Strong Consistency with Limited Information for the Binary Hypergraph Stochastic Block Model

Yongce Li

Advisor: Professor Ioana Dumitriu

Honors Thesis

June 2024

Contents

1 Introduction

Community detection deals with the identification of clusters within a network where vertices are more densely connected internally than with the rest of the network. It is a fundamental task in network analysis and has many applications in social networks, biology, natural language processing, etc. For instance, Facebook can be viewed as a big network where users are represented as vertices and friendships as edges. In this context, community detection can help identify different communities such as family members, school friends, and work colleagues based on their friendship connections.

In real-world scenarios, interactions often involve more than two vertices, such as in group emails, research collaborations, and biological interactions. To address this complexity, this thesis focuses on the community detection problem on hypergraphs, where hyperedges can contain more than two vertices. The main problem we aim to solve is: given the connection information of a hypergraph (represented by the adjacency matrix or combinatorial Laplacian), can we accurately recover its clusters?

In the case of simple graphs, the adjacency matrix provides complete information about the edges. However, for hypergraphs, the adjacency matrix only offers aggregated information about connectivity, making the problem more challenging. Therefore, this thesis aims to develop algorithms capable of recovering the cluster structure of vertices given only the limited connection information available in hypergraphs.

1.1 Graphs and Hypergraphs

We first define graphs, hypergraphs, and the associated matrices. In this paper, we only consider simple undirected graphs and hypergraphs with unordered edges and without loops.

Definition 1.1. *(Graph)*
A graph is a pair $G = (V, E)$, where V is a set whose elements are called vertices, and $E \subseteq \{\{v_1, v_2\} \mid v_1, v_2 \in V \text{ and } v_1 \neq v_2\}$ is a set of unordered pairs of vertices, whose elements are called edges. The degree of a vertex $v \in V$ is the number of edges in G that contain v.

Definition 1.2. *(Adjacency matrix of a graph)*
For the graph $G = (V, E), |V| = n$, the adjacency matrix of G is a $n \times n$ symmetric matrix with

$$A_{ij} = \begin{cases} 1 \text{ if } (v_i, v_j) \in E \\ 0 \text{ if } (v_i, v_j) \notin E \end{cases}$$

In an ordinary graph, an edge only connects two different vertices. A hyperedge is a generalization of edge that can connect more than two vertices. Based on hyperedge, we define hypergraphs.

Definition 1.3. *(Hypergraph)*
A Hypergraph is a pair $H = (V, E)$, where V is a set whose elements are called vertices, and E is a non-empty set of hyperedges such that for each $e \in E$, e is a subset of V. Specifically, if $|e| = m$ for every $e \in E$, we call H a m-uniform hypergraph. The degree of a vertex $v \in V$ is the number of hyperedges in H that contain v.

Remark 1.4. *All vertices in a hyperedge are distinct; the hypergraph contains no loops.*

A hypergraph is a collection of m-uniform hypergraphs, that is, $H = \bigcup_m H_m$ with $E = \bigcup_m E_m$. One can associate an m-uniform hypergraph $H_m = ([n], E_m)$ to an order-m symmetric tensor $A^{(m)}$, where the entry $A^{(m)}_{v_1,\dots,v_m}$ denotes the presence of the m-hyperedge $e = \{v_1, \dots, v_m\}$, i.e.,

$$A^{(m)}_e := A^{(m)}_{v_1,\dots,v_m} = \mathbb{1}_{\{e \in E_m\}}, \quad \{v_1, \dots, v_m\} \subset [n].$$

However, for a non-uniform hypergraph, one is not able to aggregate information of each uniform layer because the tensors are of different orders. It's sometimes easier to work with the adjacency matrix defined below.

Definition 1.5. *(Adjacency matrix of a hypergraph)*
For the hypergraph $H = (V, E), |V| = n$, the adjacency matrix of H is a $n \times n$ symmetric matrix A with

$$A_{ij} := \mathbb{1}_{\{i \neq j\}} \cdot \sum_{m} \sum_{e \in E(H_m), e \supset \{i,j\}} \mathbb{1}_e,$$

Another important matrix is the Combinatorial Laplacian (or Laplacian matrix), which is often used to study the spectral properties of graphs and hypergraphs, where the second smallest eigenvalue of the Laplacian (or the Fiedler value) provides information about the graph's connectivity and is commonly used in partitioning problems. See [Chu97] for a detailed analysis.

Definition 1.6. *(Combinatorial Laplacian)*
For any graph $G = (V, E)$ and hypergraph $H = (V_H, E_H)$, the combinatorial Laplacian is defined as

$$L = D - A$$

where A is the adjacency matrix associated with G and H, and D is the degree matrix with $D_{ii} = \sum_j A_{ij}$.

Remark 1.7. *Hyperedges of size m containing v_i are counted $(m-1)$ times in D_{ii}.*

The normalized Laplacian matrix is a variant of the Laplacian matrix that utilizes a normalization factor $D^{-1/2}$ to adjust the influence of each vertex based on its degree, hence counterbalancing the impact of high-degree vertices in the analysis.

Definition 1.8. *For any graph $G = (V, E)$ and hypergraph $H = (V_H, E_H)$, the normalized Laplacian is defined as*

$$\mathcal{L} = I - D^{-1/2} A D^{-1/2}$$

where A is the adjacency matrix associated with G and H, and D is the degree matrix with $D_{ii} = \sum_j A_{ij}$.

1.2 Stochastic Block Model and HSBM

A Stochastic Block Model is a type of probabilistic graphical model that is used to generate random graphs based on a partitioning of the vertex set into blocks (or communities).

Definition 1.9. *(Stochastic Block Model, SBM)*
Let $G = (V, E)$ be a graph on n vertices, where $V = [n]$ is composed of K disjoint blocks, i.e., $V = \cup_{k=1}^{K} V_k$. The proportion of each block can be denoted by $\alpha_k = |V_k|/|V|$ and we define the vector $\alpha = (\alpha_1, \ldots, \alpha_K)$ with $\|\alpha\|_1 = 1$. Let $\sigma \in [K]^n$ denote the membership vector of the vertices, i.e., $\sigma(v) = k$ if the vertex v belongs to block V_k. Let each entry of the membership vector σ be sampled independently under α. Let $\mathcal{P} \in \mathbb{R}^{K \times K}$ be a symmetric matrix, and each possible edge $e = \{v_1, v_2\}$ is generated with probability $\mathbb{P}(\mathbb{1}_e = 1) = \mathcal{P}_{\sigma(v_1), \sigma(v_2)}$. We denote this distribution on the set of graphs by

$$(\sigma, G) \sim \mathrm{SBM}(n, \alpha, \mathcal{P}).$$

In this paper, we are particularly interested in binary m-uniform hypergraph stochastic block model.

Definition 1.10. *(Binary m-uniform HSBM).*
Let $\sigma = (\sigma_1, \ldots, \sigma_n) \in \{\pm 1\}^n$ denote the block assignment vector on vertex set $V = [n]$, where σ is chosen uniformly at random among all vectors satisfying $\mathbb{1}_n^\top \sigma = 0$ to ensure a binary balanced cluster distribution. Let $m \in \mathbb{N}$ be some fixed integer. The m-uniform hypergraph $H_m = ([n], E_m, p_m, q_m)$ is drawn in the following manner: each m-hyperedge $e := \{v_1, \ldots, v_m\} \subset [n]$ is sampled independently with probability p_m if $\sigma_{v_1} = \ldots = \sigma_{v_m}$, otherwise with probability q_m. We denote this distribution on the set of hypergraphs by

$$(\sigma, H_m) \sim \mathrm{HSBM}(n, m, p_m, q_m).$$

The goal of community detection is to recover σ by observing G or H, up to some level of accuracy. We next define different regimes of recovery for the community detection problem.

Definition 1.11. *(mismatch ratio)*
For an estimation σ^* of σ, we define the mismatch ratio

$$\eta_n := \eta(\boldsymbol{z}, \hat{\sigma}) = \frac{1}{n} \inf_{\pi \in \mathcal{S}_K} \sum_{i=1}^{n} \mathbb{1}\left(\sigma_i^* \neq \pi\left(\sigma_i\right)\right)$$

where $\mathcal{S}_K$ denotes the group of all permutations on $[K]$.

Based on the mismatch ratio, we have four types of recovery: exact recovery which requires the entire partition of the graph to be correctly recovered; almost exact recovery which allows for a vanishing number of misclassified vertices; partial recovery which allows for a constant fraction of misclassified vertices; and weak recovery which requires that the estimation is better than random guess.

Definition 1.12. *(Recovery)*
(1) Exact recovery (strong consistency): $\mathbb{P}\left(\eta_n = 0\right) \geq 1 - o(1)$.
(2) Almost exact recovery (weak consistency): $\mathbb{P}\left(\eta_n = o(1)\right) \geq 1 - o(1)$.
(3) Partial recovery: $\mathbb{P}\left(\eta_n \leq 1 - \gamma\right) \geq 1 - o(1)$ for $\gamma \in \left(\|\alpha\|_2^2, 1\right)$.
(4) Weak recovery (detection): $\mathbb{P}\left(\eta_n \leq 1 - \|\alpha\|_2^2 - \Omega(1)\right) \geq 1 - o(1)$.

The regime for exact recovery requires that the degree of each vertex grow logarithmically with the number of total vertices, or otherwise the exact recovery can not be achieved statistically because there will be isolated vertices in the graph, with high probability. In the rest of the paper, we assume the model suits the regime for exact recovery.

Suppose A is an adjacency matrix sampled from binary m-uniform hypergraph stochastic block model HSBM(n, m, p_m, q_m). Without loss of generality, we assume the first $n/2$ nodes form one community and the second half form the other community. Let $A^* = \mathbb{E}[A]$ be the expectation of A, and then we have

$$A^* := \mathbb{E}[A] = \left[\begin{array}{cc} p & q \\ q & p \end{array}\right] \otimes \left(\mathbb{1}_{n/2} \mathbb{1}_{n/2}^{\top}\right), \tag{1}$$

$$p := p_m \binom{n/2-2}{m-2} + \left(\binom{n-2}{m-2} - \binom{n/2-2}{m-2}\right) q_m, \tag{2}$$

$$q := q_m \binom{n-2}{m-2}, \tag{3}$$

where

$$p_m = \frac{a_m}{\binom{n-1}{m-1}} \log n, \quad q_m = \frac{b_m}{\binom{n-1}{m-1}} \log n, \quad a_m, b_m = \Omega(1).$$

Remark 1.13. *Recovery can be thought as finding a very close approximation to u_2^* since the sign of u_2^* exactly recovers the model.*

Notice that an m-uniform hypergraph can be accurately defined through a tensor $P \in \mathbb{R}^{k^m}$, where $P[i_1, i_2, \ldots, i_m] = 1$ if there is a hyperedge involving the $i_1, i_2, \ldots, i_m$ vertices, and 0 otherwise. Given an adjacency tensor, we can recover the full information about the hypergraph. But can we recover the HSBM if the information available is restricted to an aggregated matrix, like the adjacency matrix and combinatorial Laplacian? This is the question we aim to answer.

1.3 Recent Works

In this section, we introduce different lines of research on weak recovery, partial recovery, exact recovery, and the symmetric stochastic block model (SSBM). By definition, an algorithm can solve weak recovery (or detection) if its estimated community labels assigned to the vertices are correct significantly more often than what would be expected by random chance.

First note that for weak recovery, the vertex has constant expected degree or the problem will be trivially solved by applying the degree variations. [Dec+11] first conjectured the existence of the phase transition phenomenon for the weak recovery problem of $\mathrm{SSBM}(n, k, a/n, b/n)$ and the information-computation gap at 4 communities in the symmetric case. That is, weak recovery can be efficiently achieved for any $k \geq 2$ if and only if the Kesten-Stigum(KS) threshold

$$\mathrm{SNR} = \frac{(a-b)^2}{k(a+(k-1)b)} > 1,$$

and for $k \geq 4$, it is possible to solve weak recovery information-theoretically for some $\mathrm{SNR} < 1$. [Mas14] first showed that KS-threshold can be achieved efficiently when $k = 2$. The impossibility part of the conjecture for binary symmetric communities is proved in [MNS15], which shows by information-theoretic means that when $(a-b)^2 \leq 2(a+b)$, it is impossible to achieve weak recovery, because the SBM is indistinguishable from the Erdős-Rényi model with edge probability $(a+b)/(2n)$. While weak recovery allows for a non-trivial fraction of misclassified vertices, partial recovery allows for a constant fraction of misclassified vertices. In the symmetric $\mathrm{SSBM}(n, 2, a/n, b/n)$, the regime for partial recovery takes place when $\mathrm{SNR} = \frac{(a-b)^2}{2(a+b)} = O(1)$. One main goal is to identify the optimal tradeoff between the misclassification of vertices and SNR, where [YP14] and [CRV15] show that the upper bound on the fraction of incorrectly recovered vertices is of the form $C \exp(-c\mathrm{SNR})$ when SNR is large. Furthermore, [YP14] and [MNS14] showed that almost exact recovery is solvable in $\mathrm{SSBM}(n, 2, a_n/n, b_n/n)$ if and only if

$$\frac{(a_n - b_n)^2}{2\left(a_n + b_n\right)} = \omega(1).$$

In 2014, [MNS14] and [ABH15] found that exact recovery problem for binary symmetric communities also has a phase transition, $|\sqrt{a} - \sqrt{b}| > \sqrt{2}$, in the logarithmic regime, shown to be also efficiently achievable. For the hypergraph case, more recently, [DW23] identified the exact recovery threshold in the non-uniform case, and proposed a two-stage algorithm achieving exact recovery down to the information-theoretic threshold. For the binary symmetric non-uniform case, [Wan23] shows the impossibility part under the information-theoretic threshold, and [GJ23] shows that spectral algorithm on the adjacency matrix can achieve exact recovery under the min-bisection threshold, which is different from and above the threshold introduced in [Wan23].

1.4 Notations

Before diving into the proof, we first introduce some notations which will be used throughout this paper. For any real numbers $a, b \in \mathbb{R}$, we denote $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Let $\mathrm{sgn} : \mathbb{R} \to \{\pm 1, 0\}$ be the function defined by $\mathrm{sgn}(x) = 1$ if $x > 0$, $\mathrm{sgn}(x) = -1$ if $x < 0$, and $\mathrm{sgn}(x) = 0$ if $x = 0$. We also extend the definition to vectors.

For any vector $x \in \mathbb{C}^n$, we define infinity norm $\|x\|_\infty = \max_i |x_i|$ and 2-norm $\|x\| = \sqrt{\sum_{i=1}^n x_i^2}$. For any matrix $A \in \mathbb{C}^{n \times m}$, let $A_{i\cdot}$ and $A_{\cdot j}$ denote its i-th row and j-th column respectively. Let $\|A\| = \max_{\|x\|=1} \|Ax\|$ denote the spectral norm, $\|A\|_F := \sqrt{\sum_{i,j} |A_{ij}|^2}$ denote the Frobenius norm and $\|A\|_{2,\infty} = \max_{\|x\|=1} \|Ax\|_\infty = \max_i \|A_{i\cdot}\|$ denote the two-to-infinity norm. We denote its conjugate transpose by A^H and its Moore-Penrose inverse by A^+. We denote by $\mathbb{1}_n$ the $n \times 1$ vector with all entries being 1 and let $J_n = \mathbb{1}_n \mathbb{1}_n^\top$ be the $n \times n$ matrix of all ones. Let $d_{\min}$, $d_{\max}$ be the minimum and the maximum degree of the vertices correspondingly. Let A^*, D^*, L^* be the expected adjacency matrix, degree matrix, and combinatorial Laplacian, and u_2^* be the eigenvector corresponding to the second smallest eigenvalue of L^*. Furthermore, we use the Bachmann-Landau notation $o(\cdot), O(\cdot), \omega(\cdot), \Omega(\cdot), \Theta(\cdot)$ etc. throughout the paper.

2 Main Results

Algorithm 1 Spectral Clustering algorithm

```
Input: Adjacency matrix $A$
Compute the combinatorial Laplacian $L = D - A$.
Find the eigenvector $u_2$ corresponding to the second smallest eigenvalue $\lambda_2$ of $L$.
Obtain the partitioning based on $\text{sgn}(u_2)$.
```

The main goal of this paper is to show that the spectral clustering algorithm for the combinatorial Laplacian achieves strong consistency for the model $\text{HSBM}(n, m, p_m, q_m)$ with $p_m = \frac{a_m}{\binom{n-1}{m-1}} \log n, \quad q_m = \frac{b_m}{\binom{n-1}{m-1}} \log n$ when the min-bisection threshold for strong consistency is established:

$$I(m, a_m, b_m) = \max_{t \geq 0} \frac{1}{2^{m-1}} \left[a_m \left(1 - e^{-(m-1)t}\right) + b_m \sum_{r=1}^{m-1} \binom{m-1}{r} \left(1 - e^{-(m-1-2r)t}\right) \right] > 1.$$

Following a similar route as in [DLS21] where the result was established in the graph case, our proof contains two main parts. First, we show that the second eigenvalue of the combinatorial Laplacian, λ_2, is well "separated" from $\lambda_1 = 0$ and λ_3. This is to ensure the stability of the algorithm, based on the fact that the second eigenvector can be computed accurately. Second, we show that strong consistency can be achieved by approximating u_2 with a $\tilde{u}_2$ such that:

1. The entrywise error between u_2 and $\tilde{u}_2$ is negligible.
2. The entries of $\tilde{u}_2$ exactly recover $\text{HSBM}(n, m, p_m, q_m)$.

3 Eigenvalue Separation

In this section, we show that λ_2 is well "separated" from $\lambda_1 = 0$ and λ_3 by finding the lower and upper bound for λ_2 and a lower bound for λ_3. This section is to ensure that the eigenvector we computed actually corresponds to u_2^*.

3.1 Lower Bound for $\lambda_3(L)$

In this section, we would like to find a lower bound for $\lambda_3(L)$. Our goal is to prove Lemma 3.1, which shows that $\lambda_3(L)$ is bounded from below with probability based on a_m and b_m. With the help of Lemma 3.3, we can show that with high probability, $\lambda_3(L)$ is bounded from below when above the information-theoretic threshold.

Lemma 3.1. *(Lower bound for the third eigenvalue in the critical regime.)*
Suppose $p_m = \frac{a_m}{\binom{n-1}{m-1}} \log n, \quad q_m = \frac{b_m}{\binom{n-1}{m-1}} \log n$. *Then for any* $\xi > 0$ *and* $\epsilon > 0$, *there exists* $C = C(\xi, a_m, b_m, \epsilon) > 0$ *such that*

$$\lambda_3(L) \geq (m-1)\gamma \log n - (m-1)(\xi + \epsilon) \log n$$

where $\gamma = \frac{1}{2^{m-1}} a_m + (1 - \frac{1}{2^{m-1}}) b_m$, *with probability at least* $1 - Cn^{-f(\xi; a_m, b_m)}$.

We introduce the following lemmas to help us prove Lemma 3.1.

Lemma 3.2. *Let A be the adjacency matrix of* $\text{HSBM}(n, m, p_m, q_m)$ *where* $p_m = \frac{a_m}{\binom{n-1}{m-1}} \log n, \quad q_m = \frac{b_m}{\binom{n-1}{m-1}} \log n$. *Then for any* $0 < \xi < \gamma$, $\gamma = \frac{1}{2^{m-1}} a_m + (1 - \frac{1}{2^{m-1}}) b_m$, *we have*

$$\mathbb{P}\bigg(d_{min} \geq (m-1)\gamma \log n - (m-1)\xi \log n\bigg) \geq 1 - 2n^{-f(\xi; a_m, b_m)}$$

for n larger than a constant $N = N(a_m, b_m)$. Here

$$f(\xi; a_m, b_m) = \gamma(1 - \frac{\xi}{\gamma}) \log(1 - \frac{\xi}{\gamma}) + \xi - 1.$$

The function f characterizes a trade-off between the perturbation of $d_{\min}$ and its probability. Note when ξ is sufficiently close to 0, f will eventually be negative, then Lemma 3.2 loses its usefulness. However, when above the information-theoretic treshold for strong consistency, we can ensure that $d_{\min}$ is well controlled from below.

Lemma 3.3. *When above the information-theoretic threshold for strong consistency, that is, when $\mathbb{D}_{GH}^{(m)} = \frac{1}{2^{m-1}} \left(\sqrt{a_m} - \sqrt{b_m}\right)^2 > 1$, there exists $0 < \xi < \frac{1}{2^{m-1}} a_m + \left(1 - \frac{1}{2^{m-1}}\right) b_m$ such that $f(\xi; a_m, b_m) > 0$.*

Lemma 3.4. *(Weyl's inequality)*
Let $A, B \in \mathbb{R}^{m \times n}$ be two real $m \times n$ matrices, then

$$|\sigma_i(A + B) - \sigma_i(A)| \leq \|B\|$$

for every $1 \leq i \leq m \wedge n$. Furthermore, if $m = n$ and $A, B \in \mathbb{R}^{n \times n}$ are real symmetric, then $|\lambda_i(A + B) - \lambda_i(A)| \leq \|B\|$ for all $1 \leq i \leq n$.

Lemma 3.5. *[DW23] For each $2 \leq m \leq M$, let $H_m = ([n], E_m)$ be an inhomogeneous m-uniform Erdős-Rényi hypergraph associated with a probability tensor $\mathcal{Q}^{(m)}$ and an adjacency tensor $\mathcal{A}^{(m)}$ such that each m-hyperedge $e = \{i_1, i_2, \ldots, i_m\} \subset [n]$ appears with probability*

$$\mathbb{P}\left(\mathcal{A}_e^{(m)} = 1\right) = \mathcal{Q}_{i_1,\ldots,i_m}^{(m)} = \left[\binom{n-1}{m-1}\right]^{-1} d_{i_1,\ldots,i_m}^{(m)}.$$

Denote $d_{\max}^{(m)} := \max_{i_1,\ldots,i_m \in [n]} d_{i_1,\ldots,i_m}^{(m)}$. Let $H = \bigcup_{m=2}^{M} H_m$ be the inhomogeneous non-uniform Erdős-Rényi hypergraph and define $d_{\max} := \sum_{m \in \mathcal{M}} d_{\max}^{(m)}$. Suppose that

$$d_{max} := \sum_{m \in \mathcal{M}} d_{\max}^{(m)} \geq c \log n,$$

for some constant $c > 0$, then with probability at least $1 - 2n^{-10} - 2e^{-n}$, the adjacency matrix A of H satisfies

$$\|A - A^*\| \leq \mathrm{C} \cdot \sqrt{d_{\max}},$$

where constant $\mathrm{C}_: = 10M^2 + 2\beta$ with $\beta = \beta_0 \sqrt{\beta_1} + M$, and β_0, β_1 satisfying

$$\beta_0 = 16 + 32M\left(1 + e^2\right) + 1792\left(1 + e^{-2}\right) M^2, \quad M^{-1}\beta_1 \log\left(M^{-1}\beta_1\right) - M^{-1}\beta_1 + 1 > 11/c.$$

Proof of Lemma 3.1. By Weyl's inequality (Lemma 3.4) and Lemma 3.5, we have

$$\begin{aligned}
\lambda_3(L) &\geq \lambda_3(L^*) + \lambda_{\min}(L - L^*) \\
&\geq \lambda_3(L^*) + \lambda_{\min}(D - D^*) - \|A - A^*\| \\
&= \frac{n(p+q)}{2} + \lambda_{\min}(D) - \frac{n(p+q)}{2} - \|A - A^*\| \\
&= \lambda_{\min}(D) - \|A - A^*\| \\
&= d_{\min} - O(\sqrt{\log n})
\end{aligned}$$

where p, q are the values defined in equation (2) and (3).
Therefore, to bound $\lambda_3(L)$ from below, we need to determine a lower bound for $d_{\min}$.

By Lemma 3.2, for n large enough

$$\mathbb{P}\Bigg(d_{\min} \geq (m-1)\gamma\log n - (m-1)\xi\log n\Bigg) \geq 1 - 2n^{-f(\xi;a_m,b_m)}.$$

Then by Lemma 3.5,

$$\mathbb{P}\Bigg(\|A - A^*\| \leq C_1\sqrt{\log n}\Bigg) \geq 1 - 2n^{-10} - 2e^{-n}.$$

Therefore we have

$$\mathbb{P}\Bigg(\lambda_3(L) \geq (m-1)\gamma\log n - (m-1)(\xi+\epsilon)\log n\Bigg) \geq 1 - Cn^{-f(\xi;a_m,b_m)}.$$

□

In the remaining of this subsection, we prove Lemma 3.2 and Lemma 3.3. In Lemma 3.2, we use Poisson approximation to control $d_{\min}$, and Lemma 3.3 shows the feasibility of Lemma 3.2 under the critical regime.

Proof of Lemma 3.2. For m-uniform hypergraphs, let E be the set of all possible m-uniform edges,

$$\begin{aligned} d_i = \sum_{j=1}^{n} A_{ij} &= \sum_{j=1}^{n}\sum_{\substack{i,j\in e\\ e\in E}} \mathbb{1}_e = \sum_{\substack{i\in e\\ e\in E}}(m-1)\mathbb{1}_e \\ &= (m-1)\left(\sum_{\substack{i\in e\\ e\in E\\ \forall j\in e, \sigma(i)=\sigma(j)}} \mathbb{1}_e + \sum_{\substack{i\in e\\ e\in E\\ \exists j\in e, \sigma(i)\neq\sigma(j)}} \mathbb{1}_e\right). \end{aligned} \tag{4}$$

Notice that

$$\sum_{\substack{i\in e\\ e\in E\\ \forall j\in e, \sigma(i)=\sigma(j)}} \mathbb{1}_e \sim \text{Binomial}\left(\binom{n/2-1}{m-1}, p_m\right)$$

and

$$\sum_{\substack{i\in e\\ e\in E\\ \exists j\in e, \sigma(i)\neq\sigma(j)}} \mathbb{1}_e \sim \text{Binomial}\left(\binom{n-1}{m-1} - \binom{n/2-1}{m-1}, q_m\right).$$

So we can control the minimum degree in the critical regime by the following Poisson approximation to binomials.

Lemma 3.6. *Let $X \sim$ Binomial$(\binom{n/2-1}{m-1}, p_m)$ and $Y \sim$ Binomial$(\binom{n-1}{m-1} - \binom{n/2-1}{m-1}, q_m)$ for n even. Suppose $p_m = \frac{a_m}{\binom{n-1}{m-1}}\log n$ and $q_m = \frac{b_m}{\binom{n-1}{m-1}}\log n$ for constants a_m and b_m. Let*

$$\gamma = \frac{1}{2^{m-1}}a_m + (1 - \frac{1}{2^{m-1}})b_m,$$

then for every $k \leq \gamma\log n$,

$$\mathbb{P}(X + Y = k) \leq (1 + o(1))n^{-\gamma}\frac{(\gamma\log n)^k}{k!}.$$

Proof of Lemma 3.6. For $k \le \gamma \log n$,

$$\begin{aligned}\mathbb{P}(X=k) &= \binom{\binom{n/2-1}{m-1}}{k} p_m^k (1-p_m)^{\binom{n/2-1}{m-1}-k} \\ &\le (1+o(1)) \frac{\left(\frac{1}{2^{m-1}}\frac{n^{m-1}}{(m-1)!}\right)^k}{k!} \left(\frac{a_m(m-1)!\log n}{n^{m-1}}\right)^k \left(1-\frac{a_m(m-1)!\log n}{n^{m-1}}\right)^{\frac{1}{2^{m-1}}\frac{n^{m-1}}{(m-1)!}-k} \\ &\le (1+o(1)) n^{-\frac{a_m}{2^{m-1}}} \frac{\left(\frac{1}{2^{m-1}}a_m \log n\right)^k}{k!}\end{aligned}$$

where the last inequality is due to

$$\begin{aligned}\left(1-\frac{a_m(m-1)!\log n}{n^{m-1}}\right)^{\frac{1}{2^{m-1}}\frac{n^{m-1}}{(m-1)!}-k} &\le (1+o(1)) \exp\left(-\frac{a_m(m-1)!\log n}{n^{m-1}}\left(\frac{1}{2^{m-1}}\frac{n^{m-1}}{(m-1)!}-k\right)\right) \\ &= (1+o(1)) \exp\left(-\frac{a_m}{2^{m-1}}\log n\right) \\ &= (1+o(1)) n^{-\frac{a_m}{2^{m-1}}}.\end{aligned}$$

Similarly,

$$\mathbb{P}(Y=k) \le (1+o(1)) n^{-\left(1-\frac{1}{2^{m-1}}\right)b_m} \frac{\left((1-\frac{1}{2^{m-1}})b_m \log n\right)^k}{k!}. \tag{5}$$

Finally note that

$$\begin{aligned}\mathbb{P}(X+Y=k) &= \sum_{l=0}^{k} \mathbb{P}(X=l)\mathbb{P}(Y=k-l) \\ &\le (1+o(1)) n^{-\gamma} \frac{(\gamma \log n)^k}{k!}. \end{aligned}\tag{6}$$

□

Lemma 3.7.
(i) (Chernoff) Let $\{X_i\}_{i=1}^n$ *be independent variables. Assume* $0 \le X_i \le 1$ *for each* i*. Let* $X = X_1 + \cdots + X_n$ *and* $\mu = \mathbb{E}X$*. Then for any* $t > 0$*,*

$$\mathbb{P}\left(|X-\mu| \ge t\right) \le 2\exp\left(-\frac{t^2}{2\mu+t}\right).$$

As a result, for any $r > 0$*, there exists* $C = C(r) > 0$ *such that*

$$\mathbb{P}\left(|X-\mu| \ge C\left(\log n + \sqrt{\mu \log n}\right)\right) \le 2n^{-r}.$$

(ii) (Bennett) Let $X \sim$ *Poisson*(λ)*. Then for any* $0 < x < \lambda$*,*

$$\mathbb{P}\left(X \le \lambda - x\right) \le \exp\left(-\frac{x^2}{2\lambda}h\left(-\frac{x}{\lambda}\right)\right),$$

where $h(u) = 2u^{-2}\left((1+u)\log(1+u) - u\right)$*.*

With the help of Poisson approximation, we can now prove Lemma 3.2.

Let d_i be the degree of the ith node. Let X be a Poisson variable with mean $\gamma = \frac{1}{2^{m-1}}a_m + (1-\frac{1}{2^{m-1}})b_m$. Then by Lemma 3.6 and 3.7, for n large enough, we have

$$\begin{aligned}\mathbb{P}\Big(d_i \le (m-1)\gamma\log n - (m-1)\xi\log n\Big) &\le 2\mathbb{P}\Big(X \le \gamma\log n - \xi\log n\Big) \\ &\le 2n^{-f(\xi;a_m,b_m)-1}.\end{aligned}$$

Taking union bound yields

$$\mathbb{P}\bigg(d_{\min} \geq (m-1)\gamma \log n - (m-1)\xi \log n\bigg) \geq 1 - 2n^{-f(\xi;a_m,b_m)}$$

and this finishes the proof of Lemma 3.1. □

Now we continue to prove Lemma 3.3, which shows that when above the information-theoretic threshold, $\lambda_3(L)$ is bounded from below as in Lemma 3.1 with high probability.

Proof of Lemma 3.3. Let $\gamma = \frac{1}{2^{m-1}}a_m + (1 - \frac{1}{2^{m-1}})b_m$, then

$$f(\xi;\gamma) = \gamma(1 - \frac{\xi}{\gamma})\log(1 - \frac{\xi}{\gamma}) + \xi - 1.$$

Note that

$$\frac{\partial f}{\partial \xi} = -\log(1 - \frac{\xi}{\gamma}) > 0$$

when $\xi < \gamma$. We choose $\xi^* = \frac{1}{2^{m-1}}(a_m - b_m)$, it remains to show that $f(\xi^*;\gamma) > 0$ when above the information-theoretic threshold. From $\mathbb{D}_{GH}^{(m)} = \frac{1}{2^{m-1}}\left(\sqrt{a_m} - \sqrt{b_m}\right)^2 > 1$, we have

$$a_m + b_m > 2\sqrt{a_m b_m} + 2^{m-1}$$

Note that

$$\begin{aligned} f(\xi^*;\gamma) &= \gamma(1 - \frac{\xi^*}{\gamma})\log(1 - \frac{\xi^*}{\gamma}) + \xi^* - 1 \\ &\geq b_m \left(\log\left(\frac{b_m}{\frac{1}{2^{m-1}}a_m + (1 - \frac{1}{2^{m-1}})b_m}\right) + \frac{1}{2^{m-2}}\sqrt{\frac{a_m}{b_m}} - \frac{1}{2^{m-2}}\right) \\ &= b_m \left(\frac{1}{2^{m-2}}\sqrt{\frac{a_m}{b_m}} - \log\left(\frac{a_m}{2^{m-1}b_m} + 1 - \frac{1}{2^{m-1}}\right) - \frac{1}{2^{m-2}}\right). \end{aligned}$$

Let $x = \frac{a_m}{b_m}$, by differentiation, we have

$$\frac{1}{2^{m-2}}\sqrt{x} - \log(\frac{x}{2^{m-1}} + 1 - \frac{1}{2^{m-1}}) - \frac{1}{2^{m-2}} > 0$$

when $x > 1$. This completes the proof of Lemma 3.3. □

3.2 Upper Bound for $\lambda_2(L)$

In this section, we aim to find an upper bound for $\lambda_2(L)$. By min-max principle, we can bound $\lambda_2(L)$ from above using d_{out} and a small error $\frac{2}{n}\langle d_{out} - d_{out}^*, 1_n\rangle$ where $(d_{out})_i = \sum_{\sigma(i)\neq\sigma(j)} A_{ij}$. We further show that $d_{\text{out}} = O(\log n)$ and the error is negligible compared to d_{out} .

Lemma 3.8. *(Upper bound for the second eigenvalue in the critical regime.)*
Suppose $p_m = \frac{a_m}{\binom{n-1}{m-1}}\log n, \quad q_m = \frac{b_m}{\binom{n-1}{m-1}}\log n$ *where* $I_{GH}^{(m)} = \frac{1}{2^{m-1}}(\sqrt{a_m} - \sqrt{b_m})^2 > 1$. *Then*

$$\lambda_2(L) \leq (m-1)b_m \log n + O(\log n/n).$$

We introduce the following Lemmas to help us prove Lemma 3.8.

Lemma 3.9. *Let* A *be an instance of the m-uniform hypergraph* $\mathcal{G}(n, m, p_m, q_m)$. *We define* $d_{out} \in \mathbb{R}^n$ *to be the vector with the ith element being* $(d_{out})_i = \sum_{\sigma(i)\neq\sigma(j)} A_{ij}$, *and define* $d_{out}^* = \mathbb{E}(d_{out})$. *Then*

$$\lambda_2(L) \leq (m-1)b_m \log n + \frac{2}{n}\langle d_{out} - d_{out}^*, 1_n\rangle.$$

Lemma 3.10. *If $b_m = \Omega(1)$, then for any $r > 0$ there exists $C = C(b_m, r, m)$ such that*

$$|\langle d_{out} - d^*_{out}, 1_n \rangle| = O(\log n).$$

Proof of Lemma 3.8. By Lemmas 3.9 and 3.10, we can easily see that Lemma 3.8 holds. □

Proof of Lemma 3.9. By the min-max principle

$$\begin{aligned}
\lambda_2(L) &= \min_{V \in \mathcal{V}_t} \max_{x \in V \setminus \{0\}} \frac{\langle x, Lx \rangle}{\langle x, x \rangle} \\
&\leq \max_{\substack{x \in \text{span}\{1_n, u^*_2\} \\ \|x\|=1}} \langle x, Lx \rangle \\
&= \langle u^*_2, Lu^*_2 \rangle \qquad (7) \\
&= \langle u^*_2, (D - A)u^*_2 \rangle \\
&= \frac{2}{n} \left(2 \sum_{i=1}^{\frac{n}{2}} \sum_{j=\frac{n}{2}+1}^{n} A_{ij} \right) \\
&= \frac{2}{n} \langle d_{out}, 1_n \rangle \\
&= \frac{2}{n} d^*_{out} + \frac{2}{n} \langle d_{out} - d^*_{out}, 1_n \rangle \\
&= (m-1) b_m \log n + \frac{2}{n} \langle d_{out} - d^*_{out}, 1_n \rangle \qquad (8)
\end{aligned}$$

(7) is due to $L1_n = 0$ and $1_n \perp u^*_2$,
(8) is due to

$$\begin{aligned}
d^*_{out} &= 2 \sum_{i=1}^{\frac{n}{2}} \sum_{j=\frac{n}{2}+1}^{n} \mathbb{E}(A_{ij}) \\
&= \frac{n^2}{2} \binom{n-2}{m-2} \frac{b_m}{\binom{n-1}{m-1}} \log n \\
&= \frac{n}{2} (m-1) b_m \log n
\end{aligned}$$

□

Proof of Lemma 3.10. Note that

$$\langle d_{out} - d^*_{out}, 1_n \rangle = 2 \sum_{i=1}^{\frac{n}{2}} \sum_{j=\frac{n}{2}+1}^{n} (A_{ij} - \mathbb{E}(A_{ij}))$$

$$2 \sum_{i=1}^{\frac{n}{2}} \sum_{j=\frac{n}{2}+1}^{n} A_{ij} = \sum_{r=1}^{m-1} \sum_{\substack{e \text{ is like} \\ (r, m-r)}} \mathbb{1}_e r(m-r).$$

Let

$$X_r = \sum_{\substack{e \text{ is like} \\ (r, m-r)}} \mathbb{1}_e r(m-r) = r(m-r) \sum_{\substack{e \text{ is like} \\ (r, m-r)}} \mathbb{1}_e$$

Notice that the above $\mathbb{1}_e$s are independent random variables, and $\mathbb{E}(\mathbb{1}_e) = \frac{b_m \log n}{\binom{n-1}{m-1}}$. By Chernoff inequality, we have

$$|X_r - \mu| = O\left(\log n + \sqrt{\frac{(\log n)^2}{n^{m-1}}} \right) = O(\log n).$$

Since m is a finite number, by taking a finite weighted sum, we get

$$\langle d_{out} - d_{out}^*, 1_n \rangle = O(\log n).$$

□

3.3 Lower Bound for $\lambda_2(L)$

In this section, we prove Lemma 3.11 which shows that $\lambda_2(L)$ is well controlled away from $\lambda_1 = 0$.

Lemma 3.11. *(Lower bound for the second eigenvalue in the critical regime.)*

$$\lambda_2(L) \geq (m-1)b_m \log n - O\left(\sqrt{\frac{\log n}{n}}\right)$$

We introduce the following lemmas to help us prove Lemma 3.11.

Lemma 3.12. *Let A be an instance of the m-uniform hypergraph $\mathcal{G}(n, m, p_m, q_m)$. We define $d_{out} \in \mathbb{R}^n$ to be the vector with the ith element being $(d_{out})_i = \sum_{\sigma(i) \neq \sigma(j)} A_{ij}$, and define $d_{out}^* = \mathbb{E}(d_{out})$. Let $d_{max} = \max_i d_i$, then*

$$P\left(\|d_{out} - d_{out}^*\| \geq C\sqrt{n d_{max}}\right) \leq 1 - 2n^{-10} - 2e^{-n}.$$

The generalized Davis-Kahan theorem provides bounds on the distance between the eigenspaces of two matrices based on the difference between the matrices themselves. To give the generalized Davis-Kahan theorem, we consider the following setup.

Consider the generalized eigenvalue problem $Mu = \lambda N u$ where M is Hermitian and N is Hermitian positive definite. It has the same eigenpairs as the problem $N^{-1}Mu = \lambda u$. Consider splitting the spectrum into two parts Λ_1, Λ_2. Let X be the matrix that has the eigenvectors of $N^{-1}M$ as columns. Then $N^{-1}M$ is diagonalizable and can be written as

$$N^{-1}M = X\Lambda X^{-1} = X_1 \Lambda_1 Y_1^H + X_2 \Lambda_2 Y_2^H$$

where

$$X^{-1} = \begin{pmatrix} X_1 & X_2 \end{pmatrix}^{-1} = \begin{pmatrix} Y_1^H \\ Y_2^H \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \Lambda_1 & \\ & \Lambda_2 \end{pmatrix}.$$

Lemma 3.13. *[DLS21] (Generalized Davis-Kahan theorem).*

Suppose $\delta = \min_i \left|(\Lambda_2)_{i,i} - \hat{\lambda}\right|$ is the absolute separation of $\hat{\lambda}$ from Λ_2, then for any vector $\hat{u}$ we have

$$\|P\hat{u}\| \leq \frac{\sqrt{\kappa(N)} \left\|\left(N^{-1}M - \hat{\lambda} I\right)\hat{u}\right\|}{\delta}$$

where $P = \left(Y_2^+\right)^H Y_2^H = I - \left(X_1^+\right)^H X_1^H$ is the orthogonal projection matrix onto the orthogonal complement of the column space of $X_1, \kappa(N) = \|N\| \cdot \left\|N^{-1}\right\|$ is the condition number of N and $Y_2^+$is the Moore-Penrose inverse of Y_2. Additionally, When $N = I$ and $(\hat{\lambda}, \hat{u})$ is the eigenpair of a matrix $\hat{M}$, we have

$$\sin\theta \leq \frac{\|(M - \hat{M})\hat{u}\|}{\delta},$$

where θ is the canonical angle between $\hat{u}$ and the column space of X_1.

Lemma 3.14. *Let u_2 be the eigenvector of L that corresponds to $\lambda_2(L)$. Then*

$$P\left(\|u_2 - u_2^*\| \leq C_1 \frac{1}{\sqrt{\log n}}\right) \geq 1 - C_2 n^{-f(\xi; a_m, b_m)}.$$

Assuming the 3 Lemmas above hold, we now prove Lemma 3.11.

Proof of Lemma 3.11. Let u_2 be the eigenvector of L that corresponds to $\lambda_2(L)$, we have

$$\begin{aligned}\lambda_2(L) = \langle u_2, Lu_2\rangle &= \langle (u_2 - u_2^*) + u_2^*, L\left((u_2 - u_2^*) + u_2^*\right)\rangle \\ &= \langle u_2^*, Lu_2^*\rangle + 2\left\langle u_2 - u_2^*, Lu_2^*\right\rangle + \langle u_2 - u_2^*, L\left(u_2 - u_2^*\right)\rangle \\ &\geq \langle u_2^*, Lu_2^*\rangle + 2\left\langle u_2 - u_2^*, Lu_2^*\right\rangle \\ &\geq (m-1)b_m \log n + \frac{2}{n}\left\langle d_{\text{out}} - d_{\text{out}}^*, \mathbb{1}_n\right\rangle - 2\left\|u_2 - u_2^*\right\|\left\|Lu_2^*\right\| \\ &= (m-1)b_m \log n + \frac{2}{n}\left\langle d_{\text{out}} - d_{\text{out}}^*, \mathbb{1}_n\right\rangle - \frac{4}{\sqrt{n}}\left\|u_2 - u_2^*\right\|\left\|d_{\text{out}}\right\|.\end{aligned}$$

By Lemma 3.10, we have $\langle d_{\text{out}} - d_{\text{out}}^*, \mathbb{1}_n\rangle = O(\log n)$. By Lemma 3.12, we have $\|d_{out} - d_{out}^*\| = O(\sqrt{n \log n})$. By Lemma 3.14, we have $\|u_2 - u_2^*\| = O\left(\frac{1}{\sqrt{\log n}}\right)$. Therefore, we have

$$\lambda_2(L) \geq (m-1)b_m \log n - O\left(\sqrt{\frac{\log n}{n}}\right).$$

□

Now, we return to prove the Lemmas 3.12 and 3.14.

Proof of Lemma 3.12. Let A_{out} denote the matrix after removing all A_{ij} where $\sigma(i) = \sigma(j)$ from A.

$$\begin{aligned}P\left(\|d_{out} - d_{out}^*\| \geq C\sqrt{nd_{max}}\right) &= P\left(\|(A_{out} - A_{out}^*)\mathbb{1}_n\| \geq C\sqrt{nd_{max}}\right) \\ &\leq P\left(\|A_{out} - A_{out}^*\| \geq C\sqrt{d_{max}}\right) \\ &\leq 1 - 2n^{-10} - 2e^{-n}.\end{aligned}$$

Step 2 is due to the definition of spectral norm and step 3 is derived from Lemma 3.5. Since $d_{max} = O(\log n)$, we have $\|d_{out} - d_{out}^*\| = O\left(\sqrt{n \log n}\right)$. □

Proof of Lemma 3.14. Let θ be the angle between u_2 and u_2^*. Assume $\theta \in [0, \pi/2]$, because otherwise just let $u_2 := -u_2$. Let $N = I$, $M = L$, $\hat{u} = u_2^*$, $\hat{\lambda} = \lambda_2\left(L^*\right)$, $X_1 = \left[\begin{array}{cc}\frac{1}{\sqrt{n}}\mathbb{1}_n & u_2\end{array}\right]$ and P be the projection matrix onto the orthogonal complement of X_1 in Lemma 3.13, we get

$$\|Pu_2^*\| = \sin(\theta) \leq \frac{\left\|\left(L - L^*\right)u_2^*\right\|}{\delta} = \frac{2\left\|d_{\text{out}} - d_{\text{out}}^*\right\|}{\delta\sqrt{n}},$$

where $\delta = \lambda_3(L) - \lambda_2\left(L^*\right)$ which we for now assume to be positive. Therefore

$$\|u_2 - u_2^*\| = \sqrt{2 - 2\cos(\theta)} \leq \sqrt{2}\sin(\theta) \leq \frac{2\sqrt{2}\left\|d_{\text{out}} - d_{\text{out}}^*\right\|}{\delta\sqrt{n}}.$$

It remains to find a lower bound for δ. Note that by Lemma 3.1 and Lemma 3.3, if $p_m = \frac{a_m}{\binom{n-1}{m-1}}\log n$, $\quad q_m = \frac{b_m}{\binom{n-1}{m-1}}\log n$, and $\mathbb{D}_{GH}^{(m)} = \frac{1}{2^{m-1}}(\sqrt{a_m} - \sqrt{b_m})^2 > 1$, we have

$$\mathbb{P}\Bigg(\lambda_3(L) \geq (m-1)b_m \log n + (m-1)\epsilon \log n\Bigg) \geq 1 - C_1 n^{-f(\xi; a_m, b_m)}.$$

Also, we have

$$\lambda_2(L^*) = nq = nq_m\binom{n-2}{m-2} = (1 + o(1))(m-1)b_m \log n.$$

Therefore

$$P\big(\delta \geq (m-1)\epsilon \log n\big) \geq 1 - C_1 n^{-f(\xi;a_m,b_m)}.$$

By Lemma 3.12

$$P\left(\|u_2 - u_2^*\| \leq C_2 \frac{1}{\sqrt{\log n}}\right) \geq 1 - C_3 n^{-f(\xi;a_m,b_m)}.$$

□

4 Strong Consistency

In this section, we show that the sign of the second eigenvector of the combinatorial Laplacian exactly recovers the partition of HSBM (n, m, p_m, q_m) under the min-bisection threshold. Notice that the labels of the partitions are binary, and the only thing matters is the partition while the labels can be used interchangeably. Therefore, in this section, any statement involving eigenvectors are up to sign, meaning that for any eigenvector u, either u or $-u$ will suit the statement. For example, the expression $\|u - v\|$ should be understood as $\min_{s\in\{\pm1\}} \|su - v\|$.

4.1 Sketch of Proof

The main goal of this section is to prove Lemma 4.1, which shows that the second eigenvector of the combinatorial Laplacian can be used to achieve strong consistency down to the min-bisection threshold $I(m, a_m, b_m) > 1$.

Lemma 4.1. *Let u_2 be the second eigenvector of the combinatorial Laplacian L, $p_m = \frac{a_m}{\binom{n-1}{m-1}} \log n$, $q_m = \frac{b_m}{\binom{n-1}{m-1}} \log n$, and $I(m, a_m, b_m) > 1$. Then there exists $\eta > 0$ and $s \in \{\pm 1\}$ such that with probability $1 - o(1)$,*

$$\sqrt{n}\,(su_2)_i \geq \eta \text{ for } i \leq \frac{n}{2}$$

and

$$\sqrt{n}\,(su_2)_i \leq -\eta \text{ for } i \geq \frac{n}{2} + 1.$$

The Lemma can be derived from the following two statements: we approximate u_2 with $(D - \lambda_2(L)I)^{-1}Au_2^*$ and show that with probability $1 - o(1)$
(i) $\left\|u_2 - (D - \lambda_2 I)^{-1} Au_2^*\right\|_\infty = o(1/\sqrt{n})$;
(ii) $\operatorname{sgn}\left((D - \lambda_2(L)I)^{-1} Au_2^*\right)$ exactly recovers the communities and

$$\left|\left((D - \lambda_2(L)I)^{-1} Au_2^*\right)_i\right| \geq \frac{\eta}{\sqrt{n}}$$

for all i and some $\eta > 0$.

We first look at (ii): Notice that $d_{max} = O(\log n)$ and $\lambda_2(L) = O(\log n)$ from Lemma 3.8, thus we have $d_{max} - \lambda_2(L) = O(\log n)$. It remains to show that $\|Au_2^*\|_\infty = \Omega\left(\frac{\log n}{\sqrt{n}}\right)$. In this step we use a Lemma from [GJ23] to help us prove the lower bound.

Lemma 4.2. *[GJ23] Let $m \in \{2, 3, \ldots\}$, and $a_m > b_m > 0$, such that $I(m, a_m, b_m) > 1$. Let A be the adjacency matrix of G where $G \sim$ HSBM (n, m, a_m, b_m). Then there exists a constant $\epsilon := \epsilon(m, a_m, b_m) > 0$ such that for any fixed $i \in [n]$, with probability at least $1 - o\left(n^{-1}\right)$,*

$$\sum_{j\in[n]} A_{ij}\sigma^*(i)\sigma^*(j) \geq \epsilon \log n.$$

Notice the i^{th} row of $|Au_2^*|$ is simply $\frac{1}{\sqrt{n}} \sum_{j\in[n]} A_{ij}\sigma^*(i)\sigma^*(j)$. Therefore, by Lemma 4.2, we have $\|Au_2^*\|_\infty = \Omega\left(\frac{\log n}{\sqrt{n}}\right)$, and thus complete the proof of (ii).

For (i), by definition of u_2, we have $u_2 = (D - \lambda_2(L)I)^{-1}Au_2$. Plug in (i), we get

$$\|u_2 - (D - \lambda_2(L)I)^{-1}Au_2^*\|_\infty = \|(D - \lambda_2(L)I)^{-1}A(u_2 - u_2^*)\|_\infty.$$

From Lemma 3.2 and Lemma 3.8, we have $\|\left(D - \lambda_2 I\right)^{-1}\|_\infty = O\left(\frac{1}{\log n}\right)$. It remains to show that

$$\|A(u_2 - u_2^*)\|_\infty = o\left(\frac{\log n}{\sqrt{n}}\right).$$

We use the following important row-concentration property to help us prove the upper bound.

Lemma 4.3. *[GJ23] (Row-Concentration Property of the adjacency matrix)*
Let $v \in \mathbb{R}^n$ be a fixed vector. Suppose $p \geq \max_i p_i$ and constant $c_0 > 0$. Then

$$\mathbb{P}\left(|(A - A^*)_{k\cdot}\, v| \leq \|v\|_\infty \frac{2 + 8m/c_0}{1 \vee \log \frac{\sqrt{n}\|v\|_\infty}{\|v\|}} n \binom{n-2}{m-2} p\right) \geq 1 - O\left(\frac{1}{n^4}\right).$$

Lemma 4.3 is probabilistic, and it requires independence between A and v, while A and u_2 are actually dependent from each other. In this step, we use the powerful leave-one-out technique ([ZB18], [Abb+20], [DLS21]) to help us construct independence. The idea is we construct a matrix $A^{(k)}$ where its k^{th} row and column is replaced by A^* and keep other elements unchanged. Denote $L^{(k)}$ as the corresponding Laplacian and u_2^k as the second eigenvector of $L^{(k)}$. In this case, the k^{th} row of A^*, denoted by $A^*_{k\cdot}$, is now independent of $(u_2^{(k)} - u_2^*)$. Then by triangle inequality, for the k^{th} entry of $A(u_2 - u_2^*)$, we have

$$|A_{k\cdot}(u_2 - u_2^*)| \leq |A_{k\cdot}(u_2 - u_2^{(k)})| + |A_{k\cdot}(u_2^{(k)} - u_2^*)|.$$

$|A_{k\cdot}(u_2 - u_2^{(k)})|$ can be bounded with the generalized Davis-Kahan theorem (Lemma 3.13) and $|A_{k\cdot}(u_2^{(k)} - u_2^*)|$ can be bounded with the row-concentration property (Lemma 4.3).

4.2 Proof

Let $A^{(k)}$ be the matrix that $A^{(k)}_{ij} = A^*_{ij}$ when i or j equals k and otherwise $A^{(k)}_{ij} = A_{ij}$. Let $D^{(k)}, L^{(k)}$ be the corresponding degree matrix and unnormalized Laplacian matrix of $A^{(k)}$. Let u_2 be the eigenvector of L that corresponds to the second smallest eigenvalue $\lambda_2(L)$. Let $u_2^{(k)}$ be the eigenvector of $L^{(k)}$ that corresponds to the second smallest eigenvalue $\lambda_2\left(L^{(k)}\right)$.

Lemma 4.4. *There exists $\xi = \xi(a_m, b_m) > 0, C_1, C_2 > 0$ depending on a_m, b_m and ξ, such that $f(\xi; a_m, b_m) > 0$ and*

$$\mathbb{P}\left(\max_{1\leq k\leq n}\left\|u_2 - u_2^{(k)}\right\| \leq C_1\left\|u_2\right\|_\infty\right) \geq 1 - \left(C_2(a_m, b_m, \xi)n^{-f(\xi;a_m,b_m)} \wedge (2n^{-10} - 2e^{-n})\right).$$

Proof. In Lemma 3.13 we let $M = L^{(k)}, N = I, \hat{u} = u_2, \hat{\lambda} = \lambda_2(L), X_1 = \begin{bmatrix} \frac{1}{\sqrt{n}}\mathbb{1}_n & u_2 \end{bmatrix}$. Then up to sign of eigenvectors,

$$\left\|u_2 - u_2^{(k)}\right\| \leq \frac{\sqrt{2}\left\|\left(L^{(k)} - L\right)u_2\right\|}{\delta_k}$$

where $\delta_k = \lambda_3\left(L^{(k)}\right) - \lambda_2(L)$. We first use Weyl's theorem (Lemma 3.4) to bound $\lambda_3\left(L^{(k)}\right)$ from below. The proof is similar to Lemma 3.1. We note that by the construction of $A^{(k)}$, the (k,k)-entry of $\left(D^{(k)} - D^*\right)$ is 0 and the (i,i)-entry $(i \neq k)$ only differs from $(d_i - d_i^*)$ by at most $O(\sqrt{\log n})$ with probability at least $1 - 2n^{-10} - 2e^{-n}$ by Lemma 3.5. Thus by Lemma 3.4, Lemma 3.1, Lemma 3.3, and the union bound, there exists $0 < \xi < \frac{1}{2^{m-1}}(a_m - b_m)$ such that $f(\xi; a_m, b_m) > 0$ and

$$\begin{aligned}
\min_{1\le k\le n} \lambda_3\left(L^{(k)}\right) &\ge \min_{1\le k\le n}\left(\lambda_3(L^*) + \lambda_{\min}(L^{(k)} - L^*)\right) \\
&\ge \lambda_3\left(L^*\right) + \min_{1\le k\le n}\left\{\lambda_{\min}\left(D^{(k)} - D^*\right) - \left\|A^{(k)} - A^*\right\|\right\} \\
&\ge \lambda_3\left(L^*\right) + \min\left\{\lambda_{\min}\left(D - D^*\right) - O(\sqrt{\log n}), 0\right\} - \max_{1\le k\le n}\left\|A^{(k)} - A^*\right\| \\
&= \min\left\{d_{\min} - O(\sqrt{\log n}), \frac{(p+q)n}{2}\right\} - \max_{1\le m\le n}\left\|A^{(k)} - A^*\right\| \\
&\ge (m-1)b_m \log n + \epsilon_1(a_m, b_m, \xi)\log n
\end{aligned}$$

with probability at least $1 - \left(C_1(a_m, b_m, \xi)n^{-f(\xi; a_m, b_m)} \wedge (2n^{-10} - 2e^{-n})\right)$. Using this in conjunction with Lemma 3.1, we have

$$\mathbb{P}\left(\min_{1\le k\le n}\delta_k \ge \epsilon_2(a_m, b_m, \xi)\log n\right) \ge 1 - \left(C_1(a_m, b_m, \xi)n^{-f(\xi; a_m, b_m)} \wedge (2n^{-10} - 2e^{-n})\right).$$

To bound the numerator, we consider bounding the k^{th} entry of $\left(L^{(k)} - L\right)u_2$ and the other entries separately. Let $v = \left(L^{(k)} - L\right)u_2$ then

$$|v_k| = \left|\left(L^{(k)} - L\right)_{k\cdot} u_2\right| = |(L^* - L)_{k\cdot}\, u_2| \le ||L^* - L||_\infty ||u_2||_\infty.$$

For $i \ne k$,

$$\begin{aligned}
\left(\sum_{i\ne k} v_i^2\right)^{1/2} &= \left(\sum_{i\ne k}\left(A^*_{ik} - A_{ik}\right)^2\left(u_2^{(k)} - u_2^{(i)}\right)^2\right)^{1/2} \\
&\le 2\left\|u_2\right\|_\infty \left(\sum_{i\ne k}\left(A^*_{ik} - A_{ik}\right)^2\right)^{1/2} \\
&\le 2\left\|u_2\right\|_\infty \left\|A^* - A\right\|.
\end{aligned}$$

Therefore by Lemma 3.5,

$$\max_{1\le k\le n}\left\|\left(L^{(k)} - L\right)u_2\right\| \le \left(\left\|L^* - L\right\|_\infty + 2\left\|A - A^*\right\|\right)\left\|u_2\right\|_\infty = O(\log n \left\|u_2\right\|_\infty)$$

with probability at least $1 - \left(C_1(a_m, b_m, \xi)n^{-f(\xi; a_m, b_m)} \wedge (2n^{-10} - 2e^{-n})\right)$. This concludes the proof. □

Lemma 4.5. *For $u_2^{(k)}$ defined as above*

$$\max_{1\le k\le n}\left|(A - A^*)_{k\cdot}\left(u_2^{(k)} - u_2^*\right)\right| = O\left(\frac{\log n}{\log\log n}\left\|u_2\right\|_\infty\right).$$

Proof. Note that $u_2^{(k)}$ is computed without using $A_{k\cdot}$, thus $(A - A^*)_{k\cdot}$ and $u_2^{(k)} - u_2^*$ are independent. We directly apply the row-concentration property (Lemma 4.3), letting $v = u_2^{(k)} - u_2^*$ and $\varphi(t) = (1 \vee \log(1/t))^{-1}$, for $t > 0$

$$\max_{1\le k\le n}\left|(A - A^*)_{k\cdot}\left(u_2^{(k)} - u_2^*\right)\right| = O\left(\max_{1\le k\le n}\|v\|_\infty \varphi\left(\frac{\|v\|}{\sqrt{n}\|v\|_\infty}\right)\log n\right).$$

Notice that $\varphi(t)$ is non-decreasing, $\varphi(t)/t$ is non-increasing and $\lim_{t\to 0}\varphi(t) = 0$. We set $x = \sqrt{n}\|v\|_\infty, y = \|v\|$, $\gamma = 1/\sqrt{\log n}$ and

$$(*) = \|v\|_\infty \varphi\left(\frac{\|v\|}{\sqrt{n}\|v\|_\infty}\right)\log n.$$

When $y/x \geq \gamma$ we have

$$(*) = \frac{\log n}{\sqrt{n}} \cdot y \cdot \frac{x}{y} \varphi\left(\frac{y}{x}\right) \leq \frac{\log n}{\sqrt{n}} \cdot \frac{y}{\gamma} \varphi(\gamma).$$

When $y/x \leq \gamma$ we have

$$(*) = \frac{\log n}{\sqrt{n}} \cdot x\varphi\left(\frac{y}{x}\right) \leq \frac{\log n}{\sqrt{n}} \cdot x\varphi(\gamma).$$

Thus for any $x, y > 0$ we always have

$$(*) \leq \frac{\log n}{\sqrt{n}} \cdot \left(x\varphi(\gamma) + \frac{y}{\gamma}\varphi(\gamma)\right).$$

Lemma 3.14 and Lemma 4.4 give

$$\begin{aligned}
\max_{1\leq k\leq n} x &= \sqrt{n} \max_{1\leq k\leq n} \left\| u_2^{(k)} - u_2^* \right\|_\infty \\
&\leq \sqrt{n} \left(\max_{1\leq k\leq n} \left\| u_2^{(k)} - u_2 \right\| + \|u_2\|_\infty + \|u_2^*\|_\infty \right) \\
&= \sqrt{n} \cdot O\left(\|u_2\|_\infty\right)
\end{aligned}$$

and

$$\max_{1\leq k\leq n} y = \max_{1\leq k\leq n} \left\| u_2^{(k)} - u_2^* \right\| \leq \max_{1\leq k\leq n} \left\| u_2^{(k)} - u_2 \right\| + \|u_2 - u_2^*\| = O\left(\|u_2\|_\infty + \frac{1}{\sqrt{\log n}} \right).$$

Therefore

$$\begin{aligned}
\max_{1\leq k\leq n} \left| (A - A^*)_{k\cdot} \left(u_2^{(k)} - u_2^* \right) \right| &= \frac{\log n}{\sqrt{n}} O\left(\max_{1\leq k\leq n} \left\{ x\varphi(\gamma) + \frac{y}{\gamma}\varphi(\gamma) \right\} \right) \\
&= \frac{\log n}{\sqrt{n}} O\left(\sqrt{n}\, \|u_2\|_\infty \varphi(\gamma) + \frac{\|u_2\|_\infty}{\gamma}\varphi(\gamma) + \varphi(\gamma) \right) \\
&= \frac{\log n}{\sqrt{n}} O\left(\frac{\sqrt{n}}{\log\log n} \|u_2\|_\infty + \frac{\sqrt{\log n}}{\log\log n} \|u_2\|_\infty + \frac{1}{\log\log n} \right) \\
&= O\left(\frac{\log n}{\log\log n} \|u_2\|_\infty \right).
\end{aligned}$$

□

Lemma 4.6. *There exist C_1, $C_2 > 0$ depending on a_m, b_m, ξ such that*

$$\mathbb{P}\left(\|A\left(u_2 - u_2^*\right)\|_\infty \leq C_1 \frac{\log n}{\sqrt{n}\log\log n} \right) \geq 1 - C_2 n^{-f(\xi; a_m, b_m)}.$$

Proof.

$$\begin{aligned}
\|A\left(u_2 - u_2^*\right)\|_\infty &= \max_{1\leq k\leq n} \left| A_{k\cdot}\left(u_2 - u_2^*\right) \right| \\
&\leq \max_{1\leq k\leq n} \left| A_{k\cdot} \left(u_2 - u_2^{(k)} \right) \right| + \max_{1\leq k\leq n} \left| A_{k\cdot} \left(u_2^{(k)} - u_2^* \right) \right| \\
&\leq \max_{1\leq k\leq n} \|A\|_{2,\infty} || u_2 - u_2^{(k)} \| + \max_{1\leq k\leq n} \left| A_{k\cdot}^* \left(u_2^{(k)} - u_2^* \right) \right| \qquad (9) \\
&\quad + \max_{1\leq k\leq n} \left| (A - A^*)_{k\cdot} \left(u_2^{(k)} - u_2^* \right) \right|
\end{aligned}$$

For the first term of (9), by Lemma 3.5 and 4.4 we have

$$\begin{aligned}\max_{1\le k\le n}\|A\|_{2,\infty}\|u_2-u_2^{(k)}\| &= \|A\|_{2,\infty}\max_{1\le k\le n}\|u_2-u_2^{(k)}\|\\ &\le \left(\|A^*\|_{2,\infty}+\|A-A^*\|\right)\max_{1\le k\le n}\|u_2-u_2^{(k)}\|\\ &= O(\sqrt{\log n}\|u_2\|_\infty).\end{aligned}$$

For the second term of (9), by Lemma 3.14, 4.4, we have

$$\begin{aligned}\max_{1\le k\le n}\left|A^*_{k\cdot}\left(u_2^{(k)}-u_2^*\right)\right| &\le \max_{1\le k\le n}\|A^*\|_{2,\infty}\left\|u_2^{(k)}-u_2^*\right\|\\ &\le \|A^*\|_{2,\infty}\left(\max_{1\le k\le n}\left\|u_2-u_2^{(k)}\right\|+\|u_2-u_2^*\|\right)\\ &= \frac{\log n}{\sqrt{n}}\cdot O\left(\|u_2\|_\infty+\frac{1}{\sqrt{\log n}}\right)\\ &= O\left(\frac{\log n}{\sqrt{n}}\|u_2\|_\infty+\frac{\sqrt{\log n}}{\sqrt{n}}\right).\end{aligned}$$

For the third term of (9), by Lemma 4.5, we have

$$\max_{1\le k\le n}\left|(A-A^*)_{k\cdot}\left(u_2^{(k)}-u_2^*\right)\right| = O\left(\frac{\log n}{\log\log n}\|u_2\|_\infty\right).$$

Combining the three terms together, we have

$$\|A(u_2-u_2^*)\|_\infty = O\left(\frac{\log n}{\log\log n}\|u_2\|_\infty\right).$$

By definition of u_2,

$$\begin{aligned}(D-A)u_2 &= \lambda_2(L)u_2\\ (D-\lambda_2(L)I)u_2 &= Au_2\\ u_2 &= (D-\lambda_2(L)I)^{-1}Au_2.\end{aligned}$$

So

$$\begin{aligned}\|u_2\|_\infty &= \|(D-\lambda_2(L)I)^{-1}Au_2\|_\infty\\ &\le \|(D-\lambda_2(L)I)^{-1}Au_2^*\|_\infty+\|(D-\lambda_2(L)I)^{-1}A(u_2-u_2^*)\|_\infty. \end{aligned}\tag{10}$$

For the first term of (10), by Lemma 3.2 and Lemma 3.9 we have

$$\begin{aligned}\|(D-\lambda_2(L)I)^{-1}Au_2^*\|_\infty &= O\left(\|(D-\lambda_2(L)I)^{-1}\|_\infty\cdot\|A\|_\infty\cdot\|u_2^*\|_\infty\right)\\ &= O\left(\left|\frac{1}{d_{\min}-\lambda_2(L)}\right|\|A\|_\infty\cdot\|u_2^*\|_\infty\right)\\ &= O\left(\frac{1}{\log n}\log n\frac{1}{\sqrt{n}}\right)\\ &= O\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

For the second term of (10),

$$\begin{aligned}\|(D-\lambda_2(L)I)^{-1}A(u_2-u_2^*)\|_\infty &= O(\|(D-\lambda_2(L)I)^{-1}\|_\infty\cdot\|A(u_2-u_2^*)\|_\infty)\\ &= O\left(\left|\frac{1}{d_{\min}-\lambda_2(L)}\right|\|A(u_2-u_2^*)\|_\infty\right)\\ &= O\left(\frac{1}{\log n}\frac{\log n}{\log\log n}\|u_2\|_\infty\right)\\ &= O\left(\frac{\|u_2\|_\infty}{\log\log n}\right).\end{aligned}$$

Thus

$$\|u_2\|_\infty = O\left(\frac{1}{\sqrt{n}}\right)$$

and

$$\|A\left(u_2 - u_2^*\right)\|_\infty = O\left(\frac{\log n}{\sqrt{n}\log\log n}\right).$$

□

Lemma 4.7. *$(D-\lambda_2(L)I)^{-1}Au_2^*$ is a good approximation to u_2, that is, with probability $1-o(1)$,*

$$\|u_2 - (D-\lambda_2(L)I)^{-1}Au_2^*\|_\infty = o\left(\frac{1}{\sqrt{n}}\right).$$

Proof. Note that $(D-\lambda_2(L)I)u_2 = Au_2$, thus

$$\begin{aligned}\|u_2 - (D-\lambda_2(L)I)^{-1}Au_2^*\|_\infty &= \|(D-\lambda_2(L)I)^{-1}A(u_2-u_2^*)\|_\infty \\ &\leq \|(D-\lambda_2(L)I)^{-1}\|_\infty \cdot \|A(u_2-u_2^*)\|_\infty. \qquad (11)\end{aligned}$$

From Lemma 3.1 and Lemma 3.8, we have $d_{\min} \geq (m-1)(b_m+\epsilon)\log n$ and $\lambda_2(L) \leq (m-1)b_m\log n + O(\log n/n)$. Thus for the first multiplicative term of (11), we have

$$\|(D-\lambda_2(L)I)^{-1}\|_\infty = O\left(\frac{1}{d_{\min}-\lambda_2(L)}\right) = O\left(\frac{1}{\log n}\right)$$

with high probability.

From Lemma 4.6, for the second multiplicative term of (11), we have

$$\|A(u_2-u_2^*)\|_\infty = O\left(\frac{\log n}{\sqrt{n}\log\log n}\right) = o\left(\frac{\log n}{\sqrt{n}}\right)$$

with high probability.

Thus

$$\|u_2 - (D-\lambda_2(L)I)^{-1}Au_2^*\|_\infty = o\left(\frac{1}{\sqrt{n}}\right).$$

□

Lemma 4.8. *$sgn((D-\lambda_2(L)I)^{-1}Au_2^*)$ exactly recovers the HSBM and*

$$\|(D-\lambda_2(L)I)^{-1}Au_2^*\|_\infty = \Omega\left(\frac{1}{\sqrt{n}}\right).$$

Proof. We have $d_{max} - \lambda_2(L) = O\left(\log n\right)$.
For Au_2^*, by Lemma 4.2, when $i \leq \frac{n}{2}$, with probability $1-o(n^{-1})$

$$(Au_2^*)_i = \frac{1}{\sqrt{n}}\sum_{j\in[n]} A_{ij}\sigma^*(i)\sigma^*(j) \geq \epsilon\frac{\log n}{\sqrt{n}}.$$

Similarly, when $\frac{n}{2}+1 \leq i \leq n$, with probability $1-o(n^{-1})$, we have

$$(Au_2^*)_i = -\frac{1}{\sqrt{n}}\sum_{j\in[n]} A_{ij}\sigma^*(i)\sigma^*(j) \leq -\epsilon\frac{\log n}{\sqrt{n}}.$$

Thus, with probability $1-o(n^{-1})$, for some constant $\epsilon_1, \epsilon_2, \eta_1 > 0$, we have

$$\left|((D-\lambda_2(L)I)^{-1}Au_2^*)_i\right| \geq \left|\frac{1}{d_{max}-\lambda_2(L)}(Au_2^*)_i\right| = \left|\frac{\epsilon_1}{\log n}\cdot\frac{\epsilon_2\log n}{\sqrt{n}}\right| = \frac{\eta_1}{\sqrt{n}}$$

where $\eta_1 = |\epsilon_1\epsilon_2|$. Therefore,

$$\|(D-\lambda_2(L)I)^{-1}Au_2^*\|_\infty = \Omega\left(\frac{1}{\sqrt{n}}\right).$$

□

5 Conclusion

In this thesis, we proved that when above the min-bisection threshold, we can use the sign of the eigenvector, u_2, corresponding to the second smallest eigenvalue of the combinatorial Laplacian, $\lambda_2(L)$, to exactly recover the m-uniform binary hypergraph stochastic block model. First, we prove that the eigenvalue $\lambda_2(L)$ is well separated from $\lambda_1 = 0$ and λ_3 with high probability to ensure that u_2 can be computed accurately. Second, we approximate u_2 with $(D - \lambda_2(L)I)^{-1}Au_2^*$. By showing that the entrywise error between u_2 and the appoximation is small, and that the approximation can exactly recover the HSBM, we conclude that u_2 of L can exactly recover the HSBM. One potential future work is to extend the result to non-uniform HSBM, which has a more complex hyperedge dependency structure. Note that our result is valid down to the min-bisection threshold. While [Wan23] showed that below the information-theoretic threshold, exact recovery is not achievable for binary HSBM, whether spectral algorithm or semi-definite programming can achieve exact recovery when in-between these thresholds remains an open problem.

6 Acknowledgement

I would like to express my great appreciation to my advisor, Professor Ioana Dumitriu, who introduced me to graph theory and the stochastic block model. Thank you for guiding me through my independent study over the past few years. Every conversation with you has deepened my knowledge. I am grateful for your explanations of difficult concepts in various papers and for your guidance throughout this thesis. I also want to thank Dr. Haixiao Wang for the helpful discussions and valuable feedback on my thesis.

References

[Abb+20] Emmanuel Abbe et al. "Entrywise eigenvector analysis of random matrices with low expected rank". In: *Annals of statistics* 48.3 (2020), p. 1452.

[ABH15] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. "Exact recovery in the stochastic block model". In: *IEEE Transactions on information theory* 62.1 (2015), pp. 471–487.

[Chu97] Fan RK Chung. *Spectral graph theory.* Vol. 92. American Mathematical Soc., 1997.

[CRV15] P Chin, A Rao, and V Vu. "Stochastic block model and community detection in the sparse graphs: A spectral algorithm with optimal rate of recovery. Preprint". In: *arXiv preprint arXiv:1501.05021* (2015).

[Dec+11] Aurelien Decelle et al. "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications". In: *Physical review E* 84.6 (2011), p. 066106.

[DLS21] Shaofeng Deng, Shuyang Ling, and Thomas Strohmer. "Strong consistency, graph laplacians, and the stochastic block model". In: *Journal of Machine Learning Research* 22.117 (2021), pp. 1–44.

[DW23] Ioana Dumitriu and Haixiao Wang. *Exact recovery for the non-uniform Hypergraph Stochastic Block Model.* 2023. arXiv: `2304.13139 [math.ST]`.

[GJ23] Julia Gaudio and Nirmit Joshi. "Community Detection in the Hypergraph SBM: Exact Recovery Given the Similarity Matrix". In: *Proceedings of Thirty Sixth Conference on Learning Theory.* Ed. by Gergely Neu and Lorenzo Rosasco. Vol. 195. Proceedings of Machine Learning Research. PMLR, Dec. 2023, pp. 469–510. URL: `https://proceedings.mlr.press/v195/gaudio23a.html`.

[Mas14] Laurent Massoulié. "Community detection thresholds and the weak Ramanujan property". In: *Proceedings of the forty-sixth annual ACM symposium on Theory of computing.* 2014, pp. 694–703.

[MNS14] Elchanan Mossel, Joe Neeman, and Allan Sly. "Consistency thresholds for binary symmetric block models". In: *arXiv preprint arXiv:1407.1591* 3.5 (2014).

[MNS15] Elchanan Mossel, Joe Neeman, and Allan Sly. "Reconstruction and estimation in the planted partition model". In: *Probability Theory and Related Fields* 162 (2015), pp. 431–461.

[Wan23] Haixiao Wang. *Strong consistency and optimality of spectral clustering in symmetric binary non-uniform Hypergraph Stochastic Block Model.* 2023. arXiv: `2306.06845 [math.ST]`.

[YP14] Se-Young Yun and Alexandre Proutiere. "Accurate community detection in the stochastic block model via spectral algorithms". In: *arXiv preprint arXiv:1412.7335* (2014).

[ZB18] Yiqiao Zhong and Nicolas Boumal. "Near-optimal bounds for phase synchronization". In: *SIAM Journal on Optimization* 28.2 (2018), pp. 989–1016.