Electronic Journal of Statistics
Vol. 18 (2024) 2798–2823
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2271

Strong consistency guarantees for clustering high-dimensional bipartite graphs with the spectral method

Guillaume Braun

RIKEN AIP, Tokyo
e-mail: guillaume.braun@riken.jp

Abstract: We investigate the problem of clustering bipartite graphs using a simple spectral method within the framework of the Bipartite Stochastic Block Model (BiSBM), a popular model for bipartite graphs having a community structure. Our focus lies in the high-dimensional setting where the number n_1 of rows, and n_2 of columns, of the associated adjacency matrix differ significantly. A recent study by [4] has established a sufficient and necessary condition related to the sparsity level $p_{\max}$ of the bipartite graph, enabling the recovery of the latent partition of the rows. In their work, [4] introduces an iterative method that extends the approach proposed by [26] to achieve the stated recovery goal. However, empirical results suggest that the subsequent refinement algorithm does not significantly enhance the performance of the spectral method, indicating that the spectral method achieves exact recovery within the same regime as the refinement method. We establish this claim by deriving new entrywise bounds on the eigenvectors of the similarity matrix utilized by the spectral method. Our analysis extends the framework of [23], which is limited to symmetric matrices with restricted dependencies. As a critical technical step, we also derive an improved concentration inequality tailored for similarity matrices.

MSC2020 subject classifications: Primary 62H30.
Keywords and phrases: Spectral method, clustering, bipartite stochastic block model.

Received April 2023.

1. Introduction

Bipartite graphs are a convenient way to represent the relationships between objects of two types. One can find examples of applications in many fields, such as e-commerce with customers and products [20], finance with investors and assets [28], and biology with plants of pollinators networks [31]. These networks are often large and sparse, characterized by an associated adjacency matrix with a notable imbalance between the number of rows and columns (e.g., one can have a dataset with far more products than customers).

To extract meaningful insights from these networks, clustering methods are commonly employed. Spectral Clustering (SC) is a particularly popular approach due to its efficiency in terms of computational complexity and statistical accuracy. However, existing consistency guarantees for SC are weak, requiring stronger conditions on the sparsity level of the bipartite graph than the provably optimal algorithms analyzed by [4] and [26]. Despite this, experimental results in the aforementioned works show that SC performs almost as well as the provably optimal algorithms, even with worse theoretical guarantees.

In this work, we address this gap by showing that SC achieves exact recovery under the BiSBM, an asymmetric extension of the Stochastic Block Model (SBM) commonly used to evaluate the performance of clustering algorithms for bipartite graphs. We demonstrate that SC achieves exact recovery whenever the sparsity level of the bipartite graph p_{max} (i.e., the maximal probability of observing an edge in the bipartite graph) satisfies $p_{max}^2 \gtrsim \frac{\log n_1}{n_1 n_2}$. According to [4], this is a necessary and sufficient condition to recover the rows partition exactly. Hence, SC is optimal, up to a constant factor. We leave the characterization of the precise constant necessary for exact recovery as future work.

1.1. Main contributions

Our primary contributions are summarized below:

- We demonstrate that the spectral method achieves exact recovery of the rows partition whenever $p_{\max}^2 \gtrsim \frac{\log n_1}{n_1 n_2}$, establishing its optimality, up to a constant factor. To accomplish this, we extend to similarity matrices the entrywise concentration bounds for eigenvectors of [23], originally developed for matrices with independent entries or limited dependencies.
- Central to our proof is an improved concentration bound for similarity matrices. We derive this result by adapting the combinatorial argument presented by [16], demonstrating the concentration of adjacency matrices sampled from the generalized Erdös-Renyi model.
- Our analysis applies to the rank-deficient connectivity matrices, allowing for the partial removal of the "spectral gap condition. This condition, common in the analysis of spectral methods, typically requires that the matrices of interest satisfy a rank condition to ensure the existence of a spectral gap. Our approach aligns with recent works by [25] and [32].

1.2. Related work

Bipartite graphs and spectral clustering The recent work of [4] confirmed the conjecture of [26] that $p_{max}^2 \gtrsim \frac{\log n_1}{n_1 n_2}$ is both a necessary and sufficient condition for exact recovery of the rows partition under the high-dimensional BiSBM, where $n_1 \ll n_2$. This threshold can be attained through the use of generalized power methods proposed in the aforementioned articles. However, existing strong consistency guarantees for spectral clustering (SC) require stronger assumptions. For instance, when specialized to the setting of [26] (a special case of our more general model), the result of [5] holds only when the sparsity level satisfies $p_{max}^2 \gtrsim \frac{\log^2 n_1}{n_1 n_2}$. In cases where $p_{max}^2 \gtrsim \frac{\log n_1}{n_1 n_2}$, SC is only guaranteed to achieve a weak form of consistency, see Proposition 1 in [4]. The work of [17] also demonstrated that when $p_{max}^2 \gtrsim \frac{1}{n_1 n_2}$, it is possible to recover a proportion of the row clusters through an SBM reduction. However, this represents the weakest existing recovery guarantee, and our focus is on achieving exact recovery. The recent work of [32] proposed an improved analysis of the spectral method for asymmetric matrices with independent entries, but their bound becomes trivial in the high-dimensional regime $n_1 \ll n_2$ that we are interested in. This regime where $n_2/n_1 \to \infty$ was also considered in the recent work of [29] for matrix completion in the challenging regime where only partial recovery is possible.

Entrywise concentration bounds for eigenvectors In recent years, spectral algorithms have demonstrated success in achieving exact recovery in various community detection tasks across diverse settings. Examples include the Stochastic Block Model (SBM) [3], the Contextual SBM [2], the Censored Block Model [12], Hierarchical SBM [24], and uniform Hypergraph SBM [18]. Spectral methods have also found applications in other estimation problems, such as group synchronization [11], ranking [8], or planted subgraph detection [13].

To establish these results, it is typically crucial to derive entrywise eigenvector concentration bounds. In this work, we adopt the framework developed by [23], which integrates techniques for obtaining deterministic perturbation bounds [15, 6, 10] with methods relying on certain stochastic properties of the noise [3, 7, 14].

1.3. Notations

We use lowercase letters $(\epsilon, a, b, \dots)$ to denote scalars and vectors, except for universal constants that will be denoted by $c_1, c_2, \dots$ for lower bounds, and $C_1, C_2, \dots$ for upper bounds and some random variables. Occasionally, we employ the notation $a_n \lesssim b_n$ (or $a_n \gtrsim b_n$) for sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ if there is a constant $C > 0$ such that $a_n \leq C b_n$ (resp. $a_n \geq C b_n$) for all 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. $a_n = \Theta(b_n)$).

Matrices will be denoted by uppercase letters. The i-th row of a matrix A will be denoted as $A_{i:}$. The column j of A will be denoted by $A_{:j}$, and the (i,j)th entry by A_{ij}. The transpose of A is denoted by $A^\top$ and $A_{:j}^\top$ corresponds to the j-th row of $A^\top$ by convention. I_k denotes the $k \times k$ identity matrix. For matrices, we use $||\cdot||$ to denote the spectral norm (or Euclidean norm in the case of vectors), and $||\cdot||_F$ to denote the Frobenius norm. The set of vectors $x \in \mathbb{R}^d$ such that $\|x\| = 1$ is denoted by $\mathbb{S}^{d-1}$. The number of non-zeros entries of a matrix A will be denoted by $\texttt{nnz}(A)$.

2. Model and algorithm description

2.1. The bipartite stochastic block model (BiSBM)

The BiSBM is a direct extension of the SBM [19] to bipartite graphs. The model depends on 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 $z(i) = z_i = k$ (resp. $z'(i) = z'_i = l$) where k (resp. l) is the unique column index satisfying $(Z_1)_{ik} = 1$ (resp. $(Z_2)_{il} = 1$).
- 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

$$P = (p_{ij})_{i \in [n_1], j \in [n_2]} := Z_1 \Pi (Z_2)^\top \in [0,1]^{n_1 \times n_2}.$$

A graph $\mathcal{G}$ is distributed according to BiSBM (Z_1, Z_2, Π) if the entries of the corresponding bipartite adjacency matrix $A \in \{0,1\}^{n_1 \times n_2}$ 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. The sparsity level of the graph is denoted by $p_{max} = \max_{i,j} p_{ij}$. We make the following assumptions on the model.

Assumption A1 (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 $\alpha \geq 1$ such that for all $k \in [K]$ and $l \in [L]$ we have*

$$\frac{n_1}{\alpha K} \leq |\mathcal{C}_k| \leq \frac{\alpha n_1}{K} \text{ and } \frac{n_2}{\alpha L} \leq |\mathcal{C}'_l| \leq \frac{\alpha n_2}{L}.$$

We will consider throughout this work the parameters α, K and L as constants. We won't keep track in the stated bounds of the dependencies in these parameters.

We will rely on the following assumption to ensure the communities are well separated.

Assumption A2 (Communities are well separated)**.** *Let $U^* \Lambda^* (U^*)^\top$ be the spectral decomposition of $PP^\top$. All the communities are well separated if the following assumptions are satisfied.*

1. *The smallest non zero eigenvalue of* $\Pi\Pi^\top$*, denoted by* $\lambda_{min}(\Pi\Pi^\top)$*, satisfies* $\lambda_{min}(\Pi\Pi^\top) \gtrsim p_{\max}^2$.
2. *There exists a constant* $c_1 > 0$ *such that for all* $i, j \in [n_1]$ *with* $z_i \neq z_j$ *we have* $\left\|U^*_{i:} - U^*_{j:}\right\| \geq \frac{c_1}{\sqrt{n}}$.

Remark 1. *This assumption doesn't require that* $\Pi\Pi^\top$ *is full rank contrary to classical assumptions used for analyzing spectral clustering. For example, consider the setting where* $K = 2 = L$*, the communities are exactly balanced and*

$$\Pi = \Pi^\top = \begin{pmatrix} p & cp \\ cp & c^2 p \end{pmatrix}$$

where p *is the sparsity parameter and* $c > 0$ *is a constant. Observe that*

$$PP^\top = \frac{n_2}{2} Z_1 \Pi\Pi^\top Z_1^\top = \frac{n_1 n_2}{4} W \Pi\Pi^\top W^\top$$

where $W = \sqrt{\frac{2}{n_1}} Z_1$ *has orthonormal columns. The SVD decomposition of* $\Pi\Pi^\top$ *is given by* $(1+c^2)^2 p^2 VV^\top$ *where* $V = (\frac{1}{\sqrt{1+c^2}}, \frac{c}{\sqrt{1+c^2}})^\top$*. Hence,* $PP^\top = U\Lambda U^\top$ *where* $U^* = WV$ *and* $\Lambda = n_1 n_2 (1+c^2)^2 p^2/4$ *corresponds to the only non-zero singular value of* $PP^\top$*. It is easy to check that for* $i \in \mathcal{C}_1$ *and* $j \in \mathcal{C}_2$ *we have*

$$\left\|U^*_{i:} - U^*_{j:}\right\| = \frac{2|1-c|}{\sqrt{(1+c^2)n_1}}.$$

The quality of the clustering is evaluated through the **misclustering rate** r defined by

$$r(\hat{z}, z) = \frac{1}{n} \min_{\pi \in \mathfrak{S}} \sum_{i \in [n]} \mathbf{1}_{\{\hat{z}(i) \neq \pi(z(i))\}}, \tag{2.1}$$

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

2.2. Algorithm description

In the high-dimensional and sparse setting where $n_1 \ll n_2$ and $n_1 n_2 p_{max}^2$ is of order $\log n_1$, there is no hope of recovering the columns partition Z_2. To see that, observe that the probability that all the elements of a given column are zero is of order $(1 - p_{max})^{n_1} \approx e^{-n_1 p_{max}}$ and is close to one in the considered sparsity regime. So, it is natural to form the similarity matrix $AA^\top$ and compute the top-K eigenspace of this similarity matrix[1]. Unfortunately, the diagonal

[1]Note that similarity matrices are semi-definite positive and eigenvalues correspond to singular values.

elements of $AA^\top$ create an important bias: $(AA^\top)_{ii}$ is typically of order $n_2 p_{max}$ while the diagonal entries of the corresponding population similarity matrix are of order $n_2 p_{max}^2$. To avoid this issue, one can remove the diagonal of $AA^\top$ and obtain a matrix B. In this work, we consider a slightly different variant of the spectral methods proposed by [4, 26, 17]. See Algorithm 1 for a complete description of the method.

Algorithm 1 Spectral method on $\mathcal{H}(AA^\top)$ (`Spec`)

Input: The number of communities K, the rank r of $\Pi\Pi^\top$ and the adjacency matrix A.

1: Form the diagonal hollowed Gram matrix $B := \mathcal{H}(AA^\top)$ where $\mathcal{H}(X) = X - \text{diag}(X)$.
2: Compute the matrix $U \in \mathbb{R}^{n_1 \times r}$ whose columns correspond to the top r-eigenvectors of B.
3: Apply approximate $(1 + 2/e + \epsilon)$ approximate `k-medians` on the rows of U and obtain a partition $z^{(0)}$ of $[n_1]$ into K communities.

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

When the rank of $\Pi\Pi^\top$ is not known, we propose `AdaSpec` (see Algorithm 2), an adaptive version of Algorithm 1.

Algorithm 2 Adaptive spectral method on $\mathcal{H}(AA^\top)$ (`AdaSpec`)

Input: The number of communities K, a threshold $T > 0$, and the adjacency matrix A.

1: Form the diagonal hollowed Gram matrix $B := \mathcal{H}(AA^\top)$ where $\mathcal{H}(X) = X - \text{diag}(X)$.
2: Let $\hat{r} \in [K]$ be the largest index such that the difference between two consecutive eigenvalues are larger than some threshold T

$$\hat{r} := \arg\max\{r \in [K] : \lambda_r(B) - \lambda_{r+1}(B) > T\}.$$

3: Compute the matrix $U \in \mathbb{R}^{n_1 \times r}$ whose columns correspond to the top r-eigenvectors of B.
4: Apply $(1 + 2/e + \epsilon)$ approximate `k-medians` 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 The cost of computing B is $O(n_1 \texttt{nnz}(A))$ and for U is[2] $O(n_1^2 K \log n_1)$. Applying the $(1 + 2/e + \epsilon)$ approximate `k-medians` has a complexity $O(f(K,\epsilon) n_1^{O(1)})$ where $f(K,\epsilon) = (\epsilon^{-2} K \log K)^K$, see [9]. Here we used (approximate) `k-medians` because it can be linked easily with $\ell_{2\to\infty}$ perturbation bounds (see Lemma 5.1 in [23]). But we could also apply (approximate) `k-means` as a rounding step and use results from [30], Section 2.4 for the analysis. Depending on the rounding step used, the dependencies in some model parameters such as the number of communities K can change.

[2] 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.

3. Main results

First, we derive a new concentration bound for the similarity matrix B. It improves the upper-bound $\sqrt{n_1 n_2 p_{max}^2} \vee \log n_1$ used in [4] to $\sqrt{n_1 n_2 p_{max}^2}$ when $n_1 n_2 p_{max}^2 \gtrsim \log n_1$. This improvement of a $\sqrt{\log n_1}$ factor is essential to show that `Spec` achieves exact recovery in the challenging parameter regime where $n_1 n_2 p_{max}^2$ is of order $\log n_1$.

Theorem 1. *Let $B = \mathcal{H}(AA^\top)$ where $A \sim BiSBM(n_1, n_2, K, L, \Pi)$ with $p_{max}^2 = \frac{c_{n_1}}{n_1 n_2}$ where $\log n_1 \lesssim c_{n_1} \ll \log^2 n_1$, $n_2 \gtrsim n_1 \log^2 n_1 c_{n_1}$, and $n_2 = O(n_1^\beta)$ for some constant $\beta > 0$. Then with probability at least $1 - n_1^{-\Theta(1)}$*

$$\|B - \mathbb{E}(B)\| \lesssim \sqrt{n_1 n_2 p_{max}^2}.$$

Remark 2. *By comparison with the concentration inequality established in [4], the only interesting sparsity regime is when $p_{max} \asymp \sqrt{\frac{c_{n_1}}{n_1 n_2}}$ where $\log n_1 \lesssim c_{n_1} \ll \log^2 n_1$. Note that in this case, $n_1 p_{max} = \sqrt{\frac{n_1 c_{n_1}}{n_2}} = O(\log^{-1} n_1)$ since $n_2 \gtrsim n_1 \log^2 n_1 c_{n_1}$. Similarly, $n_2 p_{max}^2 \asymp \frac{c_{n_1}}{n_1} = o(1)$. These two facts will be used repeatedly in the proofs. In particular, the condition $n_1 p_{max} = O(\log^{-1} n_1)$ is critical to ensure that for all $l \in [n_2]$, $\|A_{:l}\|^2 \leq C$ w.h.p. We, however, believe that the condition $n_2 \geq \log^2 n_1 c_{n_1}$ is an artifact of the proof and could be relaxed for example by removing, or pruning, the columns of A that have a large norm.*

Remark 3. *By using this concentration inequality, one could improve the conditions of applicability of Proposition 1 and Theorem 2 in [4]. For example, Proposition 1 requires that $n_1 n_2 p_{max}^2 \geq C \log n_1$ for a constant $C > 0$ large enough. But by using the concentration inequality of Theorem 1, we would only require $n_1 n_2 p_{max}^2 \geq c \log n_1$ for an arbitrary constant $c > 0$. See also Remark 8 in [4].*

Finally, we show that `Spec` achieves exact recovery by proving an $\ell_{2\to\infty}$ concentration bound for the top $-r$ eigenspace U of B. Let us denote the $\ell_{2\to\infty}$ distance between two matrices of eigenvectors U and $U^* \in \mathbb{R}^{n\times r}$ by

$$d_{2\to\infty}(U, U^*) = \inf_{O \in \mathbb{R}^{r\times r}, O^\top O = I} \|UO - U^*\|_{2\to\infty}$$

where $\|A\|_{2\to\infty} = \max_{\|x\|=1} \|Ax\|_\infty$.

Theorem 2. *Assume that $A \sim BiSBM(n_1, n_2, K, L, \Pi)$ with $p_{max}^2 = \frac{c_{n_1}}{n_1 n_2}$ where $\log n_1 \lesssim c_{n_1} \ll \log^2 n_1$, $n_2 \gtrsim n_1 \log^2 n_1 c_{n_1}$, and $n_2 = O(n_1^\beta)$ for some constant $\beta > 0$. Let $U\Lambda U^\top$ (resp. $U^*\Lambda^* U^{*\top}$) be the spectral decomposition of $B = \mathcal{H}(AA^\top)$ (resp. $B^* = PP^\top$). Then there exists a constant $c > 0$ (that can be made arbitrarily small if C is chosen large enough) such that with probability at least $1 - n^{-\Theta(1)}$*

$$d_{2\to\infty}(U, U^*) \leq \frac{c}{\sqrt{n_1}}.$$

Corollary 1. *Under the assumptions of Theorem 2,* **Spec** *achieves exact recovery with probability at least* $1 - n^{-\Theta(1)}$.

Corollary 2. *Under the assumptions of Theorem 2,* **AdaSpec** *achieves exact recovery with probability at least* $1 - n^{-\Theta(1)}$*, with* $T = n_1 n_2 p_{max}^2 / \log\log n_1$.

4. Proof of Theorem 1

The proof strategy is based on the combinatorial argument developed by [16]. Let us denote

$$\mathcal{E} = \left\{ \max_{l\in[n_2]} \sum_{i\in[n_1]} A_{il} \le C_{col} \right\}$$

where $C_{col} > 0$ is a constant appropriately large. By Chernoff bound (additive form) and a union bound, we obtain

$$\mathbb{P}(\mathcal{E}^c) \le n_2 e^{-\frac{C_{col}^2}{2n_1 p_{max}}} \le e^{\log n_2 - C_{col}^2 \Omega(\log n_1)} \le e^{-\Omega(\log n_1)}$$

since by assumptions $\log n_2 \le \beta \log n_1$ and $(n_1 p_{max})^{-1} \gtrsim \sqrt{\frac{n_2}{n_1 \log n_1}} \gtrsim \log n_1$. By choosing C_{col} large enough, we can ensure that $\mathcal{E}$ occurs with probability at least $1 - n_1^{-3}$.

Step 1 A standard $\epsilon-$ net argument with the Euclidean norm (see e.g. Lemma 2 and 3 in [21]) shows that for all $0 < \epsilon < 1/2$ there exists a $\epsilon-$ net $\mathcal{N}$ of $\mathbb{S}^{n_1-1}$ such that $|\mathcal{N}| \le (1 + \frac{2}{\epsilon})^{n_1}$ and

$$\|B - \mathbb{E}(B)\| \le \frac{1}{1-2\epsilon} \sup_{x\in\mathcal{N}} \left| x^\top (B - \mathbb{E}(B)) x \right|.$$

In the following, we will fix $\epsilon = 1/4$, in particular $|\mathcal{N}| \le 9^{n_1}$.

Step 2 To bound the previous quantity, let us introduce for all $x \in \mathbb{S}^{n_1-1}$ the set of "light pairs"

$$\mathcal{L}(x) = \{(i,j) \in [n_1] \times [n_1] : |x_i x_j| \le \sqrt{\frac{n_2}{n_1}} p_{max}\}$$

and the set of "heavy pairs"

$$\mathcal{H}(x) = [n_1] \times [n_1] \setminus \mathcal{L}(x).$$

When clear from the context, we will omit the dependency in x in the notations of the previous sets and write $\mathcal{L}$ and $\mathcal{H}$ instead of $\mathcal{L}(x)$ and $\mathcal{H}(x)$

We have

$$\sup_{x\in\mathcal{N}} \left| x^\top (B - \mathbb{E}(B)) x \right| \le \underbrace{\sup_{x\in\mathcal{N}} \left| \sum_{(i,j)\in\mathcal{L}} x_i x_j B_{ij} - x^\top \mathbb{E} B x \right|}_{(T1)} + \underbrace{\sup_{x\in\mathcal{N}} \left| \sum_{(i,j)\in\mathcal{H}} x_i x_j B_{ij} \right|}_{(T2)}.$$

Step 3 We are going to bound (T1) w.h.p. Observe that

$$(T1) \le \underbrace{\left|\sum_{(i,j)\in\mathcal{L}} x_i x_j (B_{ij} - \mathbb{E}B_{ij})\right|}_{(E1)} + \underbrace{\left|\sum_{(i,j)\in\mathcal{H}} x_i x_j \mathbb{E}B_{ij}\right|}_{(E2)}.$$

It is easy to bound the deterministic quantity (E2)

$$\begin{aligned}(E2) &\le \sum_{(i,j)\in\mathcal{H}} \mathbb{E}B_{ij} \frac{(x_i x_j)^2}{|x_i x_j|} \\ &\le \sqrt{\frac{n_1}{n_2}} p_{max}^{-1} n_2 p_{max}^2 \sum_{(i,j)\in\mathcal{H}} (x_i x_j)^2 \\ &\le \sqrt{n_1 n_2} p_{max} \sum_{i\in[n_1]} x_i^2 \sum_{j\in[n_1]} x_j^2 \\ &= \sqrt{n_1 n_2} p_{max}.\end{aligned}$$

The upper-bound of (E1) follows from Lemma 6 (see the appendix) that gives

$$\mathbb{P}\left(\mathcal{E} \cap \{|\sum_{(i,j)\in\mathcal{L}} x_i x_j (B_{ij} - \mathbb{E}B_{ij})| \ge C_1 \sqrt{n_1 n_2} p_{max}\}\right) \le e^{-11n_1}$$

for some constant $C_1 > 1$ large enough. Since $|\mathcal{N}| \le e^{9n_1}$ according to Step 1, we obtain by a union bound argument that

$$\begin{aligned}&\mathbb{P}\left(\mathcal{E} \cap \sup_{x\in\mathcal{N}} \left|\sum_{(i,j)\in\mathcal{L}} x_i x_j B_{ij} - x^\top \mathbb{E}Bx\right| > 2C_1 \sqrt{n_1 n_2} p_{max}\right) \\ &\le \mathbb{P}\left(\mathcal{E} \cap \sup_{x\in\mathcal{N}} \left|\sum_{(i,j)\in\mathcal{L}} x_i x_j (B_{ij} - \mathbb{E}B_{ij})\right| > C_1 \sqrt{n_1 n_2} p_{max}\right) \\ &\le |\mathcal{N}| e^{-11n_1} \le e^{-2n_1}.\end{aligned}$$

Step 4 We will now bound the term involving the heavy pairs $(T2)$. First, one needs to control the sum of the entries of each row and column of B.

Lemma 1. *Under the assumptions of Theorem 1, there exists a constant $C_2 > 0$ such that*

$$\mathbb{P}\left(\mathcal{E} \cap \{\max_{i\in[n_1]} \sum_{j\in[n_1]} B_{ij} = \max_{j\in[n_1]} \sum_{i\in[n_1]} B_{ij} \ge C_2 n_2 n_1 p_{max}^2\}\right) \le e^{-\Theta(n_1 n_2 p_{max}^2)}.$$

Proof. Fix $i \in [n_1]$. We have $S = \sum_j B_{ij} = \langle A_{i:}, \sum_{j \neq i} A_{j:} \rangle$. One can use a similar approach as in Lemma 7 (see the appendix) with sets $I = \{i\}$ and $J = [n_1] \setminus I$ and apply Bennett's inequality. We conclude by using a union bound and the fact that $n_1 e^{-\Theta(n_1 n_2 p_{max}^2)} = e^{-\Theta(n_1 n_2 p_{max}^2)}$ since $n_1 n_2 p_{max}^2 \gtrsim \log n_1$. Furthermore, B is symmetric; hence, the row sums correspond to column sums. □

Then, we need to show that the matrix B satisfies w.h.p. the discrepancy property defined below, with appropriate parameters.

Definition 1. *Let M be a $n \times n$ matrix with non-negative entries. For every $S, T \subset [n]$, let $e_M(S,T)$ denote the number of edges between S and T*

$$e_M(S,T) = \sum_{i \in S} \sum_{j \in T} M_{ij}.$$

We say that M obeys the discrepancy property DP $(\delta, \kappa_1, \kappa_2)$ with parameters $\delta > 0$, $\kappa_1 > 0$ and $\kappa_2 \geq 0$ if for all non-empty $S, T \subset [n]$, at least one of the following properties hold

1. $e_M(S,T) \leq \kappa_1 \delta |S||T|$;
2. $e_M(S,T) \log \frac{e_M(S,T)}{\delta |S||T|} \leq \kappa_2 \left(|S| \vee |T|\right) \log \frac{en_1}{|S| \vee |T|}$.

If one can show that B satisfies w.h.p. $DP(\delta, \kappa_1, \kappa_2)$ where κ_1, $\kappa_2 > 0$ are absolute constants and $\delta = n_2 p_{max}^2$, then Lemma B.4 in [21] would imply that w.h.p.

$$(T2) \lesssim \sqrt{n_1 n_2} p_{max}.$$

Consequently, to bound (T2) on the event $\mathcal{E}$, it is sufficient to show that w.h.p. B satisfies $DP(\delta, \kappa_1, \kappa_2)$. W.l.o.g., one can assume that $|S| \leq |T|$. If $|T| \geq \frac{n_1}{e}$ then Lemma 1 leads to

$$\frac{e_B(S,T)}{\delta |S||T|} \leq \frac{|S| C_2 n_2 n_1 p_{max}^2}{n_2 p_{max}^2 |S| n_1 / e} \leq C_2 e$$

with probability at least $1 - e^{\Theta(n_1 n_2 p_{max}^2)}$. Let us write $e_B(S,T) = \sum_{i,j} w_{ij} \langle A_{i:}, A_{j:} \rangle$ where $w_{ii} = 0$ and $w_{ij} = \mathbf{1}_{i \in S} \mathbf{1}_{j \in T}$. Following the proof of Theorem 5.2 in [22] (see also Lemma 4.2 in the supplementary material of [22]), for a given constant $c^* > 0$ let us define $t(T,S)$ as the unique solution of

$$t \log t = c^* \frac{|T| \log \frac{en_1}{|T|}}{\delta |S||T|}.$$

Let $C^* > 0$ a given constant and define

$$k(T,S) = \max\left(t(T,S), C^*\right).$$

We have by Lemma 7 in the appendix

$$\mathbb{P}\left(\mathcal{E} \cap \{e_B(S,T) \geq k(T,S) \delta |S||T|\}\right) \leq e^{-\frac{c}{2}|T| \log \frac{en_1}{|T|}}$$

and thus

$$\begin{aligned}
&\mathbb{P}\left(\{\exists S, T \subset [n_1], |S| \le |T| : e_B(S,T) \ge k(T,S)\delta|S||T|\} \cap \mathcal{E}\right) \\
&\le \sum_{I,J:|I|\le|J|<n_1/e} e^{-\frac{c}{2}|T|\log\frac{en_1}{|T|}} \\
&\le \sum_{s\le t\le n_1/e}\ \sum_{|S|=s,|T|=t} e^{-\frac{c}{2}|T|\log\frac{en_1}{|T|}} \\
&\le \sum_{s\le t\le n_1/e} \binom{n_1}{s}\binom{n_1}{t} e^{-\frac{c}{2}t\log\frac{en_1}{t}} \\
&\le \sum_{s\le t\le n_1/e} \left(\frac{en_1}{s}\right)^s \left(\frac{en_1}{t}\right)^t e^{-\frac{c}{2}t\log\frac{en_1}{t}} \\
&\le \sum_{s\le t\le n_1/e} e^{-\frac{c}{2}t\log\frac{en_1}{t}+t+s+t\log\frac{n_1}{t}+s\log\frac{n_1}{s}} \\
&\le \sum_{s\le t\le n_1/e} e^{-\frac{c}{2}t\log\frac{en_1}{t}+2t+2t\log\frac{n_1}{t}} \\
&\le \sum_{s\le t\le n_1/e} e^{-\frac{c-8}{2}t\log\frac{en_1}{t}} \\
&\le \sum_{s\le t\le n_1/e} n_1^{-\frac{c-8}{2}} \\
&\le n_1^{-\frac{c-12}{2}}
\end{aligned}$$

by using repeatedly the fact that $t\log\frac{n_1}{t}$ is increasing on $[1, \frac{n_1}{e}]$. By choosing a constant $c > 12$, we can show by using the same argument as in Lemma 4.2 in [22] that w.h.p. B satisfies $DP(n_2 p_{max}^2, \kappa_1, \kappa_2)$ for some constants κ_1 and κ_2.

Conclusion We have shown that

$$\begin{aligned}
\mathbb{P}(\|B-\mathbb{E}B\| \gtrsim \sqrt{n_1 n_2} p_{max}) &\le \mathbb{P}(\mathcal{E} \cap \{\|B-\mathbb{E}B\| \gtrsim \sqrt{n_1 n_2} p_{max}\}) + n_1^{-3} \\
&\le \mathbb{P}(\mathcal{E} \cap \{\sup_{x\in\mathcal{N}} \left|x^\top (B-\mathbb{E}(B))x\right| \gtrsim \sqrt{n_1 n_2} p_{max}\}) + n_1^{-3} \\
&\le n_1^{-\Omega(1)}
\end{aligned}$$

by Step 1, Step 2, Step 3, and Step 4.

5. Entrywise analysis of the spectral method

To show that the spectral method achieves exact recovery, we need to derive $\ell_{2\to\infty}$ eigenspace perturbation bound. Unfortunately, existing results only apply to symmetric matrices with independent entries or weak dependencies (see Section 7 in [23]) and cannot be directly applied to our setting. We propose an extension of the main result of [23] to the hollowed Gram matrix B considered in this work. We believe that our result can be extended to more general Gram matrices or kernel matrices.

5.1. Notations and preliminary results

First, let us introduce some notation. Let $\tilde{B}^* = \mathcal{H}(PP^\top)$ and $B^* = PP^\top$. Let $\lambda_1 \geq \ldots \geq \lambda_r$ (resp. $\lambda_1^* \geq \ldots \geq \lambda_r^*$) be the top $-r$ eigenvalues of B (resp. B^*) and U (resp. U^*) the corresponding matrix of eigenvectors.

The spectral decomposition of the matrices B and B^* is given by

$$B = U\Lambda U^\top, \quad B^* = U^*\Lambda^* U^{*\top}$$

where U (resp. U^*) is the matrix formed by the eigenvectors of B (resp. B^*) and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$ (resp. $\Lambda^* = \operatorname{diag}(\lambda_1^*, \ldots, \lambda_r^*)$) is the diagonal matrix of non-zero eigenvalues of B (resp. B^*). The noise $E = B - B^*$ can be further decomposed as

$$\underbrace{\mathcal{H}((A-P)(A-P)^\top)}_{\tilde{E}} + \underbrace{\mathcal{H}(P(A-P)^\top + (A-P)P^\top)}_{E'} + \underbrace{\tilde{B}^* - B^*}_{E''}.$$

First, let us establish analogous results to the Conditions (A2) and (A3) in [23].

Lemma 2. *Under the assumption of Theorem 2, there is an absolute constant $C_1 > 0$, such that for any $W \in \mathbb{R}^{n\times r}$, the following inequalities hold with probability at least $1 - n_1^{-\Theta(1)}$.*

1. $\|\Lambda - \Lambda^*\| \leq C_1\sqrt{n_1 n_2 p_{max}^2}$,
2. $\|E\|_{2\to\infty} \leq C_1\sqrt{n_1 n_2 p_{max}^2}$,
3. $\|EU^*\| \leq C_1\sqrt{n_1 n_2 p_{max}^2}$,
4. $\max_i \|E_{i:}W\| \leq b_\infty(\delta)\,\|W\|_{2\to\infty} + b_2(\delta)\,\|W\|$, *where* $b_\infty(\delta) = C_1\frac{R(\delta)}{\log R(\delta)}$ *and* $b_2(\delta) = C_1\frac{\sqrt{n_2 p_{max}^2}R(\delta)}{\log R(\delta)}$ *with* $R(\delta) = \log(n_1/\delta) + r$ *and* $\delta = n_1^{-c}$ *for some constant* $c > 0$.
5. $\max_i \|E_{i:}U^*\| \lesssim \|U\|_{2\to\infty}\log n_1 + \sqrt{p_{max}\log n_1}$.

Proof. By Theorem 1, we have with probability at least $1 - n_1^{-\Theta(1)}$

$$\left\|\tilde{E} + E'\right\| = \left\|\mathcal{H}(AA^\top) - \mathcal{H}(PP^\top)\right\| \lesssim \sqrt{n_1 n_2 p_{max}^2}.$$

Also, by definition, $\max_i \|P_{i:}\|^2 \leq n_2 p_{max}^2$ so

$$\|E''\| = \max_i \|P_{i:}\|^2 = o(\sqrt{n_1 n_2 p_{max}^2}).$$

By consequence, we have shown that

$$\|E\| \lesssim \sqrt{n_1 n_2 p_{max}^2}. \tag{5.1}$$

Proof of 1 This is a direct consequence of Weyl's inequality and (5.1):

$$\|\Lambda - \Lambda^*\| \leq \|B - B^*\| \lesssim \sqrt{n_1 n_2 p_{max}^2}.$$

Proof of 2 It follows from the fact that $\|E\|_{2\to\infty} \leq \|E\|$.

Proof of 3 It is a direct consequence of the sub-multiplicativity of the norm and the fact that $\|U^*\| \leq 1$.

Proof of 4 By Proposition 2.2 in [23], if we can show that for any $\delta \in (0,1)$ and vector $w \in \mathbb{R}^{n_1}$ there exist $a_\infty(\delta), a_2(\delta) > 0$ such that with probability at least $1-\delta$, for each $i \in [n_1]$

$$E_{i:}w \leq a_\infty(\delta) \|w\|_\infty + a_2(\delta) \|w\|,$$

then we obtain the stated result for $b_\infty(\delta) = 2a_\infty(\frac{\delta}{5^r n_1})$ and $b_2(\delta) = 2a_2(\frac{\delta}{5^r n_1})$. Fix $w \in \mathbb{R}^{n_1}, i \in [n_1]$ and let us denote $R = A - P$. Consider $S = \tilde{E}_{i:}w = \sum_{j\in[n_2]\setminus\{i\}} \langle R_{i:}, R_{j:}\rangle w_j$. Conditionally on $R_{i:}$, this is a sum of independent and centered r.v.s. By using Lemma F.3 in [23] with weights $\tilde{w}_{jl} = R_{il}w_j$ we obtain that conditionally on R_i the following holds with probability at least $1-\delta$

$$S \leq f(\delta) \left(\|\tilde{w}\|_\infty + \sqrt{\sum_{j,l} \tilde{w}_{jl}^2 p_{jl}} \right)$$

where $f(\delta) = \frac{2\log(1/\delta)}{F^{-1}(2\log(1/\delta)}$ and $F(t) = t^2 e^t$. But $\|\tilde{w}\|_\infty \leq \|w\|_\infty$ and

$$\|\tilde{w}\|_F^2 = \sum_{j,l} R_{il}^2 w_j^2 = \|w\|^2 \|R_{i:}\|^2.$$

Besides, with probability at least $1 - e^{\Theta(n_2 p_{max})}$, $\|R_{i:}\|^2 \leq C n_2 p_{max}$ by Hoeffding's inequality. Therefore with probability at least $1 - \delta - e^{-\Theta(n_2 p_{max})}$

$$S \leq f(\delta) \left((\|w\|_\infty + \sqrt{n_2 p_{max}^2} \|w\| \right).$$

It remains to bound $S' = E'_{i:}w = \sum_{j\in[n_2]\setminus\{i\}}(\langle R_{i:}, P_{j:}\rangle + \langle P_{i:}, R_{j:}\rangle) w_j$ and $S'' = E''_{i:}w$. We have

$$S'' = \|P_{i:}\|^2 w_i \leq \|w\|_\infty n_2 p_{max}^2 = o(\|w\|_\infty).$$

Also observe that $= \sum_{j\in[n_2]\setminus\{i\}} \langle R_{i:}, P_{j:}\rangle w_j = \sum_{j\neq i,l} R_{jl} w_j P_{il}$, so we can apply Lemma F.3 in [23] with weights $(w_j P_{il})_{j\neq i,l}$. We obtain

$$S'' \leq f(\delta) p_{max} (\|w\|_\infty + \|w\| \sqrt{n_2 p_{max}})$$

with probability at least $1-\delta$. A similar result holds for $S' = \sum_{j\neq i,l} \langle P_{i:}, R_{j:}\rangle w_j$: we can apply again Lemma F.3 in [23] with weights $(P_{jl} w_j)_{j,l}$ and obtain

$$S' \leq f(\delta) p_{max} (\|w\|_\infty + \|w\| \sqrt{n_2 p_{max}}).$$

So we can choose $a_\infty(\delta) = f(\delta)$ and $a_2(\delta) = f(\delta)\sqrt{n_2 p_{max}^2}$. One can check that, as in Lemma 3.1 in [23], $b_\infty(\delta) = \frac{4R(\delta)}{\log R(\delta)}$ and $b_2(\delta) = 4\frac{\sqrt{n_2 p_{max}^2} R(\delta)}{\log R(\delta)}$. Also note that $n_2 p_{max} \gtrsim \log n_1$ by assumption so if we choose $\delta = n_1^{-c}$ for an appropriate constant $c > 0$, the term $e^{-\Theta(n_2 p_{max})}$ will be negligible compared to δ.

Proof of 5 It also relies on Proposition 2.2 in [23] but with a different choice of $\gamma \approx \log^{-1} n_1$ to obtain a trade-off between $b_\infty(\delta) \|U^*\|_{2\to\infty} \approx \frac{b_\infty(\delta)}{\sqrt{n_1}}$ and $b_2(\delta) \|U^*\| = b_2(\delta)$. See also Lemma 3.3 in [23]. Since the proof is similar to the previous point, we omitted it. □

5.2. A new decoupling argument

The main difficulty to adapting Theorem 2.3 and 2.5 in [23] comes from their decoupling assumption (A1) which requires the existence of a matrix $B^{(i)}$ (typically obtained by replacing the i-th row and column of B by zeros or the expectation of the entries) such that for any $\delta \in (0,1)$

$$d_{TV}(\mathbb{P}_{(B_i, B^{(i)})}, \mathbb{P}_{B_i} \times \mathbb{P}_{B^{(i)}}) \leq \frac{\delta}{n}. \tag{5.2}$$

If the matrix B had independent entries it would be straightforward to satisfy this condition, but in our setting, it is not clear how to obtain such a general result. Consequently, we adopted a different approach that avoids bounding the total variation distance between two probability distributions.

Let us define $B^{(i)} \in \mathbb{R}^{n_1 \times n_1}$ the matrix obtained by replacing the i-th row and columns of B by zeros. We have

$$\left\|B^{(i)} - B\right\| \leq \|B_{i:}\| \leq \|E_{i:}\| + \|B^*_{i:}\| \lesssim \sqrt{n_1 n_2 p_{max}^2}$$

with probability at least $1 - e^{\Theta(n_1 n_2 p_{max}^2)}$, since $\|B^*_{i:}\| \leq \sqrt{n_1} n_2 p_{max}^2$, $n_2 p_{max}^2 = o(1)$, and $\|E_{i:}\| \leq \|E\| \leq \sqrt{n_1 n_2 p_{max}^2}$. We also have by definition

$$\begin{aligned}
\left\|(B^{(i)} - B)U)\right\| &\leq \|B_{i:}U\| + \left\|U_{i:}^\top B_{i:}\right\| \\
&\leq \|B_{i:}U\| + \|B_{i:}\| \|U_{i:}\| \\
&\leq \|(BU)_{i:}\| + \|B_{i:}\| \|U_{i:}\| \\
&\leq \|U_{i:}\Lambda\| + \|B_{i:}\| \|U_{i:}\| \\
&\leq (\|\Lambda\| + \|B_{i:}\|) \|U_{i:}\| .
\end{aligned}$$

Thanks to assumptions A1 and A2 we have $\lambda_r^* \gtrsim n_1 n_2 p_{max}^2$. By consequence, we have w.h.p.

$$\frac{\left\|(B^{(i)} - B)U)\right\|}{\lambda_r^*} \lesssim (1 + \frac{1}{\sqrt{n_1 n_2 p_{max}^2}}) \|U_{i:}\| \lesssim \|U\|_{2\to\infty} .$$

The previous inequalities correspond to the Condition (C0) used in the proof of Theorem 2.3 in [23] (cf. section A.2). They are summarized in the following lemma.

Lemma 3. *The following inequalities hold with probability at least* $1 - e^{\Theta(n_1 n_2 p_{max}^2)}$

1. $\left\|B^{(i)} - B\right\| \lesssim \sqrt{n_1 n_2 p_{max}^2}$,
2. $\frac{\left\|(B^{(i)}-B)U)\right\|}{\lambda_r^*} \lesssim \|U_{i:}\| \lesssim \|U\|_{2\to\infty}$.

Steps I and II of the proof of Theorem 2.3 [23] are deterministic and still hold in our setting (see the discussion in Section 5.3). The only step that uses the decoupling argument is step III where one needs to bound $\left\|E_{i:}(U^{(i)}H^{(i)} - U^*)\right\|$ where $U^{(i)}$ is the matrix formed by the eigenvectors of $B^{(i)}$ and $H^{(i)} \in \mathbb{R}^{r\times r}$ is the orthogonal matrix that best aligns $U^{(i)}$ and U^* (i.e. minimizes $\left\|U^{(i)}H - U^*\right\|$).

Lemma 4. *Let $W^{(i)} \in \mathbb{R}^{n_1\times K}$ be a matrix that only depends on $B^{(i)}$. Under the assumptions on Theorem 2, it holds with probability at least $1 - n_1^{-c'}$ for some constant $c' > 0$ that for all $i \in [n_1]$*

$$\left\|E_{i:}W^{(i)}\right\| \lesssim \frac{\log n_1}{\log\log n_1}\left\|W^{(i)}\right\|_{2\to\infty} + \frac{\sqrt{n_2}p_{max}\log n_1}{\log\log n_1}\left\|W^{(i)}\right\|.$$

Proof. Recall that $E = \tilde{E} + E' + E''$. By triangular inequality

$$\left\|E_{i:}W^{(i)}\right\| \le \left\|\tilde{E}_{i:}W^{(i)}\right\| + \left\|E'_{i:}W^{(i)}\right\| + \left\|E''_{i:}W^{(i)}\right\|.$$

We will first handle the first term. Let us denote $R = A - P$ and consider

$$S = \tilde{E}_{i:}w^{(i)} = \sum_{j\in[n_1]\setminus\{i\}, l\in[n_2]} R_{il}R_{jl}w_j^{(i)}$$

where $w^{(i)} \in \mathbb{R}^{n_1}$ depends on $A_{-i} = (A_{i'j})_{i'\neq i, j\in[n_2]}$.

Conditionally on A_{-i}, S is a weighted sum of independent and centered Bernoulli's r.v. Hence, by Lemma F.3 in [23] with $\delta = n_1^{-c}$, and weights $\tilde{w}_{jl} = R_{jl}w_j^{(i)}$ we obtain

$$\mathbb{P}\left(S \gtrsim \frac{\log n_1}{\log\log n_1}\left(\|\tilde{w}\|_\infty + \sqrt{p_{max}}\,\|\tilde{w}\|_F\right)\middle| A_{-i}\right) \le n^{-c}.$$

Since $R_{jl} \le 1$, we have $\|\tilde{w}\|_\infty \le \|w\|_\infty$. By definition we have

$$\|\tilde{w}\|_F^2 = \sum_{j\neq i, l} R_{jl}^2 (w_j^{(i)})^2.$$

Fact With probability at least $1 - e^{-\Theta(n_2 p_{max})}$

$$\max_{j\neq k}\sum_l R_{jl}^2 \lesssim n_2 p_{max}.$$

Proof of the Fact.. We have $R_{jl}^2 \le 1$ and $\mathrm{Var}(\sum_l R_{jl}^2) \le 2n_2p_{max}$. Hence by Bernstein inequality,

$$\mathbb{P}\left(\left|\sum_l R_{jl}^2 - \sum_l \mathbb{E}R_{jl}^2\right| \gtrsim n_2 p_{max}\right) \le e^{-\Theta(n_2 p_{max})}.$$

We can conclude by a union bound and the fact that $n_2 p_{max} \gtrsim \log n_1$ by assumptions on the sparsity level p_{max} and $n_2 \gtrsim n_1 \log^2 n_1$. ☐

Let us denote by Ω_1 the event under which the inequality of the previous fact holds. Note that this event only depends on A_{-i}. We have $\mathbb{P}(\Omega_1^c) \leq e^{-\Theta(n_2 p_{max})}$. Observe that under Ω_1 we have

$$\|\tilde{w}\| \lesssim \sqrt{n_2 p_{max}} \left\| w^{(i)} \right\|.$$

By consequence,

$$\begin{aligned}
&\mathbb{P}\left(S \gtrsim \frac{\log n_1}{\log\log n_1} \left(\left\| w^{(i)} \right\|_\infty + \sqrt{n_2} p_{max} \left\| w^{(i)} \right\| \right) \right) \\
&\leq \mathbb{P}\left(S \gtrsim \frac{\log n_1}{\log\log n_1} \left(\left\| w^{(i)} \right\|_\infty + \sqrt{n_2} p_{max} \left\| w^{(i)} \right\| \right) \middle| \Omega_1 \right) + e^{-\Theta(n_2 p_{max})} \\
&\leq \mathbb{E}_{\Omega_1} \mathbb{P}\left(S \gtrsim \frac{\log n_1}{\log\log n_1} \left(\left\| w^{(i)} \right\|_\infty + \sqrt{n_2} p_{max} \left\| w^{(i)} \right\| \right) \middle| A_{-i} \right) + e^{-\Theta(n_2 p_{max})} \\
&\leq \mathbb{E}_{\Omega_1} \mathbb{P}\left(S \gtrsim \frac{\log n_1}{\log\log n_1} \left(\|\tilde{w}\|_\infty + \sqrt{p_{max}} \|\tilde{w}\|_F \right) \middle| A_{-i} \right) + e^{-\Theta(n_2 p_{max})} \\
&\leq n_1^{-c} + e^{-\Theta(n_2 p_{max})}
\end{aligned}$$

where $\mathbb{E}_{\Omega_1}$ denotes the expectation over A_{-i} conditioned on Ω_1.

The other terms $E'_{i:} w^{(i)}$ and $E''_{i:} w^{(i)}$ can be handled in a similar way. They are easier to treat because one doesn't need to use a conditioning argument since $E'_{i:}$, $E''_{i:}$ are independent of A_{-i}.

We have by definition

$$E''_{i:} w^{(i)} \leq n_2 p_{max}^2 \left\| w^{(i)} \right\|_\infty \ll \frac{\log n_1}{\log\log n_1} \left\| w^{(i)} \right\|_\infty.$$

Also we can decompose

$$E'_{i:} w^{(i)} = \underbrace{\sum_{j \neq i,l} R_{il} P_{jl} w_j^{(i)}}_{S_1} + \underbrace{\sum_{j \neq i,l} P_{il} R_{jl} w_j^{(i)}}_{S_2}.$$

S_1 is a sum of n_2 weighted independent Bernoulli's r.v. with weights given by $w_l = \sum_{j \neq i} P_{jl} w_j^{(i)}$. Lemma F.3 in [23] gives with probability at least $1 - n_1^{-c}$

$$\begin{aligned}
S_1 &\lesssim \frac{\log n_1}{\log\log n_1} \left(\|w\|_\infty + \sqrt{p_{max}} \|w\| \right) \\
&\lesssim \frac{\log n_1}{\log\log n_1} \left(n_1 p_{max} \left\| w^{(i)} \right\|_\infty + \sqrt{n_1} p_{max} \left\| w^{(i)} \right\| \right).
\end{aligned}$$

By a similar argument, we can show that with probability at least $1 - n_1^{-c}$

$$S_2 \lesssim \frac{\log n_1}{\log\log n_1} \left(p_{max} \left\| w^{(i)} \right\|_\infty + \sqrt{n_2} p_{max}^{1.5} \left\| w^{(i)} \right\| \right).$$

Consequently, with probability at least $1 - O(n_1^{-c})$,

$$E_{i:}w^{(i)} \lesssim \frac{\log n_1}{\log\log n_1}\left(\left\|w^{(i)}\right\|_\infty + \sqrt{n_2}p_{max}\left\|w^{(i)}\right\|\right).$$

Then, by using Proposition 2.2 (ϵ-net argument) in [23] we obtain that with probability at least $1 - O(n_1^{-c})$

$$\left\|E_{i:}W^{(i)}\right\| \lesssim \frac{\log n_1}{\log\log n_1}\left\|W^{(i)}\right\|_{2\to\infty} + \frac{\sqrt{n_2}p_{max}\log n_1}{\log\log n_1}\left\|W^{(i)}\right\|. \tag{5.3}$$

Once we have obtained this inequality, the proof of Step III is the same as in [23]. □

5.3. Proof of Theorem 2

First, we will extend Theorem 2.3 in [23]. We will use the same notations as in [23] to make the adaptation easier. Let $\Delta^* = \lambda^*_{min}$ be the effective eigengap (it corresponds with the definition in [23], with $s = 0$). In our setting, the condition number $\bar{\kappa}$ only depends on K and L and hence is considered as a constant. We have shown in Section 5.1 and 5.2 that the following conditions (partially matching the assumptions (A1)-(A4) in [23]) hold with $\delta = n_1^{-q}$ for some constant $q > 0$.

Condition C1. *There exists a constant $C_1 > 0$ such that with probability at least $1 - O(n_1^{-q})$ the following conditions hold*

1. $\left\|B^{(i)} - B\right\| \le L_1(\delta) := C_1\sqrt{n_1 n_2 p_{max}^2}$,
2. $\frac{\left\|(B^{(i)}-B)U)\right\|}{\lambda_r^*} \le C_1 \left\|U\right\|_{2\to\infty}$.

The above inequalities show that in the notation of Assumption (A1) in [23], we may choose $L_2(\delta), L_3(\delta)$, and $\kappa(\Lambda^*) = \bar{\kappa}$ as bounded functions.

Condition C2. *There exists a constant $C_2 > 0$ such that with probability at least $1 - O(n_1^{-q})$ the following inequalities hold*

1. $\|\Lambda - \Lambda^*\| \le \lambda_-(\delta) := C_2\sqrt{n_1 n_2 p_{max}^2}$,
2. $\|EU^*\| \le E_+(\delta) := C_2\sqrt{n_1 n_2 p_{max}^2}$,
3. $\|E\|_{2\to\infty} \le E_\infty(\delta) = C_2\sqrt{n_1 n_2 p_{max}^2}$.

Condition C3. *For any $i \in [n_1]$ and fixed matrix $W \in \mathbb{R}^{n_1 \times r}$,*

$$\|E_{i:}W\| \le b_\infty(\delta)\|W\|_{2\to\infty} + b_2(\delta)\|W\|, \text{ with probability at least } 1 - O(n_1^{-q})$$

where $b_\infty(\delta) \lesssim \frac{\log n_1}{\log\log n_1}$ *and* $b_2(\delta) \lesssim \frac{\sqrt{n_2}p_{max}\log n_1}{\log\log n_1}$.

Condition C4. *We have $\Delta^* \ge 4(\sigma(\delta) + L_1(\delta) + \lambda_-(\delta))$ where $\sigma(\delta) = E_\infty(\delta) + b_\infty(\delta) + b_2(\delta) + E_+(\delta)$.*

Theorem 3. *Let $\delta = n_1^{-q}$ for some constant $q > 0$. Then under conditions C1-C4 and the assumptions of Theorem 2, there exists a constant $C_3 > 0$ such that with probability at least $1 - O(n_1^{-q})$*

$$d_{2\to\infty}(U, BU^*(\Lambda^*)^{-1}) \leq \frac{C_3}{\Delta^*}\sigma(\delta)\left(\|U^*\|_{2\to\infty} + \frac{\|EU^*\|_{2\to\infty}}{\lambda^*_{min}}\right) + \frac{C_3}{\Delta^*}\left(\frac{E_+(\delta)b_2(\delta)}{\lambda^*_{min}} + \frac{E_+(\delta)}{\sqrt{n_1}}\right).$$

Proof. We cannot directly apply Theorem 2.3 in [23] because Condition C1 doesn't include the condition stated in (5.2). But this condition is only used in the Step III. of Theorem 2.3 where one needs to control $\left\|E_{i:}(U^{(i)}H^{(i)} - U^*)\right\|$. We used a different argument to control this quantity in Section 5.2 and we obtained by equation (5.3)

$$\left\|E_{i:}(U^{(i)}H^{(i)} - U^*)\right\| \leq b_\infty(\delta)\left\|(U^{(i)}H^{(i)} - U^*)\right\|_{2\to\infty} + b_2(\delta)\left\|(U^{(i)}H^{(i)} - U^*)\right\|.$$

This concludes Step III in Theorem 2.3 in [23]. □

Corollary 3. *Under the same assumption as in Theorem 3, there is a constant $c > 0$ (possibly depending on q) such that with probability at least $1 - O(n_1^{-q})$*

$$d_{2\to\infty}(U, U^*) \leq \frac{c}{\sqrt{n_1}}.$$

Proof. By triangular inequality

$$d_{2\to\infty}(U, U^*) \leq d_{2\to\infty}(U, BU^*(\Lambda^*)^{-1}) + d_{2\to\infty}(BU^*(\Lambda^*)^{-1}, U^*).$$

Notice that $U^* = B^*U^*(\Lambda^*)^{-1}$, so

$$d_{2\to\infty}(BU^*(\Lambda^*)^{-1}, U^*) \leq \left\|(B - B^*)U^*(\Lambda^*)^{-1}\right\|_{2\to\infty} \leq \frac{\|EU^*\|_{2\to\infty}}{\lambda^*_{min}}.$$

We can bound $\|EU^*\|_{2\to\infty}$ by using inequality 5) in Lemma 2. We obtain that with probability at least $1 - n_1^{-q}$

$$\|EU^*\|_{2\to\infty} \lesssim \log n_1 \|U^*\|_{2\to\infty} + \sqrt{p_{max}\log n_1}.$$

Hence, with probability at least $1 - n_1^{-q}$,

$$\frac{\|EU^*\|_{2\to\infty}}{\lambda^*_{min}} \lesssim \frac{\log n_1}{n_1 n_2 p_{max}^2}\frac{1}{\sqrt{n_1}}.$$

It is easy to check that

$$\sigma(\delta) = o(\log n_1)$$
$$\frac{E_+(\delta)b_2(\delta)}{\Delta^*\lambda^*_{min}} = o\left(\frac{1}{\sqrt{n_1}}\right)$$

$$\frac{E_+(\delta)}{\Delta^*\sqrt{n_1}} = o\left(\frac{1}{\sqrt{n_1}}\right).$$

By consequence, triangular inequality and Theorem 3 implies that w.h.p.

$$d_{2\to\infty}(U, U^*) \leq \frac{c}{\sqrt{n_1}}$$

for a constant $c > 0$ that can be made small enough if the constant C such that $n_1 n_2 p_{max}^2 \geq C \log n_1$ is chosen large enough. □

5.4. Proof of Corollary 1

The proof is standard, but for completeness, we outline it. First, we need to relate the `k-medians` algorithm with the $\ell_{2\to\infty}$ perturbation bounds. It can be done by the following lemma.

Lemma 5 ([23]). *Let $U, U^* \in \mathbb{R}^{n\times r}$ be two matrices with orthonormal columns. Then the `k-medians` algorithm exactly recovers the clusters $\mathcal{C}_1, \dots, \mathcal{C}_K$ if*

$$d_{2\to\infty}(U, U^*) \leq \frac{1}{6\alpha} \min_{i,j\in[n_1]:z_i\neq z_j} \left\|U^*_{i:} - U^*_{j:}\right\|.$$

Since by Theorem 2 we have $d_{2\to\infty}(U, U^*) \leq \frac{c}{\sqrt{n_1}}$ and by Assumption A2 $\min_{i,j\in[n_1]:z_i\neq z_j} \left\|U^*_{i:} - U^*_{j:}\right\| \geq \frac{c_1}{\sqrt{n_1}}$, the assumption of Lemma 5 holds whenever $c_1/6\alpha > c$.

5.5. Proof of Corollary 2

It is sufficient to show that w.h.p. we have $\hat{r} = r$. But this is a straightforward consequence of Weyl's inequality and the fact that $\|B - B^*\| \lesssim \sqrt{n_1 n_2 p_{max}^2}$.

Appendix A: General concentration inequalities

In this section, we provide proof of the lemmas stated in the main text.

Lemma 6. *Assume that the assumption of Theorem 1 are satisfied. Let us denote $S = \sum_{i,j\in[n_1]} w_{ij}\langle A_{i:}, A_{j:}\rangle$ where $w_{ii} = 0$ for all i, and $w_{ij} = x_i x_j \mathbf{1}_{(i,j)\in\mathcal{L}(x)}$ where $\|x\| = 1$ and $\mathcal{L}(x)$ is the set of light pairs as defined in the proof of Theorem 1. In particular, $\|w\| \leq 1$, $\|w\|_\infty \leq \sqrt{\frac{n_2}{n_1}} p_{max}$. Recall that*

$$\mathcal{E} = \left\{\max_{l\in[n_2]} \sum_{i\in[n_1]} A_{il} \leq C_{col}\right\}.$$

We have

$$\mathbb{P}\left(\mathcal{E} \cap \{|S - \mathbb{E}S| \gtrsim \sqrt{n_1 n_2} p_{max}\}\right) \leq e^{-11n_1}.$$

Proof. We will use a similar decoupling approach as the one used in the proof of Hanson-Wright inequality, see [27]. Let $(\delta_i)_{i\in[n_1]}$ be independent Bernoulli's r.v. with parameter 1/2 and let us define the set of indices

$$\Lambda_\delta = \{i \in [n_1] : \delta_i = 1\}$$

and the random variable

$$S_\delta = \sum_{i,j} \delta_i(1-\delta_j)w_{ij}\langle A_{i:}, A_{j:}\rangle = \sum_{i\in\Lambda_\delta}\left\langle A_{i:}, \sum_{j\in\Lambda_\delta^c} w_{ij}A_{j:}\right\rangle.$$

Note that $\mathbb{E}_\delta S_\delta = S/4$. To simplify the notations we will denote by $\mathbb{E}_{\Lambda^c}(.)$ (resp. $\mathbb{E}_{\Lambda}(.)$) the expectation over $(A_{i:})_{i\in\Lambda_\delta^c}$ conditionally on δ and $(A_{i:})_{i\in\Lambda_\delta}$, S_δ (resp. the expectation over $(A_{i:})_{i\in\Lambda_\delta}$ conditionally on δ and $(A_{i:})_{i\in\Lambda_\delta^c}$, S_δ).

Upper bound of the m.g.f. of S_δ conditionally on Λ_δ Conditionally on δ and $(A_{i:})_{i\in\Lambda_\delta}$, S_δ is a weighted sum of independent Bernoulli's r.v:

$$S_\delta = \sum_{l\in[n_2]}\sum_{j\in\Lambda_\delta^c} A_{jl}\left(\sum_{i\in\Lambda_\delta} w_{ij}A_{il}\right).$$

Hence, for all $t > 0$ we have

$$\begin{aligned}
&\log\mathbb{E}_{\Lambda^c}\left(e^{t(S_\delta-\mathbb{E}_{\Lambda^c}(S_\delta))}\right) = \\
&\quad \log\mathbb{E}_{\Lambda^c}\left(e^{tS_\delta}\right) - \sum_{i\in\Lambda_\delta}\sum_{j\in\Lambda_\delta^c}\sum_{l\in[n_2]} w_{ij}tA_{il}p_{jl} \\
&\quad = \sum_{j\in\Lambda_\delta^c}\sum_{l\in[n_2]}\left(\log(e^{t\sum_{i\in\Lambda_\delta} w_{ij}A_{il}}p_{jl} + 1 - p_{jl}) - t\sum_{i\in\Lambda_\delta} w_{ij}A_{il}p_{jl}\right) \\
&\quad \le \sum_{j\in\Lambda_\delta^c}\sum_{l\in[n_2]}\left(p_{jl}(e^{t\sum_{i\in\Lambda_\delta} w_{ij}A_{il}} - 1) - t\sum_{i\in\Lambda_\delta} w_{ij}A_{il}p_{jl}\right) \\
&\qquad\qquad (\log(1+x)\ge x,\ \text{for all } x > -1) \\
&\quad \le p_{max}t^2\sum_{j\in\Lambda_\delta^c}\sum_{l\in[n_2]}\frac{e^{t\|A_{:l}\|^2\|w\|_\infty}}{2}\left(\sum_{i\in\Lambda_\delta} A_{il}w_{ij}\right)^2. \\
&\qquad\qquad (\text{by Taylor-Lagrange formula})
\end{aligned}$$

In order to upper-bound this m.g.f, it is necessary to control $\sum_{i\in[n_1]} A_{il}$ for each l. Let us define the event

$$\mathcal{E}_1 = \left\{\max_{l\in\Lambda_\delta}\sum_{i\in[n_1]} A_{il} \le C_{col}\right\}.$$

We have

$$\mathbf{1}_{\mathcal{E}_1} \log \mathbb{E}_{\Lambda^c} \left(e^{t(S_\delta - \mathbb{E}_{\Lambda^c}(S_\delta))} \right) \leq 0.5 p_{max} t^2 e^{t C_{col} \sqrt{\frac{n_2}{n_1}} p_{max}} \times \sum_{l \in [n_2]} \sum_{j \in \Lambda_\delta^c} \mathbf{1}_{\{\sum_i A_{il} \leq C_{col}\}} \left(\sum_{i \in \Lambda_\delta} A_{il} w_{ij} \right)^2 . \quad \text{(A.1)}$$

Let us denote $Z_l = \sum_{j \in \Lambda_\delta^c} \mathbf{1}_{\{\sum_i A_{il} \leq C_{col}\}} (\sum_{i \in \Lambda_\delta} A_{il} w_{ij})^2$.

Control of $\sum_l Z_l$ Conditionnaly on δ, $(Z_l)_l$ are independent r.v., so one can use Bernstein inequality to control the deviation of these r.v. from their expectation. First, observe that

$$\begin{aligned} \mathbb{E}(Z_l|\delta) &\leq \sum_{i,j,i'} \mathbb{E}(A_{il} A_{jl}) |w_{ij} w_{i'j}| \\ &\leq p_{max}^2 n_1^3 \frac{n_2}{n_1} p_{max}^2 + p_{max} n_1^2 \frac{n_2}{n_1} p_{max}^2 && (\|w\|_\infty^2 \leq \tfrac{n_2}{n_1} p_{max}^2) \\ &\lesssim n_1 n_2 p_{max}^2 . && (n_1 p_{max} = o(1)) \end{aligned}$$

Similarly, one can control the variance of Z_l by developing the square

$$\begin{aligned} \mathbb{E}(Z_l^2|\delta) &\leq \mathbb{E} \left(\sum_{j,j'} (\sum_{i \in \Lambda_\delta} A_{il} w_{ij})^2 (\sum_{i' \in \Lambda_\delta} A_{il} w_{ij'})^2 \right) \\ &\leq \sum_{j,j'} \sum_{i_1, i_1'} \sum_{i_2, i_2'} \mathbb{E}(A_{i_1 l} A_{i_1' l} A_{i_2 l} A_{i_2' l}) |w_{i_1 j} w_{i_2 j} w_{i_1' j} w_{i_2' j}| \\ &\lesssim \left(\sqrt{\frac{n_2}{n_1}} p_{max} \right)^4 (p_{max}^4 n_1^6 + p_{max}^3 n_1^5 + p_{max}^2 n_1^4 + p_{max} n_1^3) \\ &\lesssim \left(\sqrt{\frac{n_2}{n_1}} p_{max} \right)^4 p_{max} n_1^3 \\ &\lesssim n_1 n_2^2 p_{max}^5 . \end{aligned}$$

Also, notice that the positive r.v. $(Z_l)_l$ can be bounded as follows

$$\begin{aligned} Z_l &\leq \mathbf{1}_{\{\sum_i A_{il} \leq C_{col}\}} \sum_{i,i',j \in [n_1]} A_{il} A_{i'l} |w_{ij} w_{i'j}| \\ &\leq \mathbf{1}_{\{\sum_i A_{il} \leq C_{col}\}} \frac{n_2}{n_1} p_{max}^2 \sum_{i,i',j \in [n_1]} A_{il} A_{i'l} \\ &\leq C_{col}^2 n_2 p_{max}^2 . \end{aligned}$$

By Bernstein's inequality, we obtain

$$\mathbb{P} \left(\left| \sum_l Z_l - \mathbb{E}(Z_l) \right| \gtrsim n_1 n_2 p_{max}^2 \middle| \delta \right) \leq e^{-\Omega \left(\min \left(\frac{n_1^2 n_2^2 p_{max}^4}{n_1 n_2^2 p_{max}^5}, \frac{n_1 n_2 p_{max}^2}{C_{col}^2 n_2 p_{max}^2} \right) \right)} \leq e^{-\Omega(n_1)} .$$

Let us denote the event $\mathcal{E}_2 = \{\sum_l Z_l \lesssim n_1 n_2 p_{max}^2\}$.

Control of the m.g.f of S_δ conditionally on Λ_δ By plugin the previous step into (A.1), we obtain that on the event $\mathcal{E}_1 \cap \mathcal{E}_2$

$$\log \mathbb{E}_{\Lambda^c}\left(e^{t(S_\delta - \mathbb{E}_{\Lambda^c}(S_\delta))}\right) \lesssim \left(p_{max} t^2 e^{tC_{col}\sqrt{\frac{n_2}{n_1}}p_{max}} n_1 n_2 p_{max}^2\right).$$

For the choice $t = \sqrt{\frac{n_1}{n_2 p_{max}^2}}$ this upper bounds simplifies to

$$\mathbb{E}_{\Lambda^c}\left(e^{t(S_\delta - \mathbb{E}_{\Lambda^c}(S_\delta))}\right) \le e^{O(n_1^2 p_{max})} \le e^{o(n_1)} \tag{A.2}$$

since $n_1^2 p_{max} = o(n_1)$.

Control of $\mathbb{E}_{\Lambda^c} S_\delta - \mathbb{E}_A S_\delta$ Let us define the event

$$\mathcal{E}_3 = \left\{|\mathbb{E}_{\Lambda^c} S_\delta - \mathbb{E}_A S_\delta| \le C_2 \sqrt{n_1 n_2} p_{max}\right\}.$$

We have for all $t > 0$

$$\begin{aligned}\log \mathbb{E}(e^{t(\mathbb{E}_{\Lambda^c} S_\delta - \mathbb{E}_A S_\delta)}|\delta) &= \sum_{i,j.l} \log(e^{tw_{ij}p_{jl}} p_{il} + 1 - p_{il}) - tw_{ij}p_{jl}p_{il} \\ &\le \frac{n_2 p_{max}^3}{2} t^2 e^{t\sqrt{\frac{n_2}{n_1}}p_{max}^2}.\end{aligned}$$

By using Chernoff bound and the choice $t = \frac{1}{p_{max}^2}\sqrt{\frac{n_1}{n_2}}$ we obtain $\mathbb{P}(\mathcal{E}_3^c|\delta) \le e^{-cn_1}$ for some constant $c > 0$ that can be made large enough depending on the choice of C_2.

Conclusion Note that for all δ, $\mathcal{E} \subset \mathcal{E}_1$ where $\mathcal{E}_1$ only depends on $(A_{i:})_{i \in \Lambda_\delta}$. For any fixed δ, we have for $t = \sqrt{\frac{n_1}{n_2 p_{max}^2}}$ and $C_2 > 0$ large enough

$$\begin{aligned}
&\mathbb{P}(\{S_\delta - \mathbb{E}_A S_\delta \ge 2C_2\sqrt{n_1 n_2}p_{max}\} \cap \mathcal{E}|\delta) \\
&\quad = \mathbb{E}(\mathbb{P}_{\Lambda_\delta}(\{S_\delta - \mathbb{E}_A S_\delta \ge 2C_2\sqrt{n_1 n_2}p_{max}\} \cap \mathcal{E})|\delta) \\
&\quad \le \mathbb{E}(\mathbf{1}_{\mathcal{E}_1}\mathbb{P}_{\Lambda_\delta}(\{S_\delta - \mathbb{E}_A S_\delta \ge 2C_2\sqrt{n_1 n_2}p_{max}\}|\delta) \\
&\quad \le \mathbb{E}(\mathbf{1}_{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3}\mathbb{P}_{\Lambda_\delta}(S_\delta - \mathbb{E}_A S_\delta \ge 2C_2\sqrt{n_1 n_2}p_{max})|\delta) + e^{-cn_1} \\
&\quad \le \mathbb{E}(\mathbf{1}_{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3}\mathbb{P}_{\Lambda_\delta}(S_\delta - \mathbb{E}_{\Lambda^c} S_\delta \ge C_2\sqrt{n_1 n_2}p_{max})|\delta) + e^{-cn_1} \\
&\quad \le \mathbb{E}\left(\mathbf{1}_{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3} e^{-tC_2\sqrt{n_1 n_2}p_{max}} \mathbb{E}_{\Lambda^c} e^{t(S_\delta - \mathbb{E}_{\Lambda^c} S_\delta)} \middle| \delta\right) + e^{-cn_1} \\
&\qquad\qquad\qquad\qquad \text{(by using Chernoff bound)} \\
&\quad \le e^{-tC_2\sqrt{n_1 n_2}p_{max}} e^{0.01n_1} + e^{-cn_1} \qquad \text{(by using (A.2))} \\
&\quad \le e^{-C_2 n_1 + 0.01 n_1} + e^{-cn_1} \qquad \text{(by replacing } t \text{ by its value)} \\
&\quad \lesssim e^{-cn_1}.
\end{aligned}$$

By a union bound we obtain

$$\mathbb{P}\left(\exists \delta, \mathcal{E} \cap \{S_\delta - \mathbb{E}_A S_\delta \le 2C_2\sqrt{n_1 n_2}p_{max}\}\right) \lesssim 2^{n_1} e^{-cn_1} \lesssim e^{-c' n_1}$$

for a constant $c' > 0$. It follows that, on $\mathcal{E}$, with probability at least $1 - e^{-c' n_1}$

$$S - \mathbb{E}(S) = 4\mathbb{E}_\delta(S_\delta - \mathbb{E}_A S_\delta) \leq 8C_2\sqrt{n_1 n_2} p_{max}.$$

The stated result of the Lemma follows by the symmetry of S (one can replace w_{ij} by $-w_{ij}$). Note that the value of c' depends only on the constants in the events we conditioned on. So, we obtain $c' > 11$ by choosing such constants large enough. □

Lemma 7. *Assume that the assumptions of Theorem 1 are satisfied. Let $I, J \subset [n_1]$ with $0 < |I| \leq |J|$. Consider $S = \sum_{i,j\in[n_1]} w_{ij}\langle A_{i:}, A_{j:}\rangle$ where $w_{ii} = 0$ for all i, and $w_{ij} = \mathbf{1}_{i\in I}\mathbf{1}_{j\in J}$. Let $c^*, C^* > 0$ large enough constants. Recall that*

$$\kappa(I,J) = \max\left(t(I,J), C^*\right)$$

where $t(I,J)$ is the unique solution of the equation

$$t\log t = c^* \frac{|J|\log\frac{en_1}{|J|}}{\delta|I||J|}.$$

Then there exists a constant $c > 0$ that can be chosen large enough such that

$$\mathbb{P}\left(\mathcal{E} \cap \{S \geq \kappa(I,J)|I||J|n_2 p_{max}^2\}\right) \leq e^{-\frac{c}{2}|J|\log(\frac{en_1}{|J|})}.$$

Proof. On the event $\mathcal{E}$, we can rewrite S as a sum of n_2 independent r.v. as follows

$$S = \sum_l \underbrace{\mathbf{1}_{\{\sum_i A_{il} \leq C_{col}\}} \sum_{\substack{i\in I, j\in J\\ i\neq j}} A_{il}A_{jl}}_{Z_l}\,.$$

Notice that $\sum_l \mathbb{E}(Z_l) \leq |I||J|n_2 p_{max}^2$, so to prove the lemma, it is sufficient to show that $\sum_l Z_l$ concentrates at an appropriate rate around its expectation. Toward this end, we will apply Bennett's inequality. By definition, Z_l is bounded by C_{col}^2. Moreover, we have

$$\begin{aligned}\mathbb{E}(Z_l^2) &\leq \sum_{i\neq j, i'\neq j'} \mathbb{E}(A_{il}A_{jl}A_{i'l}A_{j'l}) \\ &\lesssim |I|^2|J|^2 p_{max}^4 + |I||J|^2 p_{max}^3 + |I||J|p_{max}^2 \lesssim |I||J|p_{max}^2.\end{aligned}$$

By consequence,

$$\mathbb{P}\left(S - \mathbb{E}(S) \geq t\right) \leq e^{-\frac{\sigma^2}{a}h(\frac{at}{\sigma^2})}$$

where $a = C_{col}^2$, $\sigma^2 = \sum_l \mathbb{E}(Z_l - \mathbb{E}Z_l)^2 \lesssim n_2 p_{max}^2|I||J| = \delta|I||J|$ and $h(u) = (u+1)\log(u+1) - u$. For the choice $t = k(I,J)\delta|I||J|$ we obtain

$$\mathbb{P}\left(S - \mathbb{E}(S) \geq t\right) \leq e^{-\frac{c}{2}|J|\log(\frac{en_1}{|J|})}.$$

Indeed, first consider the case where $t = t(I,J)\delta|I||J|$. W.l.o.g. we can assume that C_{col}^2 is large enough so that $at(I,J)\delta|I||J|/\sigma^2 \geq t(I,J)$. Since h is increasing we obtain $h(at/\sigma^2) \geq h(t(I,J))$. By definition of t, in this case $t(I,J) \geq C^*$. Hence

$$h(t(I,J)) \gtrsim g(t(I,J)) = c^* \frac{|J| \log \frac{en_1}{|J|}}{\delta|I||J|}$$

where $g(u) = u \log u$. By consequence, $\frac{\sigma^2}{a} h(\frac{at}{\sigma^2}) \gtrsim c^*|J| \log \frac{en_1}{|J|}$. The case where $t = C^*\delta|I||J|$ is similar. We have $h(at/\sigma^2)$ that is of constant order (the constant can be made large enough if C^* is large), so

$$\frac{\sigma^2}{a} h\left(\frac{at}{\sigma^2}\right) \gtrsim \delta|I||J|.$$

But by definition of t, we have $\delta|I||J| \geq c^*|J| \log \frac{en_1}{|J|}$. □

Acknowledgments

The work leading to the preliminary version of this manuscript was carried out while G.B. was a PhD student at Inria Lille in the MODAL team. We are grateful to Hemant Tyagi for providing valuable feedback on the preliminary version of the manuscript. Special appreciation goes to Yizhe Zhu for meticulously identifying inaccuracies within the proofs presented in the preceding version of the manuscript, contributing significantly to its improvement. We would also like to express our sincere gratitude to the anonymous reviewers whose insightful comments and constructive feedback helped in refining the manuscript.

References

[1] Abbe, E. (2018). Community detection and stochastic block models. *Foundations and Trends® in Communications and Information Theory* **14** 1–162. MR3827065

[2] Abbe, E., Fan, J. and Wang, K. (2022). An ℓ_p theory of PCA and spectral clustering. *The Annals of Statistics* **50** 2359–2385. MR4474494

[3] Abbe, E., Fan, J., Wang, K. and Zhong, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics* **48** 1452–1474. MR4124330

[4] Braun, G. and Tyagi, H. (2023). Minimax optimal clustering of bipartite graphs with a generalized power method. *Information and Inference: A Journal of the IMA* **12** 1830–1866. MR4598789

[5] Cai, C., Li, G., Chi, Y., Poor, H. V. and Chen, Y. (2021). Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *The Annals of Statistics* **49** 944–967. MR4255114

[6] Cape, J., Tang, M. and Priebe, C. E. (2019). The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics* **47**. MR3988761

[7] Cape, J., Tang, M. and Priebe, C. E. (2019). Signal-plus-noise matrix models: eigenvector deviations and fluctuations. *Biometrika* **106** 243–250. MR3912394

[8] Chen, Y., Fan, J., Ma, C. and Wang, K. (2019). Spectral method and regularized MLE are both optimal for top-K ranking. *The Annals of Statistics* **47**. MR3953449

[9] Cohen-Addad, V., Gupta, A., Kumar, A., Lee, E. and Li, J. (2019). Tight FPT approximations for k-median and k-means. In *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019). Leibniz International Proceedings in Informatics* **132** 42:1–42:14. MR3984859

[10] Damle, A. and Sun, Y. (2020). Uniform bounds for invariant subspace perturbations. *SIAM Journal on Matrix Analysis and Applications* **41** 1208–1236. MR4138209

[11] d'Aspremont, A., Cucuringu, M. and Tyagi, H. (2021). Ranking and synchronization from pairwise measurements via SVD. *Journal of Machine Learning Research* **22** 1–63. MR4253712

[12] Dhara, S., Gaudio, J., Mossel, E. and Sandon, C. *Spectral recovery of binary censored block models* In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)* 3389–3416. MR4415171

[13] Dhara, S., Gaudio, J., Mossel, E. and Sandon, C. (2022). Spectral algorithms optimally recover planted sub-structures. *arXiv:2203.11847*.

[14] Eldridge, J., Belkin, M. and Wang, Y. (2018). Unperturbed: spectral analysis beyond Davis-Kahan. In *Proceedings of Algorithmic Learning Theory. Proceedings of Machine Learning Research* **83** 321–358. MR3857310

[15] Fan, J., Wang, W. and Zhong, Y. (2016). An ℓ_∞ eigenvector perturbation bound and its application. *J. Mach. Learn. Res.* **18** 207:1–207:42. MR3827095

[16] Feige, U. and Ofek, E. (2005). Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms* **27** 251–275. MR2155709

[17] Florescu, L. and Perkins, W. (2016). Spectral thresholds in the bipartite stochastic block model. In *29th Annual Conference on Learning Theory* **49** 943–959.

[18] Gaudio, J. and Joshi, N. (2023). Community detection in the hypergraph SBM: Exact recovery given the similarity matrix. In *Proceedings of Thirty Sixth Conference on Learning Theory* **195** 469–510. PMLR.

[19] Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 109–137. MR0718088

[20] Huang, Z., Zeng, D. D. and Chen, H. (2007). Analyzing consumer-product graphs: Empirical findings and applications in recommender systems. *Management Science* **53** 1146–1164.

[21] Lee, J., Kim, D. and Chung, H. W. (2020). Robust hypergraph clustering via convex relaxation of truncated MLE. *IEEE Journal on Selected Areas in Information Theory* **1** 613–631.

[22] Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605

[23] LEI, L. (2020). Unified $\ell_{2\to\infty}$ eigenspace perturbation theory for symmetric random matrices. *arXiv:1909.04798*.

[24] LEI, L., LI, X. and LOU, X. (2020). Consistency of spectral clustering on hierarchical stochastic block models. *arXiv:2004.14531*.

[25] LÖFFLER, M., ZHANG, A. Y. and ZHOU, H. H. (2021). Optimality of spectral clustering in the Gaussian mixture model. *The Annals of Statistics* **49** 2506–2530. MR4338373

[26] NDAOUD, M., SIGALLA, S. and TSYBAKOV, A. B. (2022). Improved clustering algorithms for the bipartite stochastic block model. *IEEE Transactions on Information Theory* **68** 1960–1975. MR4395508

[27] RUDELSON, M. and VERSHYNIN, R. (2013). Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability* **18** 1–9. MR3125258

[28] SQUARTINI, T., ALMOG, A., CALDARELLI, G., VAN LELYVELD, I., GARLASCHELLI, D. and CIMINI, G. (2017). Enhanced capital-asset pricing model for the reconstruction of bipartite financial networks. *Phys. Rev. E* **96** 032315.

[29] STEPHAN, L. and ZHU, Y. (2023). A non-backtracking method for long matrix and tensor completion. *arXiv:2304.02077*.

[30] SU, L., WANG, W. and ZHANG, Y. (2020). Strong consistency of spectral clustering for stochastic block models. *IEEE Transactions on Information Theory* **66** 324–338. MR4053397

[31] YOUNG, J.-G., VALDOVINOS, F. and NEWMAN, M. (2021). Reconstruction of plant–pollinator networks from observational data. *Nature Communications* **12** 3911.

[32] ZHANG, A. Y. and ZHOU, H. H. (2022). Leave-one-out singular subspace perturbation analysis for spectral clustering. *arXiv:2205.14855*.