Proceedings of Machine Learning Research vol 195:1–42, 2023 36th Annual Conference on Learning Theory

Community Detection in the Hypergraph SBM: Optimal Recovery Given the Similarity Matrix

Julia Gaudio JULIA.GAUDIO@NORTHWESTERN.EDU
Northwestern University

Nirmit Joshi NIRMIT@TTIC.EDU
Toyota Technological Institute at Chicago

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Community detection is a fundamental problem in network science. In this paper, we consider community detection in hypergraphs drawn from the *hypergraph stochastic block model* (HSBM), with a focus on exact community recovery. We study the performance of polynomial-time algorithms which operate on the *similarity matrix* W, where W_{ij} reports the number of hyperedges containing both i and j. Under this information model, Kim, Bandeira, and Goemans determined the information-theoretic threshold for exact recovery in the logarithmic degree regime, and proposed a semidefinite programming relaxation which they conjectured to be optimal. In this paper, we confirm this conjecture. We also design a simple and highly efficient spectral algorithm with nearly linear runtime and show that it achieves the information-theoretic threshold. Moreover, the spectral algorithm also succeeds in denser regimes and is considerably more efficient than previous approaches, establishing it as the method of choice. Our analysis of the spectral algorithm crucially relies on strong *entrywise* bounds on the eigenvectors of W. Our bounds are inspired by the work of Abbe, Fan, Wang, and Zhong, who developed entrywise bounds for eigenvectors of symmetric matrices with independent entries. Despite the complex dependency structure in similarity matrices, we prove similar entrywise guarantees.

Keywords: Community Detection, Hypergraph Stochastic Block Model, Spectral Algorithm

1. Introduction

Community detection is the problem of partitioning a network into densely connected clusters. As a fundamental network science problem, community detection arises in numerous applications: sociology (Goldenberg et al., 2010; Newman et al., 2002), protein interactions (Chen and Yuan, 2006; Marcotte et al., 1999), image applications (Shi and Malik, 2000), natural language processing (Ball et al., 2011), webpage sorting (Kumar et al., 1999) and many more. In 1983, Holland et al. (1983) introduced the *stochastic block model* (SBM), a probabilistic generative model for networks with community structure. Since then, community detection in the SBM has been intensely studied in the probability, statistics, and theoretical computer science communities (Dyer and Frieze, 1989; McSherry, 2001; Decelle et al., 2011; Mossel et al., 2013; Massoulié, 2014; Abbe and Sandon, 2015a; Abbe et al., 2016; Abbe and Sandon, 2015b); also see Abbe (2017) for a survey.

In this paper, we consider an extension of the SBM to *hypergraphs*. A hypergraph is a generalization of a graph that captures higher-order interactions. For example, an academic co-authorship network may be modeled as a hypergraph, where each hyperedge represents the author list of a paper. Formally, a hypergraph is specified by a set of vertices V and a set of hyperedges E. Each hyperedge $e \in E$ is a subset of V. We specialize to *uniform* hypergraphs, where each hyperedge contains the same number of vertices. A *d-uniform hypergraph* satisfies $|e| = d$ for all $e \in E$ (in particular, a graph is 2-uniform hypergraph). We then say that the hypergraph has *order* d.

We now describe the Hypergraph Stochastic Block Model (HSBM), which was first proposed by Ghoshdastidar and Dukkipati (2014). We consider the version with two balanced communities and equal inter-community edge probabilities. The model is specified by its order d and two parameters $1 \geq p_n > q_n > 0$. First, a community assignment vector $\sigma^* \in \{\pm 1\}^n$ is sampled uniformly at random from the set $\{\sigma \in \{\pm 1\}^n : \langle \mathbf{1}_n, \sigma \rangle = 0\}$[1], where $\mathbf{1}_n \in \mathbb{R}^n$ is the vector of all ones.

Conditioned on σ^*, we sample a hypergraph $G = ([n], E)$ as follows. Each $e = \{i_1, i_2, \ldots, i_d\} \in \binom{[n]}{d}$ appears as a hyperedge independently with probability

$$\mathbb{P}(e \in E) = \begin{cases} p_n & \sigma^*(i_1) = \sigma^*(i_1) = \cdots = \sigma^*(i_d) \\ q_n & \text{otherwise.} \end{cases}$$

We then write $G \sim \text{HSBM}(d, n, p_n, q_n)$. Throughout, we consider d to be a constant and denote $\mathcal{E} := \binom{[n]}{d}$ to be the set of all possible hyperedges. We use the parametrization:

$$p_n = \alpha f_n \quad \text{and} \quad q_n = \beta f_n; \tag{1}$$

where either ($f_n = o(1)$ and $\alpha > \beta > 0$) or ($f_n = 1$ and $1 \geq \alpha > \beta > 0$) for constants α, β.

We are interested in algorithms that recover all of the vertex labels. More formally, we say that an estimator $\hat{\sigma}_n$ *achieves exact recovery* if

$$\lim_{n \to \infty} \mathbb{P}(\hat{\sigma}_n \in \{\pm \sigma_n^*\}) = 1.$$

For clarity of presentation, we typically drop the dependence on n.

The limiting regime for the exact recovery problem is $f_n = \Theta\left({}^{\log n}\!/_{n^{d-1}}\right)$. That is, when $f_n = o\left({}^{\log n}\!/_{n^{d-1}}\right)$, exact recovery is not possible statistically. This is because the hypergraph will have isolated vertices with a high probability, which are impossible to classify information theoretically. On the other hand, when $f_n = \omega\left({}^{\log n}\!/_{n^{d-1}}\right)$, there are efficient algorithms for exact recovery (Chien et al., 2019). In the logarithmic degree regime, we typically parametrize as $f_n = {}^{\log n}\!/\binom{n-1}{d-1}$. In the corresponding model $\text{HSBM}(d, n, \alpha f_n, \beta f_n)$, there is a precise information theoretic threshold, determined by Kim et al. (2018). If $I_{\text{full}}(d, \alpha, \beta) := \frac{1}{2^{d-1}}\left(\sqrt{\alpha} - \sqrt{\beta}\right)^2 < 1$, then exact recovery is impossible, while if $I_{\text{full}}(d, \alpha, \beta) > 1$, then there are polynomial-time algorithms for the exact recovery problem developed by Ghoshdastidar and Dukkipati (2015a,b, 2017); Ahn et al. (2018); Chien et al. (2019), culminating in the work of Zhang and Tan (2022), whose algorithm applies to a general class of HSBMs. Their results essentially show that there is no statistical-computational gap for the exact recovery problem.

While the work of Zhang and Tan (2022) resolves the question of poly-time exact recovery even for general HSBMs, the algorithm uses the full hypergraph information. Unfortunately, storing the full hypergraph information can be prohibitively expensive; in the regime where $p_n, q_n = \Theta(1)$, it requires $\Theta(n^d)$ space. A natural question arises: is there some other way of storing the hypergraph information which uses less space than storing the full information, while still being powerful enough for exact recovery? One candidate information model is the so-called *similarity matrix*.

1. For simplicity of exposition, we assume n is even, and thus, each community has exactly $n/2$ vertices.

Definition 1 (Similarity Matrix) *Let $G = ([n], E)$ be a hypergraph on n vertices. The similarity matrix of G is the zero-diagonal matrix W whose entries are*

$$W_{ij} = |\{e \in E : \{i, j\} \subset e\}|$$

for $i \neq j$. In other words, W_{ij} counts the number of edges which contain both i and j. We also write $W = \mathcal{S}(G)$ to define the similarity matrix transformation.

Even in the worst case when $p_n, q_n = \Theta(1)$, the similarity matrix requires only $O(n^2 \log n)$ space to store. Recent works (Lee et al., 2020; Kim et al., 2018; Cole and Zhu, 2020) have considered the exact recovery problem, given the similarity matrix $W = \mathcal{S}(G)$, where $G \sim \text{HSBM}(d, n, p_n, q_n)$. Lee et al. (2020) showed that the asymptotic regime under which the similarity matrix is powerful enough for exact recovery is given by $p_n - q_n = \Omega\left(\sqrt{p_n \log n / n^{d-1}}\right)$. In the logarithmic degree regime where $f_n = \log n / \binom{n-1}{d-1}$ and $G \sim \text{HSBM}(d, n, \alpha f_n, \beta f_n)$, (Kim et al., 2018, Theorem 3) determined the information-theoretic threshold for exact recovery given $W = \mathcal{S}(G)$. Letting

$$I(d, \alpha, \beta) = \max_{t \geq 0} \frac{1}{2^{d-1}} \left[\alpha \left(1 - e^{-(d-1)t}\right) + \beta \sum_{r=1}^{d-1} \binom{d-1}{r} \left(1 - e^{-(d-1-2r)t}\right) \right], \qquad (2)$$

exact recovery given W is not possible if $I(d, \alpha, \beta) < 1$. The MLE estimator corresponds to solving a min-bisection problem, which is NP-hard. Therefore, Kim et al. (2018) additionally proposed a semidefinite programming (SDP) relaxation, showing that it succeeds in exact recovery if $I_{\text{SDP}}(d, \alpha, \beta) > 1$, where

$$I_{\text{SDP}}(d, \alpha, \beta) = \frac{(d-1)}{2^{2d}} \cdot \frac{(\alpha - \beta)^2}{\left(\alpha \frac{d}{2^d} + \beta \left(1 - \frac{d}{2^d}\right)\right)}.$$

Note that $I_{\text{SDP}}(d, \alpha, \beta) > I(d, \alpha, \beta)$ for $d \geq 3$; see Figure 1 for an illustration. The authors conjectured that the SDP algorithm is optimal; i.e. it succeeds whenever $I(d, \alpha, \beta) > 1$ (Kim et al., 2018, Conjecture 1.2). This leaves an open question:

Can this information-theoretic threshold for exact recovery from a similarity matrix be achieved by a polynomial time algorithm? Or is there a computational-statistical gap?

We also investigate the question of whether using the similarity matrix enables efficient, nearly linear algorithms. Motivated by the success of spectral algorithms for other community detection problems (Abbe et al., 2020; Deng et al., 2021; Dhara et al., 2022a,b), and given their efficiency, we ask:

Can we design a spectral algorithm with nearly linear runtime that achieves the information-theoretic threshold from the similarity matrix?

1.1. Our Contributions

Our first main contribution is to show that the SDP algorithm succeeds whenever $I(d, \alpha, \beta) > 1$ (Theorem 2), establishing that there is no computational-statistical gap. We note that there has been some follow-up work on community recovery from the similarity matrix (Lee et al., 2020; Cole and

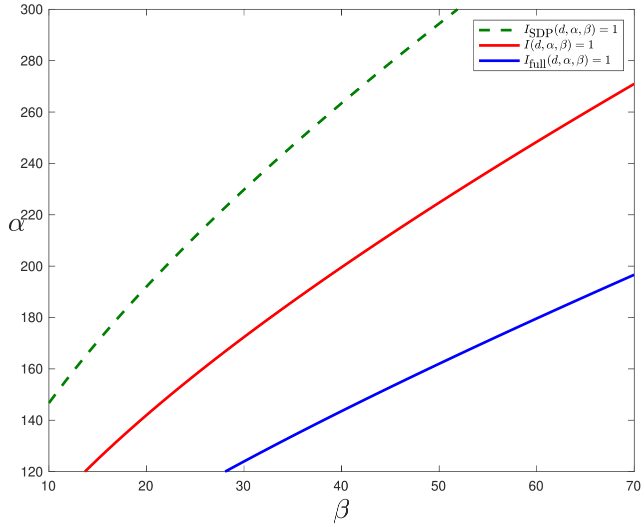

Figure 1: Visualization of $I_{\text{SDP}}(d,\alpha,\beta)=1$, $I(d,\alpha,\beta)=1$ and $I_{\text{full}}(d,\alpha,\beta)=1$ when $d=6$.

Zhu, 2020), but to our knowledge, we are the first to resolve (Kim et al., 2018, Conjecture 1.2). We also show that the SDP is robust to monotone adversarial changes on the similarity matrix.

Our second main contribution is a simple spectral algorithm based on the similarity matrix that also achieves the information-theoretic threshold in the logarithmic degree regime (Theorem 4). Our algorithm determines communities based on the signs of the entries of the second eigenvector of the similarity matrix and does not require any clean-up. Furthermore, the algorithm can be implemented in $O(n\log^2 n)$ time in the logarithmic degree regime, using the fast eigenvector computation method of Garber et al. (2016). When $f_n=\omega(\log n/n^{d-1})$, we may subsample the hypergraph to return to the logarithmic degree regime. Since the subsampling procedure takes $O(n\log n)$ time, the overall runtime is still $O(n\log^2 n)$. Table 1 summarizes the runtime of the algorithm, compared to other approaches in the literature (allowing for subsampling), including results coming after our initial posting on arXiv. These works include Wang et al. (2023), which proposed the Projected Tensor Power Method. The algorithm achieves the exact recovery threshold $I_{\text{full}}(d,\alpha,\beta)>1$, with a runtime of $O(n\log^2(n)/\log\log n)$, when initialized from a labeling satisfying a certain partial recovery condition (Wang et al., 2023, Theorem 4.1). Most recently, Dumitriu and Wang (2023) proposed a two-stage algorithm achieving exact recovery down to the information-theoretic threshold. Dumitriu and Wang (2023) also identified the recovery threshold in the non-uniform case, and showed that their algorithm achieves the information-theoretic threshold in that case also.

We also compare runtimes when only the similarity matrix is available. In this situation, we cannot subsample down to the logarithmic degree regime, due to the loss of hyperedge information. Nevertheless, we show that our spectral algorithm achieves exact recovery directly from $W=\mathcal{S}(G)$, where $G\sim\text{HSBM}(d,n,p_n,q_n)$, in all the regimes of p_n,q_n captured by (1) as long as $f_n=\omega(\log n/n^{d-1})$ (Theorem 5). Moreover, we show that its runtime is $O(n^2\log n)$ in the worst-case; i.e. $p_n,q_n=\Theta(1)$. See Table 2.

Strategy (Reference)	Runtime (all regimes)
Spectral (This work)	$O(n \log^2 n)$
SDP (Kim et al., 2018)	$\tilde{O}(n^{3.5})$ (Jiang et al., 2020)
Spectral+refinement (Chien et al., 2019)	$O(n^3 \log n)$
Projected Tensor Power Method (Wang et al., 2023)	$O(n \log^2(n))$ [2]
Spectral + refinement (Dumitriu and Wang, 2023)	$O(n \log^2(n))$

Table 1: Runtime comparison when the full hypergraph is known.

Strategy (Reference)	Runtime (Logarithmic degree regime)	Runtime (All regimes)
Spectral (This work)	$O(n \log^2 n)$	$O(n^2 \log n)$
SDP (Kim et al., 2018)	$\tilde{O}(n^{3.5})$ (Jiang et al., 2020)	

Table 2: Runtime comparison when only the similarity matrix is known.

In order to analyze our spectral algorithm, inspired by the work of Abbe et al. (2020), we develop ℓ_∞ (entrywise) bounds for the eigenvectors of similarity matrices of a large class of random hypergraph models (Theorem 7), which may be of independent interest. Roughly speaking, we show that an eigenvector u_k of a random similarity matrix W is close to its first order approximation Wu_k^*/λ_k^* in the ℓ_∞ norm under mild conditions, where (λ_k^*, u_k^*) is the corresponding eigenpair of $\mathbb{E}[W]$. Our result addresses two important questions raised by Abbe et al. (2020) by: (1) providing an example of entrywise eigenvector approximation beyond symmetric matrices with independent entries, (2) expanding the class of graph-based matrices for which entrywise eigenvector guarantees are known, beyond adjacency matrices (Abbe et al., 2020) and Laplacian matrices (Deng et al., 2021).

Organization. Section 2 contains our main results. We give proof outlines of the SDP and spectral algorithm results in Section 3 and 4, respectively. Directions for future work are proposed in Section 5. The proofs of our main results are provided in the appendix.

1.2. Further Related Work

Other recovery problems. While we only focus on exact recovery in this work, partial recovery (recovering a non-trivial constant fraction of the community labels) and almost exact recovery (recovering all but a vanishing fraction of the community labels) have also been studied in the context of the HSBM (Angelini et al., 2015; Zhang et al., 2022; Dumitriu et al., 2021; Zhen and Wang, 2022; Ke et al., 2019). Community recovery in the non-uniform HSBM has been studied (Ghoshdastidar and Dukkipati, 2017; Dumitriu et al., 2021; Alaluusua et al., 2023), with sharp results in the exact recovery problem given by the very recent work of Dumitriu and Wang (2023). Additionally, Dumitriu and Wang (2023) study almost exact recovery in non-uniform HSBM and show that it is possible whenever the maximal expected degree diverges, and give an efficient algorithm for the same.

2. Given a suitable initialization that satisfies a certain partial recovery criterion, their algorithm runs in $O(\frac{n \log^2 n}{\log \log n})$ time; see (Wang et al., 2023, Theorem 4.1). However, to the best of our knowledge, this criterion can only be achieved in $O(n \log^2 n)$ time using current methods (Dumitriu and Wang, 2023).

Spectral methods. Spectral algorithms have been very successful in statistical inference problems, e.g. community detection (Newman, 2006; Rohe et al., 2011; McSherry, 2001; Abbe et al., 2020; Dhara et al., 2022c; Deng et al., 2021), the Planted Clique Problem (Alon et al., 1998), clustering (Von Luxburg, 2007; Ng et al., 2001), dimensionality reduction (Belkin and Niyogi, 2003) and many more; see (Chen et al., 2021) for a survey on the topic.

Follow-up work. Wang (2023) generalized our spectral method to the non-uniform case, where the input is the sum of similarity matrices of all orders. Interestingly, the spectral method achieves the information-theoretic threshold under this information model, while the natural SDP (Wang, 2023, Algorithm 3) does not. Therefore, the simultaneous optimality of spectral algorithms and SDPs is unique to the uniform case. Alaluusua et al. (2023) considered a multi-layer version of the HSBM, where each layer corresponds to an independent, regular HSBM. Their work provides a sufficient condition for exact recovery from the aggregated similarity matrix (the sum of the similarity matrices of all layers), using an SDP approach.

1.3. Notation

For any real numbers $a, b \in R$, we denote $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Let $\operatorname{sgn} : \mathbb{R} \to \{\pm 1\}$ be the function defined by $\operatorname{sgn}(x) = 1$ if $x \geq 0$ and $\operatorname{sgn}(x) = -1$ if $x < 0$. We also extend the definition to vectors; let $\operatorname{sgn} : \mathbb{R}^n \to \{\pm 1\}^n$ be the map defined by applying the sign function componentwise.

We use the notation $\mathbb{R}_+ = (0, \infty)$. For $n \in \mathbb{N}$, we write $[n] = \{1, 2, \dots, n\}$. We use the Bachmann–Landau notation $o(.)$, $O(.)$, $\omega(.)$, $\Omega(.)$, $\Theta(.)$ etc. throughout the paper. For nonnegative sequences $(a_n)_{n \geq 1}$ and $(b_n)_{n \geq 1}$, we write $a_n \lesssim b_n$ to mean $a_n \leq C b_n$ for some constant $C > 0$. The notation $\asymp$ is similar, hiding two constants in upper and lower bounds. Moreover, we write $a_n \approx b_n$ as a shorthand for $\lim_{n \to \infty} \frac{a_n}{b_n} = 1$.

For any two vectors $x, y \in \mathbb{R}^n$, $\langle x, y \rangle$ represents the standard inner product in $\mathbb{R}^n$; we define $\|x\|_2 = (\sum_{i=1}^n x_i^2)^{1/2}$, $\|x\|_1 = \sum_{i=1}^n |x_i|$, and $\|x\|_\infty = \max_i |x_i|$. For any matrix $M \in \mathbb{R}^{n \times n}$, $M_{i \cdot}$ refers to its i-th row, which is a row vector, and $M_{\cdot i}$ refers to its i-th column, which is a column vector. The matrix spectral norm is $\|M\|_2 = \sup_{\|x\|_2 = 1} \|Mx\|_2$, the matrix $2 \to \infty$ norm is $\|M\|_{2 \to \infty} = \sup_{\|x\|_2 = 1} \|Mx\|_\infty = \sup_i \|M_{i \cdot}\|_2$, and the the matrix Frobenius norm is $\|M\|_F = (\sum_{i=1}^n \sum_{j=1}^n M_{ij}^2)^{1/2}$.

2. Main Results

Recall that $\sigma_n^* \in \{\pm 1\}^n$ denotes the true community assignment vector. Let G be a hypergraph on n vertices, and let $W = \mathcal{S}(G)$ be its similarity matrix. Kim et al. (2018) proposed an SDP for exact recovery (Algorithm 1). Our first main result states that the SDP relaxation achieves exact recovery down to the information-theoretic threshold in the logarithmic degree regime.

Theorem 2 *Fix $d \in \{2, 3, \dots\}$ and $\alpha > \beta > 0$ such that $I(d, \alpha, \beta) > 1$. Let $f_n = \log n / \binom{n-1}{d-1}$. Suppose $G \sim \mathrm{HSBM}(d, n, \alpha f_n, \beta f_n)$, and let $W = \mathcal{S}(G)$. Let $\hat{X}$ be the optimal solution to (3) with input W. Then $\hat{X} = \sigma^* {\sigma^*}^\top$ with probability $1 - o(1)$. It follows that*

$$\lim_{n \to \infty} \mathbb{P}\left(\hat{\sigma}_{\text{SDP}} \in \{\pm \sigma^*\}\right) = 1.$$

Algorithm 1 SDP recovery algorithm (Kim et al., 2018)

Input: An $n \times n$ similarity matrix W

Output: An estimate of community assignments

1: Solve the following SDP, where $X \in \mathbb{R}^{n \times n}$.

$$\begin{aligned} &\max \ \langle W, X \rangle \\ &\text{subject to } X_{ii} = 1 \text{ for all } i \in [n] \\ &\langle X, \mathbf{1}\mathbf{1}^\top \rangle = 0, \ X \succeq 0. \end{aligned} \tag{3}$$

2: Let $\hat{X}$ be the optimal solution, and let $\hat{X} = \sum_{i=1}^n \lambda_i v_i v_i^\top$ denote the eigendecomposition of $\hat{X}$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

3: Return $\hat{\sigma}_{\text{SDP}} = \operatorname{sgn}(v_1)$.

Note that the SDP algorithm also works in denser regimes of (1) (Lee et al., 2020). We also establish that the SDP continues to achieve exact recovery even under a *monotone adversary* model.

Lemma 3 *Consider the modified SDP based on* (3)*, with W replaced by $\widetilde{W}$ such that $\widetilde{W}_{ij} \geq W_{ij}$ if $\sigma^*(i) = \sigma^*(j)$, and $\widetilde{W}_{ij} \leq W_{ij}$ if $\sigma^*(i) \neq \sigma^*(j)$. If $I(d, \alpha, \beta) > 1$, then $X^* := \sigma^* {\sigma^*}^\top$ is the unique optimal solution to the modified SDP.*

Having shown the optimality of the SDP relaxation, we propose a spectral algorithm.

Algorithm 2 Spectral recovery algorithm

Input: An $n \times n$ similarity matrix W

Output: An estimate of community assignments

1: Compute the second eigenpair of W, denoted by (λ_2, u_2), where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

2: Return $\hat{\sigma}_{\text{spec}} = \operatorname{sgn}(u_2)$.

We first establish the optimality of Algorithm 2 in the logarithmic degree regime.

Theorem 4 *Fix $d \in \{2, 3, \dots\}$ and $\alpha > \beta > 0$ such that $I(d, \alpha, \beta) > 1$. Let $f_n = \log n / \binom{n-1}{d-1}$. Suppose $G \sim \mathrm{HSBM}(d, n, \alpha f_n, \beta f_n)$, and let $W = \mathcal{S}(G)$. Let u_2 be the second eigenvector of W. Then with probability $1 - o(1)$, there exist $s \in \{\pm 1\}$ and $\eta = \eta(d, \alpha, \beta) > 0$ such that*

$$\sqrt{n} \min_{i \in [n]} s \sigma^*(i) u_{2,i} > \eta.$$

As a result, the estimator $\hat{\sigma}_{\text{spec}}$ produced by Algorithm 2 on input W achieves exact recovery.

We also show that the spectral algorithm succeeds in all the super-logarithmic degree regimes in (1).

Theorem 5 *Fix $d \in \{2, 3, \dots\}$. Let p_n and q_n be parameterized according to* (1) *for some f_n and constants $\alpha > \beta > 0$. Suppose $G \sim \mathrm{HSBM}(d, n, \alpha f_n, \beta f_n)$, and let $W = \mathcal{S}(G)$. If $f_n = \omega(\log n / n^{d-1})$, then the estimator $\hat{\sigma}_{\text{spec}}$ produced by Algorithm 2 on input W achieves exact recovery; i.e.*

$$\lim_{n \to \infty} \mathbb{P}\left(\hat{\sigma}_{\text{spec}} \in \{\pm \sigma^*\}\right) = 1.$$

Our proofs of Theorem 4 and 5 crucially rely on *entrywise* bounds on the second eigenvector of the similarity matrix. To this end, we develop entrywise bounds on the eigenvectors of similarity matrices of a generic family of random hypergraphs (Definition 6).

Definition 6 (General random hypergraph) *Let $d \in \{2, 3, \ldots\}$, $n \in \mathbb{N}$, and $p \in [0,1]^{\binom{[n]}{d}}$. Define $H(d,n,p)$ to be the distribution on d-uniform hypergraphs with n vertices, where each edge $e \in \binom{[n]}{d}$ appears in the hypergraph with probability p_e, independently.*

We analyze the eigenvectors of $W = \mathcal{S}(G)$, where $G \sim H(d,n,p)$. Let $(\lambda_i, u_i)_{i=1}^n$ denote the eigenpairs of W, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Let $W^* = \mathbb{E}[W]$, with ordered eigenpairs $(\lambda_i^*, u_i^*)_{i=1}^n$. We use the convention $\lambda_0 = \lambda_0^* = +\infty$ and $\lambda_{n+1} = \lambda_{n+1}^* = -\infty$. We then define the following eigengap quantity:

$$\Delta_k^* = \min\{\lambda_{k-1}^* - \lambda_k^*, \lambda_k^* - \lambda_{k+1}^*\}.$$

Our entrywise guarantee requires a spectral separation assumption. The assumption easily holds for similarity matrices of (general) HSBMs, in all the parameter regimes we are interested in.

A1 (Spectral separation) *There is a sequence $\{\mu_n\}$ in $(0, \infty)$ such that*

$$\max\{p_e : e \in \mathcal{E}\} \leq \mu_n, \text{ and } n\binom{n-2}{d-2}\mu_n \geq c_0 \log n,$$

for some constant $c_0 > 0$. Moreover, there is a constant $c_1 \geq 1$ such that

$$\frac{1}{c_1} n^{d-1}\mu_n \leq |\lambda_k^*|, |\Delta_k^*| \leq c_1 n^{d-1}\mu_n; \text{ i.e. } |\lambda_k^*|, |\Delta_k^*| = \Theta(n^{d-1}\mu_n).$$

Under this assumption, we state our entrywise guarantee.

Theorem 7 *Let $k \in \mathbb{N}$ and $d \in \{2, 3, \ldots\}$ be constants. Let $p \in [0,1]^{\binom{[n]}{d}}$, such that Assumption 1 holds for some μ_n and constants $c_0, c_1 > 0$. Let $G \sim H(d,n,p)$, and $W = \mathcal{S}(G)$. Then with probability $1 - O(n^{-3})$,*

$$\min_{s^* \in \{\pm 1\}} \left\| u_k - s^* \frac{W u_k^*}{\lambda_k^*} \right\|_\infty \leq \frac{c \, \|u_k^*\|_\infty}{\log \log n},$$

where $c > 0$ is some constant depending only on d, c_0, and c_1.

Remark 8 *We remark that when $d = 2$, the similarity matrix of a graph is just its adjacency matrix and our Theorem 7 recovers the entrywise bounds of Abbe et al. (2020). Moreover, $I(2, \alpha, \beta) = (\sqrt{\alpha} - \sqrt{\beta})^2/2$, which is the information-theoretic threshold for exact recovery in the SBM setting (Abbe et al., 2015). Therefore, our Theorem 2 and Theorem 4, respectively, recover (Hajek et al., 2016a, Theorem 2) and (Abbe et al., 2020, Theorem 3.2).*

3. Optimality of the SDP Relaxation

Showing optimality of the SDP. We use a dual certificate strategy as in Kim et al. (2018). The dual of (3) is given by

$$\begin{aligned} \min \ & \operatorname{trace}(D) \\ \text{subject to } & D \text{ is } n \times n \text{ diagonal matrix}, \nu \in \mathbb{R}, \\ & D + \nu \mathbf{1}\mathbf{1}^\top - W \succeq 0. \end{aligned} \tag{4}$$

The form of the dual motivates the following sufficient condition, whose proof we include for completeness (see Appendix B).

Lemma 9 *Suppose there is a diagonal matrix $D \in \mathbb{R}^{n\times n}$ and $\nu \in \mathbb{R}$ such that the following holds. Letting $S \triangleq D + \nu \mathbf{1}\mathbf{1}^\top - W$, the matrix S is positive semidefinite, its second-smallest eigenvalue $\lambda_{n-1}(S)$ is strictly positive, and $S\sigma^* = 0$. Then $X^* := \sigma^*\sigma^{*\top}$ is the unique optimal solution to* (3).

To apply Lemma 9, we let D be the diagonal matrix whose diagonal entries are specified by

$$D_{ii} = \sum_{j\in[n]} W_{ij}\sigma^*(i)\sigma^*(j). \tag{5}$$

Setting $\nu = 1$, write $S = D + \mathbf{1}\mathbf{1}^\top - W$. By construction, we have $S\sigma^* = 0$. It remains to show

$$\mathbb{P}\left(\inf_{x\perp\sigma^*:\|x\|_2=1} x^\top S x > 0\right) = 1 - o(1). \tag{6}$$

This is where our proof diverges from Kim et al. (2018); rather than showing (6), Kim et al. (2018) proceed through a different sufficient condition. Using steps similar to the proof of (Hajek et al., 2016a, Theorem 2), we show that for all $x \perp \sigma^*$ such that $\|x\|_2 = 1$,

$$x^\top S x \geq \min_{i\in[n]} D_{ii} - \|W - W^*\|_2, \tag{7}$$

where W^* is the expected value of W, conditioned on σ^*. It remains to (1) lower-bound D_{ii} for each $i \in [n]$ and (2) upper-bound $\|W - W^*\|_2$.

To lower-bound D_{ii}, we condition on σ^* and establish concentration of D_{ii} around its mean. To see why D_{ii} should be positive and bounded away from zero, it helps to rewrite (5) as follows:

$$D_{ii} = \sum_{j\in[n]:\sigma^*(i)=\sigma^*(j)} W_{ij} - \sum_{j\in[n]:\sigma^*(i)\neq\sigma^*(j)} W_{ij}.$$

While the values $\{W_{ij}\}_{j=1}^n$ are dependent, they are functions of *independent* random variables (namely, the hyperedge random variables). After re-expressing D_{ii} in terms of the underlying hyperedge random variables, the proof proceeds by a Chernoff-style argument (Lemma 12). Whenever $I(d, \alpha, \beta) > 1$, we establish the existence of $\epsilon > 0$ such that for each i, $D_{ii} \geq \epsilon \log n$ with probability $1 - o(n^{-1})$. A union bound then implies $\min_{i\in[n]} D_{ii} \geq \epsilon \log n$ with high probability. Next, we need a tail bound on $\|W - W^*\|_2$. While sharp concentration results are known Lee et al. (2020), we note that we can also bound $\mathbb{E}\left[\|W - W^*\|_2\right]$ using much simpler techniques.

Theorem 10 (Spectral norm expectation) *Let $d \in \{2, 3, \dots\}$ be fixed. Let $p \in [0,1]^{\binom{[n]}{d}}$, where $\max_e p_e \leq c_0 \log n / \binom{n-1}{d-1}$ for some constant $c_0 > 0$. Let $G \sim H(d,n,p)$, and $W = \mathcal{S}(G)$ whose expectation is W^*. Then there exists a constant $c := c(d, c_0) > 0$ such that*

$$\mathbb{E}\left[\|W - W^*\|_2\right] \leq c\sqrt{\log n}.$$

Markov's inequality immediately implies the desired tail bound. Returning to (7), we see that $x^\top S x > 0$ simultaneously for all x satisfying $x \perp \sigma^*$, $\|x\|_2 = 1$, with high probability, concluding (6).

Spectral norm concentration. We highlight our proof technique for bounding $\mathbb{E}\left[\|W - W^*\|_2\right]$ (Theorem 10). Similar bounds are well-known for the spectral norm $\|A - \mathbb{E}[A]\|_2$ in the case where A is a symmetric matrix with independent, bounded entries and suitably bounded expectation (Feige and Ofek, 2005; Lei and Rinaldo, 2015; Hajek et al., 2016a). The first step in our proof is inspired by the symmetrization argument of Hajek et al. (2016a). Let R be a symmetric tensor of order d and dimension n with independent Rademacher entries. Let $G \circ R$ be the hypergraph where each hyperedge is independently assigned to a $+1$ or -1 label. Let $\mathcal{S}(G \circ R)$ denote the corresponding similarity matrix, where the (i,j) entry is the sum of ± 1-weighted hyperedges containing i, j. We show that

$$\mathbb{E}\left[\|W - W^*\|_2\right] \leq 2\mathbb{E}\left[\|\mathcal{S}(G \circ R)\|_2\right].$$

A coupling argument then allows us to replace G by $G^{(1)} \sim \text{HSBM}(d, n, p_{\max}, p_{\max})$, where $p_{\max} = \max_e p_e$. Unlike the matrix $W - W^*$, the matrix $S(G^{(1)} \circ R)$ has entries with identical distributions. However, the entries are dependent.

Our next goal is to create independence. We invoke Jensen's inequality, establishing

$$\mathbb{E}\left[\|W - W^*\|_2\right] \leq 2\mathbb{E}\left[\left\|\sum_{m=1}^{K} \mathcal{S}(G^{(m)} \circ R^{(m)})\right\|_2\right],$$

where each $G^{(m)}$ is an independent copy of $G^{(1)}$, each $R^{(m)}$ is an independent copy of R, and $K = d^2 - d$. Note that $\sum_{m=1}^{K} \mathcal{S}(G^{(m)} \circ R^{(m)})$ is a sum of *independent* matrices with *dependent* entries. For any $m \in [K]$, observe that a given hyperedge random variable affects exactly $2 \times \binom{d}{2} = K$ entries of $\mathcal{S}(G^{(m)} \circ R^{(m)})$. By adding K independent copies, we can rearrange the underlying hyperedge random variables to achieve $\sum_{m=1}^{K} \mathcal{S}(G^{(m)} \circ R^{(m)}) = \sum_{k=1}^{K} C^{(k)}$, where $\sum_{k=1}^{K} C^{(k)}$ is a sum of *dependent* matrices with *independent* entries. Then

$$\mathbb{E}[\|W - W^*\|_2] \leq 2\mathbb{E}\left[\left\|\sum_{k=1}^{K} C^{(k)}\right\|_2\right] \leq 2\sum_{k=1}^{K} \mathbb{E}[\|C^{(k)}\|_2].$$

The final summation is then straightforwardly bounded using (Hajek et al., 2016a, Theorem 5); see Appendix D for a complete proof.

4. Analysis of the Spectral Algorithm

Correctness of the spectral algorithm. Recall that u_2 denotes the second eigenvector of $W = \mathcal{S}(G)$. Since our algorithm determines communities based on the signs of u_2, we need precise

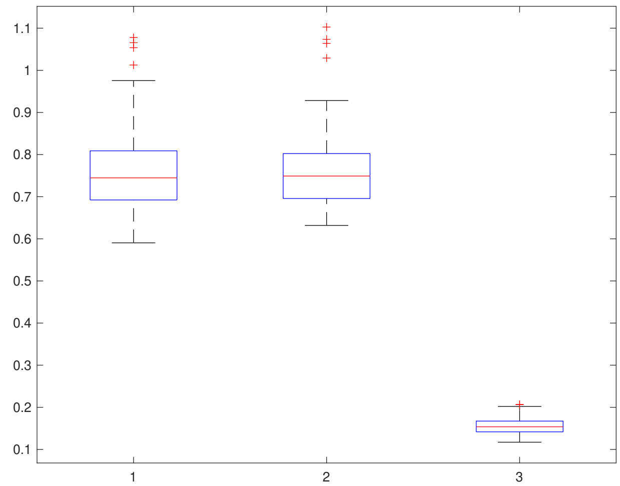

Figure 2: Consider HSBM$(d, n, \alpha f_n, \beta f_n)$ for $d = 4$, $n = 1000$, $\alpha = 50$, $\beta = 10$ and $f_n = \log n / \binom{n-1}{d-1}$. The boxplots show three different errors over 100 realizations: (1)$\sqrt{n}\, \|u_2 - u_2^*\|_\infty$, (2)$\sqrt{n}\|u_2^* - W u_2^*/\lambda_2^*\|_\infty$, and (3) $\sqrt{n}\|u_2 - W u_2^*/\lambda_2^*\|_\infty$.

bounds on each entry of u_2. A natural strategy would be to compare u_2 to u_2^*, since $u_2^* = \frac{1}{\sqrt{n}}\sigma^*$ due to the block structure of W^*. Unfortunately, $\|u_2 - u_2^*\|_\infty$ is too large for our purposes, but still u_2 recovers communities by sign. To gain intuition for this behavior, write

$$u_2 - u_2^* = \left(\frac{W u_2^*}{\lambda_2^*} - u_2^*\right) + \left(u_2 - \frac{W u_2^*}{\lambda_2^*}\right).$$

The first term on the right-hand side is the main term, while the second represents a smaller-order term (see Figure 2). Such behavior was also observed in the SBM setting by Abbe et al. (2020).

Our first step is to apply Theorem 7, showing that

$$\min_{s^* \in \{\pm 1\}} \left\| u_2 - s^* \frac{W u_2^*}{\lambda_2^*} \right\|_\infty = o(1/\sqrt{n}) \tag{8}$$

(see Corollary 13). Therefore, if we can show that the vector $\frac{W u_2^*}{\lambda_2^*}$ has the same signs as σ^* (up to a global sign flip), then the same is true for u_2. Our goal is then to show that σ^* and $W u_2^*/\lambda_2^*$ have the same signs, i.e.

$$\min_{i \in [n]} \sigma^*(i) \left(\frac{W u_2^*}{\lambda_2^*}\right)_i > 0.$$

Fixing the orientation $u_2^* = \frac{1}{\sqrt{n}}\sigma^*$, we obtain that for $i \in [n]$,

$$\sigma^*(i) \left(\frac{W u_2^*}{\lambda_2^*}\right)_i = \frac{\sigma^*(i) \sum_{j \in [n]} W_{ij} \sigma^*(j)}{\lambda_2^* \sqrt{n}} = \frac{D_{ii}}{\lambda_2^* \sqrt{n}},$$

where D_{ii} is defined in (5).

In the logarithmic degree regime, we have that $\lambda_2^* = \Theta(\log n)$. Moreover, as in the SDP analysis, we note that for $\epsilon > 0$ sufficiently small, $\min_{i\in[n]} D_{ii} \geq \epsilon \log n$ with high probability, whenever $I(d, \alpha, \beta) > 1$. Therefore, the vector $\frac{Wu_2^*}{\lambda_2^*}$ has the same signs as σ^*. Moreover, the entries are of order $1/\sqrt{n}$; in turn, (8) implies that u_2 also has the same signs as σ^*, up to a global sign flip, implying Theorem 4. Similarly, when $f_n = \omega(\log n/n^{d-1})$, the we have that $\lambda_2^* = \Theta(n^{d-1} f_n)$. In this case, using a Chernoff-style bound, we establish the existence of $\epsilon > 0$ such that $\min_{i\in[n]} D_{ii} \geq \epsilon\, n^{d-1} f_n$ with high probability for all $\alpha > \beta > 0$, giving us Theorem 5.

Runtime analysis. Observe that Step 2 of Algorithm 2 requires only $O(n)$ time. The bottleneck is the time required for Step 1 (to compute the second eigenvector of W). In order to compute the eigenpairs, one can use the power method on the matrix W, which computes the top eigenpair (λ_1, u_1) first. The method converges in $O(\log(n)/\delta)$ iterations (see Garber et al. (2016)), where $\delta = \frac{\lambda_1 - \lambda_2}{\lambda_1}$ is the relative eigengap. In our setting, both $\lambda_1 - \lambda_2$ and λ_1 are $\Theta(n^{d-1} f_n)$. This is because the eigengap $\Delta_1^* = \lambda_1^* - \lambda_2^*$ of W^* is $\Theta(n^{d-1} f_n)$ and $\|W - W^*\|_2 = o(n^{d-1} f_n)$ with high probability using Lee et al. (2020). We conclude that $\delta = \Theta(1)$ and the power method converges in $O(\log n)$ iterations. In each iteration, it needs to multiply a vector with W. The cost of this operation depends on the sparsity of W.

In the logarithmic degree regime, we have $O(n \log n)$ edges in the hypergraph, and thus, W has at most $O(n \log n)$ non-zero entries, which is the effective cost of a matrix-vector multiplication. Therefore, the total time to compute the first eigenpair (λ_1, u_1) of W is $O(n \log^2 n)$. To obtain the second eigenvector, we can deflate W by subtracting $\lambda_1 u_1 u_1^\top$. The new relative eigengap is $\frac{\lambda_2 - \lambda_3}{\lambda_2} = \Theta(1)$ since the other eigenvalues of W are close to 0. We emphasize that, even though $W - \lambda_1 u_1 u_1^\top$ may have n^2 non-zero entries, a matrix-vector multiplication can still be done in $O(n \log n)$ time by multiplying the vector with W and $\lambda_1 u_1 u_1^\top$ separately and then taking the difference. Thus, the power method requires $O(n \log^2 n)$ time to obtain the second eigenvector.

In super-logarithmic degree regimes, a matrix-vector multiplication may take up to $O(n^2)$ time. Therefore, the total runtime is $O(n^2 \log n)$ in the worst-case.

Entrwise eigenvector analysis. Abbe et al. (2020) introduced a powerful entrywise eigenvector bound, which has been used to show the optimality of spectral algorithms without the need of a clean-up stage (Abbe et al., 2020; Dhara et al., 2022a,b,c). Unfortunately, the entrywise bound (Abbe et al., 2020, Theorem 2.1) does not apply to $W = \mathcal{S}(G)$, since W violates a certain independence assumption. The independence assumption is critically used in a *leave-one-out* argument; we therefore carefully adapt this step. It also remains to prove a certain *row concentration* property of the matrix W.

For simplicity, let $\lambda = \lambda_k$, $\lambda^* = \lambda_k^*$ and $\Delta_k^* = \Delta^*$. For clarity of presentation, we assume the orientation $\langle u, u^* \rangle \geq 0$, and make similar simplifying assumptions throughout the outline. Our goal is then to show

$$\left\| u - \frac{Wu^*}{\lambda^*} \right\|_\infty \lesssim \frac{\|u^*\|_\infty}{\log \log n}.$$

We first relate λ to λ^*. By Weyl's inequality, $|\lambda - \lambda^*| \leq \|W - W^*\|_2$. In turn, (Lee et al., 2020, Theorem 4) implies $\|W - W^*\|_2 \leq \gamma \lambda^*$ with high probability, for certain $\gamma = \gamma_n = o(1)$. It then follows that $|\lambda^{-1} - \lambda^{*-1}| \lesssim \gamma |\lambda^*|^{-1}$. Using this observation along with the triangle inequality, we can show

$$\begin{aligned}\left\|u - \frac{Wu^*}{\lambda^*}\right\|_\infty &= \left\|\frac{Wu}{\lambda} - \frac{Wu^*}{\lambda} + \frac{Wu^*}{\lambda} - \frac{Wu^*}{\lambda^*}\right\|_\infty \le \left|\frac{1}{\lambda} - \frac{1}{\lambda^*}\right| \|Wu^*\|_\infty + \frac{1}{|\lambda|}\|W(u-u^*)\|_\infty \\ &\lesssim \frac{1}{|\lambda^*|}\left(\gamma \|Wu^*\|_\infty + \|W(u-u^*)\|_\infty\right).\end{aligned} \tag{9}$$

Note that u^* is a deterministic vector. Therefore, in order to bound the term $\|Wu^*\|_\infty$, we derive a row concentration result (Lemma 15). For a fixed vector $v \in \mathbb{R}^n$, our row concentration result controls $\|Wv\|_\infty$ in terms of both $\|v\|_\infty$ and $\|v\|_2$.

Since u depends on W, the second term in (9) requires a different strategy than the first. We therefore apply the leave-one-out technique, motivated by other works using a similar strategy (Bean et al., 2013; Javanmard et al., 2016; Zhong and Boumal, 2018; Abbe et al., 2020). Bounding $\|W(u-u^*)\|_\infty$ reduces to bounding $|[W(u-u^*)]_m|$ for each $m \in [n]$. To this end, we fix $m \in [n]$ and define a random matrix $W^{(m)}$ which is independent of the m-th row and column of W. Let $G^{(m)}$ be the hypergraph formed from G by deleting all hyperedges containing m, and let $W^{(m)} = \mathcal{S}(G^{(m)})$ be its similarity matrix. Let $u^{(m)}$ be the k-th eigenvector of $W^{(m)}$. Applying the triangle and Cauchy–Schwarz inequalities, we obtain

$$\begin{aligned}|[W(u-u^*)]_m| &\le \left|W_{m\cdot}(u-u^{(m)})\right| + \left|W_{m\cdot}(u^{(m)}-u^*)\right| \\ &\le \|W\|_{2\to\infty}\|u-u^{(m)}\|_2 + |W_{m\cdot}(u^{(m)}-u^*)|.\end{aligned} \tag{10}$$

Observe that $W_{m\cdot}$ and $u^{(m)} - u^*$ are independent by the leave-one-out construction. Therefore, we can bound the second term in (10) using our row concentration result. In order to bound $\|u-u^{(m)}\|_2$, we apply a version of the Davis–Kahan $\sin(\theta)$ theorem (Davis and Kahan, 1970), which yields

$$\|u-u^{(m)}\|_2 \lesssim \frac{\|(W-W^{(m)})u\|_2}{\Delta^*}.$$

Our analysis so far is almost identical to that of Abbe et al. (2020); the main difference arises here. Note that the m-th row and column $W - W^{(m)}$ are the same as those of W. Since a hyperedge containing the vertex m contributes to other entries of W, there will be additional non-zero entries outside of the m-th row and column. Thus it is harder to find tight bounds on the quantities that essentially are of main interest: $\|W - W^{(m)}\|_2$ and $\|(W-W^{(m)})u\|_2$. This complication arises whenever $d \ge 3$, and is absent from the analysis of Abbe et al. (2020), due to their independence assumption. However, we are still able to prove similar probabilistic bounds for these quantities via a series of careful (and non-trivial) lemmas that uses structural properties of the similarity matrix along with the spectral norm concentration of Lee et al. (2020) to yield the final sharp bounds.

5. Discussion and future work

This paper contributes to a line of work which studies random graph inference problems under restricted information. This paper considers *aggregated* information through the similarity matrix transformation. Other recent works consider *noisy* or *censored* (Abbe et al., 2014; Hajek et al., 2016b; Dhara et al., 2022a,b,c)] information models. While the similarity matrix is a lossy representation of the full hypergraph information, it retains most of the information about the *latent* community structure. That is, the similarity matrix is sufficient for exact recovery in all denser regimes, and even in the logarithmic degree regime, up to the threshold $I(d, \alpha, \beta) = 1$ which is slightly worse than the threshold given the full information (see Figure 1). In Appendix A, we further investigate the recovery problem given the adjacency matrix $A \in \{0,1\}^{n \times n}$, where $A_{ij} = 1$ if i, j appear together in some hyperedge. In the logarithmic degree regime, our results suggest that the adjacency matrix preserves much of the latent community structure, while in higher degree regimes the adjacency matrix becomes uninformative.

Our work shows that, in the logarithmic degree regime, the spectral algorithm recovers communities from the similarity matrix whenever the MLE does. Abbe et al. (2020), who observed the same phenomenon both in the SBM exact recovery and $\mathbb{Z}_2$-synchronization problems, noted that such a phenomenon may be more general. Indeed, our result adds to a growing list of examples where a spectral algorithm matches the MLE (Abbe et al., 2020; Deng et al., 2021; Dhara et al., 2022c,b,a). All of these results crucially rely on entrywise eigenvector bounds. Moreover, we note that a sharp phase transition occurs at $I(d, \alpha, \beta) = 1$, reiterating a common theme in community detection problems. In particular, the spectral algorithm succeeds with high probability if $I(d, \alpha, \beta) > 1$, while any algorithm fails with high probability if $I(d, \alpha, \beta) < 1$.

Finally, the spectral algorithm is advantageous even when the full information is given, due to its computational efficiency compared to existing approaches. To our knowledge, this constitutes the first algorithm with a nearly linear runtime for exact recovery in the HSBM.

We include some directions for future work.

1. Given that our spectral algorithm is highly efficient, can we generalize this approach beyond the binary symmetric communities case? Another such widely studied model is the Planted Dense Subgraph (PDS) (Hajek et al., 2016a; Dhara et al., 2022b), which has a natural hypergraph variant. It can be verified that the entrywise guarantees of eigenvectors (Theorem 7) apply. Due to the non-symmetric nature of the problem, we believe that taking some linear combination of the two leading eigenvectors would lead to an *optimal* spectral algorithm.

2. We show that the SDP is robust to a monotone adversary who operates on the similarity matrix. What can we say about an adversary who instead operates on the hypergraph directly, by adding intra-community edges and removing inter-community edges?

3. Can we find a simple proof for the sharp concentration $\|W - W^*\|_2$, by leveraging our bound on $\mathbb{E}\left[\|W - W^*\|_2\right]$ and using a concentration argument? While Talagrand's inequality leads to a sharp bound in the $d = 2$ case (Hajek et al., 2016a), it becomes vacuous in the $d \geq 3$ case.

4. What is the precise recovery threshold given the adjacency matrix A, in the logarithmic degree regime? Does the same spectral strategy (of recovering communities based on the signs of the entries of the second eigenvector of A) achieve the information-theoretic limit?

Acknowledgments

J.G. was partially supported by NSF CCF-2154100. N.J. was partially supported by NSF CAREER Award 1652491. The authors thank the anonymous reviewers for their helpful feedback.

References

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

Emmanuel Abbe and Colin Sandon. 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*, pages 670–688, 2015a. doi: 10.1109/FOCS.2015.47.

Emmanuel Abbe and Colin Sandon. Recovering communities in the general stochastic block model without knowing the parameters. In *Advances in Neural Information Processing Systems*, volume 28, 2015b.

Emmanuel Abbe, Afonso S Bandeira, Annina Bracher, and Amit Singer. Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery. *IEEE Transactions on Network Science and Engineering*, 1(1):10–22, 2014.

Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on information theory*, 62(1):471–487, 2015.

Emmanuel Abbe, Afonso S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62:471–487, 2016.

Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics*, 48 3:1452–1474, 2020.

Kwangjun Ahn, Kangwook Lee, and Changho Suh. Hypergraph spectral clustering in the weighted stochastic block model. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):959–974, 2018.

Kalle Alaluusua, Konstantin Avrachenkov, BR Vinay Kumar, and Lasse Leskelä. Multilayer hypergraph clustering using the aggregate similarity matrix. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 83–98. Springer, 2023.

Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. *Random Structures & Algorithms*, 13(3-4):457–466, 1998.

Maria Chiara Angelini, Francesco Caltagirone, Florent Krzakala, and Lenka Zdeborová. Spectral detection on sparse hypergraphs. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 66–73. IEEE, 2015.

Brian Ball, Brian Karrer, and M. E. J. Newman. Efficient and principled method for detecting communities in networks. *Phys. Rev. E*, 84:036103, Sep 2011. doi: 10.1103/PhysRevE.84.036103.

Derek Bean, Peter J. Bickel, Noureddine El Karoui, and Bin Yu. Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568, 2013. doi: 10.1073/pnas.1307845110.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Jingchun Chen and Bo Yuan. Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290, 07 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl370.

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.

I Eli Chien, Chung-Yi Lin, and I-Hsiang Wang. On the minimax misclassification ratio of hypergraph community detection. *IEEE Transactions on Information Theory*, 65(12):8095–8118, 2019.

Sam Cole and Yizhe Zhu. Exact recovery in the hypergraph stochastic block model: A spectral algorithm. *Linear Algebra and its Applications*, 593:45–73, 2020.

Chandler Davis and William Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7:1–46, 1970.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

Shaofeng Deng, Shuyang Ling, and Thomas Strohmer. Strong consistency, graph Laplacians, and the stochastic block model. *J. Mach. Learn. Res.*, 22(117):1–44, 2021.

Souvik Dhara, Julia Gaudio, Elchanan Mossel, and Colin Sandon. Spectral recovery of binary censored block models. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3389–3416. SIAM, 2022a.

Souvik Dhara, Julia Gaudio, Elchanan Mossel, and Colin Sandon. Spectral algorithms optimally recover (censored) planted dense subgraphs. *arXiv preprint arXiv:2203.11847*, 2022b.

Souvik Dhara, Julia Gaudio, Elchanan Mossel, and Colin Sandon. The power of two matrices in spectral algorithms. *arXiv preprint arXiv:2210.05893*, 2022c.

Ioana Dumitriu and Haixiao Wang. Exact recovery for the non-uniform hypergraph stochastic block model. *arXiv preprint arXiv:2304.13139*, 2023.

Ioana Dumitriu, Haixiao Wang, and Yizhe Zhu. Partial recovery and weak consistency in the non-uniform hypergraph stochastic block model. *arXiv preprint arXiv:2112.11671*, 2021.

Martin E. Dyer and Alan M. Frieze. The solution of some random NP-hard problems in polynomial expected time. *Journal of Algorithms*, 10(4):451–489, 1989.

Uriel Feige and Joe Kilian. Heuristics for semirandom graph problems. *Journal of Computer and System Sciences*, 63(4):639–671, 2001.

Uriel Feige and Robert Krauthgamer. Finding and certifying a large hidden clique in a semirandom graph. *Random Structures & Algorithms*, 16(2):195–208, 2000.

Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

Dan Garber, Elad Hazan, Chi Jin, Cameron Musco, Praneeth Netrapalli, Aaron Sidford, et al. Faster eigenvector computation via shift-and-invert preconditioning. In *International Conference on Machine Learning*, pages 2626–2634. PMLR, 2016.

Debarghya Ghoshdastidar and Ambedkar Dukkipati. Consistency of spectral partitioning of uniform hypergraphs under planted partition model. *Advances in Neural Information Processing Systems*, 27, 2014.

Debarghya Ghoshdastidar and Ambedkar Dukkipati. A provable generalized tensor spectral method for uniform hypergraph partitioning. In *International Conference on Machine Learning*, pages 400–409. PMLR, 2015a.

Debarghya Ghoshdastidar and Ambedkar Dukkipati. Spectral clustering using multilinear SVD: Analysis, approximations and applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015b.

Debarghya Ghoshdastidar and Ambedkar Dukkipati. Consistency of spectral hypergraph partitioning under planted partition model. *The Annals of Statistics*, 45(1):289–315, 2017.

Anna Goldenberg, Alice X. Zheng, Stephen E. Fienberg, and Edoardo M. Airoldi. A survey of statistical network models. *Found. Trends Mach. Learn.*, 2(2):129–233, feb 2010. ISSN 1935-8237. doi: 10.1561/2200000005.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5):2788–2797, 2016a.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *IEEE Transactions on Information Theory*, 62(10):5918–5937, 2016b.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Adel Javanmard, Andrea Montanari, and Federico Ricci-Tersenghi. Phase transitions in semidefinite relaxations. *Proceedings of the National Academy of Sciences*, 113(16):E2218–E2223, 2016. doi: 10.1073/pnas.1523097113.

Haotian Jiang, Tarun Kathuria, Yin Tat Lee, Swati Padmanabhan, and Zhao Song. A faster interior point method for semidefinite programming. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 910–918. IEEE, 2020.

Zheng Tracy Ke, Feng Shi, and Dong Xia. Community detection for hypergraph networks via regularized tensor power iteration. *arXiv preprint arXiv:1909.06503*, 2019.

Chiheon Kim, Afonso S Bandeira, and Michel X Goemans. Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach. *arXiv preprint arXiv:1807.02884*, 2018.

Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the web for emerging cyber-communities. *Computer Networks*, 31(11):1481–1493, 1999. ISSN 1389-1286. doi: https://doi.org/10.1016/S1389-1286(99)00040-7.

Jeonghwan Lee, Daesung Kim, and Hye Won Chung. Robust hypergraph clustering via convex relaxation of truncated MLE. *IEEE Journal on Selected Areas in Information Theory*, 1(3):613–631, 2020.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

Edward M. Marcotte, Matteo Pellegrini, Ho-Leung Ng, Danny W. Rice, Todd O. Yeates, and David Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, 1999. doi: 10.1126/science.285.5428.751.

Laurent Massoulié. Community detection thresholds and the weak Ramanujan property. In *Symposium on Theory of Computing, STOC 2014*, page 694–703, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327107. doi: 10.1145/2591796.2591857.

Frank McSherry. Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 529–537. IEEE, 2001.

Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture, 2013.

M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proc. Natl. Acad. Sci. USA*, 99:2566–2572, 2002.

Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4), aug 2011. doi: 10.1214/11-aos887.

Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. doi: 10.1109/34.868688.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17:395–416, 2007.

Haixiao Wang. Strong consistency and optimality of spectral clustering in symmetric binary non-uniform hypergraph stochastic block model. *arXiv preprint arXiv:2306.06845*, 2023.

Jinxin Wang, Yuen-Man Pun, Xiaolu Wang, Peng Wang, and Anthony Man-Cho So. Projected tensor power method for hypergraph community recovery. *International Conference on Machine Learning*, 2023.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102:315–323, 2014.

Erchuan Zhang, David Suter, Giang Truong, and Syed Zulqarnain Gilani. Sparse hypergraph community detection thresholds in stochastic block model. *Advances in Neural Information Processing Systems*, 35:34012–34023, 2022.

Qiaosheng Zhang and Vincent YF Tan. Exact recovery in the general hypergraph stochastic block model. *IEEE Transactions on Information Theory*, 69(1):453–471, 2022.

Yaoming Zhen and Junhui Wang. Community detection in general hypergraph via graph embedding. *Journal of the American Statistical Association*, pages 1–10, 2022.

Yiqiao Zhong and Nicolas Boumal. Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization*, 28(2):989–1016, 2018. doi: 10.1137/17M1122025.

Appendix A. Similarity Matrix vs. Adjacency Matrix Information Models

In this section, we study the exact recovery problem given only the adjacency matrix $A \in \{0,1\}^{n\times n}$, where $A_{ij} = 1$ whenever $i, j \in [n]$ appear together in some hyperedge in G. For example, an academic collaboration network can be modeled as a hypergraph, where each hyperedge represents the authors of a paper. However, academic collaboration networks typically only record pairwise information, describing whether two researchers have co-authored a paper together–motivating the need to study this problem.

Logarithmic degree regime. We first consider $f_n = \log n / \binom{n-1}{d-1}$. We conjecture that the exact recovery problem given A exhibits a sharp recovery threshold, analogously to the exact recovery problem given W. Additionally, we expect the spectral algorithm that recovers communities based on the second eigenvector of A to be optimal. In Figure 3, we present our empirical findings from simulations.

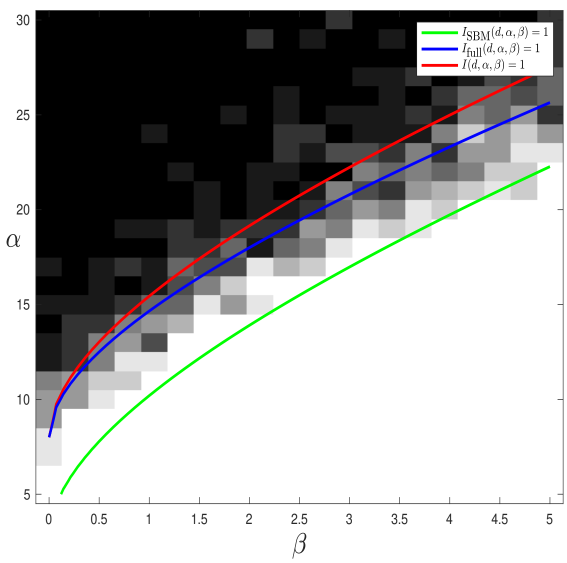

Figure 3: Visualizing the heat map of success of spectral algorithm on A alongside $I(d, \alpha, \beta) > 1$.

Fix $d = 4$ and $n = 500$. Let $G \sim \mathrm{HSBM}(d, n, \alpha f_n, \beta f_n)$ for different values of (α, β) and $W = \mathcal{S}(G)$. Then, the adjacent matrix $A = \min\{W, \mathbf{1}\mathbf{1}^\top\}$. We run the spectral algorithm for various values $\alpha > \beta$. We report the proportion of success, namely $\hat{\sigma} \in \{\pm\sigma^*\}$, out of 30 independent runs. Darker pixels represent higher chances of success in the heat map. We juxtapose this with the thresholds $I(d, \alpha, \beta) = 1$ and $I_{\text{full}}(d, \alpha, \beta) = 1$. We find all three of them to be relatively close to each other, which suggests that the matrix A still continues to retain much of the information about the ground truth community structure.

Since A is an adjacency matrix, it is also natural to compare it to an SBM G' which produces an adjacency matrix A' with the same marginal edge probabilities. Of course, A has a complicated dependency structure, while the entries of A' are independent, conditioned on the community structure. We plot the recovery threshold for the SBM in green. More precisely, we transform the parameters (α, β) to corresponding SBM parameters (α', β') by solving the following. The intra-community probability

$$p'_n = \frac{\alpha' \log n}{n} = 1 - (1 - \alpha f_n)^{\binom{n/2-2}{d-2}} (1 - \beta f_n)^{\binom{n-2}{d-2} - \binom{n/2-2}{d-2}},$$

and inter-community probability

$$q'_n = \frac{\beta' \log n}{n} = 1 - (1 - \beta f_n)^{\binom{n-2}{d-2}}.$$

This gives us α' and β' as functions of α, β and d. The line plots $I_{\text{SBM}}(d, \alpha, \beta) = (\sqrt{\alpha'} - \sqrt{\beta'})^2/2 = 1$ on the (α, β) plane. This line falls below the recovery threshold corresponding to the full hypergraph information, which implies that the threshold corresponding to recovery from A is strictly higher than the threshold corresponding to recovery from A'. In other words, even though A and A' have the same marginal entry distributions, the dependency structure in A makes recovery strictly harder.

From a memory and runtime standpoint, recovery using A or W is asymptotically equivalent in the logarithmic degree regime, since with high probability each entry of W is upper bounded by $4d = \Theta(1)$ (see Section D, Equation 41). On the other hand, there are regimes in which the similarity matrix approach succeeds while the adjacency matrix approach fails, which we discuss below.

Denser regimes. Fix d and $\alpha > \beta > 0$. Consider regimes such that $f_n = \omega(\log n / n^{d-2})$ and $G \sim (d, n, \alpha f_n, \beta f_n)$. For each pair $i, j \in [n]$, the number of edges in involving (i, j) grows with n. Therefore, with high probability, the adjacency matrix A is simply the matrix of all ones (up to the diagonal). More precisely, $\mathbb{P}\left(A = (\mathbf{1}\mathbf{1}^\top - \mathbf{I})\right) = 1 - o(1)$, rendering the adjacency matrix uninformative.

In these regimes, the adjacency matrix information model (or its variant where $A = \min\{W, c\mathbf{1}\mathbf{1}^\top\}$ for any constant $c > 0$) requires only up to $O(n^2)$ space to maintain, but it does not preserve any information. On the other hand, community recovery given the similarity matrix is still possible (Theorem 5).

Appendix B. Proofs from Section 3: Optimality of the SDP Relaxation

We begin by noting a simple corollary of our spectral norm concentration theorem (Theorem 10) in the context of the HSBM.

Corollary 11 *Fix $d \in \{2, 3, \dots\}$ and $\alpha > \beta > 0$. Let $f_n = \log n / \binom{n-1}{d-1}$ and $G \sim \text{HSBM}(d, n, \alpha f_n, \beta f_n)$. Let $W = \mathcal{S}(G)$ and $W^* = \mathbb{E}[W \mid \sigma^*]$. Then*

$$\mathbb{P}\left(\|W - W^*\|_2 \leq \log^{3/4} n\right) \geq 1 - o(1).$$

Proof By Theorem 10, $\mathbb{E}[\|W - W^*\|_2] \leq c\sqrt{\log n}$, for some $c > 0$ that depends on α and d. Therefore, using Markov's inequality,

$$\mathbb{P}\left(\|W - W^*\|_2 > \log^{3/4} n\right) \leq \frac{\mathbb{E}[\|W - W^*\|_2]}{\log^{3/4} n} \leq \frac{c\sqrt{\log n}}{\log^{3/4} n} = o(1).$$

■

We now show Lemma 9, which characterizes sufficient conditions under which $X^* = \sigma^* {\sigma^*}^\top$ is the unique solution of the SDP (3), using a dual certificate strategy.

Proof [Proof of Lemma 9] We first show that $X^* = \sigma^* {\sigma^*}^\top$ is an optimal solution. For any X satisfying the constraints in (3),

$$\begin{aligned}
\langle W, X\rangle &\leq \langle W, X\rangle + \langle S, X\rangle && \text{(since } S, X \succeq 0\text{)}\\
&= \langle W, X\rangle + \langle D + \nu \mathbf{1}\mathbf{1}^\top - W, X\rangle = \langle D, X\rangle && \text{(since } X \text{ is primal feasible)}\\
&= \langle D, X^*\rangle && \text{(as } X^*_{ii} = X_{ii} = 1 \text{ for all } i \in [n]\text{)}\\
&= \langle W + S - \nu \mathbf{1}\mathbf{1}^\top, X^*\rangle = \langle W, X^*\rangle + \langle S, X^*\rangle = \langle W, X^*\rangle. \\
& && \text{(since } \langle S, X^*\rangle = (\sigma^*)^\top S \sigma^* = 0\text{)}
\end{aligned}$$

Therefore it only remains to establish the uniqueness of X^*. Consider an optimal solution $\tilde{X}$. Then

$$\begin{aligned}
\langle S, \tilde{X}\rangle = \langle D + \nu \mathbf{1}\mathbf{1}^\top - W, \tilde{X}\rangle &= \langle D - W, \tilde{X}\rangle && \text{(since } \langle \mathbf{1}\mathbf{1}^\top, \tilde{X}\rangle = 0\text{)}\\
&= \langle D - W, X^*\rangle && \text{(as } \langle W, X^*\rangle = \langle W, \tilde{X}\rangle \text{ and } X^*_{ii} = \tilde{X}_{ii} = 1\text{)}\\
&= \langle S, X^*\rangle = 0. && \text{(using } S\sigma^* = 0\text{)}
\end{aligned}$$

Since $\tilde{X} \succeq 0$ and $S \succeq 0$ and $\lambda_{n-1}(S) > 0$, we obtain that $\tilde{X}$ is also a rank-1 matrix and hence it must be a multiple of $\sigma^* {\sigma^*}^\top$. Moreover, as $\tilde{X}_{ii} = 1$ for all $i \in [n]$, it must be that $\tilde{X} = X^* = \sigma^* {\sigma^*}^\top$. ■

We now focus on showing that the choice of D mentioned in (5) and $\nu = 1$ satisfy the conditions in Lemma 9 with high probability whenever $I(d, \alpha, \beta) > 1$ to show Theorem 2. Towards this, we prove a lemma that plays an important role in proving the theorem. Roughly speaking, it provides a probabilistic lower bound on D_{ii} defined in (5) for any $i \in [n]$.

Lemma 12 *Let $d \in \{2, 3, \dots\}$, and $\alpha > \beta > 0$, such that $I(d, \alpha, \beta) > 1$. Let $f_n = \log n / \binom{n-1}{d-1}$ and $W = \mathcal{S}(G)$ where $G \sim \mathrm{HSBM}(d, n, \alpha f_n, \beta f_n)$. Then there exists a constant $\epsilon := \epsilon(d, \alpha, \beta) > 0$ such that for any fixed $i \in [n]$, with probability at least $1 - o(n^{-1})$,*

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

Proof Fix $i \in [n]$, and let $X \triangleq \sum_{j \in [n]} W_{ij} \sigma^*(i) \sigma^*(j)$. Let $\mathcal{E} := \binom{[n]}{d}$ be the set of possible edges. For each $e \in \mathcal{E}$, let A_e be the indicator that edge e is present. Let $\mathcal{E}^{(i)} := \{e \in \mathcal{E} : i \in e\}$ represent the set of potential edges incident on i.

For any edge $e \in \mathcal{E}^{(i)}$, let $n_i(e) := |\{j \in e \setminus \{i\} : \sigma^*(i) \neq \sigma^*(j)\}|$ be the number of vertices in e that belong to the opposite community as i. We can rewrite X as follows

$$\begin{aligned}
X &= \sum_{e \in \mathcal{E}^{(i)}} \sum_{j \in e, j \neq i} \sigma^*(i) \sigma^*(j) A_e \\
&= \sum_{e \in \mathcal{E}^{(i)}} \left((d - 1 - n_i(e)) - n_i(e)\right) A_e \\
&= \sum_{e \in \mathcal{E}^{(i)}} \left(d - 1 - 2n_i(e)\right) A_e.
\end{aligned}$$

Next, observe that for $r \in \{0, 1, \ldots, d-1\}$, the set $\{e \in \mathcal{E}^{(i)} : n_i(e) = r\}$ has cardinality

$$N_r := \binom{n/2}{r}\binom{n/2-1}{d-1-r}.$$

Let $\{Y_r\}_{r=0}^{d-1}$ be independent random variables, where $Y_r \sim \mathrm{Bin}(N_r, q_r)$, with

$$q_r = \begin{cases} \alpha \log n / \binom{n-1}{d-1}, & \text{if } r = 0; \\ \beta \log n / \binom{n-1}{d-1}, & \text{if } 1 \le r \le d-1. \end{cases}$$

We then further rewrite X as follows:

$$X = \sum_{r=0}^{d-1} \sum_{e \in \mathcal{E}^{(i)}} \mathbb{1}\{n_i(e) = r\}\,(d-1-2n_i(e))\,A_e,$$

so that X is equal to $\sum_{r=0}^{d-1}(d-1-2r)Y_r$ in distribution.

Let $h_r = d-1-2r$. Fix $\epsilon \in \mathbb{R}$ and $t \ge 0$. Exponentiating and applying Markov's inequality,

$$\begin{aligned} \mathbb{P}\left(X \le \epsilon \log n\right) &\le \mathbb{P}\left(e^{-tX} \ge e^{-t\epsilon \log n}\right) \\ &\le \frac{\mathbb{E}\left[e^{-tX}\right]}{e^{-t\epsilon\log n}} \\ &= e^{t\epsilon \log n}\,\mathbb{E}\left[e^{-t\sum_{r=0}^{d-1} h_r Y_r}\right] \\ &= e^{t\epsilon \log n}\prod_{r=0}^{d-1}\mathbb{E}\left[e^{-th_rY_r}\right] \\ &= e^{t\epsilon \log n}\prod_{r=0}^{d-1}\left(1 - q_r(1-e^{-th_r})\right)^{N_r} \\ &\le \exp\left(t\epsilon\log n - \sum_{r=0}^{d-1} N_r q_r (1-e^{-h_r t})\right). \end{aligned} \tag{11}$$

Here, the second equality is due to independence of the Y_r random variables, and the final step uses $1 - x \le e^{-x}$. Next,

$$\begin{aligned} N_r = \binom{n/2}{r}\binom{n/2-1}{d-1-r} &= \frac{\binom{n/2}{r}\binom{n/2-1}{d-1-r}}{\binom{n-1}{d-1}}\binom{n-1}{d-1} \\ &= (1+o(1))\frac{(n/2)^r}{r\,!}\cdot\frac{(n/2)^{d-1-r}}{(d-1-r)\,!}\cdot\frac{(d-1)!}{n^{d-1}}\cdot\binom{n-1}{d-1} \\ &= (1+o(1))\frac{1}{2^{d-1}}\binom{d-1}{r}\binom{n-1}{d-1}. \end{aligned}$$

Substituting into (11), we obtain

$$\mathbb{P}\left(X \le \epsilon \log n\right) \le \exp\left(t\epsilon\log n - \frac{(1+o(1))}{2^{d-1}}\left(\alpha(1-e^{-(d-1)t}) + \sum_{r=1}^{d-1}\beta\binom{d-1}{r}(1-e^{-(d-1-2r)t})\right)\log n\right).$$

Let $t = t^*(d, \alpha, \beta)$, where

$$t^*(d,\alpha,\beta) = \arg\max_{t \geq 0} \ \frac{1}{2^{d-1}} \left(\alpha(1 - e^{-(d-1)t}) + \sum_{r=1}^{d-1} \beta \binom{d-1}{r} (1 - e^{-(d-1-2r)t}) \right) := \arg\max_{t\geq 0} \psi(t).$$

We then obtain

$$\begin{aligned} \mathbb{P}\left(X \leq \epsilon \log n\right) &\leq \exp\bigl(t^* \epsilon \log n - I(d,\alpha,\beta)\log n + o(\log n)\bigr) \\ &\leq n^{-I(d,\alpha,\beta) + t^*\epsilon + o(1)}. \end{aligned}$$

Note that $t^* \neq 0$ as

$$\lim_{t \to 0^+} \psi'(t) = \frac{1}{2^{d-1}} \left(\alpha(d-1) + \beta \sum_{r=1}^{d-1} \binom{d-1}{r} (d-1-2r) \right) = (\alpha - \beta)(d-1)/2^{d-1} > 0.$$

Furthermore, since $I(d,\alpha,\beta) > 1$, one can choose $\epsilon = \epsilon(d,\alpha,\beta) > 0$ sufficiently small such that

$$\mathbb{P}\left(\sum_{j \in [n]} W_{ij} \sigma^*(i) \sigma^*(j) \leq \epsilon \log n \right) = o(n^{-1}).$$

■

Finally, we make some important observations about the structure of W^*. Observe that W^* has a *block* structure (up to the diagonal entries). In particular, W^* is a zero diagonal symmetric matrix whose non-diagonal entries are given by

$$W_{ij}^* = \begin{cases} p' \triangleq \binom{n/2-2}{d-2} \alpha f_n + \left(\binom{n-2}{d-2} - \binom{n/2-2}{d-2} \right) \beta f_n, & \text{if } \sigma^*(i) = \sigma^*(j); \\ q' \triangleq \binom{n-2}{d-2} \beta f_n, & \text{if } \sigma^*(i) \neq \sigma^*(j). \end{cases} \tag{12}$$

Observe that W^* can be decomposed as

$$W^* = \left(\frac{p' + q'}{2} \right) \mathbf{1}\mathbf{1}^\top + \left(\frac{p' - q'}{2} \right) \sigma^* {\sigma^*}^\top - p' \mathbf{I}. \tag{13}$$

Proof [Proof of Theorem 2] The proof uses ideas from the proof of (Hajek et al., 2016a, Theorem 2). Let $\nu = 1$ and $S \triangleq D + \nu \mathbf{1}\mathbf{1}^\top - W = D + \mathbf{1}\mathbf{1}^\top - W$. The goal is to show that S satisfies the conditions mentioned in Lemma 9 with high probability whenever $I(d,\alpha,\beta) > 1$. Observe that, by definition of D in (5), for any $i \in [n]$, we have $D_{ii}\sigma^*(i) = \sum_{j \in [n]} W_{ij} \sigma^*(j)$; i.e. $D\sigma^* = W\sigma^*$. Therefore, using the fact that $\langle \mathbf{1}, \sigma^* \rangle = 0$, we get

$$S\sigma^* = D\sigma^* + \mathbf{1}\mathbf{1}^\top \sigma^* - W\sigma^* = 0.$$

Therefore, it remains to show that

$$\mathbb{P}\left(\left\{ \inf_{x \perp \sigma^* : \|x\|_2 = 1} x^\top S x > 0 \right\} \right) \geq 1 - o(1).$$

For any $x \perp \sigma^*$ such that $\|x\|_2 = 1$,

$$\begin{aligned}
x^\top S x &= x^\top D x + x^\top \mathbf{1}\mathbf{1}^\top x - x^\top (W - W^*) x - x^\top W^* x \\
&= x^\top D x + (\mathbf{1}^\top x)^2 - x^\top (W - W^*) x - \left(\frac{p' - q'}{2}\right)(x^\top \sigma^*)^2 - \left(\frac{p' + q'}{2}\right)(\mathbf{1}^\top x)^2 + p' \\
&\qquad \text{(using (13) to substitute } W^*\text{)} \\
&= x^\top D x + \left(1 - (p' + q')/2\right)(\mathbf{1}^\top x)^2 - x^\top (W - W^*) x + p' \qquad \text{(since } x \perp \sigma^*\text{)} \\
&\geq x^\top D x + p' - x^\top (W - W^*) x \qquad \text{(since } p', q' = \Theta(\log n / n) \text{ are vanishing)} \\
&\geq \min_{i \in [n]} D_{ii} - \|W - W^*\|_2 . \qquad \text{(by the definition of } \|.\|_2 \text{ for matrices and the fact } p' \geq 0\text{)}
\end{aligned}$$

We now use Lemma 12 and take a union bound over i to obtain $\min_{i \in [n]} D_{ii} \geq \epsilon \log n$ with probability $1 - o(1)$. Moreover, applying Corollary 11, $\|W - W^*\|_2 \leq \log^{3/4} n$ with probability $1 - o(1)$. Therefore, one can conclude that $x^\top S x \geq \epsilon \log n - \log^{3/4} n > 0$ for any x such that $\|x\|_2 = 1$ and $x \perp \sigma^*$, completing the proof. ■

We additionally show that the SDP is robust to a monotone adversary (Lemma 3). Here, we consider an adversary who can increase the value of W_{ij} for any (i, j) in the same community, and decrease the value of W_{ij} for any (i, j) in opposite communities. The robustness of SDPs to monotone adversaries is well-known (see e.g. Feige and Krauthgamer (2000); Feige and Kilian (2001)). While the monotone adversary appears to provide helpful information, spectral algorithms generally fail under such a semirandom model.

Proof [Proof of Lemma 3] Let X be a feasible solution to the modified SDP, and let X^* be the unique optimal solution to (3), which is guaranteed by Theorem 2. Due to uniqueness of X^*, we have $\langle W, X \rangle < \langle W, X^* \rangle$. Since $X \succeq 0$, we can write

$$X = \sum_{l=1}^{n} \lambda_l v_l v_l^\top$$

as its eigendecomposition, where $\lambda_l \geq 0$ for all l. Then by the Cauchy–Schwarz inequality,

$$\begin{aligned}
X_{ij}^2 &= \left(\sum_{l=1}^{n} \lambda_l v_{l,i} v_{l,j}\right)^2 \\
&\leq \left(\sum_{l=1}^{n} \lambda_l v_{l,i}^2\right)\left(\sum_{l=1}^{n} \lambda_l v_{l,j}^2\right) \\
&= X_{ii} \cdot X_{jj} \\
&= 1.
\end{aligned}$$

Therefore, $|X_{ij}| \leq 1$ for all i, j, which implies

$$\langle \widetilde{W} - W, X \rangle \leq \sum_{i,j \in [n]} |\widetilde{W}_{ij} - W_{ij}| = \langle \widetilde{W} - W, X^* \rangle.$$

Consequently,

$$\langle \widetilde{W}, X \rangle = \langle W, X \rangle + \langle \widetilde{W} - W, X \rangle < \langle W, X^* \rangle + \langle \widetilde{W} - W, X^* \rangle = \langle \widetilde{W}, X^* \rangle,$$

establishing the unique optimality of X^*. ■

One could also consider another natural monotone adversary that operates directly on the underlying hypergraph instead of on its similarity matrix. More specifically, the adversary is allowed to add intra-community edges and delete cross-community edges. It is not immediately clear whether an SDP algorithm would be robust to such an adversary.

Appendix C. Proofs from Section 4: Analysis of the Spectral Algorithm

In Section C.1, we prove the optimality of the spectral algorithm in the logarithmic degree regime (Theorem 4). In Section C.2, we establish its correctness in denser regimes (Theorem 5). Finally, in Section C.3, we present the proof of our general entrywise eigenvector bound (Theorem 7).

C.1. Proving Theorem 4

We first derive a corollary of Theorem 7, specific to HSBMs.

Corollary 13 *Fix $d \in \{2, 3, \dots\}$. Choose any $\alpha > \beta > 0$ and f_n according to (1). Let $G \sim \mathrm{HSBM}(d, n, \alpha f_n, \beta f_n)$ and $W = \mathcal{S}(G)$. If $f_n = \Omega(\log n / n^{d-1})$, then with probability at least $1 - O(n^{-3})$,*

$$\min_{s^* \in \{\pm 1\}} \left\| u_2 - s^* \frac{W u_2^*}{\lambda_2^*} \right\|_\infty \leq \frac{c}{\sqrt{n} \log \log n},$$

where $c := c(d, \alpha, \beta)$ is some positive constant that only depends on d, α, and β.

Proof Since each hyperedge exists with probability at most αf_n, we can set μ_n as αf_n. Moreover, we have that $n\binom{n-1}{d-1}\mu_n \geq c_0 \log n$ when $f_n = \Omega(\log n / n^{d-1})$, for some c_0 that depends on d and α. We now verify that Assumption 1 holds. Let W^* be the expectation of W conditioned on σ^*, whose entries are p' and q' as given by (12). By (13), $W^* + p'\mathbf{I}$ is a rank-2 matrix. Its non-zero eigenvalues are $(p' + q')n/2$ and $(p' - q')n/2$. Accounting for the diagonal matrix $p'I$, the eigenvalues of W^* are given by

$$\lambda_1^* = (1 + o(1))(p' + q')n/2 \text{ and } \lambda_2^* = (1 + o(1))(p' - q')n/2.$$

Since $p' \asymp q' \asymp n^{d-2} f_n$, we have $\lambda_1^*, \lambda_2^* \asymp n^{d-1} f_n$. Furthermore,

$$\Delta_2^* = \min\{\lambda_1^* - \lambda_2^*, \lambda_2^* - 0\} = (1 + o(1)) \min\left\{q'n, \frac{(p' - q')n}{2}\right\} \asymp n^{d-1} f_n,$$

hiding constants in α, β and d. Therefore, Theorem 7 applies.

It remains to verify that $\|u_2^*\|_\infty = O(1/\sqrt{n})$. We already know that the second eigenvector of $W^* + p'\mathbf{I}$ is $\frac{1}{\sqrt{n}}\sigma^*$ by (13). Hence

$$u_2^* = \frac{1 + o(1)}{\sqrt{n}} \sigma^*. \tag{14}$$

■

We next prove the optimality of the spectral algorithm in the logarithmic degree regime (Theorem 4).

Proof [Proof of Theorem 4] Recall that $f_n = \log n / \binom{n-1}{d-1}$. Let us fix $s = s^*$ for which Corollary 13 holds. Using the corollary, with probability $1 - o(1)$,

$$\sqrt{n} \min_{i \in [n]} s\sigma^*(i) u_{2,i} \geq \sqrt{n} \min_{i \in [n]} s^2 \sigma^*(i)(Wu_2^*)_i / \lambda_2^* - c(\log\log n)^{-1}, \tag{15}$$

where c is defined in Corollary 13. Note that $s^2 = 1$. Also, using (14),

$$\sqrt{n}\sigma^*(i)(Wu_2^*)_i = (1 + o(1)) \sum_{j \in [n]} W_{ij}\sigma^*(i)\sigma^*(j).$$

By Lemma 12, if $I(d, \alpha, \beta) > 1$, then there exists a positive constant $\epsilon(d, \alpha, \beta) > 0$ such that for a fixed $i \in [n]$, $\sum_{j \in [n]} W_{ij}\sigma^*(i)\sigma^*(j) \geq \epsilon \log n$ with probability $1 - o(n^{-1})$. Therefore, a union bound implies that with probability $1 - o(1)$,

$$\sqrt{n} \min_{i \in [n]} s^2 \sigma^*(i)(Wu_2^*)_i \geq (1 + o(1))\epsilon \log n.$$

Since $\lambda_2^* \asymp \log(n)$ when $f_n = \log n / \binom{n-1}{d-1}$, (15) implies that there exists $\eta > 0$ such that

$$\sqrt{n} \min_{i \in [n]} s^* \sigma^*(i)(u_2)_i \geq (1 + o(1))\epsilon \log n / \lambda_2^* - c(\log\log n)^{-1} > \eta,$$

with probability $1 - o(1)$, concluding the proof. ∎

C.2. Proving Theorem 5

In this subsection, we prove the correctness of the spectral algorithm (Algorithm 2) in super-logarithmic degree regimes. By Corollary 13, we already know that the entrywise bounds hold in these regimes. Therefore, it remains to show that each entry of Wu_2^*/λ_2^* is sufficiently bounded away from zero with high probability. In order to achieve this, we show the following lemma, which is similar in spirit to Lemma 12 but also captures denser regimes.

Lemma 14 *Let $d \in \{2, 3, \dots\}$. Let p_n and q_n be parameterized according to (1) for some f_n and constants $\alpha > \beta > 0$. Let $W = \mathcal{S}(G)$ where $G \sim \text{HSBM}(d, n, \alpha f_n, \beta f_n)$. If $f_n = \omega(\log n / n^{d-1})$, then there exists a constant $\epsilon := \epsilon(d, \alpha, \beta) > 0$ such that for any fixed $i \in [n]$, with probability at least $1 - O(n^{-4})$,*

$$\sum_{j \in [n]} W_{ij}\sigma^*(i)\sigma^*(j) \geq \epsilon \cdot n^{d-1} f_n.$$

Proof Let $X \triangleq \sum_{j \in [n]} W_{ij}\sigma^*(i)\sigma^*(j)$. Define

$$N_r = \binom{n/2}{r}\binom{n/2 - 1}{d - 1 - r}$$
$$q_r = \begin{cases} \alpha f_n, & \text{if } r = 0 \\ \beta f_n, & \text{if } 1 \leq r \leq d - 1 \end{cases}$$
$$h_r = d - 1 - 2r.$$

Using identical steps as in the proof of Lemma 12, we can show

$$\mathbb{P}\left(X \leq \epsilon n^{d-1} f_n\right) \leq \exp\left(t\epsilon n^{d-1} f_n - \sum_{r=0}^{d-1} N_r q_r (1 - e^{-h_r t})\right). \tag{16}$$

Further analyzing the asymptotic behavior of N_r, we see that

$$N_r = (1 + o(1))\frac{1}{2^{d-1}}\binom{d-1}{r}\binom{n-1}{d-1} = (1 + o(1))\frac{\binom{d-1}{r} n^{d-1}}{2^{d-1}(d-1)!}.$$

Substituting into (16), we obtain

$$\begin{aligned}&\mathbb{P}\left(X \leq \epsilon\, n^{d-1} f_n\right)\\ &\leq \exp\left\{\left[t\epsilon - \frac{(1+o(1))}{2^{d-1}(d-1)!}\left(\alpha(1 - e^{-(d-1)t}) + \sum_{r=1}^{d-1} \beta\binom{d-1}{r}(1 - e^{-(d-1-2r)t})\right)\right] n^{d-1} f_n\right\}.\end{aligned} \tag{17}$$

Letting $t = t^\star(d, \alpha, \beta) > 0$ as in the proof of Lemma 12, we obtain

$$\mathbb{P}\left(X \leq \epsilon\, n^{d-1} f_n\right) \leq \exp\left(\left(t^* \epsilon - \frac{I(d, \alpha, \beta)}{(d-1)!} + o(1)\right) n^{d-1} f_n\right).$$

Therefore, for $\epsilon = \epsilon(d, \alpha, \beta)$ sufficiently small, there exists $\delta = \delta(d, \alpha, \beta) > 0$ such that

$$\mathbb{P}\left(X \leq \epsilon\, n^{d-1} f_n\right) \leq e^{-\delta\, n^{d-1} f_n}$$

for n sufficiently large. Finally, using $f_n = \omega(\log n / n^{d-1})$, we obtain

$$\mathbb{P}\left(\sum_{j \in [n]} W_{ij} \sigma^*(i) \sigma^*(j) \leq \epsilon\, n^{d-1} f_n\right) \leq e^{-\delta \cdot \omega(\log n)} \leq e^{-4 \log n} = O(n^{-4}).$$

■

We now combine this with entrywise bounds on the eigenvector u_2 in Corollary 13 to show our theorem.

Proof [Proof of Theorem 5] Since $f_n = \omega(\log n / n^{d-1})$, Corollary 13 holds for some $s^* \in \{\pm 1\}$. Fixing $s = s^*$, and using the corollary we get that with probability $1 - O(n^{-3})$,

$$\sqrt{n} \min_{i \in [n]} s\sigma^*(i) u_{2,i} \geq \sqrt{n} \min_{i \in [n]} s^2 \sigma^*(i) (W u_2^*)_i / \lambda_2^* - c(\log \log n)^{-1}, \tag{18}$$

where c is the constant from Corollary 13. As $s \in \{\pm 1\}$, we have that $s^2 = 1$. Also, using (14),

$$\sqrt{n}\sigma^*(i)(W u_2^*)_i = (1 + o(1)) \sum_{j \in [n]} W_{ij} \sigma^*(i) \sigma^*(j).$$

By Lemma 14, since $\alpha > \beta > 0$, there exists a positive constant $\epsilon(d, \alpha, \beta) > 0$ such that for a fixed $i \in [n]$, $\sum_{j \in [n]} W_{ij}\sigma^*(i)\sigma^*(j) \geq \epsilon\, n^{d-1} f_n$ with probability $1 - O(n^{-4})$. Therefore, taking a union bound, we obtain that with probability $1 - O(n^{-3})$,

$$\sqrt{n} \min_{i \in [n]} s^2 \sigma^*(i)(W u_2^*)_i \geq (1 + o(1))\epsilon\, n^{d-1} f_n.$$

Finally, note that $\lambda_2^* \asymp n^{d-1} f_n$. Therefore, (18) implies that with probability $1 - O(n^{-3})$

$$\sqrt{n} \min_{i \in [n]} s^* \sigma^*(i)(u_2)_i \geq (1 + o(1))\, \epsilon\, n^{d-1} f_n / \lambda_2^* - c(\log \log n)^{-1} > \eta,$$

for some $\eta(d, \alpha, \beta) > 0$, yielding the desired result. ■

C.3. Entrywise analysis

We begin by recalling the setup of Theorem 7. For simplicity, let $\lambda = \lambda_k$, $\lambda^* = \lambda_k^*$ and $\Delta^* = \Delta_k^*$, dropping the subscript k. Let $s = \mathrm{sgn}(\langle u, u^* \rangle)$, so that $\langle su, u^* \rangle \geq 0$. Also, for any fixed $m \in [n]$, let $G^{(m)}$ denote the hypergraph formed from G by deleting all the edges incident on m. Let $W^{(m)} = \mathcal{S}(G^{(m)})$, and let $(\lambda^{(m)}, u^{(m)})$ be th k-th eigenpair of $W^{(m)}$. Let $s^{(m)} = \mathrm{sgn}\langle u^{(m)}, u^* \rangle$, so that $\langle s^{(m)} u^{(m)}, u^* \rangle \geq 0$. The notation $\lesssim$ and $\asymp$ hide constants in d, c_0, and c_1 (defined as in Theorem 7) throughout this section. Before proving the theorem, we require some additional observations and results. We begin by making a simple observation about the deterministic matrix W^*.

Observation 1 $\|W^*\|_{2\to\infty} \leq \sqrt{n}\binom{n-2}{d-2}\mu_n$.

Proof Recall that $\max_e p_e \leq \mu_n$. Since each entry W_{ij} is a sum of $\binom{n-2}{d-2}$ Bernoulli random variables, $\mathbb{E}[W_{ij}] = W_{ij}^* \leq \binom{n-2}{d-2}\mu_n$. Therefore,

$$\|W^*\|_{2\to\infty} = \max_i \|W_{i\cdot}^*\|_2 \leq \sqrt{n} \max |W_{ij}^*| \leq \sqrt{n}\binom{n-2}{d-2}\mu_n.$$

■

Define the function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\varphi(x) = \frac{2 + 8d/c_0}{(1 \vee \log(1/x))}, \tag{19}$$

where c_0 is as in the statement of Theorem 7. We note the following properties of $\varphi(\cdot)$.

Observation 2 *$\varphi(x)$ is non-decreasing and $\varphi(x)/x$ is non-increasing on $\mathbb{R}_+$.*

The next result provides a probabilistic upper bound on the inner product of a row of $W - W^*$ and a fixed vector v, in terms of the function $\varphi(\cdot)$. Intuitively, one can think that the rate of growth of $\varphi(\cdot)$ essentially controls the strength of the concentration bound. Formally,

Lemma 15 (Row concentration) *For any $m \in [n]$ and any fixed non-zero $v \in \mathbb{R}^n$,*

$$\mathbb{P}\left(|(W - W^*)_{m\cdot} v| \leq \|v\|_\infty\, \varphi\left(\frac{\|v\|_2}{\sqrt{n}\,\|v\|_\infty}\right) n\binom{n-2}{d-2}\mu_n\right) \geq 1 - O\left(\frac{1}{n^4}\right).$$

Proof Recall the definition of $\mathcal{E}^{(m)} = \{e \in \mathcal{E} : m \in e\}$. Let $v \in \mathbb{R}^n$ be a fixed vector. Let $X \triangleq (W - W^*)_{m \cdot} v$. It is convenient to rewrite X as a sum of random variables $\{A_e\}_{e \in \mathcal{E}^{(m)}}$, where A_e is the indicator random variable associated with the hyperedge e:

$$X = \sum_{e \in \mathcal{E}^{(m)}} \left(\sum_{j \in e \setminus \{m\}} v_j \right) (A_e - \mathbb{E}[A_e]).$$

Without loss of generality, assume that $\|v\|_\infty = \frac{1}{d}$ (otherwise, v may be scaled). By Markov's inequality, for any $\delta, t > 0$,

$$\mathbb{P}\left(X \geq \delta\right) \leq \mathbb{P}\left(e^{tX} \geq e^{t\delta}\right) \leq e^{-t\delta} \prod_{e \in \mathcal{E}^{(m)}} \mathbb{E}\left(e^{t(\sum_{j \in e \setminus \{m\}} v_j)(A_e - \mathbb{E}[A_e])}\right). \tag{20}$$

For $e \in \mathcal{E}^{(m)}$, we can bound the logarithm of the moment generating function as follows:

$$\begin{aligned}
&\log\left(\mathbb{E}\left[e^{t(\sum_{j \in e \setminus \{m\}} v_j)(A_e - \mathbb{E}[A_e])}\right]\right) \\
&= \log\left(\mathbb{E}\left[e^{t(\sum_{j \in e \setminus \{m\}} v_j)A_e}\right]\right) - t\left(\sum_{j \in e \setminus \{m\}} v_j\right)\mathbb{E}[A_e] \\
&= \log\left(1 - p_e + p_e e^{t\sum_{j \in e \setminus \{m\}} v_j}\right) - tp_e\left(\sum_{j \in e \setminus \{m\}} v_j\right) \\
&\leq p_e\left(e^{t\sum_{j \in e \setminus \{m\}} v_j} - 1\right) - tp_e \sum_{j \in e \setminus \{m\}} v_j,
\end{aligned} \tag{21}$$

where we have used the fact that $\log(1+x) \leq x$ for $x > 1$ in the last step. Next, we use the fact that $e^x \leq 1 + x + \frac{x^2}{2}e^r$ for $|x| \leq r$ to further upper-bound (21) by

$$\begin{aligned}
&p_e\left(1 + t\sum_{j \in e \setminus \{m\}} v_j + \frac{e^{td\|v\|_\infty}}{2} \cdot t^2 \left(\sum_{j \in e \setminus \{m\}} v_j\right)^2 - 1\right) - tp_e \sum_{j \in e \setminus \{m\}} v_j \\
&= p_e \frac{e^t}{2} \cdot t^2 \left(\sum_{j \in e \setminus \{m\}} v_j\right)^2 \leq \frac{e^t p_{\max} t^2}{2} \cdot d \sum_{j \in e \setminus \{m\}} {v_j}^2,
\end{aligned}$$

where $\max_e p_e \triangleq p_{\max}$.

Substituting our bounds on the log of moment generating functions into (20), we obtain

$$\begin{aligned}
\log(\mathbb{P}\left(X \geq \delta\right)) &\leq -t\delta + \frac{e^t p_{\max} d}{2} t^2 \sum_{e \in \mathcal{E}^{(m)}} \sum_{j \in e \setminus \{m\}} {v_j}^2 \\
&\leq -t\delta + \frac{e^t p_{\max} d}{2} \cdot t^2 \binom{n-2}{d-2} \|v\|_2^2,
\end{aligned}$$

where the last step follows from the fact that each $j \neq m$ appears in $\binom{n-2}{d-2}$ potential hyperedges in $\mathcal{E}^{(m)}$. Let $t = 1 \bigvee \log\left(\frac{\sqrt{n}}{d\|v\|_2}\right)$. Using the fact that $(1 \vee \log x)^2 \leq x$ for $x \geq 1$,

$$\log(\mathbb{P}\left(X \geq \delta\right)) \leq -t\delta + \frac{e^t p_{\max} d}{2} \frac{\sqrt{n}}{d\left\|v\right\|_2} \binom{n-2}{d-2} \left\|v\right\|_2^2.$$

Observe that $\|v\|_2 \leq \sqrt{n}\left\|v\right\|_\infty = \sqrt{n}/d$, so that $\log\left(\frac{\sqrt{n}}{d\|v\|_2}\right) \geq 0$. Therefore, $e^t = e^{1 \vee \log(\frac{\sqrt{n}}{d\|v\|_2})} \leq e^{1+\log(\frac{\sqrt{n}}{d\|v\|_2})} \leq \frac{e\sqrt{n}}{d\|v\|_2}$. Hence

$$\log(\mathbb{P}\left(X \geq \delta\right)) \leq -t\delta + \frac{e p_{\max} d}{2} \cdot \frac{\sqrt{n}}{d\left\|v\right\|_2} \frac{\sqrt{n}}{d\left\|v\right\|_2} \binom{n-2}{d-2} \left\|v\right\|_2^2 = -t\delta + \frac{e p_{\max} n\binom{n-2}{d-2}}{2d}.$$

Let $a = 8d/c_0$ and set $\delta = \ t^{-1}d^{-1}(2+a)p_{\max} n\binom{n-2}{d-2}$. We then obtain the bound

$$\log(\mathbb{P}\left(X \geq \delta\right)) \leq -\frac{(2+a)p_{\max} n}{d}\binom{n-2}{d-2} + \frac{e p_{\max} n\binom{n-2}{d-2}}{2d} \leq -\frac{a p_{\max} n\binom{n-2}{d-2}}{d}.$$

By replacing v with $-v$, we obtain a similar bound for the lower tail. The union bound gives

$$\mathbb{P}\left(|X| \geq \delta\right) \leq 2\exp\left(-\frac{a p_{\max} n\binom{n-2}{d-2}}{d}\right).$$

Substituting in the value of t and using Assumption 1 that $n\binom{n-2}{d-2}\mu_n \geq c_0 \log n$,

$$\mathbb{P}\left(|(W-W^*)_{m\cdot}v| \geq \frac{(2+a)n\binom{n-2}{d-2}\mu_n}{d\left(1 \bigvee \log\left(\frac{\sqrt{n}}{d\|v\|_2}\right)\right)}\right) \leq 2e^{-\frac{a c_0 \log n}{d}}.$$

Finally, substituting the value of a,

$$\mathbb{P}\left(|(W-W^*)_{m\cdot}v| \leq \frac{(2+8d/c_0)n\binom{n-2}{d-2}\mu_n}{d\left(1 \bigvee \log\left(\frac{\sqrt{n}}{d\|v\|_2}\right)\right)}\right) \geq 1-2n^{-4}.$$

Recalling that $\left\|v\right\|_\infty = 1/d$ yields:

$$\mathbb{P}\left(|(W-W^*)_{m\cdot}v| \leq \frac{(2+\frac{8d}{c_0})\left\|v\right\|_\infty n\binom{n-2}{d-2}\mu_n}{1 \bigvee \log\left(\frac{\sqrt{n}\|v\|_\infty}{\|v\|_2}\right)}\right) \geq 1-2n^{-4}.$$

Since $\varphi\left(\frac{\|v\|_2}{\sqrt{n}\|v\|_\infty}\right) = (2+8d/c_0)/(1 \vee \log\left(\frac{\sqrt{n}\|v\|_\infty}{\|v\|_2}\right)$, the lemma follows. ■

Lemma 16 $\left\|W - W^{(m)}\right\|_2 \lesssim \ \|W\|_{2\to\infty}$.

Proof Note that $W - W^{(m)}$ is the similarity matrix of the graph with edges present only from $\mathcal{E}^{(m)}$. Consider any $i \neq m$. Since for any $e \in \mathcal{E}^{(m)}$ with $\{i, m\} \subseteq e$, the edge e contributes to exactly $(d-1)$ entries of the i-th row of $W - W^{(m)}$, we have

$$\|(W - W^{(m)})_{i\cdot}\|_1 = (d-1)W_{im}. \tag{22}$$

Therefore $\left\|(W - W^{(m)})_{i\cdot}\right\|_2 \leq \left\|(W - W^{(m)})_{i\cdot}\right\|_1 \leq (d-1)W_{im}$. Hence we get

$$\begin{aligned}
\|W - W^{(m)}\|_2 &\leq \|W - W^{(m)}\|_F \\
&= \sqrt{\sum_{i\in[n]} \|(W - W^{(m)})_{i\cdot}\|_2^2} \\
&\leq \sqrt{\|W_{m\cdot}\|_2^2 + \sum_{i\in[n]\setminus\{m\}} (d-1)^2 W_{im}^2} \\
&\lesssim \sqrt{\|W_{m\cdot}\|_2^2} \;\leq\; \|W\|_{2\to\infty}\,.
\end{aligned}$$

■

We now record a sharp spectral norm concentration result of Lee, Kim, and Chung (Lee et al., 2020), which will play a crucial role in our further analysis.

Lemma 17 *(Lee et al., 2020, Special case of Theorem 4) Fix $d \in \{2, 3, \ldots\}$. Let $p \in [0,1]^{\binom{[n]}{d}}$ be such that Assumption 1 holds. Let $G \sim H(d, n, p)$ and $W = \mathcal{S}(G)$. Then there exists a constant $C = C(d, c_0) > 0$ such that*

$$\mathbb{P}\left(\|W - W^*\|_2 \leq C\sqrt{n\binom{n-2}{d-2}\mu_n}\right) \geq 1 - O(n^{-11}).$$

Let us define a parameter γ, which controls the concentration in the analysis from here.

$$\gamma = \gamma_n := \frac{C}{\sqrt{n\binom{n-2}{d-2}\mu_n}} \bigvee \frac{1}{\sqrt{n}},$$

where $C = C(d, c_0) > 0$ is the constant from Lemma 17. Recalling the definition of $\varphi(\cdot)$ (Equation 19), observe that

$$\gamma = o(1) \;\text{ and }\; \varphi(\gamma) \lesssim \left(\frac{1}{1 \vee \log\sqrt{n\binom{n-2}{d-2}\mu_n}} \bigvee \frac{1}{1 \vee \log\sqrt{n}}\right) \lesssim \frac{1}{\log\log n} = o(1), \tag{23}$$

where we used Assumption 1 that $n\binom{n-2}{d-2}\mu_n \geq c_0 \log n$. We define the following event:

$$F_0 := \left\{\|W - W^*\|_2 \leq \gamma \cdot n\binom{n-2}{d-2}\mu_n\right\}.$$

By Lemma 17,

$$\mathbb{P}\left(F_0^{\mathrm{c}}\right) = \mathbb{P}\left(\|W - W^*\|_2 > \gamma \cdot n\binom{n-2}{d-2}\mu_n\right)$$

$$\leq \mathbb{P}\left(\|W - W^*\|_2 > C\sqrt{n\binom{n-2}{d-2}\mu_n}\right) \qquad \text{(by definition of } \gamma)$$
$$= O(n^{-11}). \qquad \text{(using Lemma 17)}$$

Therefore, $\mathbb{P}(F_0) \geq 1 - O(n^{-11})$. We now derive some bounds on important quantities conditioned on the above event.

Lemma 18 *Conditioned on F_0, we have $|\lambda^*| \asymp |\lambda| \asymp n^{d-1}\mu_n$.*

Proof Conditioned on the event F_0, Weyl's inequality implies

$$|\lambda - \lambda^*| \leq \|W - W^*\|_2 \leq \gamma \cdot n\binom{n-2}{d-2}\mu_n = o(n^{d-1}\mu_n),$$

where the last inequality follows since $n\binom{n-2}{d-2} = \Theta(n^{d-1})$ and $\gamma = o(1)$. Therefore, $\lambda \in \lambda^* \pm o(n^{d-1}\mu_n)$. Thus, Assumption 1 (i.e. $|\lambda^*| \asymp n^{d-1}\mu_n$) then ensures that $|\lambda^*| \asymp |\lambda| \asymp n^{d-1}\mu_n$. ■

Lemma 19 *Conditioned on F_0,*

$$\|W\|_{2\to\infty} \lesssim \gamma \cdot n^{d-1}\mu_n \quad \text{and for all } m \in [n], \quad \left\|W^{(m)} - W^*\right\|_2 \lesssim \gamma \cdot n^{d-1}\mu_n.$$

Proof By the triangle inequality,

$$\begin{aligned}
\|W\|_{2\to\infty} &\leq \|W - W^*\|_2 + \|W^*\|_{2\to\infty} \\
&\leq \left(\frac{C}{\sqrt{n\binom{n-2}{d-2}\mu_n}} + \frac{1}{\sqrt{n}}\right) n\binom{n-2}{d-2}\mu_n \qquad \text{(using Lemma 17 and Observation 1)} \\
&\lesssim \gamma \cdot n^{d-1}\mu_n \qquad \text{(by definition of } \gamma \text{ and } n\tbinom{n-2}{d-2} = \Theta(n^{d-1}))
\end{aligned}$$

Similarly, using the triangle inequality and Lemma 16,

$$\|W^{(m)} - W^*\|_2 \leq \|W - W^*\|_2 + \|W - W^{(m)}\|_2 \lesssim \|W - W^*\|_2 + \|W\|_{2\to\infty} \lesssim \gamma \cdot n^{d-1}\mu_n.$$

■

Having derived the above bounds, we now bound the ℓ_2 and ℓ_∞ norms of $(s^{(m)}u^{(m)} - u^*)$ and $(su - s^{(m)}u^{(m)})$ using two variants of the Davis and Kahan $\sin(\theta)$ theorem. For completeness, we include the less well-known variant here, which is a special case of (Deng et al., 2021, Theorem 3).

Proposition 20 (Generalized Davis and Kahan sin(θ) theorem (Deng et al., 2021)) *Let $M \in \mathbb{R}^{n\times n}$ be a symmetric matrix and let X be the matrix that has the eigenvectors of M as columns. Then M can be decomposed as $M = X\Lambda X^\top = X_1\Lambda_1 X_1{}^\top + X_2\Lambda_2 X_2{}^\top$, where $X = [X_1\ X_2]$ and $\Lambda = \begin{bmatrix}\Lambda_1 & 0 \\ 0 & \Lambda_2\end{bmatrix}$. Suppose $\delta = \min_i |(\Lambda_2)_{ii} - \hat{\lambda}|$ is the absolute separation of some $\hat{\lambda}$ from Λ_2, then for any vector $\hat{u}$ we have*

$$\sin(\theta) \leq \frac{\|(M - \hat{\lambda}\mathbf{I})\,\hat{u}\|_2}{\delta},$$

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

Lemma 21 *Conditioned on F_0,*

$$\max_{m\in[n]} \|s^{(m)}u^{(m)} - u^*\|_2 \lesssim \gamma, \tag{24}$$

$$\max_{m\in[n]} \|su - s^{(m)}u^{(m)}\|_2 \lesssim (\gamma \wedge \|u\|_\infty), \tag{25}$$

$$\max_{m\in[n]} \|s^{(m)}u^{(m)} - u^*\|_\infty \lesssim (\|u\|_\infty + \|u^*\|_\infty). \tag{26}$$

Proof For any $m \in [n]$, we apply a variant of the Davis–Kahan $\sin(\theta)$ theorem (Yu et al., 2014, Corollary 3) to get

$$\|s^{(m)}u^{(m)} - u^*\|_2 \;\le\; \frac{2^{3/2}\left\|W^{(m)} - W^*\right\|_2}{\Delta^*} \;\lesssim\; \frac{\gamma \cdot n^{d-1}\mu_n}{n^{d-1}\mu_n} = \gamma,$$

where the second inequality follows from Lemma 19 and Assumption 1, concluding the proof of (24). Similarly, we also get

$$\|su - s^{(m)}u^{(m)}\|_2 \;\le\; \|su - u^*\|_2 + \|s^{(m)}u^{(m)} - u^*\|_2 \;\lesssim\; \frac{\|W - W^*\|_2}{\Delta^*} + \gamma \;\lesssim\; \gamma. \tag{27}$$

Therefore, $\langle su, s^{(m)}u^{(m)}\rangle \ge 0$. Let θ denote the angle between su and $s^{(m)}u^{(m)}$. Then

$$\|su - s^{(m)}u^{(m)}\|_2 \le \sqrt{\|u\|_2^2 + \|u^{(m)}\|_2^2 - 2\|u\|_2\|u^{(m)}\|_2 \cos\theta} \le \sqrt{2 - 2\cos^2\theta} = \sqrt{2}\sin\theta.$$

We then apply Proposition 20 with $M = W^{(m)}, X_1 = [u^{(m)}]$ and $(\hat{\lambda}, \hat{u}) = (\lambda, u)$:

$$\begin{aligned}
\|su - s^{(m)}u^{(m)}\|_2 \le \sqrt{2}\sin\theta &\le \frac{\sqrt{2}\left\|(W^{(m)} - \lambda\mathbf{I})u\right\|_2}{|\lambda^{(m)}_{k+1} - \lambda| \wedge |\lambda - \lambda^{(m)}_{k-1}|} \\
&\le \frac{\sqrt{2}\left\|(W^{(m)} - W)u\right\|_2}{(\lambda_{k+1} - \lambda_k) \wedge (\lambda_k - \lambda_{k-1}) - \|W - W^{(m)}\|_2} \\
&\lesssim \frac{\left\|(W - W^{(m)})u\right\|_2}{\Delta^* - 2\|W - W^*\|_2 - \|W - W^{(m)}\|_2}.
\end{aligned}$$

Using Assumption 1, the definition of F_0, and Lemmas 16, 19, we can lower-bound the denominator by

$$\Delta^* - 2\|W - W^*\|_2 - \|W - W^{(m)}\|_2 \gtrsim (1-\gamma)n^{d-1}\mu_n.$$

Since $\gamma = o(1)$ by (23), we obtain

$$\|su - s^{(m)}u^{(m)}\|_2 \lesssim \frac{\|(W - W^{(m)})u\|_2}{n^{d-1}\mu_n}.$$

Let $v = (W - W^{(m)})u$. We have already seen that the m-th row of $W - W^{(m)}$ is the same as W, so that $v_m = \lambda u_m$. Therefore, we bound the m-th entry of v and the rest of its entries separately. Formally,

$$|v_m| = |[Wu]_m| \le |\lambda||u_m| \le |\lambda|\,\|u\|_\infty\,,$$

$$|v_i| = |[(W - W^{(m)})u]_i| \leq \|(W - W^{(m)})_{i\cdot}\|_1 \, \|u\|_\infty \lesssim W_{im} \, \|u\|_\infty \,, \text{ for } i \neq m,$$

where the last inequality follows from (22). Therefore,

$$\|v\|_2 \;\lesssim\; \|u\|_\infty \sqrt{\lambda^2 + \sum_{i \neq m} W_{im}^2} \lesssim \|u\|_\infty \; \sqrt{|\lambda^*|^2 + \|W\|_{2\to\infty}^2}, \tag{28}$$

where the last step follows from Lemma 18. Substituting the bound (28):

$$\|su - s^{(m)}u^{(m)}\|_2 \;\lesssim\; \frac{\|v\|_2}{n^{d-1}\mu_n} \;\lesssim\; \frac{\|u\|_\infty \sqrt{|\lambda^*|^2 + \|W\|_{2\to\infty}^2}}{n^{d-1}\mu_n} \;\lesssim\; \|u\|_\infty \,, \tag{29}$$

where the last inequality follows from Lemma 19 and Assumption 1. Observe that (27) and (29) together imply (25). Finally, to prove (26), we apply the triangle inequality again and use (29).

$$\|s^{(m)}u^{(m)} - u^*\|_\infty \leq \|su - s^{(m)}u^{(m)}\|_2 + \|u\|_\infty + \|u^*\|_\infty \lesssim \|u\|_\infty + \|u^*\|_\infty \,,$$

concluding the proof. ■

Lemma 22 *With probability at least* $1 - O(n^{-3})$*,*

$$\|(W - W^*)u^*\|_\infty \lesssim \|u^*\|_\infty \, n^{d-1}\mu_n \quad \text{and} \quad \|Wu^*\|_\infty \lesssim \|u^*\|_\infty \, n^{d-1}\mu_n.$$

Proof To prove the first inequality, we apply Lemma 15. For each $m \in [n]$, we have

$$|(W - W^*)_{m\cdot}u^*| \leq \varphi\left(\frac{\|u^*\|_2}{\sqrt{n}\,\|u^*\|_\infty}\right) \|u^*\|_\infty \, n\binom{n-2}{d-2}\mu_n$$

with probability $1 - O(n^{-4})$. Using the monotonicity of φ (Observation 2) and $\|u^*\|_2 \leq \sqrt{n}\,\|u\|_\infty$, we obtain

$$|(W - W^*)_{m\cdot}u^*| \lesssim \varphi(1)\, \|u^*\|_\infty \, n^{d-1}\mu_n$$

with probability $1 - O(n^{-4})$. Note that $\varphi(1) = O(1)$. Taking a union bound over all $m \in [n]$, we have $\|(W - W^*)u^*\|_\infty \lesssim \|u^*\|_\infty \, n^{d-1}\mu_n$, with probability at least $1 - O(n^{-3})$.

To get the second statement, we apply the first statement to show that with probability $1 - O(n^{-3})$,

$$\|Wu^*\|_\infty \leq \|W^*u^*\|_\infty + \|(W - W^*)u^*\|_\infty \lesssim |\lambda^*|\,\|u^*\|_\infty + \|u^*\|_\infty \, n^{d-1}\mu_n \lesssim \|u^*\|_\infty \, n^{d-1}\mu_n.$$

The last inequality in the above follows from Assumption 1. ■

Lemma 23 *With probability at least* $1 - O(n^{-3})$*,*

$$\max_{m \in [n]} |W_{m\cdot}(s^{(m)}u^{(m)} - u^*)| \lesssim (\gamma + \varphi(\gamma))(\|u\|_\infty + \|u^*\|_\infty)n^{d-1}\mu_n.$$

Proof We denote $(s^{(m)}u^{(m)} - u^*)$ by $v^{(m)}$ for notational convenience. Recall that $\mathbb{P}(F_0) \geq 1 - O(n^{-11})$ by Lemma 12. Conditioned on F_0, for all $m \in [n]$

$$\begin{aligned}
|W_{m\cdot}v^{(m)}| &\leq |W^*_{m\cdot}v^{(m)}| + |(W - W^*)_{m\cdot}v^{(m)}| && \text{(by the triangle inequality)} \\
&\leq \|W^*\|_{2\to\infty} \|v^{(m)}\|_2 + |(W - W^*)_{m\cdot}v^{(m)}| && \text{(by the Cauchy–Schwarz inequality)} \\
&\lesssim \sqrt{n} \cdot n^{d-2}\mu_n \cdot \gamma + |(W - W^*)_{m\cdot}v^{(m)}| && \text{(by Observation 1 and (24))} \\
&= \frac{n^{d-1}\mu_n \gamma}{\sqrt{n}} + |(W - W^*)_{m\cdot}v^{(m)}| \\
&\leq \gamma \|u^*\|_\infty n^{d-1}\mu_n + |(W - W^*)_{m\cdot}v^{(m)}|. && (30)
\end{aligned}$$

We now focus on bounding the second term. We know that $v^{(m)}$ is independent of the randomness in the m-th row of W. Therefore, by the row concentration result (Lemma 15), for a fixed $m \in [n]$

$$|(W - W^*)_{m\cdot}v^{(m)}| \leq \|v^{(m)}\|_\infty \varphi\left(\frac{\|v^{(m)}\|_2}{\sqrt{n}\|v^{(m)}\|_\infty}\right) n\binom{n-2}{d-2}\mu_n, \tag{31}$$

holds with probability at least $1 - O(n^{-4})$. Let F be the event that (31) holds simultaneously for all $m \in [n]$. By a union bound, $\mathbb{P}(F) \geq 1 - O(n^{-3})$. Conditioned on the event F, let us consider two different cases.

Case 1: Suppose $\frac{\|v^{(m)}\|_2}{\sqrt{n}\|v^{(m)}\|_\infty} \leq \gamma$. Under this case, we use the fact that $\varphi(x)$ is non-decreasing (Observation 2) in (31) to get:

$$|(W - W^*)_{m\cdot}v^{(m)}| \leq \varphi(\gamma)\|v^{(m)}\|_\infty n\binom{n-2}{d-2}\mu_n.$$

Case 2: Suppose $\frac{\|v^{(m)}\|_2}{\sqrt{n}\|v^{(m)}\|_\infty} > \gamma$. In this case, multiplying and dividing (31) by $\frac{\|v^{(m)}\|_2}{\sqrt{n}}$,

$$\begin{aligned}
|(W - W^*)_{m\cdot}v^{(m)}| &\leq \frac{\|v^{(m)}\|_2}{\sqrt{n}} \frac{\sqrt{n}\|v^{(m)}\|_\infty}{\|v^{(m)}\|_2} \varphi\left(\frac{\|v^{(m)}\|_2}{\sqrt{n}\|v^{(m)}\|_\infty}\right) \cdot n\binom{n-2}{d-2}\mu_n \\
&\leq \frac{\varphi(\gamma)}{\gamma} \frac{\|v^{(m)}\|_2}{\sqrt{n}} \cdot n\binom{n-2}{d-2}\mu_n,
\end{aligned}$$

where we have used the fact that $\varphi(x)/x$ is non-increasing (Observation 2). Combining both cases, for all $m \in [n]$:

$$|(W - W^*)_{m\cdot}v^{(m)}| \leq \varphi(\gamma) \cdot n\binom{n-2}{d-2}\mu_n \left(\|v^{(m)}\|_\infty \vee \frac{\|v^{(m)}\|_2}{\gamma\sqrt{n}}\right).$$

Substituting bounds from (24) and (26) in the above, and then using the fact that $\|u^*\|_\infty \geq \|u^*\|_2 / \sqrt{n} = 1/\sqrt{n}$, we obtain

$$|(W - W^*)_{m\cdot}v^{(m)}| \lesssim \varphi(\gamma) \cdot n^{d-1}\mu_n \left((\|u\|_\infty + \|u^*\|_\infty) \vee \frac{\gamma}{\gamma\sqrt{n}}\right) \lesssim \varphi(\gamma)\left(\|u\|_\infty + \|u^*\|_\infty\right) n^{d-1}\mu_n.$$

Finally, we substitute into (30):

$$\begin{aligned}\max_{m\in[n]} |W_{m\cdot}v^{(m)}| &\lesssim \gamma \left\|u^*\right\|_\infty n^{d-1}\mu_n + \varphi(\gamma)(\left\|u\right\|_\infty + \left\|u^*\right\|_\infty)n^{d-1}\mu_n \\ &\leq (\gamma+\varphi(\gamma))(\left\|u\right\|_\infty + \left\|u^*\right\|_\infty)n^{d-1}\mu_n,\end{aligned}$$

concluding the proof. ■

We finally prove Theorem 7.

Proof [Proof of Theorem 7] Recall the definition of the event F_0 and $\mathbb{P}(F_0) \geq 1-O(n^{-11})$ (Lemma 17). Conditioned on F_0, we have $|\lambda| \asymp |\lambda^*|$ by Lemma 18; we use this throughout the proof.

We first bound $\left\|u\right\|_\infty$ in terms of $\left\|u^*\right\|_\infty$ as we need our final bounds only in terms of the latter. By the triangle inequality,

$$\begin{aligned}\left\|u\right\|_\infty = \left\|\frac{sWu}{\lambda}\right\|_\infty &\leq \left\|\frac{Wu^*}{\lambda}\right\|_\infty + \left\|\frac{W(su-u^*)}{\lambda}\right\|_\infty \lesssim \frac{1}{|\lambda^*|}\left(\left\|Wu^*\right\|_\infty + \max_{m\in[n]}|W_{m\cdot}(su-u^*)|\right) \\ &\leq \frac{1}{|\lambda^*|}\left(\left\|Wu^*\right\|_\infty + \max_{m\in[n]}|W_{m\cdot}(su-s^{(m)}u^{(m)})| + \max_{m\in[n]}|W_{m\cdot}(s^{(m)}u^{(m)}-u^*)|\right) \\ &\leq \frac{1}{|\lambda^*|}\left(\left\|Wu^*\right\|_\infty + \left\|W\right\|_{2\to\infty}\max_{m\in[n]}\left\|su-s^{(m)}u^{(m)}\right\|_2 + \max_{m\in[n]}|W_{m\cdot}(s^{(m)}u^{(m)}-u^*)|\right).\end{aligned}$$

We substitute the derived bounds from Lemma 22 in the first term, Lemmas 19 and 21 in the second term, and Lemma 23 in the third term. In particular, we obtain that with probability at least $1 - O(n^{-3})$,

$$\begin{aligned}\left\|u\right\|_\infty &\lesssim \frac{1}{|\lambda^*|}\Big(\left\|u^*\right\|_\infty n^{d-1}\mu_n + (\gamma\cdot n^{d-1}\mu_n)\left\|u\right\|_\infty + (\gamma+\varphi(\gamma))(\left\|u\right\|_\infty + \left\|u^*\right\|_\infty)n^{d-1}\mu_n\Big) \\ &\lesssim \Big((1+\gamma+\varphi(\gamma))\left\|u^*\right\|_\infty + (\gamma+\varphi(\gamma)\left\|u\right\|_\infty\Big). \qquad \text{(using Assumption 1)}\end{aligned}$$

Using the fact that $\gamma, \varphi(\gamma) = o(1)$ from (23), we have that for some constant $c_3 > 0$,

$$\begin{aligned}\left\|u\right\|_\infty &\leq c_3\left\|u^*\right\|_\infty + o(1)\left\|u\right\|_\infty \\ (1-o(1))\left\|u\right\|_\infty &\leq c_3\left\|u^*\right\|_\infty \\ \left\|u\right\|_\infty &\leq \frac{c_3}{1-o(1)}\left\|u^*\right\|_\infty \lesssim \left\|u^*\right\|_\infty.\end{aligned} \tag{32}$$

Similarly, conditioned on F_0, we bound the quantity of interest. Using the triangle inequality,

$$\begin{aligned}\left\|su - \frac{Wu^*}{\lambda^*}\right\|_\infty &= \left\|\frac{sWu}{\lambda} - \frac{Wu^*}{\lambda} + \frac{Wu^*}{\lambda} - \frac{Wu^*}{\lambda^*}\right\|_\infty \\ &\leq \left|\frac{1}{\lambda} - \frac{1}{\lambda^*}\right|\left\|Wu^*\right\|_\infty + \frac{1}{|\lambda|}\left\|W(su-u^*)\right\|_\infty.\end{aligned}$$

Recalling that $|\lambda| \asymp |\lambda^*|$,

$$\left\|su - \frac{Wu^*}{\lambda^*}\right\|_\infty \lesssim \frac{1}{|\lambda^*|}\left(\gamma\left\|Wu^*\right\|_\infty + \left\|W(su-u^*)\right\|_\infty\right)$$

$$\lesssim \frac{1}{|\lambda^*|}\left(\gamma\left\|Wu^*\right\|_\infty + \|W\|_{2\to\infty}\max_{m\in[n]}\|su - s^{(m)}u^{(m)}\|_2 + \max_{m\in[n]}|W_{m\cdot}(s^{(m)}u^{(m)} - u^*)|\right).$$

We again substitute the bounds from Lemma 22 in the first term, Lemma 19 and 21 in the second term, and Lemma 23 in the third term. In particular, we obtain that with probability at least $1 - O(n^{-3})$,

$$\begin{aligned}
\left\|su - \frac{Wu^*}{\lambda^*}\right\|_\infty &\lesssim \frac{1}{|\lambda^*|}\Big(\gamma\left\|u^*\right\|_\infty n^{d-1}\mu_n + (\gamma\cdot n^{d-1}\mu_n)\left\|u\right\|_\infty + (\gamma+\varphi(\gamma))(\left\|u\right\|_\infty + \left\|u^*\right\|_\infty)n^{d-1}\mu_n\Big) \\
&\lesssim \Big((\gamma+\varphi(\gamma))\left\|u^*\right\|_\infty + (\gamma+\varphi(\gamma)\left\|u\right\|_\infty\Big) && \text{(using Assumption 1)} \\
&\lesssim (\gamma+\varphi(\gamma))\left\|u^*\right\|_\infty && \text{(using } \|u\|_\infty \lesssim \|u^*\|_\infty \text{ as shown in (32))} \\
&\lesssim \frac{\left\|u^*\right\|_\infty}{\log\log n}. && (\gamma = O(\varphi(\gamma)) \text{ and using (23))}
\end{aligned}$$

This is precisely the bound we need, concluding the proof. ■

Appendix D. Spectral norm concentration for similarity matrices

In order to prove Theorem 10, we require the following result.

Lemma 24 *Let A be a random matrix with independent entries, where $A_{ij} \in [a, b]$ for two constants $a < b$. Suppose $\mathbb{E}\left[|A_{ij}|\right] \le q$ for all i, j, where $\frac{c_2\log n}{n} \le q \le 1 - c_3$ for arbitrary constants $c_2, c_3 > 0$. Then, there exists a constant $c' := c'(c_2, c_3, a, b) > 0$ such that*

$$\mathbb{E}\left[\left\|A - \mathbb{E}\left[A\right]\right\|_2\right] \le c'\sqrt{nq}.$$

Proof We use ideas from (Dhara et al., 2022c, Lemma 4.5), who showed a similar result for zero-diagonal, symmetric matrices with independent entries. We first construct a symmetric matrix B using A to reduce it to the symmetric case. Let

$$B = \begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix}.$$

Fix a vector $x \in \mathbb{R}^n$ such that $\left\|x\right\|_2 = 1$, and consider the vector $y \in \mathbb{R}^{2n}$ such that $y = \begin{bmatrix} \mathbf{0}_n \\ x \end{bmatrix}$. Observe that $\left\|y\right\|_2 = 1$. Moreover,

$$(B - \mathbb{E}[B])y = \begin{bmatrix} (A - \mathbb{E}[A])x \\ \mathbf{0}_n \end{bmatrix}.$$

Therefore,

$$\left\|(B - \mathbb{E}[B])y\right\|_2 = \left\|(A - \mathbb{E}[A])x\right\|_2,$$

so that

$$\|(A - \mathbb{E}[A])x\|_2 = \|(B - \mathbb{E}[B])y\|_2 \le \|B - \mathbb{E}[B]\|_2.$$

Since x was arbitrary, we have shown $\left\|A - \mathbb{E}\left[A\right]\right\|_2 \le \left\|B - \mathbb{E}\left[B\right]\right\|_2$, and therefore

$$\mathbb{E}[\left\|A - \mathbb{E}\left[A\right]\right\|_2] \le \mathbb{E}[\left\|B - \mathbb{E}\left[B\right]\right\|_2]. \tag{33}$$

Thus, it suffices to bound $\mathbb{E}[\|B - \mathbb{E}[B]\|_2]$. Let $B^+ = \max\{B, 0\}$, where the maximum is taken entrywise. Similarly, let $B^- = -\min\{B, 0\}$. Then we can write $B = B^+ - B^-$. By the triangle inequality,

$$\mathbb{E}[\|B - \mathbb{E}[B]\|_2] \le \mathbb{E}\left[\|B^+ - \mathbb{E}[B^+]\|_2\right] + \mathbb{E}\left[\|B^- - \mathbb{E}[B^-]\|_2\right]. \tag{34}$$

Observe that B^+ and B^- are nonnegative, zero-diagonal, symmetric matrices with independent entries. Also, for all i, j,

$$\max\{\mathbb{E}[B^+_{ij}], \mathbb{E}[B^-_{ij}]\} \le \mathbb{E}[|B_{ij}|] \le \max \mathbb{E}[|A_{ij}|] \le q.$$

If $b \le 0$, then $\|B^+ - \mathbb{E}[B^+]\|_2 = 0$. Otherwise it follows from (Hajek et al., 2016a, Theorem 5) that there exists $c^+ > 0$ such that

$$\mathbb{E}\left[\frac{1}{b}(\|B^+ - \mathbb{E}[B^+]\|_2)\right] \le c^+\sqrt{\frac{nq}{b}}.$$

Similarly, if $a \ge 0$, then $\|B^- - \mathbb{E}[B^-]\|_2 = 0$. Otherwise, we again use (Hajek et al., 2016a, Theorem 5) to conclude that there exists a constant $c^- > 0$ such that

$$\mathbb{E}\left[\frac{1}{|a|}(\|B^- - \mathbb{E}[B^-]\|_2)\right] \le c^-\sqrt{\frac{nq}{|a|}}.$$

Combining these with (33) and (34) we get

$$\mathbb{E}[\|A - \mathbb{E}[A]\|_2] \le c'\sqrt{nq},$$

where $c' = c^+\sqrt{\max\{b, 0\}} + c^-\sqrt{|\min\{a, 0\}|}$. ■

Proof [Proof of Theorem 10] The symbols $\lesssim$ and $\asymp$ hide constants in d, c_0 throughout the proof. Our goal is to bound $\mathbb{E}[\|\mathcal{S}(G) - \mathbb{E}[\mathcal{S}(G)]\|_2]$. Let G' be an independent copy of G. Observe that for a fixed matrix X, the function $f(Y) = \|X - Y\|_2$ is convex. By Jensen's inequality,

$$\mathbb{E}[\|\mathcal{S}(G) - \mathbb{E}[\mathcal{S}(G)]\|_2] = \mathbb{E}\left[\|\mathcal{S}(G) - \mathbb{E}[\mathcal{S}(G')]\|_2\right] \le \mathbb{E}\left[\|\mathcal{S}(G) - \mathcal{S}(G')\|_2\right].$$

We can extend the definition of $\mathcal{S}$ so that $\mathcal{S}(G - G') = \mathcal{S}(G) - \mathcal{S}(G')$; i.e. $G - G'$ is a "hypergraph" with edges labeled by $\{1, 0, -1\}$.

Let R be a symmetric tensor of order d and dimension n with independent Rademacher entries; i.e. the entries $\{R(i_1, i_2, \ldots, i_d) : i_1 \le i_2 \le \cdots \le i_d\}$ are mutually independent. Let $\circ$ denote the edge-wise product. Since $G - G'$ has the same distribution as $(G - G') \circ R$, we obtain

$$\mathbb{E}[\|\mathcal{S}(G) - \mathbb{E}[\mathcal{S}(G)]\|_2] \le \mathbb{E}\left[\|\mathcal{S}(G - G')\|_2\right] = \mathbb{E}\left[\|\mathcal{S}((G - G') \circ R)\|_2\right] \le 2\mathbb{E}[\|\mathcal{S}(G \circ R)\|_2], \tag{35}$$

where the last inequality follows from the triangle inequality. Let $p_{\max} \triangleq c_0 \log n / \binom{n-1}{d-1}$ for simplicity. Consider the hypergraph G_+ that is coupled to G as follows. The hypergraph G_+ does not contain any edge that appears in G. Each edge e that is not present in G is present in G_+ with probability $\frac{p_{\max} - p_e}{1 - p_e}$ (independently across edges). Letting $G^{(1)} \sim \mathrm{HSBM}(d, n, p_{\max}, p_{\max})$, we see that $(G + G_+) \circ R$ has the same distribution as $G^{(1)} \circ R$. Also, $\mathbb{E}[\mathcal{S}(G_+ \circ R) \mid G] = 0$. Using these observations along with Jensen's inequality, we obtain

$$\mathbb{E}[\|\mathcal{S}(G \circ R)\|_2] = \mathbb{E}[\|\mathcal{S}(G \circ R) + \mathbb{E}[\mathcal{S}(G_+ \circ R) \mid G]\|_2] \le \mathbb{E}[\|\mathcal{S}(G \circ R) + \mathcal{S}(G_+ \circ R)\|_2]$$

$$= \mathbb{E}\left[\|\mathcal{S}((G + G_+) \circ R)\|_2\right] = \mathbb{E}\left[\|\mathcal{S}(G^{(1)} \circ R)\|_2\right].$$

Substituting this into (35), we have

$$\mathbb{E}\left[\|\mathcal{S}(G) - \mathbb{E}\left[\mathcal{S}(G)\right]\|_2\right] \le 2\mathbb{E}\left[\|S(G^{(1)} \circ R)\|_2\right]. \tag{36}$$

Note that, even though $\mathcal{S}(G^{(1)} \circ R)$ has identically distributed entries, they are still dependent; we want a matrix with independent entries instead to apply Lemma 24. To achieve this, we use a somewhat involved symmetrization argument. For simplicity, consider the case when $d = 2$ (which reduces to the adjacency matrix), where we have independent entries up to the symmetry. In other words, each independent edge random variable appears exactly twice in the matrix. In this situation, (Hajek et al., 2016a, Theorem 5) uses the standard symmetrization technique by adding another independent copy of $\mathcal{S}(G^{(1)} \circ R)$ and rearranging random variables to create independence. However, for a general d, this is not sufficient. But observe that the random variable (label) associated with any fixed edge e is added to exactly $K := 2 \cdot \binom{d}{2} = d^2 - d$ entries of the similarity matrix when we apply the map $\mathcal{S}(\cdot)$. Thus, we instead add K independent copies to create enough independence and show how to rearrange the hyperedge random variables such that each matrix has fully independent entries after the rearrangement.

More formally, let $G^{(m)}$ and $R^{(m)}$ be independent copies of $G^{(1)}$ and R respectively for $m \in [K]$. Note that $\mathbb{E}\left[\mathcal{S}(G^{(m)} \circ R^{(m)})\right]$ is the zero matrix. Thus, adding the zero matrix, and then using Jensen's inequality we get

$$\begin{aligned} \mathbb{E}\left[\|\mathcal{S}(G^{(1)} \circ R)\|_2\right] &= \mathbb{E}\left[\left\|\mathcal{S}(G^{(1)} \circ R^{(1)}) + \sum_{m=2}^{K} \mathbb{E}\left[\mathcal{S}(G^{(m)} \circ R^{(m)})\right]\right\|_2\right] \\ &\le \mathbb{E}\left[\left\|\sum_{m=1}^{K} \mathcal{S}(G^{(m)} \circ R^{(m)})\right\|_2\right]. \end{aligned} \tag{37}$$

Observe that $\sum_{m=1}^{K} \mathcal{S}(G^{(m)} \circ R^{(m)})$ is the sum of *independent* copies of random matrices with *dependent* entries. The goal is to re-express this same quantity as the sum of *dependent* matrices with fully *independent* entries. To this end, let us consider the following construction. Let $L(e)$ be a fixed-ordered list of locations to which the random variable associated with the edge e is added. Formally, $L(e) := (\mathsf{op}_e^{(1)}, \dots, \mathsf{op}_e^{(K)})$ is an ordered list of all ordered pairs of e; i.e. $(i, j) : i, j \in e$.

For $e \in \mathcal{E}$ and $m, \ell \in [K]$ let $X^{(e,m,\ell)}$ be the $n \times n$ matrix which has only one non-zero entry:

$$X_{ij}^{(e,m,\ell)} = \begin{cases} A_e^{(m)} \circ R_e^{(m)} & (i, j) = \mathsf{op}_e^{(\ell)} \\ 0 & \text{otherwise,} \end{cases}$$

where $A_e^{(m)}$ denotes the indicator random variable associated with the edge e in $G^{(m)}$. By construction,

$$\sum_{m=1}^{K} \mathcal{S}(G^{(m)} \circ R^{(m)}) = \sum_{m=1}^{K} \sum_{e \in \mathcal{E}} \sum_{\ell=1}^{K} X^{(e,m,\ell)} \tag{38}$$

We will now re-express this summation as the sum of *dependent* matrices with *independent* entries. Let $Y = \sum_{i=1}^{\binom{n-2}{d-2}} X_i Z_i$, where $X_i \sim \text{Bern}(p_{\max})$ and $Z_i \sim \text{Rad}$ are independent. Let D be a diagonal matrix whose diagonal entries are i.i.d. copies of Y. Let us consider matrices $C^{(1)}, \dots, C^{(K)}$, where

$$C^{(1)} = \sum_{e \in \mathcal{E}} \sum_{\ell=1}^{K} X^{(e,\ell,\ell)} + D$$

$$C^{(k)} = \sum_{e \in \mathcal{E}} \sum_{\ell=1}^{K} X^{(e,(\ell+k-1)\%K,\ell)} + (-1)^{k+1} D$$

Here $\%$ represents the standard modulo operation, except $K\%K$ is K instead of 0. Intuitively, $C^{(1)}$ is the matrix for which the K locations corresponding to a given edge "consult" K independent copies of G. Observe that all the entries of $C^{(1)}$ are independent copies of Y. To ensure that we add the random variable associated with an edge of any given copy of G at all K locations in the similarity matrix, we iterate over the list in a cyclic manner when constructing $C^{(2)}, \ldots, C^{(K)}$. Note that D has the same distribution as $-D$. Therefore, from the symmetry all $C^{(k)}$s have the same distribution. Moreover,

$$\begin{aligned}
\sum_{k=1}^{K} C^{(k)} &= \sum_{k=1}^{K} \sum_{e \in \mathcal{E}} \sum_{\ell=1}^{K} X^{(e,(\ell+k-1)\%K,\ell)} + (-1)^{k+1} D \\
&= \sum_{e \in \mathcal{E}} \sum_{\ell=1}^{K} \sum_{k=1}^{K} X^{(e,(\ell+k-1)\%K,\ell)} + (-1)^{k+1} D \\
&= \sum_{e \in \mathcal{E}} \sum_{\ell=1}^{K} \sum_{m=1}^{K} X^{(e,m,\ell)} \\
&= \sum_{m=1}^{K} \mathcal{S}(G^{(m)} \circ R^{(m)}),
\end{aligned} \tag{39}$$

where the last step follows from (38). Therefore, combining (36), (37), and (39)

$$\mathbb{E}\left[\|W - W^*\|_2\right] \leq 2\,\mathbb{E}\left[\left\|\sum_{k=1}^{K} C^{(k)}\right\|_2\right] \leq 2K\,\mathbb{E}\left[\|C\|_2\right], \tag{40}$$

where C has the same distribution as $C^{(1)}$. Each entry of C is an independent copy of Y. Let F be the event that all the entries of C are in the range $[-4d, 4d]$. By a union bound,

$$\mathbb{P}\left(F^{c}\right) \leq\, n^2 \cdot \mathbb{P}\left(|Y| \geq 4d\right) \leq\, n^2 \cdot \mathbb{P}\left(\sum_{i=1}^{\binom{n-2}{d-2}} X_i \geq 4d\right).$$

Since $\sum_{i=1}^{\binom{n-2}{d-2}} X_i \sim \text{Bin}\left(\binom{n-2}{d-2}, p_{\max}\right)$, by (Mitzenmacher and Upfal, 2017, Theorem 4.4, Equation 4.1)

$$\mathbb{P}\left(F^{c}\right) \leq n^2 \cdot \mathbb{P}\left(\text{Bin}\left(\binom{n-2}{d-2}, \frac{c_0 \log n}{\binom{n-1}{d-1}}\right) \geq 4d\right) \leq n^2 \frac{e^{4d}}{\Theta(n/\log n)^{4d}} = O\left(\frac{1}{n^{4d-3}}\right). \tag{41}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\|C\|_2\right] &= \mathbb{P}\left(F\right)\mathbb{E}\left[\|C\|_2 \mid F\right] + \mathbb{P}\left(F^{\mathrm{c}}\right)\mathbb{E}\left[\|C\|_2 \mid F^{\mathrm{c}}\right] \\
&\leq \mathbb{E}\left[\|C\|_2 \mid F\right] + O\left(\frac{1}{n^{4d-3}}\right)\mathbb{E}[\|C\|_F \mid F^{\mathrm{c}}] \\
&\leq \mathbb{E}\left[\|C\|_2 \mid F\right] + O\left(\frac{1}{n^{4d-3}}\right) n\binom{n-2}{d-2} \\
&= \mathbb{E}\left[\|C\|_2 \mid F\right] + O\left(n^{3-4d+1+d-2}\right) \\
&\leq \mathbb{E}\left[\|C\|_2 \mid F\right] + o(1), \qquad (42)
\end{aligned}$$

where the second inequality uses the fact that each entry of C is at most $\binom{n-2}{d-2}$. Thus, it is only left to show $\mathbb{E}\left[\|C\|_2 \mid F\right] = O(\sqrt{\log n})$. Note that the entries of C are independent even after conditioning on F. Moreover, the entries are bounded in $[-4d, 4d]$. Thus

$$\begin{aligned}
\mathbb{E}\left[|C_{ij}| \mid F\right] &= \sum_{k=1}^{4d} k \cdot \mathbb{P}\left(|C_{ij}| = k \mid F\right) \leq \sum_{k=1}^{4d} k \cdot \frac{\mathbb{P}\left(F \cap |C_{ij}| = k\right)}{\mathbb{P}\left(F\right)} = \sum_{k=1}^{4d} k \cdot \frac{\mathbb{P}\left(|C_{ij}| = k\right)}{1 - o(1)} \\
&\lesssim \sum_{k=1}^{4d} k \cdot \mathbb{P}\left(|C_{ij}| = k\right) \leq \mathbb{E}\left[|C_{ij}|\right] \leq \binom{n-2}{d-2} p_{\max} = \frac{\binom{n-2}{d-2} c_0 \log n}{\binom{n-1}{d-1}} \lesssim \frac{\log n}{n}.
\end{aligned}$$

Therefore, by Lemma 24 and the fact that $\mathbb{E}\left[C \mid F\right]$ is the zero matrix, we obtain

$$\mathbb{E}\left[\|C\|_2 \mid F\right] = \mathbb{E}[\|C - \mathbb{E}\left[C \mid F\right]\|_2 \mid F] \leq c' \sqrt{n \frac{\log n}{n}} \leq c' \sqrt{\log n}.$$

Substituting this back in (40) and using (42),

$$\mathbb{E}\left[\|W - W^*\|_2\right] \leq 2K\, \mathbb{E}[\|C\|_2] \leq 2K\, \mathbb{E}[\|C\|_2 \mid F] + o(1) \leq 2Kc' \sqrt{\log n} + o(1) \leq c\sqrt{\log n},$$

for some c that depends on d and c_0, concluding the proof. ■