Combinatorics, Probability and Computing (2025), **34**, pp. 1–51
doi:10.1017/S0963548324000166

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Partial recovery and weak consistency in the non-uniform hypergraph stochastic block model

Ioana Dumitriu[1], Hai-Xiao Wang[1] and Yizhe Zhu[2]

[1]Department of Mathematics, University of California, San Diego, La Jolla, CA, USA and [2]Department of Mathematics, University of California, Irvine, CA, USA
Corresponding author: Hai-Xiao Wang; Email: h9wang@ucsd.edu

(Received 31 August 2022; revised 21 March 2024; accepted 25 March 2024; first published online 9 October 2024)

Abstract

We consider the community detection problem in sparse random hypergraphs under the non-uniform hypergraph stochastic block model (HSBM), a general model of random networks with community structure and higher-order interactions. When the random hypergraph has bounded expected degrees, we provide a spectral algorithm that outputs a partition with at least a γ fraction of the vertices classified correctly, where $\gamma \in (0.5, 1)$ depends on the signal-to-noise ratio (SNR) of the model. When the SNR grows slowly as the number of vertices goes to infinity, our algorithm achieves weak consistency, which improves the previous results in Ghoshdastidar and Dukkipati ((2017) *Ann. Stat.* **45**(1) 289–315.) fornon-uniform HSBMs.

Our spectral algorithm consists of three major steps: (1) Hyperedge selection: select hyperedges of certain sizes to provide the maximal signal-to-noise ratio for the induced sub-hypergraph; (2) Spectral partition: construct a regularised adjacency matrix and obtain an approximate partition based on singular vectors; (3) Correction and merging: incorporate the hyperedge information from adjacency tensors to upgrade the error rate guarantee. The theoretical analysis of our algorithm relies on the concentration and regularisation of the adjacency matrix for sparse non-uniform random hypergraphs, which can be of independent interest.

Keywords: Hypergraph; community detection; stochastic block model

2020 MSC Codes: Primary: 05C80; Secondary: 60B20

1. Introduction

Clustering is one of the central problems in network analysis and machine learning [59, 60, 65]. Many clustering algorithms make use of graph models, which represent pairwise relationships among data. A well-studied probabilistic model is the stochastic block model (SBM), which was first introduced in [39] as a random graph model that generates community structure with given ground truth for clusters so that one can study algorithm accuracy. The past decades have brought many notable results in the analysis of different algorithms and fundamental limits for community detection in SBMs in different settings [20, 37, 54, 70]. A major breakthrough was the proof of phase transition behaviours of community detection algorithms in various connectivity regimes [2, 5, 12, 52, 55, 57, 58]. See the survey [1] for more references.

Hypergraphs can represent more complex relationships among data [10, 11], including recommendation systems [13, 49], computer vision [34, 73], and biological networks [53, 68], and they have been shown empirically to have advantages over graphs [79]. Besides community detection

© The Author(s), 2024. Published by Cambridge University Press.

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

problems, sparse hypergraphs and their spectral theory have also found applications in data science [38, 40, 80], combinatorics [26, 29, 66], and statistical physics [14, 64].

With the motivation from a broad set of applications, many efforts have been made in recent years to study community detection on random hypergraphs. The hypergraph stochastic block model (HSBM), as a generalisation of graph SBM, was first introduced and studied in [31]. In this model, we observe a random uniform hypergraph where each hyperedge appears independently with some given probability depending on the community structure of the vertices in the hyperedge.

Succinctly put, the HSBM recovery problem is to find the ground truth clusters either approximately or exactly, given a sample hypergraph and estimates of model parameters. We may ask the following questions about the quality of the solutions (see [1] for further details in the graph case).

1. **Exact recovery (strong consistency):** With high probability, find all clusters exactly (up to permutation).
2. **Almost exact recovery (weak consistency):** With high probability, find a partition of the vertex set such that at most $o(n)$ vertices are misclassified.
3. **Partial recovery:** Given a fixed $\gamma \in (0.5, 1)$, with high probability, find a partition of the vertex set such that at least a fraction γ of the vertices are clustered correctly.
4. **Weak recovery (detection):** With high probability, find a partition correlated with the true partition.

For exact recovery of uniform HSBMs, it was shown that the phase transition occurs in the regime of logarithmic expected degrees in [16, 17, 50]. The thresholds are given for binary [30, 43] and multiple [77] community cases, by generalising the techniques in [2–4]. After our work appeared on arXiv, thresholds for exact recovery on non-uniform HSBMs were given by [25, 71]. Strong consistency on the degree-corrected non-uniform HSBM was studied in [24]. Spectral methods were considered in [6, 16, 21, 30, 75, 77], while *semidefinite programming* methods were analysed in [8, 43, 46]. Weak consistency for HSBMs was studied in [16, 17, 32, 33, 42].

For detection of the HSBM, the authors of [9] proposed a conjecture that the phase transition occurs in the regime of constant expected degrees. The positive part of the conjecture for the binary and multi-block case was solved in [62] and [67], respectively. Their algorithms can output a partition better than a random guess when above the *Kesten-Stigum* threshold, but can not guarantee the correctness ratio. [35, 36] proved that detection is impossible and the *Kesten-Stigum* threshold is tight for m-uniform hypergraphs with binary communities when $m = 3, 4$, while KS threshold is not tight when $m \geq 7$, and some regimes remain unknown.

1.1. Non-uniform hypergraph stochastic block model

The non-uniform HSBM was first studied in [32], which removed the uniform hypergraph assumption in previous works, and it is a more realistic model to study higher-order interaction on networks [51, 73]. It can be seen as a superposition of several uniform HSBMs with different model parameters. We first define the uniform HSBM in our setting andextend it to non-uniform hypergraphs.

Definition 1.1 (Uniform HSBM). *Let $V = \{V_1, \dots V_k\}$ be a partition of the set $[n]$ into k blocks of size $\frac{n}{k}$ (assuming n is divisible by k). Let $m \in \mathbb{N}$ be some fixed integer. For any set of m distinct vertices $i_1, \dots i_m$, a hyperedge $\{i_1, \dots i_m\}$ is generated with probability $a_m/\binom{n}{m-1}$ if vertices $i_1, \dots i_m$ are in the same block; otherwise with probability $b_m/\binom{n}{m-1}$. We denote this distribution on the set of m-uniform hypergraphs as*

$$H_m \sim \mathrm{HSBM}_m\left(\frac{n}{k}, \frac{a_m}{\binom{n}{m-1}}, \frac{b_m}{\binom{n}{m-1}}\right). \tag{1}$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

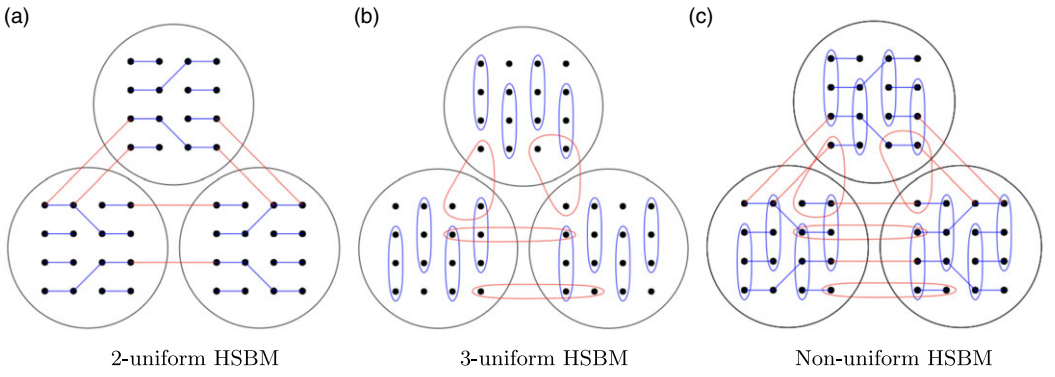


Figure 1. An example of non-uniform HSBM sampled from model 1.2.

Definition 1.2 (Non-uniform HSBM). *Let $H=(V,E)$ be a non-uniform random hypergraph, which can be considered as a collection of m-uniform hypergraphs, i.e., $H=\bigcup_{m=2}^{M} H_m$ with each H_m sampled from* (1).

Examples of 2-uniform and 3-uniform HSBM, and an exampleof non-uniform HSBM with $\mathcal{M}=\{2,3\}$ and $k=3$ is displayed in Fig. 1a, b, c respectively.

1.2. Main results

To illustrate our main results, we first introduce the concepts γ*-correctness* and *sigal-to-noise* ratio to measure the accuracy of the obtained partitions.

Definition 1.3 (γ-correctness). *Suppose we have k disjoint blocks $V_1,\dots,V_k$. A collection of subsets $\widehat{V}_1,\dots,\widehat{V}_k$ of V is γ-correct if $|V_i\cap\widehat{V}_i|\geq\gamma|V_i|$ for all $i\in[k]$.*

Definition 1.4. *For model 1.2 under Assumption 1.5, we define the signal-to-noise ratio* (SNR) *as*

$$\mathrm{SNR}_{\mathcal{M}}(k):=\frac{\left[\sum_{m\in\mathcal{M}}(m-1)\left(\frac{a_m-b_m}{k^{m-1}}\right)\right]^2}{\sum_{m\in\mathcal{M}}(m-1)\left(\frac{a_m-b_m}{k^{m-1}}+b_m\right)}. \tag{2}$$

Let $\mathcal{M}_{\max}$ denote the maximum element in the set $\mathcal{M}$. The following constant $C_{\mathcal{M}}(k)$ is used to characterise the accuracy of the clustering result,

$$C_{\mathcal{M}}(k):=\frac{[\nu^{\mathcal{M}_{\max}-1}-(1-\nu)^{\mathcal{M}_{\max}-1}]^2}{2^3\cdot(\mathcal{M}_{\max}-1)^2}\cdot\left(\mathbb{1}_{\{k=2\}}+\frac{1}{2^{2\mathcal{M}_{\max}}}\cdot\mathbb{1}_{\{k\geq3\}}\right) \tag{3}$$

Note that a non-uniform HSBM can be seen as a collection of noisy observations for the same underlying community structure through several uniform HSBMs of different orders. A possible issue is that some uniform hypergraphs with small SNR might not be informative (if we observe an m-uniform hypergraph with parameters $a_m=b_m$, including hyperedge information from it ultimately increases the noise). To improve our error rate guarantees, we start by adding a pre-processing step (Algorithm 3) for hyperedge selection according to SNR and then apply the algorithm on the sub-hypergraph with maximal SNR.

We state the following assumption that will be used in our analysis of Algorithms 1 ($k=2$) and 2 ($k\geq3$).

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

Algorithm 1. Binary Partition

Require: The adjacency tensors $\mathcal{A}^{(m)}$, a_m, b_m for $m \in \{2, \cdots, M\}$.

Ensure: The estimated sets $\widehat{V}_1, \widehat{V}_2$.

1: Run **Algorithm 3 Pre-processing** to obtain subset $\mathcal{M}$ which achieves maximal SNR.

2: Randomly colour all hyperedges red or blue with equal probability.

3: Run **Algorithm 7 Spectral Partition** on the red hypergraph and output V_1', V_2'.

4: Run **Algorithm 8 Correction** on the blue hypergraph and output $\widehat{V}_1, \widehat{V}_2$.

return The estimated sets $\widehat{V}_1, \widehat{V}_2$.

Assumption 1.5. *For each $m \in \mathcal{M}$, assume a_m, b_m are constants independent of n, and $a_m \geq b_m$. Let $\mathcal{M}_{\max}$ denote the maximum element in the set $\mathcal{M}$. Given $\nu \in (1/k, 1)$, assume that there exists a universal constant C and some ν-dependent constant $C_\nu > 0$, such that*

$$d := \sum_{m \in \mathcal{M}} (m-1)a_m \geq C\,, \tag{4a}$$

$$\sum_{m \in \mathcal{M}} (m-1)(a_m - b_m) \geq C_\nu \sqrt{d} \cdot k^{\mathcal{M}_{\max}-1} \cdot \left(2^3 \cdot \mathbb{1}_{\{k=2\}} + \sqrt{\log\left(\frac{k}{1-\nu}\right)} \cdot \mathbb{1}_{\{k \geq 3\}}\right). \tag{4b}$$

One does not have to take too large a C for (4a); for example, $C = (2^{1/\mathcal{M}_{\max}} - 1)^{-1/3}$ should suffice, but even smaller C may work. Both of the two inequalities above constant prevent the hypergraph from being too sparse, while (4b) also requires that the difference between *in-block* and *across-blocks* densities is large enough. The choices of C, C_ν and their relationship will be discussed in Remark 5.15.

1.2.1. The 2-block case

We start with Algorithm 1, which outputs a γ-correct partition when the non-uniform HSBM H is sampled from model 1.2 with only 2 communities. Inspired by the innovative graph algorithm in [19], we generalise it to non-uniform hypergraphs while we provide a complete and detailed analysis at the same time.

Theorem 1.6 ($k = 2$). *Let $\nu \in (0.5, 1)$ and $\rho = 2\exp\left(-C_{\mathcal{M}}(2) \cdot \mathrm{SNR}_{\mathcal{M}}(2)\right)$ with $\mathrm{SNR}_{\mathcal{M}}(k)$, $C_{\mathcal{M}}(k)$ defined in (2), (3), and let $\gamma = \max\{\nu,\ 1 - 2\rho\}$. Then under Assumption 1.5, Algorithm 1 outputs a γ-correct partition for sufficiently large n with probability at least $1 - O(n^{-2})$.*

1.2.2. The k-block case

For the multi-community case ($k \geq 3$), another algorithm with more subroutines is developed in Algorithm 2, which outputs a γ-correct partition with high probability. We state the result as follows.

Theorem 1.7 ($k \geq 3$). *Let $\nu \in (1/k, 1)$ and $\rho = \exp\left(-C_{\mathcal{M}}(k) \cdot \mathrm{SNR}_{\mathcal{M}}(k)\right)$ with $\mathrm{SNR}_{\mathcal{M}}(k)$, $C_{\mathcal{M}}(k)$ defined in (2), (3), and let $\gamma = \max\{\nu,\ 1 - k\rho\}$. Then under Assumption 1.5, Algorithm 2 outputs a γ-correct partition for sufficiently large n with probability at least $1 - O(n^{-2})$.*

The time complexities of Algorithms 1 and 2 are $O(n^3)$, with the bulk of time spent in Stage 1 by the spectral method.

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

Algorithm 2. General Partition

Require: The adjacency tensors $\mathcal{A}_m$, k, a_m, b_m for $m \in \mathcal{M}$.

Ensure: The estimated sets $\widehat{V}_1, \cdots, \widehat{V}_k$.

1: Run **Algorithm 3 Pre-processing** to obtain subset $\mathcal{M}$ which achieves the maximal SNR.

2: Randomly colour all hyperedges red or blue with equal probability.

3: Randomly partition V into 2 disjoint subsets Y and Z by assigning $+1$ or -1 to each vertex with equal probability.

4: Let $\mathbf{B}$ denote the adjacency matrix of the bipartite hypergraph between Y and Z consisting only of the red hyperedges, with rows indexed by Z and columns indexed by Y.

5: Run **Algorithm 4 Spectral Partition** on the red hypergraph and output $U_1', \cdots, U_k'$.

This step only uses the red hyperedges between vertices in Y and Z and outputs approximate clusters for $U_i := V_i \cap Z$, with $i = 1, \ldots, k$

6: Run **Algorithm 5 Correction** on the red hypergraph and output $\widehat{U}_1, \cdots, \widehat{U}_k$.

7: Run **Algorithm 6 Merging** on the blue hypergraph and output $\widehat{V}_1, \cdots, \widehat{V}_k$.

This step only uses the blue hyperedges between vertices in Y and Z and assigns the vertices in Y to an appropriate approximate cluster.

return The estimated sets $\widehat{V}_1, \ldots, \widehat{V}_k$.

To the best of our knowledge, Theorems 1.6 and 1.7 are the first results for partial recovery of non-uniform HSBMs. When the number of blocks is 2, Algorithm 1 guarantees a better error rate for partial recovery as in Theorem 1.6. This happens because Algorithm 1 does not need the merging routine in Algorithm 2: if one of the communities is obtained, then the other one is also obtained via the complement.

Remark 1.8. Taking $\mathcal{M} = \{2\}$, Theorem 1.7 can be reduced to [19, Lemma 9] for the graph case. The failure probability $O(n^{-2})$ can be decreased to $O(n^{-p})$ for any $p > 0$, as long as one is willing to pay the price by increasing the constants C, C_v in (4a), (4b).

Our Algorithms 1 and 2 can be summarised in 3 steps:

1. **Hyperedge selection**: select hyperedges of certain sizes to provide the maximal signal-to-noise ratio (SNR) for the induced sub-hypergraph.
2. **Spectral partition**: construct a regularised adjacency matrix and obtain an approximate partition based on singular vectors (first approximation).
3. **Correction and merging**: incorporate the hyperedge information from adjacency tensors to upgrade the error rate guarantee (second, better approximation).

The algorithm requires the input of model parameters a_m, b_m, which can be estimated by counting cycles in hypergraphs as shown in [55, 74]. Estimation of the number of blocks can be done by counting the outliers in the spectrum of the non-backtracking operator, e.g., as shown (for different regimes and different parameters) in [9, 44, 63, 67].

1.2.3. Weak consistency

Throughout the proofs for Theorems 1.6 and 1.7, we make only one assumption on the growth or finiteness of d and $\mathrm{SNR}_{\mathcal{M}}(k)$, and it happens in estimating the failure probability as noted in Remark 1.10. Consequently, the corollary below follows, which covers the case when d and $\mathrm{SNR}_{\mathcal{M}}(k)$ grow with n.

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

Corollary 1.9 (Weak consistency). *For fixed M and k, if $\mathrm{SNR}_{\mathcal{M}}(k)$ defined in (2) goes to infinity as $n \to \infty$ and $\mathrm{SNR}_{\mathcal{M}}(k) = o(\log n)$, then with probability $1 - O(n^{-2})$, Algorithms 1 and 2 output a partition with only $o(n)$ misclassified vertices.*

The paper [32] also proves weak consistency for non-uniform HSBMs, but in a much denser regime than we do here ($d = \Omega(\log^2(n))$, instead of $d = \omega(1)$, as in Corollary 1.9). In fact, we now know that strong consistency should be achievable in this denser regime, as [25] shows. When restricting to the uniform HSBM case, Corollary 1.9 achieves weak consistency under the same sparsity condition as in [7].

Remark 1.10. To be precise, Algorithms 1 and 2 work optimally in the $\mathrm{SNR}_{\mathcal{M}} = o(\log n)$ regime. When $\mathrm{SNR}_{\mathcal{M}}(k) = \Omega(\log n)$, it implies that $\rho = n^{-\Omega(1)}$, and one may have $e^{-n\rho} = \Omega(1)$ in (31), which may not decrease to 0 as $n \to \infty$. Therefore the theoretical guarantees of Algorithms 5 and 6 may not remain valid. This, however, should not matter: in the regime when $\mathrm{SNR}_{\mathcal{M}}(k) = \Omega(\log n)$, strong (rather than weak) consistency is expected, as per [25]. Therefore, the regime of interest for weak consistency is $\mathrm{SNR}_{\mathcal{M}} = o(\log n)$.

1.3. Comparison with existing results

Although many algorithms and theoretical results have been developed for hypergraph community detection, most of them are restricted to uniform hypergraphs, and few results are known for non-uniform ones. We will discuss the most relevant results.

In [42], the authors studied the degree-corrected HSBM with general connection probability parameters by using a tensor power iteration algorithm and Tucker decomposition. Their algorithm achieves weak consistency for uniform hypergraphs when the average degree is $\omega(\log^2 n)$, which is the regime complementary to the regime we studied here. They discussed a way to generalise the algorithm to non-uniform hypergraphs, but the theoretical analysis remains open. The recent paper [78] analysed non-uniform hypergraph community detection by using hypergraph embedding and optimisation algorithms and obtained weak consistency when the expected degrees are of $\omega(\log n)$, again a complementary regime to ours. Results on spectral norm concentration of sparse random tensors were obtained in [23, 40, 47, 61, 80], but no provable tensor algorithm in the bounded expected degree case is known. Testing for the community structure for non-uniform hypergraphs was studied in [41, 74], which is a problem different from community detection.

In our approach, we relied on knowing the tensors for each uniform hypergraph. However, in computations, we only ran the spectral algorithm on the adjacency matrix of the entire hypergraph since the stability of tensor algorithms does not yet come with guarantees due to the lack of concentration, and for non-uniform hypergraphs, $M - 1$ adjacency tensors would be needed. This approach presented the challenge that, unlike for graphs, the adjacency matrix of a random non-uniform hypergraph has dependent entries, and the concentration properties of such a random matrix were previously unknown. We overcame this issue and proved concentration bounds from scratch down to the bounded degree regime. Similar to [28, 45], we provided here a regularisation analysis by removing rows in the adjacency matrix with large row sums (suggestive of large degree vertices) and proving a concentration result for the regularised matrix down to the bounded expected degree regime (see Theorem 3.3).

In terms of partial recovery for hypergraphs, our results are new, even in the uniform case. In [7, Theorem 1], for uniform hypergraphs, the authors showed detection (not partial recovery) is possible when the average degree is $\Omega(1)$; in addition, the error rate is not exponential in the model parameters, but only polynomial. Here, we mention two more results for the graph case. In the arbitrarily slowly growing degrees regime, it was shown in [27, 76] that the error rate in (2) is optimal up to a constant in the exponent. In the bounded expected degrees regime, the authors in [18, 56] provided algorithms that can asymptotically recover the optimal fraction of vertices,

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

when the signal-to-noise ratio is large enough. It's an open problem to extend their analysis to obtain a minimax error rate for hypergraphs.

In [32], the authors considered weak consistency in a non-uniform HSBM model with a spectral algorithm based on the hypergraph Laplacian matrix, and showed that weak consistency is achievable if the expected degree is of $\Omega(\log^2 n)$ with high probability [31, Theorem 4.2]. Their algorithm can't be applied to sparse regimes straightforwardly since the normalised Laplacian is not well-defined due to the existence of isolated vertices in the bounded degree case. In addition, our weak consistency results obtained here are valid as long as the expected degree is $\omega(1)$ and $o(\log n)$, which is the entire set of problems on which weak consistency is expected. By contrast, in [32], weak consistency is shown only when the minimum expected degree is $\Omega(\log^2(n))$, which is a regime complementary to ours and where exact recovery should (in principle) be possible: for example, this is known to be an exact recovery regime in the uniform case [17, 43, 46, 77].

In subsequent works [25, 71] we proposed algorithms to achieve weak consistency. However, their methods can not cover the regime when the expected degree is $\Omega(1)$ due to the lack of concentration. Additionally, [72] proposed *Projected Tensor Power Method* as the refinement stage to achieve strong consistency, as long as the first stage partition is partially correct, as ours.

1.4. Organization of the paper

In Section 2, we include the definitions of adjacency matrices of hypergraphs. The concentration results for the adjacency matrices are provided in Section 3. The algorithms for partial recovery are presented in Section 4. The proof for the correctness of our algorithms for Theorem 1.7 and Corollary 1.9 are given in Section 5. The proof of Theorem 1.6, as well as the proofs of many auxiliary lemmas and useful lemmas in the literature, are provided in the supplemental materials.

2. Preliminaries

Definition 2.1 (Adjacency tensor). *Given an m-uniform hypergraph $H_m = ([n], E_m)$, we can associate to it an order-m adjacency tensor $\mathcal{A}^{(m)}$. For any m-hyperedge $e = \{i_1, \ldots, i_m\}$, let $\mathcal{A}^{(m)}_e$ denote the corresponding entry $\mathcal{A}^{(m)}_{[i_1,\ldots,i_m]}$, such that*

$$\mathcal{A}^{(m)}_e := \mathcal{A}^{(m)}_{[i_1,\ldots,i_m]} = \mathbb{1}_{\{e\in E_m\}}\,. \tag{5}$$

Definition 2.2 (Adjacency matrix). *For the non-uniform hypergraph H sampled from model 1.2, let $\mathcal{A}^{(m)}$ be the order-m adjacency tensor corresponding to the underlying m-uniform hypergraph for each $m \in \mathcal{M}$. The adjacency matrix $\mathbf{A} := [\mathbf{A}_{ij}]_{n\times n}$ of the non-uniform hypergraph H is defined by*

$$\mathbf{A}_{ij} = \mathbb{1}_{\{i\neq j\}} \cdot \sum_{m\in\mathcal{M}} \sum_{\substack{e\in E_m \\ \{i,j\}\subset e}} \mathcal{A}^{(m)}_e\,. \tag{6}$$

We compute the expectation of $\mathbf{A}$ first. In each m-uniform hypergraph H_m, two distinct vertices $i, j \in V$ with $i \neq j$ are picked arbitrarily since our model does not allow for loops. Assume for a moment $\frac{n}{k} \in \mathbb{N}$, then the expected number of m-hyperedge containing i and j can be computed as follows.

- If i and j are from the same block, the m-hyperedge is sampled with probability $a_m/\binom{n}{m-1}$ when the other $m-2$ vertices are from the same block as i, j, otherwise with probability $b_m/\binom{n}{m-1}$. Then

$$\alpha_m := \mathbb{E}\mathbf{A}_{ij} = \binom{\frac{n}{k}-2}{m-2}\frac{a_m}{\binom{n}{m-1}} + \left[\binom{n-2}{m-2} - \binom{\frac{n}{k}-2}{m-2}\right]\frac{b_m}{\binom{n}{m-1}}\,.$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

- If i and j are not from the same block, we sample the m-hyperedge with probability $b_m/\binom{n}{m-1}$, and

$$\beta_m := \mathbb{E}\mathbf{A}_{ij} = \binom{n-2}{m-2}\frac{b_m}{\binom{n}{m-1}}.$$

By assumption $a_m \geq b_m$, then $\alpha_m \geq \beta_m$ for each $m \in \mathcal{M}$. Summing over m, the *expected adjacency* matrix under the k-block non-uniform HSBM can be written as

$$\mathbb{E}\mathbf{A} = \begin{bmatrix} \alpha\mathbf{J}_{\frac{n}{k}} & \beta\mathbf{J}_{\frac{n}{k}} & \cdots & \beta\mathbf{J}_{\frac{n}{k}} \\ \beta\mathbf{J}_{\frac{n}{k}} & \alpha\mathbf{J}_{\frac{n}{k}} & \cdots & \beta\mathbf{J}_{\frac{n}{k}} \\ \vdots & \vdots & \ddots & \vdots \\ \beta\mathbf{J}_{\frac{n}{k}} & \beta\mathbf{J}_{\frac{n}{k}} & \cdots & \alpha\mathbf{J}_{\frac{n}{k}} \end{bmatrix} - \alpha\mathbf{I}_n, \tag{7}$$

where $\mathbf{J}_{\frac{n}{k}} \in \mathbb{R}^{\frac{n}{k}\times\frac{n}{k}}$ denotes the all-one matrix and

$$\alpha := \sum_{m\in\mathcal{M}} \alpha_m, \quad \beta := \sum_{m\in\mathcal{M}} \beta_m. \tag{8}$$

Lemma 2.3. *The eigenvalues of* $\mathbb{E}\mathbf{A}$ *are given below:*

$$\begin{aligned} \lambda_1(\mathbb{E}\mathbf{A}) &= \frac{n}{k}(\alpha + (k-1)\beta) - \alpha, \\ \lambda_i(\mathbb{E}\mathbf{A}) &= \frac{n}{k}(\alpha - \beta) - \alpha, \quad 2 \leq i \leq k, \\ \lambda_i(\mathbb{E}\mathbf{A}) &= -\alpha, \qquad\qquad k+1 \leq i \leq n. \end{aligned}$$

Lemma 2.3 can be verified via direct computation. Lemma 2.4 is used for approximately equi-partitions, meaning that eigenvalues of $\widetilde{\mathbb{E}\mathbf{A}}$ can be approximated by eigenvalues of $\mathbb{E}\mathbf{A}$ when n is sufficiently large.

Lemma 2.4. *For any partition* $(V_1, \ldots, V_k)$ *of* V *where* $n_i := |V_i|$, *consider the following matrix*

$$\widetilde{\mathbb{E}\mathbf{A}} = \begin{bmatrix} \alpha\mathbf{J}_{n_1} & \beta\mathbf{J}_{n_1\times n_2} & \cdots & \beta\mathbf{J}_{n_1\times n_{k-1}} & \beta\mathbf{J}_{n_1\times n_k} \\ \beta\mathbf{J}_{n_2\times n_1} & \alpha\mathbf{J}_{n_2} & \cdots & \beta\mathbf{J}_{n_2\times n_{k-1}} & \beta\mathbf{J}_{n_2\times n_k} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \beta\mathbf{J}_{n_{k-1}\times n_1} & \beta\mathbf{J}_{n_{k-1}\times n_2} & \cdots & \alpha\mathbf{J}_{n_{k-1}} & \beta\mathbf{J}_{n_{k-1}\times n_k} \\ \beta\mathbf{J}_{n_k\times n_1} & \beta\mathbf{J}_{n_k\times n_2} & \cdots & \beta\mathbf{J}_{n_k\times n_{k-1}} & \alpha\mathbf{J}_{n_k} \end{bmatrix} - \alpha\mathbf{I}_n.$$

Assume that $n_i = \frac{n}{k} + O(\sqrt{n}\log n)$ *for all* $i \in [k]$. *Then, for all* $1 \leq i \leq k$,

$$\frac{|\lambda_i(\widetilde{\mathbb{E}\mathbf{A}}) - \lambda_i(\mathbb{E}\mathbf{A})|}{|\lambda_i(\mathbb{E}\mathbf{A})|} = O\Big(n^{-\frac{1}{4}}\log^{\frac{1}{2}}(n)\Big).$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

Note that both ($\widetilde{\mathbb{E}\mathbf{A}} + \alpha \mathbf{I}_n$) and ($\mathbb{E}\mathbf{A} + \alpha \mathbf{I}_n$) are rank k matrices, then $\lambda_i(\widetilde{\mathbb{E}\mathbf{A}}) = \lambda_i(\mathbb{E}\mathbf{A}) = -\alpha$ for all $(k+1) \le i \le n$. At the same time, SNR in (2) is related to the following quantity

$$\frac{[\lambda_2(\mathbb{E}\mathbf{A})]^2}{\lambda_1(\mathbb{E}\mathbf{A})} = \frac{[(n-k)\alpha - n\beta]^2}{k[(n-k)\alpha + n(k-1)\beta]} = \frac{\left[\sum_{m\in\mathcal{M}} (m-1)\left(\frac{a_m - b_m}{k^{m-1}}\right)\right]^2}{\sum_{m\in\mathcal{M}} (m-1)\left(\frac{a_m - b_m}{k^{m-1}} + b_m\right)}(1+o(1)).$$

When $\mathcal{M} = \{2\}$ and k is fixed, SNR in (2) is equal to $\frac{(a-b)^2}{k[a+(k-1)b]}$, which corresponds to the SNR for the undirected graph in [19], see also [1, Section 6].

3. Spectral norm concentration

The correctness of Algorithms 2 and 1 relies on the concentration of the adjacency matrix of H. The following two concentration results for general random hypergraphs are included, which are independent of HSBM model. The proofs are deferred to Section A.

Theorem 3.1. *Let $H = \bigcup_{m=2}^{M} H_m$, where $H_m = ([n], E_m)$ is an Erdős-Rényi inhomogeneous hypergraph of order m for each $m \in \{2, \cdots, M\}$. Let $\mathcal{T}^{(m)}$ denote the probability tensor such that $\mathcal{T}^{(m)} = \mathbb{E}\mathcal{A}^{(m)}$ and $\mathcal{T}^{(m)}_{[i_1,\ldots,i_m]} = d_{[i_1,\ldots,i_m]}/\binom{n}{m-1}$, denoting $d_m = \max d_{[i_1,\ldots,i_m]}$. Suppose for some constant $c > 0$,*

$$d := \sum_{m=2}^{M} (m-1)\cdot d_m \ge c \log n\,. \tag{9}$$

Then for any $K > 0$, there exists a constant $C = 512M(M-1)(K+6)\,[2 + (M-1)(1+K)/c]$ such that with probability at least $1 - 2n^{-K} - 2e^{-n}$, the adjacency matrix $\mathbf{A}$ of H satisfies

$$\|\mathbf{A} - \mathbb{E}\mathbf{A}\| \le C\sqrt{d}\,. \tag{10}$$

The inequality (10) can be reduced to the result for graph case obtained in [28, 48] by taking $\mathcal{M} = \{2\}$. The result for a uniform hypergraph is obtained in [46]. Note that d is a fixed constant in our community detection problem, thus the Assumption 3.1 does not hold and the inequality (9) cannot be directly applied. However, we can still prove a concentration bound for a regularised version of $\mathbf{A}$, following the same strategy of the proof for Theorem 3.1.

Definition 3.2 (Regularized matrix). *Given any $n \times n$ matrix $\mathbf{A}$ and an index set $\mathcal{I}$, let $\mathbf{A}_{\mathcal{I}}$ be the $n \times n$ matrix obtained from $\mathbf{A}$ by zeroing out the rows and columns not indexed by $\mathcal{I}$. Namely,*

$$(\mathbf{A}_{\mathcal{I}})_{ij} = \mathbb{1}_{\{i,j\in\mathcal{I}\}} \cdot \mathbf{A}_{ij}\,. \tag{11}$$

Since every hyperedge of size m containing i is counted $(m-1)$ times in the i-th row sum of $\mathbf{A}$, the i-th row sum of $\mathbf{A}$ is given by

$$\mathrm{row}(i) := \sum_j \mathbf{A}_{ij} := \sum_j \mathbb{1}_{\{i\neq j\}} \sum_{m\in\mathcal{M}} \sum_{\substack{e\in E_m \\ \{i,j\}\subset e}} \mathcal{A}^{(m)}_e = \sum_{m\in\mathcal{M}} (m-1) \sum_{e\in E_m:\, i\in e} \mathcal{A}^{(m)}_e.$$

Theorem 3.3 is the concentration result for the regularised $\mathbf{A}_{\mathcal{I}}$, by zeroing out rows and columns corresponding to vertices with high row sums.

Theorem 3.3. *Following all the notations in Theorem 3.1, for any constant $\tau > 1$, define*

$$\mathcal{I} = \{i \in [n] : \mathrm{row}(i) \le \tau d\}.$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

Let $\mathbf{A}_{\mathcal{I}}$ *be the regularised version of* $\mathbf{A}$*, as in Definition 3.2. Then for any* $K > 0$*, there exists a constant* $C_\tau = 2((5M+1)(M-1) + \alpha_0\sqrt{\tau})$ *with* $\alpha_0 = 16 + \frac{32}{\tau}(1+e^2) + 128M(M-1)(K+4)\left(1+\frac{1}{e^2}\right)$*, such that* $\|(\mathbf{A} - \mathbb{E}\mathbf{A})_{\mathcal{I}}\| \le C_\tau\sqrt{d}$ *with probability at least* $1 - 2(e/2)^{-n} - n^{-K}$.

4. Algorithms blocks

In this section, we are going to present the algorithmic blocks constructing our main partition method (Algorithm 2): pre-processing (Algorithm 3), initial result by spectral method (Algorithm 4), correction of blemishes via majority rule (Algorithm 5), and merging (Algorithm 6).

Algorithm 3. Pre-processing

1: For each subset $\mathcal{S} \subset \{2, \cdots, M\}$, let $H_{\mathcal{S}} = \bigcup_{m\in\mathcal{S}} H_m$ denote the restriction of the non-uniform hypergraph H on $\mathcal{S}$. Compute SNR of $H_{\mathcal{S}}$, denoted by

$$\mathrm{SNR}_{\mathcal{S}} := \frac{\left[\sum_{m\in\mathcal{S}}(m-1)\left(\frac{a_m - b_m}{k^{m-1}}\right)\right]^2}{\sum_{m\in\mathcal{S}}(m-1)\left(\frac{a_m-b_m}{k^{m-1}} + b_m\right)}.$$

2: Among all the $\mathcal{S}$, find the subset $\mathcal{M}$ such that

$$\mathcal{M} := \arg\max_{\mathcal{S}\subset\{2,\cdots,M\}} \mathrm{SNR}_{\mathcal{S}},$$

with $\mathcal{M}_{\max}$ denoting its maximal element.

return $\mathcal{M}$.

Algorithm 4. Spectral Partition

1: Randomly label vertices in Y with $+1$ and -1 sign with equal probability, and partition Y into 2 disjoint subsets Y_1 and Y_2.

2: Let $\mathbf{B}_1$ (resp. $\mathbf{B}_2$) denote the adjacency matrices with all vertices in $Z \cup Y_1$, with rows indexed by Z and columns indexed by Y_1 (resp. Y_2). Pad $\mathbf{B}_1, \mathbf{B}_2$ with zeros to obtain the $n \times n$ matrices $\mathbf{A}_1$ and $\mathbf{A}_2$.

3: Let $d := \sum_{m\in\mathcal{M}}(m-1)a_m$. Zero out all the rows and columns of $\mathbf{A}_1$ corresponding to vertices whose row sum is bigger than $20\mathcal{M}_{\max}d$, to obtain the matrix $(\mathbf{A}_1)_{\mathcal{I}_1}$.

4: Find the space $\mathbf{U}$, spanned by the first k left singular vectors of $(\mathbf{A}_1)_{\mathcal{I}_1}$.

5: Randomly sample $s = 2k\log^2 n$ vertices from Y_2 without replacement. Denote the corresponding columns in $\mathbf{A}_2$ by $\boldsymbol{a}_{i_1}, \ldots, \boldsymbol{a}_{i_s}$. For each $i \in \{i_1, \cdots, i_s\}$, project $\boldsymbol{a}_i - \overline{\boldsymbol{a}}$ onto $\mathbf{U}$, where the elements in vector $\overline{\boldsymbol{a}} \in \mathbb{R}^n$ is defined by $\overline{\boldsymbol{a}}(j) = \mathbb{1}_{j\in Z}\cdot(\overline{\alpha} + \overline{\beta})/2$, with $\overline{\alpha}, \overline{\beta}$ defined in (17).

6: For each projected vector $P_{\mathbf{U}}(\boldsymbol{a}_i - \overline{\boldsymbol{a}})$, identify the top $n/(2k)$ coordinates in value as a set U_i. Discard half of the s sets U_i, those with the lowest blue hyperedge density in them.

7: Sort the remaining sets according to blue hyperedge density and identify first k distinct subsets $U_1', \cdots, U_k'$ such that $|U_i' \cap U_j'| < \lceil\frac{(1-\nu)n}{k}\rceil$ if $i \neq j$.

return $U_1', \cdots, U_k'$.

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

Algorithm 5. Correction

1: For each $u \in Z$, add u to $\widehat{U}_i$ if the number of red hyperedges, which contains u with the remaining vertices located in vertex set U_i', is at least μ_{C} in (29). Break ties arbitrarily.

return $\widehat{U}_1, \cdots, \widehat{U}_k$.

Algorithm 6. Merging

1: For all $u \in Y$, add u to $\widehat{V}_i$ if the number of blue hyperedges, which contains u with the remaining vertices located in vertex set $\widehat{U}_i$, is at least μ_{M} in (36). Label the conflicts arbitrarily.

return The estimated sets $\widehat{V}_1, \cdots, \widehat{V}_k$.

4.1. Three or more blocks ($k \geq 3$)

The proof of Theorem 1.7 is structured as follows.

Lemma 4.1. *Under the assumptions of Theorem 1.7, Algorithm 4 outputs a ν-correct partition $U_1', \cdots, U_k'$ of $Z = (Z \cap V_1) \cup \cdots \cup (Z \cap V_k)$ with probability at least $1 - O(n^{-2})$.*

Lines 4 and 6 contribute most complexity in Algorithm 4, requiring $O(n^3)$ and $O(n^2 \log^2(n))$ each (technically, one should be able to get away with $O(n^2 \log(1/\varepsilon))$ in line 4, for some desired accuracy ε to get the singular vectors). We will conservatively estimate the time complexity of Algorithm 4 as $O(n^3)$.

Lemma 4.2. *Under the assumptions of Theorem 1.7, for any ν-correct partition $U_1', \cdots, U_k'$ of $Z = (Z \cap V_1) \cup \cdots \cup (Z \cap V_k)$ and the red hypergraph over Z, Algorithm 5 computes a γ_{C}-correct partition $\widehat{U}_1, \cdots, \widehat{U}_k$ with probability $1 - O(e^{-n\rho})$, while $\gamma_{\mathrm{C}} = \max\{\nu, 1 - k\rho\}$ with $\rho := k \exp\big(-C_{\mathcal{M}}(k) \cdot \mathrm{SNR}_{\mathcal{M}}(k)\big)$ where $\mathcal{M}$ is obtained from Algorithm 3, and $\mathrm{SNR}_{\mathcal{M}}(k)$ and $C_{\mathcal{M}}(k)$ are defined in (2), (3).*

Lemma 4.3. *Given any ν-correct partition $\widehat{U}_1, \cdots, \widehat{U}_k$ of $Z = (Z \cap V_1) \cup \cdots \cup (Z \cap V_k)$ and the blue hypergraph between Y and Z, with probability $1 - O(e^{-n\rho})$, Algorithm 6 outputs a γ-correct partition $\widehat{V}_1, \cdots, \widehat{V}_k$ of $V_1 \cup V_2 \cup \cdots \cup V_k$, while $\gamma = \max\{\nu, 1 - k\rho\}$.*

The time complexities of Algorithms 5 and 6 are $O(n)$, since each vertex is adjacent to only constant many hyperedges.

4.2. The binary case ($k = 2$)

The spectral partition step is given in Algorithm 7, and the correction step is given in Algorithm 8.

Lemma 4.4. *Under the conditions of Theorem 1.6, the Algorithm 7 outputs a ν-correct partition V_1', V_2' of $V = V_1 \cup V_2$ with probability at least $1 - O(n^{-2})$.*

Lemma 4.5. *Given any ν-correct partition V_1', V_2' of $V = V_1 \cup V_2$, with probability at least $1 - O(e^{-n\rho})$, the Algorithm 8 computes a γ-correct partition $\widehat{V}_1, \widehat{V}_2$ with $\gamma = \{\nu, 1 - 2\rho\}$ and $\rho = 2\exp(-C_{\mathcal{M}}(2) \cdot \mathrm{SNR}_{\mathcal{M}}(2))$, where $\mathrm{SNR}_{\mathcal{M}}(2)$ and $C_{\mathcal{M}}(2)$ are defined in (2), (3).*

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

Algorithm 7. Spectral Partition

1: Zero out all the rows and the columns of $\mathbf{A}$ corresponding to vertices with row sums greater than $20\mathcal{M}_{\max}d$, to obtain the regularised matrix $\mathbf{A}_{\mathcal{I}}$.
2: Find the subspace $\mathbf{U}$, which is spanned by the eigenvectors corresponding to the largest two eigenvalues of $\mathbf{A}_{\mathcal{I}}$.
3: Compute $P_{\mathbf{U}}\mathbf{1}_n$, the projection of all-ones vector onto $\mathbf{U}$.
4: Let $\boldsymbol{v}$ be the unit vector in $\mathbf{U}$ perpendicular to $P_{\mathbf{U}}\mathbf{1}_n$.
5: Sort the vertices according to their values in $\boldsymbol{v}$. Let $V_1' \subset V$ be the corresponding top $n/2$ vertices, and $V_2' \subset V$ be the remaining $n/2$ vertices.
return V_1', V_2'.

Algorithm 8. Correction

1: For any $v \in V_1'$, label v "bad" if the number of blue hyperedges, which contains v with the remaining vertices in V_2' is at least μ_{C}, otherwise "good".
2: Do the same for $v \in V_2'$.
3: If $v \in V_i'$ is good, assign it to $\widehat{V}_i$, otherwise $\widehat{V}_{3-i}$.
return $\widehat{V}_1$, $\widehat{V}_2$.

5. Algorithm's correctness

We are going to present the correctness of Algorithm 2 in this section. The correctness of Algorithm 1 is deferred to Section C. We first introduce some definitions.

Vertex set splitting and adjacency matrix.
In Algorithm 2, we first randomly partition the vertex set V into two disjoint subsets Z and Y by assigning $+1$ and -1 to each vertex independently with equal probability. Let $\mathbf{B} \in \mathbb{R}^{|Z|\times|Y|}$ denote the submatrix of $\mathbf{A}$, while $\mathbf{A}$ was defined in (6), where rows and columns of $\mathbf{B}$ correspond to vertices in Z and Y respectively. Let n_i denote the number of vertices in $Z \cap V_i$, where V_i denotes the true partition with $|V_i| = \frac{n}{k}$ for all $i \in [k]$, then n_i can be written as a sum of independent Bernoulli random variables, i.e.,

$$n_i = |Z \cap V_i| = \sum_{v \in V_i} \mathbb{1}_{\{v \in Z\}}\,, \tag{12}$$

and $|Y \cap V_i| = |V_i| - |Z \cap V_i| = \frac{n}{k} - n_i$ for each $i \in [k]$.

Definition 5.1. *The splitting $V = Z \cup Y$ is perfect if $|Z \cap V_i| = |Y \cap V_i| = n/(2k)$ for all $i \in [k]$. And the splitting $Y = Y_1 \cup Y_2$ is perfect if $|Y_1 \cap V_i| = |Y_2 \cap V_i| = n/(4k)$ for all $i \in [k]$.*

However, the splitting is *imperfect* in most cases since the size of Z and Y would not be exactly the same under the independence assumption. The random matrix $\mathbf{B}$ is parameterised by $\{\boldsymbol{\mathcal{A}}^{(m)}\}_{m\in\mathcal{M}}$ and $\{n_i\}_{i=1}^k$. If we take expectation over $\{\boldsymbol{\mathcal{A}}^{(m)}\}_{m\in\mathcal{M}}$ given the block size information $\{n_i\}_{i=1}^k$, then it gives rise to the expectation of the *imperfect* splitting, denoted by $\widetilde{\mathbf{B}}$,

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

$$\widetilde{\mathbf{B}} := \begin{bmatrix} \alpha \mathbf{J}_{n_1 \times (\frac{n}{k} - n_1)} & \beta \mathbf{J}_{n_1 \times (\frac{n}{k} - n_2)} & \cdots & \beta \mathbf{J}_{n_1 \times (\frac{n}{k} - n_k)} \\ \beta \mathbf{J}_{n_2 \times (\frac{n}{k} - n_1)} & \alpha \mathbf{J}_{n_2 \times (\frac{n}{k} - n_2)} & \cdots & \beta \mathbf{J}_{n_2 \times (\frac{n}{k} - n_k)} \\ \vdots & \vdots & \ddots & \vdots \\ \beta \mathbf{J}_{n_k \times (\frac{n}{k} - n_1)} & \beta \mathbf{J}_{n_k \times (\frac{n}{k} - n_2)} & \cdots & \alpha \mathbf{J}_{n_k \times (\frac{n}{k} - n_k)} \end{bmatrix},$$

where α, β are defined in (8). In the *perfect* splitting case, the dimension of each block is $n/(2k) \times n/(2k)$ since $\mathbb{E}n_i = n/(2k)$ for all $i \in [k]$, and the expectation matrix $\overline{\mathbf{B}}$ can be written as

$$\overline{\mathbf{B}} := \begin{bmatrix} \alpha \mathbf{J}_{\frac{n}{2k}} & \beta \mathbf{J}_{\frac{n}{2k}} & \cdots & \beta \mathbf{J}_{\frac{n}{2k}} \\ \beta \mathbf{J}_{\frac{n}{2k}} & \alpha \mathbf{J}_{\frac{n}{2k}} & \cdots & \beta \mathbf{J}_{\frac{n}{2k}} \\ \vdots & \vdots & \ddots & \vdots \\ \beta \mathbf{J}_{\frac{n}{2k}} & \beta \mathbf{J}_{\frac{n}{2k}} & \cdots & \alpha \mathbf{J}_{\frac{n}{2k}} \end{bmatrix}.$$

In Algorithm 4, Y_1 is a random subset of Y obtained by selecting each element with probability $1/2$ independently, and $Y_2 = Y \setminus Y_1$. Let n_i' denote the number of vertices in $Y_1 \cap V_i$, then n_i' can be written as a sum of independent Bernoulli random variables,

$$n_i' = |Y_1 \cap V_i| = \sum_{v \in V_i} \mathbf{1}_{\{v \in Y_1\}}, \tag{13}$$

and $|Y_2 \cap V_i| = |V_i| - |Z \cap V_i| - |Y_1 \cap V_i| = n/k - n_i - n_i'$ for all $i \in [k]$.

Induced sub-hypergraph.

Definition 5.2 (Induced Sub-hypergraph). *Let $H = (V, E)$ be a non-uniform random hypergraph and $S \subset V$ be any subset of the vertices of H. Then the induced sub-hypergraph $H[S]$ is the hypergraph whose vertex set is S and whose hyperedge set $E[S]$ consists of all of the edges in E that have all endpoints located in S.*

Let $H[Y_1 \cup Z]$(resp. $H[Y_2 \cup Z]$) denote the induced sub-hypergraph on vertex set $Y_1 \cup Z$ (resp. $Y_2 \cup Z$), and $\mathbf{B}_1 \in \mathbb{R}^{|Z| \times |Y_1|}$ (resp. $\mathbf{B}_2 \in \mathbb{R}^{|Z| \times |Y_2|}$) denote the adjacency matrices corresponding to the sub-hypergraphs, where rows and columns of $\mathbf{B}_1$ (resp. $\mathbf{B}_2$) are corresponding to elements in Z and Y_1 (resp., Z and Y_2). Therefore, $\mathbf{B}_1$ and $\mathbf{B}_2$ are parameterised by $\{\boldsymbol{\mathcal{A}}^{(m)}\}_{m \in \mathcal{M}}$, $\{n_i\}_{i=1}^k$ and $\{n_i'\}_{i=1}^k$, and the entries in $\mathbf{B}_1$ are independent of the entries in $\mathbf{B}_2$, due to the independence of hyperedges. If we take expectation over $\{\boldsymbol{\mathcal{A}}^{(m)}\}_{m \in \mathcal{M}}$ conditioning on $\{n_i\}_{i=1}^k$ and $\{n_i'\}_{i=1}^k$, then it gives rise to the expectation of the *imperfect* splitting, denoted by $\widetilde{\mathbf{B}}_1$,

$$\widetilde{\mathbf{B}}_1 := \begin{bmatrix} \widetilde{\alpha}_{11} \mathbf{J}_{n_1 \times n_1'} & \cdots & \widetilde{\beta}_{1k} \mathbf{J}_{n_1 \times n_k'} \\ \vdots & \ddots & \vdots \\ \widetilde{\beta}_{k1} \mathbf{J}_{n_k \times n_1'} & \cdots & \widetilde{\alpha}_{kk} \mathbf{J}_{n_k \times n_k'} \end{bmatrix}, \tag{14}$$

where

$$\widetilde{\alpha}_{ii} := \sum_{m \in \mathcal{M}} \left\{ \binom{n_i + n_i' - 2}{m-2} \frac{a_m - b_m}{\binom{n}{m-1}} + \binom{\sum_{l=1}^k (n_l + n_l') - 2}{m-2} \frac{b_m}{\binom{n}{m-1}} \right\}, \tag{15a}$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

$$\widetilde{\beta}_{ij} := \sum_{m\in\mathcal{M}} \frac{\binom{\sum_{l=1}^{k}(n_l+n_l')-2}{m-2}}{\binom{n}{m-1}} \frac{b_m}{}, \quad i\neq j,\ i,j\in[k]\,. \tag{15b}$$

The expectation of the *perfect* splitting, denoted by $\overline{\mathbf{B}}_1$, can be written as

$$\overline{\mathbf{B}}_1 := \begin{bmatrix} \overline{\alpha}\mathbf{J}_{\frac{n}{2k}\times\frac{n}{4k}} & \overline{\beta}\mathbf{J}_{\frac{n}{2k}\times\frac{n}{4k}} & \dots & \overline{\beta}\mathbf{J}_{\frac{n}{2k}\times\frac{n}{4k}} \\ \overline{\beta}\mathbf{J}_{\frac{n}{2k}\times\frac{n}{4k}} & \overline{\alpha}\mathbf{J}_{\frac{n}{2k}\times\frac{n}{4k}} & \dots & \overline{\beta}\mathbf{J}_{\frac{n}{2k}\times\frac{n}{4k}} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{\beta}\mathbf{J}_{\frac{n}{2k}\times\frac{n}{4k}} & \overline{\beta}\mathbf{J}_{\frac{n}{2k}\times\frac{n}{4k}} & \dots & \overline{\alpha}\mathbf{J}_{\frac{n}{2k}\times\frac{n}{4k}} \end{bmatrix}, \tag{16}$$

where

$$\overline{\alpha} := \sum_{m\in\mathcal{M}} \left\{ \binom{\frac{3n}{4k}-2}{m-2} \frac{a_m-b_m}{\binom{n}{m-1}} + \binom{\frac{3n}{4}-2}{m-2} \frac{b_m}{\binom{n}{m-1}} \right\}, \quad \overline{\beta} := \sum_{m\in\mathcal{M}} \binom{\frac{3n}{4}-2}{m-2} \frac{b_m}{\binom{n}{m-1}}\,. \tag{17}$$

The matrices $\widetilde{\mathbf{B}}_2, \overline{\mathbf{B}}_2$ can be defined similarly, since dimensions of $|Y_2\cap V_i|$ are also determined by n_i and n_i'. Obviously, $\overline{\mathbf{B}}_2 = \overline{\mathbf{B}}_1$ since $\mathbb{E}n_i' = \mathbb{E}(n/k - n_i - n_i') = n/(4k)$ for all $i\in[k]$.

Fixing Dimensions.
The dimensions of $\widetilde{\mathbf{B}}_1$ and $\widetilde{\mathbf{B}}_2$, as well as blocks they consist of, are not deterministic – since n_i and n_i', defined in (12) and (13) respectively, are sums of independent random variables. As such, we cannot directly compare them. In order to overcome this difficulty, we embed $\mathbf{B}_1$ and $\mathbf{B}_2$ into the following $n\times n$ matrices:

$$\mathbf{A}_1 := \begin{bmatrix} \mathbf{0}_{|Z|\times|Z|} & \mathbf{B}_1 & \mathbf{0}_{|Z|\times|Y_2|} \\ \mathbf{0}_{|Y|\times|Z|} & \mathbf{0}_{|Y|\times|Y_1|} & \mathbf{0}_{|Y|\times|Y_2|} \end{bmatrix}, \quad \mathbf{A}_2 := \begin{bmatrix} \mathbf{0}_{|Z|\times|Z|} & \mathbf{0}_{|Z|\times|Y_1|} & \mathbf{B}_2 \\ \mathbf{0}_{|Y|\times|Z|} & \mathbf{0}_{|Y|\times|Y_1|} & \mathbf{0}_{|Y|\times|Y_2|} \end{bmatrix}. \tag{18}$$

Note that $\mathbf{A}_1$ and $\mathbf{A}_2$ have the same size. Also by definition, the entries in $\mathbf{A}_1$ are independent of the entries in $\mathbf{A}_2$. If we take expectation over $\{\boldsymbol{\mathcal{A}}^{(m)}\}_{m\in\mathcal{M}}$ conditioning on $\{n_i\}_{i=1}^k$ and $\{n_i'\}_{i=1}^k$, then we obtain the expectation matrices of the *imperfect* splitting, denoted by $\widetilde{\mathbf{A}}_1$(resp. $\widetilde{\mathbf{A}}_2$), written as

$$\widetilde{\mathbf{A}}_1 := \begin{bmatrix} \mathbf{0}_{|Z|\times|Z|} & \widetilde{\mathbf{B}}_1 & \mathbf{0}_{|Z|\times|Y_2|} \\ \mathbf{0}_{|Y|\times|Z|} & \mathbf{0}_{|Y|\times|Y_1|} & \mathbf{0}_{|Y|\times|Y_2|} \end{bmatrix}, \quad \widetilde{\mathbf{A}}_2 := \begin{bmatrix} \mathbf{0}_{|Z|\times|Z|} & \mathbf{0}_{|Z|\times|Y_1|} & \widetilde{\mathbf{B}}_2 \\ \mathbf{0}_{|Y|\times|Z|} & \mathbf{0}_{|Y|\times|Y_1|} & \mathbf{0}_{|Y|\times|Y_2|} \end{bmatrix}. \tag{19}$$

The expectation matrix of the *perfect* splitting, denoted by $\overline{\mathbf{A}}_1$(resp. $\overline{\mathbf{A}}_2$), can be written as

$$\overline{\mathbf{A}}_1 := \begin{bmatrix} \mathbf{0}_{\frac{n}{2}\times\frac{n}{2}} & \overline{\mathbf{B}}_1 & \mathbf{0}_{\frac{n}{2}\times\frac{n}{4}} \\ \mathbf{0}_{\frac{n}{2}\times\frac{n}{2}} & \mathbf{0}_{\frac{n}{2}\times\frac{n}{4}} & \mathbf{0}_{\frac{n}{2}\times\frac{n}{4}} \end{bmatrix}, \quad \overline{\mathbf{A}}_2 := \begin{bmatrix} \mathbf{0}_{\frac{n}{2}\times\frac{n}{2}} & \mathbf{0}_{\frac{n}{2}\times\frac{n}{4}} & \overline{\mathbf{B}}_2 \\ \mathbf{0}_{\frac{n}{2}\times\frac{n}{2}} & \mathbf{0}_{\frac{n}{2}\times\frac{n}{4}} & \mathbf{0}_{\frac{n}{2}\times\frac{n}{4}} \end{bmatrix}. \tag{20}$$

Obviously, $\widetilde{\mathbf{A}}_i$ and $\widetilde{\mathbf{B}}_i$(resp. $\overline{\mathbf{A}}_i$ and $\overline{\mathbf{B}}_i$) have the same non-zero singular values for $i=1,2$. In the remaining of this section, we will deal with $\widetilde{\mathbf{A}}_i$ and $\overline{\mathbf{A}}_i$ instead of $\widetilde{\mathbf{B}}_i$ and $\overline{\mathbf{B}}_i$ for $i=1,2$.

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

5.1. Spectral partition: Proof of Lemma 4.1

5.1.1. Proof outline

Recall that $\mathbf{A}_1$ is defined as the adjacency matrix of the induced sub-hypergraph $H[Y_1 \cup Z]$ in equation (5.2). Consequently, the index set should contain information only from $H[Y_1 \cup Z]$. Define the index sets

$$\mathcal{I} = \left\{ i \in [n] : \mathrm{row}(i) \le 20\mathcal{M}_{\max} d \right\}, \quad \mathcal{I}_1 = \left\{ i \in [n] : \mathrm{row}(i)\Big|_{Y_1 \cup Z} \le 20\mathcal{M}_{\max} d \right\},$$

where $d = \sum_{m \in \mathcal{M}} (m-1)a_m$, and $\mathrm{row}(i)\big|_{Y_1 \cup Z}$ is the row sum of i on $H[Y_1 \cup Z]$. We say $\mathrm{row}(i)\big|_{Y_1 \cup Z} = 0$ if $i \notin Y_1 \cup Z$, and for vertex $i \in Y_1 \cup Z$,

$$\mathrm{row}(i)\Big|_{Y_1 \cup Z} := \sum_{j=1}^{n} \sum_{m \in \mathcal{M}} \sum_{\substack{e \in E_m[Y_1 \cup Z] \\ \{i,j\} \subset e}} \mathcal{A}_e^{(m)} = \sum_{m \in \mathcal{M}} (m-1) \sum_{\substack{e \in E_m[Y_1 \cup Z] \\ \{i,j\} \subset e}} \mathcal{A}_e^{(m)}.$$

As a result, the matrix $(\mathbf{A}_1)_{\mathcal{I}_1}$ is obtained by restricting $\mathbf{A}_1$ on index set $\mathcal{I}_1$. The next 4 steps guarantee that Algorithm 4 outputs a ν-correct partition.

(i) Find the singular subspace $\mathbf{U}$ spanned by the first k left singular vectors of $(\mathbf{A}_1)_{\mathcal{I}_1}$.

(ii) Randomly pick $s = 2k \log^2 n$ vertices from Y_2 and denote the corresponding columns in $\mathbf{A}_2$ by $\boldsymbol{a}_{i_1}, \ldots, \boldsymbol{a}_{i_s}$. Project each vector $\boldsymbol{a}_i - \overline{\boldsymbol{a}}$ onto the singular subspace $\mathbf{U}$, with $\overline{\boldsymbol{a}} \in \mathbb{R}^n$ defined by $\overline{\boldsymbol{a}}(j) = \mathbb{1}_{j \in Z} \cdot (\overline{\alpha} + \overline{\beta})/2$, where $\overline{\alpha}, \overline{\beta}$ were defined in (17).

(iii) For each projected vector $P_{\mathbf{U}}(\boldsymbol{a}_i - \overline{\boldsymbol{a}})$, identify the top $n/(2k)$ coordinates in value and place the corresponding vertices into a set U_i'. Discard half of the obtained s subsets, those with the lowest *blue* edge densities.

(iv) Sort the remaining sets according to blue hyperedge density and identify k distinct subsets $U_1', \cdots, U_k'$ such that $|U_i' \cap U_j'| < \lceil (1-\nu)n/k \rceil$ if $i \neq j$.

Based on the 4 steps above in Algorithm 4, the proof of Lemma 4.1 is structured in 4 parts.

(i) Let $\widetilde{\mathbf{U}}$ denote the subspace spanned by first k left singular vectors of $\widetilde{\mathbf{A}}_1$ defined in (19). Subsection 5.1.2 shows that the subspace angle between $\mathbf{U}$ and $\widetilde{\mathbf{U}}$ is smaller than any $c \in (0, 1)$ as long as a_m, b_m satisfy certain conditions depending on c.

(ii) The vector $\widetilde{\boldsymbol{\delta}}_i$, defined in (22), reflects the underlying true partition $Z \cap V_{k(i)}$ for each $i \in [s]$, where $k(i)$ denotes the membership of vertex i. Subsection 5.1.3 shows that $\overline{\boldsymbol{\delta}}_i$, an approximation of $\widetilde{\boldsymbol{\delta}}_i$ defined in (23), can be recovered by the projected vector $P_{\mathbf{U}}(\boldsymbol{a}_i - \overline{\boldsymbol{a}})$, since projection error $\|P_{\mathbf{U}}(\boldsymbol{a}_i - \overline{\boldsymbol{a}}) - \overline{\boldsymbol{\delta}}_i\|_2 < c\|\overline{\boldsymbol{\delta}}_i\|_2$ for any $c \in (0, 1)$ if a_m, b_m satisfy the desired property in part (i).

(iii) Subsection 5.1.4 indicates that the coincidence ratio between the remaining sets and the true partition is at least ν, after discarding half of the sets with the lowest blue edge densities.

(iv) Lemma 5.13 proves that we can find k distinct subsets U_i' within $k \log^2 n$ trials with high probability.

5.1.2. Bounding the angle between $\mathbf{U}$ *and* $\widetilde{\mathbf{U}}$

The angle between subspaces $\mathbf{U}$ and $\widetilde{\mathbf{U}}$ is defined as

$$\sin \angle(\mathbf{U}, \widetilde{\mathbf{U}}) := \|P_{\mathbf{U}} - P_{\widetilde{\mathbf{U}}}\|.$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

A natural idea is to apply Wedin's sin Θ Theorem (Lemma D.7). Lemma 5.3 indicates that the difference of $\sigma_i(\widetilde{\mathbf{A}}_1)$ and $\sigma_i(\overline{\mathbf{A}}_1)$ is relatively small, compared to $\sigma_i(\overline{\mathbf{A}}_1)$.

Lemma 5.3. *Let $\sigma_i(\overline{\mathbf{A}}_1)$(resp. $\sigma_i(\widetilde{\mathbf{A}}_1)$) denote the singular values of $\overline{\mathbf{A}}_1$ (resp. $\widetilde{\mathbf{A}}_1$) for all $i \in [k]$, where the matrices $\overline{\mathbf{A}}_1$ and $\widetilde{\mathbf{A}}_1$ are defined in (20) and (19) respectively. Then*

$$\sigma_1(\overline{\mathbf{A}}_1) = \frac{n\left[\overline{\alpha} + (k-1)\overline{\beta}\right]}{2\sqrt{2}k} = \frac{n}{2\sqrt{2}k} \sum_{m\in\mathcal{M}} \left[\binom{\frac{3n}{4k} - 2}{m-2} \frac{a_m - b_m}{\binom{n}{m-1}} + k \binom{\frac{3n}{4} - 2}{m-2} \frac{b_m}{\binom{n}{m-1}} \right],$$

$$\sigma_i(\overline{\mathbf{A}}_1) = \frac{n(\overline{\alpha} - \overline{\beta})}{2\sqrt{2}k} = \frac{n}{2\sqrt{2}k} \sum_{m\in\mathcal{M}} \binom{\frac{3n}{4k} - 2}{m-2} \frac{a_m - b_m}{\binom{n}{m-1}}, 2 \le i \le k,$$

$$\sigma_i(\overline{\mathbf{A}}_1) = 0, k+1 \le i \le n.$$

with $\overline{\alpha}$, $\overline{\beta}$ defined in (17). Moreover, with probability at least $1 - 2k\exp(-k\log^2(n))$,

$$\frac{|\sigma_i(\overline{\mathbf{A}}_1) - \sigma_i(\widetilde{\mathbf{A}}_1)|}{\sigma_i(\overline{\mathbf{A}}_1)} = O\left(n^{-\frac{1}{4}} \log^{\frac{1}{2}}(n)\right).$$

Therefore, with Lemma 5.3, we can write $\sigma_i(\widetilde{\mathbf{A}}_1) = \sigma_i(\overline{\mathbf{A}}_1)(1 + o(1))$. Define $\mathbf{E}_1 := \mathbf{A}_1 - \widetilde{\mathbf{A}}_1$ and its restriction on $\mathcal{I}_1$ as

$$(\mathbf{E}_1)_{\mathcal{I}_1} := (\mathbf{A}_1 - \widetilde{\mathbf{A}}_1)_{\mathcal{I}_1} = (\mathbf{A}_1)_{\mathcal{I}_1} - (\widetilde{\mathbf{A}}_1)_{\mathcal{I}_1}, \tag{21}$$

as well as $\mathbf{\Delta}_1 := (\widetilde{\mathbf{A}}_1)_{\mathcal{I}_1} - \widetilde{\mathbf{A}}_1$. Then $(\mathbf{A}_1)_{\mathcal{I}_1} - \widetilde{\mathbf{A}}_1$ is decomposed as

$$(\mathbf{A}_1)_{\mathcal{I}_1} - \widetilde{\mathbf{A}}_1 = [(\mathbf{A}_1)_{\mathcal{I}_1} - (\widetilde{\mathbf{A}}_1)_{\mathcal{I}_1}] + [(\widetilde{\mathbf{A}}_1)_{\mathcal{I}_1} - \widetilde{\mathbf{A}}_1] = (\mathbf{E}_1)_{\mathcal{I}_1} + \mathbf{\Delta}_1.$$

Lemma 5.4. *Let $d = \sum_{m\in\mathcal{M}} (m-1)a_m$, where $\mathcal{M}$ is obtained from Algorithm 3. There exists a constant $C_1 \ge (2^{1/\mathcal{M}_{\max}} - 1)^{-1/3}$ such that if $d \ge C_1$, then with probability at least $1 - \exp\left(-d^{-2}n/\mathcal{M}_{\max}\right)$, no more than $d^{-3}n$ vertices have row sums greater than $20\mathcal{M}_{\max}d$.*

Lemma 5.4 shows that the number of high-degree vertices is relatively small. Consequently, Corollary 5.5 indicates $\|\mathbf{\Delta}_1\| \le \sqrt{d}$ with high probability.

Corollary 5.5. *Assume $d \ge \max\{C_1, \sqrt{2}\}$, where C_1 is the constant in Lemma 5.4, then $\|\mathbf{\Delta}_1\| \le \sqrt{d}$ with probability at least $1 - \exp\left(-d^{-2}n/\mathcal{M}_{\max}\right)$.*

Proof of Corollary 5.5. Note that $n - |\mathcal{I}| \le d^{-3}n$ and $\mathcal{I} \subset \mathcal{I}_1$, then $n - |\mathcal{I}_1| \le d^{-3}n$. From Lemma 5.4, there are at most $d^{-3}n$ vertices with row sum greater than $20\mathcal{M}_{\max}d$ in the adjacency matrix $\mathbf{A}_1$, then the matrix $\mathbf{\Delta}_1 = (\widetilde{\mathbf{A}}_1)_{\mathcal{I}_1} - \widetilde{\mathbf{A}}_1$ has at most $2d^{-3}n^2$ non-zero entries. Every entry of $\widetilde{\mathbf{A}}_1$ in (19) is bounded by α, then,

$$\begin{aligned}
\|\mathbf{\Delta}_1\| \le \|\mathbf{\Delta}_1\|_{\mathrm{F}} &= \|(\widetilde{\mathbf{A}}_1)_{\mathcal{I}_1} - \widetilde{\mathbf{A}}_1\|_{\mathrm{F}} \\
&\le \sqrt{2d^{-3}n^2}\,\alpha = \sqrt{2d^{-3}}n \sum_{m\in\mathcal{M}} \left[\binom{\frac{n}{k} - 2}{m-2} \frac{a_m - b_m}{\binom{n}{m-1}} + \binom{n-2}{m-2} \frac{b_m}{\binom{n}{m-1}} \right] \\
&\le \sqrt{2d^{-3}} \sum_{m\in\mathcal{M}} (m-1)a_m \le \sqrt{2d^{-1}} \le \sqrt{d}.
\end{aligned}$$

□

Moreover, taking $\tau = 20\mathcal{M}_{\max}$, $K = 3$ in Theorem 3.3, with probability at least $1 - n^{-2}$

$$\|(\mathbf{E}_1)_{\mathcal{I}_1}\| \le C_3\sqrt{d}, \tag{22}$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

where constant C_3 depends on $\mathcal{M}_{\max}$. Together with upper bounds for $\|(\mathbf{E}_1)_{\mathcal{I}_1}\|$ and $\|\mathbf{\Delta}_1\|$, Lemma 5.6 shows that the angle between $\mathbf{U}$ and $\widetilde{\mathbf{U}}$ is relatively small with high probability.

Lemma 5.6. *For any $c \in (0,1)$, there exists some constant C_2 such that, if*

$$\sum_{m\in\mathcal{M}} (m-1)(a_m - b_m) \geq C_2 k^{\mathcal{M}_{\max}-1}\sqrt{d}\,,$$

then $\sin\angle(\mathbf{U},\widetilde{\mathbf{U}}) \leq c$ *with probability* $1-n^{-2}$. *Here* $\angle(\mathbf{U},\widetilde{\mathbf{U}})$ *is the angle between* $\mathbf{U}$ *and* $\widetilde{\mathbf{U}}$.

Proof of Lemma 5.6. From (22) and Corollary 5.5, with probability at least $1-n^{-2}$,

$$\|(\mathbf{A}_1)_{\mathcal{I}_1} - \widetilde{\mathbf{A}}_1\| \leq \|(\mathbf{E}_1)_{\mathcal{I}_1}\| + \|\mathbf{\Delta}_1\| \leq (C_3+1)\sqrt{d}.$$

Since $\sigma_{k+1}(\widetilde{\mathbf{A}}_1) = 0$, using Lemma 5.3 to approximate $\sigma_k(\widetilde{\mathbf{A}}_1)$, we obtain

$$\begin{aligned}
&\sigma_k(\widetilde{\mathbf{A}}_1) - \sigma_{k+1}(\widetilde{\mathbf{A}}_1) = \sigma_k(\widetilde{\mathbf{A}}_1) = (1+o(1))\sigma_k(\overline{\mathbf{A}}_1) \geq \frac{1}{2}\sigma_k(\overline{\mathbf{A}}_1)\\
&\geq \frac{n}{4\sqrt{2}k}\sum_{m\in\mathcal{M}} \binom{\frac{3n}{4k}-2}{m-2} \frac{a_m - b_m}{\binom{n}{m-1}} \geq \frac{1}{8k}\sum_{m\in\mathcal{M}} \left(\frac{3}{4k}\right)^{m-2}(m-1)(a_m-b_m)\\
&\geq \frac{1}{8k}\left(\frac{1}{2k}\right)^{\mathcal{M}_{\max}-2}\sum_{m\in\mathcal{M}}(m-1)(a_m-b_m) \geq \frac{C_2\sqrt{d}}{2^{\mathcal{M}_{\max}+1}}\,.
\end{aligned}$$

Then for any $c \in (0,1)$, we can find $C_2 = [2^{\mathcal{M}_{\max}+2}(C_3+1)/c]$ such that $\|(\mathbf{A}_1)_{\mathcal{I}_1} - \widetilde{\mathbf{A}}_1\| \leq (1-1/\sqrt{2})\sigma_k(\widetilde{\mathbf{A}}_1)$. By Wedin's Theorem (Lemma D.7), the angle $\angle(\mathbf{U},\widetilde{\mathbf{U}})$ is bounded by

$$\sin\angle(\mathbf{U},\widetilde{\mathbf{U}}) := \|P_{\mathbf{U}} - P_{\widetilde{\mathbf{U}}}\| \leq \frac{\sqrt{2}\|(\mathbf{A}_1)_{\mathcal{I}_1} - \widetilde{\mathbf{A}}_1\|}{\sigma_k(\widetilde{\mathbf{A}}_1)} \leq \frac{\sqrt{2}(C_3+1)\sqrt{d}}{C_2\sqrt{d}/2^{\mathcal{M}_{\max}+1}} = \frac{\sqrt{2}}{2}c < c\,.$$

□

5.1.3. Bound the projection error

Randomly pick $s = 2k\log^2 n$ vertices from Y_2. Let $\boldsymbol{a}_{i_1},\ldots,\boldsymbol{a}_{i_s}$, $\widetilde{\boldsymbol{a}}_{i_1},\ldots,\widetilde{\boldsymbol{a}}_{i_s}$, $\overline{\boldsymbol{a}}_{i_1},\ldots,\overline{\boldsymbol{a}}_{i_s}$ and $\boldsymbol{e}_{i_1},\ldots,\boldsymbol{e}_{i_s}$ be the corresponding columns of $\mathbf{A}_2$, $\widetilde{\mathbf{A}}_2$, $\overline{\mathbf{A}}_2$ and $\mathbf{E}_2 := \mathbf{A}_2 - \widetilde{\mathbf{A}}_2$ respectively, where $\mathbf{A}_2$, $\widetilde{\mathbf{A}}_2$ and $\overline{\mathbf{A}}_2$ were defined in (18), (19) and (20). Let $k(i)$ denote the membership of vertex i. Note that entries of vector $\widetilde{\boldsymbol{a}}_i$ are $\widetilde{\alpha}_{ii}$, $\widetilde{\beta}_{ij}$ or 0, according to the membership of vertices in Z, where $\widetilde{\alpha}_{ii}$, $\widetilde{\beta}_{ij}$ were defined in (15a), (15b). Then the corresponding vector $\widetilde{\boldsymbol{\delta}}_i \in \mathbb{R}^n$ with the entries given by

$$\widetilde{\boldsymbol{a}}_i(j) = \begin{cases} \widetilde{\alpha}_{ii}, & \text{if } j \in Z\cap V_{k(i)}\\ \widetilde{\beta}_{ij}, & \text{if } j \in Z\setminus V_{k(i)}\\ 0\,, & \text{if } j \in Y \end{cases}, \quad \widetilde{\boldsymbol{\delta}}_i(j) = \begin{cases} (\widetilde{\alpha}_{ii} - \widetilde{\beta}_{ij})/2 > 0, & \text{if } j \in Z\cap V_{k(i)}\\ (\widetilde{\beta}_{ij} - \widetilde{\alpha}_{ii})/2 < 0, & \text{if } j \in Z\setminus V_{k(i)}\\ 0, & \text{if } j \in Y \end{cases}, \tag{23}$$

can be used to recover the vertex set $Z\cap V_{k(i)}$ based on the sign of elements in $\widetilde{\boldsymbol{\delta}}_i$. However, it is hard to handle with $\widetilde{\boldsymbol{\delta}}_i$ due to the randomness of $\widetilde{\alpha}_{ii}$, $\widetilde{\beta}_{ij}$ originated from n_i and n_i'. Note that n_i and n_i' concentrate around $n/(2k)$ and $n/(4k)$ respectively as shown in Lemma 5.3. Thus a good approximation of $\widetilde{\boldsymbol{\delta}}_i$, which rules out randomness of n_i and n_i', was given by $\overline{\boldsymbol{\delta}}_i := \overline{\boldsymbol{a}}_i - \overline{\boldsymbol{a}}$, with entries given by $\overline{\boldsymbol{a}}(j) := \mathbb{1}_{j\in Z}\cdot(\overline{\alpha}+\overline{\beta})/2$, where $\overline{\alpha}$ and $\overline{\beta}$ were defined in (17), and

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

$$\overline{\boldsymbol{a}}_i(j) = \begin{cases} \overline{\alpha}, & \text{if } j \in Z \cap V_{k(i)} \\ \overline{\beta}, & \text{if } j \in Z \setminus V_{k(i)} \\ 0, & \text{if } j \in Y \end{cases}, \quad \overline{\boldsymbol{\delta}}_i(j) = \begin{cases} (\overline{\alpha} - \overline{\beta})/2 > 0, & \text{if } j \in Z \cap V_{k(i)} \\ (\overline{\beta} - \overline{\alpha})/2 < 0, & \text{if } j \in Z \setminus V_{k(i)} \\ 0, & \text{if } j \in Y \end{cases}. \tag{24}$$

By construction, $\overline{\boldsymbol{\delta}}_i$ identifies vertex set $Z \cap V_{k(i)}$ in the case of *perfect splitting* for any $i \in \{i_1, \cdots, i_s\} \cap Y_2 \cap V_{k(i)}$. However, the access to $\overline{\boldsymbol{\delta}}_i$ is limited in practice, thus the projection $P_{\mathbf{U}}(\boldsymbol{a}_i - \overline{\boldsymbol{a}})$ is used instead as an approximation of $\overline{\boldsymbol{\delta}}_i$. Lemma 5.7 proves that at least half of the projected vectors have small projection errors.

Lemma 5.7. *For any $c \in (0, 1)$, there exist constants C_1 and C_2 such that if $d > C_1$ and*

$$\sum_{m \in \mathcal{M}} (m-1)(a_m - b_m) > C_2 k^{\mathcal{M}_{\max}} \sqrt{d},$$

then among all projected vectors $P_{\mathbf{U}}(\boldsymbol{a}_i - \overline{\boldsymbol{a}})$ for $i \in \{i_1, \cdots, i_s\} \cap Y_2$, with probability $1 - O(n^{-k})$, at least half of them satisfy

$$\|P_{\mathbf{U}}(\boldsymbol{a}_i - \overline{\boldsymbol{a}}) - \overline{\boldsymbol{\delta}}_i\|_2 < c\, \|\overline{\boldsymbol{\delta}}_i\|_2. \tag{25}$$

Proof Lemma 5.7. Note that $\overline{\boldsymbol{\delta}}_i = P_{\overline{\mathbf{U}}} \overline{\boldsymbol{\delta}}_i$, where $\overline{\mathbf{U}}$ is spanned by the first k left singular vectors of $\overline{\mathbf{A}}_1$ with $\operatorname{rank}(\overline{\mathbf{A}}_1) = k$, and $\overline{\mathbf{A}}_1$, $\overline{\mathbf{A}}_2$ preserve the same eigen-subspace. The approximation error between $P_{\mathbf{U}}(\boldsymbol{a}_i - \overline{\boldsymbol{a}})$ and $\overline{\boldsymbol{\delta}}_i$ can be decomposed as

$$\begin{aligned} P_{\mathbf{U}}(\boldsymbol{a}_i - \overline{\boldsymbol{a}}) - \overline{\boldsymbol{\delta}}_i &= P_{\mathbf{U}}\big[(\boldsymbol{a}_i - \widetilde{\boldsymbol{a}}_i) + (\widetilde{\boldsymbol{a}}_i - \overline{\boldsymbol{a}}_i) + (\overline{\boldsymbol{a}}_i - \overline{\boldsymbol{a}})\big] - P_{\overline{\mathbf{U}}} \overline{\boldsymbol{\delta}}_i \\ &= P_{\mathbf{U}} \boldsymbol{e}_i + P_{\mathbf{U}}(\widetilde{\boldsymbol{a}}_i - \overline{\boldsymbol{a}}_i) + (P_{\mathbf{U}} - P_{\overline{\mathbf{U}}}) \overline{\boldsymbol{\delta}}_i \,. \end{aligned}$$

Then by triangle inequality,

$$\|P_{\mathbf{U}}(\boldsymbol{a}_i - \overline{\boldsymbol{a}}) - \overline{\boldsymbol{\delta}}_i\|_2 \le \|P_{\mathbf{U}} \boldsymbol{e}_i\|_2 + \|P_{\mathbf{U}}(\widetilde{\boldsymbol{a}}_i - \overline{\boldsymbol{a}}_i)\|_2 + \|P_{\mathbf{U}} - P_{\overline{\mathbf{U}}}\| \cdot \|\overline{\boldsymbol{\delta}}_i\|_2 \,.$$

Note that $\|\overline{\boldsymbol{\delta}}_i\| = O(n^{-\frac{1}{2}})$ and n_i' concentrates around $n/(4k)$ for each $i \in [k]$ with deviation at most $\sqrt{n} \log(n)$, then by definitions of $\overline{\alpha}$ and $\overline{\beta}$ in (17),

$$\|P_{\mathbf{U}}(\widetilde{\boldsymbol{a}}_i - \overline{\boldsymbol{a}}_i)\|_2 \le \|\widetilde{\boldsymbol{a}}_i - \overline{\boldsymbol{a}}_i\|_2 = O\Big(\big[k\sqrt{n} \log(n) (\overline{\alpha} - \overline{\beta})^2\big]^{\frac{1}{2}}\Big) = O[n^{-\frac{3}{4}} \log^{\frac{1}{2}}(n)] = o(\|\overline{\boldsymbol{\delta}}_i\|_2).$$

Meanwhile, by an argument similar to Lemma 5.6, it can be proved that $\sin \angle(\mathbf{U}, \overline{\mathbf{U}}) < c/2$ for any $c \in (0, 1)$, if constants C_1, C_2 are chosen properly, hence $\|P_{\mathbf{U}} - P_{\overline{\mathbf{U}}}\| \cdot \|\overline{\boldsymbol{\delta}}_i\|_2 < \frac{c}{2} \|\overline{\boldsymbol{\delta}}_i\|_2$. Lemma 5.8 shows that at least half of the indices from $\{i_1, \cdots, i_s\} \cap Y_2$ satisfy $\|P_{\mathbf{U}} \boldsymbol{e}_i\|_2 < \frac{c}{2} \|\overline{\boldsymbol{\delta}}_i\|_2$, which completes the proof. □

Lemma 5.8. *Let $d = \sum_{m \in \mathcal{M}_{\max}} (m-1) a_m$. For any $c \in (0, 1)$, with probability $1 - O(n^{-k \log n})$, at least $\frac{s}{2}$ of the vectors $\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_s}$ satisfy*

$$\|P_{\mathbf{U}} \boldsymbol{e}_i\|_2 \le 2\sqrt{kd(\mathcal{M}_{\max} + 2)/n} < \frac{c}{2} \|\overline{\boldsymbol{\delta}}_i\|_2 \,, \quad i \in \{i_1, \cdots, i_s\} \subset Y_2 \,.$$

Definition 5.9. *The vector $\boldsymbol{a}_i$ satisfying (25) is referred as good vector. The index of the good vector is hence referred to as a good vertex.*

To avoid introducing extra notations, let $\boldsymbol{a}_{i_1}, \ldots, \boldsymbol{a}_{i_{s_1}}$ denote *good* vectors with $i_1, \cdots, i_{s_1}$ denoting good indices. Lemma 5.8 indicates that the number of good vectors is at least $s_1 \ge \frac{s}{2} = k \log^2 n$.

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

5.1.4. Accuracy

We are going to prove the accuracy of the initial partition obtained from Algorithm 4. Lemmas 5.10, 5.11 and 5.13 are crucial in proving our results. We present the proof logic first and defer the Lemma statements later.

For each projected vector $P_{\mathbf{U}}(\boldsymbol{a}_i - \overline{\boldsymbol{a}})$, let U_i' denote the set of its largest $\frac{n}{2k}$ coordinates, where $i \in \{i_1, \cdots, i_s\}$ and $s = 2k\log^2 n$. Note that vector $\overline{\boldsymbol{\delta}}_{i_j}$ in (23) only identifies blocks $V_{k(i_j)}$ and $V \setminus V_{k(i_j)}$, which can be regarded as clustering two blocks with different sizes. By Lemma 5.7, good vectors satisfy $\|P_{\mathbf{U}}(\boldsymbol{a}_{i_j} - \overline{\boldsymbol{a}}) - \overline{\boldsymbol{\delta}}_{i_j}\|_2 < c\,\|\overline{\boldsymbol{\delta}}_{i_j}\|_2$ for any $c \in (0,1)$. Then by Lemma 5.10 (after proper normalisation), for a good index i_j, the number of vertices in U'_{i_j} clustered correctly is at least $(1 - \frac{4}{3}kc^2)\frac{n}{k}$. By choosing $c = \sqrt{3(1-\nu)/(8k)}$, the condition $|U'_{i_j} \cap V_i| > (1+\nu)/2|U'_{i_j}|$ in part (ii) of Lemma 5.11 is satisfied. In Lemma 5.6, we choose

$$C_2 = 2^{\mathcal{M}_{\max}+2}(C_3+1)/c = 2^{\mathcal{M}_{\max}+2}(C_3+1)\sqrt{8k/(3-3\nu)}\ , \tag{26}$$

where C_3 defined in (22). Hence, with high probability, all good vectors have at least μ_{T} blue hyperedges (we call this "*high blue hyperedge density*"). From Lemma 5.8, at least half of the selected vectors are good. Then, in Algorithm 4, throwing out half of the obtained sets U_i' (those with the lowest blue hyperedge density) guarantees that the remaining sets are good.

Recall that, by choosing constant appropriately, we can make the subspace angle $\sin\angle(\mathbf{U}, \overline{\mathbf{U}}) < c$ for any $c \in (0,1)$ ($\overline{\mathbf{U}}$ is spanned by the first k left singular vectors of $\overline{\mathbf{A}}_1$). Then for each vector $\overline{\boldsymbol{\delta}}_{i_1}, \cdots, \overline{\boldsymbol{\delta}}_{i_k}$ with each i_j selected from different vertex set V_j, there is a vector $P_{\mathbf{U}}(\boldsymbol{a}_{i_j} - \overline{\boldsymbol{a}})$ in $\mathbf{U}$ arbitrarily close to $\overline{\boldsymbol{\delta}}_{i_j}$, which was proved by Lemma 5.7. From (i) of Lemma 5.11, so obtained U'_{i_j} must satisfy $|U'_{i_j} \cap V_j| \geq \nu|U'_{i_j}|$ for each $j \in [k]$. The remaining thing is to select k different U'_{i_j} with each of them concentrating around distinct V_j for each $j \in [k]$. This problem is equivalent to finding k vertices in Y_2, each from a different partition class, which can be done with $k\log^2(n)$ samplings as shown in Lemma 5.13.

To summarise, this section is a more precise and quantitative version of the following argument: with high probability,

$$\left\{ i_j : \exists i \text{ s.t. } |U'_{i_j} \cap V_i| \geq \frac{1+\nu}{2}\frac{n}{k} \right\} \subset \left\{ i_j : U'_{i_j} \text{ has} \geq \mu_T \text{ blue hyperedges} \right\} \subset \left\{ i_j : \exists i \text{ s.t. } |U'_{i_j} \cap V_i| \geq \nu\frac{n}{k} \right\}.$$

Lemma 5.10 (Adapted from Lemma 23 in [19]). *Suppose n, k are such that $\frac{n}{k} \in \mathbb{N}$. Let $\boldsymbol{v}, \bar{\boldsymbol{v}} \in \mathbb{R}^n$ be two unit vectors, and let $\bar{\boldsymbol{v}}$ be such that $\frac{n}{k}$ of its entries are equal to $\frac{1}{\sqrt{n}}$ and the rest are equal to $-\frac{1}{\sqrt{n}}$. If $\sin\angle(\bar{\boldsymbol{v}}, \boldsymbol{v}) < c \leq 0.5$, then $\boldsymbol{v}$ contains at least $(1 - \frac{4}{3}kc^2)\frac{n}{k}$ positive entries $\boldsymbol{v}_i$ such that $\bar{\boldsymbol{v}}_i$ is also positive.*

Lemma 5.11. *Suppose that we are given a set $X \subset Z$ with size $|X| = n/(2k)$. Define*

$$\mu_1 := \frac{1}{2}\sum_{m\in\mathcal{M}} m(m-1)\left\{\left[\binom{\frac{\nu n}{2k}}{m} + \binom{\frac{(1-\nu)n}{2k}}{m}\right]\frac{a_m - b_m}{\binom{n}{m-1}} + \binom{\frac{n}{2k}}{m}\frac{b_m}{\binom{n}{m-1}}\right\},$$

$$\mu_2 := \frac{1}{2}\sum_{m\in\mathcal{M}} m(m-1)\left\{\left[\binom{\frac{(1+\nu)n}{4k}}{m} + (k-1)\binom{\frac{(1-\nu)n}{4k(k-1)}}{m}\right]\frac{a_m - b_m}{\binom{n}{m-1}} + \binom{\frac{n}{2k}}{m}\frac{b_m}{\binom{n}{m-1}}\right\},$$

and $\mu_{\mathrm{T}} := (\mu_1 + \mu_2)/2 \in [\mu_1, \mu_2]$. There is a constant $c > 0$ depending on k, a_m, ν such that for sufficiently large n,

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

(i) If $|X \cap V_i| \le \nu|X|$ for each $i \in [k]$, then with probability $1 - e^{-cn}$, the number of blue hyperedges in the hypergraph induced by X is at most μ_T.

(ii) Conversely, if $|X \cap V_i| \ge \frac{1+\nu}{2}|X|$ for some $i \in \{1, \dots, k\}$, then with probability $1 - e^{-cn}$, the number of blue hyperedges in the hypergraph induced by X is at least μ_T.

Remark 5.12. Lemma 5.11 is reduced to [19, Lemma 31] when $\mathcal{M} = \{2\}$.

Lemma 5.13. *Through random sampling without replacement in Step 6 of Algorithm 4, we can find at least k indices $i_1, \dots, i_k$ in Y_2 among $k \log^2 n$ samples such that with probability $1 - n^{-\Omega(\log n)}$,*

$$|U'_{i_j} \cap U'_{i_l}| \le (1-\nu)\frac{n}{k}, \text{ for any } j, l \in [k] \text{ with } j \ne l.$$

5.2. Local correction: Proof of Lemma 4.2

For notation convenience, let $U_i := Z \cap V_i$ denote the intersection of Z and true partition V_i for all $i \in [k]$. In Algorithm 2, we first colour the hyperedges with red and blue with equal probability. By running Algorithm 4 on the red hypergraph, we obtain a ν-correct partition $U'_1, \dots, U'_k$, i.e.,

$$|U_i \setminus U'_i| \le (1-\nu) \cdot |U'_i| = (1-\nu) \cdot \frac{n}{2k}, \quad \forall i \in [k]. \tag{27}$$

In the rest of this subsection, we condition on the event that (27) holds true.

Consider a hyperedge $e = \{i_1, \cdots, i_m\}$ in the underlying m-uniform hypergraph. If vertices $i_1, \cdots, i_m$ are from the same block, then e is a red hyperedge with probability $a_m/2\binom{n}{m-1}$; if vertices $i_1, \cdots, i_m$ are not from the same block, then e is a red hyperedge with probability $b_m/2\binom{n}{m-1}$. The presence of those two types of hyperedges can be denoted by

$$T_e^{(a_m)} \sim \text{Bernoulli}\left(\frac{a_m}{2\binom{n}{m-1}}\right), \quad T_e^{(b_m)} \sim \text{Bernoulli}\left(\frac{b_m}{2\binom{n}{m-1}}\right),$$

respectively. For any finite set S, let $[S]^l$ denote the family of l-subsets of S, i.e., $[S]^l = \{Z | Z \subseteq S, |Z| = l\}$. Consider a vertex $u \in U_1 := Z \cap V_1$. The weighted number of red hyperedges, which contains $u \in U_1$ with the remaining vertices in U'_j, can be written as

$$S'_{1j}(u) := \sum_{m \in \mathcal{M}} (m-1) \cdot \left\{ \sum_{e \in \mathcal{E}_{1,j}^{(a_m)}} T_e^{(a_m)} + \sum_{e \in \mathcal{E}_{1,j}^{(b_m)}} T_e^{(b_m)} \right\}, \quad u \in U_1, \tag{28}$$

where $\mathcal{E}_{1,j}^{(a_m)} := E_m([U_1]^1, [U_1 \cap U'_j]^{m-1})$ denotes the set of m-hyperedges with one vertex from $[U_1]^1$ and the other $m-1$ from $[U_1 \cap U'_j]^{m-1}$, while $\mathcal{E}_{1,j}^{(b_m)} := E_m\Big([U_1]^1, [U'_j]^{m-1} \setminus [U_1 \cap U'_j]^{m-1}\Big)$ denotes the set of m-hyperedges with one vertex in $[U_1]^1$ while the remaining $m-1$ vertices in $[U'_j]^{m-1} \setminus [U_1 \cap U'_j]^{m-1}$(not all m vertices are from V_1) with their cardinalities

$$|\mathcal{E}_{1,j}^{(a_m)}| = \binom{|U_1 \cap U'_j|}{m-1}, \quad |\mathcal{E}_{1,j}^{(b_m)}| = \left[\binom{|U'_j|}{m-1} - \binom{|U_1 \cap U'_j|}{m-1}\right].$$

We multiply $(m-1)$ in (28) as weight since the rest $m-1$ vertices are all located in U'_j, which can be regarded as u's neighbours in U'_j. According to the fact $|U'_j \cap U_j| \ge (\nu n/2k)$ in (27) and $|U'_j| = n/(2k)$ for $j \in [k]$,

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

$$\left|\mathcal{E}_{1,1}^{(a_m)}\right| \geq \binom{\frac{\nu n}{2k}}{m-1}, \quad \left|\mathcal{E}_{1,j}^{(a_m)}\right| \leq \binom{\frac{(1-\nu)n}{2k}}{m-1}, \ j \neq 1 .$$

To simplify the calculation, we take the lower and upper bound of $\left|\mathcal{E}_{1,1}^{(a_m)}\right|$ and $\left|\mathcal{E}_{1,j}^{(a_m)}\right|$ $(j \neq 1)$ respectively. By taking expectation with respect to $T_e^{(a_m)}$ and $T_e^{(b_m)}$, then for any $u \in U_1$, we have

$$\mathbb{E}S'_{11}(u) = \sum_{m\in\mathcal{M}} (m-1) \cdot \left[\binom{\frac{\nu n}{2k}}{m-1} \frac{a_m - b_m}{2\binom{n}{m-1}} + \binom{\frac{n}{2k}}{m-1} \frac{b_m}{2\binom{n}{m-1}} \right],$$

$$\mathbb{E}S'_{1j}(u) = \sum_{m\in\mathcal{M}} (m-1) \cdot \left[\binom{\frac{(1-\nu)n}{2k}}{m-1} \frac{a_m - b_m}{2\binom{n}{m-1}} + \binom{\frac{n}{2k}}{m-1} \frac{b_m}{2\binom{n}{m-1}} \right], \quad j \neq 1 .$$

By assumptions in Theorem 1.7, $\mathbb{E}S'_{11}(u) - \mathbb{E}S'_{1j}(u) = \Omega(1)$. Define

$$\mu_{\mathrm{C}} := \frac{1}{2} \sum_{m\in\mathcal{M}} (m-1) \cdot \left\{ \left[\binom{\frac{\nu n}{2k}}{m-1} + \binom{\frac{(1-\nu)n}{2k}}{m-1} \right] \frac{a_m - b_m}{2\binom{n}{m-1}} + 2 \cdot \binom{\frac{n}{2k}}{m-1} \frac{b_m}{2\binom{n}{m-1}} \right\} . \tag{29}$$

In Algorithm 5, vertex u is assigned to $\widehat{U}_i$ if it has the maximal number of neighbours in U'_i. If $u \in U_1$ is mislabelled, then one of the following events must happen:

- $S'_{11}(u) \leq \mu_{\mathrm{C}}$, meaning that u was mislabelled by Algorithm 5.
- $S'_{1j}(u) \geq \mu_{\mathrm{C}}$ for some $j \neq 1$, meaning that u survived Algorithm 5 without being corrected.

Lemma 5.14 shows that the probabilities of those two events can be bounded in terms of the SNR.

Lemma 5.14. *For sufficiently large n and any $u \in U_1 = Z \cap V_1$, we have*

$$\rho'_1 := \mathbb{P}\left(S'_{11}(u) \leq \mu_{\mathrm{C}}\right) \leq \rho , \quad \rho'_j := \mathbb{P}\left(S'_{1j}(u) \geq \mu_{\mathrm{C}}\right) \leq \rho , \ (j \neq 1), \tag{30}$$

where $\rho := \exp(-C_{\mathcal{M}} \cdot \mathrm{SNR}_{\mathcal{M}})$ with $\mathrm{SNR}_{\mathcal{M}}$ and $C_{\mathcal{M}}$ defined in (2).

As a result, the probability that either of those events happened is bounded by ρ. The number of mislabelled vertices in U_1 after Algorithm 5 is at most

$$R_1 = \sum_{t=1}^{|U_1|} \Gamma_t + \sum_{j=2}^{k} \sum_{t=1}^{|U_1 \cap U'_j|} \Lambda_t ,$$

where Γ_t (resp. Λ_t) are i.i.d indicator random variables with mean ρ'_1 (resp. $\rho'_j, j \neq 1$). Then

$$\mathbb{E}R_1 \leq \frac{n}{2k}\rho'_1 + \sum_{j=2}^{k} \frac{(1-\nu)n}{2k} \rho'_j \leq \frac{n}{2k} \cdot k\rho = \frac{n\rho}{2} .$$

Let $t_1 := n\rho/2$, where ν denotes the correctness after Algorithm 4, then by Chernoff bound (Lemma D.1),

$$\mathbb{P}\left(R_1 \geq n\rho\right) = \mathbb{P}\left(R_1 - n\rho/2 \geq t_1\right) \leq \mathbb{P}\left(R_1 - \mathbb{E}R_1 \geq t_1\right) \leq e^{-ct_1} = O(e^{-n\rho}) . \tag{31}$$

Then with probability $1 - O(e^{-n\rho})$, the fraction of mislabelled vertices in U_1 is smaller than $k\rho$, i.e., the correctness of U_1 is at least $\gamma_{\mathrm{C}} := \max\{\nu , 1 - k\rho\}$. Therefore, Algorithm 5 outputs a γ_{C}-correct partition $\widehat{U}_1, \cdots , \widehat{U}_k$ with probability $1 - O(e^{-n\rho})$.

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

5.3. Merging: Proof of Lemma 4.3

By running Algorithm 5 on the red hypergraph, we obtain a γ_C-correct partition $\widehat{U}_1, \cdots, \widehat{U}_k$ where $\gamma_C := \max\{\nu, 1 - k\rho\} \geq \nu$, i.e.,

$$|U_j \cap \widehat{U}_j| \geq \nu \cdot |\widehat{U}_j| = \frac{\nu n}{2k}, \quad \forall j \in [k]. \tag{32}$$

In the rest of this subsection, we shall condition on this event and abbreviate $Y \cap V_l$ by $W_l := Y \cap V_l$. The failure probability of Algorithm 6 is estimated by the presence of hyperedges between vertex sets Y and Z.

Consider a hyperedge $e = \{i_1, \cdots, i_m\}$ in the underlying m-uniform hypergraph. If vertices $i_1, \cdots, i_m$ are all from the same cluster V_l, then the probability that e is an existing blue edge conditioning on the event that e is not a red edge is

$$\psi_m := \mathbb{P}\left[e \text{ is a blue edge} \,\middle|\, e \text{ is not a red edge}\right] = \frac{\frac{a_m}{2\binom{n}{m-1}}}{1 - \frac{a_m}{2\binom{n}{m-1}}} = \frac{a_m}{2\binom{n}{m-1}}(1 + o(1)), \tag{33}$$

and the presence of e can be represented by an indicator random variable $\zeta_e^{(a_m)} \sim \text{Bernoulli}(\psi_m)$. Similarly, if vertices $i_1, \cdots, i_m$ are not all from the same cluster V_l, the probability that e is an existing blue edge conditioning on the event that e is not red

$$\phi_m := \mathbb{P}\left[e \text{ is a blue edge} \,\middle|\, e \text{ is not a red edge}\right] = \frac{\frac{b_m}{2\binom{n}{m-1}}}{1 - \frac{b_m}{2\binom{n}{m-1}}} = \frac{b_m}{2\binom{n}{m-1}}(1 + o(1)), \tag{34}$$

and the presence of e can be represented by an indicator variable $\xi_e^{(b_m)} \sim \text{Bernoulli}(\phi_m)$.

For any vertex $w \in W_l := Y \cap V_l$ with fixed $l \in [k]$, we want to compute the number of hyperedges containing w with all remaining vertices located in vertex set $\widehat{U}_j$ for some fixed $j \in [k]$. Following a similar argument given in Subsection 5.2, this number can be written as

$$\widehat{S}_{lj}(w) := \sum_{m \in \mathcal{M}} (m-1) \cdot \left\{ \sum_{e \in \widehat{\mathcal{E}}_{l,j}^{(a_m)}} \zeta_e^{(a_m)} + \sum_{e \in \widehat{\mathcal{E}}_{l,j}^{(b_m)}} \xi_e^{(b_m)} \right\}, \quad w \in W_l, \tag{35}$$

where $\widehat{\mathcal{E}}_{l,j}^{(a_m)} := E_m([W_l]^1, [U_l \cap \widehat{U}_j]^{m-1})$ denotes the set of m-hyperedges with 1 vertex from $[W_l]^1$ and the other $m-1$ vertices from $[U_l \cap \widehat{U}_j]^{m-1}$, while $\widehat{\mathcal{E}}_{l,j}^{(b_m)} := E_m([W_l]^1, [\widehat{U}_j]^{m-1} \setminus [U_l \cap \widehat{U}_j]^{m-1})$ denotes the set of m-hyperedges with 1 vertex in $[W_l]^1$ while the remaining $m-1$ vertices are in $[\widehat{U}_j]^{m-1} \setminus [U_l \cap \widehat{U}_j]^{m-1}$, with their cardinalities

$$|\widehat{\mathcal{E}}_{l,j}^{(a_m)}| = \binom{|U_l \cap \widehat{U}_j|}{m-1}, \quad |\widehat{\mathcal{E}}_{l,j}^{(b_m)}| = \left[\binom{|\widehat{U}_j|}{m-1} - \binom{|U_l \cap \widehat{U}_j|}{m-1}\right].$$

Similarly, we multiply $(m-1)$ in (35) as weight since the rest $m-1$ vertices can be regarded as u's neighbours in $\widehat{U}_j$. By accuracy of Algorithm 5 in (32), $|\widehat{U}_j \cap U_j| \geq \nu n/(2k)$, then

$$|\widehat{\mathcal{E}}_{l,l}^{(a_m)}| \geq \binom{\frac{\nu n}{2k}}{m-1}, \quad |\widehat{\mathcal{E}}_{l,j}^{(a_m)}| \leq \binom{\frac{(1-\nu)n}{2k}}{m-1}, \ j \neq l.$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

Taking expectation with respect to $\zeta_e^{(a_m)}$ and $\xi_e^{(b_m)}$, for any $w \in W_l$, we have

$$\mathbb{E}\widehat{S}_{ll}(w) = \sum_{m\in\mathcal{M}} (m-1)\cdot\left[\binom{\frac{\nu n}{2k}}{m-1}(\psi_m-\phi_m)+\binom{\frac{n}{2k}}{m-1}\phi_m\right],$$

$$\mathbb{E}\widehat{S}_{lj}(w) = \sum_{m\in\mathcal{M}} (m-1)\cdot\left[\binom{\frac{(1-\nu)n}{2k}}{m-1}(\psi_m-\phi_m)+\binom{\frac{n}{2k}}{m-1}\phi_m\right],\ j\neq l\,.$$

By assumptions in Theorem 1.7, $\mathbb{E}\widehat{S}_{ll}(w) - \mathbb{E}\widehat{S}_{lj}(w) = \Omega(1)$. We define

$$\mu_{\mathrm{M}} := \frac{1}{2}\sum_{m\in\mathcal{M}} (m-1)\cdot\left\{\left[\binom{\frac{\nu n}{2k}}{m-1}+\binom{\frac{(1-\nu)n}{2k}}{m-1}\right](\psi_m-\phi_m)+2\binom{\frac{n}{2k}}{m-1}\phi_m\right\}. \tag{36}$$

After Algorithm 6, if a vertex $w \in W_l$ is mislabelled, one of the following events must happen

- $\widehat{S}_{ll}(w) \le \mu_{\mathrm{M}}$, which implies that u was mislabelled by Algorithm 6.
- $\widehat{S}_{lj}(w) \ge \mu_{\mathrm{M}}$ for some $j \neq l$, which implies that u survived Algorithm 6 without being corrected.

By an argument similar to Lemma 5.14, we can prove that for any $w \in W_l$,

$$\hat{\rho}_l := \mathbb{P}(\widehat{S}_{ll}(w) \le \mu_{\mathrm{M}}) \le \rho\,, \quad \hat{\rho}_j := \mathbb{P}(\widehat{S}_{lj}(w) \ge \mu_{\mathrm{M}}) \le \rho\,,\ (j\neq l),$$

where $\rho := \exp(-C_{\mathcal{M}}\cdot \mathrm{SNR}_{\mathcal{M}})$. The misclassified probability for $w \in W_l$ is upper bounded by $\sum_{j=1}^{k}\hat{\rho}_j \le k\rho$. The number of mislabelled vertices in W_l is at most $R_l = \sum_{t=1}^{|W_l|}\Gamma_t$, where Γ_t are i.i.d indicator random variables with mean $k\rho$ and $\mathbb{E}R_l \le n/(2k)\cdot k\rho = n\rho/2$. Let $t_l := n\rho/2$, by Chernoff bound (Lemma D.1),

$$\mathbb{P}\left(R_l \ge n\rho\right) = \mathbb{P}\left(R_l - n\rho/2 \ge t_l\right) \le \mathbb{P}\left(R_l - \mathbb{E}R_l \ge t_l\right) \le e^{-ct_l} = O(e^{-n\rho})\,.$$

Hence with probability $1 - O(e^{-n\rho})$, the fraction of mislabelled vertices in W_l is smaller than $k\rho$, i.e., the correctness in W_l is at least $\gamma_{\mathrm{M}} := \max\{\nu,\ 1-k\rho\}$.

5.4. Proof of Theorem 1.7

Now we are ready to prove Theorem 1.7. The correctness of Algorithms 5 and 6 are denoted by γ_{C} and γ_{M} respectively, then with probability at least $1 - O(e^{-n\rho})$, the correctness γ of Algorithm 2 is $\gamma := \min\{\gamma_{\mathrm{C}}, \gamma_{\mathrm{M}}\} = \max\{\nu,\ 1-k\rho\}$. We will have $\gamma = 1-k\rho$ if $\nu \le 1-k\rho$, equivalently,

$$\mathrm{SNR}_{\mathcal{M}}(k) \ge \frac{1}{C_{\mathcal{M}}}\log\left(\frac{k}{1-\nu}\right), \tag{37}$$

otherwise $\gamma = \nu$. The inequality (37) holds since

$$\mathrm{SNR}_{\mathcal{M}}(k) = \frac{\left[\sum_{m\in\mathcal{M}}(m-1)\left(\frac{a_m-b_m}{k^{m-1}}\right)\right]^2}{\sum_{m\in\mathcal{M}}(m-1)\left(\frac{a_m-b_m}{k^{m-1}}+b_m\right)}$$

$$\ge \frac{[\sum_{m\in\mathcal{M}}(a_m-b_m)]^2}{k^{2\mathcal{M}_{\max}-2}(\mathcal{M}_{\max}-1)d} \ge \frac{(C_\nu)^2}{\mathcal{M}_{\max}-1}\log\left(\frac{k}{1-\nu}\right) \ge \frac{1}{C_{\mathcal{M}}}\log\left(\frac{k}{1-\nu}\right).$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

where the first two inequalities hold since $d := \sum_{m\in\mathcal{M}} (m-1)a_m$ and Condition (4b), while the last inequality holds by taking $C_\nu \geq \max\{\sqrt{(\mathcal{M}_{\max}-1)/C_{\mathcal{M}}}, C_2\}$ with C_2 defined in (26).

Remark 5.15. The lower bound C in (4a) comes from the requirement in Lemma 5.4 that only a few high-degree vertices be deleted. The constant C_ν in (4b) comes from the requirement in Lemma 5.6 that the subspace angle is small. When C is not so large (or the hypergraph is too sparse), one could still achieve good accuracy γ if C_ν is large enough (the difference between a_m and b_m is large enough).

Remark 5.16. Condition (37) indicates that the improvement of accuracy from local refinement (Algorithms 5 and 6) will be guaranteed when $\mathrm{SNR}_{\mathcal{M}}(k)$ is large enough. If $\mathrm{SNR}_{\mathcal{M}}(k)$ is small, we use correctness of Algorithm 4 instead, i.e., $\gamma = \nu$, to represent the correctness of Algorithm 2.

5.5. *Proof of Corollary 1.9*

For any fixed $\nu \in (1/k, 1)$, $\mathrm{SNR}_{\mathcal{M}}(k) \to \infty$ implies $\rho \to 0$ and

$$d = \sum_{m\in\mathcal{M}} (m-1)a_m \to \infty.$$

Since

$$\frac{\sum_{m\in\mathcal{M}} (m-1)(a_m-b_m)^2}{\sum_{m\in\mathcal{M}} (m-1)a_m} \geq \frac{\sum_{m\in\mathcal{M}} (m-1)(a_m-b_m)^2}{\sum_{m\in\mathcal{M}} (m-1)k^{m-1}(a_m+k^{m-1}b_m)} = \mathrm{SNR}_{\mathcal{M}}(k)\,,$$

Condition (4b) is satisfied. Applying Theorem 1.7, we find $\gamma = 1 - o(1)$, which implies weak consistency. The constraint of $\mathrm{SNR}_{\mathcal{M}}(k) = o(\log n)$ is used in the proof of Lemma 4.2, see Remark 1.10.

Acknowledgements

I.D. and H.W. are partially supported by NSF DMS-2154099. I.D. and Y.Z. acknowledge support from NSF DMS-1928930 during their participation in the programme Universality and Integrability in Random Matrix Theory and Interacting Particle Systems hosted by the Mathematical Sciences Research Institute in Berkeley, California during the Fall semester of 2021. Y.Z. is partially supported by NSF-Simons Research Collaborations on the Mathematical and Scientific Foundations of Deep Learning. Y.Z. thanks Zhixin Zhou for his helpful comments. The authors thank the anonymous reviewers for detailed comments and suggestions which greatly improve the presentation of this work.

References

[1] Abbe, E. (2018) Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18**(177) 1–86.
[2] Abbe, E., Bandeira, A. S. and Hall, G. (2016) Exact recovery in the stochastic block model. *IEEE Trans. Inf. Theory* **62**(1) 471–487.
[3] Abbe, E., Fan, J., Wang, K. and Zhong, Y. (2020) Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Stat.* **48**(3) 1452.
[4] Abbe, E. and Sandon, C. (2015) 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*, IEEE, pp. 670–688.
[5] Abbe, E. and Sandon, C. (2018) Proof of the achievability conjectures for the general stochastic block model. *Commun. Pure Appl. Math.* **71**(7) 1334–1406.
[6] Ahn, Kwangjun, Lee, Kangwook and Suh, Changho (2016) Community recovery in hypergraphs. In IEEE, pp. 657–663. 2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton).

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

[7] Ahn, K., Lee, K. and Suh, C. (2018) Hypergraph spectral clustering in the weighted stochastic block model. *Ieee J Sel. Top. Signal Process.* **12**(5) 959–974.

[8] Alaluusua, K., Avrachenkov, K. , Vinay Kumar, B. R. and Leskelä, L. (2023) Multilayer hypergraph clustering using the aggregate similarity matrix, pp. 83–98.

[9] Angelini, M. C., Caltagirone, F., Krzakala, F. and Zdeborová, L. (2015) Spectral detection on sparse hypergraphs. In IEEE, pp. 66–73, 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton).

[10] Battiston, F., Cencetti, G., Iacopini, I., Latora, V., Lucas, M., Patania, A., Young, J.-G. and Petri, G. (2020) Networks beyond pairwise interactions: structure and dynamics. *Phys. Rep.* **874** 1–92.

[11] Benson, A. R., Gleich, D. F. and Leskovec, J. (2016) Higher-order organization of complex networks. *Science* **353**(6295) 163–166.

[12] Bordenave, C., Lelarge, M. and Massoulié, L. (2018) Nonbacktracking spectrum of random graphs: Community detection and nonregular Ramanujan graphs. *Ann. Probab.* **46**(1) 1–71.

[13] Bu, J., Tan, S., Chen, C., Wang, C., Wu, H., Zhang, L. and He, X. (2010) Music recommendation by unified hypergraph: Combining social media information and music content. In Proceedings of the 18th ACM International Conference on Multimedia, pp. 391–400.

[14] Cáceres, E., Misobuchi, A. and Pimentel, R. (2021) Sparse SYK and traversable wormholes. *J. High Energy Phys.* **2021**(11) 1–32.

[15] Chen, Y., Chi, Y., Fan, J. and Ma, C. (2021) Spectral methods for data science: A statistical perspective. *Found. Trends® Mach. Learn.* **14**(5) 566–806.

[16] Chien, I., Lin, Chung-Yi and Wang, I.-Hsiang (2018) Community detection in hypergraphs: Optimal statistical limit and efficient algorithms. In *International Conference on Artificial Intelligence and Statistics*, pp. 871–879.

[17] Eli Chien, I., Lin, C.-Y. and Wang, I.-H. (2019) On the minimax misclassification ratio of hypergraph community detection. *IEEE Trans. Inf. Theory* **65**(12) 8095–8118.

[18] Chin, B. and Sly, A. (2021) Optimal reconstruction of general sparse stochastic block models. arXiv preprint arXiv:2111.00697.

[19] Chin, P., Rao, A. and Van, V. (2015) Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Conference on Learning Theory*, pp. 391–423.

[20] Coja-Oghlan, A. (2010) Graph partitioning via adaptive spectral techniques. *Comb. Probab. Comput.* **19**(2) 227–284.

[21] Cole, S. and Zhu, Y. (2020) Exact recovery in the hypergraph stochastic block model: A spectral algorithm. *Linear Algebra Appl.* **593** 45–73.

[22] Cook, N., Goldstein, L. and Johnson, T. (2018) Size biased couplings and the spectral gap for random regular graphs. *Ann. Probab.* **46**(1) 72–125.

[23] Cooper, J. (2020) Adjacency spectra of random and complete hypergraphs. *Linear Algebra Appl.* **596** 184–202.

[24] Deng, C., Xu, X.-J. and Ying, S. (2024) Strong consistency of spectral clustering for the sparse degree-corrected hypergraph stochastic block model. *IEEE Trans. Inf. Theory* **70**(3) 1962–1977.

[25] Dumitriu, I. and Wang, H. (2023) Optimal and exact recovery on general non-uniform hypergraph stochastic block model. arXiv preprint arXiv:2304.13139.

[26] Dumitriu, I. and Zhu, Y. (2021) Spectra of random regular hypergraphs. *Electron. J. Comb.* **28**(3) P3.36.

[27] Fei, Y. and Chen, Y. (2020) Achieving the bayes error rate in synchronization and block models by SDP, robustly. *IEEE Trans. Inf. Theory* **66**(6) 3929–3953.

[28] Feige, Uriel and Ofek, Eran (2005) Spectral techniques applied to sparse random graphs. *Random Struct Algor* **27**(2) 251–275.

[29] Friedman, Joel and Wigderson, Avi (1995) On the second eigenvalue of hypergraphs. *Combinatorica* **15**(1) 43–65.

[30] Gaudio, Julia and Joshi, Nirmit (2023) Community detection in the hypergraph sbm: Optimal recovery given the similarity matrix. (Annual Conference Computational Learning Theory

[31] Ghoshdastidar, D. and Dukkipati, A. (2014) Consistency of spectral partitioning of uniform hypergraphs under planted partition model. In *Advances in Neural Information Processing Systems*, pp. 397–405.

[32] Ghoshdastidar, D. and Dukkipati, A. (2017) Consistency of spectral hypergraph partitioning under planted partition model. *Ann. Stat.* **45**(1) 289–315.

[33] Ghoshdastidar, D. and Dukkipati, A. (2017) Uniform hypergraph partitioning: Provable tensor methods and sampling techniques. *J. Mach. Learn. Res.* **18**(1) 1638–1678.

[34] Govindu, V. M. (2005) A tensor decomposition for geometric grouping and segmentation. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), IEEE, pp. 1150–1157, vol. **1**.

[35] Gu, Y. and Pandey, A. (2024) Community detection in the hypergraph stochastic block model and reconstruction on hypertrees. arXiv preprint arXiv:2402.06856.

[36] Gu, Y. and Polyanskiy, Y. (2023) Weak recovery threshold for the hypergraph stochastic block model. In The Thirty Sixth Annual Conference on Learning Theory, PMLR, pp. 885–920.

[37] Guédon, O. and Vershynin, R. (2016) Community detection in sparse networks via Grothendieck's inequality. *Probab. Theory Relat.* **165**(3-4) 1025–1049.

[38] Harris, K. D. and Zhu, Y. (2021) Deterministic tensor completion with hypergraph expanders. *SIAM J. Math. Data Sci.* **3**(4) 1117–1140.

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

[39] Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983) Stochastic blockmodels: First steps. *Soc. Networks* **5**(2) 109–137.
[40] Jain, P. and Oh, S. (2014) Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, pp. 1431–1439.
[41] Jin, J., Ke, T. and Liang, J. (2021) Sharp impossibility results for hyper-graph testing. In *Advances in Neural Information Processing Systems, 34*.
[42] Ke, Z. T., Shi, F. and Xia, D. (2019) Community detection for hypergraph networks via regularized tensor power iteration. arXiv preprint arXiv:1909.06503.
[43] Kim, C., Bandeira, A. S. and Goemans, M. X. (2018) Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach. arXiv preprint arXiv:1807.02884.
[44] Le, C. M. and Levina, E. (2022) Estimating the number of communities by spectral methods. *Electron. J. Stat.* **16**(1) 3315–3342.
[45] Le, C. M., Levina, E. and Vershynin, R. (2017) Concentration and regularization of random graphs. *Random Struct. Algorithms* **51**(3) 538–561.
[46] Lee, J., Kim, D. and Chung, H. W. (2020) Robust hypergraph clustering via convex relaxation of truncated mle. *IEEE J. Selected Areas Inf. Theory* **1**(3) 613–631.
[47] Lei, J., Chen, K. and Lynch, B. (2020) Consistent community detection in multi-layer network data. *Biometrika* **107**(1) 61–73.
[48] Lei, J. and Rinaldo, A. (2015) Consistency of spectral clustering in stochastic block models. *Ann. Stat.* **43**(1) 215–237.
[49] Li, L. and Li, T. (2013) News recommendation via hypergraph learning: Encapsulation of user behavior and news content. In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 305–314.
[50] Lin, C.-Y. , Eli Chien, I. and Wang, I.-H. (2017) On the fundamental statistical limit of community detection in random hypergraphs. In *2017 IEEE International Symposium on Information Theory (ISIT)*, IEEE, pp. 2178–2182.
[51] Lung, R. I., Gaskó, N. and Suciu, M. A. (2018) A hypergraph model for representing scientific output. *Scientometrics* **117**(3) 1361–1379.
[52] Massoulié, L. (2014) Community detection thresholds and the weak Ramanujan property. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, ACM, pp. 694–703.
[53] Michoel, T. and Nachtergaele, B. (2012) Alignment and integration of complex networks by hypergraph-based spectral clustering. *Phys. Rev. E* **86**(5) 056111.
[54] Montanari, A. and Sen, S. (2016) Semidefinite programs on sparse random graphs and their application to community detection. In Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing, pp. 814–827.
[55] Mossel, E., Neeman, J. and Sly, A. (2015) Reconstruction and estimation in the planted partition model. *Probab. Theory Relat.* **162**(3-4) 431–461.
[56] Mossel, E., Neeman, J. and Sly, A. (2016) Belief propagation, robust reconstruction and optimal recovery of block models. *Ann. Appl. Probab.* **26**(4) 2211–2256.
[57] Mossel, E., Neeman, J. and Sly, A. (2016) Consistency thresholds for the planted bisection model. *Electron. J. Probab.* **21**(21) 1–24.
[58] Mossel, E., Neeman, J. and Sly, A. (2018) A proof of the block model threshold conjecture. *Combinatorica* **38**(3) 665–708.
[59] Newman, M. E. J., Watts, D. J. and Strogatz, S. H. (2002) Random graph models of social networks. *Proc. Natl. Acad. Sci.* **99**(suppl 1) 2566–2572.
[60] Ng, A. Y., Jordan, M. I. and Weiss, Y. (2002) On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pp. 849–856.
[61] Nguyen, N. H., Drineas, P. and Tran, T. D. (2015) Tensor sparsification via a bound on the spectral norm of random tensors. *Inf. Inference J. IMA* **4**(3) 195–229.
[62] Pal, S. and Zhu, Y. (2021) Community detection in the sparse hypergraph stochastic block model. *Random Struct. Algorithms* **59**(3) 407–463.
[63] Saade, A., Krzakala, F. and Zdeborová, L. (2014) Spectral clustering of graphs with the bethe hessian. In *Advances in Neural Information Processing Systems, 27*.
[64] Sen, S. (2018) Optimization on sparse random hypergraphs and spin glasses. *Random Struct. Algorithms* **53**(3) 504–536.
[65] Shi, J. and Malik, J. (2000) Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(8) 888–905.
[66] Soma, T. and Yoshida, Y. (2019) Spectral sparsification of hypergraphs. In Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, pp. 2570–2581.
[67] Stephan, L. and Zhu, Y. (2024) Sparse random hypergraphs: Non-backtracking spectra and community detection. *Inf. Inference J. IMA* **13**(1) iaae004.
[68] Tian, Z., Hwang, T. and Kuang, R. (2009) A hypergraph-based learning algorithm for classifying gene expression and arrayCGH data with prior knowledge. *Bioinformatics* **25**(21) 2831–2838.
[69] Vershynin, R. (2018) *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge University Press, Cambridge Series in Statistical and Probabilistic Mathematics.
[70] Van, V. (2018) A simple svd algorithm for finding hidden partitions. *Comb. Probab. Comput.* **27**(1) 124–140.
[71] Wang, H. (2023) Fundamental limits and strong consistency of binary non-uniform hypergraph stochastic block models.

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

[72] Wang, J., Pun, Y.-M., Wang, X., Wang, P. and So, A. M.-C. (2023) Projected tensor power method for hypergraph community recovery. In International Conference on Machine Learning, PMLR, pp. 36285–36307.
[73] Wen, L., Du, D., Li, S., Bian, X. and Lyu, S. (2019) Learning non-uniform hypergraph for multi-object tracking. In Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8981–8988, vol. **33**.
[74] Yuan, M., Liu, R., Feng, Y. and Shang, Z. (2022) Testing community structure for hypergraphs. *Ann. Stat.* **50**(1) 147–169.
[75] Yuan, M., Zhao, B. and Zhao, X. (2021) Community detection in censored hypergraph. *Statistica Sinica*. arXiv preprint arXiv:2111.03179.
[76] Zhang, A. Y. and Zhou, H. H. (2016) Minimax rates of community detection in stochastic block models. *Ann. Stat.* **44**(5) 2252–2280.
[77] Zhang, Q. and Tan, V. Y. F. (2023) Exact recovery in the general hypergraph stochastic block model. *IEEE Trans. Inf. Theory* **69**(1) 453–471.
[78] Zhen, Y. and Wang, J. (2022) Community detection in general hypergraph via graph embedding. *J. Am. Stat. Assoc.*, 1–10.
[79] Zhou, D., Huang, J. and Schölkopf, B. (2007) Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems*, pp. 1601–1608.
[80] Zhou, Z. and Zhu, Y. (2021) Sparse random tensors: Concentration, regularization and applications. *Electron. J. Stat.* **15**(1) 2483–2516.

Appendix A. Proof of Theorems 3.1 and 3.3

A.1. Discretization

To prove Theorem 3.1, we start with a standard ε-net argument.

Lemma A.1 (Lemma 4.4.1 in [69]). *Let* $\mathbf{W}$ *be any Hermitian* $n \times n$ *matrix and let* $\mathcal{N}_\varepsilon$ *be an* ε*-net on the unit sphere* $\mathbb{S}^{n-1}$ *with* $\varepsilon \in (0,1)$*, then* $\|\mathbf{W}\| \le \frac{1}{1-\varepsilon} \sup_{\boldsymbol{x}\in\mathcal{N}_\varepsilon} |\langle \mathbf{W}\boldsymbol{x}, \boldsymbol{x}\rangle|$.

By [69, Corollary 4.2.13], the size of $\mathcal{N}_\varepsilon$ is bounded by $|\mathcal{N}_\varepsilon| \le (1+2/\varepsilon)^n$. We would have $\log|\mathcal{N}| \le n\log(5)$ when $\mathcal{N}$ is taken as an $(1/2)$-net of $\mathbb{S}^n$. Define $\mathbf{W} := \mathbf{A} - \mathbb{E}\mathbf{A}$, then $\mathbf{W}_{ii} = 0$ for each $i \in [n]$ by the definition of adjacency matrix in equation (7), and we obtain

$$\|\mathbf{A} - \mathbb{E}\mathbf{A}\| = \|\mathbf{W}\| \le 2 \sup_{\boldsymbol{x}\in\mathcal{N}} |\langle \mathbf{W}\boldsymbol{x}, \boldsymbol{x}\rangle| . \tag{A.1}$$

For any fixed $\boldsymbol{x} \in \mathbb{S}^{n-1}$, consider the *light* and *heavy* pairs as follows.

$$\mathcal{L}(\boldsymbol{x}) = \left\{(i,j) : |\boldsymbol{x}_i\boldsymbol{x}_j| \le \frac{\sqrt{d}}{n}\right\}, \quad \mathcal{H}(\boldsymbol{x}) = \left\{(i,j) : |\boldsymbol{x}_i\boldsymbol{x}_j| > \frac{\sqrt{d}}{n}\right\}, \tag{A.2}$$

where $d = \sum_{m=2}^{M}(m-1)d_m$. Thus by the triangle inequality,

$$|\langle \boldsymbol{x}, \mathbf{W}\boldsymbol{x}\rangle| \le \left|\sum_{(i,j)\in\mathcal{L}(\boldsymbol{x})} \mathbf{W}_{ij}\boldsymbol{x}_i\boldsymbol{x}_j\right| + \left|\sum_{(i,j)\in\mathcal{H}(\boldsymbol{x})} \mathbf{W}_{ij}\boldsymbol{x}_i\boldsymbol{x}_j\right|,$$

and by equation (A.1),

$$\|\mathbf{A} - \mathbb{E}\mathbf{A}\| \le 2 \sup_{\boldsymbol{x}\in\mathcal{N}} \left|\sum_{(i,j)\in\mathcal{L}(\boldsymbol{x})} \mathbf{W}_{ij}\boldsymbol{x}_i\boldsymbol{x}_j\right| + 2 \sup_{\boldsymbol{x}\in\mathcal{N}} \left|\sum_{(i,j)\in\mathcal{H}(\boldsymbol{x})} \mathbf{W}_{ij}\boldsymbol{x}_i\boldsymbol{x}_j\right| . \tag{A.3}$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

A.2. Contribution from light pairs

For each m-hyperedge $e \in E_m$, we define $\mathcal{W}_e^{(m)} := \mathcal{A}_e^{(m)} - \mathbb{E}\mathcal{A}_e^{(m)}$. Then for any fixed $\boldsymbol{x} \in \mathbb{S}^{n-1}$, the contribution from light couples can be written as

$$\sum_{(i,j)\in\mathcal{L}(\boldsymbol{x})} \mathbf{W}_{ij}\boldsymbol{x}_i\boldsymbol{x}_j = \sum_{(i,j)\in\mathcal{L}(\boldsymbol{x})} \left(\sum_{m=2}^{M} \sum_{\substack{e\in E_m \\ \{i,j\}\subset e}} \mathcal{W}_e^{(m)} \right) \boldsymbol{x}_i\boldsymbol{x}_j$$

$$= \sum_{m=2}^{M} \sum_{e\in E_m} \mathcal{W}_e^{(m)} \left(\sum_{\substack{(i,j)\in\mathcal{L}(\boldsymbol{x}) \\ i\neq j,\ \{i,j\}\subset e}} \boldsymbol{x}_i\boldsymbol{x}_j \right) = \sum_{m=2}^{M} \sum_{e\in E_m} \mathcal{Y}_e^{(m)}, \tag{A.4}$$

where the constraint $i \neq j$ comes from the fact $\mathbf{W}_{ii} = 0$ and we denote

$$\mathcal{Y}_e^{(m)} := \mathcal{W}_e^{(m)} \left(\sum_{\substack{(i,j)\in\mathcal{L}(\boldsymbol{x}) \\ i\neq j,\ \{i,j\}\subset e}} \boldsymbol{x}_i\boldsymbol{x}_j \right).$$

Note that $\mathbb{E}\mathcal{Y}_e^{(m)} = 0$, and by the definition of light pair equation (A.2),

$$|\mathcal{Y}_e^{(m)}| \le m(m-1)\sqrt{d}/n \le M(M-1)\sqrt{d}/n, \quad \forall m \in \{2, \cdots, M\}.$$

Moreover, equation (A.4) is a sum of independent, mean-zero random variables, and

$$\sum_{m=2}^{M} \sum_{e\in E_m} \mathbb{E}[(\mathcal{Y}_e^{(m)})^2] := \sum_{m=2}^{M} \sum_{e\in E_m} \left[\mathbb{E}[(\mathcal{W}_e^{(m)})^2] \left(\sum_{\substack{(i,j)\in\mathcal{L}(\boldsymbol{x}) \\ i\neq j,\{i,j\}\subset e}} \boldsymbol{x}_i\boldsymbol{x}_j \right)^2 \right]$$

$$\le \sum_{m=2}^{M} \sum_{e\in E_m} \left[\mathbb{E}[\mathcal{A}_e^{(m)}] \cdot m(m-1) \left(\sum_{\substack{(i,j)\in\mathcal{L}(x) \\ i\neq j,\{i,j\}\subset e}} \boldsymbol{x}_i^2\boldsymbol{x}_j^2 \right) \right]$$

$$\le \sum_{m=2}^{M} \frac{d_m \cdot m(m-1)}{\binom{n}{m-1}} \binom{n}{m-2} \sum_{(i,j)\in[n]^2} \boldsymbol{x}_i^2\boldsymbol{x}_j^2$$

$$\le \sum_{m=2}^{M} \frac{d_m m(m-1)^2}{n-m+2} \le \frac{2}{n} \sum_{m=2}^{M} d_m (m-1)^3 \le \frac{2d(M-1)^2}{n},$$

when $n \ge 2m-2$, where $d_m = \max d_{[i_1,\ldots,i_m]}$ and $d = \sum_{m=2}^{M} (m-1)d_m$. Then Bernstein's inequality (Lemma D.3) implies that for any $\alpha > 0$,

$$\mathbb{P}\left(\left| \sum_{(i,j)\in\mathcal{L}(\boldsymbol{x})} \mathbf{W}_{ij}\boldsymbol{x}_i\boldsymbol{x}_j \right| \ge \alpha\sqrt{d} \right) = \mathbb{P}\left(\left| \sum_{m=2}^{M} \sum_{e\in E_m} \mathcal{Y}_e^{(m)} \right| \ge \alpha\sqrt{d} \right)$$

$$\le 2\exp\left(-\frac{\frac{1}{2}\alpha^2 d}{\frac{2d}{n}(M-1)^2 + \frac{1}{3}(M-1)M\frac{\sqrt{d}}{n}\alpha\sqrt{d}} \right) \le 2\exp\left(-\frac{\alpha^2 n}{4(M-1)^2 + \frac{2\alpha(M-1)M}{3}} \right).$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

Therefore by taking a union bound,

$$\mathbb{P}\left(\sup_{\boldsymbol{x}\in\mathcal{N}}\left|\sum_{(i,j)\in\mathcal{L}(\boldsymbol{x})}\mathbf{W}_{ij}\boldsymbol{x}_i\boldsymbol{x}_j\right|\geq\alpha\sqrt{d}\right)\leq|\mathcal{N}|\cdot\mathbb{P}\left(\left|\sum_{(i,j)\in\mathcal{L}(\boldsymbol{x})}\mathbf{W}_{ij}\boldsymbol{x}_i\boldsymbol{x}_j\right|\geq\alpha\sqrt{d}\right)$$
$$\leq 2\exp\left(\log(5)\cdot n-\frac{\alpha^2 n}{4(M-1)^2+\frac{2\alpha(M-1)M}{3}}\right)\leq 2e^{-n}, \tag{A.5}$$

where we choose $\alpha=5M(M-1)$ in the last line.

A.3. Contribution from heavy pairs

Note that for any $i\neq j$,

$$\mathbb{E}\mathbf{A}_{ij}\leq\sum_{m=2}^{M}\binom{n-2}{m-2}\frac{d_m}{\binom{n}{m-1}}\leq\sum_{m=2}^{M}\frac{(m-1)d_m}{n}=\frac{d}{n}. \tag{A.6}$$

and

$$\left|\sum_{(i,j)\in\mathcal{H}(\boldsymbol{x})}\mathbb{E}\mathbf{A}_{ij}\boldsymbol{x}_i\boldsymbol{x}_j\right|=\left|\sum_{(i,j)\in\mathcal{H}(\boldsymbol{x})}\mathbb{E}\mathbf{A}_{ij}\frac{\boldsymbol{x}_i^2\boldsymbol{x}_j^2}{\boldsymbol{x}_i\boldsymbol{x}_j}\right| \tag{A.7}$$
$$\leq\sum_{(i,j)\in\mathcal{H}(\boldsymbol{x})}\frac{d}{n}\frac{\boldsymbol{x}_i^2\boldsymbol{x}_j^2}{|\boldsymbol{x}_i\boldsymbol{x}_j|}\leq\sqrt{d}\sum_{(i,j)\in\mathcal{H}(\boldsymbol{x})}\boldsymbol{x}_i^2\boldsymbol{x}_j^2\leq\sqrt{d}.$$

Therefore it suffices to show that, with high probability,

$$\sum_{(i,j)\in\mathcal{H}(\boldsymbol{x})}\mathbf{A}_{ij}\boldsymbol{x}_i\boldsymbol{x}_j=O\left(\sqrt{d}\right). \tag{A.8}$$

Here we use the discrepancy analysis from [22, 28]. We consider the weighted graph associated with the adjacency matrix $\mathbf{A}$.

Definition A.2 (Uniform upper tail property, **UUTP**). *Let* $\mathbf{M}$ *be an* $n\times n$ *random symmetric matrix with non-negative entries and* $\mathbf{Q}$ *be an* $n\times n$ *symmetric matrix with entries* $\mathbf{Q}_{ij}\in[0,a]$ *for all* $i,j\in[n]$. *Define*

$$\mu:=\sum_{i,j=1}^{n}\mathbf{Q}_{ij}\mathbb{E}\mathbf{M}_{ij},\quad\tilde{\sigma}^2:=\sum_{i,j=1}^{n}\mathbf{Q}_{ij}^2\mathbb{E}\mathbf{M}_{ij}.$$

We say that $\mathbf{M}$ *satisfies the uniform upper tail property* $\mathbf{UUTP}(c_0,\gamma_0)$ *with* $c_0>0$, $\gamma_0\geq 0$, *if for any* $a,t>0$,

$$\mathbb{P}\left(f_{\mathbf{Q}}(\mathbf{M})\geq(1+\gamma_0)\mu+t\right)\leq\exp\left(-c_0\frac{\tilde{\sigma}^2}{a^2}h\left(\frac{at}{\tilde{\sigma}^2}\right)\right).$$

where function $f_{\mathbf{Q}}(\mathbf{M}):\mathbb{R}^{n\times n}\mapsto\mathbb{R}$ *is defined by* $f_{\mathbf{Q}}(\mathbf{M}):=\sum_{i,j=1}^{n}\mathbf{Q}_{ij}\mathbf{M}_{ij}$ *for* $\mathbf{M}\in\mathbb{R}^{n\times n}$, *and function* $h(x):=(1+x)\log(1+x)-x$ *for all* $x>-1$.

Lemma A.3. *Let* $\mathbf{A}$ *be the adjacency matrix of non-uniform hypergraph* $H=\bigcup_{m=2}^{M}H_m$, *then* $\mathbf{A}$ *satisfies* $\mathbf{UUTP}(c_0,\gamma_0)$ *with* $c_0=[M(M-1)]^{-1}$, $\gamma_0=0$.

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

Proof of Lemma A.3. Note that

$$f_{\mathbf{Q}}(\mathbf{A}) - \mu = \sum_{i,j=1}^{n} \mathbf{Q}_{ij}(\mathbf{A}_{ij} - \mathbb{E}\mathbf{A}_{ij}) = \sum_{i,j=1}^{n} \mathbf{Q}_{ij}\mathbf{W}_{ij}$$
$$= \sum_{i,j=1}^{n} \mathbf{Q}_{ij}\left(\sum_{m=2}^{M} \sum_{\substack{e\in E_m \\ i\neq j,\, \{i,j\}\subset e}} \mathcal{W}_e^{(m)}\right) = \sum_{m=2}^{M} \sum_{e\in E_m} \mathcal{W}_e^{(m)}\left(\sum_{\{i,j\}\subset e,\, i\neq j} \mathbf{Q}_{ij}\right) = \sum_{m=2}^{M} \sum_{e\in E_m} \mathcal{Z}_e^{(m)},$$

where $\mathcal{Z}_e^{(m)} = \mathcal{W}_e^{(m)}\left(\sum_{\{i,j\}\subset e, i\neq j} \mathbf{Q}_{ij}\right)$ are independent centred random variables upper bounded by $|\mathcal{Z}_e^{(m)}| \leq \sum_{\{i,j\}\subset e,\, i\neq j} \mathbf{Q}_{ij} \leq M(M-1)a$ for each $m \in \{2, \ldots, M\}$ since $\mathbf{Q}_{ij} \in [0, a]$. Moreover, the variance of the sum can be written as

$$\sum_{m=2}^{M} \sum_{e\in E_m} \mathbb{E}(\mathcal{Z}_e^{(m)})^2 = \sum_{m=2}^{M} \sum_{e\in E_m} \mathbb{E}(\mathcal{W}_e^{(m)})^2 \left(\sum_{\{i,j\}\subset e,\, i\neq j} \mathbf{Q}_{ij}\right)^2$$
$$\leq \sum_{m=2}^{M} \sum_{e\in E_m} \mathbb{E}[\mathcal{A}_e^{(m)}] \cdot m(m-1) \sum_{\{i,j\}\subset e, i\neq j} \mathbf{Q}_{ij}^2 \leq M(M-1) \sum_{i,j=1}^{n} \mathbf{Q}_{ij}^2 \mathbb{E}\mathbf{A}_{ij} = M(M-1)\tilde{\sigma}^2.$$

where the last inequality holds since by definition $\mathbb{E}\mathbf{A}_{ij} = \sum_{m=2}^{M} \sum_{\substack{e\in E_m \\ \{i,j\}\subset e}} \mathbb{E}[\mathcal{A}_e^{(m)}]$. Then by Bennett's inequality Lemma D.4, we obtain

$$\mathbb{P}(f_{\mathbf{Q}}(\mathbf{A}) - \mu \geq t) \leq \exp\left(-\frac{\tilde{\sigma}^2}{M(M-1)a^2} h\left(\frac{at}{\tilde{\sigma}^2}\right)\right)$$

where the inequality holds since the function $x \cdot h(1/x) = (1+x)\log(1+1/x) - 1$ is decreasing with respect to x. □

Definition A.4 (Discrepancy property, **DP**). *Let* $\mathbf{M}$ *be an* $n \times n$ *matrix with non-negative entries. For* $S, T \subset [n]$, *define* $e_{\mathbf{M}}(S, T) = \sum_{i\in S, j\in T} \mathbf{M}_{ij}$. *We say* $\mathbf{M}$ *has the discrepancy property with parameter* $\delta > 0$, $\kappa_1 > 1, \kappa_2 \geq 0$, *denoted by* $\mathbf{DP}(\delta, \kappa_1, \kappa_2)$, *if for all non-empty* $S, T \subset [n]$, *at least one of the following hold:*

(1) $e_{\mathbf{M}}(S, T) \leq \kappa_1 \delta |S||T|$;
(2) $e_{\mathbf{M}}(S, T) \cdot \log\left(\frac{e_{\mathbf{M}}(S,T)}{\delta|S|\cdot|T|}\right) \leq \kappa_2(|S| \vee |T|) \cdot \log\left(\frac{en}{|S|\vee|T|}\right)$.

Lemma A.5 shows that if a symmetric random matrix $\mathbf{A}$ satisfies the upper tail property $\mathbf{UUTP}(c_0, \gamma_0)$ with parameter $c_0 > 0, \gamma_0 \geq 0$, then the discrepancy property holds with high probability.

Lemma A.5 (Lemma 6.4 in [22]). *Let* $\mathbf{M}$ *be an* $n \times n$ *symmetric random matrix with non-negative entries. Assume that for some* $\delta > 0$, $\delta > 0$, $\mathbb{E}\mathbf{M}_{ij} \leq \delta$ *for all* $i, j \in [n]$ *and* $\mathbf{M}$ *has* $\mathbf{UUTP}(c_0, \gamma_0)$ *with parameter* $c_0, \gamma_0 > 0$. *Then for any* $K > 0$, *the discrepancy property* $\mathbf{DP}(\delta, \kappa_1, \kappa_2)$ *holds for* $\mathbf{M}$ *with probability at least* $1 - n^{-K}$ $\kappa_1 = e^2(1+\gamma_0)^2, \kappa_2 = \frac{2}{c_0}(1+\gamma_0)(K+4)$.

When the discrepancy property holds, then deterministically the contribution from heavy pairs is $O(\sqrt{d})$, as shown in the following lemma.

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

Lemma A.6 (Lemma 6.6 in [22]). *Let* $\mathbf{M}$ *be a non-negative symmetric* $n \times n$ *matrix with all row sums bounded by* d. *Suppose* $\mathbf{M}$ *has* $\mathbf{M}$ *has* $\mathbf{DP}(\delta, \kappa_1, \kappa_2)$ *with* $\delta = Cd/n$ *for some* $C > 0, \kappa_1 > 1$, $\kappa_2 \geq 0$. *Then for any* $x \in \mathbb{S}^{n-1}$,

$$\left| \sum_{(i,j)\in\mathcal{H}(x)} \mathbf{M}_{ij} x_i x_j \right| \leq \alpha_0 \sqrt{d},$$

where $\alpha_0 = 16 + 32C(1+\kappa_1) + 64\kappa_2(1 + \frac{2}{\kappa_1 \log \kappa_1})$.

Lemma A.7 proves that $\mathbf{A}$ has bounded row and column sums with high probability.

Lemma A.7. *For any* $K > 0$, *there is a constant* $\alpha_1 > 0$ *such that with probability at least* $1 - n^{-K}$,

$$\max_{1\leq i\leq n} \sum_{j=1}^{n} \mathbf{A}_{ij} \leq \alpha_1 d \tag{A.9}$$

with $\alpha_1 = 4 + \frac{2(M-1)(1+K)}{3c}$ *and* $d \geq c \log n$.

Proof. For a fixed $i \in [n]$,

$$\sum_{j=1}^{n} \mathbf{A}_{ij} = \sum_{m=2}^{M} \sum_{e\in E_m : i\in e} (m-1)\mathcal{A}_e^{(m)}, \quad \sum_{j=1}^{n} (\mathbf{A}_{ij} - \mathbb{E}\mathbf{A}_{ij}) = \sum_{m=2}^{M} \sum_{e\in E_m : i\in e} (m-1)\mathcal{W}_e^{(m)},$$

$$\sum_{j=1}^{n} \mathbb{E}\mathbf{A}_{ij} \leq \sum_{m=2}^{M} \binom{n}{m-1} \frac{(m-1)d_m}{\binom{n}{m-1}} = d,$$

$$\sum_{m=2}^{M} (m-1)^2 \sum_{e\in E_m : i\in e} \mathbb{E}[(\mathcal{W}_e^{(m)})^2] \leq \sum_{m=2}^{M} (m-1)^2 \sum_{e\in E_m : i\in e} \mathbb{E}[\mathcal{A}_e^{(m)}] \leq (M-1)d.$$

Then for $\alpha_1 = 4 + \frac{2(M-1)(1+K)}{3c}$, by Bernstein's inequality, with the assumption that $d \geq c \log n$,

$$\begin{aligned} \mathbb{P}\left(\sum_{j=1}^{n} \mathbf{A}_{ij} \geq \alpha_1 d \right) &\leq \mathbb{P}\left(\sum_{j=1}^{n} \mathbf{A}_{ij} - \mathbb{E}\mathbf{A}_{ij} \geq (\alpha_1 - 1)d \right) \\ &\leq \exp\left(- \frac{\frac{1}{2}(\alpha_1 - 1)^2 d^2}{(M-1)d + \frac{1}{3}(M-1)(\alpha_1 - 1)d} \right) \leq n^{-\frac{3c(\alpha_1-1)^2}{(M-1)(2\alpha_1+4)}} \leq n^{-1-K}. \end{aligned} \tag{A.10}$$

Taking a union bound over $i \in [n]$, then equation (A.9) holds with probability $1 - n^{-K}$. □

Now we are ready to obtain equation (A.8).

Lemma A.8. *For any* $K > 0$, *there is a constant* β *depending on* K, c, M *such that with probability at least* $1 - 2n^{-K}$,

$$\left| \sum_{(i,j)\in\mathcal{H}(x)} \mathbf{A}_{ij} x_i x_j \right| \leq \beta \sqrt{d}. \tag{A.11}$$

Proof. By Lemma A.3, $\mathbf{A}$ satisfies $\mathbf{UUTP}(\frac{1}{M(M-1)}, 0)$. From equation (A.6) and Lemma A.5, the property $\mathbf{DP}(\delta, \kappa_1, \kappa_2)$ holds for $\mathbf{A}$ with probability at least $1 - n^{-K}$ with

$$\delta = \frac{d}{n}, \quad \kappa_1 = e^2, \quad \kappa_2 = 2M(M-1)(K+4).$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

Let $\mathcal{E}_1$ be the event that $\mathbf{DP}(\delta, \kappa_1, \kappa_2)$ holds for $\mathbf{A}$. Let $\mathcal{E}_2$ be the event that all row sums of $\mathbf{A}$ are bounded by $\alpha_1 d$. Then $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - 2n^{-K}$. On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, by Lemma A.6, equation (A.11) holds with $\beta = \alpha_0 \alpha_1$, where

$$\alpha_0 = 16 + 32(1 + e^2) + 128M(M-1)(K+4)(1+e^{-2}), \quad \alpha_1 = 4 + \frac{2(M-1)(1+K)}{3c}.$$

□

A.4. Proof of Theorem 3.1

Proof. From equation (A.5), with probability at least $1 - 2e^{-n}$, the contribution from light pairs in equation (A.3) is bounded by $2\alpha\sqrt{d}$ with $\alpha = 5M(M-1)$. From equations (A.7) and (A.11), with probability at least $1 - 2n^{-K}$, the contribution from heavy pairs in equation (A.3) is bounded by $2\sqrt{d} + 2\beta\sqrt{d}$. Therefore with probability at least $1 - 2e^{-n} - 2n^{-K}$,

$$\|\mathbf{A} - \mathbb{E}\mathbf{A}\| \leq C_M \sqrt{d},$$

where C_M is a constant depending only on c, K, M such that $C_M = 2(\alpha + 1 + \beta)$. In particular, we can take $\alpha = 5M(M-1)$, $\beta = 512M(M-1)(K+5)\left(2 + \frac{(M-1)(1+K)}{c}\right)$, and $C_M = 512M(M-1)(K+6)\left(2 + \frac{(M-1)(1+K)}{c}\right)$. This finishes the proof of Theorem 3.1. □

A.5. Proof of Theorem 3.3

Let $\mathcal{S} \subset [n]$ be any given subset. From equation (A.5), with probability at least $1 - 2e^{-n}$,

$$\sup_{\boldsymbol{x} \in \mathcal{N}} \left| \sum_{(i,j) \in \mathcal{L}(\boldsymbol{x})} \left(\mathbf{A}_{\mathcal{S}} - \mathbb{E}\mathbf{A}_{\mathcal{S}} \right)_{ij} \boldsymbol{x}_i \boldsymbol{x}_j \right| \leq 5M(M-1)\sqrt{d}. \tag{A.12}$$

Since there are at most 2^n many choices for $\mathcal{S}$, by taking a union bound, with probability at least $1 - 2(e/2)^{-n}$, we have for all $\mathcal{S} \subset [n]$, equation (A.12) holds. In particular, by taking $\mathcal{S} = \mathcal{I} = \{i \in [n] : \mathrm{row}(i) \leq \tau d\}$, with probability at least $1 - 2(e/2)^{-n}$, we have

$$\sup_{\boldsymbol{x} \in \mathcal{N}} \left| \sum_{(i,j) \in \mathcal{L}(\boldsymbol{x})} [(\mathbf{A} - \mathbb{E}\mathbf{A})_{\mathcal{I}}]_{ij} \boldsymbol{x}_i \boldsymbol{x}_j \right| \leq 5M(M-1)\sqrt{d}. \tag{A.13}$$

Similar to equation (A.7), deterministically,

$$\left| \sum_{(i,j) \in \mathcal{H}(\boldsymbol{x})} [(\mathbb{E}\mathbf{A})_{\mathcal{I}}]_{ij} \boldsymbol{x}_i \boldsymbol{x}_j \right| \leq (M-1)\sqrt{d}. \tag{A.14}$$

Next we show the contribution from heavy pairs for $\mathbf{A}_{\mathcal{I}}$ is bounded.

Lemma A.9. *For any $K > 0$, there is a constant β_τ depending on K, c, M, τ such that with probability at least $1 - n^{-K}$,*

$$\left| \sum_{(i,j) \in \mathcal{H}(\boldsymbol{x})} [(\mathbf{A})_{\mathcal{I}}]_{ij} \boldsymbol{x}_i \boldsymbol{x}_j \right| \leq \beta_\tau \sqrt{d}. \tag{A.15}$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

Proof. Note that $\mathbf{A}$ satisfies $\mathbf{UUTP}\left(\frac{1}{M(M-1)}, 0\right)$ from Lemma A.3. According to Lemma A.5, with probability at least $1 - n^{-K}$, $\mathbf{DP}(\delta, \kappa_1, \kappa_2)$ holds for $\mathbf{A}$ with

$$\delta = \frac{d}{n}, \quad \kappa_1 = e^2, \quad \kappa_2 = 2M(M-1)(K+4).$$

The $\mathbf{DP}(\delta, \kappa_1, \kappa_2)$ property holds for $\mathbf{A}_{\mathcal{I}}$ as well, since $\mathbf{A}_{\mathcal{I}}$ is obtained from $\mathbf{A}$ by restricting to $\mathcal{I}$. Note that all row sums in $\mathbf{A}_{\mathcal{I}}$ are bounded by τd. By Lemma A.6,

$$\left| \sum_{(i,j)\in\mathcal{H}(\boldsymbol{x})} [\mathbf{A}_{\mathcal{I}}]_{ij} \boldsymbol{x}_i \boldsymbol{x}_j \right| \le \alpha_0 \sqrt{\tau d}, \tag{A.16}$$

where we can take $\alpha_0 = 16 + \frac{32}{\tau}(1+e^2) + 128M(M-1)(K+4)\left(1+\frac{1}{e^2}\right)$. □

We can then take $\beta_\tau = \alpha_0\sqrt{\tau}$ in equation (A.15). Therefore, combining equations (A.13), (A.14), (A.16), with probability at least $1 - 2(e/2)^{-n} - n^{-K}$, there exists a constant C_τ depending only on τ, M, K such that $\|(\mathbf{A} - \mathbb{E}\mathbf{A})_{\mathcal{I}}\| \le C_\tau\sqrt{d}$, where $C_\tau = 2((5M+1)(M-1) + \alpha_0\sqrt{\tau})$. This finishes the proof of Theorem 3.3.

Appendix B. Technical Lemmas

B.1. Proof of Lemma 2.4

Proof. By Weyl's inequality (Lemma D.5), the difference between eigenvalues of $\widetilde{\mathbb{E}\mathbf{A}}$ and $\mathbb{E}\mathbf{A}$ can be upper bounded by

$$\begin{aligned} |\lambda_i(\widetilde{\mathbb{E}\mathbf{A}}) - \lambda_i(\mathbb{E}\mathbf{A})| &\le \|\widetilde{\mathbb{E}\mathbf{A}} - \mathbb{E}\mathbf{A}\|_2 \le \|\widetilde{\mathbb{E}\mathbf{A}} - \mathbb{E}\mathbf{A}\|_{\mathrm{F}} \\ &\le \left[2k \cdot \frac{n}{k} \cdot \sqrt{n}\log(n) \cdot (\alpha-\beta)^2\right]^{1/2} = O\Big(n^{3/4}\log^{1/2}(n)(\alpha-\beta)\Big). \end{aligned}$$

The lemma follows, as $\lambda_i(\mathbb{E}\mathbf{A}) = \Omega(n(\alpha-\beta))$ for all $1 \le i \le k$. □

B.2. Proof of Lemma 5.3

Proof. We first compute the singular values of $\overline{\mathbf{B}}_1$. From equation (16), the rank of matrix $\overline{\mathbf{B}}_1$ is k, and the least non-trivial singular value of $\overline{\mathbf{B}}_1$ is

$$\sigma_k(\overline{\mathbf{B}}_1) = \frac{n}{2\sqrt{2k}}(\overline{\alpha} - \overline{\beta}) = \frac{n}{2\sqrt{2k}} \sum_{m\in\mathcal{M}} \binom{\frac{3n}{4k}-2}{m-2} \frac{a_m - b_m}{\binom{n}{m-1}},$$

where $\mathcal{M}$ is obtained from Algorithm 3. By the definition of $\overline{\mathbf{A}}_1$ in equation (20), the least non-trivial singular value of $\overline{\mathbf{A}}_1$ is

$$\sigma_k(\overline{\mathbf{A}}_1) = \sigma_k(\overline{\mathbf{B}}_1) = \frac{n}{2\sqrt{2k}}(\overline{\alpha} - \overline{\beta}) = \frac{n}{2\sqrt{2k}} \sum_{m\in\mathcal{M}} \binom{\frac{3n}{4k}-2}{m-2} \frac{a_m - b_m}{\binom{n}{m-1}}.$$

Recall that n_i, defined in equation (12), denotes the number of vertices in $Z \cap V_i$, which can be written as $n_i = \sum_{v\in V_i} \mathbf{1}_{\{v\in Z\}}$. By Hoeffding's Lemma D.2,

$$\mathbb{P}\left(\left|n_i - \frac{n}{2k}\right| \ge \sqrt{n}\log(n)\right) \le 2\exp\left(-k\log^2(n)\right).$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

Similarly, n_i', defined in equation (13), satisfies

$$\mathbb{P}\left(\left|n_i' - \frac{n}{4k}\right| \geq \sqrt{n}\log(n)\right) \leq 2\exp\left(-k\log^2(n)\right).$$

As defined in equations (14) and (16), both $\widetilde{\mathbf{B}}_1$ and $\overline{\mathbf{B}}_1$ are deterministic block matrices. Then with probability at least $1 - 2k\exp\left(-k\log^2(n)\right)$, the dimensions of each block inside $\widetilde{\mathbf{B}}_1$ and $\overline{\mathbf{B}}_1$ are approximately the same, with deviations up to $\sqrt{n}\log(n)$. Consequently, the matrix $\widetilde{\mathbf{A}}_1$, which was defined in equation (19), can be treated as a perturbed version of $\overline{\mathbf{A}}_1$. By Weyl's inequality (Lemma D.5), for any $i \in [k]$,

$$\begin{aligned}|\sigma_i(\overline{\mathbf{B}}_1) - \sigma_i(\widetilde{\mathbf{B}}_1)| &= |\sigma_i(\overline{\mathbf{A}}_1) - \sigma_i(\widetilde{\mathbf{A}}_1)| \leq \|\overline{\mathbf{A}}_1 - \widetilde{\mathbf{A}}_1\|_2 \leq \|\overline{\mathbf{A}}_1 - \widetilde{\mathbf{A}}_1\|_{\mathrm{F}} \\ &\leq \left[2k \cdot \frac{n}{k} \cdot \sqrt{n}\log(n) \cdot (\overline{\alpha} - \overline{\beta})^2\right]^{1/2} = O\left(n^{3/4}\log^{1/2}(n) \cdot (\overline{\alpha} - \overline{\beta})\right).\end{aligned}$$

As a result, with probability at least $1 - 2k\exp\left(-k\log^2(n)\right)$, we have

$$\frac{|\sigma_k(\overline{\mathbf{A}}_1) - \sigma_k(\widetilde{\mathbf{A}}_1)|}{\sigma_k(\overline{\mathbf{A}}_1)} = \frac{|\sigma_k(\overline{\mathbf{B}}_1) - \sigma_k(\widetilde{\mathbf{B}}_1)|}{\sigma_k(\overline{\mathbf{B}}_1)} = O\left(n^{-1/4}\log^{1/2}(n)\right).$$

□

B.3. Proof of Lemma 5.4

Proof. Without loss of generality, we can assume $\mathcal{M} = \{2, \ldots, M\}$. If $\mathcal{M}$ is a subset of $\{2, \ldots, M\}$, we can take $a_m = b_m = 0$ for $m \notin \mathcal{M}$. Note that in fact, if the best SNR is obtained when $\mathcal{M}$ is a strict subset, we can substitute $\mathcal{M}_{\max}$ for M.

Let $X \subset V$ be a subset of vertices in hypergraph $H = (V, E)$ with size $|X| = cn$ for some $c \in (0, 1)$ to be decided later. Suppose X is a set of vertices with high degrees that we want to zero out. We first count the m-uniform hyperedges on X separately, then weight them by $(m-1)$, and finally sum over m to compute the row sums in $\mathbf{A}$ corresponding to each vertex in X. Let $E_m(X)$ denote the set of m-uniform hyperedges with all vertices located in X, and $E_m(X^c)$ denote the set of m-uniform hyperedges with all vertices in $X^c = V \setminus X$, respectively. Let $E_m(X, X^c)$ denote the set of m-uniform hyperedges with at least 1 endpoint in X and 1 endpoint in X^c. The relationship between total row sums and the number of non-uniform hyperedges in the vertex set X can be expressed as

$$\sum_{v \in X} \mathrm{row}(v) \leq \sum_{m=2}^{M} (m-1)\Big(m|E_m(X)| + (m-1)|E_m(X, X^c)|\Big) \tag{B.1}$$

If the row sum of each vertex $v \in X$ is at least $20Md$, where $d = \sum_{m=2}^{M}(m-1)a_m$, it follows

$$\sum_{m=2}^{M} (m-1)\Big(m|E_m(X)| + (m-1)|E_m(X, X^c)|\Big) \geq cn \cdot (20Md). \tag{B.2}$$

Then either

$$\sum_{m=2}^{M} m(m-1)|E_m(X)| \geq 4Mcnd, \quad \text{or} \quad \sum_{m=2}^{M} (m-1)^2|E_m(X, X^c)| \geq 16Mcnd.$$

B.3.1. Concentration of $\sum_{m=2}^{M} m(m-1)|E_m(X)|$.

Recall that $\left|E_m(X)\right|$ denotes the number of m-uniform hyperedges with all vertices located in X, which can be viewed as the sum of independent Bernoulli random variables $T_e^{(a_m)}$ and $T_e^{(b_m)}$ given by

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

$$T_e^{(a_m)} \sim \text{Bernoulli}\left(\frac{a_m}{\binom{n}{m-1}}\right), \quad T_e^{(b_m)} \sim \text{Bernoulli}\left(\frac{b_m}{\binom{n}{m-1}}\right). \tag{B.3}$$

Let $\{V_1, \ldots, V_k\}$ be the true partition of V. Suppose that there are $\eta_i cn$ vertices in block $V_i \cap X$ for each $i \in [k]$ with restriction $\sum_{i=1}^k \eta_i = 1$, then $\left|E_m(X)\right|$ can be written as

$$\left|E_m(X)\right| = \sum_{e \in E_m(X, a_m)} T_e^{(a_m)} + \sum_{e \in E_m(X, b_m)} T_e^{(b_m)},$$

where $E_m(X, a_m) := \cup_{i=1}^k E_m(V_i \cap X)$ denotes the union for sets of hyperedges with all vertices in the same block $V_i \cap X$ for some $i \in [k]$, and

$$E_m(X, b_m) := \ E_m(X) \setminus E_m(X, a_m) = E_m(X) \setminus \left(\cup_i^k E_m(V_i \cap X)\right)$$

denotes the set of hyperedges with vertices crossing different $V_i \cap X$. We can compute the expectation of $\left|E_m(X)\right|$ as

$$\mathbb{E}|E_m(X)| = \sum_{i=1}^k \binom{\eta_i cn}{m} \frac{a_m - b_m}{\binom{n}{m-1}} + \binom{cn}{m} \frac{b_m}{\binom{n}{m-1}}. \tag{B.4}$$

Then

$$\sum_{m=2}^M m(m-1) \cdot \mathbb{E}|E_m(X)| = \sum_{m=2}^M m(m-1) \left[\sum_{i=1}^k \binom{\eta_i cn}{m} \frac{a_m - b_m}{\binom{n}{m-1}} + \binom{cn}{m} \frac{b_m}{\binom{n}{m-1}} \right]. \tag{B.5}$$

As $\sum_{i=1}^k \eta_i = 1$, it follows that $\sum_{i=1}^k \binom{\eta_i cn}{m} \le \binom{cn}{m}$ by induction, thus

$$\frac{a_m - b_m}{\binom{n}{m-1}} \sum_{i=1}^k \binom{\eta_i cn}{m} + \frac{b_m}{\binom{n}{m-1}} \binom{cn}{m} = \frac{a_m}{\binom{n}{m-1}} \sum_{i=1}^k \binom{\eta_i cn}{m} + \frac{b_m}{\binom{n}{m-1}} \left(\binom{cn}{m} - \sum_{i=1}^k \binom{\eta_i cn}{m} \right)$$

where both terms on the right are positive numbers. Using this and taking $b_m = a_m$, we obtain the following upper bound for all n,

$$\sum_{m=2}^M m(m-1)\mathbb{E}|E_m(X)| \le \sum_{m=2}^M m(m-1) \binom{cn}{m} \frac{a_m}{\binom{n}{m-1}} \le cn \sum_{m=2}^M (m-1) a_m = cnd\,.$$

Note that $\sum_{m=2}^M m(m-1)|E_m(X)|$ is a weighted sum of independent Bernoulli random variables (corresponding to hyperedges), each upper bounded by M^2. Also, its variance is bounded by

$$\begin{aligned} \sigma^2 := \ \mathbb{V}\text{ar}\left(\sum_{m=2}^M m(m-1)|E_m(X)|\right) &= \sum_{m=2}^M m^2(m-1)^2 \mathbb{V}\text{ar}\left(|E_m(X)|\right) \\ &\le \sum_{m=2}^M m^2(m-1)^2 \mathbb{E}|E_m(X)| \le M^2 cnd. \end{aligned}$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

We can apply Bernstein's Lemma D.3 and obtain

$$\begin{aligned}
&\mathbb{P}\left(\sum_{m=2}^{M} m(m-1)|E_m(X)| \geq 4Mcnd\right) \\
&\leq \mathbb{P}\left(\sum_{m=2}^{M} m(m-1)(|E_m(X)| - \mathbb{E}|E_m(X)|) \geq 3Mcnd\right) \\
&\leq \exp\left(-\frac{(3Mcnd)^2}{M^2cnd + M^2cnd/3}\right) \leq \exp\left(-6cnd\right). \qquad \text{(B.6)}
\end{aligned}$$

B.3.2. Concentration of $\sum_{m=2}^{M}(m-1)^2|E_m(X, X^c)|$.

For any finite set S, let $[S]^j$ denote the family of j-subsets of S, i.e., $[S]^j = \{Z | Z \subseteq S, |Z| = j\}$. Let $E_m([Y]^j, [Z]^{m-j})$ denote the set of m-hyperedges, where j vertices are from Y and $m-j$ vertices are from Z within each m-hyperedge. We want to count the number of m-hyperedges between X and X^c, according to the number of vertices located in X^c within each m-hyperedge. Suppose that there are j vertices from X^c within each m-hyperedge for some $1 \leq j \leq m-1$.

(i) Assume that all those j vertices are in the same $[V_i \setminus X]^j$. If the remaining $m-j$ vertices are from $[V_i \cap X]^{m-j}$, then this m-hyperedge is connected with probability $a_m/\binom{n}{m-1}$, otherwise $b_m/\binom{n}{m-1}$. The number of this type m-hyperedges can be written as

$$\sum_{i=1}^{k}\left[\sum_{e\in\mathcal{E}_{j,i}^{(a_m)}} T_e^{(a_m)} + \sum_{e\in\mathcal{E}_{j,i}^{(b_m)}} T_e^{(b_m)}\right],$$

where $\mathcal{E}_{j,i}^{(a_m)} := E_m([V_i \cap X^c]^j, [V_i \cap X]^{m-j})$, and

$$\mathcal{E}_{j,i}^{(b_m)} := E_m\Big([V_i \cap X^c]^j,\ [X]^{m-j} \setminus [V_i \cap X]^{m-j}\Big)$$

denotes the set m-hyperedges with j vertices in $[V_i \cap X^c]^j$ and the remaining $m-j$ vertices in $[X]^j \setminus [V_i \cap X]^j$. We compute all possible choices and upper bound the cardinality of $\mathcal{E}_{j,i}^{(a_m)}$ and $\mathcal{E}_{j,i}^{(b_m)}$ by

$$\left|\mathcal{E}_{j,i}^{(a_m)}\right| \leq \binom{(\frac{1}{k}-\eta_i c)n}{j}\binom{\eta_i cn}{m-j}, \quad \left|\mathcal{E}_{j,i}^{(b_m)}\right| \leq \binom{(\frac{1}{k}-\eta_i c)n}{j}\left[\binom{cn}{m-j} - \binom{\eta_i cn}{m-j}\right].$$

(ii) If those j vertices in $[V \setminus X]^j$ are not in the same $[V_i \cap X]^j$ (which only happens $j \geq 2$), then the number of this type hyperedges can be written as $\sum_{e\in\mathcal{E}_j^{(b_m)}} T_e^{(b_m)}$, where

$$\begin{aligned}
\mathcal{E}_j^{(b_m)} &:= E_m\Big([V \setminus X]^j \setminus \big(\cup_{i=1}^{k}[V_i \setminus X]^j\big),\ [X]^{m-j}\Big), \\
\left|\mathcal{E}_j^{(b_m)}\right| &\leq \left[\binom{(1-c)n}{j} - \sum_{i=1}^{k}\binom{(\frac{1}{k}-\eta_i c)n}{j}\right]\binom{cn}{m-j}.
\end{aligned}$$

Therefore, $|E_m(X, X^c)|$ can be written as a sum of independent Bernoulli random variables,

$$|E_m(X, X^c)| = \sum_{j=1}^{m-1}\sum_{i=1}^{k}\left[\sum_{e\in\mathcal{E}_{j,i}^{(a_m)}} T_e^{(a_m)} + \sum_{e\in\mathcal{E}_{j,i}^{(b_m)}} T_e^{(b_m)}\right] + \sum_{j=2}^{m-1}\sum_{e\in\mathcal{E}_j^{(b_m)}} T_e^{(b_m)}. \qquad \text{(B.7)}$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

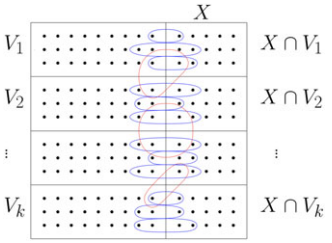


Figure B.1. Comparison of f_c and g_c.

Then the expectation can be rewritten as

$$\begin{aligned}
&\mathbb{E}\left(|E_m(X, X^c)|\right) \\
&= \sum_{j=1}^{m-1}\sum_{i=1}^{k}\binom{(\frac{1}{k}-\eta_i c)n}{j}\left\{\binom{\eta_i cn}{m-j}\frac{a_m}{\binom{n}{m-1}} + \left[\binom{cn}{m-j}-\binom{\eta_i cn}{m-j}\right]\frac{b_m}{\binom{n}{m-1}}\right\} \\
&\quad + \sum_{j=1}^{m-1}\left[\binom{(1-c)n}{j} - \sum_{i=1}^{k}\binom{(\frac{1}{k}-\eta_i c)n}{j}\right]\binom{cn}{m-j}\frac{b_m}{\binom{n}{m-1}} \qquad \text{(B.8)} \\
&= \sum_{j=1}^{m-1}\sum_{i=1}^{k}\binom{(\frac{1}{k}-\eta_i c)n}{j}\binom{\eta_i cn}{m-j}\frac{a_m - b_m}{\binom{n}{m-1}} + \sum_{j=1}^{m-1}\binom{(1-c)n}{j}\binom{cn}{m-j}\frac{b_m}{\binom{n}{m-1}} \\
&= \sum_{i=1}^{k}\left[\binom{\frac{n}{k}}{m} - \binom{\eta_i cn}{m} - \binom{(\frac{1}{k}-\eta_i c)n}{m}\right]\frac{a_m - b_m}{\binom{n}{m-1}} + \left[\binom{n}{m} - \binom{cn}{m} - \binom{(1-c)n}{m}\right]\frac{b_m}{\binom{n}{m-1}},
\end{aligned}$$

where we used the fact $\binom{(1-c)n}{1} = \sum_{i=1}^{k}\binom{(1/k-\eta_i c)n}{1}$ in the first equality and Vandermonde's identity $\binom{n_1+n_2}{m} = \sum_{j=0}^{m}\binom{n_1}{j}\binom{n_2}{m-j}$ in last equality. Note that

$$f_c := \binom{n}{m} - \binom{cn}{m} - \binom{(1-c)n}{m}$$

counts the number of subsets of V with m elements such that at least one element belongs to X and at least one element belongs to X^c. On the other hand,

$$g_c = \sum_{i=1}^{k}\left[\binom{\frac{n}{k}}{m} - \binom{\eta_i cn}{m} - \binom{(\frac{1}{k}-\eta_i c)n}{m}\right].$$

counts the number of subsets of V with m elements such that all elements belong to a single V_i, and given such an i, that at least one element belongs to $X \cap V_i$ and at least one belongs to $X^c \cap V_i$.

As Fig. B.1 shows, g_c only counts the blue pairs while f_c counts red pairs in addition. By virtue of the fact that there are fewer conditions imposed on the sets included in the count for f_c, we must have $f_c \geq g_c$. Thus, rewriting equation (B.8), we obtain

$$\mathbb{E}(|E_m(X, X^c)|) = g_c\frac{a_m}{\binom{n}{m-1}} + (f_c - g_c)\frac{b_m}{\binom{n}{m-1}}.$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

Since both terms in the above sum are positive, we can upper bound by taking $a_m = b_m$ to obtain

$$\mathbb{E}\left(|E_m(X, X^c)|\right) \leq f_c \frac{a_m}{\binom{n}{m-1}} = \left[\binom{n}{m} - \binom{cn}{m} - \binom{(1-c)n}{m}\right] \frac{a_m}{\binom{n}{m-1}}.$$

By summing over m, the expectation of $\sum_{m=2}^{M} (m-1)^2 |E_m(X, X^c)|$ satisfies

$$\begin{aligned}\sum_{m=2}^{M} (m-1)^2 \cdot \mathbb{E}\left(|E_m(X, X^c)|\right) &\leq \sum_{m=2}^{M} (m-1)^2 \left[\binom{n}{m} - \binom{cn}{m} - \binom{(1-c)n}{m}\right] \frac{a_m}{\binom{n}{m-1}}, \\ &\leq 2n \sum_{m=2}^{M} (1 - c^m - (1-c)^m)(m-1)a_m \leq 8Mcnd,\end{aligned}$$

where the last upper inequality holds when $c \in (0, 2^{1/M} - 1]$, since

$$\begin{aligned}&[(1-c)+c]^m - c^m - (1-c)^m \\ &= \binom{m}{1}(1-c)^{m-1}c + \binom{m}{2}(1-c)^{m-2}c^2 + \cdots + \binom{m}{m-1}(1-c)^1 c^{m-1} \\ &\leq \binom{m}{1}c + \binom{m}{2}c^2 + \cdots + \binom{m}{m-1}c^{m-1} \leq (1+c)^m - 1 \leq 2mc, \end{aligned} \tag{B.9}$$

where the last inequality holds by the following Claim.

Claim B.1. *Let $m \geq 2$ be some finite integer. Then for $0 < c <= 2^{1/m} - 1$, it follows that $(1+c)^m - 1 \leq 2mc$.*

Proof of the Claim We finish the proof by induction. First, the argument $(1+c)^j - 1 \leq 2jc$ holds true for the base cases $j = 1, 2$. Suppose that the argument holds for the case $j \geq 2$. For the case $j + 1 \leq m$, it follows that

$$(1+c)^{j+1} - 1 = (1+c)^j + c(1+c)^j - 1 \leq 2jc + c(1+c)^j \leq 2(j+1)c,$$

where the last inequality holds true if $c(1+c)^j \leq 2c$, and it holds since $c \leq 2^{1/m} - 1 \leq 2^{1/j} - 1$ for all $j \leq m$. □

Similarly, we apply Bernstein Lemma D.3 again with $K = M^2$, $\sigma^2 \leq 8M^3cnd$ and obtain

$$\begin{aligned}&\mathbb{P}\left(\sum_{m=2}^{M} (m-1)^2 |E_m(X, X^c)| \geq 16Mcnd\right) \\ &\leq \mathbb{P}\left(\sum_{m=2}^{M} (m-1)^2 (|E_m(X, X^c)| - \mathbb{E}|E_m(X, X^c)|) \geq 8Mcnd\right) \leq \exp\left(-6cnd/M\right). \end{aligned} \tag{B.10}$$

By the binomial coefficient upper bound $\binom{n}{k} \leq (\frac{en}{k})^k$ for $1 \leq k \leq n$, there are at most

$$\binom{n}{cn} \leq \left(\frac{e}{c}\right)^{cn} = \exp\left(-c(\log c - 1)n\right) \tag{B.11}$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

many subsets X of size $|X| = cn$. Let d be sufficiently large so that $d^{-3} \le c_0$. Substituting $c = d^{-3}$ in equation (B.11), we have

$$\binom{n}{d^{-3}n} \le \exp\left[3d^{-3}\log(d)n\right].$$

Taking $c = d^{-3}$ in equations (B.6) and (B.10), we obtain

$$\mathbb{P}\left(\sum_{m=2}^{M}(m-1)(m|E_m(X)| + (m-1)|E_m(X, X^c)|) \ge 20Md^{-2}n\right) \le 2\exp\left(-2d^{-2}n/M\right).$$

Taking a union bound over all possible X with $|X| = d^{-3}n$, we obtain with probability at least $1 - 2\exp(3d^{-3}\log dn - 2d^{-2}n/M) \le 1 - 2\exp(-d^{-2}n/M)$, no more than $d^{-3}n$ many vertices have total row sum greater than $20Md$. Note that we have imposed the condition that $c = d^{-3} \in (0, 2^{1/M} - 1]$ in (B.9), thus $d \ge (2^{1/M} - 1)^{-1/3}$, producing the lower bound in Assumption 1.5. □

B.4. Proof of Lemma 5.8

Proof. Note that $\mathbf{U}$ is spanned by first k singular vectors of $(\mathbf{A}_1)_{\mathcal{I}_1}$. Let $\{\boldsymbol{u}_i\}_{i=1}^k$ be an orthonormal basis of the $\mathbf{U}$, then the projection $P_{\mathbf{U}} := \sum_{l=1}^k \langle \boldsymbol{u}_l, \cdot\rangle \boldsymbol{u}_l$. Let $k(i)$ index the membership of vertex i. For each fixed $i \in V_{k(i)} \cap Y_2 \cap \{i_1, \cdots, i_s\}$,

$$P_{\mathbf{U}}\boldsymbol{e}_i = \sum_{l=1}^{k}\langle \boldsymbol{u}_l, \boldsymbol{e}_i\rangle \boldsymbol{u}_l, \quad \|P_{\mathbf{U}}\boldsymbol{e}_i\|_2^2 = \sum_{l=1}^{k}\langle \boldsymbol{u}_l, \boldsymbol{e}_i\rangle^2.$$

As a consequence of independence between entries in $\mathbf{A}_1$ and entries in $\mathbf{A}_2$, defined in equation (18), it is known that $\{\boldsymbol{u}_l\}_{l=1}^k$ and $\boldsymbol{e}_i$ are independent of each other, since $\boldsymbol{e}_i$ are columns of $\mathbf{E}_2 := \mathbf{A}_2 - \widetilde{\mathbf{A}}_2$. If the expectation is taken over $\{\boldsymbol{\mathcal{A}}^{(m)}\}_{m\in\mathcal{M}}$ conditioning on $\{\boldsymbol{u}_l\}_{l=1}^k$, then

$$\mathbb{E}_{\{\boldsymbol{\mathcal{A}}^{(m)}\}_{m\in\mathcal{M}}}\left[\langle \boldsymbol{u}_l, \boldsymbol{e}_i\rangle \middle| \{\boldsymbol{u}_l\}_{l=1}^k\right] = \sum_{j=1}^{n}\boldsymbol{u}_l(j)\cdot\mathbb{E}\left(\left[(\mathbf{A}_2)_{ji} - (\mathbb{E}\mathbf{A}_2)_{ji}\right]\right) = 0,$$

$$\mathbb{E}_{\{\boldsymbol{\mathcal{A}}^{(m)}\}_{m\in\mathcal{M}}}\left[\|P_{\mathbf{U}}\boldsymbol{e}_i\|_2^2 \middle| \{\boldsymbol{u}_l\}_{l=1}^k\right] = \sum_{l=1}^{k}\mathbb{E}_{\{\boldsymbol{\mathcal{A}}^{(m)}\}_{m\in\mathcal{M}}}\left[\langle \boldsymbol{u}_l, \boldsymbol{e}_i\rangle^2 \middle| \{\boldsymbol{u}_l\}_{l=1}^k\right],$$

where $\mathcal{M}$ is obtained from Algorithm 3. Expand each $\langle \boldsymbol{u}_l, \boldsymbol{e}_i\rangle^2$ and rewrite it into 2 parts,

$$\begin{aligned}\langle \boldsymbol{u}_l, \boldsymbol{e}_i\rangle^2 &= \sum_{j_1=1}^{k}\sum_{j_2=1}^{k}\boldsymbol{u}_l(j_1)\boldsymbol{e}_i(j_1)\boldsymbol{u}_l(j_2)\boldsymbol{e}_i(j_2)\\ &= \underbrace{\sum_{j=1}^{k}[\boldsymbol{u}_l(j)]^2[\boldsymbol{e}_i(j)]^2}_{(a)} + \underbrace{\sum_{j_1\neq j_2}\boldsymbol{u}_l(j_1)\boldsymbol{e}_i(j_1)\boldsymbol{u}_l(j_2)\boldsymbol{e}_i(j_2)}_{(b)}, \quad \forall l \in [k]. \end{aligned} \tag{B.12}$$

Part (a) is the contribution from graph, i.e., 2-uniform hypergraph, while part (b) is the contribution from m-uniform hypergraph with $m \ge 3$, which only occurs in hypergraph clustering. The expectation of part (a) in is upper bounded by α as defined in equation (8), since

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

$$\mathbb{E}_{\{\mathcal{A}^{(m)}\}_{m\in\mathcal{M}}}\left[\sum_{j=1}^{k}[\boldsymbol{u}_l(j)]^2[\boldsymbol{e}_i(j)]^2\middle|\{\boldsymbol{u}_l\}_{l=1}^{k}\right]=\sum_{j=1}^{n}[\boldsymbol{u}_l(j)]^2\cdot\mathbb{V}\mathrm{ar}\Big((\mathbf{A}_2)_{ji}\Big)$$
$$\le\sum_{j=1}^{n}[\boldsymbol{u}_l(j)]^2\cdot(\mathbb{E}\mathbf{A}_2)_{ji}\le\alpha=\sum_{m\in\mathcal{M}}\left[\binom{\frac{n}{k}-2}{m-2}\frac{a_m-b_m}{\binom{n}{m-1}}+\binom{n}{m-2}\frac{b_m}{\binom{n}{m-1}}\right]$$
$$\le\sum_{m\in\mathcal{M}}\binom{n}{m-2}\frac{a_m}{\binom{n}{m-1}}\le\frac{2}{n}\sum_{m\in\mathcal{M}}(m-1)a_m=\frac{2d}{n},\quad\forall l\in[k]$$

where $\|\boldsymbol{u}_l\|_2^2=\sum_{j=1}^{n}[\boldsymbol{u}_l(j)]^2=1$. For part (b),

$$\mathbb{E}_{\{\mathcal{A}^{(m)}\}_{m\in\mathcal{M}}}\left[\sum_{j_1\neq j_2}\boldsymbol{u}_l(j_1)\boldsymbol{e}_i(j_1)\boldsymbol{u}_l(j_2)\boldsymbol{e}_i(j_2)\middle|\{\boldsymbol{u}_l\}_{l=1}^{k}\right]$$
$$=\sum_{j_1\neq j_2}\boldsymbol{u}_l(j_1)\boldsymbol{u}_l(j_2)\mathbb{E}\left[\Big((\mathbf{A}_2)_{j_1i}-(\mathbb{E}\mathbf{A}_2)_{j_1i}\Big)\Big((\mathbf{A}_2)_{j_2i}-(\mathbb{E}\mathbf{A}_2)_{j_2i}\Big)\right]$$
$$=\sum_{j_1\neq j_2}\boldsymbol{u}_l(j_1)\boldsymbol{u}_l(j_2)\mathbb{E}\left(\sum_{m\in\mathcal{M}}\sum_{\substack{e\in E_m[Y_2\cup Z]\\{i,j_1\}\subset e}}(\mathcal{A}_e^{(m)}-\mathbb{E}\mathcal{A}_e^{(m)})\right)\left(\sum_{m\in\mathcal{M}}\sum_{\substack{e\in E_m[Y_2\cup Z]\\{i,j_2\}\subset e}}(\mathcal{A}_e^{(m)}-\mathbb{E}\mathcal{A}_e^{(m)})\right).$$

According to Definition 2.1 of the adjacency tensor, $\mathcal{A}_{e_1}^{(m)}$ and $\mathcal{A}_{e_2}^{(m)}$ are independent if hyperedge $e_1\neq e_2$, then only the terms with hyperedge $e\supset\{i,j_1,j_2\}$ have non-zero contribution. Then the expectation of part (b) can be rewritten as

$$\mathbb{E}_{\{\mathcal{A}^{(m)}\}_{m\in\mathcal{M}}}\left[\sum_{j_1\neq j_2}\boldsymbol{u}_l(j_1)\boldsymbol{e}_i(j_1)\boldsymbol{u}_l(j_2)\boldsymbol{e}_i(j_2)\middle|\{\boldsymbol{u}_l\}_{l=1}^{k}\right]$$
$$=\sum_{j_1\neq j_2}\boldsymbol{u}_l(j_1)\boldsymbol{u}_l(j_2)\sum_{m\in\mathcal{M}}\sum_{\substack{e\in E_m[Y_2\cup Z]\\{i,j_1,j_2\}\subset e}}\mathbb{E}\big(\mathcal{A}_e^{(m)}-\mathbb{E}\mathcal{A}_e^{(m)}\big)^2$$
$$\le\sum_{j_1\neq j_2}\boldsymbol{u}_l(j_1)\boldsymbol{u}_l(j_2)\sum_{m\in\mathcal{M}}\sum_{\substack{e\in E_m[Y_2\cup Z]\\{i,j_1,j_2\}\subset e}}\mathbb{E}\mathcal{A}_e^{(m)}$$
$$=\sum_{j_1\neq j_2}\boldsymbol{u}_l(j_1)\boldsymbol{u}_l(j_2)\sum_{m\in\mathcal{M}}\sum_{\substack{e\in E_m[Y_2\cup Z]\\{i,j_1,j_2\}\subset e}}\frac{a_m}{\binom{n}{m-1}}.\tag{B.13}$$

Note that $|Y_2\cup Z|\le n$, then the number of possible hyperedges e, while $e\in E_m[Y_2\cup Z]$ and $e\supset\{i,j_1,j_2\}$, is at most $\binom{n}{m-3}$. Thus equation (B.13) is upper bounded by

$$\sum_{j_1\neq j_2}\boldsymbol{u}_l(j_1)\boldsymbol{u}_l(j_2)\sum_{m\in\mathcal{M}}\binom{n}{m-3}\frac{a_m}{\binom{n}{m-1}}$$
$$\le\sum_{j_1\neq j_2}\boldsymbol{u}_l(j_1)\boldsymbol{u}_l(j_2)\sum_{m\in\mathcal{M}}\frac{(m-1)(m-2)}{(n-m)^2}a_m\le\frac{d\mathcal{M}_{\max}}{n^2}\sum_{j_1\neq j_2}\boldsymbol{u}_l(j_1)\boldsymbol{u}_l(j_2)$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

$$\leq \frac{d\mathcal{M}_{\max}}{2n^2}\sum_{j_1\neq j_2}\left([\boldsymbol{u}_l(j_1)]^2+[\boldsymbol{u}_l(j_2)]^2\right)\leq \frac{d\mathcal{M}_{\max}(n-1)}{2n^2}\left(\sum_{j_1=1}^{n}[\boldsymbol{u}_l(j_1)]^2+\sum_{j_2=1}^{n}[\boldsymbol{u}_l(j_2)]^2\right)$$
$$\leq \frac{d\mathcal{M}_{\max}}{n},$$

where $\|\boldsymbol{u}_l\|_2=1$, $d=\sum_{m\in\mathcal{M}}(m-1)a_m$. With the upper bounds for part (a) and (b) in equation (B.12), the conditional expectation of $\|P_{\mathbf{U}}\boldsymbol{e}_i\|_2^2$ is bounded by

$$\mathbb{E}_{\{\boldsymbol{\mathcal{A}}^{(m)}\}_{m\in\mathcal{M}}}\left[\|P_{\mathbf{U}}\boldsymbol{e}_i\|_2^2\middle|\{\boldsymbol{u}_l\}_{l=1}^k\right]=\sum_{l=1}^{k}\mathbb{E}_{\{\boldsymbol{\mathcal{A}}^{(m)}\}_{m\in\mathcal{M}}}\left[\langle\boldsymbol{u}_l,\boldsymbol{e}_i\rangle^2\middle|\{\boldsymbol{u}_l\}_{l=1}^k\right]\leq\frac{kd}{n}(\mathcal{M}_{\max}+2),$$

Let X_i be the Bernoulli random variable defined by

$$X_i=\mathbb{1}\{\|P_{\mathbf{U}}\boldsymbol{e}_i\|_2>2\sqrt{kd(\mathcal{M}_{\max}+2)/n)}\},\quad i\in\{i_1,\cdots,i_s\}.$$

By Markov's inequality,

$$\mathbb{E}X_i=\mathbb{P}\left(\|P_{\mathbf{U}}\boldsymbol{e}_i\|_2>2\sqrt{kd(\mathcal{M}_{\max}+2)/n}\right)\leq\frac{\mathbb{E}_{\{\boldsymbol{\mathcal{A}}^{(m)}\}_{m\in\mathcal{M}}}\left[\|P_{\mathbf{U}}\boldsymbol{e}_i\|_2^2\middle|\{\boldsymbol{u}_l\}_{l=1}^k\right]}{4kd(\mathcal{M}_{\max}+2)/n}\leq\frac{1}{4}.$$

Let $\delta:=\frac{s}{2\sum_{j=1}^{s}\mathbb{E}X_{i_j}}-1$ where $s=2k\log^2(n)$. By Hoeffding Lemma D.2,

$$\mathbb{P}\left(\sum_{j=1}^{s}X_{i_j}\geq\frac{s}{2}\right)=\mathbb{P}\left(\sum_{j=1}^{s}(X_{i_j}-\mathbb{E}X_{i_j})\geq\delta\sum_{j=1}^{s}\mathbb{E}X_{i_j}\right)$$
$$\leq\exp\left(-\frac{2\delta^2\left(\sum_{j=1}^{s}\mathbb{E}X_{i_j}\right)^2}{s}\right)=O\left(\frac{1}{n^{k\log(n)}}\right).$$

Therefore, with probability $1-O(n^{-k\log(n)})$, at least $s/2$ of the vectors $\boldsymbol{e}_{i_1},\ldots,\boldsymbol{e}_{i_s}$ satisfy

$$\|P_{\mathbf{U}}\boldsymbol{e}_i\|_2\leq 2\sqrt{kd(\mathcal{M}_{\max}+2)/n}.$$

Meanwhile, for any $c\in(0,2)$, there exists some large enough constant $C_2\geq 2^{\mathcal{M}_{\max}+1}\sqrt{(\mathcal{M}_{\max}+2)/k}/c$ such that if $\sum_{m\in\mathcal{M}}(m-1)(a_m-b_m)>C_2k^{\mathcal{M}_{\max}-1}\sqrt{d}$, then

$$\|\bar{\boldsymbol{\delta}}_i\|_2=\frac{\sqrt{n}(\bar{\alpha}-\bar{\beta})}{2}=\frac{\sqrt{n}}{2}\sum_{m\in\mathcal{M}}\binom{\frac{3n}{4k}-2}{m-2}\frac{a_m-b_m}{\binom{n}{m-1}}$$
$$=\frac{(1+o(1))}{2\sqrt{n}}\sum_{m\in\mathcal{M}}\left(\frac{3}{4k}\right)^{m-2}(m-1)(a_m-b_m)\geq\frac{k}{(2k)^{\mathcal{M}_{\max}-1}\sqrt{n}}\sum_{m\in\mathcal{M}}(m-1)(a_m-b_m),$$
$$>\frac{C_2k^{\mathcal{M}_{\max}}\sqrt{d}}{(2k)^{\mathcal{M}_{\max}-1}\sqrt{n}}\geq\frac{2^{\mathcal{M}_{\max}+1}\sqrt{(\mathcal{M}_{\max}+2)}k^{\mathcal{M}_{\max}}\sqrt{d}}{c(2k)^{\mathcal{M}_{\max}-1}\sqrt{kn}}$$
$$>\frac{2}{c}2\sqrt{\frac{kd(\mathcal{M}_{\max}+2)}{n}}>\frac{2}{c}\|P_{\mathbf{U}}\boldsymbol{e}_i\|_2.$$

□

B.5. Proof of Lemma 5.10

Proof. Split $[n]$ into V_1' and V_2' such that $V_1'=\{i|\boldsymbol{v}(i)>0\}$ and $V_2'=\{i|\boldsymbol{v}(i)\leq 0\}$. Without loss of generality, assume that the first $\frac{n}{k}$ entries of $\bar{\boldsymbol{v}}$ are positive. We can write $\boldsymbol{v}$ in terms of its orthogonal projection onto $\bar{\boldsymbol{v}}$ as

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

$$\boldsymbol{v} = c_1\bar{\boldsymbol{v}} + \boldsymbol{\varepsilon} = \left[\boldsymbol{\varepsilon}_1 + \frac{c_1}{\sqrt{n}}, \cdots, \boldsymbol{\varepsilon}_{\frac{n}{k}} + \frac{c_1}{\sqrt{n}}, \ \boldsymbol{\varepsilon}_{\frac{n}{k}+1} - \frac{c_1}{\sqrt{n}}, \cdots, \boldsymbol{\varepsilon}_n - \frac{c_1}{\sqrt{n}}\right]^{\mathsf{T}}, \tag{B.14}$$

where $\bar{\boldsymbol{v}} \perp \boldsymbol{\varepsilon}$ with $\|\boldsymbol{\varepsilon}\|_2 < c$ and $c_1 \geq \sqrt{1-c^2}$. The number of entries of $\boldsymbol{\varepsilon}$ smaller than $-\frac{\sqrt{1-c^2}}{\sqrt{n}}$ is at most $\frac{c^2}{1-c^2}n$. Note that $c_1 \geq \sqrt{1-c^2}$, so at least $\frac{n}{k} - \frac{c^2}{1-c^2}n$ indices i with $\bar{\boldsymbol{v}}_i = \frac{1}{\sqrt{n}}$ will have $\boldsymbol{v}_i > 0$, thus the ratio we are seeking is at least

$$\frac{\frac{n}{k} - \frac{c^2}{1-c^2}n}{\frac{n}{k}} = 1 - \frac{kc^2}{1-c^2} > 1 - \frac{4k}{3}c^2.$$

□

B.6. Proof of Lemma 5.11

Proof. We start with the following simple claim: for any $m \geq 2$ and any $\nu \in [1/2, 1)$,

$$\nu^m + (1-\nu)^m < \left(\frac{1+\nu}{2}\right)^m. \tag{B.15}$$

Indeed, one quick way to see this is by induction on m; we will induct from m to $m+2$. Assume the inequality is true for m; then

$$\begin{aligned}
\nu^{m+2} + (1-\nu)^{m+2} &= \nu^2\nu^m + (1-\nu)^2(1-\nu)^m \\
&= \nu^2\nu^m + (1-2\nu+\nu^2)(1-\nu)^m \ \leq \ \nu^2\nu^m + \nu^2(1-\nu)^m \\
&= \nu^2(\nu^m + (1-\nu)^m) \ < \ (\nu^2 + (1-\nu)^2)(\nu^m + (1-\nu)^m) \\
&< \left(\frac{1+\nu}{2}\right)^2 \left(\frac{1+\nu}{2}\right)^m = \left(\frac{1+\nu}{2}\right)^{m+2},
\end{aligned}$$

where we have used the induction hypothesis together with $1-2\nu \leq 0$ and $(1-\nu)^2 > 0$. After easily checking that the inequality works for $m = 2, 3$, the induction is complete. We shall now check that the quantities defined in Lemma 5.11 obey the relationship $\mu_2 \geq \mu_1$ and $\mu_2 - \mu_1 = \Omega(n)$, for n large enough. First, note that the only thing we need to check is that, for sufficiently large n,

$$\binom{\frac{\nu n}{2k}}{m} + \binom{\frac{(1-\nu)n}{2k}}{m} \leq \binom{\frac{(1+\nu)n}{4k}}{m} + (k-1)\binom{\frac{(1-\nu)n}{4k(k-1)}}{m};$$

in fact, we will show the stronger statement that for any $m \geq 2$ and n large enough,

$$\binom{\frac{\nu n}{2k}}{m} + \binom{\frac{(1-\nu)n}{2k}}{m} < \binom{\frac{(1+\nu)n}{4k}}{m}, \tag{B.16}$$

and this will suffice to see that the second part of the assertion, $\mu_2 - \mu_1 = \Omega(n)$, is also true. Asymptotically, $\binom{\frac{\nu n}{2k}}{m} \sim \frac{\nu^m}{m!}\left(\frac{n}{2k}\right)^m$, $\binom{\frac{(1-\nu)n}{2k}}{m} \sim \frac{(1-\nu)^m}{m!}\left(\frac{n}{2k}\right)^m$, and $\binom{\frac{(1+\nu)n}{4k}}{m} \sim \frac{\left(\frac{1+\nu}{2}\right)^m}{m!}\left(\frac{n}{2k}\right)^m$. Note then that equation (B.16) follows from equation (B.15).

Let $\{V_1, \ldots, V_k\}$ be the true partition of V. Recall that hyperedges in $H = \cup_{m\in\mathcal{M}} H_m$ are coloured red and blue with equal probability in Algorithm 2. Let $E_m(X)$ denote the set of blue m-uniform hyperedges with all vertices located in the vertex set X. Assume $|X \cap V_i| = \eta_i|X|$ with $\sum_{i=1}^k \eta_i = 1$. For each $m \in \mathcal{M}$, the presence of hyperedge $e \in E_m(X)$ can be represented by independent Bernoulli random variables

$$T_e^{(a_m)} \sim \text{Bernoulli}\left(\frac{a_m}{2\binom{n}{m-1}}\right), \quad T_e^{(b_m)} \sim \text{Bernoulli}\left(\frac{b_m}{2\binom{n}{m-1}}\right),$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

depending on whether e is a hyperedge with all vertices in the same block. Denote by

$$E_m(X, a_m) := \cup_{i=1}^{k} E_m(V_i \cap X)$$

the union of all m-uniform sets of hyperedges with all vertices in the same $V_i \cap X$ for some $i \in [k]$, and by

$$E_m(X, b_m) := \ E_m(X) \setminus E_m(X, a_m) = E_m(X) \setminus \Big(\cup_i^k E_m(V_i \cap X) \Big)$$

the set of m-uniform hyperedges with vertices across different blocks $V_i \cap X$. Then the cardinality $|E_m(X)|$ can be written as the

$$|E_m(X)| = \sum_{e \in E_m(X, a_m)} T_e^{(a_m)} + \sum_{e \in E_m(X, b_m)} T_e^{(b_m)},$$

and by summing over m, the weighted cardinality $|E(X)|$ is written as

$$|E(X)| := \sum_{m \in \mathcal{M}} m(m-1)|E_m(X)| = \sum_{m \in \mathcal{M}} m(m-1) \left\{ \sum_{e \in E_m(X, a_m)} T_e^{(a_m)} + \sum_{e \in E_m(X, b_m)} T_e^{(b_m)} \right\},$$

with its expectation

$$\mathbb{E}|E(X)| = \sum_{m \in \mathcal{M}} m(m-1) \left\{ \sum_{i=1}^{k} \binom{\eta_i \frac{n}{2k}}{m} \frac{a_m - b_m}{2\binom{n}{m-1}} + \binom{\frac{n}{2k}}{m} \frac{b_m}{2\binom{n}{m-1}} \right\}, \tag{B.17}$$

since

$$|E_m(X, a_m)| = \sum_{i=1}^{k} |E_m(V_i \cap X)| = \sum_{i=1}^{k} \binom{\eta_i \frac{n}{2k}}{m}, \quad |E_m(X, b_m)| = \binom{\frac{n}{2k}}{m} - \sum_{i=1}^{k} \binom{\eta_i \frac{n}{2k}}{m},$$

Next, we prove the two Statements in Lemma 5.11 separately. First, assume that $|X \cap V_i| \le \nu |X|$ (i.e., $\eta_i \le \nu$) for each $i \in [k]$. Then

$$\mathbb{E}|E(X)| \le \frac{1}{2} \sum_{m \in \mathcal{M}} m(m-1) \left\{ \left[\binom{\frac{\nu n}{2k}}{m} + \binom{\frac{(1-\nu)n}{2k}}{m} \right] \frac{a_m - b_m}{\binom{n}{m-1}} + \binom{\frac{n}{2k}}{m} \frac{b_m}{\binom{n}{m-1}} \right\} =: \mu_1 .$$

To justify the above inequality, note that since $\sum_{i=1}^{k} \eta_i = 1$, the sum $\sum_{i=1}^{k} \binom{\eta_i \frac{n}{2k}}{m}$ is maximised when all but 2 of the η_i are 0, and since all $\eta_i \le \nu$, this means that

$$\sum_{i=1}^{k} \binom{\eta_i \frac{n}{2k}}{m} \le \binom{\frac{\nu n}{2k}}{m} + \binom{\frac{(1-\nu)n}{2k}}{m}.$$

Note that $m(m-1)(T_e^{(a_m)} - \mathbb{E}T_e^{(a_m)})$ and $m(m-1)(T_e^{(b_m)} - \mathbb{E}T_e^{(b_m)})$ are independent mean-zero random variables bounded by $M(M-1)$ for all $m \in \mathcal{M}$, and $\mathbb{V}\mathrm{ar}(|E(X)|) \le M^2(M-1)^2 \mathbb{E}|E(X)| = \Omega(n)$. Recall that $\mu_{\mathrm{T}} := (\mu_1 + \mu_2)/2$. Define $t = \mu_{\mathrm{T}} - \mathbb{E}|E(X)|$, then $0 < (\mu_2 - \mu_1)/2 \le t \le \mu_{\mathrm{T}}$, hence $t = \Omega(n)$. By Bernstein's Lemma D.3, we have

$$\mathbb{P}\Big(|E(X)| \ge \mu_{\mathrm{T}} \Big) = \mathbb{P}\Big(|E(X)| - \mathbb{E}|E(X)| \ge t \Big) \le \exp\Big(-\frac{t^2/2}{\mathbb{V}\mathrm{ar}(|E(X)|) + M(M-1)t/3} \Big) = O(e^{-cn}),$$

where $c > 0$ is some constant. On the other hand, if $|X \cap V_i| \ge \frac{1+\nu}{2}|X|$ for some $i \in [k]$, then

$$\mathbb{E}|E(X)| \ge \frac{1}{2} \sum_{m \in \mathcal{M}} m(m-1) \left\{ \left[\binom{\frac{(1+\nu)n}{4k}}{m} + (k-1) \binom{\frac{(1-\nu)n}{4k(k-1)}}{m} \right] \frac{a_m - b_m}{\binom{n}{m-1}} + \binom{\frac{n}{2k}}{m} \frac{b_m}{\binom{n}{m-1}} \right\} =: \mu_2 .$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

The above can be justified by noting that at least one $|X \cap V_i| \geq \frac{1+\nu}{2}|X|$, and that the rest of the vertices will yield a minimal binomial sum when they are evenly split between the remaining V_j. Similarly, define $t = \mu_{\mathrm{T}} - \mathbb{E}|E(X)|$, then $0 < (\mu_2 - \mu_1)/2 \leq -t = \Omega(n)$, and Bernstein's Lemma D.3 gives

$$\begin{aligned}\mathbb{P}\Big(|E(X)| \leq \mu_{\mathrm{T}}\Big) &= \mathbb{P}\Big(|E(X)| - \mathbb{E}|E(X)| \leq -t\Big) \\ &\leq \exp\left(-\frac{t^2/2}{\mathbb{V}\mathrm{ar}(|E(X)|) + M(M-1)(-t)/3}\right) = O(e^{-c'n}),\end{aligned}$$

where $c' > 0$ is some other constant. □

B.7. Proof of Lemma 5.13

Proof. If vertex i is uniformly chosen from Y_2, the probability that $i \notin V_l$ for some $l \in [k]$ is

$$\mathbb{P}(i \notin V_l | i \in Y_2) = \frac{\mathbb{P}(i \notin V_l, i \in Y_2)}{\mathbb{P}(i \in Y_2)} = 1 - \frac{|V_l \cap Y_2|}{|Y_2|} = 1 - \frac{\frac{n}{k} - n_l - n_l'}{n - \sum_{t=1}^{k}(n_t + n_t')}, \quad l \in [k],$$

where n_t and n_t', defined in equations (12) and (13), denote the cardinality of $Z \cap V_t$ and $Y_1 \cap V_t$ respectively. As proved in Appendix B.2, with probability at least $1 - 2\exp(-k\log^2(n))$, we have

$$|n_t - n/(2k)| \leq \sqrt{n}\log(n) \text{ and } |n_t' - n/(4k)| \leq \sqrt{n}\log(n),$$

then $\mathbb{P}(i \notin V_l | i \in Y_2) = 1 - \frac{1}{k}\Big(1 + o(1)\Big)$. After $k\log^2 n$ samples from Y_2, the probability that there exists at least one node which belongs to V_l is at least

$$1 - \left(1 - \frac{1+o(1)}{k}\right)^{k\log^2 n} = 1 - n^{-(1+o(1))k\log\left(\frac{k}{k-1}\right)\log n}.$$

The proof is completed by a union bound over $l \in [k]$. □

B.8. Proof of Lemma 5.14

Proof. We calculate $\mathbb{P}(S_{11}'(u) \leq \mu_{\mathrm{C}})$ first. Define $t_{1\mathrm{C}} := \mu_{\mathrm{C}} - \mathbb{E}S_{11}'(u)$, then by Bernstein's inequality (Lemma D.3) and taking $K = \mathcal{M}_{\max} - 1$,

$$\begin{aligned}&\mathbb{P}\left(S_{11}'(u) \leq \mu_{\mathrm{C}}\right) = \mathbb{P}\left(S_{11}'(u) - \mathbb{E}S_{11}'(u) \leq t_{1\mathrm{C}}\right) \\ &\leq \exp\left(-\frac{t_{1\mathrm{C}}^2/2}{\mathbb{V}\mathrm{ar}[S_{11}'(u)] + (\mathcal{M}_{\max}-1)\cdot t_{1\mathrm{C}}/3}\right) \leq \exp\left(-\frac{3t_{1\mathrm{C}}^2/(\mathcal{M}_{\max}-1)}{6(\mathcal{M}_{\max}-1)\cdot\mathbb{E}S_{11}'(u) + 2t_{1\mathrm{C}}}\right) \\ &\leq \exp\left(-\frac{[(\nu)^{\mathcal{M}_{\max}-1} - (1-\nu)^{\mathcal{M}_{\max}-1}]^2}{(\mathcal{M}_{\max}-1)^2\cdot 2^{2\mathcal{M}_{\max}+3}} \cdot \frac{\left[\sum_{m\in\mathcal{M}}(m-1)\left(\frac{a_m - b_m}{k^{m-1}}\right)\right]^2}{\sum_{m\in\mathcal{M}}(m-1)\left(\frac{a_m-b_m}{k^{m-1}} + b_m\right)}\right),\end{aligned}$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

where $\mathcal{M}$ is obtained from Algorithm 3 with $\mathcal{M}_{\max}$ denoting the maximum value in $\mathcal{M}$, and the last two inequalities hold since $\mathbb{V}\mathrm{ar}[S'_{11}(u)] \le (\mathcal{M}_{\max}-1)^2\mathbb{E}S'_{11}(u)$, and for sufficiently large n,

$$\begin{aligned} t_{1\mathrm{C}} &:= \mu_{\mathrm{C}} - \mathbb{E}S'_{11}(u) = -\frac{1}{2}\sum_{m\in\mathcal{M}}(m-1)\cdot\left[\binom{\frac{\nu n}{2k}}{m-1} - \binom{\frac{(1-\nu)n}{2k}}{m-1}\right]\frac{a_m-b_m}{2\binom{n}{m-1}} \\ &\le -\frac{1}{2}\sum_{m\in\mathcal{M}}\frac{(\nu)^{m-1}-(1-\nu)^{m-1}}{2^m}\cdot(m-1)\frac{a_m-b_m}{k^{m-1}}(1+o(1)) \\ &\le -\frac{(\nu)^{\mathcal{M}_{\max}-1}-(1-\nu)^{\mathcal{M}_{\max}-1}}{2^{\mathcal{M}_{\max}+2}}\sum_{m\in\mathcal{M}}(m-1)\cdot\frac{a_m-b_m}{k^{m-1}}, \end{aligned}$$

$$\begin{aligned} &6(\mathcal{M}_{\max}-1)\mathbb{E}S'_{11}(u) + 2t_{1\mathrm{C}} = 2\mu_{\mathrm{C}} + (6\mathcal{M}_{\max}-8)\mathbb{E}S'_{11}(u) \\ &= \sum_{m\in\mathcal{M}}(m-1)\left\{\left[(6\mathcal{M}_{\max}-7)\binom{\frac{\nu n}{2k}}{m-1} + \binom{\frac{(1-\nu)n}{2k}}{m-1}\right]\frac{a_m-b_m}{2\binom{n}{m-1}}\right. \\ &\qquad \left. +6(\mathcal{M}_{\max}-1)\binom{\frac{n}{2k}}{m-1}\frac{b_m}{2\binom{n}{m-1}}\right\} \\ &= \sum_{m\in\mathcal{M}}(m-1)\left[\frac{(6\mathcal{M}_{\max}-7)\cdot(\nu)^{m-1}+(1-\nu)^{m-1}}{2^m}\cdot\frac{a_m-b_m}{k^{m-1}} + \frac{(6\mathcal{M}_{\max}-6)b_m}{2^m k^{m-1}}\right](1+o(1)) \\ &\le \sum_{m\in\mathcal{M}}\frac{(6\mathcal{M}_{\max}-7)\cdot(\nu)^{m-1}+(1-\nu)^{m-1}}{2^m}\cdot(m-1)\left(\frac{a_m-b_m}{k^{m-1}}+b_m\right)(1+o(1)) \\ &\le \frac{3(\mathcal{M}_{\max}-1)}{2}\sum_{m\in\mathcal{M}}(m-1)\left(\frac{a_m-b_m}{k^{m-1}}+b_m\right). \end{aligned}$$

Similarly, for $\mathbb{P}(S'_{1j}(u)\ge\mu_{\mathrm{C}})$, define $t_{j\mathrm{C}} := \mu_{\mathrm{C}} - \mathbb{E}S'_{1j}(u)$ for $j\neq 1$, by Bernstein's Lemma D.3,

$$\begin{aligned} &\mathbb{P}\left(S'_{1j}(u)\ge\mu_{\mathrm{C}}\right) = \mathbb{P}\left(S'_{1j}(u)-\mathbb{E}S'_{1j}(u)\ge t_{j\mathrm{C}}\right) \\ &\le \exp\left(-\frac{t_{j\mathrm{C}}^2/2}{\mathbb{V}\mathrm{ar}[S'_{1j}(u)]+(\mathcal{M}_{\max}-1)\cdot t_{j\mathrm{C}}/3}\right) \le \exp\left(-\frac{3t_{j\mathrm{C}}^2/(\mathcal{M}_{\max}-1)}{6(\mathcal{M}_{\max}-1)\cdot\mathbb{E}S'_{1j}(u)+2t_{j\mathrm{C}}}\right) \\ &\le \exp\left(-\frac{[(\nu)^{\mathcal{M}_{\max}-1}-(1-\nu)^{\mathcal{M}_{\max}-1}]^2}{(\mathcal{M}_{\max}-1)^2\cdot 2^{2\mathcal{M}_{\max}+3}}\cdot\frac{\left[\sum_{m\in\mathcal{M}}(m-1)\left(\frac{a_m-b_m}{k^{m-1}}\right)\right]^2}{\sum_{m\in\mathcal{M}}(m-1)\left(\frac{a_m-b_m}{k^{m-1}}+b_m\right)}\right). \end{aligned}$$

The last two inequalities holds since $\mathbb{V}\mathrm{ar}[S'_{1j}(u)] \le (\mathcal{M}_{\max}-1)^2\mathbb{E}S'_{1j}(u)$, and for sufficiently large n,

$$\begin{aligned} t_{j\mathrm{C}} &:= \mu_{\mathrm{C}} - \mathbb{E}S'_{1j}(u) = \frac{1}{2}\sum_{m\in\mathcal{M}}(m-1)\cdot\left[\binom{\frac{\nu n}{2k}}{m-1} - \binom{\frac{(1-\nu)n}{2k}}{m-1}\right]\frac{a_m-b_m}{2\binom{n}{m-1}} \\ &\ge \frac{(\nu)^{\mathcal{M}_{\max}-1}-(1-\nu)^{\mathcal{M}_{\max}-1}}{2^{\mathcal{M}_{\max}+2}}\sum_{m\in\mathcal{M}}(m-1)\cdot\frac{a_m-b_m}{k^{m-1}}, \end{aligned}$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

$$
\begin{aligned}
&6\mathbb{E}S'_{1j}(u) + 2t_{\mathrm{jC}} = 2\mu_{\mathrm{C}} + (6\mathcal{M}_{\max} - 8)\mathbb{E}S'_{1j}(u) \\
&= \sum_{m\in\mathcal{M}} (m-1)\left\{\left[\binom{\frac{\nu n}{2k}}{m-1} + (6\mathcal{M}_{\max}-7)\binom{\frac{(1-\nu)n}{2k}}{m-1}\right]\frac{a_m - b_m}{2\binom{n}{m-1}}\right. \\
&\qquad \left. +6(\mathcal{M}_{\max}-1)\binom{\frac{n}{2k}}{m-1}\frac{b_m}{2\binom{n}{m-1}}\right\} \\
&= \sum_{m\in\mathcal{M}} (m-1)\cdot\left(\frac{(\nu)^{m-1} + (6\mathcal{M}_{\max}-7)\cdot(1-\nu)^{m-1}}{2^m}\cdot\frac{a_m - b_m}{k^{m-1}} + \frac{(6\mathcal{M}_{\max}-6)\cdot b_m}{2^m k^{m-1}}\right)(1+o(1)) \\
&\le \sum_{m\in\mathcal{M}} \frac{(\nu)^{m-1} + (6\mathcal{M}_{\max}-7)\cdot(1-\nu)^{m-1}}{2^m}\cdot(m-1)\left(\frac{a_m - b_m}{k^{m-1}} + b_m\right)(1+o(1)) \\
&\le \frac{3(\mathcal{M}_{\max}-1)}{2}\sum_{m\in\mathcal{M}} (m-1)\left(\frac{a_m - b_m}{k^{m-1}} + b_m\right).
\end{aligned}
$$

□

Appendix C. Algorithm correctness for the binary case

We will show the correctness of Algorithm 1 and prove Theorem 1.6 in this section. The analysis will mainly follow from the analysis in Section 5. We only detail the differences.

Without loss of generality, we assume n is even to guarantee the existence of a binary partition of size $n/2$. The method to deal with the odd n case was discussed in Lemma 2.4. Then, let the index set be $\mathcal{I} = \{i \in [n] : \mathrm{row}(i) \le 20\mathcal{M}_{\max} d\}$, as shown in equation (11). Let $\boldsymbol{u}_i$ (resp. $\bar{\boldsymbol{u}}_i$) denote the eigenvector associated to $\lambda_i(\mathbf{A}_{\mathcal{I}})$ (resp. $\lambda_i(\overline{\mathbf{A}})$) for $i = 1, 2$. Define two linear subspaces $\mathbf{U} := \mathrm{Span}\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ and $\overline{\mathbf{U}} := \mathrm{Span}\{\bar{\boldsymbol{u}}_1, \bar{\boldsymbol{u}}_2\}$, then the angle between $\mathbf{U}$ and $\overline{\mathbf{U}}$ is defined as $\sin\angle(\mathbf{U}, \overline{\mathbf{U}}) := \|P_{\mathbf{U}} - P_{\overline{\mathbf{U}}}\|$, where $P_{\mathbf{U}}$ and $P_{\overline{\mathbf{U}}}$ are the orthogonal projections onto $\mathbf{U}$ and $\overline{\mathbf{U}}$, respectively.

C.1. Proof of Lemma 4.4

The strategy to bound the angle is similar to Subsection 5.1.2, except that we apply Davis-Kahan Theorem (Lemma D.6) here.

Define $\mathbf{E} := \mathbf{A} - \overline{\mathbf{A}}$ and its restriction on $\mathcal{I}$, namely $\mathbf{E}_{\mathcal{I}} := (\mathbf{A} - \overline{\mathbf{A}})_{\mathcal{I}} = \mathbf{A}_{\mathcal{I}} - \overline{\mathbf{A}}_{\mathcal{I}}$, as well as $\boldsymbol{\Delta} := \overline{\mathbf{A}}_{\mathcal{I}} - \overline{\mathbf{A}}$. Then the deviation $\mathbf{A}_{\mathcal{I}} - \overline{\mathbf{A}}$ is decomposed as

$$\mathbf{A}_{\mathcal{I}} - \overline{\mathbf{A}} = (\mathbf{A}_{\mathcal{I}} - \overline{\mathbf{A}}_{\mathcal{I}}) + (\overline{\mathbf{A}}_{\mathcal{I}} - \overline{\mathbf{A}}) = \mathbf{E}_{\mathcal{I}} + \boldsymbol{\Delta}\,.$$

Theorem 3.3 indicates $\|\mathbf{E}_{\mathcal{I}}\| \le C_3\sqrt{d}$ with probability at least $1 - n^{-2}$ when taking $\tau = 20\mathcal{M}_{\max}$, $K = 3$, where C_3 is a constant depending only on $\mathcal{M}_{\max}$. Moreover, Lemma 5.4 shows that the number of vertices with high degrees is relatively small. Consequently, an argument similar to Corollary 5.5 leads to the conclusion $\|\boldsymbol{\Delta}\| \le \sqrt{d}$ *w.h.p.* Together with upper bounds for $\|\mathbf{E}_{\mathcal{I}}\|$ and $\|\boldsymbol{\Delta}\|$, Lemma C.1 shows that the angle between $\mathbf{U}$ and $\overline{\mathbf{U}}$ is relatively small with high probability.

Lemma C.1. *For any $c \in (0, 1)$, there exists a constant C_2 depending on $\mathcal{M}_{\max}$ and c such that if*

$$\sum_{m\in\mathcal{M}} (m-1)(a_m - b_m) \ge C_2 \cdot 2^{\mathcal{M}_{\max}+2}\sqrt{d},$$

then $\sin\angle(\mathbf{U}, \overline{\mathbf{U}}) \le c$ *with probability* $1 - n^{-2}$.

Proof. First, with probability $1 - n^{-2}$, we have

$$\|\mathbf{A}_{\mathcal{I}} - \overline{\mathbf{A}}\| \le \|\mathbf{E}_{\mathcal{I}}\| + \|\boldsymbol{\Delta}\| \le (C_3 + 1)\sqrt{d}.$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

According to the definitions in equation (8), $\alpha \geq \beta$ and $\alpha = O(1/n)$, $\beta = O(1/n)$. Meanwhile, Lemma 2.3 shows that $|\lambda_2(\overline{\mathbf{A}})| = [-\alpha + (\alpha - \beta)n/2]$ and $|\lambda_3(\overline{\mathbf{A}})| = \alpha$. The

$$|\lambda_2(\overline{\mathbf{A}})| - |\lambda_3(\overline{\mathbf{A}})| = \frac{n}{2}(\alpha - \beta) - 2\alpha \geq \frac{3}{4} \cdot \frac{n}{2}(\alpha - \beta) = \frac{3n}{8} \sum_{m \in \mathcal{M}} \binom{\frac{n}{2} - 2}{m - 2} \frac{(a_m - b_m)}{\binom{n}{m-1}}$$

$$\geq \frac{1}{4} \sum_{m \in \mathcal{M}} \frac{(m-1)(a_m - b_m)}{2^{m-2}} \geq \frac{1}{2^{\mathcal{M}_{\max}}} \sum_{m \in \mathcal{M}} (m-1)(a_m - b_m) \geq 4C_2\sqrt{d}\,.$$

Then for some large enough C_2, the following condition for Davis-Kahan Theorem (Lemma D.6) is satisfied

$$\|\mathbf{A}_{\mathcal{I}} - \overline{\mathbf{A}}\| \leq (1 - 1/\sqrt{2})\left(|\lambda_2(\overline{\mathbf{A}})| - |\lambda_3(\overline{\mathbf{A}})|\right).$$

Then for any $c \in (0, 1)$, we can choose $C_2 = (C_3 + 1)/c$ such that

$$\|P_{\mathbf{U}} - P_{\overline{\mathbf{U}}}\| \leq \frac{2\|\mathbf{A}_{\mathcal{I}} - \overline{\mathbf{A}}\|}{|\lambda_2(\overline{\mathbf{A}})| - |\lambda_3(\overline{\mathbf{A}})|} \leq \frac{2(C_3+1)\sqrt{d}}{4C_2\sqrt{d}} = \frac{c}{2} \leq c\,.$$

□

Now, we focus on the accuracy of Algorithm 7, once the conditions in Lemma C.1 are satisfied.

Lemma C.2 (Lemma 23 in [19]). *If* $\sin\angle(\overline{\mathbf{U}}, \mathbf{U}) \leq c \leq \frac{1}{4}$, *there exists a unit vector* $\boldsymbol{v} \in \mathbf{U}$ *such that the angle between* $\bar{\boldsymbol{u}}_2$ *and* $\boldsymbol{v}$ *satisfies* $\sin\angle(\bar{\boldsymbol{u}}_2, \boldsymbol{v}) \leq 2\sqrt{c}$.

The desired vector $\boldsymbol{v}$, as constructed in Algorithm 7, is the unit vector perpendicular to $P_{\mathbf{U}}\mathbf{1}_n$, where $P_{\mathbf{U}}\mathbf{1}_n$ is the projection of all-ones vector onto $\mathbf{U}$. Lemmas C.1 and C.2 together give the following corollary.

Corollary C.3. *For any* $c \in (0, 1)$, *there exists a unit vector* $\boldsymbol{v} \in \mathbf{U}$ *such that the angle between* $\bar{\boldsymbol{u}}_2$ *and* $\boldsymbol{v}$ *satisfies* $\sin\angle(\bar{\boldsymbol{u}}_2, \boldsymbol{v}) \leq c < 1$ *with probability* $1 - O(e^{-n})$.

Proof. For any $c \in (0, 1)$, we could choose constants C_2, C_3 in Lemma C.1 such that $\sin\angle(\overline{\mathbf{U}}, \mathbf{U}) \leq \frac{c^2}{4} < 1$. Then by Lemma C.2, we construct $\boldsymbol{v}$ such that $\sin\angle(\bar{\boldsymbol{u}}_2, \boldsymbol{v}) \leq c$. □

Lemma C.4 (Lemma 23 in [19]). *If* $\sin\angle(\bar{\boldsymbol{u}}_2, \boldsymbol{v}) < c \leq 0.5$, *then we can identify at least* $(1 - 8c^2/3)n$ *vertices from each block correctly.*

The proof of Lemma 4.4 is completed when choosing C_2, C_3 in Lemma C.1 s.t. $c \leq \frac{1}{4}$.

C.2 Proof of Lemma 4.5

The proof strategy is similar to Subsections 5.2 and 5.3. In Algorithm 1, we first colour the hyperedges with red and blue with equal probability. By running Algorithm 8 on the red graph, we obtain a ν-correct partition V_1', V_2' of $V = V_1 \cup V_2$, i.e., $|V_l \cap V_l'| \geq \nu n/2$ for $l = 1, 2$. In the rest of the proof, we condition on this event and the event that the maximum red degree of a vertex is at most $\log^2(n)$ with probability at least $1 - o(1)$. This can be proved by Bernstein's inequality (Lemma D.3).

Similarly, we consider the probability of a hyperedge $e = \{i_1, \cdots, i_m\}$ being blue conditioning on the event that e is not a red hyperedge in each underlying m-uniform hypergraph separately. If vertices $i_1, \cdots, i_m$ are all from the same true cluster, then the probability is ψ_m, otherwise ϕ_m, where ψ_m and ϕ_m are defined in equations (29) and (30), and the presence of those hyperedges are represented by random variables $\zeta_e^{(a_m)} \sim$ Bernoulli (ψ_m), $\xi_e^{(b_m)} \sim$ Bernoulli (ϕ_m), respectively.

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

Following a similar argument in Subsection 5.2, the row sum of u can be written as

$$S'_{lj}(u) := \sum_{m\in\mathcal{M}} (m-1)\cdot \left\{ \sum_{e\in \mathcal{E}^{(a_m)}_{l,j}} \zeta_e^{(a_m)} + \sum_{e\in\mathcal{E}^{(b_m)}_{l,j}} \xi_e^{(b_m)} \right\}, \quad u\in V_l,$$

where $\mathcal{E}^{(a_m)}_{l,j} := E_m([V_l]^1, [V_l\cap V'_j]^{m-1})$ denotes the set of m-hyperedges with 1 vertex from $[V_l]^1$ and the other $m-1$ vertices from $[V_l\cap V'_j]^{m-1}$, while $\mathcal{E}^{(b_m)}_{l,j} := E_m\Big([V_l]^1,\ [V'_j]^{m-1}\setminus [V_l\cap V'_j]^{m-1}\Big)$ denotes the set of m-hyperedges with 1 vertex in $[V_l]^1$ while the remaining $m-1$ vertices in $[V'_j]^{m-1}\setminus [V_l\cap V'_j]^{m-1}$, with their cardinalities

$$|\mathcal{E}^{(a_m)}_{l,j}| \le \binom{|V_l\cap V'_j|}{m-1}, \quad |\mathcal{E}^{(b_m)}_{l,j}| \le \left[\binom{|V'_j|}{m-1} - \binom{|V_l\cap V'_j|}{m-1}\right].$$

According to the fact $|V_l\cap V'_l|\ge \nu n/2$, $|V_l|=n/2$, $|V'_l|=n/2$ for $l=1,2$, we have

$$|\mathcal{E}^{(a_m)}_{l,l}| \ge \binom{\frac{\nu n}{2}}{m-1}, \quad |\mathcal{E}^{(a_m)}_{l,j}| \le \binom{\frac{(1-\nu)n}{2}}{m-1},\ j\ne l.$$

To simplify the calculation, we take the lower and upper bound of $|\mathcal{E}^{(a_m)}_{l,l}|$ and $|\mathcal{E}^{(a_m)}_{l,j}|(j\ne l)$ respectively. Taking expectation with respect to $\zeta_e^{(a_m)}$ and $\xi_e^{(b_m)}$, for any $u\in V_l$, we have

$$\mathbb{E}S'_{ll}(u) = \sum_{m\in\mathcal{M}} (m-1)\cdot\left[\binom{\frac{\nu n}{2}}{m-1}(\psi_m-\phi_m) + \binom{\frac{n}{2}}{m-1}\phi_m\right],$$

$$\mathbb{E}S'_{lj}(u) = \sum_{m\in\mathcal{M}} (m-1)\cdot\left[\binom{\frac{(1-\nu)n}{2}}{m-1}(\psi_m-\phi_m) + \binom{\frac{n}{2}}{m-1}\phi_m\right],\ j\ne l.$$

By assumptions in Theorem 1.7, $\mathbb{E}S'_{ll}(u) - \mathbb{E}S'_{lj}(u) = \Omega(1)$. We define

$$\mu_C := \frac{1}{2}\sum_{m\in\mathcal{M}} (m-1)\cdot\left\{\left[\binom{\frac{\nu n}{2}}{m-1} + \binom{\frac{(1-\nu)n}{2}}{m-1}\right](\psi_m-\phi_m) + 2\binom{\frac{n}{2}}{m-1}\phi_m\right\}.$$

After Algorithm 6, if a vertex $u\in V_l$ is mislabelled, one of the following events must happen

- $S'_{ll}(u)\le \mu_C$,
- $S'_{lj}(u)\ge \mu_C$, for some $j\ne l$.

By an argument similar to Lemma 5.14, we can prove that

$$\rho'_1 = \mathbb{P}\left(S'_{ll}(u)\le \mu_C\right)\le \rho, \quad \rho'_2 = \mathbb{P}\left(S'_{lj}(u)\ge \mu_C\right)\le \rho,$$

where $\rho = \exp(-C_{\mathcal{M}}(2)\cdot \mathrm{SNR}_{\mathcal{M}}(2))$ and

$$C_{\mathcal{M}}(2) := \frac{[(\nu)^{\mathcal{M}_{\max}-1} - (1-\nu)^{\mathcal{M}_{\max}-1}]^2}{8(\mathcal{M}_{\max}-1)^2}, \quad \mathrm{SNR}_{\mathcal{M}}(2) := \frac{\left[\sum_{m\in\mathcal{M}}(m-1)\left(\frac{a_m-b_m}{2^{m-1}}\right)\right]^2}{\sum_{m\in\mathcal{M}}(m-1)\left(\frac{a_m-b_m}{2^{m-1}}+b_m\right)}.$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

As a result, the probability that either of those events happened is bounded by ρ. The number of mislabelled vertices in V_1 after Algorithm 5 is at most

$$R_l = \sum_{i=1}^{|V_l \setminus V_l'|} \Gamma_i + \sum_{i=1}^{|V_l \cap V_l'|} \Lambda_i ,$$

where Γ_i (resp. Λ_i) are i.i.d indicator random variables with mean ρ_1' (resp. ρ_2'). Then

$$\mathbb{E}R_l \le \frac{n}{2}\rho_1' + \frac{(1-\nu)n}{2}\rho_2' = (1-\nu/2)n\rho .$$

where ν is the correctness after Algorithm 4. Let $t_l := (1+\nu/2)n\rho$, then by Chernoff Lemma D.1,

$$\mathbb{P}(R_l \ge n\rho) = \mathbb{P}[R_l - (1-\nu/2)n\rho \ge t_l] \le \mathbb{P}(R_l - \mathbb{E}R_l \ge t_l) \le e^{-ct_l} = O(e^{-n\rho}) ,$$

which means that with probability $1 - O(e^{-n\rho})$, the fraction of mislabelled vertices in V_l is smaller than 2ρ, i.e., the correctness of V_l is at least $\gamma := \max\{\nu, 1-2\rho\}$.

Appendix D. Useful Lemmas

Lemma D.1 (Chernoff's inequality, Theorem 2.3.6 in [69]). *Let X_i be independent Bernoulli random variables with parameters p_i. Consider their sum $S_N = \sum_{i=1}^N X_i$ and denote its mean by $\mu = \mathbb{E}S_N$. Then for any $\delta \in (0,1]$,*

$$\mathbb{P}\big(|S_N - \mu| \ge \delta\mu\big) \le 2\exp(-c\delta^2\mu) .$$

Lemma D.2 (Hoeffding's inequality, Theorem 2.2.6 in [69]). *Let $X_1, \dots, X_N$ be independent random variables with $X_i \in [a_i, b_i]$ for each $i \in \{1, \dots, N\}$. Then for any, $t \ge 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^N (X_i - \mathbb{E}X_i)\right| \ge t\right) \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^N (b_i - a_i)^2}\right) .$$

Lemma D.3 (Bernstein's inequality, Theorem 2.8.4 in [69]). *Let $X_1, \dots, X_N$ be independent mean-zero random variables such that $|X_i| \le K$ for all i. Let $\sigma^2 = \sum_{i=1}^N \mathbb{E}X_i^2$. Then for every $t \ge 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^N X_i\right| \ge t\right) \le 2\exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right) .$$

Lemma D.4 (Bennett's inequality, Theorem 2.9.2 in [69]). *Let $X_1, \dots, X_N$ be independent random variables. Assume that $|X_i - \mathbb{E}X_i| \le K$ almost surely for every i. Then for any $t > 0$, we have*

$$\mathbb{P}\left(\sum_{i=1}^N (X_i - \mathbb{E}X_i) \ge t\right) \le \exp\left(-\frac{\sigma^2}{K^2} \cdot h\left(\frac{Kt}{\sigma^2}\right)\right) ,$$

where $\sigma^2 = \sum_{i=1}^N \mathrm{Var}(X_i)$, and $h(u) := (1+u)\log(1+u) - u$.

Lemma D.5 (Weyl's inequality). *Let $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{m\times n}$ be two real $m \times n$ matrices, then $|\sigma_i(\mathbf{A}+\mathbf{E}) - \sigma_i(\mathbf{A})| \le \|\mathbf{E}\|$ for every $1 \le i \le \min\{m,n\}$. Furthermore, if $m = n$ and $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{n\times n}$ are real symmetric, then $|\lambda_i(\mathbf{A}+\mathbf{E}) - \lambda_i(\mathbf{A})| \le \|\mathbf{E}\|$ for all $1 \le i \le n$.*

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

Lemma D.6 (Davis-Kahan's sin $\mathbf{\Theta}$ Theorem, Theorem 2.2.1 in [15]). *Let* $\overline{\mathbf{M}}$ *and* $\mathbf{M} = \overline{\mathbf{M}} + \mathbf{E}$ *be two real symmetric* $n \times n$ *matrices,* $n \times n$ *matrices, with eigenvalue decompositions given respectively by*

$$\overline{\mathbf{M}} = \sum_{i=1}^{n} \overline{\lambda}_i \overline{\boldsymbol{u}}_i \overline{\boldsymbol{u}}_i^{\mathsf{T}} = \begin{bmatrix} \overline{\mathbf{U}} & \overline{\mathbf{U}}_\perp \end{bmatrix} \begin{bmatrix} \overline{\mathbf{\Lambda}} & \mathbf{0} \\ \mathbf{0} & \overline{\mathbf{\Lambda}}_\perp \end{bmatrix} \begin{bmatrix} \overline{\mathbf{U}}^{\mathsf{T}} \\ \overline{\mathbf{U}}_\perp^{\mathsf{T}} \end{bmatrix},$$

$$\mathbf{M} = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^{\mathsf{T}} = \begin{bmatrix} \mathbf{U} & \mathbf{U}_\perp \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_\perp \end{bmatrix} \begin{bmatrix} \mathbf{U}^{\mathsf{T}} \\ \mathbf{U}_\perp^{\mathsf{T}} \end{bmatrix}.$$

Here, $\{\overline{\lambda}_i\}_{i=1}^n$*(resp.* $\{\lambda_i\}_{i=1}^n$*) stand for the eigenvalues of* $\overline{\mathbf{M}}$*(resp.* $\mathbf{M}$*), and* $\overline{\boldsymbol{u}}_i$*(resp.* $\boldsymbol{u}_i$*) denotes the eigenvector associated* $\overline{\lambda}_i$*(resp.* λ_i*). Additionally, for some fixed integer* $r \in [n]$*, we denote*

$$\begin{aligned} \overline{\mathbf{\Lambda}} &:= \operatorname{diag}\{\overline{\lambda}_1, \ldots, \overline{\lambda}_r\}, \quad & \overline{\mathbf{\Lambda}}_\perp &:= \operatorname{diag}\{\overline{\lambda}_{r+1}, \ldots, \overline{\lambda}_n\}, \\ \overline{\mathbf{U}} &:= [\overline{\boldsymbol{u}}_1, \ldots, \overline{\boldsymbol{u}}_r] \in \mathbb{R}^{n \times r}, \quad & \overline{\mathbf{U}}_\perp &:= [\overline{\boldsymbol{u}}_{r+1}, \ldots, \overline{\boldsymbol{u}}_n] \in \mathbb{R}^{n \times (n-r)}. \end{aligned}$$

The matrices $\mathbf{\Lambda}$, $\mathbf{\Lambda}_\perp$, $\mathbf{U}$, $\mathbf{U}_\perp$ *are defined analogously. Assume that*

$$\text{eigenvalues}(\overline{\mathbf{\Lambda}}) \subseteq [\alpha, \beta], \quad \text{eigenvalues}(\mathbf{\Lambda}_\perp) \subseteq (-\infty, \alpha - \Delta] \cup [\beta + \Delta, \infty), \quad \alpha, \beta \in \mathbb{R}, \Delta > 0,$$

and the projection matrices are given by $P_{\mathbf{U}} := \mathbf{U}\mathbf{U}^{\mathsf{T}}$, $P_{\overline{\mathbf{U}}} := \overline{\mathbf{U}}\,\overline{\mathbf{U}}^{\mathsf{T}}$*, then one has* $\|P_{\mathbf{U}} - P_{\overline{\mathbf{U}}}\| \le (2\|\mathbf{E}\|/\Delta)$*. In particular, suppose that* $|\overline{\lambda}_1| \ge |\overline{\lambda}_2| \ge \cdots \ge |\overline{\lambda}_r| \ge |\overline{\lambda}_{r+1}| \ge \cdots |\overline{\lambda}_n|$ *(resp.* $|\lambda_1| \ge \cdots \ge |\lambda_n|$*). If* $\|\mathbf{E}\| \le (1 - 1/\sqrt{2})(|\overline{\lambda}|_r - |\overline{\lambda}|_{r+1})$*, then one has*

$$\|P_{\mathbf{U}} - P_{\overline{\mathbf{U}}}\| \le \frac{2\|\mathbf{E}\|}{|\overline{\lambda}_r| - |\overline{\lambda}_{r+1}|}.$$

Lemma D.7 (Wedin's sin $\mathbf{\Theta}$ Theorem, Theorem 2.3.1 in [15]). *Let* $\overline{\mathbf{M}}$ *and* $\mathbf{M} = \overline{\mathbf{M}} + \mathbf{E}$ *be two* $n_1 \times n_2$ *real matrices and* n_2 *real matrices and* $n_1 \ge n_2$*, with SVDs given respectively by*

$$\overline{\mathbf{M}} = \sum_{i=1}^{n_1} \overline{\sigma}_i \overline{\boldsymbol{u}}_i \overline{\boldsymbol{v}}_i^{\mathsf{T}} = \begin{bmatrix} \overline{\mathbf{U}} & \overline{\mathbf{U}}_\perp \end{bmatrix} \begin{bmatrix} \overline{\mathbf{\Sigma}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \overline{\mathbf{\Sigma}}_\perp & \mathbf{0} \end{bmatrix} \begin{bmatrix} \overline{\mathbf{V}}^{\mathsf{T}} \\ \overline{\mathbf{V}}_\perp^{\mathsf{T}} \end{bmatrix}$$

$$\mathbf{M} = \sum_{i=1}^{n_1} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}} = \begin{bmatrix} \mathbf{U} & \mathbf{U}_\perp \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_\perp & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}^{\mathsf{T}} \\ \mathbf{V}_\perp^{\mathsf{T}} \end{bmatrix}.$$

Here, $\overline{\sigma}_1 \ge \cdots \ge \overline{\sigma}_{n_1}$ *(resp.* $\sigma_1 \ge \cdots \ge \sigma_{n_1}$*) stand for the singular values of* $\overline{\mathbf{M}}$*(resp.* $\mathbf{M}$*),* $\overline{\boldsymbol{u}}_i$*(resp.* $\boldsymbol{u}_i$*) denotes the left singular vector associated with the singular value* $\overline{\sigma}_i$*(resp.* σ_i*), and* $\overline{\boldsymbol{v}}_i$*(resp.* $\boldsymbol{v}_i$*) denotes the right singular vector associated with the singular value* $\overline{\sigma}_i$*(resp.* σ_i*). In addition, for any fixed integer* $r \in [n]$*, we denote*

$$\begin{aligned} \mathbf{\Sigma} &:= \operatorname{diag}\{\sigma_1, \ldots, \sigma_r\}, \quad & \mathbf{\Sigma}_\perp &:= \operatorname{diag}\{\sigma_{r+1}, \ldots, \sigma_{n_1}\}, \\ \mathbf{U} &:= [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r] \in \mathbb{R}^{n_1 \times r}, \quad & \mathbf{U}_\perp &:= [\boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_{n_1}] \in \mathbb{R}^{n_1 \times (n_1 - r)}, \\ \mathbf{V} &:= [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r] \in \mathbb{R}^{n_2 \times r}, \quad & \mathbf{V}_\perp &:= [\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_{n_2}] \in \mathbb{R}^{n_2 \times (n_2 - r)}. \end{aligned}$$

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press

The matrices $\overline{\mathbf{\Sigma}}$, $\overline{\mathbf{\Sigma}}_{\perp}$, $\overline{\mathbf{V}}$, $\overline{\mathbf{V}}_{\perp}$ *are defined analogously. If* $\mathbf{E} = \mathbf{M} - \mathbf{M} - \overline{\mathbf{M}}$ *satisfies* $\|\mathbf{E}\| \le \overline{\sigma}_r$ *- with the projection matrices* $P_{\mathbf{U}} := \mathbf{U}\mathbf{U}^{\mathsf{T}}$*, one has*

$$\max\left\{\|P_{\mathbf{U}} - P_{\overline{\mathbf{U}}}\|, \|P_{\mathbf{V}} - P_{\overline{\mathbf{V}}}\|\right\} \le \frac{\sqrt{2}\max\left\{\|\mathbf{E}^{\mathsf{T}}\overline{\mathbf{U}}\|, \|\mathbf{E}\overline{\mathbf{V}}\|\right\}}{\overline{\sigma}_r - \overline{\sigma}_{r+1} - \|\mathbf{E}\|}.$$

In particular, if $\|\mathbf{E}\| \le (1 - 1/\sqrt{2})(\overline{\sigma}_r - \overline{\sigma}_{r+1})$*, then one has*

$$\max\left\{\|P_{\mathbf{U}} - P_{\overline{\mathbf{U}}}\|, \|P_{\mathbf{V}} - P_{\overline{\mathbf{V}}}\|\right\} \le \frac{\sqrt{2}\|\mathbf{E}\|}{\overline{\sigma}_r - \overline{\sigma}_{r+1}}.$$

Cite this article: Dumitriu I, Wang H-X, and Zhu Y (2025). Partial recovery and weak consistency in the non-uniform hypergraph stochastic block model. *Combinatorics, Probability and Computing* **34**, 1–51. https://doi.org/10.1017/S0963548324000166

https://doi.org/10.1017/S0963548324000166 Published online by Cambridge University Press