^T Z\Pi Z^T x. \tag{4.8.3}$$

It implies in particular that any eigenvector of $Z\Pi Z^T$ associated with μ_k belongs to $\mathrm{Im}(Z)$, so it has a block structure. Let v be an eigenvector associated with $\mu_k(Z\Pi Z^T)$ for $1 \leqslant k \leqslant K$. Then $v = Zu$ where $u \in \mathbb{R}^K$. In particular $Z\Pi Z^T v = Zw$ where $w = \Pi Z^T Zu$. Thus

$$\begin{aligned}
\mu_k^2(Z\Pi Z^T) = \|Z\Pi Z^T v\|^2 &\geqslant n_{min}\|w\|^2 \\
&\geqslant n_{min}\lambda_K^2(\Pi)\|Z^T Zu\|^2 \\
&\geqslant n_{min}^2\lambda_K^2(\Pi)\|v\|^2
\end{aligned}$$

because the least singular value of Z is $\sqrt{n_{min}}$. Clearly, this in particular implies that $\lambda_K(Z\Pi Z^T) \geqslant n_{min}\lambda_K(\Pi)$. □

4.8.2 *Comparison between misclustering bound under MLSBM in the complete setting*

Here we compare existing bounds for the misclustering rate under the MLSBM in the complete data setting. In order to simplify the comparison between the existing bounds, we will assume that K is a constant, the communities are well balanced and $p_{max}^{(l)} \approx p_{max}$ for each l.

- **Co-regularized spectral clustering**. This algorithm was introduced by Kumar et al., 2011. It is an intermediate fusion method that aims to find the best set of eigenvectors that simultaneously approximate the set of eigenvectors associated with each individual layer. It was shown later by Paul et al., 2020 that if $Lnp_{max} \geqslant C \log n$ and $\Pi^{(l)}$ is full rank for all l, then with high probability (w.h.p)

$$r_{coreg} = O\left(\sqrt{\frac{\log n}{Lnp_{max}}}\right).$$

- **OLMF**. This estimator was discussed earlier in Section 4.5.1. It was shown by Paul et al., 2020 that if $np_{max} \geqslant C \log n$ and at least one of the matrices $\Pi^{(l)}$ is full rank then w.h.p.

$$r_{OLMF} = O\left(\frac{1}{\sqrt{np_{max}}} \max\left\{1, \frac{(\log n)^{2+\epsilon}\sqrt{\log L}}{L^{1/4}}\right\}\right).$$

- **Sum of adjacency matrices.** It was shown by Paul et al., 2020 that if $Lnp_{max} \geqslant \log n$ and $\lambda_K(\sum_l \Pi^{(l)}) \approx Lp_{max}$ then w.h.p.

$$r_{sum} = O\left(\frac{\log n}{Lnp_{max}}\right).$$

Bhattacharyya et al., 2018 showed that if $Lnp_{max} \geqslant \log n$ and $\lambda_K(\sum_l \Pi^{(l)}) \approx L\lambda_K(\Pi^{(1)})$ then w.h.p.

$$r_{sum} = O\left(\frac{1}{\sqrt{Lnp_{max}}}\right).$$

The condition $Lnp_{max} \geqslant \log n$ is not stated in Bhattacharyya et al., 2018 and is only assumed here for simplification. This last bound is better than the former in the sparse case when $Lnp_{max} \approx \log n$. But when $np_{max} \gg \log n^2$ the first bound is sharper.

- **Bias adjusted sum of the squared adjacency matrices.** Sum of adjacency matrices performs badly when some layers are associative and other disassociative. Taking the sum of the square of adjacency matrices instead permits us to overcome this issue. However the diagonal entries of these squared matrices introduce bias, so they are often removed. More involved debiasing strategies have also been considered by Zhang et al., 2022 and Giraud et al., 2019. Assume $L = O(n)$. In the sparse case when $\sqrt{L}np_{max} \geqslant C\sqrt{\log n}$ and $np_{max} = O(1)$, Lei et al., 2022 showed that w.h.p.

$$r_{sq} = O\left(\frac{1}{n} + \frac{\log n}{L(np_{max})^2}\right).$$

If $np_{max} \geqslant C\sqrt{\log n}$ they showed that w.h.p.

$$r_{sq} = O\left(\frac{\log n}{\sqrt{L}np_{max}}\right).$$

This method was also analyzed by Bhattacharyya et al., 2020. They showed that if $Lnp_{max} \geqslant C \log n$ then w.h.p.

$$r_{sq} = O\left(\frac{1}{(Lnp_{max})^{1/2}}\right).$$

5

MINIMAX OPTIMAL CLUSTERING OF HIGH-DIMENSIONAL BIPARTITE GRAPHS

This chapter is based on Braun et al., 2022. We show that a similar iterative refinement algorithm as the one used for clustering CSBM can be used to cluster bipartite graphs in the high-dimensional regime where the number of columns n_2 is larger than the number of rows n_1. Our work extends the work of Ndaoud et al., 2022 in several ways. First our algorithm can be applied to a more general bipartite graph model contrary to the one proposed by Ndaoud et al., 2022 that is limited to a particular case where the columns (resp. rows) can be partitioned into L (resp. K) communities with $K = L = 2$. Secondly we derive a minimax lower bound on the misclustering on the misclustering error when $K = L = 2$, which matches the corresponding upper bound up to a constant factor.

In Section 5.1 we introduce the problem and discuss the related work. The statistical framework is presented in Section 5.2 and the algorithms in Section 5.3. In Section 5.4 we analyze the spectral method used for initialization. In Section 5.5 we analyze the GPM algorithm used to refine the estimate of the partition. In Section 5.6 we derive a minimax lower bound on the misclustering error. Experimental results are presented in Section 5.7 and the additional proofs details are gathered in Section 5.8.

5.1 INTRODUCTION

The interactions between objects of two different types can be naturally encoded as a bipartite graph where nodes correspond to objects and edges to the links between the objects of different type. One can find examples of such data in various fields, e.g., interactions between customers and products in e-commerce (Huang et al., 2007), interactions between plants and pollinators (Young et al., 2021), investors and assets networks (Squartini et al., 2017), judges vote predictions (Guimerà et al., 2011) and constraint satisfaction problems (Feldman et al., 2015).

Clustering is one of the most important analysis tasks on bipartite graphs aimed at gathering nodes that have similar connectivity profiles. To this end, several methods have been proposed in the literature, e.g., convex optimization approaches (Lim et al., 2015), spectral methods (Zhou et al., 2019), modularity function maximization (Beckett, 2016), pseudo-likelihood (Zhou et al., 2020) and variational approaches (Keribin et al., 2015). The performance of the algo-

rithms are generally evaluated under the Bipartite Stochastic Block Model (BiSBM), a variant of the SBM, where the partitions of the rows and the columns are decoupled. In particular, edges are independent Bernoulli random variables with parameters depending only on the communities of the nodes.

When the number of rows n_1 (corresponding to type I objects) and the number of columns n_2 (corresponding to type II objects) of the adjacency matrix associated to a bipartite graph are of the same order, the BiSBM behaves similarly to the SBM. However, in the high dimensional setting where $n_2 \gg n_1$, classical methods that are applicable when n_1 is of the same order as n_2 can fail. In particular, when the bipartite graph is very sparse, it becomes impossible to consistently estimate the latent partition of the columns, whereas it is still possible to estimate the latent partition of the rows. Hence, methods based on estimating the latent partitions of both rows and columns will necessarily fail in the high dimensional regime when the bipartite graph is very sparse. However, the high dimensional regime appears in many applications, e.g. hypergraphs where the number of columns corresponds to the number of hyperedges, or in e-commerce where the number of customers could be much smaller than the number of products (or vice-versa). Hence it is important to understand how to design statistically optimal algorithms in this regime.

Recently, the work of Ndaoud et al., 2022 improved the state-of-the-art conditions for exact recovery of the latent partition of the rows under the BiSBM when $n_2 \gtrsim n_1 \log n_1$. Unfortunately, their method, which can be understood as a generalized power method, uses a centering argument that only works in the special case where there are $K = 2$ latent communities for the rows and $L = 2$ latent communities for the columns. Moreover, they only consider a Symmetric BiSBM (SBiSBM) where the edge probabilities can take two values (see Section 5.2). It is not clear how their method can be extended to the more general setting where $K \neq L \geqslant 2$, and with more general connectivity matrices. To overcome this limitation, we propose a new algorithm – also based on the generalized power method – that can be applied to general BiSBMs, and has similar theoretical guarantees when specialized to the setting of Ndaoud et al., 2022.

5.1.1 *Main contributions*

Our results can be summarized as follows.

- We present a novel iterative clustering method that can be applied to general BiSBMs, unlike the one proposed by Ndaoud et al., 2022. We analyze our algorithm under the BiSBM without restrictions on K and L, and derive an upper bound on the misclustering error. In particular, we show that our algorithm achieves exact recovery when the sparsity level p_{max} of the

graph satisfies $p_{max}^2 = \Omega(\frac{\log n_1}{n_1 n_2})$ for fixed K, L. This is the same sparsity regime obtained in Ndaoud et al., 2022 for the SBiSBM with $K = L = 2$. We remark that our bounds are non-asymptotic and showcase the full dependence on K, L.

- We derive a minimax lower bound for the misclustering error in the special case of an SBiSBM with $L = K = 2$ that matches the corresponding upper bound of our algorithm, and is the first such lower bound for this problem. It completes the work of Ndaoud et al., 2022 which only shows that an oracle version of their algorithm fails to achieve exact recovery when $p_{max}^2 \leqslant \epsilon \frac{\log n_1}{n_1 n_2}$ where ϵ is a small enough constant.
- As part of our analysis, we derive a concentration inequality for matrices with independent centered Binomial entries (see part 3 of Lemma 36) that could be of independent interest.

Our findings are complemented by numerical experiments on synthetic data.

5.1.2 *Related work*

A clustering strategy based on the MAP estimate for discrete weighted bipartite graphs that can encompass the BiSBM as a special case has recently been proposed by Jo et al., 2021. However, their method requires estimating the latent partition of the columns and hence requires the sparsity level of the graph p_{max} to satisfy $p_{max} \gtrsim \frac{\log n_2}{n_1}$. In contrast, we need a far weaker condition $p_{max} \gg \frac{1}{\sqrt{n_1 n_2}}$ in the high dimensional regime. This highlights one of the difficulties we face in the high dimensional regime – while it is impossible to correctly estimate all the model parameters, it is still possible to exactly recover the row partition. A similar phenomenon is known for Gaussian mixture models, see Ndaoud, 2018.

Our algorithm design is based on the GPM which has been applied successfully in various statistical learning problems in recent years. This includes, e.g., group synchronization (Boumal, 2016), joint alignment from pairwise difference (Chen et al., 2018), graph matching (Onaran et al., 2017), low rank matrix recovery (Chi et al., 2019) and SBM (Wang et al., 2021).

The work of Ndaoud et al., 2022 which we extend in the present paper is also based on the GPM. However, there are significant differences between the algorithms. We do not need the centering step used in (Ndaoud et al., 2022) and since we encode the community memberships in a $n_1 \times K$ matrix (instead of using the sign), our algorithm can be applied when $K > 2$. Our algorithm is closer to the one proposed by Wang et al., 2021 for clustering graphs under the SBM. In contrast to Wang et al., 2021, we do not add any constraints

on the columns, and instead of solving a linear assignment problem, we directly project on the extreme points of the unit simplex.

The GPM is also related to alternating optimization, a common strategy used to solve non-convex optimization problems in an iterative way. For example, EM-type algorithms (Dempster et al., 1977) have been used since decades. In general, these methods are not guaranteed to achieve a global optimum. However, a recent line of research has show that under various statistical models, alternating optimization can actually lead to consistent estimators (Lu et al., 2016; Chen et al., 2019; Gao et al., 2019; Han et al., 2020; Chen et al., 2021; Braun et al., 2021a). Our proof techniques are based on the work of Braun et al., 2021a which itself is based on a general framework developed by Gao et al., 2019.

5.2 THE STATISTICAL FRAMEWORK

The BiSBM is defined by the following parameters.

- A set of nodes of type I, $\mathcal{N}_1 = [n_1]$, and a set of nodes of type II, $\mathcal{N}_2 = [n_2]$.
- A partition of $\mathcal{N}_1$ into K communities $\mathcal{C}_1, \dots, \mathcal{C}_K$ and a partition of $\mathcal{N}_2$ into L communities $\mathcal{C}'_1, \dots, \mathcal{C}'_L$.
- Membership matrices $Z_1 \in \mathcal{M}_{n_1,K}$ and $Z_2 \in \mathcal{M}_{n_2,L}$ where $\mathcal{M}_{n,K}$ denotes the class of membership matrices with n nodes and K communities. Each membership matrix $Z_1 \in \mathcal{M}_{n_1,K}$ (resp. $Z_2 \in \mathcal{M}_{n_2,L}$) can be associated bijectively with a partition function $z : [n] \to [K]$ (resp. $z' : [n] \to [L]$) such that $z(i) = z_i = k$ where k is the unique column index satisfying $(Z_1)_{ik} = 1$ (resp. $(Z_2)_{ik} = 1$). To each matrix $Z_1 \in \mathcal{M}_{n_1,K}$ we can associate a matrix W by normalizing the columns of Z_1 in the ℓ_1 norm: $W = Z_1 D^{-1}$ where $D = Z_1^\top \mathbf{1}_{n_1}$. This implies that $W^\top Z_1 = I_K = Z_1^\top W$.
- A connectivity matrix of probabilities between communities

$$\Pi = (\pi_{kk'})_{k \in [K], k' \in [L]} \in [0,1]^{K \times L}.$$

Let us write $P = (p_{ij})_{i,j \in [n]} := Z_1 \Pi (Z_2)^\top \in [0,1]^{n_1 \times n_2}$. A graph $\mathcal{G}$ is distributed according to BiSBM(Z_1, Z_2, Π) if the entries of the corresponding bipartite adjacency matrix A are generated by

$$A_{ij} \overset{\text{ind.}}{\sim} \mathcal{B}(p_{ij}), \quad i \in [n_1],\ j \in [n_2],$$

where $\mathcal{B}(p)$ denotes a Bernoulli distribution with parameter p. Hence the probability that two nodes are connected depends only on their community memberships. We will frequently use the notation E for the centered noise matrix defined as $E_{ij} = A_{ij} - p_{ij}$, and denote the

maximum entry of P by $p_{max} = \max_{i,j} p_{ij}$. The latter can be interpreted as the sparsity level of the graph. We make the following assumptions on the model.

Assumption A3 (Approximately balanced communities)**.** *The communities $\mathcal{C}_1, \dots, \mathcal{C}_K$, (resp. $\mathcal{C}'_1, \dots, \mathcal{C}'_L$) are approximately balanced, i.e., there exists a constant $\alpha \geqslant 1$ such that for all $k \in [K]$ and $l \in [L]$ we have*

$$\frac{n_1}{\alpha K} \leqslant |\mathcal{C}_k| \leqslant \frac{\alpha n_1}{K} \text{ and } \frac{n_2}{\alpha L} \leqslant |\mathcal{C}'_l| \leqslant \frac{\alpha n_2}{L}.$$

Assumption A4 (Full rank connectivity matrix)**.** *The smallest eigenvalue of $\Pi\Pi^\top$, denoted by $\lambda_K(\Pi\Pi^\top)$, satisfies*

$$\lambda_K(\Pi\Pi^\top) \gtrsim p_{max}^2 \text{ and } \left\|\Pi\Pi^\top\right\| \lesssim p_{max}^2.$$

Assumption A5 (Diagonal dominance)**.** *There exist $\beta > 0$ and $\eta \geqslant 1$ such that for all $k' \neq k \in [K]$, we have*

$$(\Pi\Pi^\top)_{kk} - \alpha^2 (\Pi\Pi^\top)_{kk'} \geqslant \beta p_{max}^2$$

and

$$(\Pi\Pi^\top)_{kk} - \frac{1}{\alpha^2}(\Pi\Pi^\top)_{kk'} \leqslant \eta\beta p_{max}^2.$$

Remark 10. *When all communities $\mathcal{C}'_l$ have size equal to n_2/L, the first condition in Assumption A5 simplifies to $(\Pi\Pi^\top)_{kk} - (\Pi\Pi^\top)_{kk'} \geqslant \beta p_{max}^2$ and corresponds to a diagonal dominance assumption. The second condition in Assumption A5 is useful to show that β is (up to a parameter η that could depend on L as discussed in the remark below) the parameter that measures the minimum difference between the diagonal and off-diagonal entries of $\Pi\Pi^\top$.*

SYMMETRIC BISBM. A particular case of interest is where $L = K$ and $\Pi = (p-q)I_K + q\mathbf{1}_K\mathbf{1}_K^\top$ where $1 \geqslant p > q \geqslant 0$ and $q = cp$ for some constant $0 < c < 1$. This model will be referred to as the SBiSBM, denoted by SBiSBM(Z_1, Z_2, p, q). For $K = 2$ this corresponds to the model analyzed in the work of Ndaoud et al., 2022. Since we now have (for $k' \neq k$)

$$(\Pi\Pi^\top)_{kk} = p^2 + (K-1)q^2 \text{ and } (\Pi\Pi^\top)_{kk'} = 2qp + (K-2)q^2$$

the following observations are useful to note.

- We have $(\Pi\Pi^\top)_{kk} - (\Pi\Pi^\top)_{kk'} = (p-q)^2 \leqslant p_{max}^2$, and also $(p-q)^2 \geqslant (1-c)^2 p_{max}^2$. Hence Assumption A5 is satisfied for $\eta = \frac{1}{(1-c)^2}$ and $\beta = (1-c)^2$ in the equal-size community case ($\alpha = 1$). Additionally, $\lambda_K(\Pi\Pi^\top) = (1-c)p_{max}$.
- In the unequal-size case, we need to show choices of β, η under which Assumption A5 holds. To this end, note that

$$(\Pi\Pi^\top)_{kk} - \alpha^2(\Pi\Pi^\top)_{kk'} = p_{max}^2((1-c)^2 - (\alpha^2 - 1)K),$$

and

$$(\Pi\Pi^\top)_{kk} - \frac{1}{\alpha^2}(\Pi\Pi^\top)_{kk'} = p_{max}^2((1-c)^2 - (\frac{1}{\alpha^2} - 1)K) \leqslant p_{max}^2((1-c)^2 + (\alpha^2 - 1)K).$$

Hence if, for instance, $\alpha^2 = 1 + \frac{(1-c)^2}{2K}$, we see that Assumption A5 is satisfied for $\beta = \frac{(1-c)^2}{2}$ and $\eta = 3$.

Remark 11. *One of the parameters β or η can scale with L as shown in the following examples. Assume L is even and $\alpha = 1$. Let us consider $\mathbf{p} = (p_1, \dots, p_{L/2})$, $\mathbf{q} = (q_1, \dots, q_{L/2})$ where $p_l > q_l$ for all $l \leqslant L/2$ and*

$$\Pi = \begin{pmatrix} \mathbf{p} & \mathbf{q} \\ \mathbf{q} & \mathbf{p} \end{pmatrix} \in [0,1]^{2\times L}.$$

Then

$$\begin{aligned} L \max_l (p_l - q_l)^2 &\geqslant \min_{k\neq k'} (\Pi\Pi^\top)_{kk} - (\Pi\Pi^\top)_{kk'} \\ &= \|p - q\|^2 \\ &\geqslant L \min_l (p_l - q_l)^2, \end{aligned}$$

hence $\beta = \Theta(L)$ and $\eta = \Theta(1)$. Let us define $\mathbf{q}' = (q_1', q_2', q_3, \dots, q_{L/2})$ for $q_1' = 0.5q_1$ and $q_2' = 0.5q_2$ If we now consider

$$\Pi = \begin{pmatrix} \mathbf{p} & \mathbf{q} \\ \mathbf{q} & \mathbf{p} \\ \mathbf{q}' & \mathbf{p} \end{pmatrix} \in [0,1]^{3\times L}.$$

where q_i and p_i are as before and $q_1' \neq q_1$, $q_2' \neq q_2$. Then it is easy to check that $\max_{k\neq k'}(\Pi\Pi^\top)_{kk} - (\Pi\Pi^\top)_{kk'} \leqslant L p_{max}^2$ but

$$\min_{k\neq k'}(\Pi\Pi^\top)_{kk} - (\Pi\Pi^\top)_{kk'} = (\Pi\Pi^\top)_{22} - (\Pi\Pi^\top)_{23} = \frac{q_1^2 + q_2^2}{2}$$

is independent of L. Hence $\beta = \Theta(1)$ and $\eta = \Theta(L)$.

Remark 12. *As already mentioned in the introduction, we will focus in this work on the regime where $n_2 \gg n_1 \log n_1$ and $\sqrt{n_1 n_2} p_{max} \gtrsim \sqrt{\log n_1}$. In this parameter regime there is no hope to accurately recover Z_2 because the columns of A are too sparse. Indeed, consider the setting where $\sqrt{n_1 n_2} p_{max} \asymp \sqrt{\log n_1}$. Then in expectation, the sum of the entries of each column is $n_1 p_{max} \asymp \sqrt{\frac{n_1 \log n_1}{n_2}} \to 0$. But by analogy to the SBM, we would need a condition $n_1 p_{max} \to \infty$ in order to recover Z_2. This is actually a necessary condition obtained in a related setting by Jo et al., 2021. While it is impossible to estimate Z_2 in this sparsity regime, it is still possible to accurately estimate Z_1. We will focus on this problem from now onwards.*

5.3 ALGORITHM

5.3.1 *Initialization with a spectral method*

We can use a spectral method on the hollowed Gram matrix B to obtain a first estimate of the partition Z_1. This is similar to the algorithm in Florescu et al., 2016; we only use a different rounding step so the algorithm can be applied to general bipartite graphs with $K > 2$ communities, in contrast to Florescu et al., 2016.

Algorithm 9 Spectral method on $\mathcal{H}(AA^\top)$ (`Spec`)

Input: The number of communities K and the adjacency matrix A.

1: Form the diagonal hollowed Gram matrix $B := \mathcal{H}(AA^\top)$ where $\mathcal{H}(X) = X - \operatorname{diag}(X)$.
2: Compute the matrix $U \in \mathbb{R}^{n_1 \times K}$ whose columns correspond to the top K-eigenvectors of B.
3: Apply approximate $(1+\bar{\varepsilon})$ approximate `k-means` on the rows of U and obtain a partition $z^{(0)}$ of $[n_1]$ into K communities.

Output: A partition of the nodes $z^{(0)}$.

COMPUTATIONAL COMPLEXITY OF `spec`. The cost for computing B is $O(n_1 \operatorname{nnz}(A))$ and for U is[1] $O(n_1^2 K \log n_1)$. Applying the $(1+\bar{\varepsilon})$ approximate `k-means` has a complexity $O(2^{(K/\bar{\varepsilon})^{O(1)}} n_1 K)$, see Kumar et al., 2004. In practice, this operation is fast and the most costly operation is the computation of B which has complexity $O(n_1^2 n_2 p_{max})$, since one can show that $\operatorname{nnz}(A) = O(n_1 n_2 p_{max})$ with high probability.

5.3.2 *Iterative refinement with GPM*

Our algorithm is based on the Generalized Power Method. In contrast to the power method proposed recently by Ndaoud et al., 2022, we do not require to center the adjacency matrix A and, instead of using the sign to identify the communities, we project on $\mathcal{M}_{n_1,K}$. Consequently, our algorithm can be applied to bipartite graphs with $K > 2$ and $K \neq L$ while Ndaoud et al., 2022 require $K = L = 2$.

In the first step we form the diagonal hollowed Gram matrix B. This is natural since in the parameter regimes we are interested in, there is no hope to consistently estimate Z_2. Then, we iteratively update the current estimate of the partition $Z_1^{(t)}$ by a row-wise projection of

1 The $\log n_1$ term comes from the number of iterations needed when using the power method to compute the largest (or smallest) eigenvector of a given matrix.

$BZ_1^{(t)}$ onto the extreme points of the unit simplex of $\mathbb{R}^K$ denoted by $\mathcal{S}_K$. The projection operator $\mathcal{P} : \mathbb{R}^K \to \mathcal{S}_K$ is formally defined by

$$\mathcal{P}(x) := \arg\min_{y \in \mathcal{S}_K} \|x - y\| \text{ for all } x \in \mathbb{R}^K. \tag{5.3.1}$$

This implies that $(Z_1^{(t+1)})_{ik} = 1$ for the value of k that maximizes $B_{i:}Z_{:k}^{(t)}$. Ties are broken arbitrarily.

Algorithm 10 Generalized Power Method (GPM)

Input: Number of communities K, adjacency matrix A, an estimate of the row partition $\hat{Z}_1^{(0)}$ and, number of iterations T.

1: Form the diagonal hollowed Gram matrix $B := \mathcal{H}(AA^\top)$ where $\mathcal{H}(X) = X - \mathrm{diag}(X)$ and compute $W^{(0)} = (\hat{Z}_1^{(0)})^\top (D^{(0)})^{-1}$ where $D^{(0)} = \mathrm{diag}((\hat{Z}_1^{(0)})^\top \mathbf{1}_{n_1})$.
2: **for** $0 \leqslant t \leqslant T$ **do**
3: Update the partition $Z_1^{(t+1)} = \mathcal{P}(BW^{(t)})$ where $\mathcal{P}$ is the operator in (5.3.1) applied row-wise.
4: Compute $D^{(t+1)} = \mathrm{diag}((Z_1^{(t+1)})^\top \mathbf{1}_{n_1})$. Then form $W^{(t+1)} = (Z_1^{(t+1)})^\top (D^{(t+1)})^{-1}$.
5: **end for**

Output: A partition of the nodes $Z_1^{(T+1)}$.

COMPUTATIONAL COMPLEXITY OF GPM. Computing B has a cost $O(n_1 \mathrm{nnz}(A))$, as mentioned earlier. For each t, the cost of computing $BW^{(t)}$ is $O(n_1^2 K)$ and the cost of the projection is $O(Kn_1)$. The cost of computing $D^{(t+1)}$ and $W^{(t+1)}$ is $O(Kn_1)$. So the total cost over T iterations is $O(Tn_1^2 K)$ and doesn't depend on n_2.

Remark 13. *The recent work of Wang et al., 2021 proposed a power method for clustering under the SBM, but instead of using a projection on the simplex as we did, they add an additional constraint on the column on Z so that each cluster has the same size. They showed that computing this projection is equivalent to solving a Linear Assignment Problem (LAP). But the cost of solving this LAP is $O(K^2 n_1 \log n_1)$ whereas the cost of the projection on the simplex is $O(Kn_1)$. Moreover, their algorithm requires to know in advance the size of each cluster and it is not straightforward to extend their theoretical guarantees to the approximately balanced community setting.*

5.4 SPECTRAL INITIALIZATION

We show that Algorithm 9 can recover an arbitrary large proportion of community provided that the sparsity level p_{max} is sufficiently large, and $n_2/n_1 \to \infty$ sufficiently fast, as $n_1 \to \infty$.

Proposition 3. *Assume that* $A \sim \mathrm{BiSBM}(Z_1, Z_2, \Pi)$ *and Assumptions A3, A4 are satisfied. For any* $0 < \varepsilon < 1$*, suppose additionally that*

$$\frac{n_2}{n_1} \gtrsim \max(L \log n_1, \log^2 n_2), \quad p_{\max}^2 \geqslant \frac{KL \log n_1}{\varepsilon n_1 n_2}, \tag{5.4.1}$$

$$\frac{KL}{\log n_1} = o(1), \quad \frac{(KL)^3}{\frac{n_2}{n_1} \log n_1} = o(1) \quad \text{as } n_1 \to \infty. \tag{5.4.2}$$

Then the output $z^{(0)}$ *of Algorithm 9 satisfies with probability at least* $1 - n^{-\Omega(1)}$ *the bound* $r(z^{(0)}, z) \lesssim K\varepsilon^2$.

Proof. First, we will control the noise $\left\|B - PP^\top\right\|$ using Lemma 36. Note that the conditions in (5.4.1) imply the conditions on n_1, n_2 and $p_{\max}$ in Lemma 36. Since $\mathcal{H}(\cdot)$ is a linear operator and $B = \mathcal{H}(B)$, we have the decomposition

$$B = \mathcal{H}(PP^\top) + \mathcal{H}(PE^\top) + \mathcal{H}(EP^\top) + \mathcal{H}(EE^\top),$$

which leads to the following bound by triangle inequality

$$\begin{aligned}\left\|B - PP^\top\right\| &\leqslant \left\|\mathcal{H}(PE^\top) + \mathcal{H}(EP^\top)\right\| + \left\|\mathcal{H}(EE^\top)\right\| + \left\|\mathcal{H}(PP^\top) - PP^\top\right\| \\ &\leqslant 4\|EZ_2\| \left\|\Pi^\top Z_1^\top\right\| + \left\|\mathcal{H}(EE^\top)\right\| + \left\|\mathcal{H}(PP^\top) - PP^\top\right\|.\end{aligned} \tag{5.4.3}$$

Now let us observe that

$$\begin{aligned}\left\|\mathcal{H}(PP^\top) - PP^\top\right\| &= \left\|\mathcal{H}(Z_1 \Pi Z_2^\top Z_2 \Pi^\top Z_1^\top) - Z_1 \Pi Z_2^\top Z_2 \Pi^\top Z_1^\top\right\| \\ &= \left\|\mathrm{diag}(Z_1 \Pi Z_2^\top Z_2 \Pi^\top Z_1^\top)\right\| \\ &\leqslant \frac{\alpha n_2}{L}(L p_{\max}^2) \\ &\leqslant \alpha n_2 p_{\max}^2,\end{aligned}$$

while,

$$\left\|\Pi^\top Z_1^\top\right\| \leqslant \|Z_1\| \, \|\Pi\| \leqslant \sqrt{\frac{n_1}{K}} \|\Pi\|_F \leqslant \sqrt{n_1 L} p_{\max}.$$

Using these bounds along with the bounds on $\|EZ_2\|$ and $\left\|\mathcal{H}(EE^\top)\right\|$ from Lemma 36, and applying them in (5.4.3), it holds with probability at least $1 - n^{-\Omega(1)}$ that

$$\left\|B - PP^\top\right\| \lesssim \max(\log n_1 + \sqrt{n_1 n_2} p_{\max}) + \sqrt{n_2} n_1 p_{\max}^{1.5} + n_2 p_{\max}^2.$$

Let us denote by $\hat{U}$ (resp. U) to be the matrix of top-K eigenvectors of B (resp. $PP^\top$). Denoting $\lambda_K(PP^\top)$ to be the Kth largest eigenvalue of $PP^\top$, it is not difficult to verify that

$$\lambda_K(PP^\top) \geqslant \frac{n_2}{L} \lambda_K(Z_1)^2 \lambda_K(\Pi)^2 \gtrsim \frac{n_1 n_2 p_{max}^2}{KL}.$$

Then by the Davis-Kahan Theorem (Yu et al., 2014), there exists an orthogonal matrix $Q \in \mathbb{R}^{K\times K}$ such that

$$\|\hat{U} - UQ\| \lesssim \frac{\|B - PP^\top\|}{\lambda_K(PP^\top)} \lesssim \frac{KL\log n_1}{n_1 n_2 p_{\max}^2} + \frac{KL}{\sqrt{n_1 n_2}p_{\max}} + \frac{KL}{\sqrt{n_2 p_{\max}}} + \frac{KL}{n_1}. \tag{5.4.4}$$

where we bounded the max operation by the sum. Now on account of the conditions in (5.4.1) and (5.4.2), it is easily seen that $\frac{KL\log n_1}{n_1 n_2 p_{\max}^2} \lesssim \varepsilon$, while the remaining three terms in the RHS of (5.4.4) are $o(1)$ as $n_1 \to \infty$. Finally, we conclude by using the same argument as in the proof of Theorem 3.1 in Lei et al., 2015 to show that

$$r(\hat{z}, z) \lesssim \|\hat{U} - UQ\|_F^2 \lesssim K\|\hat{U} - UQ\|^2 \lesssim K\varepsilon^2.$$

□

Remark 14. *As shown in Section 5.7,* `Spec` *has very good empirical performance. This suggests that the previous proposition is far from being optimal and doesn't capture the true rate of convergence of this spectral method. Also note that Proposition 3 gives a meaningful bound only when* $\varepsilon = O(\frac{1}{\sqrt{K}})$.

5.5 ANALYSIS OF `gpm`

Our analysis strategy is similar to the one recently considered by Braun et al., 2021a for clustering under the contextual SBM, which in turn is based on the framework recently developed by Gao et al., 2019. There are however some additional technical difficulties due to the fact that there are more dependencies in the noise since the matrix B is a Gram matrix.

We will assume w.l.o.g. that π^* the permutation that best aligns $z^{(0)}$ with z is the identity, if not, then replace z by $(\pi^*)^{-1}(z)$. Hence there is no label switching ambiguity in the community labels of $z^{(t)}$ because they are determined from $z^{(0)}$.

First we will decompose the event "after one refinement step, the node i will be incorrectly clustered given the current estimation of the partition $z^{(t)}$ at time t" into an event independent of t, and events that depend on how close $z^{(t)}$ is from z. Then we will analyze these events separately. Finally, we will use these results to show that the error contracts at each step.

5.5.1 *Error decomposition*

By definition, a node i is misclustered at step $t+1$ if there exists a $k \neq z_i \in [K]$ such that

$$B_{i:}W_{:k}^{(t)} \geqslant B_{i:}W_{:z_i}^{(t)}. \tag{5.5.1}$$

By decomposing B as

$$B = \underbrace{\mathcal{H}(PP^\top)}_{\tilde{P}} + \underbrace{\mathcal{H}(EP^\top + PE^\top + EE^\top)}_{\tilde{E}}$$

one can show that condition (5.5.1) is equivalent to

$$\tilde{E}_{i:}(W_{:z_i} - W_{:k}) \leqslant -\Delta^2(z_i, k) + F_{ik}^{(t)} + G_{ik}^{(t)} \tag{5.5.2}$$

where

$$\begin{aligned}
\Delta^2(z_i, k) &= \tilde{P}_{i:}(W_{:z_i} - W_{:k}),\\
F_{ik}^{(t)} &= \langle \tilde{E}_{i:}(W^{(t)} - W), e_{z_i} - e_k \rangle,\\
\text{and } G_{ik}^{(t)} &= \langle \tilde{P}_{i:}(W^{(t)} - W), e_{z_i} - e_k \rangle.
\end{aligned}$$

Here $e_1, \dots, e_K$ denotes the canonical basis of $\mathbb{R}^K$. The terms $F_{ik}^{(t)}$ and $G_{ik}^{(t)}$ can be interpreted as error terms (due to $W^{(t)} \neq W$) while $\Delta^2(z_i, k)$ corresponds to the signal. Indeed, let us denote

$$Q = \Pi(Z_2^\top Z_2)\Pi^\top.$$

Then we obtain $PP^\top = Z_1 Q Z_1^\top$ which implies

$$\begin{aligned}
\Delta^2(z_i, k) = \tilde{P}_{i:}(W_{:z_i} - W_{:k}) &= Q_{z_i z_i} - Q_{z_i k} - (PP^\top)_{ii}(W_{iz_i} - W_{ik})\\
&= (1 - \frac{K}{n_1})Q_{z_i z_i} - Q_{z_i k}.
\end{aligned}$$

We have for all $k, k' \in [K]$ that

$$\frac{n_2}{\alpha L}(\Pi\Pi^\top)_{kk'} \leqslant Q_{kk'} \leqslant \frac{\alpha n_2}{L}(\Pi\Pi^\top)_{kk'}$$

which implies

$$\begin{aligned}
\frac{n_2}{\alpha L}\left((\Pi\Pi^\top)_{z_i z_i} - \alpha^2(\Pi\Pi^\top)_{z_i k}\right) \leqslant \Delta^2(z_i, k) \leqslant\\
\frac{\alpha n_2}{L}\left((\Pi\Pi^\top)_{z_i z_i} - \frac{1}{\alpha^2}(\Pi\Pi^\top)_{z_i k}\right)
\end{aligned}$$

for n_1 large enough. By Assumption A5 this implies

$$\beta\frac{n_2 p_{max}^2}{\alpha L} \leqslant \Delta^2(z_i, k) \leqslant \eta\beta\frac{\alpha n_2 p_{max}^2}{L}. \tag{5.5.3}$$

when n_1 is large enough.

5.5.2 *Oracle error*

We want to show that the condition (5.5.2) cannot occur with high probability. First we will show that this is indeed the case when we

ignore the F and G error terms; the subsequent error will be referred to as the oracle error

$$\xi(\delta) = \sum_{i=1}^{n_1} \sum_{k \in [K] \setminus z_i} \Delta^2(z_i, k) \mathbb{1}_{\Omega_{i,k}(\delta)} \text{ for } \delta \in (0,1)$$

where

$$\Omega_{i,k}(\delta) = \left\{ \tilde{E}_{i:}(W_{:z_i} - W_{:k}) \leqslant -(1-\delta)\Delta^2(z_i, k) \right\}.$$

Let us denote

$$\Delta_{min} = \min_{a \neq b \in [K]} \Delta(a, b)$$

to be the minimal separation of the parameters associated with the different communities. We will also denote by

$$\tilde{\Delta}^2 := \frac{\beta^2}{12eL\alpha^3} \frac{n_1 n_2 p_{max}^2}{KL},$$

to be the approximate signal-to-noise ratio (SNR) associated with the model.

In general the rate of decay of $\xi(\delta)$ leads to the convergence rate of iterative refinement algorithms, hence it is important to control this quantity by showing that the following condition is satisfied. Let us define

$$\tau^{(0)} = \epsilon' \delta \min(1, \beta^2) \frac{n_1 \Delta_{min}^2}{K}$$

for a small enough constant $\epsilon' > 0$. This parameter will be referred to as the minimal error required for the initial estimate of our algorithm.

Condition C4 (oracle error)**.** *Assume that there exists* $\delta \in (0,1)$ *such that*

$$\xi(\delta) \leqslant \frac{3}{4}\tau^{(0)}$$

holds with probability at least $1 - \eta_1$.

5.5.3 *Contraction of the error at each step*

Let

$$l(z, z') = \sum_{i \in [n]} \Delta^2(z_i, z'_i) \mathbb{1}_{\{z_i \neq z'_i\}}$$

be a measure of the distance between two partitions $z, z' \in [K]^n$. We want to show that $l(z^{(t)}, z)$ decreases until reaching the oracle error. To this end, we will need to control the noise level. In particular, we are going to show that the following two conditions are satisfied.

Condition C5 (F-error type)**.** *Assume that*

$$\max_{\{z^{(t)} : l(z, z^{(t)}) \leqslant \tau^{(0)}\}} \sum_{i=1}^{n} \max_{b \in [K] \setminus z_i} \frac{(F_{ib}^{(t)})^2}{\Delta^2(z_i, b) l(z, z^{(t)})} \leqslant \frac{\delta^2}{256}$$

holds with probability at least $1 - \eta_2$.

Condition C6 (G-error type). *Assume that*

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

holds uniformly on the event $\{z^{(t)} : l(z, z^{(t)}) \leqslant \tau^{(0)}\}$ *with probability at least* $1-\eta_3$.

Under these conditions, we can show that the error contracts at each step.

Theorem 11. *Assume that* $l(z^{(0)}, z) \leqslant \tau^{(0)}$. *Additionally assume that Conditions C4, C5, and C6 hold. Then with probability at least* $1 - \sum_{i=1}^{3} \eta_i$

$$l(z^{(t)}, z) \leqslant \xi(\delta) + \frac{1}{8} l(z^{(t-1)}, z), \forall t \geqslant 1. \tag{5.5.4}$$

In particular, we have for all $t \gtrsim \log(1/\delta)$ *that*

$$l(z^{(t)}, z) \lesssim \xi(\delta) + \tau^{(0)}(1/8)^{t-\Theta(\log(1/\delta))}.$$

Proof. It is an immediate adaptation of Theorem 3.1 in Gao et al., 2019. The last part can be derived in the same way as in Corollary 1 in Braun et al., 2021b. □

5.5.4 *Application to BiSBM*

When applied to the BiSBM the previous theorem leads to the following result.

Theorem 12. *Assume that* $A \sim \mathrm{BiSBM}(Z_1, Z_2, \Pi)$, $K^2 L \leqslant \tilde{\Delta}^2$, $\tilde{\Delta} \to \infty$, (5.4.1) *and* (5.4.2) *are satisfied. Then, under Assumption A3, A4 and A5, if* `GPM` *is initialized with a* $z^{(0)}$ *such that*

$$l(z, z^{(0)}) \leqslant \tau^{(0)} = \epsilon' \delta \min(1, \beta^2) \frac{n_1 \Delta_{min}^2}{K}$$

for a small enough constant $\epsilon' > 0$ *and* $\delta = \frac{1}{4\eta\alpha}$. *Then with probability at least* $1 - n_1^{-\Omega(1)}$ *we have for all* $t \gtrsim \log n_1$

$$r(z^{(t)}, z) \leqslant \exp(-(1 - o(1))\tilde{\Delta}^2).$$

In particular, if $\tilde{\Delta}^2 > \log n_1$, *we can exactly recover* Z_1.

Corollary 2. *Under the assumptions of Theorem 12,* `Spec` *returns an estimate* $z^{(0)}$ *such that*

$$l(z, z^{(0)}) \leqslant \eta\beta \frac{\alpha n_1 n_2 p_{max}^2}{L} r(z, z^{(0)}) \lesssim \eta\beta K^2 \varepsilon^2 \frac{n_1 \Delta_{min}^2}{K}$$

and hence satisfies $l(z, z^{(0)}) \leqslant \tau^{(0)}$ *for* $\varepsilon = O(\frac{\min(1,\beta^2)}{K\eta\sqrt{\eta\beta}})$.

Remark 15. *Specialized to the SBISBM case where* $K = 2 = L$, $q = cp$ *for some* $0 < c < 1$, *the condition* $\tilde{\Delta}^2 = \Omega(\log n_1)$ *needed for exact recovery in the above theorem is equivalent to the condition* $p_{max} = \Omega(\sqrt{\frac{\log n_1}{n_1 n_2}})$ *needed by Ndaoud et al., 2022. We will show in the next section that this rate is optimal.*

Remark 16. *The condition* $n_1 n_2 p_{max}^2 \gtrsim KL \log n_1$ *is needed because we used the concentration inequality* $\|\mathcal{H}(EE^\top)\| \lesssim \max(\log n_1, \sqrt{n_1 n_2} p_{max})$ *in Lemma 36. If this last condition could have been replaced by* $\|\mathcal{H}(EE^\top)\| \lesssim \sqrt{n_1 n_2} p_{max}$ *then we would only need the condition* $\tilde{\Delta} \to \infty$. *This is the principal step that needs to be improved in order to get more general weak consistency guarantees. A possible way to obtain this tighter concentration inequality would be to adapt the combinatorial method of Friedman et al., 1989 similarly to what have been done to obtain a tight concentration inequality for similarity matrices formed from an hypergraph in Lee et al., 2020.*

Proof of Theorem 12. In order to apply Theorem 4, we need to show that Conditions C4, C5 and C6 are satisfied.

ORACLE ERROR. Let $\delta = \frac{1}{4\eta\alpha}$. We obtain by Lemma 19 (see Section 5.8) and Assumption A5 that

$$\mathbb{E}(\xi(\delta)) \leqslant \alpha\eta\beta K n_1 \frac{n_2}{L} p_{max}^2 e^{-\tilde{\Delta}^2} \lesssim \frac{\alpha^4}{\beta} K^2 L \tilde{\Delta}^2 e^{-\tilde{\Delta}^2} = e^{-(1-o(1))\tilde{\Delta}^2}$$

by assumptions on $\tilde{\Delta}$ and KL. Then, by Markov inequality, we obtain

$$\mathbb{P}\left(\xi(\delta) \geqslant \exp(\tilde{\Delta})\mathbb{E}\xi(\delta)\right) \leqslant \exp(-\tilde{\Delta}).$$

Consequently we get

$$\exp(\tilde{\Delta})\mathbb{E}\xi(\delta) \leqslant \exp(-(1-o(1))\tilde{\Delta}^2)$$

and hence with probability at least $1 - \exp(-\tilde{\Delta})$

$$\xi(\delta) \leqslant \exp(-(1-o(1))\tilde{\Delta}^2) \leqslant \frac{3}{4}\tau^{(0)}$$

because $\exp(-(1-o(1))\tilde{\Delta}^2) = o(1) \ll \tau^{(0)} = \Omega(1)$ for n_1 large enough. This shows that Condition C4 is satisfied.

F-ERROR TERM. W.h.p, for all $z^{(t)}$ such that $l(z^{(t)}, z) \leqslant \tau^{(0)}$ we have

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

$$\leqslant \sum_i \left\| \tilde{E}_{i:}(W^{(t)} - W) \right\|^2 \max_{b \in [K] \setminus z_i} \frac{\|e_{z_i} - e_b\|^2}{\Delta^2(z_i, b) l(z, z^{(t)})}$$

(by Cauchy-Schwartz)

$$\leqslant \frac{2\left\|\tilde{E}(W^{(t)}-W)\right\|_F^2}{\Delta_{min}^2 l(z^{(t)},z)}$$
$$\leqslant 2\frac{\left\|\tilde{E}\right\|^2\left\|W^{(t)}-W\right\|_F^2}{\Delta_{min}^2 l(z^{(t)},z).}$$

By Lemma 36, we have w.h.p.

$$\left\|\mathcal{H}(EE^\top)\right\| \lesssim \left\|EE^\top - \mathbb{E}(EE^\top)\right\| \lesssim \max(\log n_1, \sqrt{n_1 n_2}p_{max})$$

and

$$\left\|EP^\top\right\| \leqslant \|EZ_2\|\left\|\Pi Z_1^\top\right\| \lesssim \sqrt{\frac{n_1 n_2 p_{max}}{L}}p_{max}\sqrt{\frac{n_1}{K}}$$
$$= (\sqrt{n_1 n_2}p_{max})\sqrt{\frac{n_1 p_{max}}{KL}}.$$

Since $\left\|\tilde{E}\right\| \lesssim \left\|\mathcal{H}(EE^\top)\right\| + \left\|EP^\top\right\|$ we obtain

$$\sum_{i=1}^{n}\max_{b\in[K]\setminus z_i}\frac{(F_{ib}^{(t)})^2}{\Delta^2(z_i,b)l(z,z^{(t)})}$$
$$\lesssim \frac{\max\left(\log^2 n_1, n_1 n_2 p_{max}^2, (KL)^{-1}n_1^2 n_2 p_{max}^3\right)K^3 l(z^{(t)},z)}{n_1^3\Delta_{min}^6}$$
$$\lesssim K^2\frac{\max\left(\log^2 n_1, n_1 n_2 p_{max}^2, (KL)^{-1}n_1^2 n_2 p_{max}^3\right)}{n_1^2\Delta_{min}^4}\frac{K\tau^{(0)}}{n_1\Delta_{min}^2}$$
$$\lesssim \frac{K^2L^2\max\left(\log^2 n_1, n_1 n_2 p_{max}^2, (KL)^{-1}n_1^2 n_2 p_{max}^3\right)}{\beta^2 n_1^2 n_2^2 p_{max}^4}\epsilon'\delta\beta^2$$
(by definition of $\tau^{(0)}$ and (5.5.3))
$$\lesssim \frac{K^2L^2\left(\log^2 n_1 + n_1 n_2 p_{max}^2 + (KL)^{-1}n_1^2 n_2 p_{max}^3\right)}{n_1^2 n_2^2 p_{max}^4}\epsilon'\delta.$$

By assumptions, we have $\frac{KL}{n_2 p_{max}} = o(1)$ and

$$\frac{K^2L^2\log^2 n_1}{n_1^2 n_2^2 p_{max}^4} = O(1), \quad \frac{K^2L^2}{n_1 n_2 p_{max}^2} \lesssim \frac{KL}{\log n_1} = o(1).$$

Hence for an appropriately small choice of ϵ' Condition C5 is satisfied.

G-ERROR TERM. W.h.p, for all $z^{(t)}$ such that $l(z^{(t)},z) \leqslant \tau^{(0)}$ we have

$$\frac{|G_i^{(t)}|}{\Delta^2(z_i,b)} \leqslant \sqrt{2}\frac{\left\|\tilde{P}_{i:}(W^{(t)}-W)\right\|}{\Delta_{min}^2}. \qquad \text{(by Cauchy-Schwartz)}$$

But since $\|\tilde{P_{i:}}\| \lesssim \|P_{i:}P^\top\| \leqslant \|\Pi\|^2 \|Z_2^\top Z_2\| \|Z_1\| \leqslant \frac{\alpha n_2}{L} p_{max}^2 \sqrt{\frac{n_1}{K}}$ we obtain

$$\begin{aligned}
\frac{|G_i^{(t)}|}{\Delta^2(z_i, b)} &\lesssim \frac{n_2 p_{max}^2 \sqrt{n_1} \|(W^{(t)} - W)\|}{L\sqrt{K}\Delta_{min}^2} \\
&\lesssim \frac{n_2 p_{max}^2}{L\Delta_{min}^2} \frac{Kl(z^{(t)}, z)}{n_1 \Delta_{min}^2} \qquad \text{(by Lemma 11)} \\
&\lesssim K \frac{l(z^{(t)}, z)}{\beta n_1 \Delta_{min}^2} \\
&\lesssim \frac{K\tau^{(0)}}{\beta n_1 \Delta_{min}^2}.
\end{aligned}$$

Now by choosing ϵ' to be a suitably small constant (< 1), we obtain

$$\frac{|G_i^{(t)}|}{\Delta^2(z_i, b)} \leqslant \frac{\delta}{4}.$$

This shows that Condition C6 is satisfied. □

5.6 MINIMAX LOWER BOUND

Let us denote the admissible parameters space by

$$\begin{aligned}
\Theta = \{P \in [0,1]^{n_1 \times n_2} : P = Z_1 \Pi Z_2^\top \text{ where } \Pi = q\mathbb{1}_K\mathbb{1}_K^\top + (p-q)I_K, \\
1 > p > 0, q = cp \text{ for some constant } c \in (0,1), Z_1 \in \mathcal{M}_{n_1,K}, \\
Z_2 \in \mathcal{M}_{n_2,K} \text{ with } \alpha = 1 + O\left(\sqrt{\frac{\log n_1}{n_1}}\right)\}.
\end{aligned}$$

We want to lower bound $\inf_{\hat{z}} \sup_{\theta\in\Theta} \mathbb{E}(r(\hat{z}, z))$ where $\hat{z}$ is an estimator of Z_1. For simplicity, we will assume that $K = 2$, but we believe that the same argument can be extended to SBiSBM with an arbitrary number of communities K. In the supervised case, i.e. when Z_2 is known, we can use the same strategy as the one use for the degree-corrected SBM by Gao et al., 2018 and obtain a lower bound of the order $e^{-(1+o(1))n_2(p-q)/2}$ corresponding to the failure probability of the optimal test associated to the following two hypothesis problem

$$\begin{aligned}
H_0 &: \otimes_{i=1}^{n_2/2} \mathcal{B}(p) \otimes_{i=n_2/2+1}^{n_2} \mathcal{B}(q), \text{ vs} \\
H_1 &: \otimes_{i=1}^{n_2/2} \mathcal{B}(q) \otimes_{i=n_2/2+1}^{n_2} \mathcal{B}(p).
\end{aligned}$$

However when $n_2 \gg n_1 \log n_1$ the error associated with this hypothesis testing problem is of order $\exp(-n_2(p-q)/2) \approx \exp(-\sqrt{\frac{n_2 \log n_1}{n_1}})$ but this is far smaller than $\exp(-n_1 n_2 p_{max}^2)$, the misclustering rate obtained for our algorithm. A similar phenomena appears for high-dimensional Gaussian mixture models. Indeed, as shown in Ndaoud, 2018, it is essential to capture the hardness of estimating the model

parameters in the minimax lower bound in order to get the right rate of convergence. The argument developed for obtaining a lower bound of the minimax risk in Ndaoud, 2018 relies heavily on the Gaussian assumption (they set a Gaussian prior on the model parameters and use the fact that the posterior distribution is also Gaussian) and cannot directly be extended to this setting.

Our lower bound result is summarised in the following theorem.

Theorem 13. *Suppose that* $A \sim \mathrm{SBiSBM}(Z_1, Z_2, p, q)$ *with* $K = L = 2$, $n_2 \gg n_1 \log n_1$, $n_1 n_2 p_{max}^2 \to \infty$ *and* $n_1 n_2 p_{max}^2 = O(\log n_1)$*. Then there exists a constant* $c_1 > 0$ *such that*

$$\inf_{\hat{z}} \sup_{\theta \in \Theta} \mathbb{E}(r(\hat{z}, z)) \geqslant \exp(-c_1 n_1 n_2 p_{max}^2)$$

where the infimum is taken over all measurable functions $\hat{z}$ *of* A*. Moreover, if* $n_1 n_2 p_{max}^2 = \Theta(1)$*, then* $\inf_{\hat{z}} \sup_{\theta \in \Theta} \mathbb{E}(r(\hat{z}, z)) \geqslant c_2$ *for some positive constant* c_2*.*

Remark 17. *This lower-bound shows that the rate of convergence of our estimator is optimal up to a constant factor. Indeed, for exact recovery, the minimax lower bound implies that* $n_1 n_2 p_{max}^2 \gtrsim \log n_1$ *is necessary, while for weak recovery, we need* $n_1 n_2 p_{max}^2 \to \infty$*. It also shows that when* $n_1 n_2 p_{max}^2 = O(1)$ *it is not possible to consistently estimating* Z_1*.*

Remark 18. *For simplicity, we only write the proof for* $K = L = 2$ *but we believe it can be extended to the case* $K = L > 2$ *by reducing the problem to test between two fixed communities as in Gao et al., 2018.*

Proof. The general idea of the proof is to first lower bound the minimax risk by an error occuring in a two hypothesis testing problem and then to replace this hypothesis testing problem by a simpler one. The steps are detailed below.

STEP 1. Recall that z, z' denote the partition functions associated with Z_1 and Z_2 respectively. We choose as a prior on z and z' a product of independent, centered Rademacher distributions. Since the marginals of z, z' are sign invariant, By using standard arguments the results in Gao et al., 2018 or Ndaoud, 2018 show that

$$\inf_{\hat{z}} \sup_{\theta \in \Theta} \mathbb{E}(r(\hat{z}, z)) \gtrsim \frac{2}{n_1} \sum_{i=1}^{\lceil n_1/2 \rceil} \inf_{\hat{z}_i} \mathbb{E}_{z,z'} \mathbb{E}_{A|z,z'}(\phi_i(A))$$

where $\hat{z}$ is a measurable function in A and $\phi_i(A) = \mathbb{1}_{\hat{z}_i \neq z_i}$.

STEP 2. We can write

$$\mathbb{E}_{z,z'} \mathbb{E}_{A|z,z'}(\phi_i(A)) = \mathbb{E}_{z_{-i}} \mathbb{E}_{z_i} \mathbb{E}_{z'} \mathbb{E}_{A|z,z'}(\phi_i(A)) = \underbrace{\mathbb{E}_{z_{-i}} \mathbb{E}_{z_i} \mathbb{E}_{A|z}(\phi_i(A))}_{R_i}$$

where $z_{-i} := (z_j)_{j \neq i}$ and $\mathbb{E}_{X|Y}$ means that we integrate over the random variable X conditioned on Y. Since z' is random, the entries of A are no longer independent (this is the reason why we use an uninformative prior on z'). Note that the quantity $\mathbb{E}_{z'}\mathbb{E}_{A|z,z'}(\phi_i(A)) = \mathbb{E}_{A|z}(\phi_i(A))$ only depends on $f(A|z)$ the density of A conditionally on z. Since the columns of A are independent conditionally on z (because $(z'_j)_{j=1}^{n_2}$ are independent) we have that

$$f(A|z) = \prod_j f(A_{:j}|z),$$

i.e. $f(A|z)$ is the product of the densities of the columns $A_{:j}$ conditionally on z. Let us denote $A_{-ij} := (A_{i'j})_{i' \neq i}$ and $A_{-i:} = (A_{i'j})_{i' \neq i, j}$. Now for each j we can write

$$f(A_{:j}|z) = f(A_{ij}|A_{-ij}, z) f(A_{-ij}|z).$$

Since A_{-ij} doesn't depend on z_i we have $f(A_{-ij}|z) = f(A_{-ij}|z_{-i})$. Assume that $z_i = 1$. Then $A_{ij} \sim \mathcal{B}(p)$ if $z'_j = 1$ or $A_{ij} \sim \mathcal{B}(q)$ if $z'_j = -1$. This in turn implies

$$\begin{aligned}\mathbb{P}(A_{ij} = 1|A_{-ij}, z) &= p\mathbb{P}(z'_j = 1|A_{-ij}, z) + q\mathbb{P}(z'_j = -1|A_{-ij}, z) \\ &= p\mathbb{P}(z'_j = 1|A_{-ij}, z_{-i}) + q\mathbb{P}(z'_j = -1|A_{-ij}, z_{-i})\end{aligned}$$

Let us denote by α_j the random variable $\mathbb{P}(z'_j = 1|A_{-ij}, z_{-i})$. When $z_i = -1$, similar considerations imply

$$\mathbb{P}(A_{ij} = 1|A_{-ij}, z_{-i}) = q\alpha_j + p(1 - \alpha_j).$$

This shows that the conditional distribution

$$A_{ij}|A_{-ij}, z \sim \mathcal{B}(\alpha_j p + (1 - \alpha_j)q) \quad \text{when } z_i = 1 \tag{5.6.1}$$

and

$$A_{ij}|A_{-ij}, z \sim \mathcal{B}(\alpha_j q + (1 - \alpha_j)p) \quad \text{when } z_i = -1. \tag{5.6.2}$$

We can write R_i as

$$\begin{aligned}R_i &= \mathbb{E}_{z_{-i}}\mathbb{E}_{z_i}\mathbb{E}_{A_{-i:}|z}\mathbb{E}_{A_{i:}|z,A_{-i:}}(\phi_i(A)) \\ &= \mathbb{E}_{z_{-i}}\mathbb{E}_{z_i}\mathbb{E}_{A_{-i:}|z_{-i}}\mathbb{E}_{A_{i:}|z,A_{-i:}}(\phi_i(A)) \\ &= \mathbb{E}_{z_{-i}}\mathbb{E}_{A_{-i:}|z_{-i}} \underbrace{\mathbb{E}_{z_i}\mathbb{E}_{A_{i:}|z,A_{-i:}}(\phi_i(A))}_{R'_i}.\end{aligned}$$

The term R'_i corresponds to the risk associated with the following two hypothesis testing problem (conditionally on $(A_{i'j})_{i' \neq i}$ and z_{-i})

$$H_0 : \otimes_{j=1}^{n_2}\mathcal{B}(\alpha_j p + (1 - \alpha_j)q) \text{ vs } H_1 : \otimes_{j=1}^{n_2}\mathcal{B}(\alpha_j q + (1 - \alpha_j)p). \tag{5.6.3}$$

Our goal is to replace this two-hypothesis testing problem by a simpler hypothesis test associated with a smaller error. The proof strategy is the following. When α_j is very close to $1/2$ it is not possible to

statistically distinguish $(\mathcal{B}(\alpha_j p + (1-\alpha_j)q)$ from $(\mathcal{B}(\alpha_j q + (1-\alpha_j)p)$ and these factors can be dropped. When α_j is significantly different from $1/2$, then the risk associated to the test can be lower bounded by a the risk of testing between a product of $\mathcal{B}(p)$ vs. a product of $\mathcal{B}(q)$. More precisely, we will show that the number of indices j for which α_j is significantly different from $1/2$ is of order $n_1 n_2 p$ and hence the risk is lower bounded by the one associated by the two hypothesis testing problem

$$H_0' : \otimes_{i=1}^{n_2 n_1 p} \mathcal{B}(p) \otimes_{i=1}^{n_2 n_1 p} \mathcal{B}(q) \text{ vs } H_1' : \otimes_{i=1}^{n_2 n_1 p} \mathcal{B}(q) \otimes_{i=1}^{n_2 n_1 p} \mathcal{B}(p).$$

It is well known that the error associated with the above testing problem is of the order $e^{-\Theta(n_2 n_1 p^2)}$ (see e.g. Gao et al., 2018) which corresponds to the rate of convergence of our algorithm. Let us now formalize the above argument.

STEP 3. First, let us define

$$\begin{aligned}
\mathcal{C}_+ &= \{i' \neq i : z_{i'} = 1\},\\
\mathcal{C}_- &= \{i' \neq i : z_{i'} = -1\},\\
\theta_j &= \frac{\alpha_j}{1-\alpha_j} \text{ for } j = 1 \dots n_2,\\
\epsilon &= \Theta(p\sqrt{n_1 \log n_1}) = o(1),\\
J_b &= \{j \in [n_2] : \theta_j \in [1-\epsilon, 1+\epsilon]\},\\
J_g &= \{j \in [n_2] : \theta_j \notin [1-\epsilon, 1+\epsilon]\},\\
T_j &\overset{ind.}{\sim} \mathcal{B}(\alpha_j), \text{ for all } j \in J_g,
\end{aligned}$$

and the events

$$\begin{aligned}
\mathcal{E}_1 &= \left\{ |\mathcal{C}_+| - |\mathcal{C}_-| \in [-C\sqrt{n_1 \log n_1}, C\sqrt{n_1 \log n_1}] \right\}\\
&\quad \text{(for some constant)} C > 0,\\
\mathcal{E}_2 &= \left\{ \sum_j \mathbb{1}_{\{\sum_{i' \in \mathcal{C}_+} A_{i'j} - \sum_{i' \in \mathcal{C}_-} A_{i'j} \neq 0\}} = \Theta(n_2 n_1 p) \right\},\\
\mathcal{E}_3 &= \left\{ \sum_{j \in J_b} A_{ij} = \Theta(n_2 p) \right\}.
\end{aligned}$$

We will show later that these events occur with high probability. They are useful for obtaining a lower bound on the densities $f(A_{ij}|A_{-ij}, z)$. Since we are integrating over positive functions one can write

$$R_i \geqslant \mathbb{E}_{z_{-i}} \mathbb{1}_{\mathcal{E}_1} \mathbb{E}_{A_{-i:}|z_i} \mathbb{1}_{\mathcal{E}_2} \mathbb{E}_{z_i} \mathbb{E}_{A_{i:}|z, A_{-i:}} (\mathbb{1}_{\mathcal{E}_3} \phi_i(A)). \tag{5.6.4}$$

STEP 4. For all $j \in J_b$, the set for which $\alpha_j \approx \frac{1}{2}$, we are going to lower bound the densities $f(A_{ij}|A_{-ij}, z)$ by $g(A_{ij})$ corresponding to the density of $\mathcal{B}(\frac{p+q}{2})$. A simple calculation shows that

$$\theta_j = \frac{\alpha_j}{1-\alpha_j} = \frac{\prod_{i' \neq i: z_{i'}=1} p^{A_{i'j}}(1-p)^{1-A_{i'j}} \prod_{i' \neq i: z_{i'}=-1} q^{A_{i'j}}(1-q)^{1-A_{i'j}}}{\prod_{i' \neq i: z_{i'}=1} q^{A_{i'j}}(1-q)^{1-A_{i'j}} \prod_{i' \neq i: z_{i'}=-1} p^{A_{i'j}}(1-p)^{1-A_{i'j}}}.$$

The previous expression can be rewritten as

$$\theta_j = \left(\frac{p(1-q)}{q(1-p)}\right)^{\sum_{i' \in \mathcal{C}_+} A_{i'j} - \sum_{i' \in \mathcal{C}_-} A_{i'j}} \left(\frac{1-p}{1-q}\right)^{|\mathcal{C}_+|-|\mathcal{C}_-|}.$$

We also have the relation $\alpha_j = \frac{\theta_j}{1+\theta_j}$, so α_j is close to $1/2$ if and only if θ_j is close to 1. If z_{-i} were an exactly balanced partition, i.e., $|\mathcal{C}_+| - |\mathcal{C}_-| = 0$, then $\theta_j = 1$ would be equivalent to $\sum_{i' \in \mathcal{C}_+} A_{i'j} - \sum_{i' \in \mathcal{C}_-} A_{i'j} = 0$. However the contribution of the term $\left(\frac{1-p}{1-q}\right)^{|\mathcal{C}_+|-|\mathcal{C}_-|}$ is small under $\mathcal{E}_1$. Indeed, we have

$$|\mathcal{C}_+| - |\mathcal{C}_-| \in [-C\sqrt{n_1 \log n_1}, C\sqrt{n_1 \log n_1}]$$

. Note that $\log(\frac{1-p}{1-q}) \in [-c(p-q), c(p-q)]$ by using Taylor's formula for some constant $c > 0$. Hence, under $\mathcal{E}_1$

$$\left(\frac{1-p}{1-q}\right)^{|\mathcal{C}_+|-|\mathcal{C}_-|} \in [e^{-c'\sqrt{n_1 \log n_1}p}, e^{c'\sqrt{n_1 \log n_1}p}]$$

for some constant $c' > 0$. But we have

$$\max(|e^{c'p\sqrt{n_1 \log n_1}} - 1|, |e^{-c'p\sqrt{n_1 \log n_1}} - 1|) \leqslant c'p\sqrt{n_1 \log n_1} := \epsilon.$$

Therefore, it follows that $(\frac{1-p}{1-q})^{|\mathcal{C}_+|-|\mathcal{C}_-|} \in [1-\epsilon, 1+\epsilon]$ under $\mathcal{E}_1$. It is easy to check $\theta_j \in [1-\epsilon, 1+\epsilon]$ implies $\alpha_j \in [1/2-\epsilon', 1/2+\epsilon']$ for ϵ' proportional to ϵ. Since the constant involved here doesn't matter, we won't make a distinction between ϵ and ϵ'.

Now recall that by (5.6.1) and (5.6.2) we have when $z_i = 1$

$$f(A_{ij}|A_{-ij}, z) = \alpha_j p^{A_{ij}}(1-p)^{1-A_{ij}} + (1-\alpha_j)q^{A_{ij}}(1-q)^{1-A_{ij}}$$

and when $z_i = -1$

$$f(A_{ij}|A_{-ij}, z) = \alpha_j q^{A_{ij}}(1-q)^{1-A_{ij}} + (1-\alpha_j)p^{A_{ij}}(1-p)^{1-A_{ij}}.$$

- When $z_i = 1$, we have for all $j \in [n_2]$ such that $\alpha_j \in [1/2-\epsilon, 1/2+\epsilon]$ that

$$\alpha_j p^{A_{ij}}(1-p)^{1-A_{ij}} + (1-\alpha_j)q^{A_{ij}}(1-q)^{1-A_{ij}} \geqslant (1-\epsilon)\frac{p+q}{2}\mathbb{1}_{A_{ij}=1} + (1-p\epsilon)\frac{1-p+1-q}{2}\mathbb{1}_{A_{ij}=0}$$

and

$$\prod_{j \in J_b} \alpha_j p^{A_{ij}}(1-p)^{1-A_{ij}} + (1-\alpha_j)q^{A_{ij}}(1-q)^{1-A_{ij}} \geqslant (1-\epsilon)^{\sum_{j \in J_b} A_{ij}}(1-p\epsilon)^{\sum_{j \in J_b}(1-A_{ij})} \times \prod_{j \in J_b} \frac{p^{A_{ij}}(1-p)^{1-A_{ij}} + q^{A_{ij}}(1-q)^{1-A_{ij}}}{2}.$$

- When $z_i = -1$, we have for all $j \in [n_2]$ such that $\alpha_j \in [1/2 - \epsilon, 1/2 + \epsilon]$

$$\alpha_j q^{A_{ij}}(1-q)^{1-A_{ij}} + (1-\alpha_j)p^{A_{ij}}(1-p)^{1-A_{ij}} \geqslant (1-\epsilon)\frac{p+q}{2}\mathbb{1}_{A_{ij}=1} + (1-q\epsilon)\frac{1-p+1-q}{2}\mathbb{1}_{A_{ij}=0}$$

and

$$\prod_{j\in J_b} \alpha_j q^{A_{ij}}(1-q)^{1-A_{ij}} + (1-\alpha_j)p^{A_{ij}}(1-p)^{1-A_{ij}} \geqslant (1-\epsilon)^{\sum_{j\in J_b} A_{ij}}(1-q\epsilon)^{\sum_{j\in J_b}(1-A_{ij})} \times \prod_{j\in J_b} \frac{p^{A_{ij}}(1-p)^{1-A_{ij}} + q^{A_{ij}}(1-q)^{1-A_{ij}}}{2}.$$

Since on $\mathcal{E}_3$, $\sum_{j\in J_b} A_{ij} = \Theta(n_2 p)$ we obtain that

$$(1-\epsilon)^{\sum_{j\in J_b} A_{ij}} \geqslant 1 - n_2\epsilon p = 1 - o(1)$$

and similarly

$$(1-p\epsilon)^{\sum_{j\in J_b}(1-A_{ij})} \geqslant 1 - o(1) \text{ and } (1-q\epsilon)^{\sum_{j\in J_b}(1-A_{ij})} \geqslant 1 - o(1).$$

This implies the lower bound

$$\mathbb{1}_{\mathcal{E}_3} \prod_{j\in J_b} f(A_{ij}|A_{-ij}, z) \geqslant (1-o(1))\mathbb{1}_{\mathcal{E}_3} \prod_{j\in J_b} g(A_{ij}).$$

Consequently, we obtain the lower bound

$$\begin{aligned}
&\mathbb{E}_{z_i}\mathbb{E}_{A_{i:}|z,A_{-i:}}(\mathbb{1}_{\mathcal{E}_3}\phi_i(A)) \\
&\geqslant (1-o(1))\int_{z_i}\int_{A_{i:}} \mathbb{1}_{\mathcal{E}_3}\phi_i(A)\prod_{j\in J_b} g(A_{ij})\prod_{j\in J_g} f(A_{ij}|A_{-ij},z)\mathrm{d}A_{i:}\mathrm{d}\mathbb{P}(z_i) \\
&\gtrsim \int_{z_i}\int_{(A_{ij})_{j\in J_b}}\int_{(A_{ij})_{j\in J_g}} \prod_{j\in J_b} g(A_{ij}) \left(\phi_i(A)\prod_{j\in J_g} f(A_{ij}|A_{-ij},z)\mathrm{d}(A_{ij})_{j\in J_g}\right)\mathrm{d}(A_{ij})_{j\in J_b}\mathrm{d}\mathbb{P}(z_i) \\
&\quad - \tilde{\mathbb{P}}_{A_{i:}|A_{-i:},z}(\mathcal{E}_3^c) \\
&\gtrsim \int_{z_i}\int_{(A_{ij})_{j\in J_b}} \prod_{j\in J_b} g(A_{ij}) \int_{(A_{ij})_{j\in J_g}} \left(\phi_i(A)\prod_{j\in J_g} f(A_{ij}|A_{-ij},z)\mathrm{d}(A_{ij})_{j\in J_g}\right)\mathrm{d}(A_{ij})_{j\in J_b}\mathrm{d}\mathbb{P}(z_i) \\
&\quad - \tilde{\mathbb{P}}_{A_{i:}|A_{-i:},z}(\mathcal{E}_3^c) \qquad \text{(since } J_b \text{ and } J_g \text{ are disjoint)} \\
&\gtrsim \int_{(A_{ij})_{j\in J_b}} \prod_{j\in J_b} g(A_{ij}) \underbrace{\left(\int_{z_i}\int_{(A_{ij})_{j\in J_g}} \phi_i(A)\prod_{j\in J_g} f(A_{ij}|A_{-ij},z)\mathrm{d}(A_{ij})_{j\in J_g}\right)}_{R_i''}\mathrm{d}(A_{ij})_{j\in J_b}\mathrm{d}\mathbb{P}(z_i) \\
&\quad - \tilde{\mathbb{P}}_{A_{i:}|A_{-i:},z}(\mathcal{E}_3^c) \qquad \text{(since } g(A_{ij}) \text{ is independent from } z_i)
\end{aligned}$$

where $\tilde{\mathbb{P}}_{A_{i:}|A_{-i:},z}$ is the conditional distribution corresponding to the density product

$$\prod_{j\in J_b} g(A_{ij}) \prod_{j\in J_g} f(A_{ij}|A_{-ij},z)\mathrm{d}\mathbb{P}(z_i).$$

STEP 5. R_i'' corresponds to risk associated to the testing problem (5.6.3) where the product is restricted to the set J_g. One can lower bound this risk as follows.

By definition of T_j, we have when $z_i = 1$ by equation (5.6.1)

$$f(A_{ij}|A_{-ij}, z) = \underbrace{\mathbb{E}_{T_j}\left(T_j p^{A_{ij}}(1-p)^{1-A_{ij}} + (1-T_j)q^{A_{ij}}(1-q)^{1-A_{ij}}\right)}_{:=\mathbb{E}_{T_j}\left(f_{T_j}(A_{ij})\right)}.$$

A similar result holds when $z_i = -1$. Since T_j is independent of z_i (it only depends on A_{-ij} and z_{-i} through α_j) we can write

$$R_i'' = \mathbb{E}_T \mathbb{E}_{z_i} \underbrace{\int_{(A_{ij})_{j\in J_g}} \phi_i(A) \prod_{j\in J_g} f_{T_j}(A_{ij}) d(A_{ij})_{j\in J_g}}_{R_i'''} = \mathbb{E}_T(R_i''')$$

where $T = (T_j)_{j\in J_g}$. But now R_i''' is the risk associated with the following two hypothesis testing problem

$$H_0'' : \otimes_{T_j=1} \mathcal{B}(p) \otimes_{T_j=0} \mathcal{B}(q), \text{ vs}$$
$$H_1'' : \otimes_{T_j=1} \mathcal{B}(q) \otimes_{T_j=0} \mathcal{B}(p).$$

The error associated with this test is lower bounded by the error associated with the test

$$H_0''' : \otimes_{j\in J_g} \mathcal{B}(p) \otimes_{j\in J_g} \mathcal{B}(q), \text{ vs}$$
$$H_1''' : \otimes_{j\in J_g} \mathcal{B}(q) \otimes_{j\in J_g} \mathcal{B}(p).$$

since adding more information can only decreases the error (more formally, the lower bound follows from the fact that we multiply by density functions inferior to one). Under $\mathcal{E}_2 \cap \mathcal{E}_1$ we have

$$|J_g| = \Theta(n_1 n_2 p) \quad \text{and } |J_b| = (1 - o(1))n_2. \tag{5.6.5}$$

Hence, by Lemma 4 in Gao et al., 2018 the risk is lower bounded by $R_i'' \leqslant e^{-\Theta(n_1 n_2 p^2)}$.

CONCLUSION. It remains to integrate over all the events we conditioned on. We have shown sor far that

$$\begin{aligned} R_i &\geqslant \mathbb{E}_{z_{-i}} \mathbb{1}_{\mathcal{E}_1} \mathbb{E}_{A_{-i:}|z_{-i}} \mathbb{1}_{\mathcal{E}_2} e^{-\Theta(n_1 n_2 p^2)} \\ &- \quad \mathbb{E}_{z_{-i}} \mathbb{1}_{\mathcal{E}_1} \mathbb{E}_{A_{-i:}|z_{-i}} \mathbb{1}_{\mathcal{E}_2} \tilde{\mathbb{P}}_{A_{i:}|A_{-i:},z}(\mathcal{E}_3^c) \\ &\geqslant e^{-\Theta(n_1 n_2 p^2)} \mathbb{E}_{z_{-i}}(\mathbb{1}_{\mathcal{E}_1} \mathbb{P}_{A_{-i:}|z_{-i}}(\mathcal{E}_2)) \\ &\quad - \mathbb{E}_{z_{-i}} \mathbb{1}_{\mathcal{E}_1} \mathbb{E}_{A_{-i:}|z_{-i}} \mathbb{1}_{\mathcal{E}_2} \tilde{\mathbb{P}}_{A_{i:}|A_{-i:},z}(\mathcal{E}_3^c). \end{aligned} \tag{5.6.6}$$

We can control these probabilities with the following lemma proved in the appendix, see Section 5.8.2.

Lemma 18. *We have*

1. $\mathbb{P}_{Z_{-i}}(\mathcal{E}_1) \geqslant 1 - n_1^{-\Omega(1)}$;
2. *for all realizations* Z_{-i}, $\mathbb{P}_{A_{-i:}|Z_{-i}}(\mathcal{E}_2) \geqslant 1 - e^{-\Omega(n_1 n_2 p)}$;
3. *for all* $Z_{-i} \in \mathcal{E}_1$ *and* $A_{-i:} \in \mathcal{E}_2$, $\tilde{\mathbb{P}}_{A_{i:}|A_{-i:},Z}(\mathcal{E}_3) \geqslant 1 - e^{-\Omega(n_2 p)}$.

Lemma 18 and (5.6.6) directly imply $R_i \gtrsim e^{-\Theta(n_1 n_2 p^2)}$ since

$$e^{-\Theta(n_2 p)} = o(e^{-\Theta(n_1 n_2 p^2)})$$

under the assumption on p.

When $n_1 n_2 p_{max}^2 = \Theta(1)$, we can use the same proof except that now the final two hypothesis testing problem we reduced to has a non vanishing risk (see e.g. Zhang et al., 2016a). □

5.7 NUMERICAL EXPERIMENTS

In this section, we empirically compare the performance of our algorithm (GPM) with the spectral algorithm (Spec) and the algorithm introduced by Ndaoud et al., 2022 (referred to as HL).

CASE $K = L = 2$. In this setting, we generate a SBiSBM with parameters $n_1 = 500$, $n_2 = \lceil C n_1 \log n_1 \rceil$ (where $C \geqslant 1$ is a constant), $p \in (0,1)$, and $q = cp$ where $c > 0$ is a constant. The accuracy of the clustering is measured by the NMI; it is equal to one when the partitions match exactly and is zero when they are independent. The results are averaged over 20 Monte-Carlo runs. For the experiment presented in Figure 16, we fixed $C = 10$; for the experiment in Figure 17, we fixed $C = 3$ and $c = 0.5$. For the experiment presented in Figure 18, we fixed $p = 0.01$ and $c = 0.5$. We observe that HL and GPM have similar performance in all the aforementioned experiments. Thus, there is no gain in using the specialized method HL instead of the general algorithm GPM. The spectral method Spechas only slightly worst performance than the iterative methods HL and GPM. In particular, when approaching the threshold for exact recovery, the performance gap disappears. This suggests that Spec also reaches the threshold for exact recovery. It would be interesting to obtain stronger theoretical guarantees to explain the good performance of Spec.

CASE $K = L \neq 2$. We fix $n_1 = 1000, n_2 = 10000, p = 0.05$ and $c = 0.5$ and vary K from 2 to 10. As can be seen from Figure 19, the performance of Spec decreases faster than GPM when K increases.

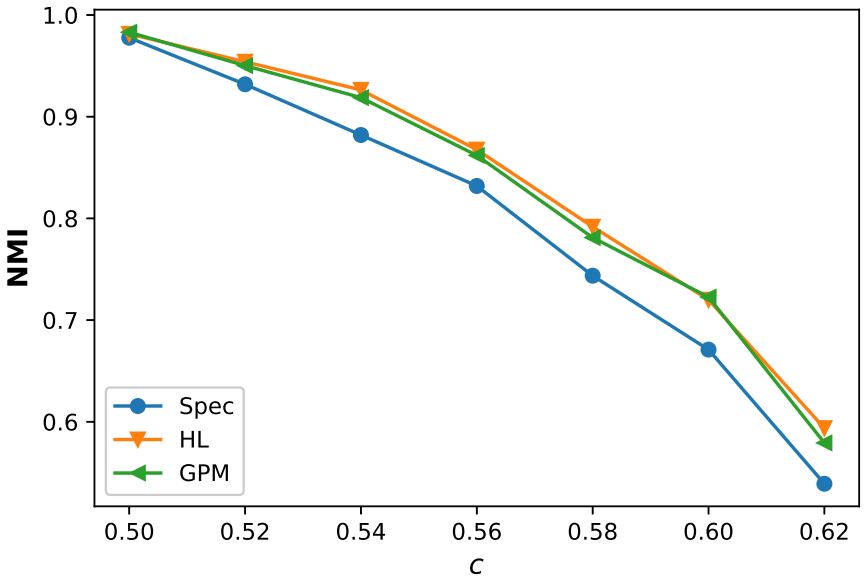

Figure 16: Relative performance of `Spec`, `HL` and `GPM` for varying c.

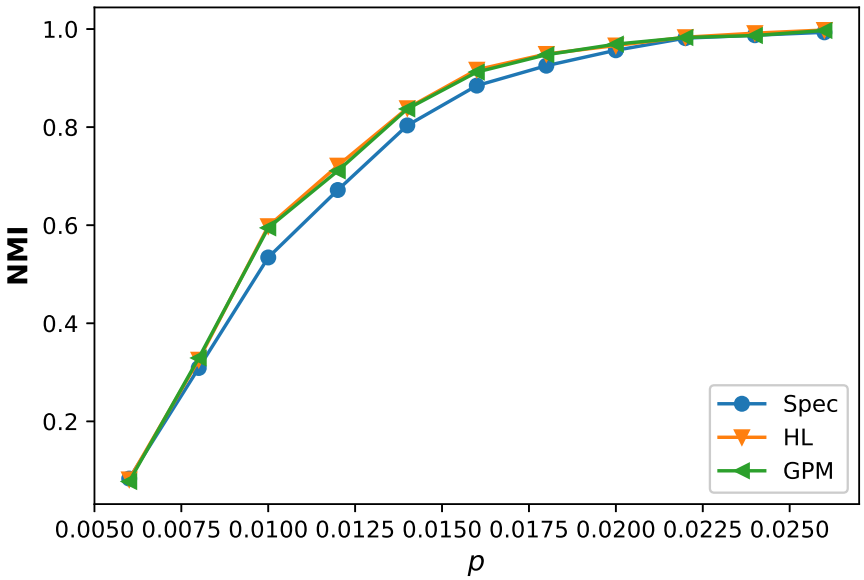

Figure 17: Relative performance of `Spec`, `HL` and `GPM` for varying p.

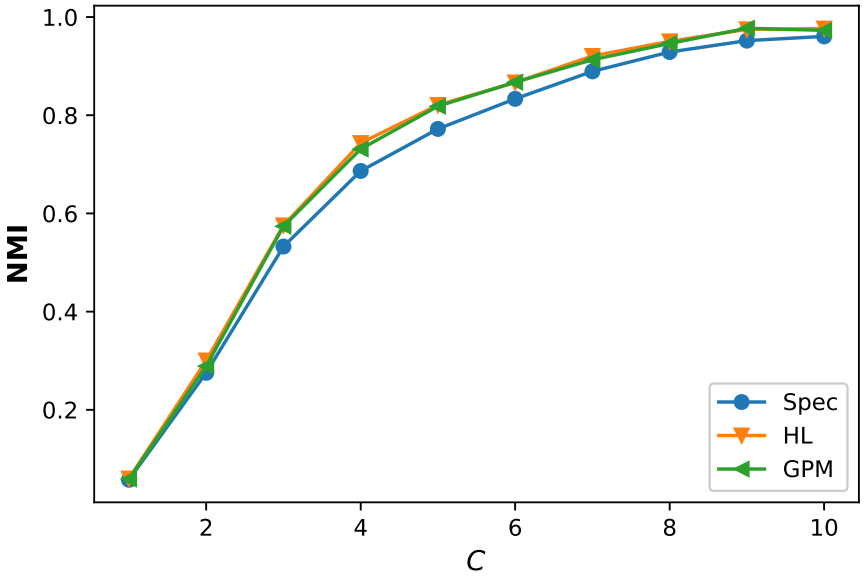

Figure 18: Relative performance of `Spec`, `HL` and `GPM` for varying C.

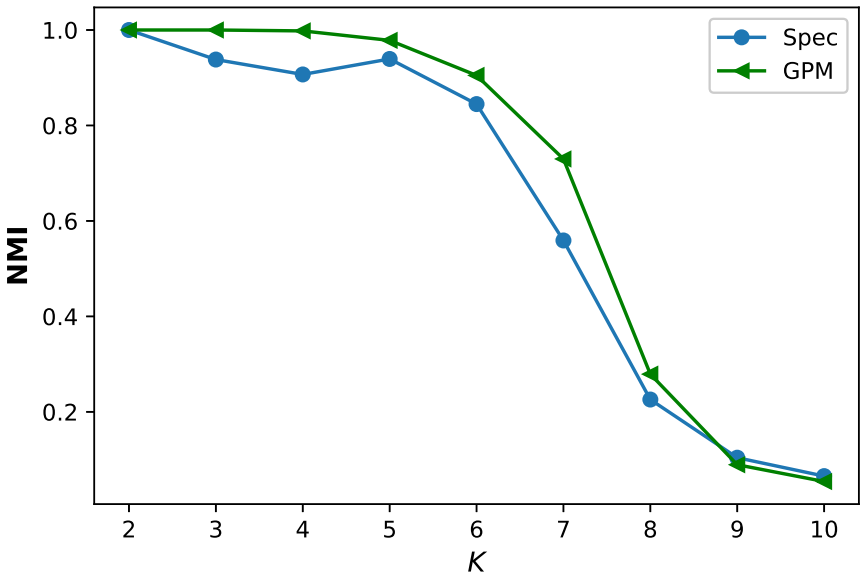

Figure 19: Relative performance of `Spec`, `HL` and `GPM` for varying K.

5.8 ADDITIONAL PROOFS

5.8.1 *Concentration of the oracle error*

Lemma 19. *Assume that the assumptions of Theorem 12 hold. Recall that*

$$\Omega_{i,k}(\delta) = \left\{\tilde{E}_{i:}(W_{:z_i} - W_{:k}) \leqslant -(1-\delta)\Delta^2(z_i,k)\right\}.$$

We have for all $0 < \delta \leqslant \frac{1}{4\eta\alpha}$, $k \neq z_i \in [K]$ *and* $i \in [n_1]$

$$\mathbb{P}(\Omega_{i,k}(\delta)) \leqslant e^{-\frac{\beta^2}{12eL\alpha^3}\frac{n_1 n_2 p_{max}^2}{KL}} = e^{-\tilde{\Delta}^2}.$$

Proof. The event $\Omega_{i,k}(\delta)$ holds if and only if

$$B_{i:}(W_{:k} - W_{:z_i}) \geqslant -\delta\Delta^2(z_i,k).$$

But we can decompose this quantity as

$$\begin{aligned} B_{i:}(W_{:k} - W_{:z_i}) &= \sum_{j\neq i} B_{ij}(W_{jk} - W_{jz_i}) \\ &= |\mathcal{C}_k|^{-1} \sum_{j\in\mathcal{C}_k\setminus\{i\}} \langle A_{i:}, A_{j:}\rangle - |\mathcal{C}_{z_i}|^{-1} \sum_{j\in\mathcal{C}_{z_i}\setminus\{i\}} \langle A_{i:}, A_{j:}\rangle \\ &= \langle A_{i:}, \tilde{A}_k - \tilde{A}_{z_i}\rangle \end{aligned}$$

where $\tilde{A}_k = \frac{1}{|\mathcal{C}_k|}\sum_{j\in\mathcal{C}_k} A_{j:}$ and $\tilde{A}_{z_i} = \frac{1}{|\mathcal{C}_{z_i}|}\sum_{j\in\mathcal{C}_{z_i}\setminus\{i\}} A_{j:}$ are independent random variables. Since the index i doesn't appear in the second sum, $\tilde{A}_{z_i}$ is independent from $A_{i:}$. By definition the entries of $\tilde{A}_k$ and $\tilde{A}_{z_i}$ are independent normalized binomial random variables whose parameters vary depending on the community associated with the entry.

We are now going to bound the moment generating function of $M := \langle A_{i:}, \tilde{A}_k - \tilde{A}_{z_i}\rangle$, conditionally on $A_{i:}$. Observe that M is a sum of independent random variables. Recall that if $X \sim \mathcal{B}(p)$ then $\mathbb{E}(e^{tX}) = (e^t p + 1 - p)$. Hence, for $t > 0$, conditionally on A_{ij}, each summand has a m.g.f equal to

$$\begin{aligned} &\log \mathbb{E}(e^{tA_{ij}(\tilde{A}_{kj} - \tilde{A}_{z_i j})}|A_{ij}) = \\ &|\mathcal{C}_k|\log(e^{\frac{tA_{ij}}{|\mathcal{C}_k|}}\Pi_{kz_j'} + 1 - \Pi_{kz_j'}) + (|\mathcal{C}_{z_i}| - 1)\log(e^{-\frac{tA_{ij}}{|\mathcal{C}_{z_i}|}}\Pi_{z_i z_j'} + 1 - \Pi_{z_i z_j'}) \\ &\leqslant |\mathcal{C}_k|\Pi_{kz_j'}(e^{\frac{tA_{ij}}{|\mathcal{C}_k|}} - 1) + (|\mathcal{C}_{z_i}| - 1)\Pi_{z_i z_j'}(e^{-\frac{tA_{ij}}{|\mathcal{C}_{z_i}|}} - 1) \end{aligned}$$

by using the fact that $\log(1+x) \leqslant x$ for all $x > -1$.

Fix $t = t^* = \epsilon\frac{n_1}{\alpha K}$ for a parameter $\epsilon \in (0,1)$ that will be fixed later. We have by Taylor Lagrange formula for all $t \leqslant \frac{n_1}{\alpha K}$

$$e^{\frac{t}{|\mathcal{C}_k|}} - 1 \leqslant \frac{t}{|\mathcal{C}_k|} + \frac{e}{2}\left(\frac{t}{|\mathcal{C}_k|}\right)^2$$

and

$$e^{\frac{-t}{|\mathcal{C}_{z_i}|}} - 1 \leqslant -\frac{t}{|\mathcal{C}_{z_i}|} + \frac{e}{2}\left(\frac{t}{|\mathcal{C}_{z_i}|}\right)^2.$$

By using these upper bounds we get

$$\log \mathbb{E}(e^{t^* A_{ij}(\tilde{A}_{kl} - \tilde{A}_{z_i j})} | A_{ij}) \leqslant \left(\Pi_{k z'_j} - \frac{|\mathcal{C}_{z_i}| - 1}{|\mathcal{C}_{z_i}|}\Pi_{z_i z'_j}\right) t^* A_{ij} + \frac{e}{2}(t^*)^2 A_{ij}\left(\frac{\Pi_{k z'_j}}{|\mathcal{C}_k|} + \frac{\Pi_{z_i z'_j}}{|\mathcal{C}_{z_i}|}\right).$$

Hence by using independence we obtain

$$\log \mathbb{E}(e^{t^* M} | A_{i:}) \leqslant \sum_{j \in [n_2]} \left[\underbrace{\left(\Pi_{k z'_j} - \frac{|\mathcal{C}_{z_i}| - 1}{|\mathcal{C}_{z_i}|}\Pi_{z_i z'_j}\right) t^* + \frac{e}{2}(t^*)^2\left(\frac{\Pi_{k z'_j}}{|\mathcal{C}_k|} + \frac{\Pi_{z_i z'_j}}{|\mathcal{C}_{z_i}|}\right)}_{t_{ij}}\right] A_{ij}.$$

Using Markov inequality and the fact that $\Delta^2(z_i, k) \leqslant \eta\beta\alpha n_2 p_{max}^2/L$ by (5.5.3) leads to

$$\begin{aligned}\mathbb{P}(\Omega_{i,k}(\delta) | A_{i:}) &\leqslant e^{\delta\epsilon\frac{n_1}{\alpha K}\Delta^2(z_i,k)}\mathbb{E}(e^{t^* M} | A_{i:}) \\ &\leqslant e^{\delta\epsilon\frac{\beta n_1 n_2 p_{max}^2}{LK}} \prod_j \mathbb{E}(e^{t_{ij} A_{ij}}).\end{aligned}$$

But since A_{ij} is a Bernoulli random variable with parameter $\Pi_{z_i z_{2j}}$ we have

$$\mathbb{E}(e^{t_{ij} A_{ij}}) = (e^{t_{ij}} - 1)\Pi_{z_i z_{2j}} + 1.$$

By our choice of t^*, $t_{ij} = O(n_1 p_{max}) = o(1)$ so that

$$\prod_j \left((e^{t_{ij}} - 1)\Pi_{z_i z_{2j}} + 1\right) \leqslant e^{\sum_j (e^{t_{ij}} - 1)\Pi_{z_i z_{2j}}}.$$

Here we use the fact that for $x_1, \ldots, x_n > -1$ we have $\prod_{i \in [n]}(1 + x_i) \leqslant e^{\sum_{i \in [n]} x_i}$. But again, by using Taylor Lagrange formula, we have $e^{t_{ij}} - 1 = t_{ij} + O(t_{ij}^2)$. Consequently

$$\prod_j \mathbb{E}(e^{t_{ij} A_{ij}}) \leqslant e^{\sum_j (t_{ij} + o(t_{ij}))\Pi_{z_i z_{2j}}}.$$

We can write $\sum_j t_{ij}\Pi_{z_i z_{2j}}$ as $t^* A_1 + \frac{e(t^*)^2}{2} A_2$ where

$$A_1 = \sum_{l \in [L]} |\mathcal{C}'_l|\left(\Pi_{z_i l}\Pi_{kl} - \Pi_{z_i l}^2\left(\frac{|\mathcal{C}_{z_i}| - 1}{|\mathcal{C}_{z_i}|}\right)\right)$$

and

$$A_2 = \sum_{l \in [L]} |\mathcal{C}'_l|\left(\frac{\Pi_{z_i l}\Pi_{kl}}{|\mathcal{C}_k|} + \Pi_{z_i l}^2\left(\frac{|\mathcal{C}_{z_i}| - 1}{|\mathcal{C}_{z_i}|^2}\right)\right).$$

By Assumption A3 and A5 we have

$$-A_1 \geqslant \frac{n_2}{\alpha L}\left(\beta p_{max}^2 + o(p_{max}^2)\right)$$

and

$$A_2 \leqslant \frac{\alpha K}{n_1} n_2 p_{max}^2$$

Let $\epsilon = \frac{\beta}{3eL\alpha}$ (recall that $t^* = \epsilon \frac{n_1}{\alpha K}$). By this choice of ϵ we have

$$\frac{e\epsilon n_1}{2\alpha K} A_2 \leqslant \frac{|A_1|}{2}.$$

Consequently we have for all $\delta \leqslant 1/(4\eta\alpha)$

$$\begin{aligned}
\mathbb{P}(\Omega_{i,k}(\delta)) &\leqslant e^{-\frac{\epsilon\beta}{2\alpha^2}\frac{n_1 n_2 p_{max}^2}{KL} + \delta\epsilon\frac{n_1}{\alpha K}\Delta^2(z_i,k)} \\
&\leqslant e^{-\frac{\epsilon\beta}{4\alpha^2}\frac{n_1 n_2 p_{max}^2}{KL}} \\
&\leqslant e^{-\frac{\beta^2}{12eL\alpha^3}\frac{n_1 n_2 p_{max}^2}{KL}}.
\end{aligned}$$

□

5.8.2 *Proof of Lemma 18*

The first inequality is a direct consequence of Hoeffding concentration inequality. For the second inequality, observe that

$$\mathbb{P}_{A_{-i:}|z_{-i}}(\mathcal{E}_2) = \mathbb{E}_{z'}\mathbb{P}_{A_{-i:}|z_{-i},z'}(\mathcal{E}_2)$$

and

$$\begin{aligned}
\mathbb{P}_{A_{-i:}|z_{-i},z'}\left(\sum_{i'\in\mathcal{C}_+} A_{i'j} - \sum_{i'\in\mathcal{C}_-} A_{i'j} = 0\right) &\geqslant \mathbb{P}(A_{i'j} = 0 \text{ for all } i') \\
&\gtrsim (1-p)^{n_1} = 1 - n_1 p.
\end{aligned}$$

To obtain an upper bound on this probability, we can use Paley-Zigmund inequality as follows. Let us denote $Z = |\sum_{i'\in\mathcal{C}_+} A_{i'j} - \sum_{i'\in\mathcal{C}_-} A_{i'j}|$. We then have

$$\begin{aligned}
\mathbb{P}_{A_{-i:}|z_{-i},z'}(Z = 0) &= 1 - \mathbb{P}_{A_{-i:}|z_{-i},z'}(Z > 0) \\
&\leqslant 1 - \mathbb{P}_{A_{-i:}|z_{-i},z'}(Z > \theta\mathbb{E}Z) \\
&\leqslant 1 - (1-\theta)^2 \frac{(\mathbb{E}_{A_{-i:}|z_{-i},z'}(Z))^2}{\mathbb{E}_{A_{-i:}|z_{-i},z'}(Z^2)}
\end{aligned}$$

for $\theta \in (0,1)$ by Paley-Zigmund inequality. It is easy to check that

$$\mathbb{E}_{A_{-i:}|z_{-i},z'}Z \gtrsim n_1(p-q) \text{ and } \mathbb{E}_{A_{-i:}|z_{-i},z'}(Z^2) \asymp n_1 p.$$

Consequently, $\mathbb{P}_{A_{-i:}|z_{-i},z'}(Z = 0) = 1 - \Theta(n_1 p)$. By using independence over j and Chernoff's multiplicative bound we obtain that

$$\mathbb{P}_{A_{-i:}|z_{-i},z'}\left(\sum_j \mathbb{1}_{\{\sum_{i' \in \mathcal{C}_+} A_{i'j} - \sum_{i' \in \mathcal{C}_-} A_{i'j} \neq 0\}} = \Theta(n_2 n_1 p)\right) \geqslant 1 - e^{-cn_2 n_1 p}$$

for any realization of z_{-i} and z'. Hence the stated bound follows.

Finally it remains to control the probability of $\mathcal{E}_3$. Under $\tilde{\mathbb{P}}_{A_{i:}|A_{-i:},z}$, for each $j \in J_b$, A_{ij} are independent and distributed as a mixture of Bernoulli of parameters p and q. The size of the set J_b is also controlled by the assumption $A_{-i:} \in \mathcal{E}_2$ (see equation (5.6.5)). Since under $\tilde{\mathbb{P}}_{A_{i:}|A_{-i:},z}$ for all $j \in J_b$, A_{ij} are independent r.v. with density g, we obtain by stochastic domination (we can replace a mixture of Bernoulli with parameters p and q by a Bernoulli of parameter p) and Chernoff's bound that $\tilde{\mathbb{P}}_{A_{i:}|A_{-i:},z}(\mathcal{E}_3) \geqslant 1 - e^{-\Theta(n_2 p)}$.

Part II

GRAPH MATCHING

6 INTRODUCTION TO GRAPH MATCHING

The objective of the *graph matching problem* is to find a bijective correspondence between the vertices of two graphs G and H such that the alignment between the edges of G and H is maximized (see Section 1.5). This problem appears in many applications such as computer vision (H. Sun et al., 2020), network de-anonymization (Narayanan et al., 2009), pattern recognition (Conte et al., 2004; Emmert-Streib et al., 2016), protein-protein interactions and computational biology (Zaslavskiy et al., 2009b; Singh et al., 2009). In computer vision, for example, it is used as a method of comparing two objects (or images) encoded as graph structures or to identify the correspondence between the points of two discretized images of the same object at different times. In network de-anonymization, the goal is to learn information about an anonymized (unlabeled) graph using a related labeled graph as a reference, exploiting their structural similarities. For example, Narayanan et al., 2006 show that it was possible to effectively de-anonymize the Netflix database using the IMDb (Internet Movie Database) as the "reference" network. It also contains the ubiquitous *graph isomorphism* as a special case.

In general, the graph matching problem is NP-hard. But under generative models such as the CoERM and CoWM (see Section 1.5 for a precise description) there exists efficient algorithms. Recently, other models of correlation have been proposed for random graphs with a latent geometric structure (Kunisky et al., 2022; Wang et al., 2022a), community structure (Racz et al., 2021) and with power law degree profile (Yu et al., 2021b).

In this thesis, we will focus on the Correlated Erdös-Renyi Model (CoERM) and the Correlated Wigner Model (CoWM). In Section 6.1, we will present the information-theoretic limits of the graph matching problem under these generative models. Then, in Section 6.2 we will discuss the principal efficient methods that have been proposed in the literature.

6.1 INFORMATION THEORETIC LIMITS OF MATCHING

The accuracy of an estimated permutation $\hat{\pi}$ to the ground-truth permutation π^* is usually measured by the fraction of well-matched nodes

$$\operatorname{overlap}(\pi, \pi^*) := \frac{|\{i \in [n] : \pi(i) = \pi^*(i)\}|}{n}. \tag{6.1.1}$$

Similarly to the clustering problem, one can have different recovery requirements.

- **Exact recovery:** $\text{overlap}(\hat{x}, x^*) = 1$ with probability $1 - o(1)$ when $n \to \infty$, i.e. one can recover exactly the latent permutation w.h.p.
- **Almost exact recovery:** $\text{overlap}(\hat{x}, x^*) = 1 - o(1)$ with probability $1 - o(1)$ when $n \to \infty$, i.e. the proportion of correctly matched nodes goes to zero when $n \to \infty$.[1]
- **Partial recovery:** $\text{overlap}(\hat{x}, x^*) \geqslant \alpha$ with probability $1 - o(1)$ when $n \to \infty$, i.e. one can correctly classify at least a proportion α of the nodes w.h.p.

6.1.1 *Correlated Wigner Model (CoWM)*

In the Gaussian setting, a sharp phase transition phenomenon occurs. There is a threshold involving n and the noise level σ above which exact recovery of x^* is possible and below which even partial recovery is impossible. Recall that the MLE is given by

$$\hat{x}_{\text{MLE}} \in \arg\max_{x \in \mathfrak{S}_n} \sum_{i,j} A_{ij} B_{x(i)x(j)}.$$

Theorem 14 (Wu et al., 2021)**.** *Assume that* $(A, B) \sim \text{CoWM}(n, \sigma, x^*)$ *and let us define* $\rho = \sqrt{1 - \sigma^2}$*. If for some constant* $\epsilon > 0$

$$\rho^2 \geqslant \frac{(4 + \epsilon) \log n}{n},$$

then $\hat{x}_{\text{MLE}} = x^*$ *w.h.p. Conversely, if for some constant* $\epsilon > 0$

$$\rho^2 \leqslant \frac{(4 - \epsilon) \log n}{n},$$

then for any estimator $\hat{x}$ *and fixed constant* $\delta > 0$, $\mathbb{P}(\text{overlap}(\hat{x}, x^*) \leqslant \delta) = 1 - o(1)$.

6.1.2 *Correlated Erdös-Renyi Model (CoERM)*

In the dense regime, the recovery of x^* exhibits a sharp phase transition similar to the Correlated Wigner model: there is no partial recovery regime.

Theorem 15 (Wu et al., 2021)**.** *Assume that* $(A, B) \sim \text{CoERM}(n, q, s, x^*)$ *where* q *is bounded away from* 1 *and* $q = n^{-o(1)}$*. If for some* $\epsilon > 0$

$$nqs^2 \geqslant \frac{(2 + \epsilon) \log n}{\log \frac{1}{q} - 1 + q},$$

1 There is an alternative definition where we only require to be able to correctly match an arbitrary large fraction of nodes. See Wu et al., 2021. Here we preferred to use an analogous definition to the one used for clustering.

then for any constant $\delta < 1$, $\mathbb{P}(\text{overlap}(\hat{x}_{MLE}, x^*) \geqslant \delta)) = 1 - o(1)$. *Conversely, if for some* $\epsilon > 0$

$$nqs^2 \leqslant \frac{(2-\epsilon)\log n}{\log \frac{1}{q} - 1 + q},$$

then for any estimator $\hat{x}$ *and any constant* $\delta > 0$, $\mathbb{P}(\text{overlap}(\hat{x}_{MLE}, x^*) \leqslant \delta) = 1 - o(1)$.

The threshold for exact-recovery in the Erdös-Renyi model when q is bounded away from 1 is given by

$$\frac{nqs^2(1-\sqrt{q})^2}{\log n} = 1$$

as shown in Theorem 4 in Wu et al., 2021.

However, in the sparse regime, there is an interesting regime where partial recovery is possible but almost exact recovery impossible.

Theorem 16 (Wu et al., 2021)**.** *Assume that* $(A, B) \sim \text{CoERM}(n, q, s, x^*)$ *where* $q = n^{-\Omega(1)}$. *If for some constant* $\epsilon > 0$

$$nqs^2 \geqslant (2+\epsilon)\max(\frac{\log n}{\log(1/q)}, 2),$$

then there exists a constant $\delta > 0$ *such that* $\mathbb{P}(\text{overlap}(\hat{x}_{MLE}, x) \geqslant \delta) = 1 - o(1)$. *Conversely, assuming* $nq = \omega(\log^2 n)$, *if for some* $\epsilon > 0$

$$nqs^2 \leqslant 1 - \epsilon,$$

then for any estimator $\hat{x}$ *and any constant* $\delta > 0$, $\mathbb{P}(\text{overlap}(\hat{x}, x) \leqslant \delta) = 1 - o(1)$.

Indeed, this theorem implies that the threshold for partial recovery is at $nqs^2 \asymp 1$, but it has been shown by Cullina et al., 2018 that one needs $nqs^2 \to \infty$ in order to get almost exact recovery. So there is a regime where partial recovery is possible but not almost exact recovery.

Remark 19. *While the statistical limits of graph matching under the* CoERM *or* CoWM *are well understood, it is not known if there are polynomial time algorithms that achieve the threshold for exact recovery. Usually the existing methods that come with consistency guarantees assume that the correlation level is asymptotically one. The only exception is the recent work of Mao et al., 2021a, that proposed a polynomial time algorithm that comes with consistency guarantees for a constant level of correlation* $s < 1$ *under the* CoERM.

6.2 MATCHING METHODS

6.2.1 *Convex Quadratic Programming (CQP) relaxation*

As saw in Chapter 1, Section 1.5, QAP is a NP-hard problem. A popular approach is to consider a relaxation of the set of permutations to a continuous domain, leading to a tractable optimization problem. There are several ways to do this, we recall some of them below. First observe that (1.5.2) can be rewritten in the following form

$$\min_{X\in\mathfrak{S}_n} \left\|A - XBX^\top\right\|_F^2 \Leftrightarrow \min_{X\in\mathfrak{S}_n} \|AX - XB\|_F^2 .$$

If we replace the set of permutations $\mathfrak{S}_n$ by the space of doubly stochastic matrices

$$\mathcal{P}_n = \{X \in \mathbb{R}^{n\times n} : X\mathbf{1}_n = \mathbf{1}_n, X^\top \mathbf{1}_n = \mathbf{1}_n, X_{ij} \geqslant 0\}$$

as proposed by Aflalo et al., 2015 we reduce the original problem to the following optimization problem

$$\arg\min_{X\in\mathcal{P}_n} \|AX - XB\|_F . \tag{6.2.1}$$

Since $\mathcal{P}_n$ is a convex set (it is the Birkhoff polytope and is the tightest convex relaxation possible), the previous optimization problem can be solved in polynomial time and leads to a solution $\tilde{X}$. Since in general $\tilde{X}$ is not a permutation, one needs to project $\tilde{X}$ on $\mathfrak{S}_n$. This can be done by solving the following LAP

$$\hat{X} = \arg\max_{X\in\mathfrak{S}_n} \mathrm{Tr}(X^\top \tilde{X}).$$

In practice, the method performs well. But it remains an open problem to derive consistency guarantees for this algorithm. Besides, the computational cost of this method can be prohibitive for large graphs.

6.2.2 *SDP relaxation*

The QAP can be rewritten in a vectorial form

$$\arg\max_{X\mathfrak{S}_n} \mathrm{Tr}((B\otimes A)vec(X)vec(X)^\top)$$

where $\otimes$ denotes the Kronecker product and $vec(X) \in \mathbb{R}^{n^2}$ is the column-wise version of X. Zhao et al., 1997 proposed to relax the constraints on X by replacing $\mathfrak{S}_n$ by $\mathcal{P}_n$ as before and replace $vec(X)vec(X)^\top$ by a matrix Q satisfying some constraints. More precisely, they consider the following optimization problem.

$$\max_{Q,P\in\mathbb{R}^{n\times n}} \mathrm{Tr}((B\otimes A)Q) \text{ s.t.} \qquad P \in \mathcal{P}_n$$

$$
\begin{aligned}
&Q - \nu ec(P)\nu ec(P)^\top \succeq 0,\\
&\sum_i Q(i,i) = I_n,\\
&\mathrm{Tr}(Q(i,j)) = 0 \text{ for } i \neq j,\\
&\mathrm{Tr}(Q(i,j)\mathbf{1}_n\mathbf{1}_n^\top) = 1 \text{ for all } i,j,\\
&\text{all entries of } Q \text{ are non-negative},\\
&(Q(i,i))_{jj} = P_{ji},
\end{aligned}
$$

where $Q(i,j) \in \mathbb{R}^{n\times n}$ is the block (i,j) of size $n \times n$ that constitutes Q. See also S. Bravo Ferreira et al., 2018 for other relaxations.

From a computational point of view, this approach is feasible since it requires solving an SDP problem in a space of dimension n^2.

6.2.3 *Spectral methods*

The spectral method GRAMPA introduced by Fan et al., 2019 can be related view as a solution of a relaxed and regularized version of the QAP for a given hyperparameter $\eta > 0$

$$\min_{X\in\mathbb{R}^{n\times n}} \|AX - XB\|_F^2 + \frac{\eta^2}{2}\|X\|_F^2 - \mathbf{1}_n^\top X\mathbf{1}_n. \tag{6.2.2}$$

Since it is a convex quadratic optimization problem, we can write the solution $\hat{X}$ in closed form. Write the spectral decomposition of A and B as

$$A = \sum_{i\in[n]} \lambda_i u_i u_i^\top \quad \text{and} \quad B = \sum_{j\in[n]} \mu_j v_j v_j^\top$$

where the eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_n$ of A (resp. $\mu_1 \geqslant \ldots \geqslant \mu_n$ of B) are ordered in decreasing order. Then, as shown in Corollary 2.2 in Fan et al., 2019,

$$\hat{X} = \sum_{i,j\in[n]} w(\lambda_i,\mu_j) u_i u_i^\top \mathbf{1}_n\mathbf{1}_n^\top v_j v_j^\top$$

where $w(x,y)$ is the Cauchy Kernel of bandwidth η defined by

$$w(x,y) = \frac{1}{(x-y)^2+\eta^2}.$$

It can be shown that under the correlated Wigner model $W(n,\sigma,x^*)$, with a noise level σ of order at most $O(\frac{1}{\log n})$, GRAMPA exactly recovers the ground-truth permutation x^* for a suitable choice of η (see Theorem 2.1 in Fan et al., 2019). In practice, GRAMPA has good performance, but its computational cost is $O(n^3)$. A faster (but less robust to noise) spectral method has also been proposed by Ganassali et al., 2022a.

Another popular approach is Umeyama's algorithm (Umeyama, 1988). Here we construct a similarity matrix $\hat{X}$ based on the absolute value of the eigenvectors of A and B. More precisely

$$\hat{X} = \sum_{i=1}^n |u_i||v_i|^\top$$

where $|v|$ denotes the entrywise absolute value of vector v.

6.2.4 *Message passing algorithm*

Let $(A, B) \sim \text{CoERM}(n, q, s, x^*)$. Recently Ganassali et al., 2022b and independently Piccioli et al., 2022 proposed a matching method based on a message passing algorithm for Correlated ER model that are sparse enough, i.e. when the sparsity q satisfies $q = \frac{\lambda}{n}$ for some constant λ. The idea, as explained in Piccioli et al., 2022, is to take advantage of the local tree-like structure of the extended neighborhood of each node in order to approximate the posterior distribution $\mathbb{P}(x^*(i) = i'|A, B)$ for each i, i' (where x^* is generated uniformly at random). If instead of conditioning on A and B we only condition on $A_{i,d}$ (resp. $B_{i',d}$) the set of nodes at distance d from i (resp i') in A (resp. B) one can show that the posterior distribution $\mathbb{P}(x^*(i) = i'|A_{i,d}, B_{i',d})$ can be written in close form when $n \to \infty$ and d is chosen such that the neighborhoods of i and i' at distance d are trees. More precisely for any trees T, T' of length d

$$\lim_{n\to\infty} \mathbb{P}(x^*(i) = i'|A_{i,d} = T, B_{i',d} = T') = n\frac{\mathbb{P}_1^{(d)}(T, T')}{\mathbb{P}_0^{(d)}(T)\mathbb{P}_0^{(d)}(T')} \quad (6.2.3)$$

where $\mathbb{P}_0^{(d)}(T)$ is the distribution of a Galton-Watson (GW) tree (of length d) with offspring given by a Poisson law with parameter λ and $\mathbb{P}_1^{(d)}(T, T')$ is the distribution associated to a colored GW tree where edges can be red, blue or bicolored. The offsprings associated with red edges (resp. blue and bicolored) are drawn by independent Poisson law with parameter $\lambda(1-s)$ (resp. $\lambda(1-s)$ and λs).

By (6.2.3), determining if the posterior distribution

$$\mathbb{P}(x^*(i) = i'|A_{i,d} = T, B_{i',d} = T')$$

is superior or inferior to $1/2$ is equivalent to testing between the distributions $\mathbb{P}_1^{(d)}(T, T')$ (T and T' are generated in a correlated way) and $\mathbb{P}_0^{(d)}(T)\mathbb{P}_0^{(d)}(T')$ (T and T' are generated independently). One can write the likelihood ratio

$$L^{(d)}(T, T') = \frac{\mathbb{P}_1^{(d)}(T, T')}{\mathbb{P}_0^{(d)}(T)\mathbb{P}_0^{(d)}(T')}$$

in a recursive form so that it can be computed in polynomial time. But in practice, the algorithm has limited applications because due to its complexity that is exponential in λ. From a theoretical point of view, Ganassali et al., 2022b have shown that a one-sided test between the two hypotheses can succeed when the Kullback-Leibler (KL) divergence between the two probabilities goes to infinity with d and $\lambda s > 1$.

6.2.5 *Signature matching*

In their groundbreaking work, Mao et al., 2021a proposed a matching method based on signatures which are statistics associated with each node that captures some of their particularities. The basic idea shares some similarities with the message passing approach: locally most of the nodes' neighborhood is a tree. Hence, in order to match nodes, it would be sufficient to match trees. Unfortunately, since these trees can have at each level number of nodes, it is not possible to compute all the possible permutations to find the one that best matches the tree. The idea is to construct for node i a partition tree rooted at i and use the information provided by the degrees of the neighbors to label the tree.

More formally, let $(A, B) \sim \mathrm{CoERM}(n, q, s, x^*)$ and G, G' the associated graphs. Given a node i and a depth m we iteratively construct sets T_s^k for $k \in [m]$ and $s \in \{-1, 1\}^k$ as follows. We start with the root $T_\emptyset^0 = \{i\}$. Assume that all T_s^k have been constructed. Let us denote by $\mathcal{N}(T)$ the set of neighbors in G of all the nodes belonging in the set T and $\mathcal{S}(i, l)$ the sets of nodes that are exactly at distance l from i in G. Then we can define the two children $T_{(s,-1)}^{k+1}$ and $T_{(s,+1)}^{k+1}$ of T_k^s by

$$T_{(s,-1)}^{k+1} = \left\{ j \in \mathcal{N}(T_s^k) \cap \mathcal{S}(i, k+1) \,:\, \deg(j) < np \right\}$$

and

$$T_{(s,+1)}^{k+1} = \left\{ j \in \mathcal{N}(T_s^k) \cap \mathcal{S}(i, k+1) \,:\, \deg(j) \geqslant np \right\}.$$

The collection of sets $(T_s^k)_{s,k}$ encodes information about the path of length m starting from i. These paths are labeled with s depending on the degrees of the nodes on the path. Based on these paths we can associate to each node i a signature $(f(i), v(i)) \in \mathbb{R}^{2^m} \times \mathbb{R}^{2^m}$ defined by $f(i) = (f_s(i))_s$ and $v(i) = (v_s(i))_s$ with

$$f_s(i) = \sum_{j \in \mathcal{N}(T_s^m) \cap \mathcal{S}(i, m+1)} (\deg(j) - 1 - np)$$

and

$$v_s(i) = np(1-p)|\mathcal{N}(T_s^m) \cap \mathcal{S}(i, m+1)|.$$

Then two nodes $i \in G$ and $i' \in G'$ are matched if

$$T := \frac{1}{|J|} \sum_{s \in J} \frac{(f_s(i) - f'_s(i'))^2}{v_s(i) + v'_s(i')} < \left(1 - \frac{1}{\sqrt{\log n}}\right)$$

where J is a subset of $\{-1, 1\}^m$ chosen uniformly at random and $(f'(i'), v'(i'))$ is the signature associated to the node i' in G'. The intuition is as follows. When there is no noise, $s = 0$, the graphs G and G' are the same and consequently if $i' = x^*(i)$ the signature of i and i' should be the same. On the other hand, if $i' \neq x^*(i)$ the signatures are likely to be different and hence the variance measured by the statistic T should be larger than in the case where $i' = x^*(i)$.

It has been shown by Mao et al., 2021a that even when the noise s is of constant order, the previously described method works. This is the first theoretical result in graph matching that allows a constant order noise, all the previous results assumed that $s = o(1)$. The computational cost of the algorithm is $O(n^2 \log n^c)$ for some constant $c > 0$. So it is faster than GRAMPA.

7

SEEDED GRAPH MATCHING WITH THE GENERALIZED POWER METHOD

This chapter is based on the article Araya et al., 2022.

7.1 INTRODUCTION

In this chapter, we will focus on the setting where side information is provided under the form of a noisy seed, which is a partial matching between the graphs. The seeded version of the problem is motivated by the fact that in many applications, a set of correctly matched vertices can be available as prior information. Several algorithms, based on different techniques, have been proposed for seeded graph matching. In Pedarsani et al., 2011; Yartseva et al., 2013, the authors use a percolation based method to "grow" the seed to recover (at least partially) the ground-truth matching. Other algorithms (Lubars et al., 2018; Yu et al., 2021a) construct a similarity matrix between the vertices of both graphs and then solve the maximum linear assignment problem (either optimally or by a greedy approach) using the similarity matrix as the cost matrix. The latter strategy has also been successfully applied in the case described below, when no side information is provided.

In this chapter we analyse the performance of the *projected power method* (PPM) for the seeded graph matching problem in the context of the correlated Wigner model. This family of iterative algorithms has recently been successfully applied to several problems in machine learning and statistics (Chen et al., 2018; Boumal, 2016; Wang et al., 2021), in clustering in particular (see Chapter 3). We prove that PPM can exactly recover the ground-truth permutation provided that a sufficiently good initial permutation is provided. Our analysis extends the analysis of the refinement algorithm Mao et al., 2021a, Alg.4 to the case of (dense) Wigner graphs and represents, to best of our knowledge, the first analysis of PPM in the dense regime. The main technical difficulty in proving the convergence of PPM lies in proving that each step of the algorithm is a contraction, which requires establishing a uniform bound for the error in a neighborhood of the ground truth. As a byproduct of our analysis, we see that PPM provides a general framework which generalizes some of the state-of-the-art algorithms in the seeded case, such as Yu et al., 2021a, Alg.1, Lubars et al., 2018, Alg.2 and Mao et al., 2021a, Alg.4.

CONTRIBUTIONS. The main contributions of this work can be summarized as follows.

- We provide (see Theorems 17, 18) exact and partial recovery guarantees under the Gaussian Wigner model when the PPM is initialized with a given data-independent seed, and only one iteration of the PPM algorithm is performed. For this result to hold, it suffices that the overlap of the seed with the ground-truth permutation is $\Omega(\sqrt{n \log n})$. This matches the best-known bound for the sparse Erdös-Renyi case (Yu et al., 2021a), for which an overlap of $\Omega(\sqrt{n \log n})$ is required to obtain exact recovery.
- We prove (see Theorem 19) that when multiple iterations are allowed, then PPM converges to the ground-truth matching in $\mathcal{O}(\log n)$ iterations provided that it is initialized with a seed with overlap $\Omega\big((1-\kappa)n\big)$, for a constant κ small enough, even if the initialization is data-dependent or adversarial. This extends the results in (Mao et al., 2021a) from the sparse Erdös-Renyi setting, to the dense Wigner case.
- We complement our theoretical results with experiments on synthetic data showing that PPM can help to significantly improve the accuracy of the matching (for correlated Wigner model) compared to that obtained by a standalone application of existing seedless methods.

7.1.1 *Related work*

PROJECTED POWER METHOD (PPM). PPM, which is also often referred to as a *generalized power method* (GPM) in the literature, is a family of iterative algorithms for solving constrained optimization problems. It has been used with success for various tasks including clustering SBM (Wang et al., 2021), group synchronization (Boumal, 2016; Gao et al., 2021), joint alignment from pairwise difference (Chen et al., 2018), low rank-matrix recovery (Chi et al., 2019) and the generalized orthogonal procrustes problem (Ling, 2021). It is a useful iterative strategy for solving non-convex optimization problems as shown in the previous chapters, but it usually requires a good enough initial estimate. The use of PPM for graph matching was first proposed and experimentally analysed in (Onaran et al., 2017) and it has been subsequently been analysed in the case of sparse Erdös-Renyi graphs in (Lubars et al., 2018; Yu et al., 2021a) (only for one iteration) and in (Mao et al., 2021a) (although the connection with PPM is not mentioned in those works).

GRAPH MATCHING. When no side information is available, several polynomial time algorithms have been proposed recently relying on

spectral methods (Umeyama, 1988; Fan et al., 2019; Ganassali et al., 2022a; Feizi et al., 2020; Cour et al., 2006), degree profiles (Ding et al., 2021; Dai et al., 2018), other vertex signatures (Mao et al., 2021a), random walk based approaches (Rohit Singh et al., 2008; Kazemi et al., 2016; Gori et al., 2004), convex and concave relaxations (Aflalo et al., 2015; Lyzinski et al., 2016; Zaslavskiy et al., 2009a), and other non-convex methods (Yu et al., 2018; Xu et al., 2019; Onaran et al., 2017). Most of the previous algorithms have theoretical guarantees only in the low noise regime. A more complete discussion of the problem can be found in Chapter 6.

SEEDED ALGORITHMS. Different algorithms have been proposed when a seed of the the form $S = \{(i, i') : i \in V(G), i' \in V(H)\}$ is given as side information (Pedarsani et al., 2011; Yartseva et al., 2013; Mossel et al., 2019; Fishkind et al., 2019; Lubars et al., 2018; Yu et al., 2021a). Many algorithms in this class work under the assumption that the information provided by the seed corresponds perfectly to the ground truth permutation, i.e., $(i, i') \in S$ if and only if $x^*(i) = i'$. Some algorithms relax this requirement allowing "noisy" seeds, where for some (i, i') in S it happens that $x^*(i) \neq i$ (Yartseva et al., 2013; Kazemi et al., 2015; Lubars et al., 2018; Yu et al., 2021a; Mao et al., 2021a). Most of the previous work on the seeded version of the problem has been devoted to the Erdös-Renyi model, under different assumptions on the sparsity. To the best of our knowledge, the state-of-art algorithm in this category is the `j-hop` algorithm (Yu et al., 2021a, Alg.1), although it shares similarities with (Lubars et al., 2018, Alg.2) and (Mao et al., 2021a, Alg.4). On the other hand, it will be evident from our analysis of PPM for graph matching that those algorithms can also be seen as examples of the PPM.

7.2 ALGORITHM

7.2.1 *Projected power method for Graph matching*

The projected power method (PPM) has been used to solve the graph matching problem, and its variants, by several authors (Onaran et al., 2017). Most of the work so far has been empirical and, to the best of our knowledge, theoretical guarantees have been obtained only in the case of sparse Erdös-Renyi graphs, such as in (Mao et al., 2021a, Thm.B) in the case of multiple iterations, and (Yu et al., 2021a; Lubars et al., 2018) in the case of one iteration. Interestingly, the connection with the PPM is not explicitly stated in any of these works.

We start by defining the projection operator onto $\mathcal{P}_n$ for a matrix $C \in \mathbb{R}^{n \times n}$. We will use the greedy maximum weight matching (GMWM) algorithm introduced in (Lubars et al., 2018), for the problem of graph matching with partially correct seeds, and subsequently

used in (Yu et al., 2021a). The steps are outlined in Algorithm 11. No-

Algorithm 11 GMWM (Greedy maximum weight matching)

Input: A cost matrix $C \in \mathbb{R}^{n\times n}$.

1: Select (i_1, j_1) such that $C_{i_1 j_1}$ is the largest entry in C (break ties arbitrarily). Define $C^{(1)} \in \mathbb{R}^{n\times n}$ by $C^{(1)}_{ij} = C_{ij}\mathbb{1}_{i\neq i_1, j\neq j_1} - \infty \cdot \mathbb{1}_{i=i_1 \text{ or } j=j_1}$.
2: **for** $k = 2$ to N **do**
3: Select (i_k, j_k) such that $C^{(k-1)}_{i_k j_k}$ is the largest entry in $C^{(k-1)}$.
4: Define $C^{(k)} \in \mathbb{R}^{n\times n}$: $C^{(k)}_{ij} = C^{(k-1)}_{ij}\mathbb{1}_{i\neq i_k, j\neq j_k} - \infty \cdot \mathbb{1}_{i=i_k \text{ or } j=j_k}$.
5: **end for**
6: Define $X \in \{0,1\}^{n\times n}$: $X_{ij} = \sum_{k=1}^{N} \mathbb{1}_{i=i_k, j=j_k}$.
7: **return** X

Output: A permutation matrix X.

tice that the original version of GMWM works by erasing the row and column of the largest entry of the matrix $C^{(k)}$ at each step k. We change this to assign $-\infty$ to each element of the row and column of the largest entry (which is equivalent), mainly to maintain the original indexation. The output of Algorithm 11 is clearly a permutation matrix, hence we define

$$\tau(C) := \{\text{Output of GMWM with input } C\} \tag{7.2.1}$$

which can be considered a projection since $\tau^2(C) = \tau(C)$ for all $C \in \mathbb{R}^{n\times n}$. Notice that, in general, the output of GMWM will be different from solving the linear assignment problem

$$\tilde{\tau}(C) := \operatorname{argmin}\{\|C - X\|_F \mid X \in \mathcal{P}_n\} = \operatorname*{argmax}_{\Pi\in\mathcal{P}_n} \langle \Pi, C\rangle_F$$

which provides an orthogonal projection, while τ corresponds to an oblique projection in general.

Algorithm 12 PPMGM (PPM for graph matching)

Require: Matrices A, B, an initial point $X^{(0)}$ and N the maximum number of iterations.

Ensure: A permutation matrix X.

1: **for** $k = 0$ to $N - 1$ **do**
2: $X^{(k+1)} \leftarrow \tau(AX^{(k)}B)$.
3: **end for**
4: **return** $X = X^{(N)}$

The PPM is outlined in Algorithm 12. Given the estimate of the permutation $X^{(k)}$ from step k, the power step corresponds to the operation $AX^{(k)}B$ while the projection step is given by the application of the projection τ on $AX^{(k)}B$. The similarity matrix $C^{k+1} := AX^{(k)}B$

is the matrix form of the left multiplication of $[X^{(k)}]$ by the matrix $B \otimes A$. Indeed, given that A and B are symmetric matrices, we have $[AX^{(k)}B] = (B \otimes A)[X^{(k)}]$, by (Schäcke, 2004, eqs. 6 and 10). All previous works related to the PPM for graph matching use $(B \otimes A)[X^{(k)}]$ in the power step, which is highly inconvenient in practice. Also, a power step of the form $AX^{(k)}B$ connects the PPM with the seeded graph matching methods proposed for correlated Erdös-Renyi graphs (Lubars et al., 2018; Yu et al., 2021a; Mao et al., 2021a) where related similarity matrices are used, thus providing a more general framework. Indeed, the set of elements correctly matched by the initial permutation $x^{(0)} \in \mathcal{S}_n$ can be considered as the seed of the problem, *i.e.*, we take the set of seeds $S := \{(i, i') : x^{(0)}(i) = i'\}$. Thus, the number of correct seeds will be the number of elements $i \in [n]$ such that $x^{(0)}(i) = x^*(i)$. Observe that the definition of the seed as a permutation is more general than a set S of bijectively pre-matched vertices, because S can be augmented (arbitrarily) to a permutation.

INITIALIZATION. We prove in Section 7.3 that Algorithm 12 recovers the ground truth permutation x^* provided that the initialization $x^{(0)}$ is sufficiently close to x^*. The initialization assumption will be written in the form

$$\|X^{(0)} - X^*\|_F \leqslant \theta\sqrt{n} \tag{7.2.2}$$

for some $\theta \in [0, \sqrt{2})$. Here, the value of θ measures how good $X^{(0)}$ is as a seed. Indeed, (7.2.2) can be equivalently stated as: the number of correct seeds is larger than $n(2 - \theta^2)$. The question of finding a good initialization method can be seen as a seedless graph matching problem, where only partial recovery guarantees are necessary. In practice, we can use existing seedless algorithms such as those in (Umeyama, 1988; Fan et al., 2019; Feizi et al., 2020) to initialize Algorithm 12. We compare different initialization methods numerically, in Section 7.5.

Remark 20 (PPM as a gradient method)**.** *The projected power method can be seen as a projected gradient ascent method for solving the MLE formulation in* (1.5.1)*. This can be reformulated in a vectorial form as*

$$\max_{[X] \in [\mathcal{P}_n]} [X]^T B \otimes A [X] \tag{7.2.3}$$

where $[\mathcal{P}_n]$ *is the set of permutation matrices in vector form. From the formulation* (7.2.3) *it is clear that the gradient of the likelihood evaluated on* $X \in \mathcal{P}_n$ *is* $2B \otimes A[X]$ *or, equivalently,* $2AXB$ *in matrix form. This interpretation of PPM has been acknowledged in the context of other statistical problems (Journée et al., 2010; Chen et al., 2018).*

Remark 21 (Optimality)**.** *Algorithms based on PPM or GPM have been shown to attain optimal, or near-optimal, statistical guarantees for several problems in statistics, including community detection (Wang et al., 2021; Wang et al., 2022b), group syncronization (Boumal, 2016; Gao et al., 2019) and generalized orthogonal procrustes problem (Ling, 2021).*

Remark 22 (Complexity). *The computational time complexity of Algorithm 12 is* $\mathcal{O}(n^\omega \log n + n^2 \log^2 n)$, *where* $\mathcal{O}(n^\omega)$ *is the matrix multiplication complexity and* $\mathcal{O}(n^2 \log n)$ *is the complexity of Algorithm 11 (Yu et al., 2021a). In (Le Gall, 2014), the authors establish the bound* $\omega \leqslant 2.373$.

7.3 MAIN RESULTS

Our goal in this section is to prove recovery guarantees for Algorithm 12 when the input matrices A, B are realizations of the correlated Wigner model. In what follows, we will assume without loss of generality that $X^* = I_n$.

7.3.1 *Exact recovery in one iteration*

For any given seed $x^{(0)}$ that is close enough to x^*, the main result of this section states that x^* is recovered exactly in one iteration of Algorithm 12 with high probability. Let us first introduce the following definition: we say that a matrix M is diagonally dominant[1] if for all i, j with $i \neq j$ we have $M_{ii} > M_{ij}$. This notion will be used in conjunction with the following lemma, its proof is in Appendix 7.8.

Lemma 20. *If a matrix* C *satisfies the diagonal dominance property, then the greedy algorithm* `GMWM` *with input* C *will return the identical permutation. Consequently, for* $C = AXB$ *and* $\Pi = \tau(C)$, *we have*

$$\mathbb{P}(\Pi \neq I_n) \leqslant \mathbb{P}(C \text{ is not diag. dominant}) \tag{7.3.1}$$

The next theorem allows us to control the probability that C is not diagonally dominant and, in turn, proves that Algorithm 12 recovers the ground truth permutation with high probability. The proof of Theorem 17 is outlined in Section 7.4.1.

Theorem 17. *Let* $A, B \sim \mathrm{CoWM}(n, \sigma, \mathrm{id})$ *and* $X \in \mathcal{P}_n$ *with* $\|X - \mathrm{Id}\|_F \leqslant \theta\sqrt{n}$, *with* $0 \leqslant \theta \leqslant \sqrt{2(1 - \frac{10}{n})}$ *and* $n \geqslant 10$. *Then the following holds.*

1. *For* $C = AXB$ *we have*

$$\mathbb{P}(C \text{ is not diag. dominant }) \leqslant 5n^2 e^{-c(\sigma)\left(1-\frac{\theta^2}{2}\right)^2 n}$$

   where $c(\sigma) = \frac{1}{384}\left(\frac{1-\sigma^2}{1+2\sigma\sqrt{1-\sigma^2}}\right)$.

2. *Denote* Π *as the output of Algorithm 12 with input* $(A, B, X^{(0)} = X, N = 1)$, *then*

$$\mathbb{P}(\Pi = \mathrm{Id}) \geqslant 1 - 5n^2 e^{-c(\sigma)\left(1-\frac{\theta^2}{2}\right)^2 n}.$$

1 This is weaker than the usual notion of diagonal dominance, where for all $i \in [n]$ $|M_{ii}| \geqslant \sum_{j \neq i} |M_{ij}|$.

In particular, if $\|X - \mathrm{Id}\|_F^2 \leqslant 2\left(n - \sqrt{\frac{1}{c(\sigma)} n \log(5n^3)}\right)$ *then*

$$\mathbb{P}(\Pi = \mathrm{Id}) \geqslant 1 - n^{-1}.$$

Remark 23. *The assumption* $\|X - \mathrm{Id}\|_F^2 \leqslant 2(n - \sqrt{\frac{1}{c(\sigma)} n \log(5n^3)})$ *can be restated as* $|S_X| \geqslant \sqrt{\frac{1}{c(\sigma)} n \log 5n^3}$*, where* S_X *is the set of fixed points of* X*. That is, for this assumption to hold, we need that* X *has a number of fixed points of order* $\Omega_\sigma(\sqrt{n \log n})$*. Also note that* $c(\sigma)$ *is decreasing with* σ*, which is consistent with the intuition that larger levels of noise make it more difficult to recover the ground truth permutation. We include a plot of* $c(\sigma)$ *(rescaled) in Figure 20.*

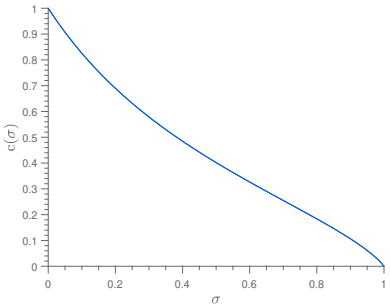

Figure 20: The constant $c(\sigma)$ (re-scaled multiplying by 384) appearing in Theorem 17.

DISCUSSION. Given an initial seed $X^{(0)} \in \mathcal{P}_n$, the case $N = 1$ in Algorithm 12 can be alternatively interpreted as the following two step process: first, compute a similarity matrix $AX^{(0)}B$ and then round the similarity matrix to an actual permutation matrix. This strategy has been frequently applied in graph matching algorithms in both the seeded and seedless case (Umeyama, 1988; Fan et al., 2019; Lubars et al., 2018; Yu et al., 2021a). In terms of the quality of the seed, Theorem 17 gives the same guarantees obtained by Yu et al., 2021a, Thm.1 which requires $\Omega(\sqrt{n \log n})$ vertices in the seed to be correctly matched. However the results of (Yu et al., 2021a) are specifically for the correlated Erdös-Renyi model.

7.3.2 *Partial recovery in one iteration*

In the partial recovery setting, we are interested in the fraction of nodes that are correctly matched. To this end let us define the following measure of performance

$$\mathrm{overlap}(\nu, \nu') := \frac{1}{n} |\{i \in [n] : \nu(i) = \nu'(i)\}| \tag{7.3.2}$$

for any pair $\nu, \nu' \in \mathcal{S}_n$. Recall that we assume that the ground truth permutation is $x^* = \mathrm{id}$ and π is the output of Algorithm 12 with

input $(A, B, X^{(0)} = X, N = 1)$ where $\Pi = \mathrm{GMWM}(AXB)$. Observe that $\mathrm{overlap}(\pi, x^* = \mathrm{id}) = s_\pi$ is the fraction of fixed points of the permutation π. It will be useful to consider the following definition. We say that C_{ij} is *row-column dominant* if $C_{ij} > C_{i'j}$ for all $i' \neq i$ and $C_{ij} > C_{ij'}$, for all $j' \neq j$. The following lemma relates the overlap of the output of `GMWM` with the property that a subset of the entries of C is row-column dominant, its proof is outlined in Appendix 7.8.

Lemma 21. *Let C be a $n \times n$ matrix with the property that there exists a set $\{i_1, \cdots, i_r\}$, with $1 \leqslant r \leqslant n$ such that C_{i_k,i_k} is row-column dominant for $k \in [r]$. Let $\pi \in \mathcal{S}_n$ be permutation corresponding to $\mathrm{GMWM}(C) \in \mathcal{P}_n$. Then it holds that $\pi(i_k) = i_k$ for $k \in [r]$ and, in consequence, the following event inclusion holds*

$$\{\mathrm{overlap}(\pi, \mathrm{id}) < r/n\} \subset \bigcap_{\substack{I_r \subset [n] \\ |I_r| = r}} \bigcup_{i \in I_r} \{C_{ii} \text{ is not row-column dominant }\}. \tag{7.3.3}$$

Equipped with this lemma, we can prove the following generalization of Theorem 17, its proof is detailed in Section 7.4.2.

Theorem 18. *Let $A, B \sim W(n, \sigma, \mathrm{id})$ and $X \in \mathcal{P}_n$ with $\|X - \mathrm{Id}\|_F \leqslant \theta\sqrt{n}$, where $0 \leqslant \theta \leqslant \sqrt{2(1 - \frac{10}{n})}$ and $n \geqslant 10$. Let $\pi \in \mathcal{S}_n$ be the output of Algorithm 12 with input $(A, B, X^{(0)} = X, N = 1)$. Then, for $r \in [n]$*

$$\mathbb{P}(\mathrm{overlap}(\pi, \mathrm{id}) > r/n) \geqslant 1 - 16rne^{-c(\sigma)\left(1 - \frac{\theta^2}{2}\right)^2 n}.$$

In particular, if $x \in \mathcal{S}_n$ is the map corresponding to X and

$$|S_X| \geqslant \sqrt{\frac{1}{c(\sigma)} n \log(16rn^2)},$$

then

$$\mathbb{P}(\mathrm{overlap}(\pi, \mathrm{id}) > r/n) \geqslant 1 - n^{-1}.$$

7.3.3 *Exact recovery after multiple iterations, uniformly in the seed*

The results in Sections 7.3.1 and 7.3.2 hold for any given seed $X^{(0)}$, and it is crucial that the seed does not depend on the graphs A, B. In this section, we provide uniform convergence guarantees for `PPMGM` which hold uniformly over all choices of the seed in a neighborhood around x^*.

Theorem 19. *Let $\sigma \in [0, 1)$, $A, B \sim W(n, \sigma, \mathrm{id})$ and let $X^{(0)}$ be a – possibly random and data dependent – permutation such that $|S_{X^{(0)}}| \geqslant (1 - \kappa)n$ for a constant $\kappa > 0$ such that $\sqrt{1 - \sigma^2} > 48\kappa$. Then by applying `PPMGM` with input $(\mathcal{H}(A), \mathcal{H}(B), X^{(0)}, N = 2\log n)$ where $\mathcal{H}(X)$ corresponds to*

the matrix X with the diagonal removed, when n is large enough, we obtain a permutation $X^{(N)}$ such that

$$\mathbb{P}(X^{(N)} \neq \mathrm{Id}) \geqslant 1 - \frac{6 \log n}{n^2}.$$

The diagonal of the adjacency matrices A and B in Algorithm 12 was removed in the above theorem only for ease of analysis. Its proof is detailed in Section 7.4.3.

Remark 24. *Contrary to our previous theorems, here the strong consistency of the estimator holds uniformly over all possible seeds that satisfy the condition $|S_{X^{(0)}}| \geqslant (1-\kappa)n$. For this reason, we need a stronger condition than $|S_X| = \Omega(\sqrt{n \log n})$ as was the case in Theorem 17. Our result is non trivial and cannot be obtained from Theorem 17 by taking a union bound. The proof relies on a decoupling technique adapted from Mao et al., 2021a that used a similar refinement method for Erdös-Renyi graphs.*

Remark 25. *Contrary to the results obtained in the seedless case that require $\sigma = o(1)$ for exact recovery (Fan et al., 2019), we can allow σ to be of constant order. The condition $\sqrt{1-\sigma^2} > 48\kappa$ seems to be far from optimal as shown in the experiments in Section 7.5. For example, PPMGM can achieve exact recovery when $\kappa = 0.08$ and $\sigma = 0.6$. But interestingly, this condition shows that when the noise σ increases, PPMGM needs a more accurate initialization, hence a larger κ, to recover the latent permutation. This is confirmed by our experiments.*

7.4 PROOF OUTLINE

7.4.1 *Proof of Theorem 17*

For $A, B \sim W(n, \sigma, \mathrm{id})$, the proof of Theorem 17 relies heavily on the concentration properties of the entries of the matrix $C = AXB$, which is the matrix that is projected by our proposed algorithm. In particular, we use the fact that C is diagonally dominant with high probability, under the assumptions of Theorem 17, which is given by the following result. Its proof is delayed to Appendix 7.7.1.

Proposition 4 (Diagonal dominance property for the matrix $C = AXB$)**.** *Let $A, B \sim W(n, \sigma, \mathrm{id})$ with correlation parameter $\sigma \in [0, 1)$ and let $X \in \mathcal{P}_n$ with S_X the set of its fixed points and $s_X := |S_X|/n$. Assume that $s_X \geqslant 10/n$ and that $n \geqslant 10$. Then the following is true.*

1. ***Noiseless case.*** *For a fixed $i \in [n]$ it holds that*

$$\mathbb{P}\big(\exists j \neq i : (AXA)_{ij} > (AXA)_{ii}\big) \leqslant 4ne^{-\frac{s_X^2}{96}n}.$$

2. *For $C = AXB$ and $i \in [n]$ it holds*

$$\mathbb{P}(\exists j \neq i : C_{ij} > C_{ii}) \leqslant 5ne^{-c(\sigma)s_X^2 n}$$

where $c(\sigma) = \frac{1}{384}\frac{1-\sigma^2}{1+2\sigma\sqrt{1-\sigma^2}}$.

With this we can proceed with the proof of Theorem 17.

Proof of Theorem 17. To prove part (i) of the theorem it suffices to notice that in Proposition 4 part (ii) we upper bound the probability that $C = AXB$ is not diagonally dominant for each fixed row. Using the union bound, summing over the n rows, we obtain the desired upper bound on the probability that C is not diagonally dominant. We now prove part (ii). Notice that the assumption $\|X - I_n\|_F \leqslant \theta\sqrt{n}$ for $\theta < \sqrt{2}$ implies that s_X is strictly positive. Moreover, from this assumption and the fact that $\|X - I_n\|_F^2 = 2(n - |S_X|)$ we deduce that

$$s_X \geqslant \left(1 - \frac{\theta^2}{2}\right). \tag{7.4.1}$$

On the other hand, we have

$$\begin{aligned}
\mathbb{P}(\Pi \neq \mathrm{Id}) &\leqslant \mathbb{P}(C \text{ is not diag.dom}) \\
&= \mathbb{P}(\exists i, j \in [n], i \neq j : C_{ii} < C_{ij}) \\
&\leqslant 5n^2 e^{-c(\sigma)s_X^2 n} \\
&\leqslant 5n^2 e^{-c(\sigma)\left(1-\frac{\theta^2}{2}\right)^2 n}
\end{aligned}$$

where we used Lemma 20 in the first inequality, Proposition 4 in the penultimate step and, (7.4.1) in the last inequality. □

7.4.1.1 *Proof of Proposition 4*

In Proposition 4 part (i) we assume that $\sigma = 0$. The following are the main steps of the proof.

1. We first prove that for all $X \in \mathcal{P}_n$ such that $s_X = |S_X|/n$ and for $i \neq j \in [n]$ the gap $C_{ii} - C_{ij}$ is of order s_X in expectation.
2. We prove that C_{ii} and C_{ij} are sufficiently concentrated around its mean. In particular, the probability that C_{ii} is smaller than $s_X/2$ is exponentially small. The same is true for the probability that C_{ij} is larger than $s_X/2$.
3. We use the fact $\mathbb{P}(C_{ii} \leqslant C_{ij}) < \mathbb{P}(C_{ii} \leqslant s_X/2) + \mathbb{P}(C_{ij} \geqslant s_X/2)$ to control the probability that C is not diagonally dominant.

The proof is mainly based upon the following two lemmas.

Lemma 22. *For the matrix $C = AXA$ and with $s_X = |S_X|/n$ we have*

$$\mathbb{E}[C_{ij}] = \begin{cases} s_X + \frac{1}{n}\mathbb{1}_{i \in S_X} & \text{for } i = j, \\ \frac{1}{n}\mathbb{1}_{X(j)=i} & \text{for } i \neq j, \end{cases}$$

and from this we deduce that for $i, j \in [n]$ with $i \neq j$

$$s_X - \frac{1}{n} \leqslant \mathbb{E}[C_{ii}] - \mathbb{E}[C_{ij}] \leqslant s_X + \frac{1}{n}.$$

Lemma 23. *Assume that* $s_X \in (10/n, 1]$ *and* $n \geqslant 10$. *Then for* $i, j \in [n]$ *with* $i \neq j$ *we have*

$$\mathbb{P}(C_{ii} \leqslant s_X/2) \leqslant 4e^{-\frac{s_X^2}{48}n}, \tag{7.4.2}$$

$$\mathbb{P}(C_{ij} \geqslant s_X/2) \leqslant 3e^{-\frac{s_X^2}{96}n}. \tag{7.4.3}$$

With this we can prove Proposition 4 part (i).

Proof of Prop. 4 (i). Define the event $\mathcal{E}_j = \{C_{ii} < \frac{s_X}{2}\} \cup \{C_{ij} > \frac{s_X}{2}\}$ and note that for $j \neq i$, we have $\{C_{ij} > C_{ii}\} \subset \mathcal{E}_j$. With this and the bounds (7.4.2) and (7.4.3) we have

$$\begin{aligned}
\mathbb{P}\big(\exists j \neq i : C_{ij} > C_{ii}\big) &= \mathbb{P}(\cup_{j\neq i}\{C_{ij} > C_{ii}\}) \\
&\leqslant \mathbb{P}(\cup_{j\neq i}\mathcal{E}_j) \\
&\leqslant \mathbb{P}(C_{ii} \leqslant \frac{s_X}{2}) + \sum_{j\neq i}\mathbb{P}(C_{ij} \geqslant \frac{s_X}{2}) \\
&\leqslant 4e^{-\frac{s_X^2}{96}n} + 3(n-1)e^{-\frac{s_X^2}{96}n} \\
&\leqslant 4ne^{-\frac{s_X^2}{96}n}.
\end{aligned}$$

□

The proof of Lemma 22 is short and we include it in the main body of the paper. On the other hand, the proof of Lemma 23 mainly uses concentration inequalities for Gaussian quadratic forms, but the details are quite technical. Hence we delay its proof to Section 7.7.1.1. Before proceeding with the proof of Lemma 22, observe that the following decomposition holds for the matrix C.

$$C_{ij} = \sum_{k,k'} A_{ik}X_{k,k'}A_{k'i} = \begin{cases} \sum_{k\in S_X} A_{ik}^2 + \sum_{k\notin S_X} A_{ik}A_{ix(k)} & \text{for } i = j, \\ \sum_{k=1}^n A_{ik}A_{x(k)j} & \text{for } i \neq j. \end{cases} \tag{7.4.4}$$

Proof of Lemma 22. From (7.4.4) we have that

$$\mathbb{E}[C_{ii}] = \sum_{k\in S_X}\mathbb{E}[A_{ik}^2] + \sum_{k\notin S_X}\mathbb{E}[A_{ik}^2] = \frac{|S_X|}{n} + \frac{\mathbb{1}_{i\in S_X}}{n}.$$

Similarly, for $j \neq i$ it holds

$$\mathbb{E}[C_{ij}] = \sum_{k=1}^n \mathbb{E}[A_{ik}A_{x(k)j}] = \frac{1}{n}\mathbb{1}_{i,j\notin S_X, x(j)=i} = \frac{\mathbb{1}_{x(j)=i}}{n}$$

from which the results follows easily. □

The proof of Proposition 4 part (ii) which corresponds to the case $\sigma \neq 0$ uses similar ideas and the details can be found in Appendix 7.7.1.2.

7.4.2 *Proof of Theorem 18*

The proof of Theorem 18 will be based on the following lemma, which extends Proposition 4.

Lemma 24. *For a fixed* $i \in [n]$*, we have*

$$\mathbb{P}(C_{ii} \text{ is not row-column dominant}) \leqslant 16ne^{-c(\sigma)s_x^2 n}.$$

The proof of Lemma 24 is included in Section 7.7.2. We now prove Theorem 18. The main idea is that for a fixed $i \in [n]$, with high probability the term C_{ii} will be the largest term in the i-th row and the i-th column, and so `GMWM` will assign $\pi(i) = i$. We will also use the following event inclusion, which is direct from (7.3.3) in Lemma 21.

$$\{\text{overlap}(\pi, \mathrm{id}) < r/n\} \subset \bigcup_{i=1}^{r} \{C_{ii} \text{ is not row-column dominant }\}. \tag{7.4.5}$$

Proof of Theorem 18 . Fix $i \in [n]$. By (7.4.5) we have that

$$\begin{aligned}
\mathbb{P}(\text{overlap}(\pi, \mathrm{id}) \leqslant r/n) &\leqslant \sum_{i=1}^{r} \mathbb{P}(C_{ii} \text{ is not row-column dominant}) \\
&\leqslant \sum_{i=1}^{r} \mathbb{P}(\exists j \neq i, \text{ s.t } C_{ij} \vee C_{ji} > C_{ii}) \\
&\leqslant 16rne^{-c(\sigma)s_x^2 n}
\end{aligned}$$

where we used Lemma 24 in the last inequality. □

Remark 26. *Notice that the RHS of* (7.4.5) *is a superset of the RHS of* (7.3.3)*. To improve this, it is necessary to include dependency information. In other words, we need to 'beat Hölder's inequality'. To see this, define*

$$E_i := \mathbb{1}_{C_{ii} \textit{ is not row-column dominant}}, \quad \varepsilon_I := \mathbb{1}_{\sum_{i \in I} E_i > 0}, \textit{ for } I \subset [n];$$

then $\varepsilon_{I'}$*, for* $I' = [r]$*, is the indicator of the event in the RHS of* (7.4.5)*. On other hand, the indicator of the event in the RHS of* (7.3.3) *is* $\prod_{I \subset [n], |I| = r} \varepsilon_I$*. If* $\mathbb{E}\left[\varepsilon_I\right]$ *is equal for all* I*, then Hölder inequality gives*

$$\mathbb{E}\Big[\prod_{I \subset [n], |I| = r} \varepsilon_I\Big] \leqslant \mathbb{E}[\varepsilon_{I'}]$$

which does not help in quantifying the difference between (7.3.3) *and* (7.4.5)*. This is not surprising as we are not taking into account the dependency between the events* ε_I *for the different sets* $I \subset [n], |I| = r$*.*

7.4.3 *Proof of Theorem 19*

The general proof idea is based on the decoupling strategy used by (Mao et al., 2021a) for Erdös-Renyi graphs. To extend their result from binary graphs to weighted graphs, we need to use an appropriate measure of similarity. For $i, i' \in [n], W \subset [n]$ and $g \in \mathcal{P}_n$, let us define

$$\langle A_{i:}, B_{i':} \rangle_{g,W} := \sum_{j \in W} A_{ig(j)} B_{i'j}$$

to be the similarity between i and i' restricted to W and measured with a scalar product depending on g (the permutation used to align A and B). When $g = \mathrm{id}$ or $W = [n]$ we will drop the corresponding subscript(s). If A and B were binary matrices, we would have the following correspondence

$$\langle A_{i:}, B_{i':} \rangle_{g,W} = |g(\mathcal{N}_A(i) \cap W) \cap \mathcal{N}_B(i')|.$$

This last quantity plays an essential role in Proposition 7.5 of (Mao et al., 2021a). Here $g(S)$ denotes the image of a set $S \subseteq [n]$ under permutation g.

STEP 1. The algorithm design relies on the fact that if the matrices A and B were correctly aligned then the correlation between $A_{i:}$ and $B_{i:}$ should be large and the correlation between $A_{i:}$ and $B_{i':}$ should be small for all $i \neq i'$. The following two Lemmas precisely quantify these correlations when the two matrices are well aligned.

Lemma 25 (Correlation between corresponding nodes). *Let* $(A, B) \sim W(n, \sigma, x^* = \mathrm{id})$ *and assume that the diagonals of* A *and* B *have been removed. Then for* n *large enough, we have with probability at least* $1 - n^{-2}$ *that*

$$\langle A_{i:}, B_{i:} \rangle \geqslant \sqrt{1-\sigma^2}(1-\epsilon_1) - \sigma\epsilon_2 \text{ for all } i \in [n],$$

where $\epsilon_1, \epsilon_2 = O(\sqrt{\frac{\log n}{n}})$.

Lemma 26 (Correlation between different nodes). *Let* $(A, B) \sim W(n, \sigma, \mathrm{id})$ *and assume that the diagonals of* A *and* B *have been removed. Then for* n *large enough, we have with probability at least* $1 - n^{-2}$ *that*

$$|\langle A_{i:}, B_{i':} \rangle| \leqslant \sqrt{1-\sigma^2}\epsilon_2 + \sigma\epsilon_3 \text{ for all } i, i' \in [n] \text{ such that } i' \neq i,$$

where $\epsilon_3 = O(\sqrt{\frac{\log n}{n}})$.

The proofs of Lemmas 25 and 26 can be found in Section 7.8.1.

STEP 2. Since the ground truth alignment between A and B is unknown, we need to use an approximate alignment (provided by $X^{(0)}$). It will suffice that $X^{(0)}$ is close enough to the ground truth permutation. This is linked to the fact that if $|S_{X^{(0)}}|$ is large enough then

the number of nodes for which there is a substantial amount of information contained in $S^c_{X^{(0)}}$ is small. This is shown in the following lemma.

Lemma 27 (Growing a subset of vertices)**.** *Let* G *a graph generated from the Wigner model with self-loops removed, associated with an adjacency matrix* A*, and let* I *be a random subset of* $[n]$ *(possibly depending on* A*) with* $|I| \geqslant (1-\kappa)n$ *where* $\kappa \in (0, 1/2)$*. Let* $\delta = 8\kappa$ *and define a random subset of vertices*

$$\tilde{I} = \{i \in [n] : \|A_{i:}\|^2_{I^c} < \delta\}.$$

Then for n *large enough, we have*

$$\mathbb{P}\left(|\tilde{I}^c| \leqslant \frac{1}{4}|I^c|\right) \geqslant 1 - e^{-c'\kappa n}$$

for some constant $c' > 0$*.*

In order to prove this lemma we will need the following decoupling lemma.

Lemma 28 (An elementary decoupling)**.** *Let* $M > 0$ *be a parameter and* G *be a weighted graph on* $[n]$*, with weights of magnitude bounded by* 1 *and without self loops, represented by an adjacency matrix* $A \in [-1, 1]^{n\times n}$*. Assume that there are two subsets of vertices* $Q, W \subset [n]$ *such that*

$$\|A_{i:}\|^2_W \geqslant M \text{ for all } i \in Q.$$

Then there are subsets $Q' \subseteq Q$ *and* $W' \subseteq W$ *such that* $Q' \cap W' = \emptyset$*,* $|Q'| \geqslant |Q|/5$ *and*

$$\|A_{i:}\|^2_{W'} \geqslant M/2 \text{ for all } i \in Q'.$$

Proof. If $|Q \setminus W| \geqslant |Q|/5$ then one can take $Q' = Q \setminus W$ and $W' = W$. So we can assume that $|Q \cap W| \geqslant 4|Q|/5$. Let $\tilde{W} := W \setminus Q$ and $\hat{Q}$ be a random subset of $Q \cap W$ where each element $j \in Q \cap W$ is selected independently with probability $1/2$ in $\hat{Q}$. Consider the random disjoint sets $\hat{Q}$ and $W' := \tilde{W} \cup ((Q \cap W) \setminus \hat{Q})$. First, we will show the following claim.

Claim 1. *For every* $i \in Q \cap W$*, we have* $\mathbb{P}(\|A_{i:}\|^2_{W'} \geqslant M/2 | i \in \hat{Q}) \geqslant 1/2$.

Indeed, we have by definition

$$\|A_{i:}\|^2_{W'} = \sum_{j\in W'} A^2_{ij} = \sum_{j\in W\cap Q} A^2_{ij}\mathbb{1}_{j\notin\hat{Q}} + \sum_{j\in\tilde{W}} A^2_{ij}.$$

By taking the expectation conditional on $i \in \hat{Q}$, we obtain

$$\mathbb{E}\left(\|A_{i:}\|^2_{W'}\middle| i \in \hat{Q}\right) = \sum_{j\in W\cap Q} \frac{A^2_{ij}}{2} + \sum_{j\in\tilde{W}} A^2_{ij} \geqslant \frac{1}{2}\sum_{j\in W} A^2_{ij} \geqslant \frac{M}{2}.$$

But since $\sum_{j\in W\cap Q} A_{ij}^2(\mathbb{1}_{j\notin\hat{Q}} - \frac{1}{2})$ is a symmetric random variable we have that

$$\mathbb{P}\left(\|A_{i:}\|_{W'}^2 \geqslant \mathbb{E}(\|A_{i:}\|_{W'}^2)\middle| i \in \hat{Q}\right) = 1/2$$

and hence

$$\mathbb{P}\left(\|A_{i:}\|_{W'}^2 \geqslant \frac{M}{2}\middle| i \in \hat{Q}\right) \geqslant 1/2.$$

Consequently, we have

$$\begin{aligned}\mathbb{E}\left(\sum_{i\in Q\cap W} \mathbb{1}_{\{\|A_{i:}\|_{W'}^2 \geqslant M/2\}}\mathbb{1}_{i\in\hat{Q}}\right) &= \sum_{i\in Q\cap W} \mathbb{P}(i\in\hat{Q})\mathbb{E}\left(\mathbb{1}_{\{\|A_{i:}\|_{W'}^2\geqslant M/2\}}\middle| i\in\hat{Q}\right)\\ &\geqslant \frac{|Q\cap W|}{4} \geqslant \frac{|Q|}{5}.\end{aligned}$$

Therefore, there is a realization Q' of $\hat{Q}$ such that Q' and W' satisfy the required conditions. □

Proof of Lemma 27. By considering sets $W = I^c$ and $Q \subset \tilde{I}^c$ we obtain the following inclusion

$$\{|\tilde{I}^c| > \frac{1}{4}|I^c|\} \subset \mathcal{E} := \{\exists\, Q, W \subset [n] : |W| \leqslant \kappa n, |Q| \geqslant |W|/4 \neq 0, \|A_{i:}\|_W^2 \geqslant \delta \text{ for all } i \in Q\}.$$

According to Lemma 28, $\mathcal{E}$ is contained in

$$\mathcal{E}' := \{\exists\, Q', W' \subset [n] : |W'| \leqslant \kappa n, |Q'| \geqslant |W|/20 \neq 0, Q'\cap W' = \emptyset, \|A_{i:}\|_{W'}^2 \geqslant \delta/2 \text{ for all } i \in Q'\}.$$

For given subsets Q' and W', the random variables $(\|A_{i:}\|_{W'}^2)_{i\in Q'}$ are independent. So, by a union bound argument we get

$$\mathbb{P}\left(|\tilde{I}^c| > \frac{1}{4}|I^c|\right) \leqslant \sum_{w=1}^{\lceil\kappa n\rceil} \sum_{|W'|=w} \sum_{k=\lceil w/20\rceil}^{n} \binom{n}{k} \mathbb{P}\left(\|A_{i:}\|_{W'}^2 \geqslant \delta/2\right)^k.$$

According to Lemma 37, for the choice $t = \kappa n$ we have for all W'

$$\mathbb{P}\left(\|A_{i:}\|_{W'}^2 \geqslant \delta/2\right) \leqslant \mathbb{P}\left(n\,\|A_{i:}\|_{W'}^2 \geqslant |W| + \sqrt{|W|t} + 2t\right) \leqslant e^{-\kappa n}.$$

Consequently, for n large enough, we have

$$\mathbb{P}\left(|\tilde{I}^c| > \frac{1}{4}|I^c|\right) \leqslant \sum_{w=1}^{\lceil\kappa n\rceil} \sum_{k=\lceil w/20\rceil}^{n} \left(\frac{en}{w}\right)^w \left(\frac{en}{k}\right)^k e^{-k\kappa n} < e^{-c\kappa n}$$

for a constant $c > 0$. Indeed, since

$$\frac{en}{ke^{\kappa n}} < 1$$

for n large enough we have

$$\sum_{k=\lceil w/20\rceil}^{n} \left(\frac{en}{k}\right)^k e^{-k\kappa n} \leqslant C\left(\frac{en}{e^{\kappa n}}\right)^{\lceil w/20\rceil}$$

by thw property of geometric series, where $C > 0$ is a constant. But by the same argument

$$\sum_{w=1}^{\lceil \kappa n \rceil} \left(\frac{en}{w}\right)^w \left(\frac{(en)^{1/20}}{e^{\kappa n/20}}\right)^w \leqslant \frac{(en)^{1/20}}{e^{\kappa n/20}} \leqslant e^{-c\kappa n}$$

where $c > 0$ is a constant. □

STEP 3. We are now in position to show that at each step the set of fixed points of the permutation obtained with PPMGM increases.

Lemma 29 (Improving a partial matching)**.** *Let* G *and* G' *two graphs as before, and* g *be a random permutation possibly depending on* G *and* G'*. Further assume that* $\sqrt{1-\sigma^2} > 48\kappa$*. Let*

$$\mathcal{E} := \{|i \in [n] : g(i) = i| \geqslant (1-\kappa)n\}$$

be the event that the number of fixed point of g *is large enough. Define a random permutation* $\tilde{g}$ *and a random set* $\tilde{J}$ *as follows. Let* $\delta = 8\kappa$*, we say that a vertex* $i \in [n]$ *belongs to* $\tilde{J}$ *if there is a unique* $i' \in [n]$ *such that*

- $\langle A_{i:}, B_{i':}\rangle_g \geqslant 3\delta$*;*
- $|\langle A_{i:}, B_{j:}\rangle_g| < 3\delta$ *for all* $j \neq i'$*;*
- $|\langle A_{j:}, B_{i':}\rangle_g| < 3\delta$ *for all* $j \neq i$*.*

Then we set $\tilde{g}(i) = i'$ *for any such pair of vertices. We complete* $\tilde{g}$ *into a permutation in an arbitrary way.*

If n *is sufficiently large and* κ *sufficiently small, we have with probability at least* $\mathbb{P}(\mathcal{E}) - \frac{3}{n^2}$*,*

$$|\{i \in [n] : \tilde{g}(i) = i\}| \geqslant \frac{n}{2} + \frac{|\{i \in [n] : g(i) = i\}|}{2}.$$

It implies in particular that the set of fixed points of $\tilde{g}$ *is strictly larger than the set of fixed points of* g*.*

Remark 27. *The description of* GMWM *doesn't involve the use of a threshold, but for the nodes that satisfy the conditions described in Lemma 29,* GMWM *provides by definition the same matching (this can be seen using the notion of row-column dominance and Lemma 21). Since the nodes that do not satisfy these conditions can be matched in an arbitrary way, we can use* GMWM *instead of the thresholding procedure and the analysis remains valid.*

Proof. Define the random sets

$$\begin{aligned} I &:= \{j \in [n] : g(j) = j\}, \\ \tilde{I} &:= \{j \in [n] : \|A_{i:}\|^2_{I^c} < \delta\}, \\ \tilde{I}' &:= \{j \in [n] : \|B_{i:}\|^2_{I^c} < \delta\}, \end{aligned}$$

where $\delta = 8\kappa$ and consider the event $\mathcal{E}' = \mathcal{E}'_1 \cap \mathcal{E}'_2 \cap \mathcal{E}'_3$ where

$$\begin{aligned}
\mathcal{E}'_1 &:= \{|\tilde{I}^c| \vee |(\tilde{I}')^c| \leqslant \frac{1}{4}|I^c|\} \\
\mathcal{E}'_2 &:= \{\forall i \in [n] : \ \langle A_{i:}, B_{i:}\rangle \geqslant 0.9\sqrt{1-\sigma^2}\} \\
\mathcal{E}'_3 &:= \{\forall i \neq i' \in [n] : \ |\langle A_{i:}, B_{i':}\rangle| < C \log n/n\}
\end{aligned}$$

for a suitably large constant $C > 0$ (which is the constant hidden in the $O(\cdot)$ symbol in Lemma 26). If n is sufficiently large and κ satisfies $\sqrt{1-\sigma^2} > 48\kappa$, one can show that

$$\mathbb{P}(\mathcal{E}' \cap \mathcal{E}) \geqslant \mathbb{P}(\mathcal{E}) - \frac{3}{n^2}$$

by combining Lemma 25, 26 and 27. Condition on any realization of G, G', g such that the event $\mathcal{E}' \cap \mathcal{E}$ holds. Let $i \in \tilde{I} \cap \tilde{I}'$. By definition of $\mathcal{E}' \cap \mathcal{E}$, we have

$$\begin{aligned}
\langle A_{i:}, B_{i:}\rangle_g &\geqslant \langle A_{i:}, B_{i:}\rangle_{g,I} - |\langle A_{i:}, B_{i:}\rangle_{g,I^c}| \\
&\geqslant \langle A_{i:}, B_{i:}\rangle - |\langle A_{i:}, B_{i:}\rangle_{I^c}| - |\langle A_{i:}, B_{i:}\rangle_{g,I^c}| \\
&\geqslant 3\delta.
\end{aligned}$$

Here we used the fact that for all permutations g, $|\langle A_{i:}, B_{i:}\rangle_{g,I^c}| \leqslant \|A_{i:}\|_{I^c} \|B_{i:}\|_{I^c}$ (because $g(I^c) = I^c$ by definition of I).

On the other hand, for every $i' \in [n] \setminus i$ we have

$$\begin{aligned}
|\langle A_{i:}, B_{i':}\rangle_g| &\leqslant |\langle A_{i:}, B_{i':}\rangle_{g,I}| + |\langle A_{i:}, B_{i':}\rangle_{g,I^c}| \\
&< 3\delta.
\end{aligned}$$

Similarly we have $|\langle A_{i':}, B_{i:}\rangle_g| < 3\delta$. Hence $\tilde{I} \cap \tilde{I}' \subset \tilde{J}$ and $\tilde{g}(i) = i$ for all $i \in \tilde{I} \cap \tilde{I}'$. Moreover we have by the first condition on $\mathcal{E}'$

$$|\tilde{I} \cap \tilde{I}'| \geqslant n - |\tilde{I}^c| - |(\tilde{I}')^c| \geqslant n - \frac{|I|^c}{2} = \frac{n}{2} + \frac{|I|}{2},$$

so the result follows. □

CONCLUSION. By Lemma 29, if the initial number of fixed points is $(1-\kappa)n$ then after one iteration step the size of the set of fixed points of the new iteration is at least $(1-\kappa/2)n$ with probability greater than $1 - \frac{3}{n^2}$. So after $2\log n$ iterations the set of fixed points has size at least $(1-\kappa/2^{2\log n})n > n-1$ with probability greater than $1 - \frac{6\log n}{n^2}$.

7.5 NUMERICAL EXPERIMENTS

In this section, we present numerical experiments to assess the performance of the PPMGM algorithm and compare it to the state-of-art algorithms for graph matching, under the correlated Wigner model. We divide this section into two parts. In Section 7.5.1 we generate

correlated Wigner graphs $A, B \sim W(n, \sigma, x^*)$ for a random permutation x^*, and apply to A, B the spectral algorithms `Grampa` (Fan et al., 2019) and the classic `Umeyama` (Umeyama, 1988), both of which work in the seedless case. As a second step, we apply algorithm `PPMGM` with the initialization given by the output of `Grampa` and `Umeyama`. We show experimentally that by applying `PPMGM` the solution obtained in both cases improves, when measured as the overlap (defined in (7.3.2)) of the output with the ground truth. We also run experiments by initializing `PPMGM` with $X^{(0)}$ randomly chosen at a certain distance of the ground truth permutation X^*. Specifically, we select $X^{(0)}$ uniformly at random from the set of permutation matrices that satisfy $\|X^{(0)} - X^*\|_F = \theta' \sqrt{n}$, and vary the value of $\theta' \in (0, 1)$.

In Section 7.5.2 we run algorithm `PPMGM` with different pairs of input matrices. We consider the Wigner correlated matrices A, B and also the pairs of matrices (A^{spar_1}, B^{spar_1}), (A^{spar_2}, B^{spar_2}) and (A^{spar_3}, B^{spar_3}), which are produced from A, B by means of a sparsification procedure (detailed in Section 7.5.2). The main idea behind this setting is that, to the best of our knowledge, the best theoretical guarantees for exact graph matching have been obtained in Mao et al., 2021a for relatively sparse Erdös-Renyi graphs. The algorithm proposed in Mao et al., 2021a has two steps, the first of which is a seedless type algorithm which produces a partially correct matching, that is later refined with a second algorithm Mao et al., 2021a, Alg.4. Their proposed algorithm `RefinedMatching` shares similarities with `PPMGM` and with algorithms `1-hop` (Lubars et al., 2018; Yu et al., 2021a) and `2-hop` (Yu et al., 2021a). Formulated as it is, `RefinedMatching` (Mao et al., 2021a) (and the same is true for `2-hop` for that matter) only accepts binary edge graphs as input and also uses a threshold-based rounding approach instead of Algorithm 11, which might be difficult to calibrate in practice. With this we address experimentally the fact that the analysis (and algorithms) in Mao et al., 2021a do not extend automatically to a simple binarization of the (dense) Wigner matrices, and that specially in high noise regimes, the sparsification strategies do not perform very well.

7.5.1 *Performance of `PPMGM`*

In Figure 21 we plot the recovery fraction, which is defined as the overlap (see (7.3.2)) between the ground truth permutation and the output of: `Grampa`, `Umeyama`, `Grampa+PPMGM`, `Umeyama+PPMGM` and `PPMGM`. The algorithms `Grampa+PPMGM` and `Umeyama+PPMGM` use the output of `Grampa` and `Umeyama` as seeds for `PPMGM`, which is performed with $N = 1$. In the algorithm `PPMGM`, we use an initial permutation $x^{(0)} \in \mathcal{S}_n$ chosen uniformly at random in the set of permutations such that $\text{overlap}\,(x^{(0)}, x^*) = 0.08$; this is referred to as '`PPMGM` rand.init'. We take $n = 800$ and consider the average overlap over 25 Monte Carlo

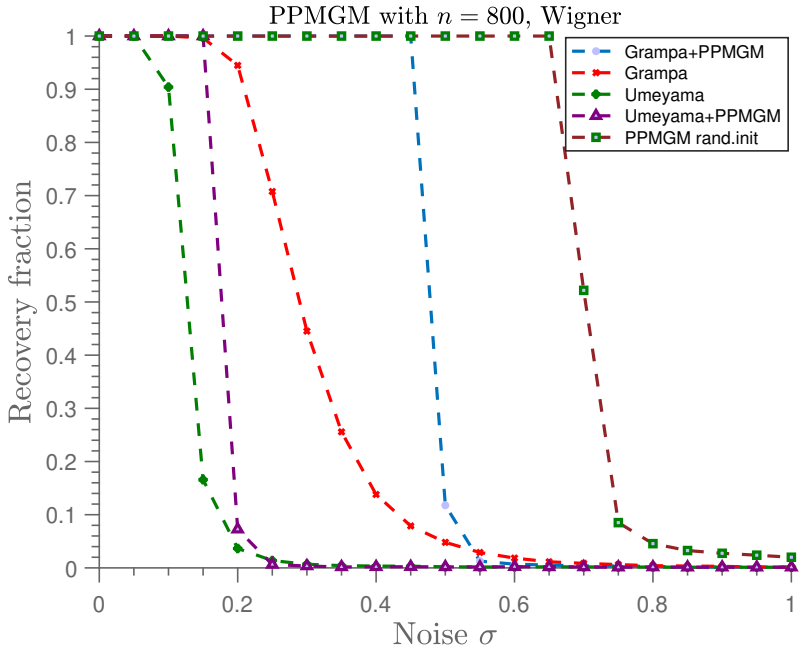

Figure 21: Performance of PPMGM as a refinement of Grampa and Umeyama algorithms, compared with PPM with a random initialization $x^{(0)}$, such that $\text{overlap}(x^{(0)}, x^*) = 0.08$.

runs. In Figure 22 we plot the performance of the PPMGM algorithm for randomly chosen seeds and with different number of correctly pre-matched vertices. More specifically, we consider an initial permutation $x_j^{(0)} \in \mathcal{S}_n$ (corresponding to initializations $X_j^{(0)} \in \mathcal{P}_n$) for $j = 1, \cdots, 4$ with $\text{overlap}(x_1^{(0)}, x^*) = 0.05$, $\text{overlap}(x_2^{(0)}, x^*) = 0.1$, $\text{overlap}(x_3^{(0)}, x^*) = 0.15$ and $\text{overlap}(x_4^{(0)}, x^*) = 0.5$. Equivalently, we have $\|X_j^{(0)} - X^*\|_F = \theta_j' \sqrt{n}$, where $\theta_j' = \sqrt{2\big(1 - \text{overlap}(x_j^{(0)}, x^*)\big)}$. Each permutation $x_j^{(0)}$ is chosen uniformly at random in the subset of permutations that satisfy each overlap condition. We observe that initializing the algorithm with an overlap of 0.1 with the ground truth permutation already produces perfect recovery in one iteration for levels of noise as high as $\sigma = 0.6$.

VARYING THE NUMBER OF ITERATIONS N. We experimentally evaluate the performance of PPMGM when varying the number of iterations N in Algorithm 12. In Figures 23 and 24 we plot the recovery rate of PPMGM, initialized with $x^{(0)}$, with an overlap of 0.1 with the ground truth. In Figure 23 we see that adding more iterations increases the performance of the algorithm for $n = 500$; however the improvement is less pronounced in the higher noise regime. In other words, the number of iterations cannot make up for the fact that the initial seed is of poor quality (relative to the noise level). We use $N = 1, 2, 4, 8, 30$ iterations and we observe a moderate gain between

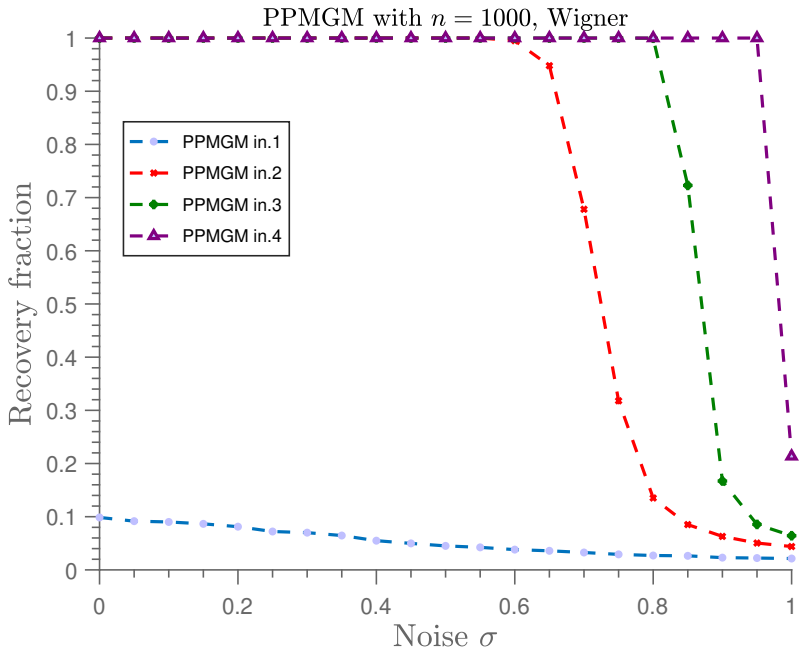

Figure 22: Performance of PPMGM with different initializations. Here in.1, in.2, in.3, in.4 corresponds to and overlap of $x^{(0)}$ with the ground truth of 0.05, 0.1, 0.15 and 0.5 respectively.

$N = 8$ and $N = 30$. In Figure 24 we use a matrix of size $n = 1000$ and we see that the difference between using $N = 1$ and $N > 1$ is even less pronounced (we omit the case of 30 iterations for readability purposes, as it is very similar to $N = 8$).

7.5.2 *Sparsification strategies*

Here we run PPMGM using different input matrices which are all transformations of the Wigner correlated matrices A, B. Specifically, we compare PPMGM with A, B as input with the application of PPMGM to three different pairs of input matrices (A^{spar_1}, B^{spar_1}), (A^{spar_2}, B^{spar_2}) and (A^{spar_3}, B^{spar_3}) that are defined as follows.

$$
\begin{aligned}
A_{ij}^{spar_1} &= \mathbb{1}_{|A_{ij}|<\tau}; \;\; B_{ij}^{spar_1} = \mathbb{1}_{|B_{ij}|<\tau}, \\
A_{ij}^{spar_2} &= A_{ij}\mathbb{1}_{|A_{ij}|<\tau}; \;\; B_{ij}^{spar_2} = B_{ij}\mathbb{1}_{|B_{ij}|<\tau}, \\
A_{ij}^{spar_3} &= A_{ij}\mathbb{1}_{|A_{ij}|\in top_k(A_{i:})}; \;\; B_{ij}^{spar_2} = B_{ij}\mathbb{1}_{|B_{ij}|\in top_k(B_{i:})},
\end{aligned}
$$

where $\tau > 0$ and for $k \in \mathbb{N}$ and a $n \times n$ matrix M, $top_k(M_{i:})$ is the set of the k largest elements (breaking ties arbitrarily) of $M_{i:}$ (the i-th row of M). The choice of the parameter τ is mainly determined by the sparsity assumptions in (Mao et al., 2021a, Thm.B), *i.e.*, if G, H are two Erdös-Renyi graphs to be matched with connection probability p

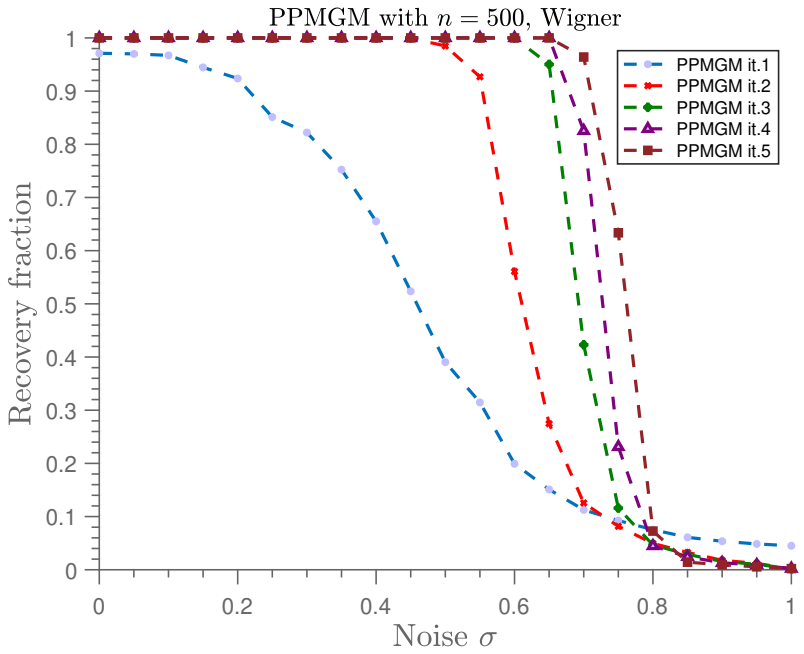

Figure 23: PPMGM with an initialization such that overlap$(x^{(0)}, x^*) = 0.1$. Here it.1, it.2, it.3, it.4, it.5 corresponds to 1, 2, 4, 8 and 30 iterations respectively.

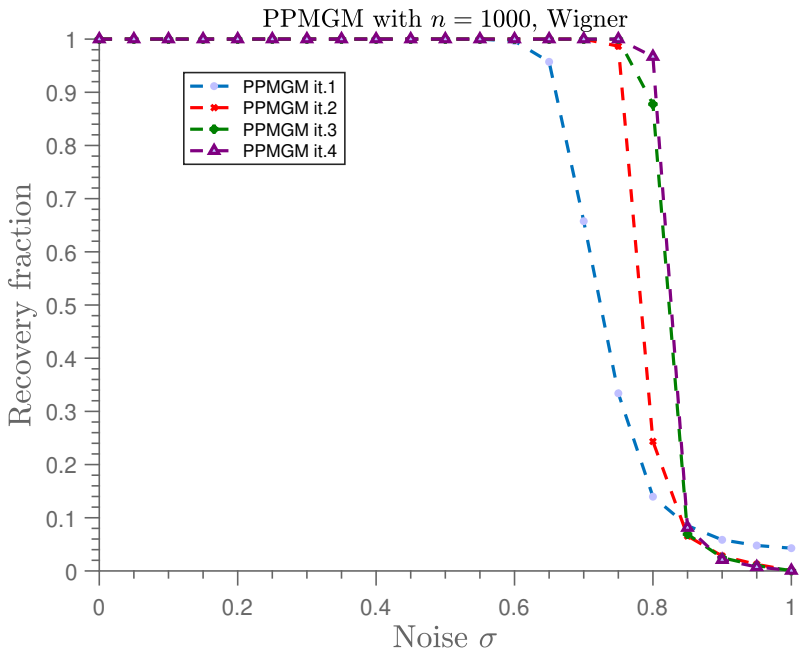

Figure 24: Here it.1, it.2, it.3, it.4 corresponds to 1, 2, 4 and 8 iterations respectively.

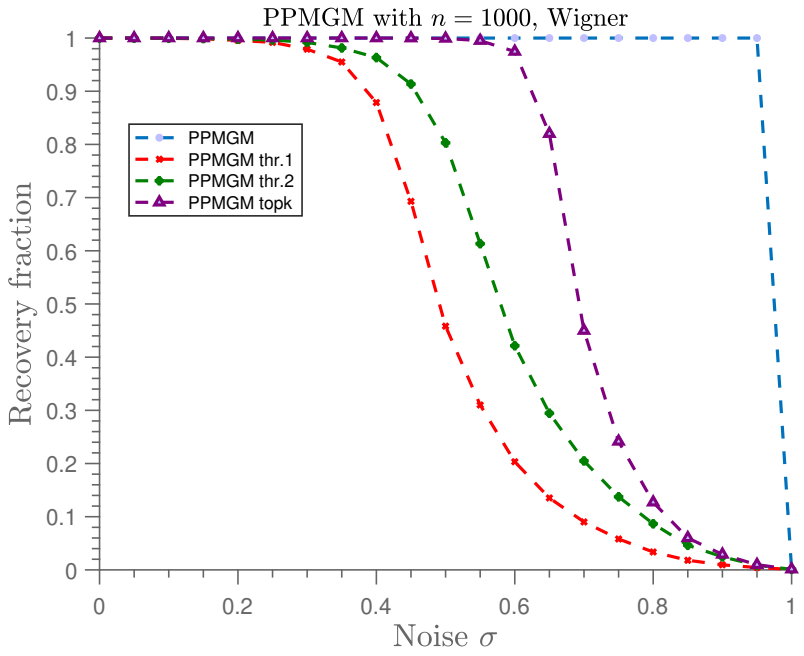

Figure 25: Initial overlap is equal to 0.5

(which is equal to q_s in the definition (1.5.3)), then the assumption is that

$$(1+\epsilon)\frac{\log n}{n} \leqslant p \leqslant n^{\frac{1}{R\log\log n}-1} \tag{7.5.1}$$

where $\epsilon > 0$ is arbitrary and R is an absolute constant. We refer the reader to (Mao et al., 2021a) for details. For each p in the range defined by (7.5.1) we solve the equation

$$\mathbb{P}(|A_{ij}| \leqslant \tau_p) = 2\Phi(-\tau_p\sqrt{n}) = p \tag{7.5.2}$$

where Φ is the standard Gaussian cdf (which is bijective so τ_p is well defined). In our experiments, we solve (7.5.2) numerically. Notice that A^{spar_1} and B^{spar_1} are sparse correlated Erdös-Renyi graphs with a correlation that depends on σ. For the value of k that defines A^{spar_3}, B^{spar_3} we choose $k = \Omega(\log n)$ or $k = \Omega(n^{o(1)})$, to maintain the sparsity degree in (7.5.1). In Figures 25 and 26 we plot the performance comparison between the `PPMGM` without sparsification, and the different sparsification strategies (initialized with overlap 0.5 and 0.1). We see that the use of the full information A, B outperforms the sparser versions in the higher noise regimes and for small overlap of the initialization. On the other hand, the performance tends to be more similar for low levels of noise and moderately large number of correct initial seeds. In theory, sparsification strategies have a moderate denoising effect (and might considerably speed up computations), but this process seems to destroy important correlation information.

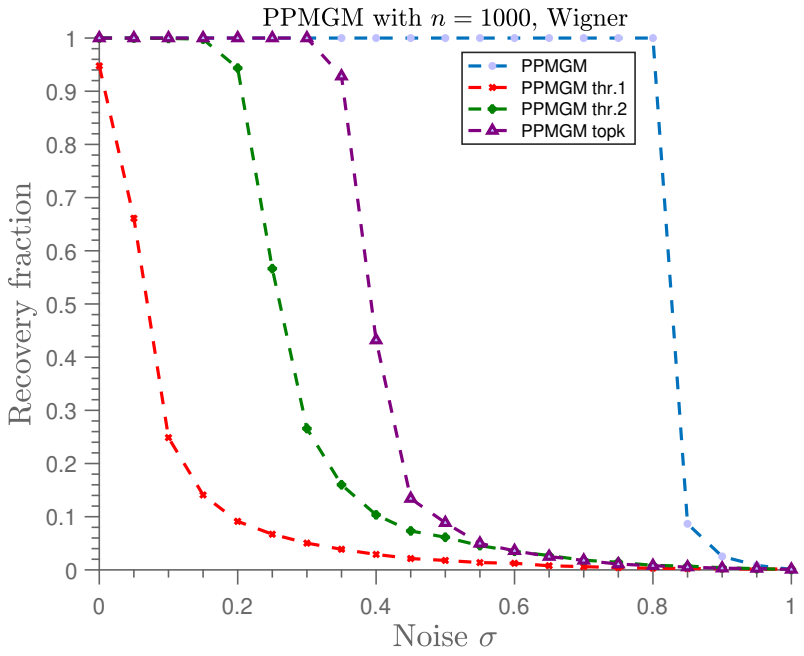

Figure 26: Initial overlap is equal to 0.1

7.5.2.1 *Choice of the sparsification parameter* τ

Solving (7.5.2) for p in the range (7.5.1) we obtain a range of possible values for the sparsification parameter τ. To choose between them, we use a simple grid search where we evaluate the recovery rate for each sparsification parameter on graphs of size $n = 1500$, and take the mean over 25 independent Monte Carlo runs. In Figure 27, we plot a heatmap with the results. We see that the best performing parameter in this experiment was for τ_5 corresponding to a probability $p_5 = 51 \times 10^{-3}$, although there is a moderate change between all the choices for p.

7.6 CONCLUDING REMARKS AND FUTUR WORK

In this work, we analysed the performance of the projected power method (proposed in Onaran et al., 2017) as a seeded graph matching algorithm, in the correlated Wigner model. We proved that for a non-data dependent seed with $\mathcal{O}(\sqrt{n \log n})$ correctly pre-assigned vertices, the PPM exactly recovers the ground truth matching in one iteration. This is analogous to the state-of-the-art results for algorithms in the case of relatively sparse correlated Erdös-Renyi graphs. We additionally proved that the PPM can exactly recover the optimal matching in $\mathcal{O}(\log n)$ iterations for a seed that contains $\Omega\big((1-\kappa)n\big)$ correctly matched vertices, for a constant $\kappa \in (0, 1)$, even if the seed can potentially be dependent on the data. For the latter result, we extended the arguments of Mao et al., 2021a from the (sparse) correlated

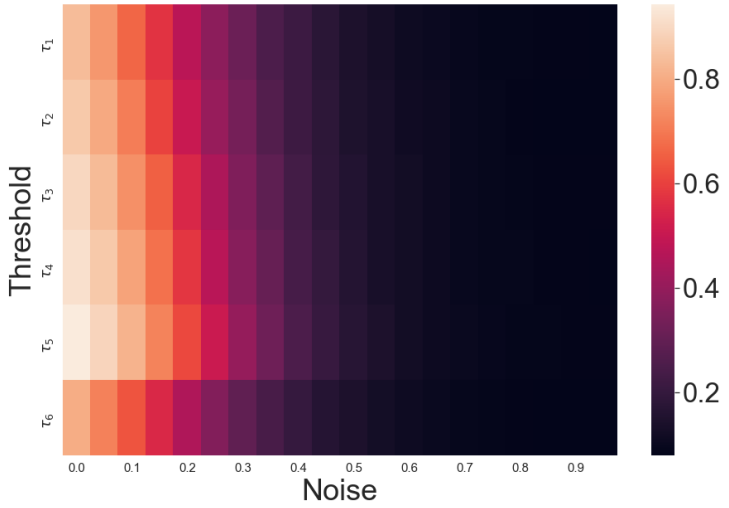

Figure 27: Heatmap for the recovery rate of `PPMGM` algorithm with input (A^{spar_1}, B^{spar_1}) for different threshold values τ_i(y axis); $i = 1, \cdots, 6$, and different values of σ (x axis). Here τ_i corresponds to the solution of (7.5.2) with $n = 1500$ and p_i for $i = 1, 2 \cdots, 6$ in a uniform grid between $p_1 = 42 \times 10^{-3}$ and $p_6 = 54 \times 10^{-3}$.

Erdös-Renyi model to the (dense) correlated Wigner case, providing a uniform control on the error when the seed contains $\Omega\big((1-\kappa)n\big)$ fixed points. This provides theoretical guarantees for the use of PPM as a refinement algorithm (or a post-processing step) for other seedless graph matching methods.

An open question is to find an efficient initialization method which outputs a permutation with order $(1-\kappa)n$ correctly matched vertices in regimes with higher σ (say for $\sigma > 1/2$). For those noise levels, spectral methods do not seem to perform well (at least in the experiments). An idea could be to adapt the results Mao et al., 2021a from the sparse Erdös-Renyi case to the Wigner case. In that paper, the authors construct for each vertex a signature containing the neighborhood information of that vertex and which is encoded as tree. Then a matching is constructed by matching those trees. It is however unclear how to adapt those results (which heavily rely on the sparsity) to the dense Wigner case.

It would also be interesting to compare in numerical experiments the performance of the algorithm proposed by Ma et al., 2021 to other state-of-the-art methods for graph matching to see if the main contribution of Mao et al., 2021a is on the theoretical analysis or if their work can also lead to an efficient algorithm in practice. We are currently working on the implementation of this algorithm and possible improvements.

7.7 ADDITIONAL PROOFS

7.7.1 *Proof of Proposition 4*

We divide the proof into two subsections. In Appendix 7.7.1.1 we prove Lemma 23 and in Appendix 7.7.1.2 we prove part (ii) of Proposition 4. Before proceeding, let us introduce and recall some notation. Define $C' := AXA$ and $C'' := AXZ$, then $C = AXB = \sqrt{1-\sigma^2}C' + \sigma C''$. Recall that for a permutation x, S_X will denote the set of fixed points of x (the set of non-zero diagonal terms of its matrix representation X) and we will often write $s_x = |S_X|/n = \mathrm{Tr}(X)/n$. We will say that a real random variable $Y \sim \chi^2_K$ if it follows a central Chi-squared distribution with K degrees of freedom.

7.7.1.1 *Proof of Lemma 23*

The proof of Lemma 23 mainly revolves around the use of concentration inequalities for quadratic forms of Gaussian random vectors. For that, it will be useful to use the following representation of the entries of C.

$$C_{ij} = \langle A_{:i}, XA_{:j} \rangle \tag{7.7.1}$$

where we recall that $A_{:k}$ represents the k-th column of the matrix A.

Proof of Lemma 23.

HIGH PROBABILITY BOUND FOR C_{ii}. Define $\tilde{a}_i$ to be a vector in $\mathbb{R}^n$ such that

$$\tilde{a}_i(k) = \begin{cases} A_{ki}, \text{ for } k \notin i, x^{-1}(i), \\ \frac{1}{\sqrt{2}} A_{ii}, \text{ for } k \in i, x^{-1}(i). \end{cases}$$

Using representation (7.7.1) we have

$$C_{ii} = \langle \tilde{a}_i, X\tilde{a}_i \rangle + Z_i$$

where

$$Z_i := \frac{1}{2} A_{ii} \big(A_{x(i)i} + A_{x^{-1}(i)i} \big).$$

It is easy to see that $\sqrt{n}\tilde{a}_i$ is a standard Gaussian vector. Using Lemma 38 we obtain

$$n\langle \tilde{a}_i, X\tilde{a}_i \rangle \overset{d}{=} \sum_{i=1}^{n_1} \mu_i g_i^2 - \sum_{i=1}^{n_2} \nu_i g_i'^2$$

where $(\mu_i)_{i=1}^{n_1}, (-\nu_i)_{i=1}^{n_2}$, (with $\mu_i \geqslant 0$, $\nu_i \geqslant 0$ and $n_1 + n_2 = n$) is the sequence of eigenvalues of $\frac{1}{2}(X + X^T)$ and $g = (g_1, \cdots, g_{n_1})$, $g' = (g'_1, \cdots, g'_{n_2})$ are two independent sets of i.i.d standard Gaussians.

Lemma 38 tell us in addition that $\|\mu\|_1 - \|\nu\|_1 = s_x n$, $\|\mu\|_2 + \|\nu\|_2 \leqslant \sqrt{2n}$ and $\|\mu\|_\infty, \|\nu\|_\infty \leqslant 1$. Using Corollary 3 (A.0.2), we obtain

$$\mathbb{P}(n\langle \tilde{a}_i, X\tilde{a}_i\rangle \leqslant s_x n - 2\sqrt{2nt} - 2t) \leqslant e^{-t} \tag{7.7.2}$$

for all $t \geqslant 0$. To obtain a concentration bound for $\mathcal{Z}_i$ we will distinguish two cases.
(a) *Case* $i \in S_X$. In this case, we have $\mathcal{Z}_i = a_i^2(i)$, which implies that $C_{ii} \geqslant \langle \tilde{a}_i, X\tilde{a}_i\rangle$. Hence

$$\mathbb{P}(nC_{ii} \leqslant s_x n - 2\sqrt{2nt} - 2t) \leqslant 2e^{-t}.$$

Replacing $t = \bar{t} := \frac{n}{2}(\sqrt{1 + \frac{s_x}{2}} - 1)^2$ in the previous expression, one can verify[2] that $\bar{t} \geqslant \frac{n}{48}s_x^2$, for $s_x \in (0, 1]$, hence

$$\mathbb{P}(C_{ii} \leqslant s_x/2) \leqslant 2e^{-\frac{s_x^2}{48}n}$$

which proves (7.4.2) in this case.
(b) *Case* $i \notin S_X$. Notice that in this case, $a_i(i)$ is independent from $(a_i(x(i)) + a_i(x^{-1}(i))$, hence $n\mathcal{Z}_i \overset{d}{=} g_1 g_2$, where g_1, g_2 are independent standard Gaussians. Using the polarization identity $g_1 g_2 = \frac{1}{4}(g_1 + g_2)^2 - \frac{1}{4}(g_1 - g_2)^2$, we obtain

$$n\mathcal{Z}_i \overset{d}{=} \frac{1}{2}(\tilde{g}_1^2 - \tilde{g}_2^2)$$

where $\tilde{g_1}, \tilde{g_2}$ are independent standard Gaussians. By Corollary 3 we have

$$\mathbb{P}\left(2n\mathcal{Z}_i \leqslant -4\sqrt{t} - 2t\right) \leqslant 2e^{-t}. \tag{7.7.3}$$

Using (7.7.2) and (7.7.3), we get

$$\mathbb{P}(nC_{ii} \leqslant s_x n - 2(\sqrt{2n} + 1)\sqrt{t} - 3t) \leqslant 4e^{-t}$$

or, equivalently

$$\mathbb{P}\left(C_{ii} \leqslant s_x - 2(\sqrt{2} + 1/\sqrt{n})\sqrt{\frac{t}{n}} - 3\frac{t}{n}\right) \leqslant 4e^{-t}. \tag{7.7.4}$$

Replacing $t = \bar{t} := \frac{n}{36}\left(\sqrt{d^2 + 6s_x} - d\right)^2$, where $d = 2(\sqrt{2} + 1/\sqrt{n})$, in the previous expression and noticing that $\bar{t} \geqslant \frac{1}{6}s_x^2 n$, we obtain the bound

$$\mathbb{P}(C_{ii} \leqslant s_x/2) \leqslant 4e^{-\frac{s_x^2}{6}n}.$$

2 Indeed, the inequality $(\sqrt{1+x} - 1)^2 \geqslant \frac{1}{6}x^2$, follows from the inequality $x^2 + (2\sqrt{6} - 6)x \leqslant 0$, which holds for $0 < x \leqslant 1$.

HIGH PROBABILITY BOUND FOR C_{ij}, $i \neq j$. Let us first define the vectors $\tilde{a}_i, \tilde{a}_j \in \mathbb{R}^n$ as

$$\tilde{a}_i(k) := \begin{cases} A_{ki}, & \text{for } k \notin \{j, x^{-1}(i)\}, \\ 0, & \text{for } k \in \{j, x^{-1}(i)\}, \end{cases}$$

and

$$\tilde{a}_j(k) := \begin{cases} A_{kj}, & \text{for } k \notin \{j, x^{-1}(i)\}, \\ 0, & \text{for } k \in \{j, x^{-1}(i)\}. \end{cases}$$

Contrary to a_i and a_j which share a coordinate, the vectors $\tilde{a}_i$ and $\tilde{a}_j$ are independent. With this notation, we have the following decomposition

$$C_{ij} = \langle \tilde{a}_i, X\tilde{a}_j \rangle + A_{ji}\Big(A_{x(j)j} + A_{x^{-1}(i)i}\Big).$$

For the first term, we will use the following polarization identity

$$\langle \tilde{a}_i, X\tilde{a}_j \rangle = \|\frac{1}{2}(\tilde{a}_i + X\tilde{a}_j)\|^2 - \|\frac{1}{2}(\tilde{a}_i - X\tilde{a}_j)\|^2. \tag{7.7.5}$$

By the independence of $\tilde{a}_i$ and $\tilde{a}_j$, it is easy to see that $\tilde{a}_i + X\tilde{a}_j$ and $\tilde{a}_i - X\tilde{a}_j$ are independent Gaussian vectors and $\mathbb{E}[\langle \tilde{a}_i, X\tilde{a}_j \rangle] = 0$. Using (7.7.5) and defining $Z_{ij} := A_{ji}\Big(A_{x(j)j} + A_{x^{-1}(i)i}\Big)n$, it is easy to see that

$$nC_{ij} \overset{d}{=} \sum_{i=1}^{n-1} \mu_i g_i^2 - \sum_{i=1}^{n-1} \nu_i g_i'^2 + Z_{ij} \tag{7.7.6}$$

where $g_1, \cdots, g_{n-1}$ and $g_1', \cdots, g_{n-1}'$ are two sets of independent standard Gaussian variables and $\mu_i, \nu_i \in \{\frac{1}{2}, \frac{3}{4}, 1\}$, for $i \in [n-1]$. The sequences $(\mu_i)_{i=1}^{n-1}, (\nu_i)_{i=1}^{n-1}$ will be characterised below, when we divide the analysis into two cases $x(j) = i$ and $x(j) \neq i$. We first state the following claim about Z_{ij}.

Claim 2. *For $i \neq j$, we have*

$$Z_{ij} \overset{d}{=} \begin{cases} q_{ij}(\zeta_1 - \zeta_2) & \text{if } x(j) \neq i, \\ 2\zeta_3 & \text{if } x(j) = i, \end{cases}$$

where ζ_1, ζ_2 and ζ_3 are independent Chi-squared random variables with one degree of freedom and

$$q_{ij} = \begin{cases} \sqrt{\frac{3}{2}} & \text{if } i \in S_X, j \notin S_X \text{ or } i \notin S_X, j \in S_X, \\ \sqrt{2} & \text{if } i, j \in S_X, \\ \frac{1}{\sqrt{2}} & \text{if } i, j \notin S_X. \end{cases}$$

We delay the proof of this claim until the end of this section. From the expression (7.7.6), we deduce that the vectors $g = (g_1, \cdots, g_{n-1})$,

$g' = (g'_1, \cdots, g'_{n-1})$ and $\mathcal{Z}_{ij}$ are independent. Hence, by Claim 2 the following decomposition holds

$$nC_{ij} \stackrel{d}{=} \sum_{i=1}^{n} \mu_i g_i^2 - \sum_{i=1}^{n} \nu_i g_i'^2$$

where

$$\mu_n = \begin{cases} q_{ij} \text{ if } x(j) \neq i, \\ 2 \text{ if } x(j) = i, \end{cases} \quad \text{and } \nu_n = \begin{cases} q_{ij} \text{ if } x(j) \neq i, \\ 0 \text{ if } x(j) = i. \end{cases}$$

Let us define $\mu := (\mu_1, \cdots, \mu_n)$ and $\nu := (\nu_1, \cdots, \nu_n)$. We will now distinguish two cases.

(a) Case $x(j) \neq i$. In this case, we can verify that one of the $\mu_1, \cdots, \mu_{n-1}$ is equal to 0 (and the same is true for the values $\nu_1, \cdots, \nu_{n-1}$). Assume without loss of generality that $\mu_1 = \nu_1 = 0$. Also, one of the following situations must happen for the sequence $\mu_2, \cdots, \mu_{n-1}$ (resp. $\nu_2, \cdots, \nu_{n-1}$): either $n-3$ of the elements of the sequence are equal to $\frac{1}{2}$ and one is equal 1 or $n-4$ are equal to $\frac{1}{2}$ and two are equal to $\frac{3}{4}$ or $n-3$ are equal to $\frac{1}{2}$ and one is equal to $\frac{3}{4}$. In either of those cases, the following is verified

$$\begin{aligned} \|\mu\|_1 - \|\nu\|_1 &= 0, \\ \|\mu\|_2 + \|\nu\|_2 &\leqslant \sqrt{2n}, \\ \|\mu\|_\infty, \|\nu\|_\infty &\leqslant \sqrt{2}, \end{aligned}$$

where the first equality comes from Lemma 22, the inequality on the norm $\|\cdot\|_2$ comes from the fact that in the worst case $\|\mu\|_2 = \|\nu\|_2 \leqslant \sqrt{\frac{n+1}{4}}$. The statement about the norm $\|\cdot\|_\infty$ can be easily seen by the definition of μ and ν. Using (A.0.1), we obtain

$$\mathbb{P}(nC_{ij} \geqslant 4\sqrt{nt} + 4t) \leqslant 2e^{-t}.$$

Replacing $t = \bar{t} := \frac{n}{4}(\sqrt{1 + \frac{s_x}{2}} - 1)^2$ in the previous expression and noticing that $\bar{t} \geqslant \frac{1}{96} s_x^2 n$ for $s_x \in (0, 1]$ leads to the bound

$$\mathbb{P}(C_{ij} \geqslant s_x/2) \leqslant 2e^{-\frac{s_x^2}{96} n}.$$

(b) Case $x(j) = i$. In this case, we have that for the sequence $\mu_1, \cdots, \mu_{n-1}$ (resp. $\nu_1, \cdots, \nu_{n-1}$): either $n-2$ of the elements of the sequence are equal to $\frac{1}{2}$ and one is equal 1 or $n-3$ are equal to $\frac{1}{2}$ and two are equal to $\frac{3}{4}$. In either case, the following holds

$$\begin{aligned} \|\mu\|_1 - \|\nu\|_1 &= 2, \\ \|\mu\|_2 + \|\nu\|_2 &\leqslant 2\sqrt{n}, \\ \|\mu\|_\infty, \|\nu\|_\infty &\leqslant 2. \end{aligned}$$

Here, the inequalities for the norms $\|\cdot\|_1, \|\cdot\|_\infty$ follow directly from the definition of μ and ν, and the inequality for $\|\cdot\|_2$ follows by the fact that, in the worst case, $\|\mu\|_2 + \|\nu\|_2 = \sqrt{\frac{n+6}{4}} + \sqrt{\frac{n+2}{4}}$. Using (A.0.1), we get

$$\mathbb{P}(nC_{ij} \geqslant 2 + 4\sqrt{nt} + 4t) \leqslant 2e^{-t}.$$

Replacing $t = \bar{t} := \frac{n}{4}(\sqrt{1 + \frac{s_x}{2} - \frac{2}{n}} - 1)^2$ in the previous expression and noticing that $\bar{t} \geqslant \frac{1}{20}s_x^2 n$ for $s_x \in (10/n, 1]$ we get

$$\mathbb{P}(C_{ij} \geqslant s_x/2) \leqslant 2e^{-\frac{s_x^2}{20} + 4/n} \leqslant 3e^{-\frac{s_x^2}{20}},$$

where we used that $n \geqslant 10$. □

Proof of Claim 2. Observe that when $x(j) = i$ (or equivalently $x^{-1}(i) = j$) we have $Z_{ij} = 2nA_{ij}^2$. Given that $i \neq j$ by assumption, it holds $A_{ij}^2 \sim \mathcal{N}(0, \frac{1}{n})$, which implies that $Z_{ij} \overset{d}{=} 2\zeta_3$ for $\zeta_3 \sim \chi_1^2$. In the case $x(j) \neq i$, let us define

$$\psi_1 := \sqrt{n}A_{ij}, \quad \psi_2 := \sqrt{n}A_{jx(j)}, \quad \psi_3 := \sqrt{n}A_{ix^{-1}(i)},$$

which are all independent Gaussians random variables. Moreover, $\psi_1 \sim \mathcal{N}(0,1)$ and

$$\psi_2 + \psi_3 \sim \begin{cases} \mathcal{N}(0,2) \text{ if } i,j \notin S_X, \\ \mathcal{N}(0,3) \text{ if } i \in S_X, j \notin S_X \text{ or } i \notin S_X, j \in S_X, \\ \mathcal{N}(0,4) \text{ if } i,j \in S_X. \end{cases}$$

Consider the case $i, j \notin S_X$. In this case, it holds

$$Z_{ij} = \sqrt{2}\psi_1\Big(\frac{\psi_2 + \psi_3}{\sqrt{2}}\Big) = \frac{1}{\sqrt{2}}\Big(\frac{\psi_1}{\sqrt{2}} + \frac{\psi_2 + \psi_3}{2}\Big)^2 - \frac{1}{\sqrt{2}}\Big(\frac{\psi_1}{\sqrt{2}} - \frac{\psi_2 + \psi_3}{2}\Big)^2.$$

Notice that $\frac{\psi_1}{\sqrt{2}} + \frac{\psi_2+\psi_3}{2}$ and $\frac{\psi_1}{\sqrt{2}} - \frac{\psi_2+\psi_3}{2}$ are independent standard normal random variables, hence $Z_{ij} \overset{d}{=} \frac{1}{\sqrt{2}}(\zeta_1 - \zeta_2)$, where ζ_1 and ζ_2 are independent χ_1^2 random variables. The proof for the other cases is analogous. □

7.7.1.2 *Proof of Proposition 4 part* (ii)

Now we consider the case where $\sigma \neq 0$. It is easy to see that here the analysis of the noiseless case still applies (up to re-scaling by $\sqrt{1-\sigma^2}$) for the matrix $C' = AXA$. We can proceed in an analogous way for the matrix $C'' = AXZ$ which will complete the analysis (recalling that $C = \sqrt{1-\sigma^2}C' + \sigma C''$).

Before we proceed with the proof, we explain how the tail analysis of entries of C' in Prop.4 part (i) helps us with the tail analysis of C''. Observe that for each $i, j \in [n]$ we have

$$C''_{ij} = \sum_{k,k'} A_{ik} X_{k,k'} Z_{k',j} = \sum_{k=1}^{n} A_{ik} Z_{\chi(k)j} = \langle A_{:i}, XZ_{:j} \rangle.$$

The term C''_{ij}, for all $i, j \in [n]$, can be controlled similarly to the term $C'_{i'j'}$ (when $i' \neq j'$). Indeed, we have the following

Lemma 30. *For $t \geqslant 0$ we have*

$$\mathbb{P}(C''_{ij} \leqslant -4\sqrt{nt} - 2t) = \mathbb{P}(C''_{ij} \geqslant 4\sqrt{nt} + 2t) \leqslant 2e^{-t}.$$

Consequently,

$$\mathbb{P}(C''_{ij} \geqslant s_x/2) \leqslant 2e^{-\frac{s_x^2}{96}n}.$$

Proof. We define $h_1 := \frac{1}{2}(A_{:i} + XZ_{:j})$ and $h_2 := \frac{1}{2}(A_{:i} - XZ_{:j})$. It is easy to see that h_1 and h_2 are two i.i.d Gaussian vectors of dimension n. By the polarization identity, we have

$$n\langle A_{:i}, XZ_{:j} \rangle = n(\|h_1\|^2 - \|h_2\|^2) \stackrel{d}{=} \sum_{i=1}^{n} \mu_i g_i^2 - \sum_{i=1}^{n} \nu_i g_i'^2$$

where $g = (g_1, \cdots, g_n)$ and $g' = (g'_1, \cdots, g'_n)$ are independent standard Gaussian vectors and the vectors $\mu = (\mu_1, \cdots, \mu_n), \nu = (\nu_1, \cdots, \nu_n)$ have positive entries that satisfy, for all $i \in [n]$, $\mu_i, \nu_i \in \{\frac{1}{\sqrt{2}}, \sqrt{\frac{3}{4}}, 1\}$. For μ_i (and the same is true for ν_i) the following two cases can happen: either $n-1$ of its entries are $1/\sqrt{2}$ and one entry takes the value 1 (when $i = j$) or $n-2$ of its entries are $1/\sqrt{2}$ and two entries take the value $\sqrt{3/4}$ (when $i \neq j$). In any of those cases, one can readily see that

$$\|\mu\|_1 = \|\nu\|_1, \;\; \|\mu\|_2 + \|\nu\|_2 \leqslant \sqrt{n}, \;\; \|\mu\|_\infty, \|\nu\|_\infty \leqslant 1.$$

Using Corollary 3 we obtain

$$\mathbb{P}\big(n(\|h_1\|^2 - \|h_2\|^2) \geqslant 4\sqrt{nt} + 2t\big) \leqslant 2e^{-t},$$
$$\mathbb{P}\big(n(\|h_1\|^2 - \|h_2\|^2) \leqslant -4\sqrt{nt} - 2t\big) \leqslant 2e^{-t}.$$

Arguing as in the proof of Proposition 4 part (i) we obtain the bound

$$\mathbb{P}(C''_{ij} \geqslant s_x/2) \leqslant 2e^{-\frac{s_x^2}{96}n}.$$

□

Now we introduce some definitions that will be used in the proof. We define $s_{\sigma,x} := \frac{1}{2}\sqrt{1-\sigma^2}s_x$, and for $\delta > 0$, $i, j \in [n]$, we define the following events

$$\mathcal{E}_\delta^i := \{\sqrt{1-\sigma^2}C'_{ii} \leqslant s_{\sigma,x} + \delta\} \cup \{\sigma C''_{ii} \leqslant -\delta\},$$

$$\mathcal{E}^{ij} := \{\sqrt{1-\sigma^2}C'_{ij} \geqslant s_{\sigma,x}/2\} \cup \{\sigma C''_{ij} \geqslant s_{\sigma,x}/2\} \text{ , for } i \neq j.$$

One can easily verify that $\{C_{ii} \leqslant s_{\sigma,x}\} \subset \mathcal{E}^i_\delta$, hence it suffices to control the probability of $\mathcal{E}^i_\delta$. For that we use the union bound and the already established bounds in Lemmas 23 and 30. To attack the off-diagonal case, we observe that the following holds $\{C_{ij} \geqslant s_{\sigma,x}\} \subset \mathcal{E}^{ij}$. The following lemma allows us to bound the probability of the events $\mathcal{E}^i_\delta$ and $\mathcal{E}^{ij}$.

Lemma 31. *Let δ be such that $0 \leqslant \delta \leqslant \frac{s_x}{2}\sqrt{1-\sigma^2}$. Then for $i, j \in [n]$ with $i \neq j$ have the following bounds*

$$\mathbb{P}(\mathcal{E}^i_\delta) \leqslant 4e^{-\frac{1}{96}(\frac{s_x}{2} - \frac{\delta}{\sqrt{1-\sigma^2}})^2 n} + 2e^{-\frac{1}{96}(\frac{\delta}{\sigma})^2 n} \tag{7.7.7}$$

$$\mathbb{P}(\mathcal{E}^{ij}) \leqslant 4e^{-\frac{1}{384}s_x^2(\frac{1-\sigma^2}{\sigma^2} \wedge 1)n}. \tag{7.7.8}$$

In particular, we have

$$\mathbb{P}(\mathcal{E}^i_{\delta_{\sigma,x}}) \leqslant 6e^{-\frac{1}{384}s_x^2(\frac{1-\sigma^2}{1+2\sigma\sqrt{1-\sigma^2}})n} \tag{7.7.9}$$

where $\delta_{\sigma,x} = \frac{\sigma\sqrt{1-\sigma^2}}{\sigma+\sqrt{1-\sigma^2}}\frac{s_x}{2}$.

Proof. Using (7.7.4), we have that

$$\mathbb{P}\Big(\sqrt{1-\sigma^2}C_{ii} \leqslant \sqrt{1-\sigma^2}\big(s_x - 2(\sqrt{2}+1/\sqrt{n})\sqrt{\frac{t}{n}} - 3\frac{t}{n}\big)\Big) \leqslant 4e^{-t}.$$

Replacing $t = \bar{t} := \frac{n}{36}\Big(\sqrt{d^2 + 6s_x - \frac{12\delta}{\sqrt{1-\sigma^2}}} - d\Big)^2$ in the previous expression, where $d = 2(\sqrt{2}+1/\sqrt{n})$, and observing that $\bar{t} \geqslant \frac{1}{6}(\frac{s_x}{2} - \frac{\delta}{\sqrt{1-\sigma^2}})^2$, which is valid for $0 \leqslant \delta \leqslant \frac{s_x}{2}\sqrt{1-\sigma^2}$, we obtain

$$\mathbb{P}\Big(\sqrt{1-\sigma^2}C'_{ii} \leqslant s_{\sigma,x} + \delta\Big) \leqslant 4e^{-\frac{1}{6}(\frac{s_x}{2} - \frac{\delta}{\sqrt{1-\sigma^2}})^2 n}.$$

Using this and Lemma 30 we have

$$\begin{aligned}\mathbb{P}(\mathcal{E}^i_\delta) &\leqslant \mathbb{P}(\sqrt{1-\sigma^2}C'_{ii} \leqslant s_{\sigma,x} + \delta) + \mathbb{P}(\sigma C''_{ii} \leqslant -\delta) \\ &\leqslant 4e^{-\frac{1}{6}(\frac{s_x}{2} - \frac{\delta}{\sqrt{1-\sigma^2}})^2 n} + 2e^{-\frac{1}{96}(\frac{\delta}{\sigma})^2 n}.\end{aligned}$$

Similarly, to prove (7.7.8) we verify that

$$\begin{aligned}\mathbb{P}(\mathcal{E}^{ij}) &\leqslant \mathbb{P}(C'_{ij} \geqslant \frac{s_x}{4}) + \mathbb{P}(C''_{ij} \geqslant \frac{\sqrt{1-\sigma^2}}{\sigma}\frac{s_x}{4}) \\ &\leqslant 2e^{-\frac{1}{384}s_x^2 n} + 2e^{-\frac{1}{384}s_x^2(\frac{1-\sigma^2}{\sigma^2})n} \\ &\leqslant 4e^{-\frac{1}{384}s_x^2(\frac{1-\sigma^2}{\sigma^2} \wedge 1)n}.\end{aligned}$$

To prove (7.7.9) it suffices to use (7.7.7) with the choice of $\delta = \delta_{\sigma,x} = \frac{\sigma\sqrt{1-\sigma^2}}{\sigma+\sqrt{1-\sigma^2}}\frac{s_x}{2}$. □

With this we prove the diagonal dominance for each fixed row of C.

Proof of Prop. 4 part (ii). Define $\tilde{\mathcal{E}}_j := \{C_{ii} \leqslant s_{\sigma,x}\} \cup \{C_{ij} \geqslant s_{\sigma,x}\}$, which clearly satisfies $\{C_{ii} \leqslant C_{ij}\} \subset \tilde{\mathcal{E}}_j$. Then by the union bound,

$$\begin{aligned}
\mathbb{P}(\cup_{j\neq i}\tilde{\mathcal{E}}_j) &\leqslant \mathbb{P}(C_{ii} \leqslant s_{\sigma,x}) + \sum_{j\neq i} \mathbb{P}(C_{ij} \geqslant s_{\sigma,x}) \\
&\leqslant \mathbb{P}(\mathcal{E}^i_{\delta_{\sigma,x}}) + \sum_{j\neq i} \mathbb{P}(\mathcal{E}^{ij}) \\
&\leqslant 6e^{-\frac{1}{384}s_x^2(\frac{1-\sigma^2}{1+2\sigma\sqrt{1-\sigma^2}})n} + 4(n-1)e^{-\frac{1}{384}s_x^2(\frac{1-\sigma^2}{\sigma^2}\wedge 1)n} \\
&\leqslant 5ne^{-\frac{1}{384}s_x^2(\frac{1-\sigma^2}{1+2\sigma\sqrt{1-\sigma^2}})n}
\end{aligned}$$

where in the third inequality we used Lemma 31, and in the last inequality we used the fact that $\frac{1-\sigma^2}{\sigma^2} \wedge 1 \geqslant \frac{1-\sigma^2}{1+2\sigma\sqrt{1-\sigma^2}}$. □

7.7.2 *Proof of Lemma 24*

The proof of Lemma 24 uses elements of the proof of Proposition 4. The interested reader is invited to read the proof of Proposition 4 first.

Proof of Lemma 24. It will be useful to first generalize our notation. For that, we denote

$$C_{ij,x} = (AXB)_{ij}, \quad C'_{ij,x} = (AXA)_{ij}, \quad C''_{ij,x} = (AXZ)_{ij}$$

for $x \in \mathcal{S}_n$, and

$$\mathcal{E}^{ij}_{x^{-1}} := \{\sqrt{1-\sigma^2}C'_{ij,x^{-1}} \geqslant s_{\sigma,x}/2\} \cup \{\sigma C''_{ij,x^{-1}} \geqslant s_{\sigma,x}/2\}$$

where x^{-1} is the inverse permutation of x. The fact that $\mathbb{P}(C_{ii,x} < C_{ij,x}) \leqslant 8e^{-c(\sigma)s_x^2 n}$ follows directly from the bound for $\tilde{\mathcal{E}}_j$ derived in the proof of Proposition 4 part (ii). To prove $\mathbb{P}(C_{ii,x} < C_{ji,x})$ notice that $C'_{ji,x} = C'_{ij,x^{-1}}$ and that $C''_{ji,x} \overset{d}{=} C''_{ij,x^{-1}}$. On the other hand, notice that $s_x = s_{x^{-1}}$ (hence $s_{\sigma,x} = s_{\sigma,x^{-1}}$). Arguing as in Lemma 31 it is easy to see that

$$\mathbb{P}(C_{ii,x} < C_{ji,x}) \leqslant 8e^{-c(\sigma)s_x^2 n}.$$

The bound on $\mathbb{P}(\exists j, \text{ s.t } C_{ij,x} \vee C_{ji,x} > C_{ii,x})$ then follows directly by the union bound. □

7.8 PROOFS OF LEMMAS 20 AND 21

Proof of Lemma 20. By assumption C is diagonally dominant, which implies that $\exists i_1$ such that $C_{i_1 i_1} = \max_{i,j} C_{ij}$ (in other words, if the

largest entry of C is in the i_1-th row, then it has to be $C_{i_1 i_1}$, otherwise it would contradict the diagonal dominance of C). In the first step of GMWM we select $C_{i_1 i_1}$, assign $\pi(i_1) = i_1$ and erase the i_1-th row and column of C. By erasing the i_1-th row and column of C we obtain a matrix which is itself diagonally dominant. So by iterating this argument we see $\exists\, i_1, \cdots, i_n \subset [n]$ such that $\pi(i_k) = i_k$, for all k, so π has to be the identical permutation. This proves that if C is diagonally dominant, then $\Pi = \mathrm{Id}$. By using the contrareciprocal, (7.3.1) follows. □

Proof of Lemma 21. We argue by contradiction. Assume that for some $1 \leqslant k \leqslant r$, we have $\pi(i_k) \neq i_k$ (and $\pi^{-1}(i_k) \neq i_k$). This means that at some some step j the algorithm selects either $C^{(j)}_{i_k \pi(i_k)}$ or $C^{(j)}_{\pi^{-1}(i_k)\pi(i_k)}$ as the largest entry, but this contradicts the row-column dominance of i_k. This proves that that if there exists a set of indices $I_r \subset [n]$ of size r such that for all $i \in I_r$, C_{ii} is row-column dominant, then that set is selected by the algorithm, which implies that $\pi(i) = i$ for $i \in I_r$, thus $\mathrm{overlap}(\pi, \mathrm{id}) \geqslant r$. (7.3.3) follows by the contrareciprocal. □

7.8.1 *Concentration inequalities used in Theorem 19*

In this section we provide proofs of Lemma's 25 and 26 used to prove Theorem 19.

Proof of Lemma's 25 and 26. Recall that $B_{ij} = \sqrt{1-\sigma^2} A_{ij} + \sigma Z_{ij}$.

STEP 1. First let us consider the terms of the form $\langle A_{i:}, A_{i:} \rangle$. We can write

$$\langle A_{i:}, A_{i:} \rangle = \sum_{i=1}^{n-1} \mu_i g_i^2$$

where g_i are independent standard Gaussian random variables and $\mu_i = 1/n$ for all i. Observe that $\|\mu\|_2 = \sqrt{\frac{n-1}{n^2}}$. By Lemma 37 we have for $i \in [n]$ and all $t > 0$

$$\mathbb{P}\left(\langle A_{i:}, A_{i:} \rangle \leqslant \frac{n-1}{n} - 2\sqrt{\frac{t(n-1)}{n^2}} \right) \leqslant e^{-t}.$$

For the choice $t = 5 \log n$ we obtain

$$\langle A_{i:}, A_{i:} \rangle \geqslant 1 - O\left(\sqrt{\frac{\log n}{n}} \right)$$

with probability at least $1 - e^{-5 \log n}$.

STEP 2. Let us consider now terms of the form $\langle A_{i:}, Z_{i:}\rangle$. We can write

$$\langle A_{i:}, Z_{i:}\rangle = \frac{1}{n}\sum_{i=1}^{n-1}(g_i g_i') = \frac{1}{n}G^\top G'$$

where $G = (g_i)_{i=1}^{n-1}$ and $G' = (g_i')_{i=1}^{n-1}$ are i.i.d. standard Gaussian random variables. We can write

$$G^\top G' = \|G\|\left(\left(\frac{G}{\|G\|}\right)^\top G'\right).$$

Since G' is invariant by rotation $(\frac{G}{\|G\|})^\top G'$ is independent from G and has distribution $\mathcal{N}(0,1)$. By Gaussian concentration inequality we hence have

$$\left(\frac{G}{\|G\|}\right)^\top G' \leqslant C\sqrt{\log n}$$

with probability at least $1 - e^{-5\log n}$ for a suitable choice of C. Similarly, by Lemma 37 we have

$$\|G\| \leqslant 2\sqrt{n}$$

with probability at least $1 - e^{-5\log n}$. Hence with probability at least $1 - 2e^{-5\log n}$ we have

$$\frac{1}{n}G^\top G' \leqslant 2C\sqrt{\frac{\log n}{n}}.$$

STEP 3. The same argument can be used to show that for $i \neq j$

$$\mathbb{P}\left(\langle A_{i:}, A_{j:}\rangle \geqslant C\sqrt{\frac{\log n}{n}}\right) \leqslant e^{-5\log n}.$$

CONCLUSION. We can conclude by using the identity

$$B_{ij} = \sqrt{1-\sigma^2}A_{ij} + \sigma Z_{ij}$$

and taking the union bound over all indices $i \neq j$. □

Part III

APPENDIX

A

GENERAL CONCENTRATION INEQUALITIES

Lemma 32 (Chernoff bound - multiplicative form)**.** *Let* $X = \sum_{i \leqslant n} X_i$ *where* $X_i \overset{ind.}{\sim} \mathcal{B}(p)$*. Then*

$$\mathbb{P}(X \leqslant (1-\delta)np) \leqslant e^{-np\delta^2/2}$$

and

$$\mathbb{P}(X \geqslant (1+\delta)np) \leqslant e^{-np\delta^2/3}$$

for all $0 < \delta < 1$*.*

Proof. See Mitzenmacher et al., 2005, Theorem 4.5 and Corollary 4.6 . □

Lemma 33 (Hoeffding inequality)**.** *Let* $X_1, \dots, X_n \overset{ind.}{\sim} \mathcal{B}(p)$*. Then*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i\in[n]} X_i - p\right| \geqslant t\right) \leqslant 2\exp(-2nt^2).$$

Proof. See Boucheron et al., 2013, Theorem 2.8. □

Lemma 34 (Matrix Bernstein inequality)**.** *Let* $X_1, \dots, X_n$ *be a sequence of independent zero-mean random matrices of size* $d_1 \times d_2$*. Suppose that* $\|X_i\| \leqslant M$ *almost surely, for all* i*. Then for all positive* t*,*

$$\mathbb{P}(\|\sum_i X_i\| \geqslant t) \leqslant (d_1 + d_2)\exp\left(-\frac{t^2}{2\sigma^2 + 2M/3t}\right)$$

where $\sigma^2 = \max(\|\sum_i \mathbb{E}(X_i X_i^*)\|, \|\sum_i \mathbb{E}(X_i^* X_i)\|)$*.*

Proof. See Tropp, 2012, Theorem 1.6 □

Lemma 35. *Assume that* $A \sim \mathrm{SBM}(Z, \Pi)$*. Let* $E = A - \mathbb{E}(A)$*. Then with probability at least* $1 - n^{-\Omega(1)}$ *the following holds.*

1. $\|E\| \leqslant \sqrt{np_{max}}$,
2. $\|EW\|_F^2 \lesssim K^2 p_{max}$.

Proof. The first inequality is a classical result used for SBM in the relatively sparse regime $p_{max} = \omega(\log n)$. It can obtained as a consequence of Remark 3.13 in Bandeira et al., 2016. The second inequality follows from

$$\|EW\|_F^2 \leqslant K\|EW\|^2 \leqslant K\|E\|^2\|W\|^2 \lesssim K^2 p_{max}.$$

□

Lemma 36. *Assume that* $A \sim \mathrm{BiSBM}(Z_1, Z_2, \Pi)$*, and recall* $E = A - \mathbb{E}(A)$*. For* n_1 *large enough, the following holds true.*

1. $\mathbb{P}(\|E\| \lesssim \sqrt{n_2 p_{max}}) \geqslant 1 - n_1^{-\Omega(1)}$ *if* $n_2 p_{max} \gtrsim \log n_2$ *and* $n_2 \geqslant n_1$.
2. $\mathbb{P}(\|EE^\top - \mathbb{E}(EE^\top)\| \lesssim \max(\log n_1, \sqrt{n_1 n_2} p_{max})) \geqslant 1 - n_1^{-\Omega(1)}$. *Moreover, this event implies* $\left\|\mathcal{H}(EE^\top)\right\| \lesssim \max(\log n_1, \sqrt{n_1 n_2} p_{max}))$.
3. $\mathbb{P}(\|EZ_2\| \lesssim \sqrt{n_1 n_2 p_{max}/L}) \geqslant 1 - n_1^{-\Omega(1)}$ *if* $n_2 p_{max} \gtrsim L \log n_1$ *and* $\sqrt{\frac{n_1 n_2 p_{max}}{L}} \gtrsim \log n_1$.

Proof. 1. The first inequality follows from classical proof techniques. We first convert E into a square symmetric matrix as follows

$$\tilde{E} = \begin{pmatrix} 0 & E \\ E^\top & 0 \end{pmatrix}.$$

It is easy to verify that $\left\|\tilde{E}\right\| = \|E\|$. Since $\tilde{E}$ is a square symmetric matrix with independent entries, we can use the result of Remark 3.13 in Bandeira et al., 2016 and obtain with probability at least $1 - n_1^{-\Omega(1)}$ that

$$\left\|\tilde{E}\right\| \lesssim \sqrt{n_2 p_{max}}$$

since $n_2 p_{max} \gtrsim \log(n_1 + n_2)$. This condition is ensured since $n_2 \geqslant n_1$ and $n_2 p_{max} \gtrsim \log n_2$.

2. The second inequality follows from the recent work of Cai et al., 2022. First observe that

$$\|\mathcal{H}(EE^\top)\| \leqslant 2\|EE^\top - \mathbb{E}(EE^\top)\|$$

since $\|\mathcal{H}(A)\| \leqslant \|A\| + \|\operatorname{diag}(A)\| \leqslant 2\|A\|$ for all square matrices A and $\mathcal{H}(EE^\top) = \mathcal{H}(EE^\top - \mathbb{E}(EE^\top))$. Theorem 4 in Cai et al., 2022 shows that

$$\mathbb{E}(\|EE^\top - \mathbb{E}(EE^\top)\|) \lesssim \max(\sqrt{n_1 n_2} p_{max}, \log n_1).$$

It is however more difficult to get a high probability bound from this last inequality since we can no longer use Talagrand's inequality as in Bandeira et al., 2016. However, we can use the moments to obtain a tail bound as in Theorem 5 in Cai et al., 2022. This theorem is stated for matrices with Gaussian entries, but if instead of Lemma 1 we use Lemma 9 of Cai et al., 2022, we obtain a similar result for bounded sub-Gaussian entries. Since the variance parameters σ_R and σ_C that appear in the statement of Theorem 5 in Cai et al., 2022 satisfy $\sigma_R^2 \leqslant n_1 p_{max}$ and $\sigma_C^2 \leqslant n_2 p_{max}$ we obtain the result.

3. The last inequality can be obtained by using the proof techniques in Bandeira et al., 2016 as follows. In order to extend the concentration result from a matrix Y with independent standard Gaussian entries to a matrix X with symmetric sub-Gaussian entries, the key is to upper bound all the moments of X_{ij} by moment of Y_{ij}. This can be done by using the boundedness of Z_{ij} as in Corollary 3.2, or the sub-Gaussian norm of X_{ij} as in Corollary 3.3 of Bandeira et al., 2016. But in our case, none of these bounds gives a good result. However, the proof of Theorem 1.1 (and its extensions) only requires control of the moments of the order $\log n_1$. For a Binomial r.v. X with parameters n_2/L and p_{max}, we have, according to Theorem 1 in Ahle, 2022, for all $c \in \mathbb{N}^*$,
$$\mathbb{E}(X^c) \leqslant (n_2 p_{max}/L)^c e^{c^2/(2n_2 p_{max}/L)}.$$
Let X' be an independent copy of X. Since $n_2 p_{max}/L \gtrsim \log n_1$ by assumption, $e^{c^2/(2n_2 p_{max}/L)} \leqslant e^{\gamma c}$ for $c \asymp \log(n_1)$ and an absolute constant $\gamma > 0$. Hence
$$\mathbb{E}((X-X')^c) \leqslant \sum_{i \leqslant c} \binom{c}{i} \mathbb{E}(X^i)\mathbb{E}(X^{c-i}) \leqslant 2^c (n_2 p_{max}/L)^c e^{\gamma c}.$$
Observe that $(2e^\gamma)^c \lesssim \mathbb{E}Y_{ij}^{2c} = O((2c)^c)$ for every c, so that for all even c we have
$$\mathbb{E}\left(\frac{L(X-X')}{2e^\gamma n_2 p_{max}}\right)^c \lesssim \mathbb{E}Y_{ij}^{2c}.$$
We can now use the same argument as in Corollary 3.2 of Bandeira et al., 2016 to conclude that the matrix M with independent entries generated with the same law as $X-X'$ satisfies with probability at least $1 - O(n_1^{\Omega(1)})$ (since $\sqrt{n_1 n_2 p_{max}/L} \gtrsim \log n_1$ by assumption)
$$\|M\| \leqslant \sqrt{n_1 n_2 p_{max}/L}.$$
When the random variables are only centered, we can use the symmetrization argument of Corollary 3.3 to finally obtain
$$\|\mathbb{E}Z_2^\top\| \lesssim \sqrt{n_1 n_2 p_{max}/L},$$
with probability at least $1 - O(n_1^{\Omega(1)})$.

□

Remark 28. *The second concentration inequality of the above lemma slightly improves Proposition 1 and Theorem 4 in Ndaoud et al., 2022. The third concentration could be of independent interest and be applied for example in multilayer network analysis where matrices with independent Binomial entries arise naturally as the sum of the adjacency matrices of the layers, see e.g. Paul et al., 2020; Braun et al., 2021b.*

The following lemma corresponds to (Laurent et al., 2000, Lemma 1.1) and controls the tails of the weighted sums of squares of Gaussian random variables.

Lemma 37 (Laurent-Massart bound)**.** *Let $X_1, \cdots, X_n$ be i.i.d standard Gaussian random variables. Let $\mu = (\mu_1, \cdots, \mu_n)$ be a vector with non-negative entries and define $\zeta = \sum_{i=1}^{n} \mu_i (X_i^2 - 1)$. Then it holds for all $t \geqslant 0$ that*

$$\mathbb{P}(\zeta \geqslant 2\|\mu\|_2 \sqrt{t} + 2\|\mu\|_\infty t) \leqslant e^{-t}$$
$$\mathbb{P}(\zeta \leqslant -2\|\mu\|_2 \sqrt{t}) \leqslant e^{-t}$$

An immediate corollary now follows.

Corollary 3. *Let $X_1, \cdots, X_{n_1}$ and $Y_1, \cdots, Y_{n_2}$ be two independent sets of i.i.d standard Gaussian random variables. Let $\mu = (\mu_1, \cdots, \mu_{n_1})$ and $\nu = (\nu_1, \cdots, \nu_{n_2})$ be two vectors with non-negative entries. Define $\zeta = \sum_{i=1}^{n_1} \mu_i X_i^2$ and $\xi = \sum_{i=1}^{n_2} \nu_i Y_i^2$. Then it holds for $t \geqslant 0$ that*

$$\mathbb{P}\big(\zeta - \xi \geqslant \|\mu\|_1 - \|\nu\|_1 + 2(\|\mu\|_2 + \|\nu\|_2)\sqrt{t} + 2\|\mu\|_\infty t\big) \leqslant 2e^{-t}, \tag{A.o.1}$$

$$\mathbb{P}\big(\zeta - \xi \leqslant \|\mu\|_1 - \|\nu\|_1 - 2(\|\mu\|_2 + \|\nu\|_2)\sqrt{t} - 2\|\nu\|_\infty t\big) \leqslant 2e^{-t}. \tag{A.o.2}$$

The next lemma give us a distributional equality for terms of the form $\langle g, Xg \rangle$ where g is a standard Gaussian vector and X is a permutation matrix.

Lemma 38. *Let $X \in \mathcal{P}_n$ and $g = (g_1, \cdots, g_n)$ be a standard Gaussian vector. Then is holds*

$$\langle g, Xg \rangle \overset{d}{=} \sum_{i=1}^{n} \lambda_i g_i'^2,$$

where λ_i are the eigenvalues of $\frac{1}{2}(X + X^T)$ and $g' = (g_1, \cdots, g_n)$ is a vector of independent standard Gaussians. Moreover, if $|S_X| = s_x n$ for $s_x \in (0, 1]$, $\mu \in \mathbb{R}^{n_1}$ is a vector containing the positive eigenvalues of $\frac{1}{2}(X + X^T)$, and $-\nu \in \mathbb{R}^{n_2}$ is a vector containing the negative eigenvalues of $\frac{1}{2}(X + X^T)$, then

$$\|\mu\|_1 - \|\nu\|_1 = s_x n,$$
$$\sqrt{n} \leqslant \|\mu\|_2 + \|\nu\|_2 \leqslant \sqrt{2n},$$
$$\|\mu\|_\infty, \|\nu\|_\infty \leqslant 1.$$

Proof. Notice that $\langle g, Xg \rangle = \langle g, \frac{1}{2}(X + X^T)g \rangle$ and given the symmetry of the matrix $\frac{1}{2}(X + X^T)$ all its eigenvalues are real. Take its SVD decomposition $\frac{1}{2}(X + X^T) = V \Lambda V^T$. We have that

$$\langle g, \frac{1}{2}(X + X^T)g \rangle = (V^T g)^T \Lambda V^T g \overset{d}{=} \sum_{i=1}^{n} \lambda_i g_i'^2$$

using the rotation invariance of the standard Gaussian vectors. Notice that

$$|S_X| = \mathrm{Tr}(X) = \mathrm{Tr}\left(\frac{1}{2}(X + X^T)\right) = \sum_{i=1}^{n} \lambda_i$$

which leads to

$$\|\mu\|_1 - \|\nu\|_1 = \sum_{i=1}^{n} \lambda_i = |S_X| = s_x n.$$

The fact that $\|\mu\|_\infty, \|\nu\|_\infty \leqslant 1$ follows easily since X is a unitary matrix. The inequality $\|\mu\|_2 + \|\nu\|_2 \geqslant \sqrt{n}$ follows from the fact that $\|\mu\|_2^2 + \|\nu\|_2^2 = n$. From the latter, we deduce that $\|\mu\|_2 + \|\nu\|_2 \leqslant \sqrt{\|\mu\|_2^2} + \sqrt{n - \|\mu\|_2^2} \leqslant 2\sqrt{\frac{n}{2}}$, and the result follows. □

BIBLIOGRAPHY

Abbe, E. and C. Sandon (2015a). "Community Detection in General Stochastic Block models: Fundamental Limits and Efficient Algorithms for Recovery." In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 670–688.

Abbe, Emmanuel (2018). "Community Detection and Stochastic Block Models." In: *Foundations and Trends® in Communications and Information Theory* 14.1-2, pp. 1–162.

Abbe, Emmanuel, Enric Boix-Adserà, Peter Ralli, and Colin Sandon (2020a). "Graph Powering and Spectral Robustness." In: *SIAM Journal on Mathematics of Data Science* 2.1, pp. 132–157.

Abbe, Emmanuel, Jianqing Fan, and Kaizheng Wang (2020b). *An ℓ_p theory of PCA and spectral clustering*. arXiv: 2006.14062 [math.ST].

Abbe, Emmanuel, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong (2020c). "Entrywise Eigenvector Analysis of Random Matrices with Low Expected Rank." In: *Annals of statistics* 48.3, pp. 1452–1474.

Abbe, Emmanuel and Colin Sandon (2015b). *Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic BP, and the information-computation gap*. arXiv: 1512.09080 [math.PR].

Abbe, Emmanuel and Colin Sandon (2016). "Achieving the KS threshold in the general stochastic block model with linearized acyclic belief propagation." In: *Advances in Neural Information Processing Systems*. Vol. 29, 1334—1342.

Aflalo, Yonathan, Alexander Bronstein, and Ron Kimmel (2015). "On convex relaxation of graph isomorphism." In: *PNAS* 112.10, pp. 2942–2947.

Ahle, Thomas D. (2022). "Sharp and simple bounds for the raw moments of the binomial and Poisson distributions." In: *Statistics & Probability Letters* 182, p. 109306.

Ahn, Kwangjun, Kangwook Lee, Hyunseung Cha, and Changho Suh (2018). "Binary Rating Estimation with Graph Side Information." In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., pp. 4272–4283.

Airoldi, Edoardo, D.M. Blei, S.E. Fienberg, and E.P. Xing (Sept. 2008). "Mixed Membership Stochastic Blockmodels." In: *Journal of Machine Learning Research* 9, pp. 1981–2014.

Araya, Ernesto, Guillaume Braun, and Hemant Tyagi (2022). *Seeded graph matching for the correlated Wigner model via the projected power method*. arXiv: 2204.04099 [math.ST].

Arroyo, Jesús and Elizaveta Levina (2021). "Overlapping Community Detection in Networks via Sparse Spectral Decomposition." In: *Sankhya A* 84.1, pp. 1–35.

Athreya, Avanti, Donniell E. Fishkind, Minh Tang, Carey E. Priebe, Youngser Park, Joshua T. Vogelstein, Keith Levin, Vince Lyzinski, Yichen Qin, and Daniel L Sussman (2018). "Statistical Inference on Random Dot Product Graphs: a Survey." In: *Journal of Machine Learning Research* 18.226, pp. 1–92.

Bandeira, Afonso S. and Ramon van Handel (July 2016). "Sharp nonasymptotic bounds on the norm of random matrices with independent entries." In: *Ann. Probab.* 44.4, pp. 2479–2506.

Beckett, Stephen J. (2016). "Improved community detection in weighted bipartite networks." In: *Royal Society Open Science* 3.1, p. 140536.

Bhattacharyya, Sharmodeep and Shirshendu Chatterjee (2018). *Spectral Clustering for Multiple Sparse Networks: I*. arXiv: 1805.10594 [stat.ME].

Bhattacharyya, Sharmodeep and Shirshendu Chatterjee (2020). *General Community Detection with Optimal Recovery Conditions for Multi-relational Sparse Networks with Dependent Layers*. arXiv: 2004.03480 [cs.SI].

Bickel, Peter, David Choi, Xiangyu Chang, and Hai Zhang (2013). "Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels." In: *The Annals of Statistics* 41.4, pp. 1922 –1943.

Binkiewicz, Norbert, J. Vogelstein, and K. Rohe (2017). "Covariate-assisted spectral clustering." In: *Biometrika* 104, pp. 361 –377.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre (2008). "Fast unfolding of communities in large networks." In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008.

Bordenave, Charles, Marc Lelarge, and Laurent Massoulié (2018). "Non-backtracking spectrum of random graphs: Community detection and nonregular Ramanujan graphs." In: *The Annals of Probability* 46.1, pp. 1–71.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities. A nonasymptotic theory of independence*. Oxford University Press, x–481 p.

Boumal, Nicolas (2016). "Nonconvex Phase Synchronization." In: *SIAM Journal on Optimization* 26.4, pp. 2355–2377.

Bouveyron, Charles, Gilles Celeux, T. Brendan Murphy, and Adrian E. Raftery (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Brault, Vincent, Christine Keribin, Gilles Celeux, and Gerard Govaert (Jan. 2014). "Estimation and Selection for the Latent Block Model

on Categorical Data." In: *Statistics and Computing* 25, pp. 1–16. DOI: 10.1007/s11222-014-9472-2.

Brault, Vincent, Christine Keribin, and Mahendra Mariadassou (2020). "Consistency and asymptotic normality of Latent Block Model estimators." In: *Electronic Journal of Statistics* 14.1, pp. 1234 –1268.

Braun, Guillaume and Hemant Tyagi (2022). *Minimax Optimal Clustering of Bipartite Graphs with a Generalized Power Method*. arXiv: 2205.12104 [math.ST].

Braun, Guillaume, Hemant Tyagi, and Christophe Biernacki (2021a). *An iterative clustering algorithm for the Contextual Stochastic Block Model with optimality guarantees*. arXiv: 2112.10467 [stat.ML].

Braun, Guillaume, Hemant Tyagi, and Christophe Biernacki (2021b). "Clustering multilayer graphs with missing nodes." In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Vol. 130, pp. 2260–2268.

Braun, Urs et al. (2015). "Dynamic reconfiguration of frontal brain networks during executive cognition in humans." In: *Proceedings of the National Academy of Sciences* 112.37, pp. 11678–11683.

Buldygin, V. and K. Moskvichova (2013). "The sub-Gaussian norm of a binary random variable." In: *Theory of Probability and Mathematical Statistics* 86, pp. 33–49.

Burer, Samuel and Renato D.C. Monteiro (2001). "A Nonlinear Programming Algorithm for Solving Semidefinite Programs via Low-rank Factorization." In: *Mathematical Programming (series B* 95, p. 2003.

Cai, T. Tony, Rungang Han, and Anru R. Zhang (2022). "On the non-asymptotic concentration of heteroskedastic Wishart-type matrix." In: *Electronic Journal of Probability* 27, pp. 1 –40.

Celeux, G. and G. Govaert (1992). "A Classification EM algorithm for clustering and two stochastic versions." In: *Computational Statistics & Data Analysis* 14, pp. 315–332.

Celisse, Alain, Jean-Jacques Daudin, and Laurent Pierre (2011). "Consistency of maximum-likelihood and variational estimators in the Stochastic Block Model." In: *Electronic Journal of Statistics* 6, pp. 1847–1899.

Chen, Xin and Anderson Y. Zhang (2021). *Optimal Clustering in Anisotropic Gaussian Mixture Models*. arXiv: 2101.05402 [math.ST].

Chen, Yuxin and Emmanuel J. Candès (2018). "The Projected Power Method: An Efficient Algorithm for Joint Alignment from Pairwise Differences." In: *Communications on Pure and Applied Mathematics* 71.8, pp. 1648–1714.

Chen, Yuxin, Yuejie Chi, Jianqing Fan, and Cong Ma (2019). "Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval." In: *Mathematical Programming* 176.1–2, 5–37.

Chi, Jocelyn, Eric Chi, and Richard Baraniuk (2015). "k -POD A Method for k -Means Clustering of Missing Data." In: *The American Statistician* 70, pp. 1–29.

Chi, Yuejie, Yue Lu, and Yuxin Chen (Aug. 2019). "Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview." In: *IEEE Transactions on Signal Processing* PP, pp. 1–1.

Cohen-Addad, Vincent, Adrian Kosowski, Frederik Mallmann-Trenn, and David Saulpic (2020). "On the Power of Louvain in the Stochastic Block Model." In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 4055–4066.

Conte, D., P. Foggia, C. Sansone, and M. Vento (2004). "Thirty Years Of Graph Matching In Pattern Recognition." In: *International Journal of Pattern Recognition and Artificial Intelligence* 18.3, pp. 265–298.

Contisciani, Martina, Eleanor Power, and Caterina De Bacco (Sept. 2020). "Community detection with node attributes in multilayer networks." In: *Scientific reports* 10, p. 15736.

Cour, Timothee, Praveen Srinivasan, and Jianbo Shi (2006). "Balanced Graph Matching." In: *Proceedings of the 19th International Conference on Neural Information Processing Systems*, 313–320.

Cucuringu, Mihai, Peter Davies, Aldo Glielmo, and Hemant Tyagi (2019). "SPONGE: A generalized eigenproblem for clustering signed networks." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89, pp. 1088–1098.

Cullina, Daniel, Negar Kiyavash, Prateek Mittal, and H. Vincent Poor (2018). *Partial Recovery of Erdős-Rényi Graph Alignment via* k*-Core Alignment*. arXiv: 1809.03553 [cs.IT].

Czekanowski, Jan (1909). "Zur Differentialdiagnose der Neandertalgruppe." In: *Korrespondenz-Blatt der Deutschen Gesellschaft für Anthropologie, Ethnologie und Urgeschichte* 40, pp. 44–47.

Dai, Osman Emre, Daniel Cullina, Negar Kiyavash, and Matthias Grossglauser (2018). *Analysis of a Canonical Labeling Algorithm for the Alignment of Correlated Erdős-Rényi Graphs*. arXiv: 1804.09758 [cs.DS].

Dasgupta, S. (2008). *The Hardness of K-means Clustering*. Technical report (University of California, San Diego. Department of Computer Science and Engineering).

Daudin, Jean-Jacques, Franck Picard, and Stéphane Robin (June 2008). "A mixture model for random graph." In: *Statistics and Computing* 18, pp. 173–183.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm." In: *Journal of the Royal Statistical Society, Series B* 39.1, pp. 1–38.

Deng, Shaofeng, Shuyang Ling, and Thomas Strohmer (2021). "Strong Consistency, Graph Laplacians, and the Stochastic Block Model." In: *Journal of Machine Learning Research* 22.117, pp. 1–44.

Deshpande, Yash, Andrea Montanari, Elchanan Mossel, and Subhabrata Sen (2018). "Contextual Stochastic Block Models." In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Vol. 31, 8590–8602.

Ding, Jian, Zongming Ma, Yihong Wu, and Jiaming Xu. (2021). "Efficient random graph matching via degree profiles." In: *Probability Theory and Related Fields* 179, 29–115.

Dong, X., P. Frossard, P. Vandergheynst, and N. Nefedov (2014). "Clustering on Multi-Layer Graphs via Subspace Analysis on Grassmann Manifolds." In: *IEEE Transactions on Signal Processing* 62.4, pp. 905–918.

Eagle, Nathan and Alex Pentland (2006). "Reality Mining: Sensing Complex Social Systems." In: *Personal Ubiquitous Comput.* 10.4, 255–268.

Emmert-Streib, Frank, Matthias Dehmer, and Yongtang Shi (2016). "Fifty Years of Graph Matching, Network Alignment and Network Comparison." In: *Inf. Sci.* 346.C, 180–197.

Emmons, Scott and Peter J. Mucha (2019). "Map equation with metadata: Varying the role of attributes in community detection." In: *Phys. Rev. E* 100 (2), p. 022301.

Fajardo-Fontiveros, Oscar, Roger Guimerà, and Marta Sales-Pardo (2022). "Node Metadata Can Produce Predictability Crossovers in Network Inference Problems." In: *Phys. Rev. X* 12 (1), p. 011010.

Fan, Zhou, Cheng Mao, Yihong Wu, and Jiaming Xu (2019). *Spectral Graph Matching and Regularized Quadratic Relaxations I: The Gaussian Model*. arXiv: 1907.08880 [stat.ML].

Feizi, Soheil, Gerald T. Quon, Mariana Recamonde-Mendoza, Muriel Médard, Manolis Kellis, and Ali Jadbabaie (2020). "Spectral Alignment of Graphs." In: *IEEE Transactions on Network Science and Engineering* 7, pp. 1182–1197.

Feldman, Vitaly, Will Perkins, and Santosh Vempala (2015). "Subsampled Power Iteration: A Unified Algorithm for Block Models and Planted CSP's." In: *Proceedings of the 28th International Conference on Neural Information Processing Systems*. Vol. 28, 2836–2844.

Fishkind, Donniell E., Sancar Adali, Heather G. Patsolic, Lingyao Meng, Digvijay Singh, Vince Lyzinski, and Carey E. Priebe (2019). "Seeded graph matching." In: *Pattern Recognition* 87, pp. 203–215.

Florescu, Laura and Will Perkins (2016). "Spectral thresholds in the bipartite stochastic block model." In: *29th Annual Conference on Learning Theory*. Vol. 49, pp. 943–959.

Fortunato, Santo (2010). "Community detection in graphs." In: *Physics Reports* 486.3, pp. 75–174.

Friedman, J., J. Kahn, and E. Szemerédi (1989). "On the Second Eigenvalue of Random Regular Graphs." In: *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*. STOC '89, 587–598.

Ganassali, Luca, Marc Lelarge, and Laurent Massoulié (2022a). "Spectral alignment of correlated Gaussian matrices." In: *Advances in Applied Probability* 54.1, 279–310.

Ganassali, Luca, Laurent Massoulié, and Marc Lelarge (2022b). "Correlation Detection in Trees for Planted Graph Alignment." In: *ITCS*, 74:1–74:8.

Gao, Chao, Zongming Ma, Anderson Y. Zhang, and Harrison H. Zhou (2018). "Community detection in degree-corrected block models." In: *The Annals of Statistics* 46.5, pp. 2153–2185.

Gao, Chao and A. Zhang (2019). "Iterative Algorithm for Discrete Structure Recovery." In: *arXiv: Statistics Theory*.

Gao, Chao and Anderson Y. Zhang (2021). *Optimal Orthogonal Group Synchronization and Rotation Group Synchronization*. arXiv: 2109.13491 [math.ST].

Gao, Chao and Anderson Y. Zhang (2022). "SDP Achieves Exact Minimax Optimality in Phase Synchronization." In: *IEEE Transactions on Information Theory* 68.8, pp. 5374–5390.

Gaucher, Solenne and Olga Klopp (2021). "Optimality of variational inference for stochasticblock model with missing links." In: *Advances in Neural Information Processing Systems*. Vol. 34, pp. 19947–19959.

Gaudio, Julia, Miklos Z. Racz, and Anirudh Sridhar (2022). *Exact Community Recovery in Correlated Stochastic Block Models*. arXiv: 2203.15736 [math.ST].

Giraud, Christophe and Nicolas Verzelen (2019). "Partial recovery bounds for clustering with the relaxed K-means." In: *Mathematical Statistics and Learning* 3, 317–374.

Gori, Marco, Marco Maggini, and Lorenzo Sarti (2004). "Graph matching using random walks." In: *in IEEE Proceedings of the International Conference on Pattern Recongition*, pp. 394–397.

Guimerà, Roger and Marta Sales-Pardo (2011). "Justice Blocks and Predictability of U.S. Supreme Court Votes." In: *PLOS ONE* 6.11, pp. 1–8.

H. Sun, W. Zhou and M. Fei (2020). "A Survey On Graph Matching In Computer Vision." In: *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 225–230.

Han, Qiuyi, Kevin Xu, and Edoardo Airoldi (2015). "Consistent Estimation of Dynamic and Multi-Layer Block Models." In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, 1511–1520.

Han, Rungang, Yuetian Luo, Miaoyan Wang, and Anru R. Zhang (2020). *Exact Clustering in Tensor Block Model: Statistical Optimality and Computational Limit*. arXiv: 2012.09996 [stat.ME].

Holland, Paul W., Kathryn Blackmond Laskey, and Samuel Leinhardt (1983). "Stochastic blockmodels: First steps." In: *Social Networks* 5.2, pp. 109 –137.

Hric, Darko, Tiago P. Peixoto, and Santo Fortunato (2016). "Network Structure, Metadata, and the Prediction of Missing Nodes and Annotations." In: *Phys. Rev. X* 6 (3), p. 031038.

Hu, Jiaxin and Miaoyan Wang (2022). "Multiway Spherical Clustering via Degree-Corrected Tensor Block Models." In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Vol. 151, pp. 1078–1119.

Hu, Menglei and Songcan Chen (2019). "One-Pass Incomplete Multi-View Clustering." In: *The Thirty-Third Conference on Artificial Intelligence*, pp. 3838–3845.

Huang, Baihe, Shunhua Jiang, Zhao Song, Runzhou Tao, and Ruizhe Zhang (2021). *Solving SDP Faster: A Robust IPM Framework and Efficient Implementation*. arXiv: 2101.08208 [math.OC].

Huang, Zan, Daniel D. Zeng, and Hsinchun Chen (2007). "Analyzing Consumer-Product Graphs: Empirical Findings and Applications in Recommender Systems." In: *Management Science* 53.7, pp. 1146–1164.

Jin, Jiashun, Zheng Tracy Ke, and Shengming Luo (2017). *Estimating network memberships by simplex vertex hunting*. arXiv: 1708.07852 [stat.ME].

Jin, Jiashun, Zheng Tracy Ke, and Shengming Luo (2021). "Optimal adaptivity of signed-polygon statistics for network testing." In: *The Annals of Statistics* 49.6, pp. 3408 –3433.

Jin, Jiashun, Zheng Tracy Ke, and Shengming Luo (2022). "Improvements on SCORE, Especially for Weak Signals." In: *Sankhya A: The Indian Journal of Statistics* 84.1, pp. 127–162.

Jing, Bing-Yi, Ting Li, Zhongyuan Lyu, and Dong Xia (2021). "Community detection on mixture multilayer networks via regularized tensor decomposition." In: *The Annals of Statistics* 49.6, pp. 3181 –3205.

Jo, Changhun and Kangwook Lee (2021). "Discrete-Valued Latent Preference Matrix Estimation with Graph Side Information." In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139, pp. 5107–5117.

Journée, Michel, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre (2010). "Generalized Power Method for Sparse Principal Component Analysis." In: *Journal of Machine Learning Research* 11, pp. 517–553.

Kaufmann, Emilie, Thomas Bonald, and Marc Lelarge (2018). "A spectral algorithm with additive clustering for the recovery of overlapping communities in networks." In: *Theor. Comput. Sci.* 742, pp. 3–26.

Kazemi, Ehsan and Matthias Grossglauser (2016). *On the Structure and Efficient Computation of IsoRank Node Similarities*. arXiv: 1602.00668 [q-bio.MN].

Kazemi, Ehsan, S. Hamed Hassani, and Matthias Grossglauser (2015). "Growing a Graph Matching from a Handful of Seeds." In: *Proceedings of the VLDB endowment* 8.10, pp. 1010–1021.

Keribin, Christine, Vincent Brault, Gilles Celeux, and Gérard Govaert (2015). "Estimation and selection for the latent block model on categorical data." In: *Statistics and Computing* 25, pp. 1201–1216.

Kim, Jungeun and Jae-Gil Lee (2015). "Community Detection in Multi-Layer Graphs: A Survey." In: *SIGMOD Record* 44, pp. 37–48.

Kivelä, Mikko, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter (2014). "Multilayer networks." In: *Journal of Complex Networks* 2.3, pp. 203–271.

Kumar, Abhishek, Piyush Rai, and Hal Daume (2011). "Co-regularized Multi-view Spectral Clustering." In: *Advances in Neural Information Processing Systems 24*, pp. 1413–1421.

Kumar, Amit, Yogish Sabharwal, and Sandeep Sen (2004). "A Simple Linear Time $(1+\epsilon)$-Approximation Algorithm for k-Means Clustering in Any Dimensions." In: *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pp. 454–462.

Kunisky, Dmitriy and Jonathan Niles-Weed (2022). "Strong recovery of geometric planted matchings." In: *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 834–876.

Larremore, Daniel, Aaron Clauset, and Caroline Buckee (2013). "A Network Approach to Analyzing Highly Recombinant Malaria Parasite Genes." In: *PLoS computational biology* 9, e1003268.

Latouche, Pierre, Etienne Birmelé, and Christophe Ambroise (2011). "Overlapping stochastic block models with application to the French political blogosphere." In: *The Annals of Applied Statistics* 5.1, pp. 309–336.

Laurent, Beatrice and Pascal Massart (2000). "Adaptive estimation of a quadratic functional by model selection." In: *Annals of statistics* 28.5, pp. 1302–1338.

Le Gall, François (2014). "Algebraic Complexity Theory and Matrix Multiplication." In: *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation*, p. 23.

Le, Can M., Elizaveta Levina, and Roman Vershynin (2017). "Concentration and regularization of random graphs." In: *Random Structures & Algorithms* 51.3, pp. 538–561.

Lee, Jeonghwan, Daesung Kim, and Hye Won Chung (2020). "Robust Hypergraph Clustering via Convex Relaxation of Truncated MLE." In: *IEEE Journal on Selected Areas in Information Theory* 1.3, pp. 613–631.

Léger, Jean-Benoist (2016). *Blockmodels: A R-package for estimating in Latent Block Model and Stochastic Block Model, with various probability functions, with or without covariates*. arXiv: 1602.07587 [stat.CO].

Lei, Jing, Kehui Chen, and Brian Lynch (Mar. 2020). "Consistent community detection in multi-layer network data." In: *Biometrika* 107, pp. 61–73.

Lei, Jing and Kevin Z. Lin (2022). "Bias-Adjusted Spectral Clustering in Multi-Layer Stochastic Block Models." In: *Journal of the American Statistical Association* 0.0, pp. 1–13.

Lei, Jing and Alessandro Rinaldo (2015). "Consistency of spectral clustering in stochastic block models." In: *Annals of Statistics* 43.1, pp. 215–237.

Lei, Lihua (2019). *Unified $\ell_{2\to\infty}$ Eigenspace Perturbation Theory for Symmetric Random Matrices*. arXiv: 1909.04798 [math.PR].

Li, Xiaodong, Yudong Chen, and Jiaming Xu (2021). "Convex Relaxation Methods for Community Detection." In: *Statistical Science* 36.1, pp. 2 –15.

Lim, Shiau Hong, Yudong Chen, and Huan Xu (2015). "A Convex Optimization Framework for Bi-Clustering." In: *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37, pp. 1679–1688.

Ling, Shuyang (2021). *Generalized Power Method for Generalized Orthogonal Procrustes Problem: Global Convergence and Optimization Landscape Analysis*. arXiv: 2106.15493 [cs.IT].

Liu, Xinwang, Miaomiao Li, Chang Tang, Jingyuan Xia, Jian Xiong, Li Liu, Marius Kloft, and En Zhu (2021). "Efficient and Effective Regularized Incomplete Multi-View Clustering." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.8, pp. 2634–2646.

Longepierre, Léa and Catherine Matias (2019). "Consistency of the maximum likelihood and variational estimators in a dynamic stochastic block model." In: *Electronic Journal of Statistics* 13.2, pp. 4157 –4223.

Lu, Chen and Subhabrata Sen (2020). *Contextual Stochastic Block Model: Sharp Thresholds and Contiguity*. arXiv: 2011.09841 [cs.SI].

Lu, Yu and Harrison H. Zhou (2016). *Statistical and Computational Guarantees of Lloyd's Algorithm and its Variants*. arXiv: 1612.02099 [math.ST].

Lubars, Joseph and R. Srikant (2018). "Correcting the Output of approximate Graph matching algorithms." In: *IEEE conference on Computer Communications*, pp. 1745–1753.

Luxburg, Ulrike (Jan. 2004). "A Tutorial on Spectral Clustering." In: *Statistics and Computing* 17, pp. 395–416.

Lyzinski, Vince, Donniell E. Fishkind, Marcelo Fiori, Joshua T. Vogelstein, Carey E. Priebe, and Guillermo Sapiro (2016). "Graph

Matching: relax at your own risk." In: *IEEE transactions on pattern analysis and machine intelligence* 38.1, pp. 60–73.

Ma, Zongming and Sagnik Nandy (2021). *Community Detection with Contextual Multilayer Networks*. arXiv: 2104.02960 [math.ST].

Majmudar, Jimit and Stephen Vavasis (2020). "Provable Overlapping Community Detection in Weighted Graphs." In: *Advances in Neural Information Processing Systems*. Vol. 33, pp. 19028–19038.

Mao, Cheng, Mark Rudelson, and Konstantin Tikhomirov (2021a). *Exact Matching of Random Graphs with Constant Correlation*. arXiv: 2110.05000 [math.ST].

Mao, Xueyu, Purnamrita Sarkar, and Deepayan Chakrabarti (2018). "Overlapping Clustering Models, and One (Class) SVM to Bind Them All." In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Vol. 31, 2130–2140.

Mao, Xueyu, Purnamrita Sarkar, and Deepayan Chakrabarti (2021b). "Estimating Mixed Memberships With Sharp Eigenvector Deviations." In: *Journal of the American Statistical Association* 116.536, pp. 1928–1940.

Mariadassou, Mahendra and Timothée Tabouy (2020). "Consistency and asymptotic normality of stochastic block models estimators from sampled data." In: *Electronic Journal of Statistics* 14.2, pp. 3672 –3704.

Mele, Angelo, Lingxin Hao, Joshua Cape, and Carey Priebe (2019). "Spectral Inference for Large Stochastic Blockmodels With Nodal Covariates." In: *SSRN Electronic Journal*, pp. 1–61.

Mitzenmacher, Michael and Eli Upfal (2005). *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press.

Mossel, Elchanan and Jiaming Xu (2019). "Seeded Graph Matching via Large Neighborhood Statistics." In: *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '19, 1005–1014.

Mouselimis, Lampros (2022). *ClusterR: Gaussian Mixture Models, K-Means, Mini-Batch-Kmeans, K-Medoids and Affinity Propagation Clustering*. R package version 1.2.6.

Narayanan, A. and V. Shmatikov (2009). "De-anonymizing Social Networks." In: *2009 30th IEEE Symposium on Security and Privacy*, pp. 173–187.

Narayanan, Arvind and Vitaly Shmatikov (2006). *How To Break Anonymity of the Netflix Prize Dataset*. arXiv: cs/0610105 [cs.CR].

Ndaoud, Mohamed (2018). *Sharp optimal recovery in the two-component Gaussian Mixture Model*. arXiv: 1812.08078 [math.ST].

Ndaoud, Mohamed, Suzanne Sigalla, and Alexandre B. Tsybakov (2022). "Improved Clustering Algorithms for the Bipartite Stochastic Block Model." In: *IEEE Transactions on Information Theory* 68.3, pp. 1960–1975.

Newman, M. E. J. and M. Girvan (2004). "Finding and evaluating community structure in networks." In: *Physical Review E* 69 (2), p. 026113.

Newman, M. and Aaron Clauset (2015). "Structure and inference in annotated networks." In: *Nature Communications* 7, p. 11863.

Onaran, Efe and Soledad Villar (2017). "Projected power iteration for network alignment." In: *Wavelets and Sparsity XVII*. Vol. 10394, pp. 298 –305.

Paul, Subhadeep and Yuguo Chen (2020). "Spectral and matrix factorization methods for consistent community detection in multi-layer networks." In: *Annals of Statistics* 48.1, pp. 230–250.

Pedarsani, Pedram and Matthias Grossglauser (2011). "On the privacy of anonymized networks." In: *In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1235–1243.

Pensky, Marianna and Teng Zhang (2019). "Spectral clustering in the dynamic stochastic block model." In: *Electron. J. Statist.* 13.1, pp. 678–709.

Piccioli, Giovanni, Guilhem Semerjian, Gabriele Sicuro, and Lenka Zdeborová (2022). "Aligning random graphs with a sub-tree similarity message-passing algorithm." In: *Journal of Statistical Mechanics: Theory and Experiment* 2022.6, p. 063401.

Racz, Miklos and Anirudh Sridhar (2021). "Correlated Stochastic Block Models: Exact Graph Matching with Applications to Recovering Communities." In: *Advances in Neural Information Processing Systems*. Vol. 34, pp. 22259–22273.

Ricci-Tersenghi, Federico, Adel Javanmard, and Andrea Montanari (2016). "Performance of a community detection algorithm based on semidefinite programming." In: *Journal of Physics: Conference Series* 699, p. 012015.

Rohit Singh, Jinbo Xu and Bonnie Berger (2008). "Global alignment of multiple protein interaction networks with application to functional orthology detection." In: *Proceedings of the National Academy of Sciences* 105.35, pp. 12763–12768.

S. Bravo Ferreira, Jose F., Yuehaw Khoo, and Amit Singer (2018). "Semidefinite Programming Approach for the Quadratic Assignment Problem with a Sparse Graph." In: *Computational Optimization and Applications* 69, pp. 677–712.

Saade, Alaa, Florent Krzakala, and Lenka Zdeborová (2014). "Spectral Clustering of graphs with the Bethe Hessian." In: *Advances in Neural Information Processing Systems 27*, pp. 406–414.

Sahni, Sartaj and Teofilo Gonzalez (1976). "P-Complete Approximation Problems." In: *Journal of the ACM* 23, 555–565.

Sarkar, Purnamrita, Y. X. Rachel Wang, and Soumendu S. Mukherjee (2021). "When random initializations help: a study of variational

inference for community detection." In: *Journal of Machine Learning Research* 22.22, pp. 1–46.

Schäcke, Kathrin (2004). *On the Kronecker Product*. Technical report. URL: https://www.math.uwaterloo.ca/~hwolkowi/henry/reports/kronthesisschaecke04.pdf.

Singh, Rohit, Jinbo Xu, and Bonnie Berger (2009). "Global alignment of multiple protein interaction networks with application to functional orthology detection." In: *PNAS* 105.35, pp. 12763–12768.

Squartini, Tiziano, Assaf Almog, Guido Caldarelli, Iman van Lelyveld, Diego Garlaschelli, and Giulio Cimini (May 2017). "Enhanced capital-asset pricing model for bipartite financial networks reconstruction." In: *Physical Review E* 96, p. 032315.

Stanley, Natalie, Thomas Bonacci, Roland Kwitt, Marc Niethammer, and Peter Mucha (2019). "Stochastic Block Models with Multiple Continuous Attributes." In: *Applied Network Science* 4, p. 54.

Tang, Wei, Zhengdong Lu, and Inderjit Dhillon (2009). "Clustering with Multiple Graphs." In: *IEEE International Conference on Data Mining*, pp. 1016–1021.

Tropp, Joel A. (2012). "User-Friendly Tail Bounds for Sums of Random Matrices." In: *Found. Comput. Math.* 12.4, 389–434.

Umeyama, Shinji (1988). "An eigendecomposition approach to weighted graph matching problems." In: *IEEE transactions on pattern analysis and machine intelligence* 10.5, pp. 695–703.

Vandenberghe, Lieven and Stephen Boyd (1996). "Semidefinite Programming." In: *SIAM Rev.* 38.1, 49–95.

Vershynin, Roman (2016). *Four lectures on probabilistic methods for data science*. arXiv: 1612.06661 [math.PR].

Wang, Haoyu, Yihong Wu, Jiaming Xu, and Israel Yolou (2022a). "Random Graph Matching in Geometric Models: the Case of Complete Graphs." In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Vol. 178, pp. 3441–3488.

Wang, Peng, Huikang Liu, Zirui Zhou, and Anthony Man-Cho So (2021). "Optimal Non-Convex Exact Recovery in Stochastic Block Model via Projected Power Method." In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139, pp. 10828–10838.

Wang, Xiaolu, Peng Wang, and Anthony Man-Cho So (2022b). "Exact Community Recovery over Signed Graphs." In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Vol. 151, pp. 9686–9710.

Wedin, Per-Åke (1972). "Perturbation bounds in connection with singular value decomposition." In: *BIT Numerical Mathematics* 12.1, pp. 99–111.

Weng, Haolei and Yang Feng (2022). "Community detection with nodal information: Likelihood and its variational approximation." In: *Stat* 11.1, e428.

Weyl, Hermann (1912). "Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung)." In: *Mathematische Annalen* 71.4, pp. 441–479.

Wu, Yihong, Jiaming Xu, and H Yu Sophie (2021). "Settling the sharp reconstruction thresholds of random graph matching." In: *2021 IEEE International Symposium on Information Theory*, pp. 2714–2719.

Wu, Yihong and Harrison H. Zhou (2022). "Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations." In: *Mathematical Statistics and Learning* 4.3, 143–220.

Xu, Hongteng, Dixin Luo, and Lawrence Carin (2019). "Scalable Gromov-Wasserstein Learning for Graph Partitioning and Matching." In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 3052–3062.

Yan, Bowei and Purnamrita Sarkar (2020). "Covariate Regularized Community Detection in Sparse Graphs." In: *Journal of the American Statistical Association* 0.0, pp. 1–12.

Yang, Jaewon, Julian McAuley, and Jure Leskovec (2013). "Community Detection in Networks with Node Attributes." In: *2013 IEEE 13th International Conference on Data Mining*, pp. 1151–1156.

Yartseva, Lyudmila and Matthias Grossglauser (2013). "On the Performance of Percolation Graph Matching." In: *Proceedings of the First ACM Conference on Online Social Networks*, 119–130.

Yin, Mingzhang, Y. X. Rachel Wang, and Purnamrita Sarkar (2020). "A Theoretical Case Study of Structured Variational Inference for Community Detection." In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Vol. 108, pp. 3750–3761.

Young, Jean-Gabriel, Fernanda Valdovinos, and M. Newman (2021). "Reconstruction of plant–pollinator networks from observational data." In: *Nature Communications* 12, p. 3911.

Yu, Liren, Jiaming Xu, and Xiaojun Lin (2021a). "Graph matching with partially-correct seeds." In: *Journal of Machine Learning Research* 22.2021, pp. 1–54.

Yu, Liren, Jiaming Xu, and Xiaojun Lin (2021b). "The Power of D-Hops in Matching Power-Law Graphs." In: *Proc. ACM Meas. Anal. Comput. Syst.* 5.2, pp. 1–43.

Yu, Tianshu, Junchi Yan, Yilin Wang, Wei Liu, and Baoxin Li (2018). "Generalizing graph matching beyond quadratic assignment model." In: *Advances in Neural Information Processing Systems* 31.12, 861–871.

Yu, Yi, Tengyao Wang, and Richard J. Samworth (2014). "A useful variant of the Davis–Kahan theorem for statisticians." In: *Biometrika* 102, pp. 315–323.

Zaslavskiy, Mikhail, Francis Bach, and Jean-Philippe Vert (2009a). "A Path Following Algorithm for the Graph Matching Problem." In:

IEEE transactions on pattern analysis and machine intelligence 31.12, pp. 60–73.

Zaslavskiy, Mikhail, Francis Bach, and Jean-Philippe Vert (2009b). "Global alignment of protein–protein interaction networks by graph matching methods." In: *Bioinformatics* 25.12, pp. 1259–1267.

Zhang, Anderson Y and Harrison H Zhou (2016a). "Minimax Rates of Community Detection in Stochastic Block Models." In: *The Annals of Statistics* 44.5, pp. 2252–2280.

Zhang, Anderson Y. and Harrison H. Zhou (2020a). "Theoretical and computational guarantees of mean field variational inference for community detection." In: *Annals of Statistics* 48.5, pp. 2575 –2598.

Zhang, Anru R., T. Tony Cai, and Yihong Wu (2022). "Heteroskedastic PCA: Algorithm, optimality, and applications." In: *The Annals of Statistics* 50.1, pp. 53 –80.

Zhang, Yuan, Elizaveta Levina, and Ji Zhu (2016b). "Community detection in networks with node features." In: *Electronic Journal of Statistics* 10.2, pp. 3153 –3178.

Zhang, Yuan, Elizaveta Levina, and Ji Zhu (2020b). "Detecting Overlapping Communities in Networks Using Spectral Methods." In: *SIAM J. Math. Data Sci.* 2, pp. 265–283.

Zhao, Qing, Stefan Karisch, Franz Rendl, and Henry Wolkowicz (Oct. 1997). "Semidefinite Programming Relaxations For The Quadratic Assignment Problem." In: *Journal of Combinatorial Optimization* 2, pp. 71–109.

Zhou, Zhixin and Arash A.Amini (2019). "Analysis of spectral clustering algorithms for community detection: the general bipartite setting." In: *Journal of Machine Learning Research* 20.47, pp. 1–47.

Zhou, Zhixin and Arash Amini (2020). "Optimal Bipartite Network Clustering." In: *Journal of Machine Learning Research* 21, pp. 1–68.

Zhu, Ziwei, Tengyao Wang, and Richard J. Samworth (2019). *High-dimensional principal component analysis with heterogeneous missingness*. arXiv: 1906.12125 [stat.ME].

Zubin, Joseph (1938). "A technique for measuring like-mindedness." In: *The Journal of Abnormal and Social Psychology* 33, pp. 508–516.