

Gap-Free Clustering: Sensitivity and Robustness of SDP

Matthew Zurek MATTHEW.ZUREK@WISC.EDU and **Yudong Chen** YUDONG.CHEN@WISC.EDU
Department of Computer Sciences, University of Wisconsin-Madison

Editors: Shipra Agrawal and Aaron Roth

Abstract

We study graph clustering in the Stochastic Block Model (SBM) in the presence of both large clusters and small, unrecoverable clusters. Previous convex relaxation approaches achieving exact recovery do not allow any small clusters of size $o(\sqrt{n})$, or require a size gap between the smallest recovered cluster and the largest non-recovered cluster. We provide an algorithm based on semidefinite programming (SDP) which removes these requirements and provably recovers large clusters regardless of the remaining cluster sizes. Mid-sized clusters pose unique challenges to the analysis, since their proximity to the recovery threshold makes them highly sensitive to small noise perturbations and precludes a closed-form candidate solution. We develop novel techniques, including a leave-one-out-style argument which controls the correlation between SDP solutions and noise vectors even when the removal of one row of noise can drastically change the SDP solution. We also develop improved eigenvalue perturbation bounds of potential independent interest. Our results are robust to certain semirandom settings that are challenging for alternative algorithms. Using our gap-free clustering procedure, we obtain efficient algorithms for the problem of clustering with a faulty oracle with superior query complexities, notably achieving $o(n^2)$ sample complexity even in the presence of a large number of small clusters. Our gap-free clustering procedure also leads to improved algorithms for recursive clustering.

Keywords: stochastic block model, semi-random models, clustering with a faulty oracle, semidefinite programming, sensitivity analysis, leave-one-out

1. Introduction

Graph clustering is a fundamental problem with various applications to the study of diverse real-world networks, such as social, biological, and information networks. A standard approach for benchmarking clustering algorithms is to consider probabilistic models for generating a graph with a planted set of unknown, ground-truth clusters. Beyond their modeling power of practical networks, the study of such models over the past decades has led to notable theoretical advancement in the interface of combinatorial optimization, statistical learning, information theory and applied probability, greatly enriching our understanding of computational and statistical tradeoffs. For a survey of the algorithmic and theoretical work in this area, see [Abbe \(2018\)](#).

In this paper, we consider the canonical model called Stochastic Block Model (SBM), also known as the Planted Partition Model ([Holland et al., 1983](#); [Condon and Karp, 2001](#)). We focus on the unbalanced clusters setting, formally defined below, where the planted clusters have different sizes.

As we discuss in greater details below, most existing results, with a few notable exceptions, require all clusters to be sufficiently large or the existence of a gap between the large and small clusters. The goal of this paper is to eliminate such assumptions.

2. Problem Setup and Prior Art

Let $[n] := \{1, \dots, n\}$ be a set of n vertices partitioned into K unknown clusters $V_1, \dots, V_K$. We observe a random graph on these vertices with adjacency matrix $A \in \{0, 1\}^{n \times n}$, where for two numbers $1 \geq p > q \geq 0$, and independently across all $i \leq j$,

$$A_{ij} \sim \begin{cases} \text{Bernoulli}(p), & \text{if } i, j \text{ belong to the same cluster,} \\ \text{Bernoulli}(q), & \text{if } i, j \text{ do not belong to the same cluster.} \end{cases}$$

and $A_{ij} = A_{ji}$ for all $i > j$. In other words, p is the intra-cluster edge probability and q is the inter-cluster edge probability. Note that p, q and K are allowed to depend on n .

We let $(s_1, \dots, s_K) = (|V_1|, \dots, |V_K|)$ be the cluster sizes, where the sizes are ordered as $s_1 \geq s_2 \geq \dots \geq s_K$. Since all our algorithms and analysis do not rely on the ordering of the vertices, we further assume, without loss of generality and for notational convenience, that the vertices are ordered by their cluster membership, so that $V_1 = \{1, \dots, s_1\}$, $V_2 = \{s_1 + 1, \dots, s_1 + s_2\}$, and so on. We encode the ground truth clustering using a matrix $Y^* \in \{0, 1\}^{n \times n}$, defined by setting $Y_{ij}^* = 1$ if and only if i and j belong to the same cluster. By our ordering assumption Y^* is block-diagonal and has the form $Y^* = \text{diag}(J_{s_1 \times s_1}, \dots, J_{s_K \times s_K})$, where $J_{m \times r}$ denotes an all-ones matrix of size $m \times r$. Note that the cluster sizes $\{s_i\}$ correspond to the eigenvalues of the Y^* . We assume that $p \geq \log n/n$, a common assumption in the literature and known to be necessary for exact recovery.

2.1. Algorithms and Prior Results for Unbalanced SBM

A plethora of algorithms have been developed for graph clustering, and many of them enjoy rigorous performance guarantees; see, for instance, [McSherry \(2001\)](#); [Bollobás and Scott \(2004\)](#); [Chen et al. \(2012\)](#); [Chaudhuri et al. \(2012\)](#); [Vu \(2018\)](#); [Abbe \(2018\)](#), and the references therein. Nearly all theoretical guarantees for the problem of exact recovery in the SBM require all clusters to be “large”, typically of size $\tilde{\Omega}(\sqrt{n})$. Under such an assumption, these algorithms recover all clusters. From theoretical and practical standpoints, it is desirable to have algorithms which, despite the presence of “small” clusters below this threshold, provably recover all sufficiently large clusters. While this is a natural goal, establishing such a result is surprisingly nontrivial. Only a few existing results allow the presence of small clusters, [Ailon et al. \(2013, 2015\)](#), [Mukherjee et al. \(2022\)](#), and [Mukherjee and Zhang \(2022\)](#). Postponing the discussion of the spectral-based methods [Mukherjee et al. \(2022\)](#) and [Mukherjee and Zhang \(2022\)](#) to after presenting our main theorem (in Section 3), we first discuss the limitations of all existing convex-relaxation-based approaches which permit small clusters, including [Ailon et al. \(2013, 2015\)](#).

For simplicity of presentation, in order to illustrate the limitations of all existing convex-relaxation-based approaches, we present only the guarantees of our own “warm-up” result which is of a very similar (but slightly improved and much simpler) form to [Ailon et al. \(2013, 2015\)](#), and defer discussion of the finer details of [Ailon et al. \(2013, 2015\)](#) to Section B.1. All our results are based on solving the following trace-regularized semi-definite program (SDP):

$$\hat{Y} = \underset{Y \in \mathbb{R}^{n \times n}}{\text{argmax}} \quad \left\langle Y, A - \frac{p+q}{2} J_{n \times n} \right\rangle - \lambda \text{Tr}(Y) \quad (1a)$$

$$\text{s.t.} \quad Y \succeq 0, \quad 0 \leq Y_{ij} \leq 1, \forall i \in [n], j \in [n] \quad (1b)$$

where λ is a regularization parameter which will be specified later. We refer to (1) as the *recovery SDP*. We note that this SDP and its variants, usually without the trace regularizer, are considered in

many prior works; we refer to the survey [Li et al. \(2021\)](#) and the references therein, as well as [Ames \(2014\)](#), [Chen and Xu \(2014\)](#), [Cai and Li \(2015\)](#), and [Amini and Levina \(2018\)](#).

Using this SDP, we can obtain the following recovery guarantee. In the sequel, we write $a_n \lesssim b_n$ or $a_n = O(b_n)$ if $a_n \leq Cb_n, \forall n$ for a universal numerical constant $C > 0$.

Theorem 1 (Informal version of Theorem 9) *If two consecutive cluster sizes $\bar{s} > \underline{s}$ satisfying*

$$(1 - \underline{s}/\bar{s})^2 (p - q)^2 \bar{s}^2 / pn \gtrsim \log n \quad (2)$$

then with high probability the solution to the recovery SDP (1) with a suitable λ is of the form

$$\hat{Y} = \text{diag} (J_{s_1 \times s_1}, \dots, J_{\bar{s} \times \bar{s}}, 0_{\underline{s} \times \underline{s}}, \dots, 0_{s_K \times s_K}). \quad (3)$$

That is, all big clusters are exactly recovered and all small clusters are completely ignored.

Like [Ailon et al. \(2013, 2015\)](#), this result allows for small clusters, but additionally assumes there exists a constant *multiplicative gap* between the sizes of large and small clusters.

We believe the essential limitation of both Theorem 1 and [Ailon et al. \(2013, 2015\)](#) that necessitates a gap assumption is not fundamentally algorithmic, but rather the form of their guarantees. Without assuming a gap in cluster sizes, there may exist clusters on the boundary of recoverability. In different typical samples from the same SBM instance, the recovery SDP (1) solution may or may not recover such clusters. Furthermore, the corresponding block of the SDP solution may not be all-one nor all-zero, making it impossible to guarantee an SDP solution of the form (3). This is demonstrated experimentally in Figure 1(b), where the largest cluster is recovered with all entries equal to 1, the middle cluster is recovered but with a block which has entries between zero and one, and the smaller clusters are ignored. Nonetheless, such a block-diagonal solution, albeit having some non-binary values, suffices to recover the large and middle clusters. In our experiments, the SDP solution always has this block-diagonal form regardless of the values of the cluster sizes. One might then hope to rigorously prove this property, which is left as an open question in [Ailon et al. \(2015\)](#).

3. Our Contributions

We completely resolve the above question by showing the recovery SDP (1) returns a desired block-diagonal matrix regardless of the distribution of cluster sizes. Our result completely eliminates the gap assumption, guaranteeing recovery of large clusters in the presence of arbitrary middle and small clusters. The middle, critically sized clusters greatly complicate the proof of this result. Not only do we no longer have a simple closed-form candidate solution for which we can verify optimality, but more significantly, threshold-sized clusters cause the SDP solution $\hat{Y}$ to be extremely sensitive to noise and thus unstable. In particular, as is seen both numerically and from our analysis, a small perturbation to the graph or cluster sizes may cause a zero block in $\hat{Y}$ to become near all-one, and vice versa. We overcome these challenges to give a very detailed description of the SDP solution. We believe our most important contributions are the ideas behind this analysis.

Algorithmically, the trace penalty term, which is ignored in most prior work, plays an important role in the recovery SDP (1). This term is not needed when all clusters are large, in which case the hard, entrywise constraint in (1b) is sufficient for controlling the diagonal entries of $\hat{Y}$. In the presence of small clusters, however, the trace penalty is essential for ensuring that the SDP solution only captures clusters with sufficient signal and ignores noise, thus yielding a block-diagonal solution.

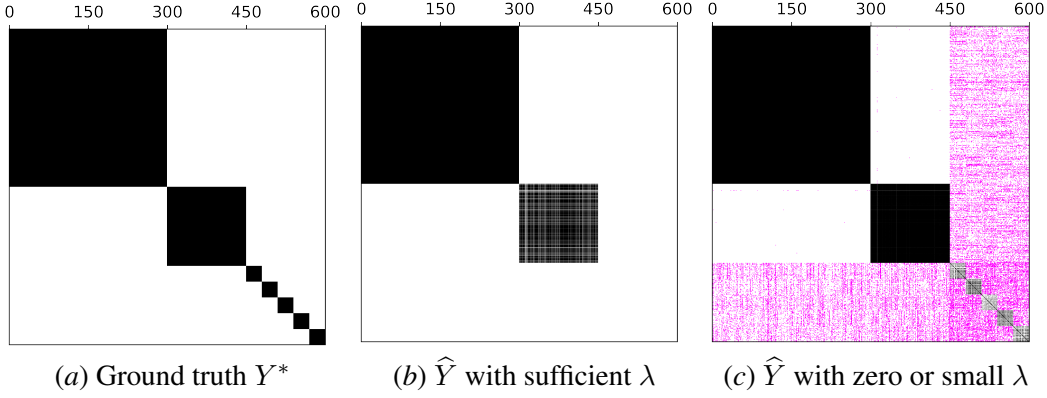


Figure 1: SDP solutions and effects of regularization λ . (a): Ground truth clusters of sizes 300, 150, and 50 ($\times 5$). (b)&(c): Solutions to SDP (1) with different λ . Nonzero off-block-diagonal entries are highlighted in pink; other entries shown in grayscale (white = 0, black = 1).

Figure 1 provides an experimental demonstration: Figure 1(c) uses an overly small λ and leads to a solution with nonzero entries in the off-block-diagonals shown in pink, making the two largest clusters nontrivial to recover. In contrast, Figure 1(b) shows a solution found with λ set sufficiently large so that all off-block-diagonal entries are zero.

We now formally present our main result, which characterizes the solution to the recovery SDP (1) in terms of the solutions to the following *oracle SDPs*, which are K smaller SDPs of the same form as the recovery SDP except that the full adjacency matrix is replaced by each of the K submatrices corresponding to only the within-cluster edges of each cluster: for $k = 1, \dots, K$, let $A^{(k)} = (A_{ij})_{i,j \in V_k} \in \mathbb{R}^{s_k \times s_k}$ and define the k th oracle SDP:

$$\hat{Y}^k = \underset{\substack{Y \in \mathbb{R}^{s_k \times s_k} \\ Y \succeq 0, 0 \leq Y \leq 1}}{\operatorname{argmax}} \left\langle Y, A^{(k)} - \frac{p+q}{2} J_{s_k \times s_k} \right\rangle - \lambda \operatorname{Tr}(Y). \quad (4)$$

Of course the formation of the oracle SDPs requires knowledge of the ground-truth clustering Y^* , so they are only usable for analysis. Now we can state our main theorem.

Theorem 2 *The following holds for some absolute constants $B, \underline{B}, \overline{B} > 2$: Fix m such that $m \geq n$ and $p \geq \frac{\log m}{n}$. Set the regularization parameter $\lambda = \kappa \sqrt{pn \log m} + \delta$ for $\kappa \geq B$ and $\delta \sim \text{Uniform}[0, 0.1]$. With probability $1 - O(m^{-3})$, the unique solution $\hat{Y}$ to recovery SDP (1) is*

$$\hat{Y} = \operatorname{diag}(\hat{Y}^1, \dots, \hat{Y}^K),$$

where $\hat{Y}^1, \dots, \hat{Y}^K$ are the unique solutions to the K oracle SDPs (4). Furthermore, for each $k \in \{1, \dots, K\}$, we have $\operatorname{rank} \hat{Y}^k \in \{0, 1\}$, and

1. if $\frac{p-q}{2} s_k \geq (1 + \frac{1}{\underline{B}}) \kappa \sqrt{pn \log m}$, then $\hat{Y}^k = J_{s_k \times s_k} = \mathbf{1}\mathbf{1}^\top$,
2. if $\frac{p-q}{2} s_k \leq (1 - \frac{1}{\underline{B}}) \kappa \sqrt{pn \log m}$, then $\hat{Y}^k = 0_{s_k \times s_k} = \mathbf{0}\mathbf{0}^\top$,
3. if $(1 - \frac{1}{\underline{B}}) \kappa \sqrt{pn \log m} < \frac{p-q}{2} s_k < (1 + \frac{1}{\underline{B}}) \kappa \sqrt{pn \log m}$, then either $\hat{Y}^k = 0$ or $\hat{Y}^k = y^k y^{k\top}$ where $y^k \geq \frac{1}{2}$ entrywise.

Theorem 2 states that the recovery SDP solution is block-diagonal, and the blocks matches the oracle SDP solutions. We also precisely characterize the oracle SDP solutions, which all have rank zero or one, and furthermore are guaranteed to be all-one if the cluster size s_k satisfies

$$(p - q)s_k \gtrsim \kappa \sqrt{pn \log m} \quad \text{or equivalently} \quad (p - q)^2 s_k^2 / pn \gtrsim \log m. \quad (5)$$

Comparing to the gap-dependent condition (2), there is no relative size gap term. With respect to the critically-sized clusters, those satisfying case 3, their oracle solution may be zero, in which case they are not recovered, but if the oracle solution is non-zero, it is all-positive, hence recoverable. Even when the nodes are not ordered, one can easily postprocess the SDP solution $\hat{Y}$ to recover the nonzero diagonal blocks promised by the theorem as the off-diagonal blocks are all zero.

Remark 3 *A few technical points about Theorem 2: Up to log factors and assuming $p/q = O(1)$, condition (5) is needed in all known polynomial-time algorithms even when all clusters have the same size (Abbe, 2018; Li et al., 2021). We include an extra parameter m for the failure probability to facilitate repeated application of the algorithm for recursive clustering and adaptive sampling. The random perturbation δ is not practically necessary, but simplifies the theorem conclusions by ensuring the oracle and recovery SDPs have unique solutions; it avoids the boundary situation that $\lambda_1(\vec{A}^{(k)}) = \lambda$ lies exactly on the recovery threshold. The number $(p + q)/2$ used in SDP (1) can be replaced by other values in the interval (q, p) and thus be treated as a tuning parameter along with λ ; similar observations were in prior work (e.g., Chen et al., 2018; Amini and Levina, 2018).*

We believe our most significant contributions are the ideas behind the proof of Theorem 2. With a more extensive discussion given in the proof sketch in Section 4, here we briefly highlight some key steps. The challenges are mainly associated with the analysis of the oracle SDPs for clusters near the threshold of recovery (case 3 of Theorem 2). The crucial step is to show that the oracle SDP has a rank-one solution, which is the case as long as the non-negativity constraints in (4) are non-binding. Defining a relaxed oracle SDP by removing these constraints, we reduce the problem to showing that the relaxed oracle SDP solutions are weakly correlated with individual rows of the random noise, suggesting a leave-one-out-style (LOO) argument. However, the naive application of LOO ideas does not work due to the instability of clusters near the recovery threshold—the removal of one row of noise can drastically change the corresponding SDP solution. To overcome this challenge we define modified LOO versions of the oracle SDPs, using an improved eigenvalue perturbation bound which may be of independent interest. We hope that some proof techniques may be useful in more generality for problems featuring mixtures of strong and weak signals without a gap between them.

Now we discuss Mukherjee et al. (2022) and Mukherjee and Zhang (2022), two recent works which also allows the presence of small clusters. Their algorithms, based on spectral clustering, also succeed without the assumption of a gap and recover clusters meeting similar to (5). In both cases, their algorithms and analysis are completely different to our SDP relaxation approach, making our work the first among a long line of convex-relaxation-based clustering methods to break the gap assumption. Beyond providing a different and arguably conceptually simpler clustering technique, our method inherits robustness properties generally unique to convex relaxation algorithms. We elaborate on this point in the following section.

3.1. Semirandom Robustness

To the best of our knowledge, no prior work has considered the semirandom model (Feige and Kilian, 2001) in the presence of small clusters, but by virtue of our convex relaxation approach, we can

tackle this challenging setting. We first define a semirandom model for the small cluster regime. We generate an adjacency matrix A' as before, but now allow a semirandom adversary to construct the observed adjacency matrix $A \in \{0, 1\}^{n \times n}$ by changing A' arbitrarily but restricted such that:

1. For any $i, j \in [n]$ such that i, j are in different clusters, $A_{ij} \leq A'_{ij}$.
2. For any $i, j \in [n]$ such that i, j are in the same cluster k , if $\frac{p-q}{2}s_k \geq \frac{3}{2}B\sqrt{pn \log n}$ then $A_{ij} \geq A'_{ij}$, otherwise $A_{ij} = A'_{ij}$.

Note B is the constant appearing in Theorem 2. Thus the semirandom adversary is allowed to arbitrarily delete edges between vertices of different clusters, and allowed to arbitrarily add edges within large clusters. We refer to this setting as the *large cluster semirandom model*.

The above model includes a still challenging *heterogeneous* setting as a special case: Suppose that for each $i, j \in [n]$ such that i and j are members of distinct clusters, there exist numbers $q_{ij} = q_{ji}$ such that $1 \geq p > q \geq q_{ij} \geq 0$, and that A is sampled such that independently across all $i \leq j$,

$$A_{ij} \sim \begin{cases} \text{Bernoulli}(p), & \text{if } i, j \text{ belong to the same cluster,} \\ \text{Bernoulli}(q_{ij}), & \text{if } i, j \text{ do not belong to the same cluster.} \end{cases}$$

By a coupling argument, there exists an adversary under which A has this distribution (Chen et al., 2014).

Despite seemingly easing the clustering task, the semirandom model disrupts many local structures of the SBM, foiling algorithms dependent on such structures. For instance, under the standard SBM, the expected degree of a node i in cluster k would satisfy the identity $\sum_{j \in [n]} \mathbb{E}A_{ij} = ps_k + q(n - s_k)$, in which case the observed degree of node i can be used to estimate the cluster size s_k . Mukherjee et al. (2022) use such a subroutine for estimating the largest cluster size (and the algorithm of Mukherjee and Zhang (2022) also requires knowledge of the largest cluster size but they do not discuss its estimation). Under the large cluster semirandom SBM, such procedures would not directly work. In contrast, the SDP (1) and our same Theorem 2 continue to work:

Theorem 4 *Under the large cluster semirandom SBM, Theorem 2 holds with $\kappa = B$ and $m = n$.*

We prove Theorem 4 in Section F, but it follows easily from the observation that the semirandom adversary cannot change the optimal solution of the SDP (1). Similar phenomena, that proofs for the SDP relaxation approach generalize (often automatically) to heterogeneous or semi-random generative models, have been observed in the literature and recognized as a testament of the robustness of the SDP approach; see Moitra et al. (2016); Amini and Levina (2018) and the references therein. Although the above large cluster semirandom model excludes *perturbations* to small clusters, all prior work on the semirandom model excluded the *presence* of small clusters whatsoever. Thus the large cluster semirandom adversary strictly generalizes previous work, and so we believe Theorem 4 is an important step in developing robust algorithms for unbalanced community detection.

3.2. Recursive Clustering

Theorem 2 demonstrates that we can recover sufficiently large clusters even in the presence of arbitrarily-sized smaller clusters. As first recognized in Ailon et al. (2015), this allows for the use of a recursive clustering procedure: whenever we recover some clusters, we can remove the corresponding nodes from the graph, thereby eliminating some noisy entries and lowering the detection threshold

for another application of the recovery procedure. Specifically in our case, recovering more nodes enables us to lower the regularization parameter λ . We can thus repeatedly apply our clustering algorithm to uncover smaller and smaller clusters, until no more can be found. We emphasize that this recursive clustering procedure operates using only one sample A from the SBM.

Here we informally discuss our recursive clustering results, and defer detailed presentation and comparison with Ailon et al. (2015) to Subsection B.3. Ailon et al. (2015) consider a setting where $n \rightarrow \infty$ and p, q are constant. They recover all but $o(n)$ nodes via a recursive clustering procedure, but since their clustering subroutine requires a multiplicative gap between cluster sizes, they can only accommodate a total number of clusters $K \lesssim \log n$ to ensure that a sufficient gap remains after each round. By contrast, our gap free result recovers $n - o(n)$ nodes even if K is polynomially large in n :

Theorem 5 (Informal version of Theorem 10) *Suppose that $K \lesssim \frac{n^{0.5-\alpha}}{\sqrt{\log n}}$ for some $\alpha < 0.5$. Then with high probability, our Algorithm 1 recovers all but $O(n^{1-2\alpha})$ nodes.*

3.3. Clustering With a Faulty Oracle

The problem of clustering with a faulty oracle was introduced by Mazumdar and Saha (2017) and has since received significant attention. It can be viewed as an adaptive sampling version of the SBM, where we try to recover clusters by querying $o(n^2)$ individual entries of the adjacency matrix. Formally, we assume there exist n vertices partitioned into K unknown clusters $V_1, \dots, V_K$ of different sizes. For some bias parameter $\delta > 0$, we assume there exists a faulty clustering oracle $\mathcal{O}$ which, upon being queried with a pair of vertices (i, j) , returns an answer distributed as

$$\mathcal{O}(i, j) \sim \begin{cases} \text{Bernoulli}(1/2 + \delta/2), & \text{if } i, j \text{ belong to the same cluster,} \\ \text{Bernoulli}(1/2 - \delta/2), & \text{if } i, j \text{ do not belong to the same cluster.} \end{cases}$$

Answers are independent for different pairs, and repeating the same query produces the same answer. Clustering with a faulty oracle has several important applications, such as crowdsourced entity resolution and edge prediction in social networks. We refer to Mazumdar and Saha (2017) for more discussion, and refer to Peng and Zhang (2021) and Mukherjee et al. (2022) for more in-depth comparison of prior work. We summarize existing algorithms which are computationally efficient and work for general K in Table 1, and defer more comments to Subsection B.4.

As identified by Mukherjee et al. (2022), all prior algorithms (except that of Mukherjee et al. (2022)) fail when the number of clusters K is large. In particular, $K = \Omega(n^{1/4})$ causes all algorithms in the first 5 rows of Table 1 to have sample complexity $\Omega(n^2)$ and a recovery threshold size of $\Omega(n)$, with smaller K still presenting issues for some algorithms. With a gap-free clustering subroutine, we provide an algorithm with a different type of guarantee which circumvents the above issues under large K . Letting s be an input parameter, our Algorithm 2 can recover all clusters of size at least s , and for an appropriate choice of s this leads to an $o(n^2)$ sample complexity:

Theorem 6 *There exist absolute constants C_1, C_2 such that for any parameter $s \geq C_2 \frac{\sqrt{n \log n}}{\delta}$, by setting $\gamma = C_2 \frac{n \log n}{s^2 \delta^2}$, with probability at least $1 - O(n^{-3})$, Algorithm 2 recovers all clusters of size at least s with query complexity $O\left(\frac{n^2 \log n}{s \delta^2} + \frac{n^4 \log^2 n}{s^4 \delta^4}\right)$.*

We formally define Algorithm 2 and give further discussion in Section B.4. All proofs for this section are provided in Section E. Note the constant C_1 is used within the algorithm. Algorithm 2 requires a

Query Complexity	Threshold	Reference
$O\left(\frac{nK \log n}{\delta^2} + \frac{nK^2 \log n}{\delta^4} \wedge \frac{K^5 \log^2 n}{\delta^8}\right)$	$\Omega\left(\frac{K^4 \log n}{\delta^2}\right)$	Mazumdar and Saha (2017)
$O\left(\frac{nK \log n}{\delta^2} + \frac{K^4 \log^2 n}{\delta^4}\right)$	$\Omega\left(\frac{n}{K}\right)$	Peng and Zhang (2021)
$O\left(\frac{nK \log n}{\delta^2} + \frac{K^{10} \log^2 n}{\delta^4}\right)$	$\Omega\left(\frac{K^4 \log n}{\delta^2}\right)$	Peng and Zhang (2021)
$O\left(\frac{n(K+\log n)}{\delta^2} + \frac{K^8 \log^3 n}{\delta^4}\right)$	$\Omega\left(\frac{K^4 \log n}{\delta^2}\right)$	Xia and Huang (2022)
$O\left(\frac{nK \log n}{\delta^2} + \frac{K^9 \log K \log^2 n}{\delta^{12}}\right)$	$\Omega\left(\frac{K^4 \log n}{\delta^6}\right)$	Pia et al. (2022)
$O\left(\frac{nK \log n}{\delta^2} + \frac{K^5 \log^2 n}{\delta^4}\right)$	$\Omega\left(\frac{K^2 \log n}{\delta^2}\right)$	Our Thm. 8
$O\left(\frac{n}{s} \frac{n \log n}{\delta^2} + \frac{n^4}{s^4} \frac{\log^2 n}{\delta^4}\right)^1$	s	Mukherjee et al. (2022), Our Thm. 6

Table 1: Comparison of algorithms for clustering with a faulty oracle which are computationally efficient and work for general K .

user-specified target cluster size s , but by using an adaptive resampling procedure, we can obtain the following instance-dependent query complexity without any prior knowledge of cluster sizes. Recall that s_1 is the largest cluster size.

Theorem 7 *There exist absolute constants C_1, C_2, C_3 such that as long as $s_1 \geq C_2 \frac{\sqrt{n \log n}}{\delta}$, with probability at least $1 - O(n^{-2})$, Algorithm 3 recovers a cluster of size at least $\frac{s_1}{C_3}$ with sample complexity $O\left(n \frac{\log n}{\delta^2} + \frac{n^4}{s_1^4} \frac{\log^2 n}{\delta^4}\right)$.*

We present Algorithm 3 in Section B.4. The strategy of obtaining instance-dependent query complexity using adaptive sampling has a crucial reliance on the gap-free recovery guarantee of Theorem 2, since our objective is to sample minimally so that only the largest cluster is recoverable, meaning the recovery procedure must not be sensitive to the presence of slightly smaller but still large clusters. We therefore hope that Theorem 2 lays the foundation for future instance-dependent algorithms for clustering with a faulty oracle, particularly in unbalanced cluster size settings.

Even if the number of clusters K is small, by using our Theorem 2 instead of gap-requiring clustering subroutines, we obtain the following result which improves previous query complexities and recovery thresholds, while also having a substantially simpler algorithm and proof. The Algorithm 4, along with further discussion of this result, is presented in Section B.4.

Theorem 8 *With probability at least $1 - O(n^{-2})$, Algorithm 4 recovers all clusters of size at least $\Omega\left(\frac{K^2 \log n}{\delta^2}\right)$ with query complexity $O\left(\frac{nK \log n}{\delta^2} + \frac{K^5 \log^2 n}{\delta^4}\right)$.*

3.4. Eigenvalue Perturbation Bounds

As elaborated in the proof sketch in Section 4, to overcome the instability of threshold-size cluster oracle SDP solutions, we modify standard leave-one-out techniques by also changing the regularization strength. This requires eigenvalue perturbation bounds which improve upon existing results and may be of independent interest. Here we compare our results to existing work Eldridge et al. (2018) in a simplified setting, and present the full versions, along with more background and detailed comparison

in our clustering context, in Section B.5. Let $M, H \in \mathbb{R}^{n \times n}$ be symmetric matrices, viewing H as a perturbation, and let v_1 be a top eigenvector of M . The classical Weyl's inequality gives that $|\lambda_1(M + H) - \lambda_1(M)| \leq \|H\|_{\text{op}}$. As argued by Eldridge et al. (2018), one can sometimes obtain a bound of $|v_1^\top H v_1|$, which can be significantly smaller than $\|H\|_{\text{op}}$ when H is a random perturbation. Assuming an eigengap $\lambda_1(M) - \lambda_2(M) \geq 2 \|H\|_{\text{op}}$, Eldridge et al. (2018) show

$$\lambda_1(M + H) \leq \lambda_1(M) + v_1^\top H v_1 + \frac{2 \|H\|_{\text{op}}^2}{\lambda_1(M) - \lambda_2(M)} \quad (6)$$

Under this same assumption, our Theorem 12 improves this to

$$\lambda_1(M + H) \leq \lambda_1(M) + v_1^\top H v_1 + \frac{2 \|H v_1\|_2^2}{\lambda_1(M) - \lambda_2(M)}. \quad (7)$$

When H is random, we often have $\|H v_1\|_2 \ll \|H\|_{\text{op}}$, making (7) a substantial improvement.

Notation For an $n \times n$ matrix M , $M^{(ij)}$ is the $s_i \times s_j$ submatrix obtained by deleting all rows and columns not associated with cluster i and all columns not associated with cluster j . Let $M^{(i)} = M^{(ii)}$. For a vector $v \in \mathbb{R}^n$, $v^{(i)} \in \mathbb{R}^{s_i}$ is obtained by deleting all entries not associated with cluster i . We use M_{ij} and v_i to denote the entries of M and v . The convention is that $v_j^{(i)} = (v^{(i)})_j$. For two matrices X and Y , $\langle X, Y \rangle := \sum_{ij} X_{ij} Y_{ij}$ denotes their inner product and $X \circ Y$ their element-wise product. $\|X\|_{\text{op}}$ is the largest singular value of X ; $\|X\|_\infty := \max_{i,j} |X_{ij}|$. Denote by $\text{diag}(M_1, \dots, M_m)$ a block-diagonal matrix with diagonal blocks $M_1, \dots, M_m$. We use $\mathbf{1}_m$ to denote a length- m vector of all ones, occasionally dropping subscripts.

4. Proof Outline of Main Theorem

In this section, we sketch proof of Theorem 2. The general strategy is a primal-dual witness approach: we construct a candidate primal solution to the recovery SDP (1) as well as dual variables which certify its optimality. The construction relies on showing that the oracle SDPs (4) have rank-one and “well-spread” solutions, which is most challenging for the mid-size oracle blocks. As detailed below, we start by observing the rank-one property holds if the lower-bound constraints $\hat{Y}^k \geq 0$ are non-binding, which is in turn closely related to controlling the correlation between the oracle SDP solution and each row of the oracle noise matrix $W^{(k)}$. This suggests a leave-one-out-style argument, where the correlation strength is quantified by the change of the solution when a row of $W^{(k)}$ is left out. However, due to the arbitrary proximity of the mid-size clusters to the recovery boundary, the standard argument of zeroing one row of noise (see, e.g., Zhong and Boumal (2018)) may have a drastically different solution. To remedy this issue, we also modify the regularization strength, and the correct modification requires improved eigenvalue perturbation bounds in order to ensure that the leave-one-out solution and the oracle SDP solution are suitably close.

In this sketch we simply set $m = n$. Our regularization parameter λ will be at least $\kappa \sqrt{pn \log n}$ for some κ which is a fixed absolute constant which we can make arbitrarily large. We will write ε in place of terms which are $\asymp \frac{1}{\kappa}$, which may be multiple (unequal) terms in the same expression.

Setting up primal-dual witness argument We seek to show the solution is block-diagonal with support contained within that of the ground truth Y^* , but we do not have closed forms for each block.

This motivates us to consider the K oracle SDPs (4) associated with each cluster and combine their respective solutions $\hat{Y}^1, \dots, \hat{Y}^K$ to form our candidate solution $\hat{Y} = \text{diag}(\hat{Y}^1, \dots, \hat{Y}^K)$.

To show optimality of $\hat{Y}$ to the recovery SDP, it suffices to find nonnegative dual variables $U, L \in \mathbb{R}^{n \times n}$ satisfying the KKT conditions:

$$U \circ (\hat{Y} - J_{n \times n}) = L \circ \hat{Y} = 0, \quad \overleftrightarrow{A} - \lambda I - U + L \preceq 0, \quad (\overleftrightarrow{A} - \lambda I - U + L) \hat{Y} = 0. \quad (8)$$

On the other hand, each oracle SDP solution $\hat{Y}^k$ must satisfy its own similar KKT conditions: there exist nonnegative dual variables $U^k, L^k \in \mathbb{R}^{s_k \times s_k}$ which, together with the oracle block $\overleftrightarrow{A}^{(k)}$, satisfy similar equations as (8). To find a solution for the recovery conditions (8), we can set $U = \text{diag}(U^1, \dots, U^K)$, and it remains to choose the non-block-diagonal entries of L . This choice cannot be made without first gaining more detailed information about the oracle solutions $\hat{Y}^k$, particularly that they are rank-one and “well-spread.” The reasons will be made more clear once we fully construct L , but for now we can motivate these requirements with the following observations: Suppose for simplicity that $K = 2$. The second equation in the oracle SDP version of (8) guarantees that each unit norm eigenvector $v \in \mathbb{R}^{s_1}$ of $\hat{Y}^1$ will satisfy $(\overleftrightarrow{A}^{(1)} - U^1 + L^1)v = \lambda v$. Now letting $\bar{v} \in \mathbb{R}^n$ be a zero-padded version of v such that $\bar{v}^{(1)} = v$, we will have

$$(\overleftrightarrow{A} - U + L)\bar{v} = \begin{bmatrix} \overleftrightarrow{A}^{(1)} - U^1 + L^1 & \overleftrightarrow{A}^{(12)} + L^{(12)} \\ \overleftrightarrow{A}^{(21)} + L^{(21)} & \overleftrightarrow{A}^{(2)} - U^2 + L^2 \end{bmatrix} \begin{bmatrix} v \\ 0 \end{bmatrix} = \begin{bmatrix} \lambda v \\ (\overleftrightarrow{A}^{(21)} + L^{(21)})v \end{bmatrix}. \quad (9)$$

The right hand side has norm at least λ , so the only way to satisfy the last equation in (8) is for $(\overleftrightarrow{A}^{(21)} + L^{(21)})v = 0$. Therefore, each eigenvector (corresponding to a nonzero eigenvalue) of $\hat{Y}^1$ constrains $L^{(21)}$, and in order for the entries of $\overleftrightarrow{A}^{(21)}v = (-\frac{p-q}{2}J_{s_2 \times s_1} + W^{(21)})v$ to concentrate around $-\frac{p-q}{2}$ so that we may choose $L^{(21)} \geq 0$, we will need $\|v\|_\infty$ to be sufficiently small.

Analyzing oracle SDPs Now we can discuss how to establish these facts about the oracle SDPs. Fix a cluster k . If the signal $\frac{p-q}{2}s_k$ is sufficiently small or large, everything is relatively simple. If $\frac{p-q}{2}s_k \leq (1-\varepsilon)\lambda$, the top eigenvalue of $\overleftrightarrow{A} - \lambda I$ will be negative, so $\hat{Y}^k$ will be zero. If $\frac{p-q}{2}s_k \geq (1+\varepsilon)\lambda$, we can apply the gap-dependent recovery Theorem 9 to guarantee that $\hat{Y}^k = J_{s_k \times s_k}$ (since, with only one cluster, we need not worry about smaller clusters being sufficiently small). This leaves the most interesting mid-size case, where $\frac{p-q}{2}s_k$ is within a small multiple of λ .

Showing relaxed oracle SDP has rank-one solution In this case, there is no closed-form candidate for $\hat{Y}^k$ independent of the realization of A ; entries may be neither 0 nor 1. Experimentally, when λ is sufficiently large, the oracle solutions have no zero entries (unless the entire block is zero), meaning the non-negativity constraints are non-binding. We consider the oracle SDP (4) with these constraints removed, and refer to the resulting program the *Relaxed Oracle SDP*. A key simplification in the relaxed oracle SDP is that the last two conditions in the oracle SDP version of (8) become

$$\overleftrightarrow{A}^{(k)} - \lambda I - \text{diag}(u^k) \preceq 0_{s_k \times s_k}, \quad (\overleftrightarrow{A}^{(k)} - \lambda I - \text{diag}(u^k))\bar{Y}^k = 0_{s_k \times s_k}. \quad (10)$$

Since $\overleftrightarrow{A}^{(k)} = \frac{p-q}{2}J + W^{(k)}$, that is a noisy perturbation of a rank-one matrix, and also since $\text{diag}(\bar{u}^k) \succcurlyeq 0$, if $\lambda > \|W^{(k)}\|_{\text{op}}$ then we will be guaranteed that $\overleftrightarrow{A}^{(k)} - \lambda I - \text{diag}(\bar{u}^k)$ has all eigenvalues strictly negative except for at most one (which by first equation in (10) must then be equal to zero). Thus we can combine this with the fact that $\bar{Y}^k \succcurlyeq 0_{s_k \times s_k}$ and second equation in (10)

to conclude that $\bar{Y}^k$ must have rank zero or one. Thus we can write $\bar{Y}^k = yy^\top$ (choosing the sign of y so that $\mathbf{1}^\top y \geq 0$). We note that this step is somewhat related to [Sagnol \(2011\)](#) which shows a similarly structured SDP has low-rank solution, but [Sagnol \(2011\)](#) assumes that the objective matrix is low-rank, while we instead use the fact that it has a low number of positive eigenvalues.

Setting up leave-one-out technique to control noise correlation To justify using the relaxed oracle SDP, we must show that $\bar{Y}^k$ or equivalently y are elementwise-non-negative, in which case the relaxed oracle SDP solution will also be feasible for the oracle SDP, and therefore the optimal solution for the oracle SDP will be $\bar{Y}^k = yy^\top$. This is easy to check when $\lambda_1(\bar{A}^{(k)}) < \lambda$, since then $\bar{Y}^k = 0_{s_k \times s_k}$, so we can focus on the case where $\lambda_1(\bar{A}^{(k)}) > \lambda$ and $\bar{Y}^k$ is nonzero. Now we argue why this reduces to bounding the correlation between y and the rows of the noise matrix $W^{(k)}$.

Since the argument is identical for all entries, we focus on showing that $y_1 \geq 0$. If $y_1 = 1$ then we are done, and otherwise either $y_1 \in (-1, 1)$ so $u_1^k = 0$, or $y_1 = -1$ and then $-u_1^k y_1 \geq 0$. By using the fact that $\bar{Y}^k$ is rank-one and rearranging the optimality condition (10), we obtain $y_1 \geq \frac{1}{\lambda} \left(\frac{p-q}{2} \mathbf{1}^\top y + w^\top y \right)$, where $w^\top$ is the first row of $W^{(k)}$. Since in a mid-size cluster $(1-\varepsilon)\lambda \leq \frac{p-q}{2} s_k \leq (1+\varepsilon)\lambda$, and it is straightforward to show $\mathbf{1}^\top y \geq s(1-\varepsilon)$, the first term on the RHS is close to one. Therefore it suffices to show that w and y are uncorrelated, specifically that $|w^\top y| \lesssim \varepsilon \lambda$.

The entries of w are bounded by 1 and each have variance $\leq p$, while y has $\|y\|_2 \leq \sqrt{s_k}$ and $\|y\|_\infty \leq 1$, so if they were independent we could apply Bernstein's inequality to conclude $|w^\top y| \lesssim \sqrt{ps_k \log n} \leq \varepsilon \lambda$. The challenge is that they are not independent. Furthermore, naively bounding $|w^\top y| \leq \|w\|_2 \|y\|_2 \lesssim (\sqrt{ps_k})(\sqrt{s_k})$ does not work, because this bound can be much larger than $\lambda = \kappa \sqrt{pn \log n}$. Still, the guiding intuition is that since y needs to depend on all rows of $W^{(k)}$, it should not be strongly correlated with any single row. This is the standard motivation for the leave-one-out technique: to formalize this intuition, we can form a LOO relaxed oracle SDP where we set the noise in the first row and column of $\bar{A}^{(k)}$ to zero, and let its solution (which is also rank-one) be $\tilde{y}\tilde{y}^\top$. Then $\tilde{y}$ will be independent of w , but we might hope it will also be very close to y , two facts which we could combine to show $|w^\top y| \approx |w^\top \tilde{y}| \lesssim \sqrt{ps_k \log n}$.

Due to the extreme noise sensitivity inherent to the mid-size clusters, this strategy will not work: after leaving out noise to get $\tilde{A}$, we may have $\lambda_1(\tilde{A}) < \lambda$ (even though $\lambda_1(\bar{A}^{(k)}) > \lambda$), which means we will have solution $\tilde{y}\tilde{y}^\top = 0_{s_k \times s_k}$ for the LOO relaxed oracle SDP. In simpler terms, the usual LOO solution $\tilde{y}$ might be very different from y . To overcome this issue, we will *slightly* reduce the regularization parameter λ for the LOO problem, so that $\lambda_1(\tilde{A}) > \tilde{\lambda}$. Of course, if $\tilde{\lambda}$ is too small then $\tilde{y}$ might be significantly different from y . Existing eigenvalue perturbation bounds from ([Eldridge et al., 2018](#), Theorem 6) are too large to use in our remaining arguments, but our Theorem 12 gives a much smaller bound which ensures $\tilde{\lambda}$ is sufficiently close to λ . For the concrete comparison between these two bounds in our setting, see Section B.5.

Showing leave-one-out solution is close to relaxed oracle SDP solution With this resolved, we write $y = \alpha \tilde{y} + \beta t$ where t is a unit vector orthogonal to $\tilde{y}$. It is easy to show $|\alpha - 1| \lesssim \varepsilon$. We use this decomposition because there is a natural method to bound the norm of the rejection $\beta = \|y - \text{proj}_{\tilde{y}} y\|_2$. One interpretation of Lagrangian relaxation is that the optimal dual variables provide “bonuses” subtracted from the objective values of points which satisfy constraints, while preserving the optimal solution and the optimal value. Forming the Lagrangian for the LOO relaxed oracle SDP (after converting to a minimization problem for consistency with optimization conventions), the objective becomes $\mathcal{L}(Y, \tilde{u}, \tilde{P}) = -\langle \tilde{A}^{(k)} - \tilde{\lambda} I, Y \rangle + \langle \text{diag}(\tilde{u}), Y - I \rangle + \langle -\tilde{P}, Y \rangle$,

where $\tilde{P} = -(\tilde{A} - \tilde{\lambda}I - \text{diag}(\tilde{u}))$ is the dual variable satisfying the KKT conditions corresponding to the PSD constraint $Y \succcurlyeq 0$ (which does not appear in other arguments because we eliminate it using the stationarity condition).

Now using optimality of $\tilde{y}\tilde{y}^\top$ and feasibility of $yy^\top$, we can derive $\langle \tilde{P}, yy^\top \rangle \leq -\langle \tilde{A} - \tilde{\lambda}I, yy^\top - \tilde{y}\tilde{y}^\top \rangle$. Similar to our earlier arguments, we can show that $\tilde{P}\tilde{y} = 0$, while all other eigenvalues of $\tilde{P}$ are positive and at least $(1 - \varepsilon)\lambda$. Thus the bonus term directly measures $y - \text{proj}_{\tilde{y}} y$ and is at least $(1 - \varepsilon)\lambda\beta^2$. We would like the RHS to involve the difference between $\tilde{A}$ and $\overleftrightarrow{A}^{(k)}$ (which we know is small), so we can add the term $-\langle \overleftrightarrow{A}^{(k)} - \lambda I, \tilde{y}\tilde{y}^\top - yy^\top \rangle$ which is non-negative because it is the suboptimality of $\tilde{y}\tilde{y}^\top$ for the relaxed oracle SDP. This yields the final perturbation bound

$$\langle \tilde{P}, yy^\top \rangle \leq \langle (\tilde{A} - \tilde{\lambda}I) - (\overleftrightarrow{A}^{(k)} - \lambda I), \tilde{y}\tilde{y}^\top - yy^\top \rangle. \quad (11)$$

By bounding the RHS of (11) in terms of β and combining with our earlier arguments, we obtain a relationship which implies that $\beta \lesssim \varepsilon$. We note that this step is the key location where we must use the improved eigenvalue perturbation bound on $|\lambda - \tilde{\lambda}|$. Because this perturbation bound (11) is essential for controlling the difference between y and $\tilde{y}$, we hope that our explanation of the simple underlying principles could be useful for other settings involving leave-one-out analysis.

Wrapping up, while we could use our bounds on α and β to show $|w^\top y| \lesssim \varepsilon\lambda$, we simply use the fact $\tilde{y}_1 \geq 1 - \varepsilon$ since there is no noise in the first row, so $y_1 \geq \alpha\tilde{y}_1 - |\beta t_1| \geq (1 - \varepsilon)(1 - \varepsilon) - \varepsilon \geq 1 - \varepsilon$. This proves that the oracle SDP solution $\hat{Y}^k$ is rank-one (when nonzero) for mid-size clusters.

Constructing dual variables to complete primal-dual witness argument Now we can complete the proof by constructing L which satisfies the KKT conditions of the recovery SDP (8) as outlined in the beginning of this proof sketch. We discuss one final technical consideration. Consider again for simplicity the case that there are $K = 2$ clusters. Above, we argued using (9) why it was essential for the non-block-diagonal entries of L to be set in such a way that whenever $yy^\top$ is a non-zero solution for the first oracle SDP, then the zero-padded vector $\bar{y}/\|y\|_2 \in \mathbb{R}^n$ (satisfying $\bar{y}^{(k)} = y$) is an eigenvector of $\overleftrightarrow{A} + L - U$. Even when the first oracle SDP is zero and we have $U^1 = L^1 = 0$, if the first cluster is near the recovery threshold, it is possible to have $\lambda_1(\overleftrightarrow{A}^{(k)}) = (1 - \varepsilon)\lambda$. Now letting $v \in \mathbb{R}^{s_1}$ be the top eigenvector of $\overleftrightarrow{A}^{(1)} - U^1 + L^1 = \overleftrightarrow{A}^{(1)}$ (and $\bar{v} \in \mathbb{R}^n$ its zero-padded version), we have a nearly identical situation to that in (9) because $(\overleftrightarrow{A} - U + L)\bar{v} = \left[(1 - \varepsilon)\lambda v \quad (\overleftrightarrow{A}^{(21)} + L^{(21)})v \right]^\top$. The vector on the right hand side has norm at least close to λ , which again places a constraint on our choice of $L^{(21)}$ as we will need to ensure that $\|(\overleftrightarrow{A}^{(21)} + L^{(21)})v\|_2$ is very small. For this reason, we will treat clusters which are slightly below the recovery threshold identically to how we treat clusters which have nonzero oracle solutions (and thus are recovered), except we must use the top eigenvector of $\overleftrightarrow{A}^{(k)}$ instead of the y from the k th oracle solution $yy^\top$ (which, by the oracle SDP version of (8), is a top eigenvector of $\overleftrightarrow{A}^{(k)} - U^k + L^k$).

Acknowledgments

Y. Chen and M. Zurek were supported in part by National Science Foundation grants CCF-2233152 and DMS-2023239. We thank Lijun Ding and Yiqiao Zhong for inspiring discussion.

References

- Emmanuel Abbe. Community Detection and Stochastic Block Models: Recent Developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018. URL <http://jmlr.org/papers/v18/16-5480.html>.
- 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. IEEE, 2015.
- 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, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*, 2017.
- Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *The Annals of Statistics*, 48(3):1452–1474, June 2020. ISSN 0090-5364, 2168-8966. doi: 10.1214/19-5AOS1854. Publisher: Institute of Mathematical Statistics.
- Naman Agarwal, Afonso S Bandeira, Konstantinos Koiliaris, and Alexandra Kolla. Multisection in the stochastic block model using semidefinite programming. In *Compressed Sensing and its Applications: Second International MATHEON Conference 2015*, pages 125–162. Springer, 2017.
- Nir Ailon, Yudong Chen, and Huan Xu. Breaking the small cluster barrier of graph clustering. In *International Conference on Machine Learning*, pages 995–1003, 2013.
- Nir Ailon, Yudong Chen, and Huan Xu. Iterative and active graph clustering using trace norm minimization without cluster size constraints. *The Journal of Machine Learning Research*, 16(1): 455–490, 2015.
- Brendan P. Ames. Guaranteed clustering and biclustering via semidefinite programming. *Mathematical Programming: Series A and B*, 147(1-2):429–465, October 2014. ISSN 0025-5610. doi: 10.1007/s10107-5013-50729-5x. URL <https://doi.org/10.1007/s10107-5013-50729-5x>.
- Arash A. Amini and Elizaveta Levina. On semidefinite relaxations for the block model. *The Annals of Statistics*, 46(1):149–179, February 2018. ISSN 0090-5364, 2168-8966. doi: 10.1214/17-5AOS1545. Publisher: Institute of Mathematical Statistics.
- Afonso S Bandeira and Ramon Van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. 2016.
- B. Bollobás and A. D. Scott. Max Cut for Random Graphs with a Planted Partition. *Combinatorics, Probability and Computing*, 13(4-5):451–474, July 2004. ISSN 0963-5483, 1469-2163. doi: 10.1017/S0963548304006303. URL http://www.journals.cambridge.org/abstract_S0963548304006303.

- Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1347–1357. IEEE, 2015.
- Stephen P. Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK ; New York, 2004. ISBN 978-0-521-83378-3.
- T. Tony Cai and Xiaodong Li. Robust and Computationally Feasible Community Detection in the Presence of Arbitrary Outlier Nodes. *The Annals of Statistics*, 43(3):1027–1059, 2015. ISSN 0090-5364. URL <https://www.jstor.org/stable/43556546>. Publisher: Institute of Mathematical Statistics.
- Francesco Caltagirone, Marc Lelarge, and Léo Miolane. Recovering asymmetric communities in the stochastic block model. *IEEE Transactions on Network Science and Engineering*, 5(3):237–246, 2017.
- Ted Carson and Russell Impagliazzo. Hill-climbing finds random planted bisections. In *Proceedings of the 12th Annual SIAM Symposium on Discrete Algorithms*. Citeseer, 2001.
- Kamalika Chaudhuri, Fan Chung, and Alexander Tsiatas. Spectral Clustering of Graphs with General Degrees in the Extended Planted Partition Model. In *Proceedings of the 25th Annual Conference on Learning Theory*, pages 35.1–35.23. JMLR Workshop and Conference Proceedings, June 2012. URL <https://proceedings.mlr.press/v23/chaudhuri12.html>. ISSN: 1938-7228.
- Yudong Chen and Jiaming Xu. Statistical-computational phase transitions in planted models: the high-dimensional setting. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML’14, pages II–244–II–252, Beijing, China, June 2014. JMLR.org.
- Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering Sparse Graphs. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://papers.nips.cc/paper_files/paper/2012/hash/1e6e0a04d20f50967c64dac2d639a577-5Abstract.html.
- Yudong Chen, Sujay Sanghavi, and Huan Xu. Improved graph clustering. *IEEE Transactions on Information Theory*, 60(10):6440–6455, 2014.
- Yudong Chen, Xiaodong Li, and Jiaming Xu. Convexified Modularity Maximization for Degree-Corrected Stochastic Block Models. *Annals of Statistics*, 46(4):1573–1602, 2018.
- 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.
- Amin Coja-Oghlan, Florent Krzakala, Will Perkins, and Lenka Zdeborová. Information-theoretic thresholds from the cavity method. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 146–157, 2017.
- Anne Condon and Richard M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Structures and Algorithms*, 18(2):116–140, 2001.

- Lijun Ding and Yudong Chen. Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Transactions on Information Theory*, 66(11):7274 – 7301, 2020.
- Justin Eldridge, Mikhail Belkin, and Yusu Wang. Unperturbed: spectral analysis beyond davis-kahan. In *Algorithmic Learning Theory*, pages 321–358. PMLR, 2018.
- Uriel Feige and Joe Kilian. Heuristics for semirandom graph problems. *Journal of Computer and System Sciences*, 63(4):639–671, 2001.
- Kasper Green Larsen, Michael Mitzenmacher, and Charalampos Tsourakakis. Clustering with a faulty oracle. In *Proceedings of The Web Conference 2020, WWW ’20*, pages 2831–2834, New York, NY, USA, April 2020. Association for Computing Machinery. ISBN 978-1-4503-7023-3. doi: 10.1145/3366423.3380045. URL <https://doi.org/10.1145/3366423.3380045>.
- Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.
- Paul W. Holland, Kathryn B. Laskey, and Samuel Leinhardt. Stochastic blockmodels: Some first steps. *Social Networks*, 5:109–137, 1983.
- Mark Jerrum and Gregory B Sorkin. The metropolis algorithm for graph bisection. *Discrete Applied Mathematics*, 82(1-3):155–175, 1998.
- Varun Jog and Po-Ling Loh. Information-theoretic bounds for exact recovery in weighted stochastic block models using the renyi divergence. *arXiv preprint arXiv:1509.06418*, 2015.
- Michael Krivelevich and Dan Vilenchik. Semirandom models as benchmarks for coloring algorithms. In *Third Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pages 211–221, 2006.
- Marc Lelarge, Laurent Massoulié, and Jiaming Xu. Reconstruction in the labelled stochastic block model. *IEEE Transactions on Network Science and Engineering*, 2(4):152–163, 2015.
- Xiaodong Li, Yudong Chen, and Jiaming Xu. Convex relaxation methods for community detection. *Statistical Science*, 36(1):2–15, 2021.
- Yu Lu and Harrison H. Zhou. Statistical and computational guarantees of Lloyd’s algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.
- Albert W Marshall, Ingram Olkin, and Barry C Arnold. Inequalities: theory of majorization and its applications. 1979.
- Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703, 2014.
- Arya Mazumdar and Barna Saha. Clustering with Noisy Queries. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/db5cea26ca37aa09e5365f3e7f5dd9eb-3Paper.pdf.

- F. McSherry. Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 529–537, October 2001. doi: 10.1109/SFCS.2001.959929. ISSN: 1552-5244.
- Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, Cambridge, 2005. doi: 10.1017/CBO9780511813603. URL <https://www.cambridge.org/core/books/probability-and-computing/3A5B47DB315FC64B9256C5C8131C5EFA>.
- Ankur Moitra, William Perry, and Alexander S. Wein. How robust are reconstruction thresholds for community detection? In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 828–841. ACM, 2016.
- Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs and their application to community detection. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 814–827, 2016.
- Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 69–75, 2015a.
- Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162:431–461, 2015b.
- Chandra Sekhar Mukherjee and Jiapeng Zhang. Detecting Hidden Communities by Power Iterations with Connections to Vanilla Spectral Algorithms, November 2022. URL <http://arxiv.org/abs/2211.03939>. arXiv:2211.03939 [math] version: 1.
- Chandra Sekhar Mukherjee, Pan Peng, and Jiapeng Zhang. Recovering unbalanced communities in the stochastic block model with application to clustering with a faulty oracle. *arXiv preprint arXiv:2202.08522*, 2022.
- Pan Peng and Jiapeng Zhang. Towards a query-optimal and time-efficient algorithm for clustering with a faulty oracle. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3662–3680. PMLR, 15–19 Aug 2021. URL <https://proceedings.mlr.press/v134/peng21a.html>.
- Amelia Perry and Alexander S Wein. A semidefinite program for unbalanced multisection in the stochastic block model. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 64–67. IEEE, 2017.
- Alberto Del Pia, Mingchen Ma, and Christos Tzamos. Clustering with Queries under Semi-Random Noise. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 5278–5313. PMLR, June 2022. URL <https://proceedings.mlr.press/v178/pia22a.html>. ISSN: 2640-3498.
- Guillaume Sagnol. A class of semidefinite programs with rank-one solutions. *Linear Algebra and its Applications*, 435(6):1446–1463, September 2011. ISSN 0024-3795. doi: 10.1016/j.laa.

2011.03.027. URL <https://www.sciencedirect.com/science/article/pii/S0024379511002515>.

Tom AB Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification*, 14(1):75–100, 1997.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018. ISBN 978-1-108-41519-4. doi: 10.1017/9781108231596. URL <https://www.cambridge.org/core/books/highdimensional-probability/797C466DA29743D2C8213493BD2D2102>.

Van Vu. A Simple SVD Algorithm for Finding Hidden Partitions. *Combinatorics, Probability and Computing*, 27(1):124–140, January 2018. ISSN 0963-5483, 1469-2163. doi: 10.1017/S0963548317000463. URL <https://www.cambridge.org/core/journals/combinatorics-probability-and-computing/article/abs/simple-svd-algorithm-for-finding-hidden-partitions/3691FFE262617493CEC469769154EA19>. Publisher: Cambridge University Press.

Jinghui Xia and Zengfeng Huang. Optimal Clustering with Noisy Queries via Multi-Armed Bandit, July 2022. URL <http://arxiv.org/abs/2207.05376>. arXiv:2207.05376 [cs].

Yiqiao Zhong and Nicolas Boumal. Near-Optimal Bounds for Phase Synchronization. *SIAM Journal on Optimization*, 28(2):989–1016, January 2018. ISSN 1052-6234. doi: 10.1137/17M1122025. URL <https://epubs.siam.org/doi/10.1137/17M1122025>. Publisher: Society for Industrial and Applied Mathematics.

Appendix A. Additional Notation

We introduce some additional notation used in the rest of the paper. For a matrix X , $\|X\|_{\text{nuc}}$ is its nuclear norm (the sum of singular values), $\|X\|_1 := \sum_{i,j} |X_{ij}|$ is its entrywise ℓ_1 norm, and $\text{supp}(X) := \{(i, j) : X_{ij} \neq 0\}$ is its support. We define $\bar{e} = \mathbf{1}/\|\mathbf{1}\|_2$ and we let e_j denote a standard basis vector with entry j , both of which are only used when their size is clear from context. For any event E we define E^c to be its complement. For any natural number h we define $[h] = \{1, \dots, h\}$.

Appendix B. Additional Discussion of Results and Related Work

B.1. Further Details of Ailon et al. and Gap-Dependent Clustering

Ailon et al. (2013, 2015) consider a convex relaxation approach based on semidefinite programming. As discussed in the introduction, their results allow for small clusters, but additionally assume there exists a constant *multiplicative gap* between the sizes of large and small clusters. Again, letting $\bar{s} > \underline{s}$ be two consecutive cluster sizes (i.e., there are no clusters with sizes in $(\underline{s}, \bar{s})$), they guarantee exact recovery of all clusters with sizes at least $\bar{s}$ as long as the following conditions are satisfied:

$$\frac{(p-q)^2 \bar{s}^2}{pn} \gtrsim \log n \quad \text{and} \quad \bar{s}/\underline{s} \gtrsim 1. \quad (12)$$

Under the above condition and after reordering nodes by cluster identity, their optimization program returns a block-diagonal solution of the form

$$\hat{Y} = \text{diag} (J_{s_1 \times s_1}, \dots, J_{\bar{s} \times \bar{s}}, 0_{\underline{s} \times \underline{s}}, \dots, 0_{s_K \times s_K}). \quad (13)$$

Now we present the formal version of our gap-dependent recovery result in Theorem 1. While inferior to our main Theorem 2, this result has a simpler proof and is in fact used as a subroutine in the proof of Theorem 2.

Theorem 9 *There exists an absolute constant C such that the following holds: If there exists some $k \in \{1, \dots, K\}$ such that*

$$\frac{p-q}{2} (s_k - s_{k+1}) \geq \kappa \sqrt{pn \log n} \quad \text{or equivalently} \quad \frac{(p-q)^2 (s_k - s_{k+1})^2}{pn} \geq 4\kappa^2 \log n \quad (14)$$

and $\kappa \geq C$, then if we set $\lambda = \kappa \sqrt{pn \log n} + \frac{p-q}{2} s_{k+1}$ in (1a), with probability at least $1 - O(n^{-3})$,

$$\hat{Y} = \text{diag} (J_{s_1 \times s_1}, \dots, J_{s_k \times s_k}, 0_{s_{k+1} \times s_{k+1}}, \dots, 0_{s_K \times s_K})$$

is the unique optimal solution to the recovery SDP (1).

This theorem improves upon the result in (12) because instead of requiring the size ratio $\bar{s}/\underline{s}$ between large and small clusters to be a large constant, it allows for the ratio to approach 1 so long as the signal-to-noise ratio, $\frac{(p-q)^2 \bar{s}^2}{pn}$ is proportionally larger.

Note that Theorem 9 also utilizes a value of λ which depends on the size of the largest small cluster s_{k+1} , which may not be straightforward to find, while Theorem 2 can be set without this information.

Now we provide a simple instance where no gap-dependent algorithm can recover any cluster. Suppose $p = 1, q = 1/2$, there are n nodes, and $K = \Theta(n^{1/3})$ clusters with sizes $1, 2^2, 3^2, \dots, K^2$. (Note $n = \sum_{k=1}^K k^2 = \Theta(K^3)$.) All known efficient algorithms have recovery thresholds requiring cluster sizes to be at least $\Theta(\sqrt{n}) = \Theta(K^{1.5})$. To recover any cluster, a gap dependent algorithm will require at least a constant multiplicative gap between some cluster of size at least $\Theta(\sqrt{n})$ and the next largest cluster. As K and n get large, the ratios between the sizes of all pairs of such consecutive clusters approaches 1, hence the requisite gap does not exist. On the other hand, our gap-free algorithm can recover all clusters larger than $\Theta(K^{1.5} \log K)$. This example can be easily generalized to cluster sizes growing like $(k^\alpha)_k$ for any $\alpha > 1$, as well as to other settings of p and q .

B.2. Additional Related Work

In proceeding sections we have provided references to SBM algorithms focusing on SDP relaxation and spectral clustering methods. There are many other approaches, such as the EM algorithm (Snijders and Nowicki, 1997) and local search methods (Carson and Impagliazzo, 2001; Jerrum and Sorkin, 1998), many of which may be coupled with an SDP-based or spectral algorithm with the latter serving as an initialization procedure (Lu and Zhou, 2016; Vu, 2018; Abbe, 2018).

Many works have considered the setting that all clusters are sufficiently large to be exactly recovered, leading to precise computational and information-theoretic thresholds in the case of two balanced clusters (Abbe et al., 2015; Mossel et al., 2015a). Much subsequent work has endeavored to obtain similar characterizations in more general settings, for instance in Abbe and Sandon (2015), Abbe et al. (2017), Perry and Wein (2017), Agarwal et al. (2017), and Jog and Loh (2015), where typical recovery conditions have the form $\frac{(p-q)^2}{pn} s_K \gtrsim \log n$ assuming $p, q \asymp \frac{\log n}{n}$, where s_K is the size of the smallest cluster. There is also a complementary weak or approximate recovery setting, where the objective is to achieve strictly better classification error than random guess, usually under the constant degree regime $p, q \asymp \frac{1}{n}$. Again precise phase transitions have been established for the two-balanced-clusters case (Massoulié, 2014; Lelarge et al., 2015; Mossel et al., 2015b), and generalizations to larger K or unbalanced settings, such as Bordenave et al. (2015), Coja-Oghlan et al. (2017), Caltagirone et al. (2017), and Montanari and Sen (2016), develop recovery conditions of the form $\frac{(p-q)^2}{pn} s_K \gtrsim 1$. While the focus of the current paper is mostly orthogonal to the above work, our result, when specialized to the setting where all clusters are large, recovers the condition $\frac{(p-q)^2}{pn} s_K \gtrsim \log n$ typical in existing work.

The semi-random SBM has been considered in many works without small clusters (Moitra et al., 2016; Krivelevich and Vilenchik, 2006; Feige and Kilian, 2001). The observation that any recovery result for semi-random SBM implies a result for SBM with heterogeneous edge probabilities is a standard reduction (Chen et al., 2014). For more on the leave-one-out approach applied to clustering and related matrix estimation problems, see Zhong and Boumal (2018); Abbe et al. (2020); Ding and Chen (2020), as well as the survey in Chen et al. (2021). The work Green Larsen et al. (2020) studies faulty oracle clustering in the setting of $K = 2$ clusters.

B.3. Discussion and Comparison of Recursive Clustering Results

To analyze the performance of recursive clustering procedures, we follow (Ailon et al., 2015, Theorems 6 and 7) and consider the following setting: p, q are considered fixed and independent of n , and $C(p, q)$ is used to denote constants which depend on p, q but not n . (Generalization to

other settings of p, q is straightforward.) We examine what assumptions on the number of clusters K suffice to ensure that the recursive clustering procedure recovers all but a vanishing fraction of nodes.

Ailon et al. (2015, Theorem 7) roughly states there exists some $C(p, q)$ such that if $K \leq \beta C(p, q) \log n$ for some $\beta \leq 1$, then a recursive clustering procedure will recover clusters containing all but $O(n^\beta) + O(1)$ nodes. Since, as discussed in Subsection 2.1, their core recovery algorithm requires a certain-sized gap, they require the bound on K to grow at most logarithmically to ensure that a sufficient gap remains after each round.

If we instead apply our gap-free clustering guarantee from Theorem 2, we obtain a significant improvement, allowing K to be polynomially large. In particular, when specialized to the setting above, Theorem 2 guarantees that there exists some $C'(p, q)$ such that (with probability $1 - O(m^{-3})$) all clusters k satisfying $s_k \geq C'(p, q)\sqrt{n \log m}$ will be recovered. Using this improved guarantee, we follow the aforementioned recursive clustering strategy to define Algorithm 1. For a square matrix X , we use X_I to denote the principal submatrix of X indexed by I . When the SDP solution $\hat{Y}$ has the form stated in Theorem 2, each cluster V_k corresponding an all-positive block $\hat{Y}_{V_k}$ is said to be recovered.

Algorithm 1 Recursive clustering using Theorem 2

Input: $n \times n$ adjacency matrix A sampled from unbalanced SBM

- 1: Set $n_1 = n$, $I_1 = \{1, \dots, n\}$, round counter $\ell = 1$
 - 2: **repeat**
 - 3: Recover clusters by applying the SDP (1) to A_{I_ℓ} , where λ is set per Theorem 2 with $m = n$
 - 4: Remove all recovered clusters of size $\geq C'(p, q)\sqrt{n_\ell \log n}$ from I_ℓ to obtain $I_{\ell+1}$
 - 5: Set $n_{\ell+1} = |I_{\ell+1}|$; increase ℓ by 1
 - 6: **until** no clusters are removed
-

We can prove the following guarantees:

Theorem 10 *Under the above setting, suppose the number of clusters K satisfies $K \leq \frac{n^{\frac{1}{2}-\alpha}}{C'(p,q)\sqrt{\log n}}$ for some $0 < \alpha < \frac{1}{2}$. Then with probability at least $1 - O(n^{-2})$:*

1. *The algorithm terminates after at most K rounds.*
2. *Whenever the algorithm terminates, there is at most $n^{1-2\alpha}$ unrecovered nodes remaining.*
3. *After ℓ rounds, there is at most $n^{1-\alpha} \sum_{i=0}^{\ell-1} 2^{-i}$ unrecovered nodes remaining.*

Note that we interpret an empty summation as zero, that is, $\sum_{\ell=0}^{-1} 2^{-\ell} = 0$. The algorithm will run for at most K rounds, but it could also be terminated early to save on computation, as part 3 of Theorem 10 guarantees that a large number of nodes will be recovered after only a few rounds. The proof of Theorem 10 is given in Section D.

B.4. Discussion and Comparison of Faulty Oracle Clustering Results

First we make some comments regarding the algorithms in Table 1. The listed algorithm of Peng and Zhang (2021) assumes that all clusters are “nearly balanced”: they have size at least $\frac{n}{bK}$ for some known constant $b > 1$; the algorithm recovers all clusters under this assumption. The algorithm of Pia et al. (2022) is robust to a semirandom version of the faulty oracle clustering model (and we refer

to their paper for its definition). Additionally, we note that the listed algorithm of Mukherjee et al. (2022) actually obtains a sample complexity of $O\left(\frac{n}{s} \frac{n \log^2 n}{\delta^2} + \frac{n^4 \log^2 n}{s^4 \delta^4}\right)$, which is slightly worse by a log factor in one term compared to our Theorem 6.

Now we formally define the Algorithm 2 behind our Theorem 6 and provide further discussion of Theorem 6.

Algorithm 2 Clustering with a faulty oracle using Theorem 2

Input: Faulty clustering oracle $\mathcal{O}$, parameter $\gamma \in (0, 1)$

- 1: Choose set $T \subseteq [n]$ by including vertex i in T with probability γ (independently)
 - 2: Query (u, v) for all pairs $u, v \in T$ to form adjacency matrix $A \in \{0, 1\}^{|T| \times |T|}$
 - 3: Recover clusters from A using the recovery SDP (1), where λ is set per Theorem 2 with $m = n$
 - 4: **for** each recovered cluster $S \subseteq T$ such that $|S| \geq \frac{C_1 \log n}{\delta^2}$ **do**
 - 5: Choose $S' \subseteq S$ with $|S'| = \frac{C_1 \log n}{\delta^2}$
 ▷ Recover remainder of cluster in $[n] \setminus T$ by majority voting using S' as voters:
 - 6: **for** each vertex $i \in [n] \setminus T$ **do**
 - 7: Query (i, v) for all $v \in S'$ and add i to S if $\frac{1}{|S'|} \sum_{v \in S'} \mathcal{O}(i, v) \geq \frac{1}{2}$
 - 8: **end for**
 - 9: **end for**
-

We prove Theorem 6 in Section E. We also note that by setting $s = \frac{n}{K}$ in this result, we recover identical query complexity and cluster recovery threshold to the algorithm of Peng and Zhang (2021) for the nearly-balanced setting listed in Table 1, except our algorithm has no restrictions on cluster sizes as opposed to requiring them to be nearly balanced. Peng and Zhang (2021) conjecture that this sample complexity is optimal among efficient algorithms (up to log factors) for recovery of clusters of size $\Omega\left(\frac{n}{K}\right)$, based on statistical physics arguments.

Our Algorithm 2 used in Theorem 6 follows a common template for algorithms in the faulty oracle clustering problem:

1. Choose a subset T of the nodes and query all pairs between them.
2. Apply a clustering algorithm on the fully observed subgraph consisting of the nodes T and the edges between them to recover subclusters.
3. Take a subset S' from each recovered subcluster of minimal size such that it can be used to test all remaining nodes $i \in [n] \setminus T$ for membership in the cluster, by querying (i, v) for all $v \in S'$ and using majority voting.

These essential steps are used in all algorithms in Table 1, with one key difference being the clustering algorithm used in step 2 (for which we use our recovery SDP and Theorem 2). Mukherjee et al. (2022) utilize their (gap-free) spectral clustering procedure and thus obtain a nearly identical result. This template also explains why all the sample complexities (for efficient algorithms) consist of two terms, with the first term roughly corresponding to the majority voting procedure and the second term roughly corresponding to the subsample of size $O(|T|^2)$, with possible additional steps for some algorithms. Notably, Xia and Huang (2022) improve upon the majority voting procedure by making a connection to best arm identification, obtaining a first term of $\frac{n(K + \log n)}{\delta^2}$ which they show is optimal.

Algorithm 2 requires a user-specified target cluster size s , which in turn determines the input parameter γ , as does (Mukherjee et al., 2022, Algorithm 5). By growing the subsample T geometri-

cally over several rounds, stopping when we first recover a (sub)cluster, we can obtain our Theorem 7 which obtains an instance-dependent query complexity.

Algorithm 3 Instance-adaptive clustering with a faulty oracle

Input: Faulty clustering oracle $\mathcal{O}$

- 1: Set round counter $r = 1$
 - 2: **repeat**
 - 3: Set $\hat{s}_r = \frac{n}{2^{r-1}}$
 - 4: Set $\gamma_r = C_2 \frac{n \log n}{\hat{s}_r^2 \delta^2}$
 - 5: Choose set $T_r \subseteq [n]$ by including vertex i in T_r with probability γ_r (independently)
 - 6: Query (u, v) for all pairs $u, v \in T_r$ to form adjacency matrix $A_r \in \{0, 1\}^{|T_r| \times |T_r|}$
 - 7: Attempt to recover subclusters from A_r using the recovery SDP (1) (with $m = n$)
 - 8: **until** at least one subcluster is recovered
 - 9: Let S be an arbitrary recovered subcluster of size at least $\frac{C_1 \log n}{\delta^2}$
 - 10: Choose $S' \subseteq S$ with $|S'| = \frac{C_1 \log n}{\delta^2}$
 ▷ Recover remainder of cluster in $[n] \setminus T_r$ by majority voting using S' as voters:
 - 11: **for** each vertex $i \in [n] \setminus T_r$ **do**
 - 12: Query (i, v) for all $v \in S'$ and add i to S if the majority of queries are 1
 - 13: **end for**
-

Finally we discuss our algorithm for the setting with a small number of clusters K which was the subject of our Theorem 8.

Algorithm 4 Clustering with a faulty oracle for small K

Input: Faulty clustering oracle $\mathcal{O}$

- 1: Set round counter $r = 1$
 - 2: Set $n_1 = n$, $K_1 = K$
 - 3: **while** $n_r \geq C_2 \frac{K_r^2 \log n}{\delta^2}$ **do**
 - 4: Set $\gamma_r = C_2 \frac{K_r^2 \log n}{n_r \delta^2}$
 - 5: Choose set $T_r \subseteq [n]$ by including vertex i in T_r with probability γ_r (independently)
 - 6: Query (u, v) for all pairs $u, v \in T_r$ to form adjacency matrix $A_r \in \{0, 1\}^{|T_r| \times |T_r|}$
 - 7: Attempt to recover subclusters from A_r using the recovery SDP (1) (with $m = n$)
 - 8: **for** each recovered cluster $S \subseteq T$ such that $|S| \geq \frac{C_1 \log n}{\delta^2}$ **do**
 - 9: Choose $S' \subseteq S$ with $|S'| = \frac{C_1 \log n}{\delta^2}$
 - 10: Recover remainder of cluster in $[n] \setminus T_r$ by majority voting using S' as voters:
 - 11: **for** each vertex $i \in [n] \setminus T_r$ **do**
 - 12: Query (i, v) for all $v \in S'$ and add i to S if the majority of queries are 1
 - 13: **end for**
 - 14: **end for**
 - 15: Increase r by 1
 - 16: Decrease K_r by the number of clusters recovered in the previous round
 - 17: Remove all nodes from clusters recovered in the previous round
 - 18: Set n_r to the remaining number of nodes
 - 19: **end while**
-

The Algorithm 4 is nearly identical to Algorithm 2 except we use a different subsampling ratio γ and repeat in at most K rounds (choosing a new subsample T each round). This use of K rounds is the reason that the second term in the query complexity is $\frac{K^5 \log^2 n}{\delta^4}$ rather than the $\frac{K^4 \log^2 n}{\delta^4}$, which Peng and Zhang (2021) conjecture is optimal for the nearly balanced setting. In the setting that all clusters are $\Omega(\frac{n}{K})$, only one round of choosing $T \subseteq [n]$ and querying all pairs $u, v \in T$ would be needed to recover a subcluster from each of the K clusters, while in general settings more rounds have the potential benefit of recovering smaller clusters after larger clusters are removed. (In fact, our Algorithm 4 would match this conjectured-optimal sample complexity in the nearly-balanced setting as it would terminate after a constant number of rounds.) In summary, our algorithm recovers smaller clusters and improves upon the second term in the sample complexity compared to all previous algorithms which work for the case of general size clusters. We also believe that the first term could be improved to $\frac{n(K+\log n)}{\delta^2}$ using ideas from Xia and Huang (2022), but our focus was on the improvement available by using a gap-free clustering subroutine.

B.5. Discussion of Eigenvalue Perturbation Bounds

As discussed in Subsection 3.4, our analysis of the SDP solutions in the presence of mid-size clusters requires sharp eigenvalue perturbation bounds which improve upon existing results. We expand on our discussion in Subsection 3.4 and provide some brief background on related eigenvalue perturbation results. We refer to Eldridge et al. (2018) for more details. Again, let $M, H \in \mathbb{R}^{n \times n}$ be symmetric matrices, viewing H as a perturbation, and let v_1 be a top eigenvector of M . If we would like to bound $|\lambda_1(M + H) - \lambda_1(M)|$, the classical Weyl's inequality gives that

$$|\lambda_1(M + H) - \lambda_1(M)| \leq \|H\|_{\text{op}}.$$

As argued in Eldridge et al. (2018), this bound can be overly pessimistic, especially when H is a random perturbation whose projection onto v_1 is small relative to the eigengap $\lambda_1(M) - \lambda_2(M)$. In such a situation, intuitively the large eigengap should ensure that the new top eigenvector of $M + H$ is close to v_1 , in which case the top eigenvalue of $M + H$ will be approximately $v_1^\top (M + H) v_1 = \lambda_1(M) + v_1^\top H v_1$, and then since H is random, $|v_1^\top H v_1|$ should be much smaller than $\|H\|_{\text{op}}$. The idea that $|\lambda_1(M + H) - \lambda_1(M)| \approx |v_1^\top H v_1|$ is stated as an informal guiding principle in Eldridge et al. (2018), and they confirm the above intuition with both experiments and theoretical results. Let $\text{Col } Q$ denotes the span of the columns of a matrix Q . They prove the following upper bound:

Theorem 11 (Eldridge et al. (2018), Theorem 6) *Let $T \in [n]$ and h be such that $|x^\top H x| \leq h$ for all $x \in \text{Col}(U_{1:T})$ with $\|x\|_2 = 1$, where $U_{1:T}$ are the top T eigenvectors of M . Then for $t \leq T$, if $\lambda_t - \lambda_{T+1} > 2\|H\|_{\text{op}} - h$,*

$$\lambda_t(M + H) \leq \lambda_t(M) + h + \frac{\|H\|_{\text{op}}^2}{\lambda_t(M) - \lambda_{T+1}(M) + h - \|H\|_{\text{op}}}.$$

We modify their argument to prove the following improvement:

Theorem 12 *Let $T \in [n]$ and h be such that $|x^\top H x| \leq h$ for all $x \in \text{Col}(U_{t:T})$ with $\|x\|_2 = 1$, where $U_{t:T}$ are eigenvectors t through T of M . Then for $t \leq T$, if $\lambda_t - \lambda_{T+1} > \|H\|_{\text{op}} - h -$*

$$2 \|HU_{t:T}\|_{\text{op}},$$

$$\lambda_t(M + H) \leq \lambda_t(M) + h + \frac{\|HU_{t:T}\|_{\text{op}}^2}{\lambda_t(M) - \lambda_{T+1}(M) - \|H\|_{\text{op}} + h + 2 \|HU_{t:T}\|_{\text{op}}}.$$

We prove Theorem 12 in Section G. Notice that compared to Theorem 11, Theorem 12 has slightly weakened assumptions, and always provides a better bound since $\|HU_{t:T}\|_{\text{op}} \leq \|H\|_{\text{op}}$. Especially when H is random and T is small, $\text{Col}(U_{t:T})$ is a small subspace and so we might hope that $\|HU_{t:T}\|_{\text{op}} \ll \|H\|_{\text{op}}$.

Now we compare these results within the context of the proof of Theorem 2, and show that the improvement from our Theorem 12 is indeed substantial. In this subsection we also write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Fix a cluster k so that $\frac{p-q}{2}s_k \asymp \kappa\sqrt{pn \log n}$. Let $w^\top$ be the first row of $W^{(k)}$ except with its first entry halved. This is chosen so that by defining $H = e_1 w^\top + w e_1^\top$, H contains the noise from the first row and column of $W^{(k)}$. Now let $M = \overleftrightarrow{A}^{(k)} - H$ (so $M + H = \overleftrightarrow{A}^{(k)}$). Our objective will be to bound $\left| \lambda_1(\overleftrightarrow{A}^{(k)}) - \lambda_1(\overleftrightarrow{A}^{(k)} - H) \right|$. By the assumption on the size of the cluster k , there is sufficient signal so that the top eigenvector v_1 of M satisfies $\|v_1\|_\infty \lesssim \frac{1}{\sqrt{s_k}}$ (by applying an ∞ -norm eigenvector perturbation bound (Chen et al., 2021)). Then

$$\left| v_1^\top H v_1 \right| = 2 |v_1^\top e_1| |w^\top v_1| \lesssim \frac{1}{\sqrt{s_k}} \sqrt{p \log n}$$

where the bound $|w^\top v_1| \lesssim \sqrt{p \log n}$ holds with high probability using the fact that w is independent from v_1 (since v_1 is the top eigenvector of a matrix where the first row and column of noise have been set to 0). The informal guiding principle from Eldridge et al. (2018) would then give us hope that $\left| \lambda_1(\overleftrightarrow{A}^{(k)}) - \lambda_1(\overleftrightarrow{A}^{(k)} - H) \right| \lesssim \frac{\sqrt{p \log n}}{\sqrt{s_k}}$, which is the order needed within our proof of Theorem 2. Unfortunately, Theorem 11 only yields that

$$\lambda_1(\overleftrightarrow{A}^{(k)}) - \lambda_1(\overleftrightarrow{A}^{(k)} - H) \lesssim \frac{\sqrt{p \log n}}{\sqrt{s_k}} + \frac{\|H\|_{\text{op}}^2}{\kappa\sqrt{pn \log n} - \|H\|_{\text{op}}} \lesssim \frac{(\sqrt{ps_k})^2}{\kappa\sqrt{pn \log n}} \lesssim \frac{\sqrt{ps_k}}{\kappa\sqrt{n \log n}}$$

since $\|H\|_{\text{op}} \asymp \|w\|_2 \asymp \sqrt{ps_k}$. However, since

$$\|H v_1\|_{\text{op}} = \left\| w e_1^\top v_1 + e_1 w^\top v_1 \right\|_2 \leq \|w\|_2 \left| e_1^\top v_1 \right| + \left| w^\top v_1 \right| \lesssim \sqrt{ps_k} \frac{1}{\sqrt{s_k}} + \sqrt{p \log n} \lesssim \sqrt{p \log n}$$

our Theorem 12 gives that

$$\lambda_1(\overleftrightarrow{A}^{(k)}) - \lambda_1(\overleftrightarrow{A}^{(k)} - H) \lesssim \frac{\sqrt{p \log n}}{\sqrt{s_k}} + \frac{(\sqrt{p \log n})^2}{\kappa\sqrt{pn \log n}} \lesssim \frac{\sqrt{p \log n}}{\sqrt{s_k}} + \frac{\sqrt{p \log n}}{\kappa\sqrt{n}} \lesssim \frac{\sqrt{p \log n}}{\sqrt{s_k}}.$$

Appendix C. Proof of Main Theorem

Our main Theorem 2 can be decomposed into two parts: first, that the oracle SDPs (4) have solutions which are either zero or rank-one and all-positive, and second, that the recovery SDP solution (1) has a block-diagonal form where the blocks are the solutions to the oracle SDPs. Thus, we prove the first part in Subsection C.1 and the second part in Subsection C.2. The standalone result on clustering

with a gap, Theorem 9, is proved in Subsection C.3. So as not to interrupt the flow of the proofs, we place most concentration inequalities or their proofs within Subsection C.4.

We define the shifted adjacency matrix $\overleftrightarrow{A} := A - \frac{p+q}{2}J$. We also define the noise matrix $W := A - \mathbb{E}A$.

The following proofs make use of dual variables for the optimization problems (1) and (4), so we stop to introduce the KKT conditions for these problems.

Lemma 13 $\hat{Y}$ is an optimal solution of the recovery SDP (1) if and only if there exist $U, L \in \mathbb{R}^{n \times n}$ such that

$$0 \leq \hat{Y} \leq 1, \hat{Y} \succcurlyeq 0 \quad (15a)$$

$$\overleftrightarrow{A} - \lambda I - U + L \preccurlyeq 0, L \geq 0, U \geq 0 \quad (15b)$$

$$L_{ij}\hat{Y}_{ij} = 0, U_{ij}(\hat{Y}_{ij} - 1) = 0 \quad \forall i, j \in [n] \quad (15c)$$

$$\left(\overleftrightarrow{A} - \lambda I - U + L\right) \hat{Y} = 0_{n \times n}. \quad (15d)$$

Likewise, for any $k \in \{1, \dots, K\}$, $\hat{Y}^k$ is an optimal solution of the k th oracle SDP (4) if and only if there exist $U^k, L^k \in \mathbb{R}^{s_k \times s_k}$ such that

$$0 \leq \hat{Y}^k \leq 1, \hat{Y}^k \succcurlyeq 0 \quad (16a)$$

$$\overleftrightarrow{A}^{(k)} - \lambda I - U^k + L^k \preccurlyeq 0, L^k \geq 0, U^k \geq 0 \quad (16b)$$

$$L_{ij}^k \hat{Y}_{ij}^k = 0, U_{ij}^k (\hat{Y}_{ij}^k - 1) = 0 \quad \forall i, j \in [s_k] \quad (16c)$$

$$\left(\overleftrightarrow{A}^{(k)} - \lambda I - U^k + L^k\right) \hat{Y}^k = 0_{s_k \times s_k} \quad (16d)$$

Proof First we consider the recovery SDP (1). First, we note that the feasible set is compact since $\{Y \in \mathbb{R}^{n \times n} : 0 \leq Y \leq 1\}$ is compact, $\{Y : Y \succcurlyeq 0\}$ is closed, and the feasible set is the intersection of these two sets. Therefore there exists an optimal solution $\hat{Y}$. By the generalized Slater's condition within (Boyd and Vandenberghe, 2004, Section 5.9.1), since the SDP (1) is convex and the feasible point $\frac{1}{2}J_{n \times n}$ satisfies $0 < \frac{1}{2}J_{n \times n} < 1$ and $\frac{1}{2}J_{n \times n} \succcurlyeq 0$, strong duality holds and thus from (Boyd and Vandenberghe, 2004, Section 5.9.2) the following KKT conditions are necessary and sufficient for $\hat{Y}$ to be an optimal solution:

$$\begin{aligned} 0 &\leq \hat{Y} \leq 1, \hat{Y} \succcurlyeq 0 \\ \overleftrightarrow{A} - \lambda I - U + L &\preccurlyeq 0, L \geq 0, U \geq 0, P \succcurlyeq 0 \\ L_{ij}\hat{Y}_{ij} &= 0, U_{ij}(\hat{Y}_{ij} - 1) = 0 \quad \forall i, j \in [n] \\ P\hat{Y} &= 0_{n \times n} \\ -\overleftrightarrow{A} + \lambda I + U - L + P &= 0 \end{aligned} \quad (17)$$

where $U, L, P \in \mathbb{R}^{n \times n}$. Now by eliminating P using (17), we obtain the desired recovery SDP KKT conditions (15). An identical argument suffices for the oracle SDP KKT conditions (16). $\blacksquare$

Lastly, we set the regularization parameter λ which appears in both SDPs (1) and (4) as $\lambda = \kappa\sqrt{pn \log m} + \delta$ for κ which is a constant which we will choose later and $\delta \sim \text{Uniform}[0, 0.1]$.

Conceptually, we can make κ arbitrarily large, and throughout the proofs we imagine any constant multiple of $1/\kappa$ as arbitrarily small. We do not use any particular properties of δ except that almost surely we will have $\kappa\sqrt{pn \log m} \leq \lambda \leq (\kappa + 1)\sqrt{pn \log m}$ and $\lambda_1(\vec{A}^{(k)}) \neq \lambda$ (for all $k \in [K]$).

C.1. Oracle SDP Solutions

This subsection is devoted to the proof of the following lemma:

Lemma 14 *There exist constants $B, \underline{B}, \bar{B} > 2$ such that if $\kappa \geq B$, with probability at least $1 - O(m^{-3})$, for each cluster k , the k th oracle SDP (4) has a unique solution $\hat{Y}^k$, and there exist dual variables U^k, L^k which satisfy the oracle KKT equations (16) such that $L^k = 0$ and $U^k \succcurlyeq 0$. Furthermore,*

1. *if $\frac{p-q}{2}s_k \geq \left(1 + \frac{1}{\underline{B}}\right)\kappa\sqrt{pn \log m}$, then $\hat{Y}^k = J_{s_k \times s_k} = \mathbf{1}\mathbf{1}^\top$,*
2. *if $\frac{p-q}{2}s_k \leq \left(1 - \frac{1}{\bar{B}}\right)\kappa\sqrt{pn \log m}$, then $\hat{Y}^k = 0_{s_k \times s_k} = 00^\top$,*
3. *if $\left(1 - \frac{1}{\bar{B}}\right)\kappa\sqrt{pn \log m} < \frac{p-q}{2}s_k < \left(1 + \frac{1}{\underline{B}}\right)\kappa\sqrt{pn \log m}$, then either $\hat{Y}^k = 0$ or $\hat{Y}^k = y^k y^{k\top}$ where $y^k \geq \frac{1}{2}$ entrywise.*

First we establish an event upon which we will condition the rest of the proof, which will be the intersection of the events described in the following Lemmas 15, 16, 17, and 18.

Lemma 15 *With probability at least $1 - O(m^{-3})$, there exists an absolute constant B_1 such that:*

1. *For all clusters k such that $\frac{p-q}{2}s_k \geq \sqrt{pn \log m}$,*

$$\left\| W^{(k)} \right\|_{\text{op}} \leq B_1 \sqrt{ps_k}. \quad (18)$$

2. *For all clusters k such that $\frac{p-q}{2}s_k \geq \sqrt{pn \log m}$, for all columns $(W^{(k)}_{:,j})_{j=1}^{s_k}$,*

$$\left\| W^{(k)}_{:,j} \right\|_2 \leq B_1 \sqrt{ps_k}. \quad (19)$$

3. *For all clusters k ,*

$$\left\| W^{(k)} \right\|_{\text{op}} \leq B_1 \sqrt{pn}. \quad (20)$$

Lemma 15 is proven in Subsection C.4.

Recall that for any $n \times n$ matrix M and any $k \in [K]$, we define $M^{(k)}$ to be the $s_k \times s_k$ principal submatrix of M corresponding to the rows and columns associated with cluster V_k . We later make use of leave-one-out (LOO) versions of the noise for analysis so we define them now: For each $k \in [K]$ and each $j \in [s_k]$, let $w^{k,j}$ be equal to $W^{(k)}_{:,j}$ (the j th column of $W^{(k)}$) but with the j th entry divided by two (so that we do not subtract $W^{(k)}_{jj}$ twice). Now define the j th LOO noise matrix $W^{k,j} := W^{(k)} - e_j w^{k,j\top} - w^{k,j} e_j^\top$ and the j th LOO adjacency matrix

$$A^{k,j} := A^{(k)} - e_j w^{k,j\top} - w^{k,j} e_j^\top = \mathbb{E}A^{(k)} + W^{k,j}.$$

We also define the j th LOO shifted adjacency matrix $\overleftrightarrow{A}^{k,j} := A^{k,j} - \frac{p+q}{2} J_{s_k \times s_k}$.

Let v^k be the top eigenvector of $\overleftrightarrow{A}^{(k)}$ and let $v^{k,j}$ be the top eigenvector of $\overleftrightarrow{A}^{k,j}$ (by top, we mean corresponding to the largest eigenvalue). Of course these eigenvectors are only defined up to a global rotation, but the following lemma resolves this ambiguity and also establishes a key property of v^k and $v^{k,j}$.

Lemma 16 *There exist absolute constants B_2, B_3 such that with probability at least $1 - O(m^{-3})$, for all clusters k such that $\frac{p-q}{2} s_k \geq B_3 \sqrt{pn \log m}$, there exists a top eigenvector v^k of $\overleftrightarrow{A}^{(k)}$ such that*

$$\left\| v^k - \frac{1}{\sqrt{s_k}} \mathbf{1} \right\|_{\infty} \leq 2B_2 \frac{\sqrt{p \log m}}{(p-q)s_k}, \quad (21)$$

and for each $j \in [s_k]$, there exists a top eigenvector $v^{k,j}$ of $\overleftrightarrow{A}^{k,j}$ such that

$$\left\| v^{k,j} - \frac{1}{\sqrt{s_k}} \mathbf{1} \right\|_{\infty} \leq 2B_2 \frac{\sqrt{p \log m}}{(p-q)s_k}. \quad (22)$$

The proof of Lemma 16 is in Subsection C.4, as are the proofs of the following Lemmas 17 and 18.

Lemma 17 *There exist absolute constants B_4, B_5 such that with probability at least $1 - O(m^{-3})$, for every cluster k such that $\frac{p-q}{2} s_k \geq B_4 \sqrt{pn \log m}$, for each $j \in [s_k]$,*

$$\left\| v^{k,j} \right\|_{\infty} \leq \frac{2}{\sqrt{s_k}} \quad (23)$$

and

$$\left| \langle w^{k,j}, v^{k,j} \rangle \right| \leq B_5 \sqrt{p \log m}. \quad (24)$$

Lemma 18 *With probability at least $1 - O(m^{-3})$, for all clusters k such that $\frac{p-q}{2} s_k \geq B_4 \sqrt{pn \log m}$,*

$$\left\| W^{(k)} \frac{1}{\sqrt{s_k}} \mathbf{1}_{s_k} \right\|_{\infty} \leq B_5 \sqrt{p \log m}.$$

Define E_1 to be the intersection of the events in Lemmas 15, 16, 17, and 18. By the union bound $\mathbb{P}(E_1) \geq 1 - O(m^{-3})$. For most of the rest of the proof we condition on the event E_1 or supersets of E_1 . Also we assume that $\kappa > B_1$, which implies $\lambda > \|W^{(k)}\|_{\text{op}}$ for each $k \in [K]$.

We split the proof into three cases related to the size of s_k corresponding to those in Lemma 14. Before considering each case, we remind the reader that we avoid the situation that $\lambda_1(\overleftrightarrow{A}^{(k)}) = \lambda$ with probability one by our addition of the small random continuous perturbation to λ .

C.1.1. ALL-ZERO SOLUTION

First we handle the easiest case where $\lambda_1 \left(\overleftrightarrow{A}^{(k)} \right) < \lambda$, which will cause the oracle solution to be zero.

Lemma 19 *On the event $E_1 \cap \left\{ \lambda_1 \left(\overleftrightarrow{A}^{(k)} \right) < \lambda \right\}$, we have that $\hat{Y}^k = 0_{s_k \times s_k}$ is the unique solution of the oracle SDP (4), and furthermore by setting $L^k = 0_{s_k \times s_k}$ and $U^k = 0_{s_k \times s_k}$, then $\hat{Y}^k, L^k, U^k$ satisfy the oracle SDP KKT conditions (16). Also, for any cluster k such that $\frac{p-q}{2} s_k \leq \left(1 - \frac{B_1}{\kappa}\right) \kappa \sqrt{pn \log m}$, $E_1 = E_1 \cap \left\{ \lambda_1 \left(\overleftrightarrow{A}^{(k)} \right) < \lambda \right\}$.*

Proof First we check $\hat{Y}^k = 0$ is the unique solution of the oracle SDP (4) when $\lambda_1 \left(\overleftrightarrow{A}^{(k)} \right) < \lambda$. The oracle SDP objective (4) is

$$\begin{aligned} \left\langle Y, A^{(k)} - \frac{p+q}{2} J_{s_k \times s_k} \right\rangle - \lambda \text{Tr}(Y) &= \left\langle Y, \overleftrightarrow{A}^{(k)} - \lambda I \right\rangle \\ &\leq \sum_{i=1}^{s_k} \lambda_i(Y) \lambda_i \left(\overleftrightarrow{A}^{(k)} - \lambda I \right) \\ &\leq \lambda_1(Y) \lambda_1 \left(\overleftrightarrow{A}^{(k)} - \lambda I \right) \end{aligned} \quad (25)$$

where we obtain the first inequality using a Hermitian trace inequality (Marshall et al., 1979, Theorem 9.H.1.g and the discussion thereafter). We obtain the second inequality from the fact that all terms are non-positive since $\lambda_i \left(\overleftrightarrow{A}^{(k)} - \lambda I \right) \leq \lambda_1 \left(\overleftrightarrow{A}^{(k)} \right) - \lambda < 0$ and $\lambda_i(Y) \geq 0$ since $Y \succeq 0$ by feasibility to the oracle SDP (4). Now expression (25) is strictly less than zero unless we have $\lambda_1(Y) = 0$, and again since $Y \succeq 0$ this implies that $Y = 0$. Therefore $\hat{Y}^k = 0$ is the unique solution of the oracle SDP.

Next we check that $L^k = 0$ and $U^k = 0$ satisfy the oracle SDP KKT conditions (16). We note that

$$\overleftrightarrow{A}^{(k)} - \lambda I - U^k + L^k = \overleftrightarrow{A}^{(k)} - \lambda I \preceq 0$$

where the PSD inequality follows from the assumption that we are in the event $\left\{ \lambda_1 \left(\overleftrightarrow{A}^{(k)} \right) < \lambda \right\}$, satisfying the first part of (16b). All other conditions are trivial to check.

Finally, note that the event E_1 contains the conclusions of Lemma 15, so by (20) we have for any cluster k that

$$\lambda_1 \left(W^{(k)} \right) \leq \left\| W^{(k)} \right\|_{\text{op}} \leq B_1 \sqrt{pn} \leq \frac{B_1}{\kappa} \kappa \sqrt{pn \log m}.$$

Therefore if $\frac{p-q}{2} s_k \leq \left(1 - \frac{B_1}{\kappa}\right) \kappa \sqrt{pn \log m}$, then by Weyl's inequality

$$\lambda_1 \left(\overleftrightarrow{A}^{(k)} \right) \leq \lambda_1 \left(\frac{p-q}{2} J_{s_k \times s_k} \right) + \lambda_1 \left(W^{(k)} \right) \leq \frac{p-q}{2} s_k + \frac{B_1}{\kappa} \kappa \sqrt{pn \log m} < \lambda.$$

■

C.1.2. ALL-ONE SOLUTION

We also provide the following lemma summarizing the case that $\frac{p-q}{2}s_k$ is large-enough for the oracle solution to be all-ones.

Lemma 20 *On the event E_1 , for all clusters k such that $\frac{p-q}{2}s_k \geq (1 + \frac{B_6}{\kappa}) \kappa \sqrt{pn \log m}$, we have that $\hat{Y}^k = J_{s_k \times s_k}$ is the unique solution of the oracle SDP (4), and furthermore there exist some U^k, L^k such that $\hat{Y}^k, L^k, U^k$ satisfy the oracle SDP KKT conditions (16), $L^k = 0$, and $U^k \succcurlyeq 0$.*

Since this essentially follows as a special case of Theorem 9 (but with a larger λ), we prove Lemma 20 in Subsection C.3.

C.1.3. NONZERO CRITICAL SIZE SOLUTION

Finally we consider the most challenging case, where

$$\left(1 - \frac{B_1}{\kappa}\right) \kappa \sqrt{pn \log m} < \frac{p-q}{2}s_k < \left(1 + \frac{B_6}{\kappa}\right) \kappa \sqrt{pn \log m} \quad (26)$$

$$\text{and } \lambda_1(\overleftrightarrow{A}^{(k)}) > \lambda. \quad (27)$$

To analyze the oracle SDP in this case, we consider a new *relaxed oracle SDP* where all lower-bound constraints are removed, which enables us to prove that its unique solution has rank 1. Then we will analyze the relaxed oracle SDP and show that its solution is non-negative, which then implies that its solution is also the solution of the corresponding oracle SDP, and therefore the original oracle SDP has a unique solution with rank ≤ 1 . First we define the k th relaxed oracle SDP:

$$\bar{Y}^k = \underset{Y \in \mathbb{R}^{s_k \times s_k}}{\operatorname{argmax}} \quad \left\langle Y, A^{(k)} - \frac{p+q}{2} J_{s_k \times s_k} \right\rangle - \lambda \operatorname{Tr}(Y) \quad (28a)$$

$$\text{s.t. } Y \succcurlyeq 0, \quad (28b)$$

$$Y_{ii} \leq 1, \forall i \in [s_k]. \quad (28c)$$

Beyond removing the lower bound constraints, the relaxed oracle SDP also has fewer upper bound constraints (28c), which when combined with the PSD constraint (28b) are equivalent to upper-bounding each entry of the solution, but having the reduced constraints are useful for showing that the relaxed oracle SDP solution $\bar{Y}^k$ must have rank 1.

To show that $\bar{Y}^k$ is entrywise positive despite the removal of the lower bound constraints, we make use of s_k *leave-one-out (LOO) relaxed oracle SDPs* wherein one row and column of the noise are zeroed out. The key property of the LOO relaxed oracle SDPs is that we will be able to show that (for each k, j) their solutions have low correlation with the corresponding left-out noise vectors $w^{k,j}$, due to their independence. We would like to define a second high-probability event E_2 under which this occurs, but we cannot simply work under the event E_1 to establish that E_2 has high probability, since E_1 is not independent of the left-out noise. To remedy this we define new high-probability events.

For each cluster k such that $\frac{p-q}{2}s_k \geq B_4 \sqrt{pn \log m}$, for each $j \in [s_k]$, define the event

$$E_1^{k,j} = \left\{ \begin{array}{l} \|W^{k,j}\|_{\text{op}} \leq B_1 \sqrt{ps_k} \\ \left\| v^{k,j} - \frac{1}{\sqrt{s_k}} \mathbf{1} \right\|_{\infty} \leq 2B_2 \frac{\sqrt{p \log m}}{(p-q)s_k} \end{array} \right\}.$$

These conditions are (18) and (22) which are both included in E_1 , so we have that $E_1^{k,j} \supseteq E_1$. Also none of these conditions involve the (k, j) th left-out noise vector $w^{k,j}$, so event $E_1^{k,j}$ is independent of $w^{k,j}$.

Before defining the LOO relaxed oracle SDP, a second issue is that even when $\lambda_1(\overleftrightarrow{A}^{(k)}) > \lambda$, we may have $\lambda_1(\overleftrightarrow{A}^{k,j}) < \lambda$ after removing noise, in which case if we used the same regularization parameter λ in the LOO relaxed oracle SDP, it would have a zero solution which would not be useful. To resolve this issue, we will later show that under the event E_1 , for all clusters k such that $\frac{p-q}{2}s_k \geq B_4\sqrt{pn \log m}$ (which includes all clusters in the critical regime (26) assuming κ is sufficiently large), $\lambda_1(\overleftrightarrow{A}^{k,j})$ is very close to $\lambda_1(\overleftrightarrow{A}^{(k)})$. This bound is of course dependent on the left-out noise $w^{k,j}$ so we state it later, but for now we mention this consideration in order to motivate our choice to define the LOO relaxed oracle SDPs with reduced regularization parameters λ^k .

For each k such that $\frac{p-q}{2}s_k \geq B_4\sqrt{pn \log m}$, we define the modified regularization parameter $\lambda^k := \lambda - \frac{B_8\sqrt{p \log m}}{\sqrt{s_k}}$. Also we add the requirement that κ is large enough so that under $E_1^{k,j}$ we are ensured $\lambda^k > \|W^{k,j}\|_{\text{op}}$. Finally we can define the leave-one-out SDPs. For all k such that $\frac{p-q}{2}s_k \geq B_4\sqrt{pn \log m}$ and each $j \in [s_k]$, we define the (k, j) th LOO relaxed oracle SDP

$$\bar{Y}^{k,j} = \underset{Y \in \mathbb{R}^{s_k \times s_k}}{\text{argmax}} \quad \left\langle Y, A^{k,j} - \frac{p+q}{2} J_{s_k \times s_k} \right\rangle - \lambda^k \text{Tr}(Y) \quad (29a)$$

$$\text{s.t.} \quad Y \succcurlyeq 0, \quad (29b)$$

$$Y_{ii} \leq 1, \forall i \in [s_k]. \quad (29c)$$

Now we are prepared to show that the relaxed oracle SDPs (28) and the LOO relaxed oracle SDPs (29) have rank-one solutions.

Lemma 21 *Under the event $E_1 \cap \left\{ \lambda_1(\overleftrightarrow{A}^{(k)}) > \lambda \right\}$, the k th relaxed oracle SDP (28) has rank-one solution $\bar{Y}^k = \bar{y}^k \bar{y}^{k\top}$ where $\mathbf{1}^\top \bar{y}^k \geq 0$, and there exists $u^k \in \mathbb{R}^{s_k}$ such that:*

$$-1 \leq \bar{y}^k \leq 1 \quad (30a)$$

$$\overleftrightarrow{A}^{(k)} - \lambda I - \text{diag}(u^k) \preccurlyeq 0, u^k \geq 0 \quad (30b)$$

$$\left((\bar{y}_i^k)^2 - 1 \right) u_i^k = 0 \quad \forall i \in [s_k] \quad (30c)$$

$$\left(\overleftrightarrow{A}^{(k)} - \lambda I - \text{diag}(u^k) \right) \bar{y}^k = 0. \quad (30d)$$

Also, for each k such that $\frac{p-q}{2}s_k \geq B_4\sqrt{pn \log m}$, for each $j \in [s_k]$, under the event $E_1^{k,j} \cap \left\{ \lambda_1(\overleftrightarrow{A}^{k,j}) > \lambda^k \right\}$, the (k, j) th LOO relaxed oracle SDP (29) has rank-one solution $\bar{Y}^{k,j} =$

$\bar{y}^{k,j} \bar{y}^{k,j \top}$ where $\mathbf{1}^\top \bar{y}^{k,j} \geq 0$, and there exists $u^{k,j} \in \mathbb{R}^{s_k}$ such that

$$-1 \leq \bar{y}^{k,j} \leq 1 \quad (31a)$$

$$\overleftrightarrow{A}^{k,j} - \lambda I - \text{diag}(u^{k,j}) \preceq 0, u^{k,j} \geq 0 \quad (31b)$$

$$\left(\left(\bar{y}_i^{k,j} \right)^2 - 1 \right) u_i^{k,j} = 0 \quad \forall i \in [s_k] \quad (31c)$$

$$\left(\overleftrightarrow{A}^{k,j} - \lambda I - \text{diag}(u^{k,j}) \right) \bar{y}^{k,j} = 0. \quad (31d)$$

Proof We start by writing the KKT conditions for the relaxed oracle SDP (28). By a nearly identical argument as in Lemma 13, for $\bar{Y}^i$ to be optimal there must exist $u^k \in \mathbb{R}^{s_k}$ such that

$$\bar{Y}^k \succcurlyeq 0, \bar{Y}_{ii}^k \leq 1 \quad \forall i \in [s_k] \quad (32a)$$

$$u^k \geq 0, \overleftrightarrow{A}^{(k)} - \lambda I - \text{diag}(u^k) \preceq 0 \quad (32b)$$

$$((\bar{Y}^k)_{ii} - 1) u_i^k = 0 \quad \forall i \in [s_k] \quad (32c)$$

$$\left(\overleftrightarrow{A}^{(k)} - \lambda I - \text{diag}(u^k) \right) \bar{Y}^k = 0_{s_k \times s_k}. \quad (32d)$$

Note $\text{diag}(u^k) \succcurlyeq 0$ since $u^k \geq 0$. Recall that we are working under an event where $\lambda_1 \left(\overleftrightarrow{A}^{(k)} \right) > \lambda$. Notice

$$\overleftrightarrow{A}^{(k)} - \lambda I - \text{diag}(u^k) = \frac{p-q}{2} J_{s_k \times s_k} + W^{(k)} - \lambda I - \text{diag}(u^k),$$

and $\frac{p-q}{2} J_{s_k \times s_k}$ is rank 1. Therefore by Weyl's inequality we have for any $i \in [s_k]$ that

$$\begin{aligned} \lambda_i \left(\overleftrightarrow{A}^{(k)} - \lambda I - \text{diag}(u^k) \right) &= \lambda_i \left(\frac{p-q}{2} J_{s_k \times s_k} + W^{(k)} - \lambda I - \text{diag}(u^k) \right) \\ &\leq \lambda_i \left(\frac{p-q}{2} J_{s_k \times s_k} \right) + \lambda_1 \left(W^{(k)} - \lambda I - \text{diag}(u^k) \right) \\ &< \lambda_i \left(\frac{p-q}{2} J_{s_k \times s_k} \right) \end{aligned}$$

because $W^{(k)} \prec \lambda I$. Since $\lambda_i \left(\frac{p-q}{2} J_{s_k \times s_k} \right) = 0$ for $i \geq 2$, we have that $\lambda_i \left(\overleftrightarrow{A}^{(k)} - \lambda I - \text{diag}(u^k) \right) < 0$ for $i \geq 2$. Now applying a Hermitian version of von Neumann's trace inequality (Marshall et al., 1979, Theorem 9.H.1.g and the discussion thereafter), we have

$$\begin{aligned} \text{Tr} \left(\overleftrightarrow{A}^{(k)} - \lambda I - \text{diag}(u^k) \right) \bar{Y}^k &\leq \sum_{i=1}^{s_k} \lambda_i \left(\overleftrightarrow{A}^{(k)} - \lambda I - \text{diag}(u^k) \right) \lambda_i(\bar{Y}^k) \\ &\leq \sum_{i=2}^{s_k} \lambda_i \left(\overleftrightarrow{A}^{(k)} - \lambda I - \text{diag}(u^k) \right) \lambda_i(\bar{Y}^k) \end{aligned} \quad (33)$$

where in the second line we use the fact that $\lambda_1 \left(\overleftrightarrow{A}^{(k)} - \lambda I - \text{diag}(u^k) \right) \leq 0$ and $\lambda_1(\bar{Y}^k) \geq 0$. Now the only way for (33) to be 0 and not < 0 is if $\lambda_i(\bar{Y}^k) = 0$ for all $i \geq 2$. Therefore $\bar{Y}^k$ has rank

at most 1 in this case. But also $\bar{Y}^k$ cannot be 0, since 0 has a lower objective value than $v^k v^{k\top}$ (the top eigenvector of $\overleftrightarrow{A}^{(k)}$) since

$$\langle v^k v^{k\top}, \overleftrightarrow{A}^{(k)} \rangle - \lambda \text{Tr}(v^k v^{k\top}) = v^{k\top} \overleftrightarrow{A}^{(k)} v^k - \lambda > 0.$$

Thus $\bar{Y}^k$ must be rank 1.

Now letting $\bar{Y}^k = \bar{y}^k \bar{y}^{k\top}$, to resolve the sign ambiguity of $\bar{y}^k$ we define $\mathbf{1}^\top \bar{y}^k \geq 0$. Finally we establish the optimality conditions (30). (32a) follows from (32a) once we note that $\bar{Y}_{ii}^k = (\bar{y}_i^k)^2$. Similarly (30c) follows from (32c). (30d) follows from (32d) by right-multiplying by $\bar{y}^k$ and noting that $\|\bar{y}^k\|_2 \neq 0$ since $\bar{y}^k \neq 0$.

A completely analogous argument suffices to prove the analogous statements about the LOO relaxed oracle SDP solutions. $\blacksquare$

Next we are able to show that the LOO relaxed oracle SDP solutions $\bar{y}^{k,j}$ have low correlation with the corresponding left-out noise vectors $w^{k,j}$.

Lemma 22 *With probability at least $1 - O(m^{-3})$, for all clusters k satisfying $\frac{p-q}{2} s_k \geq B_4 \sqrt{pn \log m}$, for each $j \in [s_k]$,*

$$|\langle w^{k,j}, \bar{y}^{k,j} \rangle| \leq B_5 \sqrt{ps_k \log m}.$$

We define the event in Lemma 22 to be E_2 . We let $E_3 = E_1 \cap E_2$, and note by the union bound that $\mathbb{P}(E_3) \geq 1 - O(m^{-3})$. From here on the proof will be deterministic. First we check that the LOO relaxed oracle problems have nonzero solutions when the corresponding relaxed oracle problem has a nonzero solution.

Lemma 23 *Under the event E_1 , for all clusters k such that $\frac{p-q}{2} s_k \geq B_4 \sqrt{pn \log m}$, for each $j \in [s_k]$,*

$$|\lambda_1(\overleftrightarrow{A}^{(k)}) - \lambda_1(\overleftrightarrow{A}^{k,j})| \leq \frac{B_8 \sqrt{p \log m}}{\sqrt{s_k}},$$

and therefore by the choice of λ^k , for any k such that $\lambda_1(\overleftrightarrow{A}^{(k)}) > \lambda$, we have $\lambda_1(\overleftrightarrow{A}^{k,j}) > \lambda^k$ for all $j \in [s_k]$.

Proof Fix a cluster k such that $\frac{p-q}{2} s_k \geq B_4 \sqrt{pn \log m}$. We can show the desired bound for arbitrary j . Let $\tilde{w} = w^{k,j}$, $\tilde{v} = v^{k,j}$, and $\tilde{W} = W^{k,j}$. Note $\|\tilde{W}\|_{\text{op}} \leq \|W^{(k)}\|_{\text{op}} \leq B_1 \sqrt{ps_k}$. Let $\Delta = \tilde{w} e_j^\top + e_j \tilde{w}^\top$ and $\tilde{A} = \overleftrightarrow{A}^{k,j} = \overleftrightarrow{A}^{(k)} - \Delta$. First we bound $|\tilde{v}^\top \Delta \tilde{v}|$ and $\|\Delta \tilde{v}\|_2$.

Using Lemma 17,

$$|\tilde{v}^\top \Delta \tilde{v}| = 2|\tilde{w}^\top \tilde{v}| |\tilde{v}_j| \leq \frac{2B_5 \sqrt{p \log m}}{\sqrt{s_k}}.$$

Next, again using Lemma 17,

$$\|\Delta \tilde{v}\|_2 = \left\| \tilde{w} \tilde{v}_j + e_j \tilde{w}^\top \tilde{v} \right\|_2 \leq \|\tilde{w}\|_2 \|\tilde{v}\|_\infty + |\tilde{w}^\top \tilde{v}| \leq B_1 \sqrt{ps_k} \frac{2}{\sqrt{s_k}} + B_5 \sqrt{p \log m} \leq B_7 \sqrt{p \log m}.$$

Now we can use these bounds in the eigenvalue perturbation inequalities. By Theorem 12 we have

$$\begin{aligned}
 \lambda_1 \left(\overleftrightarrow{A}^{(k)} \right) &= \lambda_1 \left(\tilde{A} + \Delta \right) \\
 &\leq \lambda_1 \left(\tilde{A} \right) + |\tilde{v}^\top \Delta \tilde{v}| + \frac{\|\Delta \tilde{v}\|_2^2}{\lambda_1 \left(\tilde{A} \right) - \lambda_2 \left(\tilde{A} \right) - \|\Delta\|_{\text{op}}} \\
 &\leq \lambda_1 \left(\tilde{A} \right) + \frac{2B_5 \sqrt{p \log m}}{\sqrt{s_k}} + \frac{B_7^2 p \log m}{\frac{p-q}{2} s_k - 4B_1 \sqrt{ps_k}} \\
 &\leq \lambda_1 \left(\tilde{A} \right) + \frac{2B_5 \sqrt{p \log m}}{\sqrt{s_k}} + \frac{B_7^2 p \log m}{(\kappa - B_1) \sqrt{pn \log m} - 5B_1 \sqrt{ps_k}} \\
 &\leq \lambda_1 \left(\tilde{A} \right) + \frac{2B_5 \sqrt{p \log m}}{\sqrt{s_k}} + \frac{B_7^2 p \log m}{(\kappa - 6B_1) \sqrt{pn \log m}} \\
 &\leq \lambda_1 \left(\tilde{A} \right) + \frac{2B_5 \sqrt{p \log m}}{\sqrt{s_k}} + \frac{B_7^2 p \log m}{\sqrt{pn \log m}} \\
 &\leq \lambda_1 \left(\tilde{A} \right) + \frac{B_8 \sqrt{p \log m}}{\sqrt{s_k}}.
 \end{aligned}$$

In the second inequality we used our bounds for $|\tilde{v}^\top \Delta \tilde{v}|$ and $\|\Delta \tilde{v}\|_2$, as well as the facts that $\|\Delta\|_{\text{op}} \leq 2\|\tilde{w}\|_2 \leq 2B_1 \sqrt{ps_k}$, $\lambda_1 \left(\tilde{A} \right) \geq \frac{p-q}{2} s_k - \|\tilde{W}\|_{\text{op}} \geq \frac{p-q}{2} s_k - B_1 \sqrt{ps_k}$, and $\lambda_2 \left(\tilde{A} \right) \leq \|\tilde{W}\|_{\text{op}} \leq B_1 \sqrt{ps_k}$. In the third inequality we use the fact that $\frac{p-q}{2} s_k > (1 - \frac{B_1}{\kappa}) \kappa \sqrt{pn \log m} = (\kappa - B_1) \sqrt{pn \log m}$ from (26). Finally we assume κ is large enough so that $\kappa > 6B_1$ in the fourth inequality.

For the other direction, we can simply use (Eldridge et al., 2018, Theorem 5) and our bound on $|\tilde{v}^\top \Delta \tilde{v}|$ to conclude

$$\lambda_1 \left(\overleftrightarrow{A}^{(k)} \right) = \lambda_1 \left(\tilde{A} + \Delta \right) \geq \lambda_1 \left(\tilde{A} \right) - |\tilde{v}^\top \Delta \tilde{v}| \geq \lambda_1 \left(\tilde{A} \right) - \frac{2B_5 \sqrt{p \log m}}{\sqrt{s_k}} \geq \lambda_1 \left(\tilde{A} \right) - \frac{B_8 \sqrt{p \log m}}{\sqrt{s_k}}.$$

■

Now we prove two preliminary facts about the relaxed oracle SDP and the LOO relaxed oracle SDPs. Note that Lemma 23 states that the event $E_1 \cap \left\{ \lambda_1 \left(\overleftrightarrow{A}^{(k)} \right) > \lambda \right\}$ implies the event $E_1^{k,j} \cap \left\{ \lambda_1 \left(\overleftrightarrow{A}^{k,j} \right) > \lambda^k \right\}$, which could slightly simplify the statements of Lemmas 24 and 25 below.

Lemma 24 *Suppose k satisfies the condition (26). Under the event $E_1 \cap \left\{ \lambda_1 \left(\overleftrightarrow{A}^{(k)} \right) > \lambda \right\}$, we have*

$$\left\| \overline{y}^k \right\|_2 \geq \sqrt{s_k} \left(1 - \frac{B_2}{\kappa - B_1} \right).$$

Also, under the event $E_1^{k,j} \cap \left\{ \lambda_1 \left(\overleftrightarrow{A}^{k,j} \right) > \lambda^k \right\}$, we have

$$\left\| \overline{y}^{k,j} \right\|_2 \geq \sqrt{s_k} \left(1 - \frac{B_2}{\kappa - B_1} \right).$$

Proof Recalling v^k as the top eigenvector of $\overleftrightarrow{A}^{(k)}$, $\frac{v^k}{\|v^k\|_\infty}$ is feasible as all its entries are in $[-1, 1]$. Furthermore, for rank-one solutions $xx^\top$, setting $x = \frac{v^k}{\|v^k\|_\infty}$ is optimal among all vectors x with norm $\leq \frac{1}{\|v^k\|_\infty}$ (even comparing to x with $\|x\|_\infty > 1$) since $\frac{v^k}{\|v^k\|_\infty}$ is a rescaled top eigenvector of $\overleftrightarrow{A}^{(k)}$ and $\langle \overleftrightarrow{A}^{(k)}, xx^\top \rangle = x^\top \overleftrightarrow{A}^{(k)} x$. So we must have $\|\bar{y}^k\|_2 \geq \frac{1}{\|v^k\|_\infty}$, as the optimal solution must have larger 2-norm than $\frac{v^k}{\|v^k\|_\infty}$. Now we use Lemma 16 to upper-bound $\|v^k\|_\infty$.

By (21) we have

$$\begin{aligned} \frac{1}{\|v^k\|_\infty} &\geq \frac{1}{\frac{1}{\sqrt{s_k}} + 2B_2 \frac{\sqrt{p \log m}}{(p-q)s_k}} \\ &\geq \frac{\sqrt{s_k}}{1 + 2B_2 \frac{\sqrt{ps_k \log m}}{(p-q)s_k}} \\ &\geq \sqrt{s_k} \left(1 - 2B_2 \frac{\sqrt{ps_k \log m}}{(p-q)s_k} \right) \\ &\geq \sqrt{s_k} \left(1 - \frac{B_2}{\kappa - B_1} \right) \end{aligned}$$

where we used the fact that $\frac{1}{1+x} \geq 1-x$ and then that

$$\left(1 - \frac{B_1}{\kappa} \right) \kappa \sqrt{ps_k \log m} \leq \left(1 - \frac{B_1}{\kappa} \right) \kappa \sqrt{pn \log m} < \frac{p-q}{2} s_k$$

using (26), which implies

$$\frac{2\sqrt{ps_k \log m}}{(p-q)s_k} \leq \frac{1}{\left(1 - \frac{B_1}{\kappa} \right) \kappa} = \frac{1}{\kappa - B_1}.$$

Therefore

$$\|\bar{y}^k\|_2 \geq \sqrt{s_k} \left(1 - \frac{B_2}{\kappa - B_1} \right).$$

By simply starting with (22), which is part of $E_1^{k,j}$, instead of (21), we have the same bound for $\bar{y}^{k,j}$. $\blacksquare$

Lemma 25 Suppose k satisfies the condition (26). Under the event $E_1 \cap \left\{ \lambda_1 \left(\overleftrightarrow{A}^{(k)} \right) > \lambda \right\}$, we have

$$\bar{e}^\top \bar{y}^k \geq \sqrt{s_k} \left(1 - \frac{B_9}{\kappa - B_1} \right).$$

Also, under the event $E_1^{k,j} \cap \left\{ \lambda_1 \left(\overleftrightarrow{A}^{k,j} \right) > \lambda^k \right\}$, we have

$$\bar{e}^\top \bar{y}^{k,j} \geq \sqrt{s_k} \left(1 - \frac{B_9}{\kappa - B_1} \right).$$

Proof For convenience we abbreviate $y := \bar{y}^k$. Since $\lambda_1 \left(\overleftrightarrow{A}^{(k)} - \lambda I \right) > 0$,

$$y^\top \left(\overleftrightarrow{A}^{(k)} - \lambda I \right) y = \frac{p-q}{2} (\mathbf{1}^\top y)^2 + y^\top W y - \lambda \|y\|_2^2 \geq 0$$

so starting by rearranging this inequality, we have

$$\begin{aligned} \frac{p-q}{2} (\mathbf{1}^\top y)^2 &\geq \lambda \|y\|_2^2 - y^\top W^{(k)} y \\ &\geq \lambda \|y\|_2^2 - \|W^{(k)}\|_{\text{op}} \|y\|_2^2 \\ &\geq \kappa \sqrt{pn \log m} \|y\|_2^2 - B_1 \sqrt{ps_k} \|y\|_2^2 \\ &= \left(1 - \frac{B_1 \sqrt{ps_k}}{\kappa \sqrt{pn \log m}} \right) \kappa \sqrt{pn \log m} \|y\|_2^2 \\ &\geq \left(1 - \frac{B_1}{\kappa} \right) \kappa \sqrt{pn \log m} \|y\|_2^2 \\ &\geq \left(1 - \frac{B_1}{\kappa} \right) \left(1 - \frac{B_6}{\kappa} \right) \frac{p-q}{2} s_k \|y\|_2^2 \\ &\geq \left(1 - \frac{B_1 + B_6}{\kappa} \right) \frac{p-q}{2} s_k \|y\|_2^2 \end{aligned}$$

where we used the fact that under E_1 we have $\|W^{(k)}\|_{\text{op}} \leq B_1 \sqrt{ps_k}$, and then finally that

$$\kappa \sqrt{pn \log m} \geq \frac{p-q}{2} s_k \left(1 + \frac{B_6}{\kappa} \right)^{-1} \geq \frac{p-q}{2} s_k \left(1 - \frac{B_6}{\kappa} \right)$$

which follows from condition (26) and the fact that $\frac{1}{1+x} \geq 1-x$. Dividing both sides by $\frac{p-q}{2} s_k$, noting $\bar{e} = \mathbf{1}/\sqrt{s_k}$, and using Lemma 24 we have

$$\begin{aligned} (\bar{e}^\top y)^2 &\geq \left(1 - \frac{B_1 + B_6}{\kappa} \right) \|y\|_2^2 \\ &\geq \left(1 - \frac{B_1 + B_6}{\kappa} \right) \left(\sqrt{s_k} \left(1 - \frac{B_2}{\kappa - B_1} \right) \right)^2 \\ &\geq s_k \left(1 - \frac{B_1 + B_6 + 2B_2}{\kappa - B_1} \right). \end{aligned}$$

Taking square roots and using the fact that $\sqrt{1-x} \geq 1-x$ we obtain $\bar{e}^\top y \geq \sqrt{s_k} \left(1 - \frac{B_1 + B_6 + 2B_2}{\kappa - B_1} \right)$.

A nearly identical argument works for $\|\bar{y}^{k,j}\|_2$ (note $\|W^{k,j}\|_{\text{op}} \leq B_1 \sqrt{ps_k}$ is guaranteed under $E_1^{k,j}$) except that λ^k is slightly smaller than λ . Here we can simply bound very loosely

$$\lambda^k = \lambda - \frac{B_8 \sqrt{p \log m}}{\sqrt{s_k}} \geq \kappa \sqrt{pn \log m} - B_8 \sqrt{pn \log m} = \left(1 - \frac{B_8}{\kappa} \right) \kappa \sqrt{pn \log m},$$

which will simply add a term of B_8 to the numerator above, leading to the bound

$$\bar{e}^\top \bar{y}^{k,j} \geq \sqrt{s_k} \left(1 - \frac{B_8 + B_1 + B_6 + 2B_2}{\kappa - B_1} \right).$$

Lastly we set $B_9 = B_8 + B_1 + B_6 + 2B_2$. ■

Finally we are able to show that the relaxed oracle SDP solution (28) is all-positive, and thus it is also a solution of the (un-relaxed) oracle SDP (4). For notational convenience, we will only show that the first entry of $\bar{y}^k$ is positive, since the same argument applies to all other entries. Thus we will abbreviate the rank-one relaxed oracle SDP solution by $y := \bar{y}^k$ and the rank-one 1st LOO relaxed oracle solution by $\tilde{y} := \bar{y}^{k,1}$, and similarly we let $Y := \bar{Y}^k$, $\tilde{Y} := \bar{Y}^{k,1}$, $u := u^k$, and $\tilde{u} := u^{k,j}$. We also let $\tilde{\lambda} = \lambda^k$, $\tilde{W} := W^{k,1}$, and $\tilde{w} := w^{k,1}$. Since we are working with a fixed k we will also abbreviate $s := s_k$.

Now we can show one of the key properties of the LOO relaxed oracle solution, which is that $\tilde{y}_1$ is close to 1.

Lemma 26 *Supposing cluster k satisfies the condition (26), under the event $E_3 \cap \left\{ \lambda_1 \left(\overleftrightarrow{A}^{(k)} \right) > \lambda \right\}$,*

$$\tilde{y}_1 \geq 1 - B_{10} \frac{1}{\kappa - 1}.$$

Proof By the complementary slackness conditions (31d) for the LOO relaxed oracle SDP,

$$0 = \tilde{u} \circ \tilde{y} - \frac{p-q}{2} \mathbf{1}^\top \tilde{y} - \tilde{W} \tilde{y} + \tilde{\lambda} \tilde{y}$$

so

$$\begin{aligned} \tilde{y}_1 &= \frac{1}{\tilde{\lambda}} \left(\frac{p-q}{2} \mathbf{1}^\top \tilde{y} + e_1^\top \tilde{W} \tilde{y} - \tilde{u}_1 \tilde{y}_1 \right) \\ &= \frac{1}{\tilde{\lambda}} \left(\frac{p-q}{2} \mathbf{1}^\top \tilde{y} - \tilde{u}_1 \tilde{y}_1 \right). \end{aligned}$$

Now either $\tilde{y}_1 = 1$ and we are done, or otherwise $\tilde{u}_1 \tilde{y}_1 \leq 0$ by complementary slackness, in which case we have

$$\begin{aligned} \tilde{y}_1 &\geq \frac{1}{\tilde{\lambda}} \frac{p-q}{2} \mathbf{1}^\top \tilde{y} \\ &= \frac{1}{\tilde{\lambda}} \frac{p-q}{2} \sqrt{s} \tilde{e}^\top \tilde{y} \\ &\geq \frac{1}{\kappa \sqrt{pn \log m}} \frac{p-q}{2} \sqrt{s} \sqrt{s} \left(1 - \frac{B_9}{\kappa - B_1} \right) \\ &\geq \left(1 - \frac{B_1}{\kappa} \right) \left(1 - \frac{B_9}{\kappa - B_1} \right) \\ &\geq 1 - \frac{B_1 + B_9}{\kappa - B_1} \end{aligned}$$

where we used Lemma 25, the fact that $\frac{1}{\tilde{\lambda}} \geq \frac{1}{(\kappa+1)\sqrt{pn \log m}}$ since $\tilde{\lambda} \leq \lambda$, and then the fact that $\left(1 - \frac{B_1}{\kappa} \right) \kappa \sqrt{pn \log m} < \frac{p-q}{2} s$ from condition (26). ■

Now let $y = \alpha \tilde{y} + \beta t$, where t is a unit vector orthogonal to $\tilde{y}$. Also we may assume for convenience that $\beta \geq 0$ by possibly scaling t by -1 . If we can show that α is close to 1 and β is small, then in light of Lemma 26, we will succeed in showing that y_1 is also close to 1 and thus positive. First we show α is close to 1.

Lemma 27 *Supposing cluster k satisfies the condition (26), under the event $E_3 \cap \left\{ \lambda_1 \left(\overleftrightarrow{A}^{(k)} \right) > \lambda \right\}$,*

$$1 - \frac{B_{11}}{\kappa - B_1} \leq \alpha \leq 1 + \frac{B_{11}}{\kappa - B_1}.$$

Proof By orthogonality of t and $\tilde{y}$ we have

$$\alpha = \frac{\langle \tilde{y}, y \rangle}{\|\tilde{y}\|_2^2}.$$

Now we give upper and lower bounds on α . First we prove the upper bound. Note $\langle y, \tilde{y} \rangle \leq s$ since each coordinate is bounded by 1, and using Lemma 24, we have

$$\alpha = \frac{\langle \tilde{y}, y \rangle}{\|\tilde{y}\|_2^2} \leq \frac{s}{s \left(1 - \frac{B_2}{\kappa - B_1} \right)^2} \leq \frac{1}{1 - \frac{2B_2}{\kappa - B_1}} \leq 1 + \frac{4B_2}{\kappa - B_1}$$

where we use the fact that $\frac{1}{1-a} \leq 1 + 2a$ for $a \in [0, 0.5]$, and assume $\kappa - B_1 \geq 4B_2$.

For the lower bound, write $\langle \tilde{y}, y \rangle = \frac{1}{2} \left(\|\tilde{y}\|_2^2 + \|y\|_2^2 - \|y - \tilde{y}\|_2^2 \right)$. For the terms $\|\tilde{y}\|_2^2$ and $\|y\|_2^2$ we have lower bounds from Lemma 24. Now we need to upper bound $\|y - \tilde{y}\|_2$, which we do by showing that both y and $\tilde{y}$ are close to $\mathbf{1}$ in 2-norm and then using the triangle inequality. Using the Law of Cosines again we have

$$\begin{aligned} \|y - \mathbf{1}\|_2^2 &= \|y\|_2^2 + \|\mathbf{1}\|_2^2 - 2\langle y, \mathbf{1} \rangle \\ &\leq 2s - 2\langle y, \mathbf{1} \rangle \\ &= 2s - 2\sqrt{s}\bar{e}^\top y \\ &\leq 2s - 2s \left(1 - \frac{B_9}{\kappa - B_1} \right) \\ &= 2s \frac{B_9}{\kappa - B_1} \end{aligned}$$

where the inequality steps came from the fact that $\|y\|_\infty \leq 1$ and Lemma 25. The same argument also shows that $\|\tilde{y} - \mathbf{1}\|_2^2 \leq 2s \frac{B_9}{\kappa - B_1}$. Then using the triangle inequality and AM-GM we have

$$\begin{aligned} \|y - \tilde{y}\|_2^2 &\leq 2 \left(\|y - \mathbf{1}\|_2^2 + \|\tilde{y} - \mathbf{1}\|_2^2 \right) \\ &\leq 8s \frac{B_9}{\kappa - B_1}. \end{aligned}$$

Combining with the above arguments we have that

$$\begin{aligned} \langle \tilde{y}, y \rangle &= \frac{1}{2} \left(\|\tilde{y}\|_2^2 + \|y\|_2^2 - \|y - \tilde{y}\|_2^2 \right) \\ &\geq \frac{1}{2} \left(s \left(1 - \frac{B_2}{\kappa - B_1} \right)^2 + s \left(1 - \frac{B_2}{\kappa - B_1} \right)^2 - 8s \frac{B_9}{\kappa - B_1} \right) \\ &\geq \frac{1}{2} \left(s \left(1 - \frac{2B_2}{\kappa - B_1} \right) + s \left(1 - \frac{2B_2}{\kappa - B_1} \right) - 8s \frac{B_9}{\kappa - B_1} \right) \\ &= s \left(1 - \frac{2B_2 + 4B_9}{\kappa - B_1} \right). \end{aligned}$$

Now using the bound $\|\tilde{y}\|_2^2 \leq s$, we have

$$\begin{aligned}\alpha &= \frac{\langle \tilde{y}, y \rangle}{\|\tilde{y}\|_2^2} \\ &\geq 1 - \frac{2B_2 + 4B_9}{\kappa - B_1}.\end{aligned}$$

■

Now we derive a perturbation bound which will be key in showing that β is small.

Lemma 28 *We have*

$$\left\langle \text{diag}(\tilde{u}) - (\tilde{A} - \tilde{\lambda}I), Y \right\rangle \leq \left\langle (\tilde{A} - \tilde{\lambda}I) - (\overleftarrow{A}^{(k)} - \lambda I), \tilde{Y} - Y \right\rangle. \quad (34)$$

Proof This essentially follows from the LOO relaxed oracle KKT conditions (31). Condition (31d) gives $(\text{diag}(\tilde{u}) - (\tilde{A} - \tilde{\lambda}I)) \tilde{Y} = 0$ (since $\tilde{Y} = \overline{y} \overline{y}^\top$). Also by condition (31c) and since $\tilde{u} \geq 0$ from (31b),

$$\left\langle \text{diag}(\tilde{u}), Y - \tilde{Y} \right\rangle = \left\langle \text{diag}(\tilde{u}), Y - I \right\rangle = \sum_{i=1}^s \tilde{u}_i (Y_{ii} - 1) \leq 0.$$

Using these two facts we have

$$\begin{aligned}\left\langle \text{diag}(\tilde{u}) - (\tilde{A} - \tilde{\lambda}I), Y \right\rangle &= \left\langle \text{diag}(\tilde{u}) - (\tilde{A} - \tilde{\lambda}I), Y - \tilde{Y} \right\rangle \\ &= \left\langle \tilde{A} - \tilde{\lambda}I, \tilde{Y} - Y \right\rangle + \left\langle \text{diag}(\tilde{u}), Y - \tilde{Y} \right\rangle \\ &\leq \left\langle \tilde{A} - \tilde{\lambda}I, \tilde{Y} - Y \right\rangle.\end{aligned} \quad (35)$$

Now since $\tilde{Y}$ is feasible for the relaxed oracle SDP (28) and Y is optimal, we have

$$\left\langle \overleftarrow{A}^{(k)} - \lambda I, \tilde{Y} - Y \right\rangle \leq 0 \implies \left\langle -(\overleftarrow{A}^{(k)} - \lambda I), \tilde{Y} - Y \right\rangle \geq 0.$$

Adding this to (35), we obtain (34). ■

Lemma 29 *Supposing cluster k satisfies the condition (26), under the event $E_3 \cap \left\{ \lambda_1 \left(\overleftarrow{A}^{(k)} \right) > \lambda \right\}$,*

$$\beta \leq \frac{B_{14}}{\kappa - B_{15}}$$

Proof First we relate the LHS of (34) to β . The event $E_3 \cap \left\{ \lambda_1 \left(\overleftarrow{A}^{(k)} \right) > \lambda \right\}$ implies the event $E_1^{k,j} \cap \left\{ \lambda_1 \left(\overleftarrow{A}^{k,j} \right) > \lambda^k \right\}$ by Lemma 23, so we may use (31d) which states $(\tilde{\lambda}I + \text{diag}(\tilde{u}) - \tilde{A}) \tilde{y} = 0$. Since $\tilde{y}$ is nonzero this implies that $\tilde{y}$ is an eigenvector with eigenvalue 0.

Also we use a similar calculation to that within the proof of Lemma 21 to lower bound the remaining eigenvalues of $\tilde{\lambda}I + \text{diag}(\tilde{u}) - \tilde{A}$. By Weyl's inequality for any $i \in [s_k]$ with $i < s_k$ we have

$$\begin{aligned}
 \lambda_i \left(\tilde{\lambda}I + \text{diag}(\tilde{u}) - \tilde{A} \right) &= \lambda_i \left(\tilde{\lambda}I + \text{diag}(\tilde{u}) - \tilde{W} - \frac{p-q}{2} J_{s_k \times s_k} \right) \\
 &\geq \lambda_i \left(-\frac{p-q}{2} J_{s_k \times s_k} \right) + \lambda_{s_k} \left(\tilde{\lambda}I + \text{diag}(\tilde{u}) - \tilde{W} \right) \\
 &= \lambda_{s_k} \left(\tilde{\lambda}I + \text{diag}(\tilde{u}) - \tilde{W} \right) \\
 &\geq \lambda_{s_k} \left(\tilde{\lambda}I - \tilde{W} \right) \\
 &\geq \tilde{\lambda} - \|\tilde{W}\|_{\text{op}} \\
 &\geq \kappa \sqrt{pn \log m} - \frac{B_8 \sqrt{p \log m}}{\sqrt{s_k}} - B_1 \sqrt{ps_k} \\
 &\geq (\kappa - B_8 - B_1) \sqrt{pn \log m}
 \end{aligned}$$

where we use the fact that $\lambda_i \left(-\frac{p-q}{2} J_{s_k \times s_k} \right) = 0$ for $i < s_k$.

Therefore using $y = \alpha \tilde{y} + \beta t$ in the LHS of (34) we obtain

$$\begin{aligned}
 \left\langle \text{diag}(\tilde{u}) - \left(\tilde{A} - \tilde{\lambda}I \right), Y \right\rangle &= y^\top \left(\tilde{\lambda}I + \text{diag}(\tilde{u}) - \tilde{A} \right) y \\
 &= \beta^2 t^\top \left(\tilde{\lambda}I + \text{diag}(\tilde{u}) - \tilde{A} \right) t \\
 &\geq \beta^2 (\kappa - B_8 - B_1) \sqrt{pn \log m}
 \end{aligned} \tag{36}$$

using that t is a unit vector orthogonal to $\tilde{y}$ as well as the above facts.

Next expanding the RHS of (34) we have

$$\begin{aligned}
 \left\langle \left(\tilde{A} - \tilde{\lambda}I \right) - \left(\overleftrightarrow{A}^{(k)} - \lambda I \right), \tilde{Y} - Y \right\rangle &= \left\langle -\tilde{w}e_1^\top - e_1\tilde{w}^\top + (\tilde{\lambda} - \lambda)I, \tilde{y}\tilde{y}^\top - yy^\top \right\rangle \\
 &= 2y_1\tilde{w}^\top y - 2\tilde{y}_1\tilde{w}^\top \tilde{y} + \left\langle (\tilde{\lambda} - \lambda)I, \tilde{y}\tilde{y}^\top - yy^\top \right\rangle \\
 &\leq 2y_1\tilde{w}^\top y + 2|\tilde{y}_1| |\tilde{w}^\top \tilde{y}| + |\tilde{\lambda} - \lambda| \sum_{j=1}^s |y_j^2 - \tilde{y}_j^2| \\
 &\leq 2y_1\tilde{w}^\top y + 2|\tilde{w}^\top \tilde{y}| + s|\tilde{\lambda} - \lambda| \\
 &\leq 2y_1\tilde{w}^\top y + 2B_5 \sqrt{ps \log m} + s \frac{B_8 \sqrt{p \log m}}{\sqrt{s}} \\
 &= 2y_1\tilde{w}^\top y + (2B_5 + B_8) \sqrt{ps \log m}
 \end{aligned}$$

using Lemma 22 to bound $|\tilde{w}^\top \tilde{y}|$ and Lemma 23 to bound $|\tilde{\lambda} - \lambda|$.

Now we bound the final term $2y_1 \tilde{w}^\top y$. Using $y = \alpha \tilde{y} + \beta t$ we split

$$\begin{aligned} 2y_1 \tilde{w}^\top y &= 2\langle \tilde{w} e_1^\top, y y^\top \rangle \\ &= 2\langle \tilde{w} e_1^\top, (\alpha \tilde{y} + \beta t)(\alpha \tilde{y} + \beta t)^\top \rangle \\ &= \underbrace{2\alpha^2 \tilde{y}_1 \tilde{w}^\top \tilde{y}}_{T_1} + \underbrace{2\alpha\beta t_1 \tilde{w}^\top \tilde{y}}_{T_2} + \underbrace{2\alpha\beta \tilde{y}_1 \tilde{w}^\top t}_{T_3} + \underbrace{2\beta^2 t_1 \tilde{w}^\top t}_{T_4}. \end{aligned}$$

Now we bound each term. Note that from Lemma 27 we have that $|\alpha| \leq 2$ for sufficiently large κ . Also $\beta t = y - \alpha \tilde{y}$ and $\|y\|_\infty \leq 1$, $\|\alpha \tilde{y}\|_\infty \leq |\alpha| \|\tilde{y}\|_\infty \leq 2$, so $\|\beta t\|_\infty \leq 3$. Using these as well as the aforementioned bounds on $|\tilde{w}^\top \tilde{y}|$ and the fact that $\|t\|_\infty \leq 1$ because it is a unit vector, we have

$$\begin{aligned} T_1 &\leq 2(2)^2(1)B_5\sqrt{ps_k \log m} \leq 8B_5\sqrt{ps_k \log m} \\ T_2 &\leq 2(2)(3)B_5\sqrt{ps_k \log m} \leq 12B_5\sqrt{ps_k \log m} \\ T_3 &\leq 2|\beta|(1)\|\tilde{w}\|_2 \leq 2B_1|\beta|\sqrt{ps} \\ T_4 &\leq 2|\beta|\|\beta t\|_\infty \|\tilde{w}\|_2 \leq 6B_1|\beta|\sqrt{ps}. \end{aligned}$$

Therefore

$$2y_1 \tilde{w}^\top y \leq 20B_5\sqrt{ps \log m} + 8B_1|\beta|\sqrt{ps}$$

and so combining with the earlier bound we have

$$\left\langle \left(\tilde{A} - \tilde{\lambda} I \right) - \left(\overleftrightarrow{A}^{(k)} - \lambda I \right), \tilde{Y} - Y \right\rangle \leq (22B_5 + B_8) \sqrt{ps \log m} + 8B_1|\beta|\sqrt{ps}.$$

Combining this with (34) and (36) we have

$$\begin{aligned} \beta^2 (\kappa - B_8 - B_1) \sqrt{pn \log m} &\leq \left\langle \text{diag}(\tilde{u}) - \left(\tilde{A} - \tilde{\lambda} I \right), Y \right\rangle \\ &\leq \left\langle \left(\tilde{A} - \tilde{\lambda} I \right) - \left(\overleftrightarrow{A}^{(k)} - \lambda I \right), \tilde{Y} - Y \right\rangle \\ &\leq (22B_5 + B_8) \sqrt{ps \log m} + 8B_1|\beta|\sqrt{ps} \\ &\leq (22B_5 + B_8 + 8B_1|\beta|) \sqrt{pn \log m} \\ &= (B_{12} + B_{13}|\beta|) \sqrt{pn \log m} \end{aligned}$$

which implies after rearranging and dividing by $\sqrt{pn \log m}$ that

$$\beta^2 (\kappa - B_8 - B_1) - \frac{B_{13}}{\kappa} |\beta| + \frac{B_{12}}{\kappa} \leq 0.$$

By solving for the positive root in this quadratic of $|\beta|$ we obtain that

$$\begin{aligned} |\beta| &\leq \frac{\frac{B_{13}}{\kappa} + \sqrt{\left(\frac{B_{13}}{\kappa}\right)^2 + \frac{4B_{12}}{\kappa} (\kappa - B_8 - B_1)}}{2(\kappa - B_8 - B_1)} \\ &\leq \frac{\frac{B_{13}}{\kappa} + \frac{B_{13}}{\kappa} + \sqrt{\frac{4B_{12}}{\kappa} (\kappa - B_8 - B_1)}}{2(\kappa - B_8 - B_1)} \\ &\leq \frac{\frac{2B_{13}}{\kappa} + \sqrt{\frac{4B_{12}}{\kappa} \kappa}}{2(\kappa - B_8 - B_1)} \\ &\leq \frac{B_{13} + B_{12}}{\kappa - B_8 - B_1} \end{aligned}$$

as desired. ■

Now combining Lemma 26 (which guaranteed $\tilde{y}_1$ is close to 1), Lemma 27 which bounded α , and the above Lemma 29 which bounds β , we have

$$\begin{aligned} y_1 &= \alpha \tilde{y}_1 + \beta t_1 \\ &\geq \left(1 - \frac{B_{11}}{\kappa - B_1}\right) \left(1 - \frac{B_{10}}{\kappa - 1}\right) - \frac{B_{14}}{\kappa - B_{15}} \\ &\geq 1 - \frac{B_{11}}{\kappa - B_1} - \frac{B_{10}}{\kappa - 1} - \frac{B_{14}}{\kappa - B_{15}} \end{aligned}$$

where for the second line we used the triangle inequality, the fact that β is defined to be positive, and the fact that t is a unit vector so $|t_1| \leq 1$. Now by choosing κ sufficiently large, we are guaranteed that $y_1 \geq \frac{1}{2}$. Again, since the identical argument works for all entries of y , we conclude $y \geq \frac{1}{2}$ elementwise.

Now that we have shown the k th relaxed oracle SDP (28) has rank one and nonnegative solution $y^k y^{k\top}$, we can easily obtain the desired conclusions about the k th oracle SDP (4). First, since all feasible points for the k th oracle SDP are also feasible for the k th relaxed oracle SDP, and also we have shown that the k th relaxed oracle SDP has an optimal solution $y^k y^{k\top}$ which is feasible for the k th oracle SDP, we have that $y^k y^{k\top}$ must also be optimal for the k th oracle SDP. Furthermore, we will now show that $y^k y^{k\top}$ is the unique optimal solution of the k th relaxed oracle SDP, which by this same reasoning implies it must be the unique optimal solution of the k th oracle SDP.

We know from Lemma 21 that any optimal solution to the relaxed oracle SDP (28) must be rank-one, so let one be $y' y'^\top$. Now by repeating the proof of the perturbation bound Lemma 28 up until (35), but replacing the LOO quantities $\tilde{u}, \tilde{\lambda}, \tilde{A}, \tilde{Y}$ with $u^k, \lambda, \overleftrightarrow{A}^{(k)}, \hat{Y}^k$ (which satisfy the analogous conditions (30)), and setting $Y = y' y'^\top$, we obtain

$$\left\langle \text{diag}(u^k) - \left(\overleftrightarrow{A}^{(k)} - \lambda I\right), y' y'^\top \right\rangle \leq \left\langle \overleftrightarrow{A}^{(k)} - \lambda I, \hat{Y}^k - y' y'^\top \right\rangle.$$

Since $y' y'^\top$ is optimal, we have $\left\langle \overleftrightarrow{A}^{(k)} - \lambda I, \hat{Y}^k - y' y'^\top \right\rangle = 0$. Now write $y' = \gamma y + z$, where z is orthogonal to y . Combining the previous display with the fact that $(\lambda I + \text{diag}(u) - \overleftrightarrow{A}^{(k)}) y = 0$ from (31d), we have that

$$z^\top (\lambda I + \text{diag}(u) - \overleftrightarrow{A}^{(k)}) z \leq 0$$

but we already know from the proof of Lemma 21 that $\lambda_i (\lambda I + \text{diag}(u) - \overleftrightarrow{A}^{(k)}) > 0$ for all $i < s_k$ and that the smallest eigenvalue has eigenvector y , so we must have $z = 0$. Thus $y' = \gamma y^k$, and $|\gamma| > 1$ would cause $y' y'^\top$ to be infeasible while $|\gamma| < 1$ would cause $y' y'^\top$ to be suboptimal (since $\lambda_1 (\overleftrightarrow{A}^{(k)} - \lambda I) > 0$), so we must have $\gamma \in \{-1, 1\}$ and thus $y' y'^\top = \hat{Y}^k$. Thus $\hat{Y}^k = y^k y^{k\top}$ is the unique optimal solution to the k th relaxed oracle SDP, and thus by the reasoning in the previous paragraph it is also the unique optimal solution to the k th oracle SDP.

Finally, we set $L^k = 0_{s_k \times s_k}$ and $U^k = \text{diag}(u^k)$. We have $U^k \succcurlyeq 0$ as desired since $u^k \geq 0$ from (30b), and from the conditions (30) as well as the fact that $y^k \geq 0$ elementwise, we immediately have that these choices of L^k, U^k (and $\hat{Y}^k = y^k y^{k\top}$) satisfy the oracle SDP KKT conditions (16).

In summary, all conclusions of all cases of Lemma 14 hold under the event E_3 (which is contained in the event E_1 which sufficed for the non-mid-size cases as shown in Lemmas 19 and 20), and we checked that $\mathbb{P}(E_3) \geq 1 - O(m^{-3})$ as desired. To complete our proof of Lemma 14, note we can ensure that the constants $\underline{B}, \overline{B}$ can be taken to be > 2 by requiring κ sufficiently large.

C.2. Recovery SDP Solution

Now we complete the proof of the main Theorem 2. First we assume we are in the event E_3 defined in the previous subsection, under which the main Lemma 14 from the previous subsection holds. Now we define a further subset of this event, E_4 , which will be the intersection of E_3 with the events described in the following lemmas.

For each cluster k such that $\frac{p-q}{2}s_k \geq B_4\sqrt{pn \log m}$ if the k th oracle solution $\hat{Y}^k = y^k y^{k\top}$ is nonzero we define $x^k = \frac{y^k}{\|y^k\|_2}$, and otherwise if $y^k = 0$ we define $x^k = v^k$ (a top eigenvector of $\overleftarrow{A}^{(k)}$, with properties described in Lemma 16). Our event E_4 will involve these x^k so we establish a basic fact about them.

Lemma 30 *Under the event E_3 , for all clusters k such that $\frac{p-q}{2}s_k \geq B_4\sqrt{pn \log m}$, we have*

$$\|x^k\|_\infty \leq \frac{2}{\sqrt{s_k}}.$$

Proof The case that $y^k = 0$ and $x^k = v^k$ is established by Lemma 17. By Lemma 24 if $\frac{p-q}{2}s_k < (1 + \frac{B_6}{\kappa}) \kappa\sqrt{pn \log m}$ then we will have

$$\|y^k\|_2 \geq \sqrt{s_k} \left(1 - \frac{B_2}{\kappa - B_1}\right) \geq \frac{\sqrt{s_k}}{2}$$

(by our assumptions on the size of κ) and otherwise when $\frac{p-q}{2}s_k \geq (1 + \frac{B_6}{\kappa}) \kappa\sqrt{pn \log m}$ by Lemma 20 we have $y^k = \mathbf{1}$ so $x^k = \frac{1}{\sqrt{s_k}}\mathbf{1}$ which clearly satisfies the conclusion. $\blacksquare$

Lemma 31 *With probability at least $1 - O(m^{-3})$, for all clusters k such that $\frac{p-q}{2}s_k \geq B_4\sqrt{pn \log m}$, for all clusters $j \neq k$, we have*

$$\|W^{(jk)}x^k\|_\infty \leq B_5\sqrt{p \log m}. \quad (37)$$

Lemma 32 *With probability at least $1 - O(m^{-3})$,*

$$\|W\|_{\text{op}} \leq B_1\sqrt{pn}. \quad (38)$$

Lemmas 31 and 32 are proven in Subsection C.4.

We let the event E_4 be the intersection of the event E_3 (under which Lemma 14 on the oracle SDP solutions holds) and the events described in Lemmas 31 and 32. By the union bound (recalling from the previous subsection that E_3 holds with probability at least $1 - O(m^{-3})$) we have that $\mathbb{P}(E_4) \geq 1 - O(m^{-3})$. Now we will construct a solution to the KKT equations (15) of the recovery SDP (1). The rest of the proof will be deterministic, assuming the event E_4 holds.

We define

$$\begin{aligned}\hat{Y} &= \text{diag} \left(y^1 y^{1\top}, \dots, y^K y^{K\top} \right) \\ U &= \text{diag} \left(U^1, \dots, U^K \right).\end{aligned}$$

Note that it is immediate that $\hat{Y}$ satisfies the conditions in (15a), since each block $y^i y^{i\top}$ is feasible for the oracle SDP (4) and thus satisfies $0 \leq y^i y^{i\top} \leq 1$ and $y^i y^{i\top} \succcurlyeq 0$, and a block-diagonal matrix of PSD blocks is PSD. Also from the oracle KKT conditions (16) we have that $U^i \geq 0$ and $(U_{jk}^i - 1)\hat{Y}_{jk}^i = 0 \forall j, k \in [n]$, which ensure $U \geq 0$ and the condition $(U_{jk} - 1)\hat{Y}_{jk} = 0 \forall j, k \in [n]$ from (15b) and (15c).

Finally we describe our choice of L . First we set the diagonal blocks $L^{(i)} = L^i = 0$ for all i . This guarantees that the condition $L_{jk}\hat{Y}_{jk} = 0 \forall j, k \in [n]$ from line (15c) is satisfied, since the diagonal blocks of L are 0 and the off-diagonal blocks of $\hat{Y}$ are 0.

Now we describe the off-diagonal blocks $L^{(ij)}$, which have different forms depending on the sizes of s_i and s_j , so we split into 3 cases. We only describe the cases where $i > j$, because we set $L^{(ji)} = L^{(ij)\top}$. We assume that $\kappa \geq 2B_4$, so that the conclusions of Lemma 31 hold for clusters k such that $\frac{p-q}{2}s_k \geq \frac{\kappa}{2}\sqrt{pn \log m}$.

C.2.1. BOTH CLUSTERS LARGE

If $s_i, s_j \geq \left(\frac{p-q}{2}\right)^{-1} \frac{\kappa\sqrt{pn \log m}}{2}$, we use

$$L^{(ij)} = -\mathbb{E} \overleftrightarrow{A}^{(ij)} - x_i x_i^\top W^{(ij)} - W^{(ij)} x_j x_j^\top + x_i x_i^\top W^{(ij)} x_j x_j^\top$$

This is chosen so that we have

$$\begin{aligned}\overleftrightarrow{A}^{(ij)} + L^{(ij)} &= W^{(ij)} + \mathbb{E} \overleftrightarrow{A}^{(ij)} + L^{(ij)} \\ &= W^{(ij)} - x_i x_i^\top W^{(ij)} - W^{(ij)} x_j x_j^\top + x_i x_i^\top W^{(ij)} x_j x_j^\top \\ &= (I - x_i x_i^\top) W^{(ij)} (I - x_j x_j^\top).\end{aligned}$$

Now we check non-negativity of this choice of $L^{(ij)}$. Since $\|x_i\|_\infty \leq \frac{2}{\sqrt{s_i}}$ and $\|x_j\|_\infty \leq \frac{2}{\sqrt{s_j}}$ by Lemma 30, and since $\|W^{(ji)} x_i\|_\infty, \|W^{(ij)} x_j\|_\infty \leq B_5 \sqrt{p \log m}$ by Lemma 31, we can verify

$$\begin{aligned}& \left\| -x_i x_i^\top W^{(ij)} - W^{(ij)} x_j x_j^\top + x_i x_i^\top W^{(ij)} x_j x_j^\top \right\|_\infty \\ & \leq \left\| x_i x_i^\top W^{(ij)} \right\|_\infty + \left\| W^{(ij)} x_j x_j^\top \right\|_\infty + \left\| x_i x_i^\top W^{(ij)} x_j x_j^\top \right\|_\infty \\ & \leq \|x_i\|_\infty \left\| W^{(ji)} x_i \right\|_\infty + \left\| W^{(ij)} x_j \right\|_\infty \|x_j\|_\infty + \|x_i\|_\infty \|x_j\|_\infty \left| x_i^\top W^{(ij)} x_j \right| \\ & \leq \frac{2}{\sqrt{s_i}} B_5 \sqrt{p \log m} + \frac{2}{\sqrt{s_j}} B_5 \sqrt{p \log m} + \|x_i\|_\infty \|x_j\|_\infty \|x_i\|_1 \left\| W^{(ij)} x_j \right\|_\infty \\ & \leq \frac{2}{\sqrt{s_i}} B_5 \sqrt{p \log m} + \frac{2}{\sqrt{s_j}} B_5 \sqrt{p \log m} + s_i \|x_i\|_\infty^2 \|x_j\|_\infty \left\| W^{(ij)} x_j \right\|_\infty \\ & \leq \frac{2}{\sqrt{s_i}} B_5 \sqrt{p \log m} + \frac{2}{\sqrt{s_j}} B_5 \sqrt{p \log m} + s_i \frac{4}{s_i} \frac{2}{\sqrt{s_j}} B_5 \sqrt{p \log m} \\ & \leq \left(\frac{2}{\sqrt{s_i}} + \frac{10}{\sqrt{s_j}} \right) B_5 \sqrt{p \log m}\end{aligned}$$

which is $\leq \frac{p-q}{2}$ as long as κ is sufficiently large, since by assumption s_i, s_j are each sufficiently large so that

$$\frac{p-q}{2}s_i \geq \frac{\kappa\sqrt{pn\log m}}{2} \implies \frac{p-q}{2} \geq \frac{\kappa\sqrt{pn\log m}}{2s_i} = \frac{\kappa}{2} \frac{\sqrt{p\log m}}{\sqrt{s_i}} \frac{\sqrt{n}}{\sqrt{s_i}} \geq \frac{\kappa}{2} \frac{\sqrt{p\log m}}{\sqrt{s_i}}$$

and likewise for s_j . Now noting that $-\mathbb{E} \overleftrightarrow{A}^{(ij)} = \frac{p-q}{2} J_{s_i \times s_j}$, we have that all entries of $L^{(ij)}$ are positive.

C.2.2. ONE CLUSTER LARGE, ONE CLUSTER SMALL

When $\frac{p-q}{2}s_i \geq \frac{\kappa\sqrt{pn\log m}}{2} > \frac{p-q}{2}s_j$, we use

$$L^{(ij)} = -\mathbb{E} \overleftrightarrow{A}^{(ij)} - x_i x_i^\top W^{(ij)}$$

which ensures that

$$\begin{aligned} \overleftrightarrow{A}^{(ij)} + L^{(ij)} &= W^{(ij)} + \mathbb{E} \overleftrightarrow{A}^{(ij)} + L^{(ij)} \\ &= (I - x_i x_i^\top) W^{(ij)}. \end{aligned}$$

Now we check non-negativity of this choice of $L^{(ij)}$. Again using the facts that $\|x_i\|_\infty \leq \frac{2}{\sqrt{s_i}}$ and $\|W^{(ji)}x_i\|_\infty \leq B_5\sqrt{p\log m}$, we have

$$\left\| x_i x_i^\top W^{(ij)} \right\|_\infty \leq \|x_i\|_\infty \left\| x_i^\top W^{(ij)} \right\|_\infty \leq \frac{2}{\sqrt{s_i}} B_5 \sqrt{p\log n}$$

which identically to the previous case is $\leq \frac{p-q}{2}$ for sufficiently large κ due to the assumption on the size of s_i .

C.2.3. BOTH CLUSTERS SMALL

This is the simplest case. If $s_i, s_j < \left(\frac{p-q}{2}\right)^{-1} \frac{\lambda}{2}$, we set $L^{(ij)} = -\mathbb{E} \overleftrightarrow{A}^{(ij)} = \frac{p-q}{2} J_{s_i \times s_j}$ which is obviously all non-negative. Note in this case we have

$$\overleftrightarrow{A}^{(ij)} + L^{(ij)} = \mathbb{E} \overleftrightarrow{A}^{(ij)} + W^{(ij)} - \mathbb{E} \overleftrightarrow{A}^{(ij)} = W^{(ij)}.$$

C.2.4. CHECKING KKT CONDITIONS

It remains to check that $\overleftrightarrow{A} - \lambda I - U + L \preceq 0$ and $\left(\overleftrightarrow{A} - \lambda I - U + L\right) \hat{Y} = 0_{n \times n}$.

First we check that $\left(\overleftrightarrow{A} - \lambda I - U + L\right) \hat{Y} = 0_{n \times n}$ using blockwise matrix multiplication. We have

$$\begin{aligned} \left(\left(\overleftrightarrow{A} - \lambda I - U + L\right) \hat{Y}\right)^{(ij)} &= \sum_{k=1}^K \left(\overleftrightarrow{A} - \lambda I - U + L\right)^{(ik)} \hat{Y}^{(kj)} \\ &= \left(\overleftrightarrow{A} - \lambda I - U + L\right)^{(ij)} \hat{Y}^{(jj)} \end{aligned}$$

using the fact that $\widehat{Y}^{(kj)} = 0$ for $k \neq j$.

If $i \neq j$, then $\left(\overleftrightarrow{A} - \lambda I - U + L\right)^{(ij)} = \overleftrightarrow{A}^{(ij)} + L^{(ij)}$. If $\frac{p-q}{2}s_j < \frac{\kappa\sqrt{pn\log m}}{2}$, then since $\frac{\kappa\sqrt{pn\log m}}{2} \leq (1 - \frac{B_1}{\kappa})\kappa\sqrt{pn\log m}$, by Lemma 14 $\widehat{Y}^{(jj)} = \widehat{Y}^j = 0$ so $\left(\overleftrightarrow{A} - \lambda I - U + L\right)^{(ij)}\widehat{Y}^{(jj)} = 0$ as desired. If $\frac{p-q}{2}s_j \geq \frac{\kappa\sqrt{pn\log m}}{2}$, if still $\widehat{Y}^{(jj)} = 0$ then again we are done, and otherwise in this case by construction of $L^{(ij)}$ we have $\overleftrightarrow{A}^{(ij)} + L^{(ij)} = M(I - x^j x^{j\top})$ for some M , so then

$$\left(\overleftrightarrow{A}^{(ij)} + L^{(ij)}\right)\widehat{Y}^{(jj)} = M(I - x^j x^{j\top})y^j y^{j\top} = M(y^j y^{j\top} - y^j y^{j\top}) = 0.$$

Next if $i = j$, we have $\left(\overleftrightarrow{A} - \lambda I - U + L\right)^{(ij)} = \left(\overleftrightarrow{A}^{(i)} - \lambda I - U^i + L^i\right)$. Again if $\widehat{Y}^{(jj)} = \widehat{Y}^j = 0$ then we have $\left(\overleftrightarrow{A} - \lambda I - U + L\right)^{(ij)}\widehat{Y}^{(jj)} = 0$ as desired. If $\widehat{Y}^j \neq 0$, by the KKT conditions for the oracle SDP (16d) we have $\left(\overleftrightarrow{A}^{(i)} - \lambda I - U^i + L^i\right)\widehat{Y}^i = 0$ so

$$\left(\overleftrightarrow{A} - \lambda I - U + L\right)^{(ij)}\widehat{Y}^{(jj)} = \left(\overleftrightarrow{A}^{(i)} - \lambda I - U^i + L^i\right)\widehat{Y}^i = 0$$

as desired.

Finally we verify that $\overleftrightarrow{A} - \lambda I - U + L \preceq 0$. Equivalently we check that $\overleftrightarrow{A} - U + L \preceq \lambda I$. Let $\overline{x}^i$ be a zero-padded version of x^i so that $\overline{x}^{i(i)} = x^i$ (recall that for $z \in \mathbb{R}^n$, $z^{(i)} \in \mathbb{R}^{s_i}$ extracts the entries corresponding to cluster i). Let $Q \in \mathbb{R}^{n \times r}$ be a matrix with (orthonormal) columns equal to the $\overline{x}^1, \dots, \overline{x}^r$ corresponding to the clusters large enough so that $s_1, \dots, s_r \geq \left(\frac{p-q}{2}\right)^{-1} \frac{\kappa\sqrt{pn\log m}}{2}$. Let $P = QQ^\top = \text{diag}\left(x^1 x^{1\top}, \dots, x^r x^{r\top}, 0_{s_{r+1} \times s_{r+1}}, \dots, 0_{s_K \times s_K}\right)$.

Lemma 33 *For each $i = 1, \dots, r$, we have*

$$\left(\overleftrightarrow{A} - U + L\right)\overline{x}^i = \alpha^i \overline{x}^i$$

for some $\alpha^i \leq \lambda$.

Proof First, by the zero-padding of $\overline{x}^i$, we have

$$\begin{aligned} \left(\left(\overleftrightarrow{A} - U + L\right)\overline{x}^i\right)^{(j)} &= \sum_{k=1}^K \left(\overleftrightarrow{A} - U + L\right)^{(jk)} \overline{x}^{i(k)} \\ &= \left(\overleftrightarrow{A} - U + L\right)^{(ji)} \overline{x}^{i(i)} \\ &= \left(\overleftrightarrow{A}^{(ji)} - U^{(ji)} + L^{(ji)}\right)x^i \end{aligned}$$

so it remains to show that the above is equal to $\alpha^i x^i$ for some $\alpha^i \leq \lambda$. If $i \neq j$ then $U^{(ji)} = 0$ and $\overleftrightarrow{A}^{(ji)} + L^{(ji)}$ has the form $\overleftrightarrow{A}^{(ji)} + L^{(ji)} = M(I - x^i x^{i\top})$, so in this case $\left(\overleftrightarrow{A}^{(ji)} - U^{(ji)} + L^{(ji)}\right)x^i = 0$. If $i = j$, then we have to consider the cases that $y^i = 0$ and $y^i \neq 0$ separately.

When $y^i = 0$ we define x^i as a top eigenvector v^i of $\overleftrightarrow{A}^{(i)} = \overleftrightarrow{A}^{(i)} - U^i + L^i$ (where we are using the fact from Lemma 14 that $U^i = L^i = 0$ in this case). Thus there exists some α^i such that

$\left(\overleftrightarrow{A}^{(i)} - U^i + L^i\right) x^i = \alpha^i x^i$, and furthermore by condition (16b) from the KKT conditions for the oracle SDP we have that $\alpha^i \leq \lambda$.

When $y^i \neq 0$, then by combining the condition (16d) from the oracle SDP KKT conditions with the fact that $\widehat{Y}^i = y^i y^{i\top}$, we obtain (by right-multiplying by y^i and dividing by $\|y^i\|_2^2 \neq 0$) that

$$\left(\overleftrightarrow{A}^{(i)} - \lambda I - U^i + L^i\right) y^i = 0,$$

so now dividing by $\|y^i\|_2$ and finally recall that in this case we have defined $x^i = \frac{y^i}{\|y^i\|_2}$. ■

Lemma 34 *We have*

$$\begin{aligned} (I - P) \left(\overleftrightarrow{A} - U + L \right) P &= 0, \\ P \left(\overleftrightarrow{A} - U + L \right) P &\preceq \lambda I. \end{aligned}$$

Proof Immediate from the form of P and Lemma 33. ■

Lemma 35 *We have*

$$(I - P) \left(\overleftrightarrow{A} - U + L \right) (I - P) \preceq \lambda I.$$

Proof Let $D = \text{diag}(\overleftrightarrow{A}^{(1)} - U^{(1)}, \dots, \overleftrightarrow{A}^{(r)} - U^{(r)}, \frac{p-q}{2} J_{s_{r+1}}, \dots, \frac{p-q}{2} J_{s_K})$. First we check that $(I - P)D(I - P) \preceq \frac{\lambda}{2}$. Notice that $(I - P)D(I - P)$ is also block-diagonal, with

$$((I - P)D(I - P))^{(i)} = \begin{cases} (I - x^i x^{i\top}) \left(\overleftrightarrow{A}^{(i)} - U^{(i)} \right) (I - x^i x^{i\top}) & i \leq r \\ \frac{p-q}{2} J_{s_i \times s_i} & i > r \end{cases}$$

so $\lambda_1((I - P)D(I - P))$ is bounded by the maximum eigenvalue of the blocks. We have chosen r so that for all $i > r$,

$$\lambda_1 \left(\frac{p-q}{2} J_{s_i \times s_i} \right) = \frac{p-q}{2} s_i \leq \frac{\kappa \sqrt{pn \log m}}{2} \leq \frac{\lambda}{2}.$$

For all $i \leq r$, we know x^i is an eigenvector of $\overleftrightarrow{A}^{(i)} - U^{(i)} + L^{(i)} = \overleftrightarrow{A}^{(i)} - U^{(i)}$ with eigenvalue λ from the proof of Lemma 33. By Weyl's inequality we have for all $j > 1$ that

$$\begin{aligned} \lambda_j \left(\overleftrightarrow{A}^{(i)} - U^{(i)} \right) &= \lambda_j \left(\frac{p-q}{2} J_{s_i \times s_i} + W^{(i)} - U^{(i)} \right) \\ &\leq \lambda_j \left(\frac{p-q}{2} J_{s_i \times s_i} \right) + \lambda_1 \left(W^{(i)} - U^{(i)} \right) \\ &\leq 0 + \left\| W^{(i)} \right\|_{\text{op}} \\ &\leq B_1 \sqrt{pn \log m} \\ &\leq \frac{\kappa \sqrt{pn \log m}}{2} \\ &< \lambda \end{aligned}$$

where we use the facts that $\lambda_j \left(\frac{p-q}{2} J_{s_i \times s_i} \right) = 0$ for $j > 1$, that $U^{(i)} = U^i \succcurlyeq 0$ from Lemma 14, and assume $\kappa \geq 2B_1$. Therefore x^i is a top eigenvector of $\overleftrightarrow{A}^{(i)} - U^{(i)}$ and all other eigenvalues are $\leq \frac{\kappa\sqrt{pn \log m}}{2}$. Now we can calculate that

$$\begin{aligned} \lambda_1 \left((I - x^i x^{i\top}) \left(\overleftrightarrow{A}^{(i)} - U^{(i)} \right) (I - x^i x^{i\top}) \right) &= \sup_{z: \|z\|_2 \leq 1} z^\top \left((I - x^i x^{i\top}) \left(\overleftrightarrow{A}^{(i)} - U^{(i)} \right) (I - x^i x^{i\top}) \right) z \\ &= \sup_{z: \|z\|_2 \leq 1, z \perp x^i} z^\top \left(\overleftrightarrow{A}^{(i)} - U^{(i)} \right) z \\ &= \lambda_2 \left(\overleftrightarrow{A}^{(i)} - U^{(i)} \right) \\ &\leq \frac{\kappa\sqrt{pn \log m}}{2} \end{aligned}$$

as desired, where we used the fact that $(I - x^i x^{i\top})z$ is perpendicular to x^i for any z and also always has norm $\leq \|z\|_2$ since it is an orthogonal projection. Therefore we have checked that $(I - P)D(I - P) \preccurlyeq \frac{\lambda}{2}$.

Now we check that by our construction of L we have

$$\begin{aligned} &(I - P) \left(\overleftrightarrow{A} + L - U - D \right) (I - P) \\ &= (I - P) \left(W - \text{diag} \left(W^{(1)}, \dots, W^{(r)}, 0_{s_r \times s_r}, \dots, 0_{s_K \times s_K} \right) \right) (I - P) \\ &\preccurlyeq 2 \|W\|_{\text{op}} I \\ &\preccurlyeq 2B_1 \sqrt{pn} I \\ &\preccurlyeq \frac{\kappa\sqrt{pn \log m}}{2} I. \end{aligned}$$

This completes the proof of the lemma. ■

Now we combine the results of Lemmas 34 and 35 with the conjugation rule and the fact that P is an orthogonal projection to obtain that

$$\begin{aligned} P \left(\overleftrightarrow{A} - U + L \right) P &= P^2 \left(\overleftrightarrow{A} - U + L \right) P^2 \preccurlyeq \lambda P^2 = \lambda P \\ (I - P) \left(\overleftrightarrow{A} - U + L \right) (I - P) &= (I - P)^2 \left(\overleftrightarrow{A} - U + L \right) (I - P)^2 \preccurlyeq \lambda (I - P)^2 = \lambda (I - P). \end{aligned}$$

Finally we add these two inequalities and use the fact that $(I - P) \left(\overleftrightarrow{A} - U + L \right) P = 0$ from Lemma 34 to obtain that $\overleftrightarrow{A} - U + L \preccurlyeq \lambda I$ as desired.

Thus we have checked all KKT conditions (15) so we conclude that $\hat{Y}$ is a solution to the recovery SDP (1). Finally we need to show that $\hat{Y}$ is the unique solution. We use an argument similar to the uniqueness proof for the mid-size oracle SDP solutions. We start with bound similar to the perturbation bound Lemma 28.

Lemma 36 *Suppose that Y' is feasible for the recovery SDP (1) and $\langle \overleftrightarrow{A} - \lambda I, Y' - Y \rangle = 0$. Then*

$$\langle L, Y' \rangle = 0. \tag{39}$$

Proof The key ingredients will be the recovery SDP KKT conditions (15). From (15c), we have the complementarity condition $\langle L, \hat{Y} \rangle = 0$. Also from (15d) we have $(\overleftrightarrow{A} - \lambda I - U + L) \hat{Y} = 0_{n \times n}$, from (15b) we have $\overleftrightarrow{A} - \lambda I - U + L \preceq 0$, and since Y' is feasible we have $Y' \succeq 0$, so combining these we have

$$\langle \overleftrightarrow{A} - \lambda I - U + L, Y' - \hat{Y} \rangle = \langle \overleftrightarrow{A} - \lambda I - U + L, Y' \rangle \leq 0.$$

Finally, from (15d) we have $\langle U, \hat{Y} \rangle = \langle U, \hat{Y} \rangle$ and we know $U \geq 0$ (from (15b)) and that $Y' \leq 1$ since it is feasible, so

$$\langle U, Y' - \hat{Y} \rangle = \langle U, Y' - J \rangle \leq 0.$$

Now combining all these facts, we can obtain that

$$\begin{aligned} \langle L, Y' \rangle &= \langle L, Y' - \hat{Y} \rangle \\ &= \langle \overleftrightarrow{A} - \lambda I - U + L, Y' - \hat{Y} \rangle + \langle U, Y' - \hat{Y} \rangle - \langle \overleftrightarrow{A} - \lambda I, Y' - Y \rangle \\ &\leq -\langle \overleftrightarrow{A} - \lambda I, Y' - Y \rangle \\ &= 0. \end{aligned}$$

Now since $Y' \geq 0$ by feasibility and $L \geq 0$ from (15b), $\langle L, Y' \rangle \geq 0$, so we must have $\langle L, Y' \rangle = 0$. $\blacksquare$

From the construction for L given in this subsection, all non-block-diagonal entries of L are strictly positive. Therefore if Y' is optimal for the recovery SDP, since it must have $Y' \geq 0$, we can apply Lemma 36 to immediately obtain that Y' is zero for all non-diagonal blocks (that is, $Y'^{(ij)} = 0_{s_i \times s_j}$ whenever $i \neq j$). Now since Lemma 14 states that the oracle diagonal blocks have unique solutions, we must have $Y' = \hat{Y}$, and thus $\hat{Y}$ is the unique solution to the recovery SDP. This completes the proof of Theorem 2.

C.3. Clustering With a Gap Proofs

In this subsection we prove our gap-dependent clustering result Theorem 9, and we reuse the ideas to prove Lemma 20 which handles the case that the oracle SDP will have an all-one solution. Specifically, we first prove Lemmas 37, 38, and 39, which can then be applied to easily prove Theorem 9 and Lemma 20.

Before beginning, we introduce some notation used in this subsection. We let $\bar{s} > \underline{s}$ be consecutive cluster sizes, that is there exists some r such that $\bar{s} = s_r, \underline{s} = s_{r+1}$. $\bar{s}$ will be the smallest cluster recovered in Theorem 9. We define $\bar{Y}^*$ to be a version of the ground truth Y^* which only contains clusters at least the size of $\bar{s}$, that is,

$$\bar{Y}^* = \text{diag} (J_{s_1 \times s_1}, \dots, J_{\bar{s} \times \bar{s}}, 0_{\underline{s} \times \underline{s}}, \dots, 0_{s_K \times s_K}).$$

We also define a version for only clusters at most the size of $\underline{s}$,

$$\underline{Y}^* = Y^* - \bar{Y}^* = \text{diag} (0_{s_1 \times s_1}, \dots, 0_{\bar{s} \times \bar{s}}, J_{\underline{s} \times \underline{s}}, \dots, J_{s_K \times s_K}).$$

Letting $m = \sum_{k=1}^r s_k$ be the number of nodes in the “big” clusters (at least size $\bar{s}$), we define $\underline{J} = \text{diag}(0_{m \times m}, J_{(n-m) \times (n-m)})$ and $\bar{J} = J_{n \times n} - \underline{J}$, that is,

$$\bar{J} = \begin{bmatrix} J_{m \times m} & J_{m \times (n-m)} \\ J_{(n-m) \times m} & 0_{(n-m) \times (n-m)} \end{bmatrix}, \quad \underline{J} = \begin{bmatrix} 0_{m \times m} & 0_{m \times (n-m)} \\ 0_{(n-m) \times m} & J_{(n-m) \times (n-m)} \end{bmatrix}.$$

We let $Q \in \mathbb{R}^{n \times r}$ be a matrix whose (orthonormal) columns are the r singular vectors of $\bar{Y}^*$. Then $QQ^\top = \text{diag}\left(\frac{1}{s_1}J_{s_1 \times s_1}, \dots, \frac{1}{\bar{s}}J_{\bar{s} \times \bar{s}}, 0_{\underline{s} \times \underline{s}}, \dots, 0_{s_K \times s_K}\right)$. Finally we define the projection $\mathcal{P}_\parallel : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}$ and its complement $\mathcal{P}_\perp$ by

$$\begin{aligned} \mathcal{P}_\parallel(M) &= QQ^\top M + MQQ^\top - QQ^\top MQQ^\top, \\ \mathcal{P}_\perp(M) &= M - \mathcal{P}_\parallel(M) = (I - QQ^\top)M(I - QQ^\top). \end{aligned}$$

Lemma 37 *For any Y which is feasible to the recovery SDP (1),*

$$\begin{aligned} \left\| \mathcal{P}_\perp(Y - \bar{Y}^*) \right\|_{\text{nuc}} &\leq \text{Tr}(Y - \bar{Y}^*) + \frac{1}{\bar{s}} \left\| \bar{J} \circ (Y - \bar{Y}^*) \right\|_1 \\ \text{or equivalently } -\text{Tr}(\bar{Y}^* - Y) &\geq \left\| \mathcal{P}_\perp(Y - \bar{Y}^*) \right\|_{\text{nuc}} - \frac{1}{\bar{s}} \left\| \bar{J} \circ (Y - \bar{Y}^*) \right\|_1. \end{aligned}$$

Proof Let $D := Y - \bar{Y}^*$. We have the following chain of inequalities:

$$\begin{aligned} \left\| \mathcal{P}_\perp D \right\|_{\text{nuc}} &= \text{Tr}(\mathcal{P}_\perp D) & \mathcal{P}_\perp(Y - \bar{Y}^*) = \mathcal{P}_\perp(Y) = (I - QQ^\top)Y(I - QQ^\top) \text{ is psd} \\ &= \text{Tr}\left((I - QQ^\top)D(I - QQ^\top)\right) \\ &= \text{Tr}\left((I - QQ^\top)D\right) & \text{cyclic property of trace; } (I - QQ^\top)^2 = I - QQ^\top \\ &= \text{Tr}(D) - \text{Tr}(QQ^\top D) \\ &= \text{Tr}(D) - \left\langle QQ^\top, D \right\rangle \\ &= \text{Tr}(D) - \left\langle QQ^\top, \bar{J}D \right\rangle \\ &\leq \text{Tr}(D) + \left\| QQ^\top \right\|_\infty \left\| \bar{J}D \right\|_1 \\ &\leq \text{Tr}(D) + (1/\bar{s}) \left\| \bar{J}D \right\|_1. \end{aligned}$$

■

Lemma 38 *For any Y which is feasible to the recovery SDP (1),*

$$\left\langle \bar{Y}^* - Y, \mathbb{E}A - \frac{p+q}{2}J \right\rangle \geq \frac{p-q}{2} \left\| \bar{J} \circ (Y - \bar{Y}^*) \right\|_1 - \frac{p-q}{2} \underline{s} \left\| \mathcal{P}_\perp(Y - \bar{Y}^*) \right\|_{\text{nuc}}.$$

Proof Let Y be an arbitrary feasible solution of (1). Let us write

$$\begin{aligned} \left\langle \bar{Y}^* - Y, \mathbb{E}A - \frac{p+q}{2}J \right\rangle &= \left\langle (\mathcal{P}_\parallel + \mathcal{P}_\perp)(\bar{Y}^* - Y), \mathbb{E}A - \frac{p+q}{2}J \right\rangle \\ &= \underbrace{\left\langle \bar{Y}^* - Y, \mathcal{P}_\parallel \left(\mathbb{E}A - \frac{p+q}{2}J \right) \right\rangle}_{F_1} + \underbrace{\left\langle \mathcal{P}_\perp(\bar{Y}^* - Y), \mathcal{P}_\perp \left(\mathbb{E}A - \frac{p+q}{2}J \right) \right\rangle}_{F_2}. \end{aligned}$$

We calculate that

$$\begin{aligned}\mathcal{P}_{\parallel}J &= QQ^{\top}J + JQQ^{\top} - QQ^{\top}JQQ^{\top} \\ &= \begin{bmatrix} J_{m \times m} & J_{m \times (n-m)} \\ 0_{(n-m) \times m} & 0_{(n-m) \times (n-m)} \end{bmatrix} + \begin{bmatrix} J_{m \times m} & 0_{m \times (n-m)} \\ J_{(n-m) \times m} & 0_{(n-m) \times (n-m)} \end{bmatrix} - \begin{bmatrix} J_{m \times m} & 0_{m \times (n-m)} \\ 0_{(n-m) \times m} & 0_{(n-m) \times (n-m)} \end{bmatrix} \\ &= \bar{J}.\end{aligned}$$

It follows that $\mathcal{P}_{\perp}J = J - \mathcal{P}_{\parallel}J = \underline{J}$. Also note that $\mathbb{E}A = (p - q)Y^* + qJ$. Hence

$$\mathcal{P}_{\parallel} \left(\mathbb{E}A - \frac{p+q}{2}J \right) = \mathcal{P}_{\parallel} \left((p-q)Y^* - \frac{p-q}{2}J \right) = (p-q)\bar{Y}^* - \frac{p-q}{2}\bar{J}. \quad (40)$$

and

$$\mathcal{P}_{\perp} \left(\mathbb{E}A - \frac{p+q}{2}J \right) = \mathcal{P}_{\perp} \left((p-q)Y^* - \frac{p-q}{2}J \right) = (p-q)\underline{Y}^* - \frac{p-q}{2}\underline{J}. \quad (41)$$

Using (40), the F_1 term can be written explicitly as

$$\begin{aligned}F_1 &= \left\langle \bar{Y}^* - Y, (p-q)\bar{Y}^* - \frac{p-q}{2}\bar{J} \right\rangle \\ &= \sum_{i,j} (\bar{Y}_{ij}^* - Y_{ij}) \left((p-q)\bar{Y}_{ij}^* - \frac{p-q}{2}\bar{J}_{ij} \right).\end{aligned}$$

For each $(i, j) \in [n] \times [n]$, observe that

- If $\bar{Y}_{ij}^* = 1$, then $\bar{Y}_{ij}^* - Y_{ij} \geq 0$ since $Y_{ij} \leq 1$, and $\bar{J}_{ij} = 1$. Moreover, $(p-q)\bar{Y}_{ij}^* - \frac{p-q}{2}\bar{J}_{ij} = (p-q) - \frac{p-q}{2} = \frac{p-q}{2}\bar{J}_{ij}$. It follows that

$$(\bar{Y}_{ij}^* - Y_{ij}) \left((p-q)\bar{Y}_{ij}^* - \frac{p-q}{2}\bar{J}_{ij} \right) = \frac{p-q}{2}\bar{J}_{ij} |\bar{Y}_{ij}^* - Y_{ij}|.$$

- If $\bar{Y}_{ij}^* = 0$, then $\bar{Y}_{ij}^* - Y_{ij} \leq 0$ since $Y_{ij} \geq 0$. Moreover, $(p-q)\bar{Y}_{ij}^* - \frac{p-q}{2}\bar{J}_{ij} = -\frac{p-q}{2}\bar{J}_{ij}$. It follows that

$$(\bar{Y}_{ij}^* - Y_{ij}) \left(\mathbb{E}A_{ij} - \frac{p+q}{2} \right) = \frac{p-q}{2}\bar{J}_{ij} |Y_{ij}^* - Y_{ij}|.$$

Combining, we see that

$$F_1 = \frac{p-q}{2} \sum_{i,j} |Y_{ij}^* - Y_{ij}| \cdot \bar{J}_{ij} = \frac{p-q}{2} \left\| \bar{J} \circ (\bar{Y}^* - Y) \right\|_1.$$

Turning to the F_2 term, we note that $\mathcal{P}_\perp \left(\mathbb{E}A - \frac{p+q}{2}J \right)$ is supported on $\text{supp}(\underline{J})$ thanks to (41), and $\mathcal{P}_\perp (\underline{J} \circ B) = \underline{J} \circ B$ for any matrix B . Hence

$$\begin{aligned}
 F_2 &= \left\langle \bar{Y}^* - Y, \mathcal{P}_\perp \left(\mathbb{E}A - \frac{p+q}{2}J \right) \right\rangle \\
 &= \left\langle \bar{Y}^* - Y, J \circ \mathcal{P}_\perp \left(\mathbb{E}A - \frac{p+q}{2}J \right) \right\rangle \\
 &= \left\langle J \circ (\bar{Y}^* - Y), \mathcal{P}_\perp \left(\mathbb{E}A - \frac{p+q}{2}J \right) \right\rangle \\
 &= \left\langle \mathcal{P}_\perp \left(J \circ (\bar{Y}^* - Y) \right), \mathbb{E}A - \frac{p+q}{2}J \right\rangle \\
 &= \left\langle J \circ (\bar{Y}^* - Y), \mathbb{E}A - \frac{p+q}{2}J \right\rangle \\
 &= \left\langle \underline{J} \circ (\bar{Y}^* - \hat{Y}), (p-q)\underline{Y}^* - \frac{p-q}{2}\underline{J} \right\rangle,
 \end{aligned}$$

where the last step follows from (41). Further observe that $\underline{J} \circ (\bar{Y}^* - Y)$ is non-positive on all entries, and $(p-q)\underline{Y}^* - \frac{p-q}{2}\underline{J}$ is positive on $\text{supp}(\underline{Y}^*)$ and non-positive elsewhere. Ignoring the non-negative contribution from outside $\text{supp}(\underline{Y}^*)$, we obtain that

$$\begin{aligned}
 F_2 &\geq \left\langle \underline{J} \circ (\bar{Y}^* - Y), \underline{Y}^* \circ \left((p-q)\underline{Y}^* - \frac{p-q}{2}\underline{J} \right) \right\rangle \\
 &= \left\langle \bar{Y}^* - Y, \underline{Y}^* \circ \left((p-q)\underline{Y}^* - \frac{p-q}{2}\underline{J} \right) \right\rangle \\
 &= \left\langle \bar{Y}^* - Y, \frac{p-q}{2}\underline{Y}^* \right\rangle \\
 &= \left\langle \bar{Y}^* - Y, \mathcal{P}_\perp \left(\frac{p-q}{2}\underline{Y}^* \right) \right\rangle \\
 &= \left\langle \mathcal{P}_\perp(\bar{Y}^* - Y), \frac{p-q}{2}\underline{Y}^* \right\rangle \\
 &\geq - \left\| \mathcal{P}_\perp(\bar{Y}^* - Y) \right\|_{\text{nuc}} \left\| \frac{p-q}{2}\underline{Y}^* \right\|_{\text{op}} \\
 &= - \left\| \mathcal{P}_\perp(\bar{Y}^* - Y) \right\|_{\text{nuc}} \frac{p-q}{2} \underline{s}.
 \end{aligned}$$

Combining our calculations for the F_1 and F_2 terms completes the proof. ■

Lemma 39 *For all Y feasible to the recovery SDP (1),*

$$\left| \left\langle \bar{Y}^* - Y, W \right\rangle \right| \leq \left\| \mathcal{P}_\parallel W \right\|_\infty \left\| \bar{J} \circ (Y - \bar{Y}^*) \right\|_1 + \|W\|_{\text{op}} \left\| \mathcal{P}_\perp(Y - \bar{Y}^*) \right\|_{\text{nuc}}.$$

Proof Let $D := Y - Y^*$. Since $\mathcal{P}_{\parallel} W$ is supported on $\text{supp}(\bar{J})$, we have

$$\begin{aligned}\langle \mathcal{P}_{\parallel} D, W \rangle &= \langle D, \mathcal{P}_{\parallel} W \rangle \\ &= \langle D, \bar{J} \circ (\mathcal{P}_{\parallel} W) \rangle \\ &= \langle \bar{J} \circ D, \mathcal{P}_{\parallel} W \rangle,\end{aligned}$$

hence

$$\begin{aligned}|\langle Y - Y^*, W \rangle| &= |\langle \mathcal{P}_{\parallel} D, W \rangle + \langle \mathcal{P}_{\perp} D, W \rangle| \\ &\leq \|\bar{J} \circ D\|_1 \|\mathcal{P}_{\parallel} W\|_{\infty} + \|\mathcal{P}_{\perp} D\|_{\text{nuc}} \|W\|_{\text{op}},\end{aligned}\tag{42}$$

where the last step follows from the generalized Holder's inequality since $\|\cdot\|_1$ ($\|\cdot\|_{\text{nuc}}$, resp.) is the dual norm of $\|\cdot\|_{\infty}$ ($\|\cdot\|_{\text{op}}$, resp.). $\blacksquare$

Proof of Theorem 9 First we bound $\|W\|_{\text{op}}$ and $\|\mathcal{P}_{\parallel} W\|_{\infty}$. By Lemma 32 we have that $\|W\|_{\text{op}} \leq B_1 \sqrt{pn}$ with probability at least $1 - O(n^{-3})$.

For $\|\mathcal{P}_{\parallel} W\|_{\infty}$, first fix j and i and let w be the i th row of W . First note by definition of Q that if $j \in V_{\ell}$ for some ℓ such that $s_{\ell} < \bar{s}$, then $(WQQ^{\top})_{ij} = 0$. Otherwise, letting $j \in V_{\ell}$ where $s_{\ell} \geq \bar{s}$,

$$(WQQ^{\top})_{ij} = \left\langle w^{(\ell)}, \frac{1}{s_{\ell}} \mathbf{1}_{s_{\ell}} \right\rangle = \frac{1}{\sqrt{s_{\ell}}} \left\langle w^{(\ell)}, \frac{1}{\sqrt{s_{\ell}}} \mathbf{1}_{s_{\ell}} \right\rangle.$$

By Bernstein's inequality, since all entries of w have variance $\leq p$, mean zero, and are bounded in magnitude by 1, we have

$$\begin{aligned}\mathbb{P} \left(\left| \left\langle w^{(\ell)}, \frac{1}{\sqrt{s_{\ell}}} \mathbf{1}_{s_{\ell}} \right\rangle \right| \geq t \right) &\leq 2 \exp \left(- \frac{t^2/2}{p \left\| \frac{1}{\sqrt{s_{\ell}}} \mathbf{1}_{s_{\ell}} \right\|_2^2 + \frac{t}{3} \left\| \frac{1}{\sqrt{s_{\ell}}} \mathbf{1}_{s_{\ell}} \right\|_{\infty}} \right) \\ &\leq 2 \exp \left(- \frac{t^2/2}{p + \frac{t}{3\sqrt{s_{\ell}}}} \right) \\ &= 2 \exp \left(- \frac{pB_5^2(\log n)/2}{p + \frac{B_5\sqrt{p\log n}}{3\sqrt{s_{\ell}}}} \right)\end{aligned}\tag{43}$$

where in the last line we choose $t = B_5 \sqrt{p \log n}$ (similarly to the proof of Lemma 17). We have $s_{\ell} \geq \bar{s}$, and by assumption we have $\frac{p-q}{2} (\bar{s} - \underline{s}) \geq \sqrt{pn \log n}$, which implies that

$$ps_{\ell} \geq p\bar{s} \geq \frac{p-q}{2} (\bar{s} - \underline{s}) \geq \sqrt{pn \log n} \geq \sqrt{ps_{\ell} \log n} \implies \frac{\sqrt{p \log n}}{\sqrt{s_{\ell}}} \leq p$$

so the leading term of the denominator is p , and therefore by taking B_5 large enough (the same value as in Lemma 17) we have that the expression (43) is $O(n^{-5})$. Now taking a union bound over all i and j ($O(n^2)$ pairs), we have that

$$\left\| WQQ^{\top} \right\|_{\infty} \leq \sup_{s_{\ell} \geq \bar{s}} \frac{1}{\sqrt{s_{\ell}}} B_5 \sqrt{p \log n} \leq B_5 \sqrt{\frac{p \log n}{\bar{s}}}$$

with probability at least $1 - O(n^{-3})$.

Notice that $\|QQ^\top WQQ^\top\|_\infty \leq \|WQQ^\top\|_\infty$ since

$$QQ^\top = \text{diag}\left(\frac{1}{s_1}J_{s_1 \times s_1}, \dots, \frac{1}{\bar{s}}J_{\bar{s} \times \bar{s}}, 0_{\underline{s} \times \underline{s}}, \dots, 0_{s_K \times s_K}\right)$$

which implies that $(QQ^\top WQQ^\top)_{ij}$ is an average of some entries from column j of $WQQ^\top$ (or 0). Using triangle inequality and then this fact, we have

$$\|\mathcal{P}_\parallel W\|_\infty \leq \|WQQ^\top\|_\infty + \|QQ^\top W\|_\infty + \|QQ^\top WQQ^\top\|_\infty \leq 3\|WQQ^\top\|_\infty \leq 3B_5\sqrt{\frac{p \log n}{\bar{s}}}.$$

We now assume that we are in the event that the bounds $\|W\|_{\text{op}} \leq B_1\sqrt{pn}$ and $\|\mathcal{P}_\parallel W\|_\infty \leq 3B_5\sqrt{\frac{p \log n}{\bar{s}}}$, which by the union bound occurs with probability at least $1 - O(n^{-3})$.

Starting by using the optimality of the SDP solution $\hat{Y}$, we have

$$0 \geq \left\langle \bar{Y}^* - \hat{Y}, A - \frac{p+q}{2}J \right\rangle - \lambda \text{Tr}(\bar{Y}^* - \hat{Y}) \quad (44)$$

$$= \left\langle \bar{Y}^* - \hat{Y}, \mathbb{E}A - \frac{p+q}{2}J \right\rangle + \left\langle \bar{Y}^* - \hat{Y}, W \right\rangle - \lambda \text{Tr}(\bar{Y}^* - \hat{Y}) \quad (45)$$

where recall we define $W = A - \mathbb{E}A$. Our first objective will be to lower bound the final expression by a multiple of $\|\bar{J} \circ (\hat{Y} - \bar{Y}^*)\|_1$, which will then imply that $\|\bar{J} \circ (\hat{Y} - \bar{Y}^*)\|_1 = 0$ and thus $\bar{J} \circ \hat{Y} = \bar{Y}^*$. We accomplish this by using Lemmas 37, 38, and 39. Using their bounds in (45) and combining terms we have

$$\begin{aligned} 0 &\geq \left\langle \bar{Y}^* - \hat{Y}, \mathbb{E}A - \frac{p+q}{2}J \right\rangle + \left\langle \bar{Y}^* - \hat{Y}, W \right\rangle - \lambda \text{Tr}(\bar{Y}^* - \hat{Y}) \\ &\geq \left(\frac{p-q}{2} - \|\mathcal{P}_\parallel W\|_\infty - \frac{\lambda}{\bar{s}} \right) \|\bar{J} \circ (\hat{Y} - \bar{Y}^*)\|_1 + \left(\lambda - \|W\|_{\text{op}} - \frac{p-q}{2}\underline{s} \right) \|\mathcal{P}_\perp(\hat{Y} - \bar{Y}^*)\|_{\text{nuc}} \\ &\geq \left(\frac{p-q}{2} - 3B_5\sqrt{\frac{p \log n}{\bar{s}}} - \frac{\lambda}{\bar{s}} \right) \|\bar{J} \circ (\hat{Y} - \bar{Y}^*)\|_1 \end{aligned}$$

using the fact that $\lambda = B_1\sqrt{pn} + \frac{p-q}{2}\underline{s} > \|W\|_{\text{op}} + \frac{p-q}{2}\underline{s}$. Therefore if $\frac{p-q}{2}\bar{s} > 3B_5\sqrt{p\bar{s} \log n} + \lambda$ we can conclude that $\bar{J} \circ \hat{Y} = \bar{Y}^*$.

Now we show that also $\underline{J} \circ \hat{Y} = 0$ or equivalently that $\hat{Y} = \bar{Y}^*$. Since we know $\bar{J} \circ \hat{Y} = \bar{Y}^*$, it suffices to show that for any feasible Y with $\bar{J} \circ Y = \bar{Y}^*$, we have $\left\langle Y - \bar{Y}^*, A - \frac{p+q}{2}J \right\rangle -$

$\lambda \text{Tr}(Y - \bar{Y}^*) \leq 0$. Starting with the fact that $\bar{J} \circ Y = \bar{Y}^*$,

$$\begin{aligned}
\left\langle Y - \bar{Y}^*, A - \frac{p+q}{2}J \right\rangle - \lambda \text{Tr}(Y - \bar{Y}^*) &= \left\langle \underline{J} \circ Y, \underline{J} \circ \left(A - \frac{p+q}{2}J \right) \right\rangle - \lambda \text{Tr}(\underline{J} \circ Y) \\
&= \left\langle \underline{J} \circ Y, \underline{J} \circ \frac{p-q}{2}\underline{J} \right\rangle + \langle \underline{J} \circ Y, \underline{J} \circ W \rangle - \lambda \text{Tr}(\underline{J} \circ Y) \\
&\leq \|\underline{J} \circ Y\|_{\text{nuc}} \left\| \frac{p-q}{2}\underline{J} \right\|_{\text{op}} + \|\underline{J} \circ Y\|_{\text{nuc}} \|W\|_{\text{op}} - \lambda \text{Tr}(\underline{J} \circ Y) \\
&= \left(\left\| \frac{p-q}{2}\underline{J} \right\|_{\text{op}} + \|W\|_{\text{op}} - \lambda \right) \text{Tr}(\underline{J} \circ Y) \\
&= \left(\frac{p-q}{2}\bar{s} + \|W\|_{\text{op}} - \lambda \right) \text{Tr}(\underline{J} \circ Y) \tag{46}
\end{aligned}$$

where we then used the generalized Holder's inequality and then fact that $\|\underline{J} \circ Y\|_{\text{nuc}} = \text{Tr}(\underline{J} \circ Y)$ since $\underline{J} \circ Y \succcurlyeq 0$, which follows from Schur's Lemma since $Y, \underline{J} \succcurlyeq 0$. Because $\lambda > \|W\|_{\text{op}} - \frac{p-q}{2}\bar{s}$, (46) establishes that $\left\langle Y - \bar{Y}^*, A - \frac{p+q}{2}J \right\rangle - \lambda \text{Tr}(Y - \bar{Y}^*) \leq 0$, and furthermore if $\underline{J} \circ \hat{Y} \neq 0$ then we will have $\left\langle Y - \bar{Y}^*, A - \frac{p+q}{2}J \right\rangle - \lambda \text{Tr}(Y - \bar{Y}^*) < 0$ (since then $\text{Tr}(\underline{J} \circ Y) = \|\underline{J} \circ Y\|_{\text{nuc}} > 0$), establishing uniqueness.

Finally we establish the claimed recovery condition by noting that if $\frac{p-q}{2}(\bar{s} - \underline{s}) \geq (3B_5 + B_1)\sqrt{pn \log n}$, then we will have

$$\frac{p-q}{2}\bar{s} \geq \frac{p-q}{2}\underline{s} + (3B_5 + B_1)\sqrt{pn \log n} > 3B_5\sqrt{p\bar{s} \log n} + \lambda$$

which was shown above to be sufficient. $\blacksquare$

Proof of Lemma 20 Lemma 20 is essentially a special case of Theorem 9, but since we have a slightly different choice of λ , we must repeat the proof.

Fix a cluster k such that $\frac{p-q}{2}\bar{s} \geq (1 + \frac{B_6}{\kappa})\kappa\sqrt{pn \log m}$, where we set $B_6 = 2 + 3B_5$. Since we treat $A^{(k)}$ as the entire adjacency matrix for the purposes of this proof, the conclusions of Lemmas 37, 38, and 39 become that for all Y feasible to the k th oracle SDP (4), we have

$$-\text{Tr}(J_{s_k \times s_k} - Y) \geq \|\mathcal{P}_\perp(Y - J_{s_k \times s_k})\|_{\text{nuc}} - \frac{1}{s_k} \|Y - J_{s_k \times s_k}\|_1 \tag{47}$$

$$\left\langle J_{s_k \times s_k} - Y, \mathbb{E}A^{(k)} - \frac{p+q}{2}J_{s_k \times s_k} \right\rangle \geq \frac{p-q}{2} \|Y - J_{s_k \times s_k}\|_1 \tag{48}$$

$$\left| \left\langle J_{s_k \times s_k} - Y, W^{(k)} \right\rangle \right| \leq \|\mathcal{P}_\parallel W^{(k)}\|_\infty \|Y - J_{s_k \times s_k}\|_1 + \|W^{(k)}\|_{\text{op}} \|\mathcal{P}_\perp(Y - J_{s_k \times s_k})\|_{\text{nuc}} \tag{49}$$

because now $\bar{Y}^* = \bar{J} = J_{s_k \times s_k}$, $\bar{s} = s_k$, and there are no other clusters (so $\underline{s} = 0$).

Since we still assume we are in the event E_1 , we have $\|W^{(k)}\|_{\text{op}} \leq B_1\sqrt{ps_k} \leq \lambda$ (we assume $\kappa \geq 1$ which ensures that s_k meets the condition in (18), and also that $\kappa \geq B_1$). We also have by Lemma 18 that $\left\| W^{(k)} \frac{1}{\sqrt{s_k}} \mathbf{1}_{s_k} \right\|_\infty \leq B_5\sqrt{p \log m}$ (since we also assume $\kappa \geq B_4$). Then since

$$W^{(k)} V V^\top = W^{(k)} \frac{1}{s_k} J_{s_k \times s_k} = \frac{1}{\sqrt{s_k}} W^{(k)} \frac{1}{\sqrt{s_k}} \mathbf{1}_{s_k} \mathbf{1}_{s_k}^\top$$

we have $\left\| W^{(k)} \frac{1}{s_k} J_{s_k \times s_k} \right\|_\infty \leq \frac{B_5 \sqrt{p \log m}}{\sqrt{s_k}}$. Identically to the proof of 9 we have $\|VV^\top W^{(k)} VV^\top\|_\infty \leq \|W^{(k)} VV^\top\|_\infty$, so we conclude that

$$\left\| \mathcal{P}_\parallel W^{(k)} \right\|_\infty \leq 3 \left\| W^{(k)} \frac{1}{s_k} J_{s_k \times s_k} \right\|_\infty \leq \frac{3B_5 \sqrt{p \log m}}{\sqrt{s_k}}.$$

Now like in the proof of Theorem 9, using the optimality of the k th oracle SDP solution $\hat{Y}^k$ and then plugging in the bounds (47), (48), and (49), we have

$$\begin{aligned} 0 &\geq \left\langle J_{s_k \times s_k} - \hat{Y}^k, A^{(k)} - \frac{p+q}{2} J_{s_k \times s_k} \right\rangle - \lambda \text{Tr} \left(J_{s_k \times s_k} - \hat{Y}^k \right) \\ &= \left\langle J_{s_k \times s_k} - \hat{Y}^k, \mathbb{E} A^{(k)} - \frac{p+q}{2} J_{s_k \times s_k} \right\rangle + \left\langle J_{s_k \times s_k} - \hat{Y}^k, W^{(k)} \right\rangle - \lambda \text{Tr} (J_{s_k \times s_k} - \hat{Y}^k) \\ &\geq \left(\lambda - \|W^{(k)}\|_{\text{op}} \right) \left\| \mathcal{P}_\perp (\hat{Y}^k - J_{s_k \times s_k}) \right\|_{\text{nuc}} + \left(\frac{p+q}{2} - \left\| \mathcal{P}_\parallel W^{(k)} \right\|_\infty - \frac{\lambda}{s_k} \right) \left\| \hat{Y}^k - J_{s_k \times s_k} \right\|_1 \\ &\geq \left(\frac{p+q}{2} - \frac{3B_5 \sqrt{p \log m}}{\sqrt{s_k}} - \frac{\lambda}{s_k} \right) \left\| \hat{Y}^k - J_{s_k \times s_k} \right\|_1 \end{aligned} \quad (50)$$

using the fact that $\lambda \geq \|W^{(k)}\|_{\text{op}}$. Now since $\frac{p+q}{2} s_k \geq (1 + \frac{B_6}{\kappa}) \kappa \sqrt{pn \log m} = \kappa \sqrt{pn \log m} + (2 + 3B_5) \sqrt{pn \log m}$, using the fact that we assumed $\lambda \leq (\kappa + 1) \sqrt{pn \log m}$ almost surely, we have

$$\begin{aligned} \frac{p-q}{2} s_k &\geq (\kappa + 1) \sqrt{pn \log m} + 3B_5 \sqrt{pn \log m} + \sqrt{pn \log m} \\ &> \lambda + 3B_5 \sqrt{ps_k \log m} \end{aligned}$$

and plugging this back in to (50) implies that we must have $\left\| \hat{Y}^k - J_{s_k \times s_k} \right\|_1 = 0$ and thus $\hat{Y}^k = J_{s_k \times s_k}$.

The fact that there exist some U^k, L^k which satisfy the oracle SDP KKT conditions (16) follows from Lemma 13 (as a solution of the oracle SDP KKT conditions (16) is necessary for the optimality of $\hat{Y}^k$). From the complementary slackness conditions (16c) and the fact that $\hat{Y}_{ij}^k = 1$ for all i, j , we must have $L^k = 0$. Finally, it is not immediate that $U^k \succcurlyeq 0$, but we ensure this by finding $u \in \mathbb{R}^{s_k}$ such that $\text{diag}(u) \succcurlyeq 0$ and such that $\text{diag}(u)$ also satisfies the oracle SDP KKT conditions (still with $L^k = 0$). Define u as the unique solution to $U^k \mathbf{1}_{s_k} = \text{diag}(u) \mathbf{1}_{s_k}$. Clearly $u \geq 0$ since $U^k \geq 0$, so we have $\text{diag}(u) \succcurlyeq 0$ and $\text{diag}(u) \geq 0$. Also by definition of u the complementarity condition (16d) still holds with $\text{diag}(u)$ in place of U^k , since $\hat{Y}^k = \mathbf{1}\mathbf{1}^\top$. This leaves us only to check the dual feasibility condition (16b) that $\overleftarrow{A}^{(k)} - \lambda I - \text{diag}(u) + L^k \preceq 0$. Again since (16d) holds with $\text{diag}(u)$ in place of U^k , we have that $\mathbf{1}$ is an eigenvector of $\overleftarrow{A}^{(k)} - \lambda I - \text{diag}(u) + L^k$ (with eigenvalue 0), and for all vectors v orthogonal to $\mathbf{1}$, we have

$$\begin{aligned} v^\top \left(\overleftarrow{A}^{(k)} - \lambda I - \text{diag}(u) + L^k \right) v &= v^\top \left(\frac{p-q}{2} \mathbf{1}\mathbf{1}^\top + W^{(k)} - \lambda I - \text{diag}(u) + L^k \right) v \\ &= v^\top \left(W^{(k)} - \lambda I - \text{diag}(u) \right) v \\ &\leq v^\top \left(W^{(k)} - \lambda I \right) v \\ &< 0 \end{aligned}$$

using the facts that $v^\top \mathbf{1} = 0$, $L^k = 0$, and that $\lambda > \|W^{(k)}\|_{\text{op}}$. Therefore $\overleftrightarrow{A}^{(k)} - \lambda I - \text{diag}(u) + L^k \preceq 0$ as desired. $\blacksquare$

C.4. Concentration Inequalities

Theorem 40 (Bernstein's Inequality (Vershynin, 2018)) Suppose $(X_i)_{i=1}^n$ are independent zero-mean random variables such that $|X_i| \leq M$ almost surely. Then for any $t \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{\frac{1}{2}t^2}{\sum_{i=1}^n \mathbb{E}[X_i^2] + \frac{1}{3}Mt}\right).$$

Theorem 41 ((Bandeira and Van Handel, 2016)) Suppose X is a symmetric $n \times n$ matrix with independent entries such that $\mathbb{E}X_{ij} = 0$ and $|X_{ij}| \leq B$ for all $1 \leq i, j \leq n$. Let $\nu = \max_i \sum_j \mathbb{E}X_{ij}^2$. Then there exists an absolute constant $c > 0$ such that for any $t \geq 0$,

$$\mathbb{P}\left(\|X\|_{\text{op}} \geq 4\sqrt{\nu} + t\right) \leq n \exp\left(-\frac{t^2}{cB^2}\right).$$

Proof of Lemma 15 We simply check the conditions of Theorem 41. First we establish (20), which we will actually establish by showing that $\|W\|_{\text{op}} \leq B_1\sqrt{pn}$, which immediately implies (20) (although we could simply state this bound on $\|W\|_{\text{op}}$ in Lemma 15, we want the event in the lemma to be independent of the non-block-diagonal noise entries). We have $\mathbb{E}W_{ij} = 0$, $|W_{ij}| \leq \max\{p, 1-p\} = 1-p \leq 1$, and $\mathbb{E}W_{ij}^2 = p(1-p) \leq p$, so with $\nu = np$ and $B = 1$ in Theorem 41 we have

$$\mathbb{P}\left(\|W\|_{\text{op}} \geq 4\sqrt{pn} + t\right) \leq n \exp\left(-\frac{t^2}{c}\right).$$

By our assumption that $p \geq \frac{\log m}{n}$, $\sqrt{pn} \geq \sqrt{\log m}$, so setting $t = 3\sqrt{c \log m}$ the above probability is $\leq n \exp(-9 \log m) = nm^{-9} \leq m^{-8}$ (since $n \leq m$). Now take $B_1 = 4 + 3\sqrt{c}$ to obtain (20).

To establish (18), following the same steps for a fixed cluster k such that $\frac{p-q}{2}s_k \geq \sqrt{pn \log m}$ we will obtain

$$\mathbb{P}\left(\left\|W^{(k)}\right\|_{\text{op}} \geq 4\sqrt{ps_k} + t\right) \leq s_k \exp\left(-\frac{t^2}{c}\right).$$

Now we use the assumed size of the cluster to show that $\sqrt{ps_k}$ is the dominant term when we set $t = \Theta(\sqrt{\log m})$. We have

$$ps_k \geq \frac{p-q}{2}s_k \geq \sqrt{pn \log m} \geq \log m$$

where the final inequality again uses that $p \geq \frac{\log m}{n}$, so taking square roots $\sqrt{ps_k} \geq \sqrt{\log m}$ and therefore we may again set $t = 3\sqrt{c \log m}$ and obtain a failure probability of m^{-8} .

To establish (19) note $\left\|W_{:,j}^{(k)}\right\|_2 \leq \|W^{(k)}\|_{\text{op}}$. Finally by taking a union bound, there are obviously fewer than $n \leq m$ clusters so the overall failure probability is $O(m^{-7})$. $\blacksquare$

Proof of Lemma 16 A bound of this form first appeared in [Abbe et al. \(2017\)](#), but we use the version from [\(Chen et al., 2021, Theorem 4.2\)](#). One technical issue is that [\(Chen et al., 2021, Theorem 4.2\)](#) uses a failure probability which is inverse polynomial in the size of the matrix in the theorem, which in our case will be s_i , but we seek a failure probability which is inverse polynomial in m . However, by inspecting the proof, thanks to the fact that $m \geq n \geq s_i$, the exact same argument goes through if we replace $\log s_i$ by $\log m$ in the conclusion as well as the conditions of [\(Chen et al., 2021, Theorem 4.2\)](#).

Now fix a cluster i such that $\frac{p-q}{2}s_i \geq \sqrt{pn \log m}$. As checked in the proof of Lemma 15, this implies that $\sqrt{ps_i} \geq \sqrt{\log m}$. Now we check the conditions of [\(Chen et al., 2021, Theorem 4.2\)](#). Note $A^{(i)} - \frac{p+q}{2}J_{s_i \times s_i} = W^{(i)} + \frac{p-q}{2}J_{s_i \times s_i}$, and $\frac{p-q}{2}J_{s_i \times s_i}$ is rank 1 and has top eigenvector $\frac{1}{\sqrt{s_i}}\mathbf{1}$, which has incoherence parameter $s_i \left\| \frac{1}{\sqrt{s_i}}\mathbf{1} \right\|_\infty^2 = 1$ and top eigenvalue $\frac{p-q}{2}s_i$. Each entry of $W^{(i)}$ is independent with mean zero, variance $\leq p(1-p) \leq p$, and is almost surely bounded by 1. [\(Chen et al., 2021, Theorem 4.2\)](#) requires that $\frac{1}{\sqrt{ps_i/\log m}} = O(1)$, which is satisfied due to our assumption on the size of s_i , since as mentioned this assumption implies $\sqrt{ps_i} \geq \sqrt{\log m}$.

Therefore by [\(Chen et al., 2021, Theorem 4.2\)](#) there exists an absolute constant $c > 0$ such that if $\sqrt{ps_i \log m} \leq c \frac{p-q}{2}s_i$, there exists a constant B_2 such that with probability at least $1 - O(m^{-5})$ we have

$$\left\| v^i - \frac{1}{\sqrt{s_i}}\mathbf{1} \right\|_\infty \leq 2B_2 \frac{\sqrt{p \log m}}{(p-q)s_i}.$$

We now let $B_3 = \frac{1}{c}$ and note that

$$\frac{p-q}{2}s_i \geq B_3 \sqrt{pn \log m} \implies \sqrt{ps_i \log m} \leq \sqrt{pn \log m} \leq \frac{1}{c} \frac{p-q}{2}s_i.$$

Furthermore by examining the proof of [\(Chen et al., 2021, Theorem 4.2\)](#) it is also established for all $j \in [s_i]$ that

$$\left\| v^{i,j} - \frac{1}{\sqrt{s_i}}\mathbf{1} \right\|_\infty \leq 2B_2 \frac{\sqrt{p \log m}}{(p-q)s_i}$$

(in the same event where the bound on v^i is shown to hold).

Now repeating this argument for each sufficiently large cluster and taking a union bound over all $O(n) \leq O(m)$ clusters, we obtain the desired conclusion. $\blacksquare$

Proof of Lemma 17 We let $B_4 = \max\{B_2, B_3, 1\}$. Now we fix a cluster k such that $\frac{p-q}{2}s_k \geq \max\{B_2, B_3\}\sqrt{pn \log m}$. Under the event in Lemma 16, for all $j \in [s_k]$ we have that

$$\begin{aligned} \left\| v^{k,j} - \frac{1}{\sqrt{s_k}}\mathbf{1} \right\|_\infty &\leq 2B_2 \frac{\sqrt{p \log m}}{(p-q)s_k} \\ &= \frac{1}{\sqrt{s_k}} \frac{2}{(p-q)s_k} B_2 \sqrt{ps_k \log m} \\ &\leq \frac{1}{\sqrt{s_k}}. \end{aligned}$$

Therefore we have that $\left\| v^{k,j} \right\|_\infty \leq \frac{2}{\sqrt{s_k}}$.

Now also fix j and define $F^{k,j}$ to be the event where (23) holds, and note that $F^{k,j}$ is independent of $w^{k,j}$. Therefore we can apply Bernstein's inequality conditionally on the event $F^{k,j}$ to obtain

$$\mathbb{P}\left(\left|\left\langle w^{k,j}, v^{k,j} \right\rangle\right| \geq t \middle| F^{k,j}\right) \leq 2 \exp\left(\frac{-t^2/2}{p1 + \frac{t}{3} \frac{2}{\sqrt{s_k}}}\right)$$

using the facts that $\|v^{k,j}\|_2^2 = 1$, $\|v^{k,j}\|_\infty \leq \frac{2}{\sqrt{s_k}}$ on $F^{k,j}$. Now we can take $t = B_5 \sqrt{p \log m}$ and note that as we checked in the proof of Lemma 15, the assumption that $\frac{p-q}{2} s_k \geq B_4 \sqrt{pn \log m} \geq \sqrt{pn \log m}$ implies $\sqrt{ps_k} \geq \sqrt{\log m}$ which ensures that

$$\frac{-t^2/2}{p + \frac{2t}{3\sqrt{s_k}}} \leq \frac{-pB_5^2 \log m/2}{p + p \frac{2B_5 \sqrt{\log m}}{3\sqrt{ps_k}}} \leq -\frac{B_5^2/2}{1 + \frac{2B_5}{3}} \log m$$

so we can make $\exp\left(\frac{-t^2/2}{p + \frac{2t}{3\sqrt{s_k}}}\right) \leq m^{-5}$ by taking B_5 sufficiently large. Also, by inspecting the proof of Lemma 17, the event where (22) holds occurs with probability at least $1 - O(m^{-4})$, and as shown above $F^{k,j}$ is guaranteed to hold when (22) holds, so we have that $\mathbb{P}(F^{k,j^c}) \leq O(m^{-4})$.

Therefore

$$\begin{aligned} \mathbb{P}\left(\left|\left\langle w^{k,j}, v^{k,j} \right\rangle\right| \geq B_5 \sqrt{p \log m}\right) &\leq \mathbb{P}\left(\left|\left\langle w^{k,j}, v^{k,j} \right\rangle\right| \geq B_5 \sqrt{p \log m} \middle| F^{k,j}\right) + \mathbb{P}(F^{k,j^c}) \\ &\leq O(m^{-4}) \end{aligned}$$

and now taking a union bound over at most $n \leq m$ pairs (k, j) , we obtain the desired conclusion. ■

Proof of Lemma 18 Fix a cluster k such that $\frac{p-q}{2} s_k \geq B_4 \sqrt{pn \log m}$, and let w be the j th row of $W^{(k)}$ for some fixed $j \in [s_k]$. Note that $\left\|\frac{1}{\sqrt{s_k}} \mathbf{1}_{s_k}\right\|_2 = 1$ and $\left\|\frac{1}{\sqrt{s_k}} \mathbf{1}_{s_k}\right\|_\infty = \frac{1}{\sqrt{s_k}}$, so using Bernstein's inequality the same way as in the proof of Lemma 17 we obtain

$$\mathbb{P}\left(\left|\left\langle w, \frac{1}{\sqrt{s_k}} \mathbf{1}_{s_k} \right\rangle\right| \geq B_5 \sqrt{p \log m}\right) \leq O(m^{-5}).$$

Now taking a union bound over at most $n \leq m$ pairs (k, j) we obtain the desired conclusion. ■

Proof of Lemma 22 Fix a cluster k satisfying $\frac{p-q}{2} s_k \geq B_4 \sqrt{pn \log m}$ and fix $j \in [s_k]$. Technically, we have not assumed that we are in the event $F^{k,j} := E_1^{k,j} \cap \left\{\lambda_1\left(\overleftrightarrow{A}^{k,j}\right) > \lambda^k\right\}$, which is the event that the (k, j) th LOO relaxed oracle SDP (29) is guaranteed to have rank-one solution $\overline{Y}^{k,j} = \overline{y}^{k,j} \overline{y}^{k,j^\top}$ by Lemma 21. We simply extend the definition of $\overline{y}^{k,j}$ to be 0 outside the event $F^{k,j}$. Then the desired bound trivially holds on F^{k,j^c} , so we have that

$$\begin{aligned} \mathbb{P}\left(\left|\left\langle w^{k,j}, \frac{1}{\sqrt{s_k}} \overline{y}^{k,j} \right\rangle\right| \geq t\right) &= \mathbb{P}\left(\left|\left\langle w^{k,j}, \frac{1}{\sqrt{s_k}} \overline{y}^{k,j} \right\rangle\right| \geq t \middle| F^{k,j}\right) \mathbb{P}(F^{k,j}) \\ &\leq \mathbb{P}\left(\left|\left\langle w^{k,j}, \frac{1}{\sqrt{s_k}} \overline{y}^{k,j} \right\rangle\right| \geq t \middle| F^{k,j}\right). \end{aligned}$$

Since $\|\bar{y}^{k,j}\|_\infty \leq 1$ from Lemma 21, we have $\left\|\frac{1}{\sqrt{s_k}}\bar{y}^{k,j}\right\|_\infty \leq \frac{1}{\sqrt{s_k}}$ and $\left\|\frac{1}{\sqrt{s_k}}\bar{y}^{k,j}\right\|_2 \leq 1$. Since $\bar{y}^{k,j}$ and $F^{k,j}$ are independent of $w^{k,j}$, we can apply Bernstein's inequality to obtain that

$$\begin{aligned} \mathbb{P}\left(\left|\left\langle w^{k,j}, \frac{1}{\sqrt{s_k}}\bar{y}^{k,j} \right\rangle\right| \geq t\right) &\leq \mathbb{P}\left(\left|\left\langle w^{k,j}, \frac{1}{\sqrt{s_k}}\bar{y}^{k,j} \right\rangle\right| \geq t \mid F^{k,j}\right) \\ &\leq 2 \exp\left(\frac{-t^2/2}{p1 + \frac{t}{3}\frac{1}{\sqrt{s_k}}}\right) \\ &\leq O(m^{-5}) \end{aligned}$$

by taking $t = B_5\sqrt{p\log m}$ identically to the proof of Lemma 17. Now taking a union bound over all $\leq n \leq m$ pairs (k, j) , we obtain that with probability at least $1 - O(m^{-4})$, for all k such that $\frac{p-q}{2}s_k \geq B_4\sqrt{pn\log m}$ and all $j \in [s_k]$, $\left|\left\langle w^{k,j}, \frac{1}{\sqrt{s_k}}\bar{y}^{k,j} \right\rangle\right| \leq B_5\sqrt{p\log m}$. Now multiplying both sides by $\sqrt{s_k}$, we obtain the desired conclusion. $\blacksquare$

Proof of Lemma 31 We will use Bernstein's inequality. Let S be the set of all k satisfying $\frac{p-q}{2}s_k \geq B_4\sqrt{pn\log m}$. Fix $k \in S$, fix another cluster $j \neq k$, and let w be one row of $W^{(kj)}$. Note $w \in \mathbb{R}^{s_k}$, and all entries have mean zero, variance $q(1-q) \leq p$, and magnitude bounded by $\max\{q, 1-q\} \leq 1$.

By Lemma 30 that under the event E_3 we have $\|x^k\|_\infty \leq \frac{2}{\sqrt{s_k}}$ and also note that E_3 is independent of w (since E_3 only involves noise variables which are part of the diagonal blocks $(W^{(i)})_{i \in [K]}$). Therefore Bernstein's inequality gives that

$$\begin{aligned} \mathbb{P}\left(\left|w^\top x^k\right| \geq t \mid E_3\right) &\leq 2 \exp\left(-\frac{t^2/2}{p + \frac{2t}{3\sqrt{s_k}}}\right) \\ &\leq O(m^{-5}) \end{aligned}$$

by taking $t = B_5\sqrt{p\log m}$ in the same way as in the proof of Lemma 17. Taking a union bound over all clusters $j \neq k$ all rows ($\leq n \leq m$ in total), we have

$$\mathbb{P}\left(\sup_{j \neq k} \left\|W^{(jk)}x^k\right\|_\infty \geq B_5\sqrt{p\log m} \mid E_3\right) \leq O(m^{-4})$$

and taking a union bound over all $k \in S$

$$\mathbb{P}\left(\sup_{k \in S} \sup_{j \neq k} \left\|W^{(jk)}x^k\right\|_\infty \geq B_5\sqrt{p\log m} \mid E_3\right) \leq O(m^{-3}).$$

Therefore recalling that $\mathbb{P}(E_3^c) \leq O(m^{-3})$,

$$\begin{aligned} \mathbb{P}\left(\sup_{k \in S} \sup_{j \neq k} \left\|W^{(jk)}x^k\right\|_\infty \geq B_5\sqrt{p\log n}\right) &\leq \mathbb{P}\left(\sup_{k \in S} \sup_{j \neq k} \left\|W^{(jk)}x^k\right\|_\infty \geq B_5\sqrt{p\log n} \mid E_3\right) + \mathbb{P}(E_3^c) \\ &\leq O(m^{-3}). \end{aligned}$$

$\blacksquare$

Proof of Lemma 32 During the proof of Lemma 15 we already check that with probability at least $1 - O(m^{-8})$, we have $\|W\|_{\text{op}} \leq B_1\sqrt{pn}$. $\blacksquare$

Appendix D. Proofs for Recursive Clustering

Proof of Theorem 10 First, we note that assuming the recovery SDP subroutine always succeeds, since in Algorithm 1 we only remove recovered clusters which are so large that they are guaranteed to be all-one, the sequence of clusters which will be recovered is actually known a priori. (If we removed clusters which were recovered but are mid-size, then since their recovery is highly noise-dependent, we would not be able to predict which mid-size clusters would be recovered in a given round. We believe a more careful analysis would have no trouble with this added complexity, but since the ultimate conclusions of Theorem 10 would not change, we elect to only remove clusters which are guaranteed to be all-one.)

From an analysis standpoint, we have access to the list $(s_1, \dots, s_K)$ of all true cluster sizes. Let $Z_1 = (1, \dots, K)$ represent the original SBM instance. There is obviously an algorithm which, given Z_ℓ , either recursively constructs $Z_{\ell+1}$ by removing all clusters from Z_ℓ which meet the guaranteed all-ones recovery threshold of Z_ℓ (case 1 of Theorem 2), or if there are no such clusters, simply terminates. This will produce a sequence $Z_1, \dots, Z_r$ as well as the sequence of instance sizes $n_1, \dots, n_r$. Now for each fixed $\ell \in [r]$, we apply Theorem 2 with $m = n$ to guarantee that all clusters of size large enough to meet the all-ones recovery threshold $C'(p, q)\sqrt{n_\ell \log n}$ will be successfully recovered by the recovery SDP with failure probability at most $O(n^{-3})$. Note that although their sizes are initially unknown to a user, after being recovered their sizes will be known, allowing us to verify that they meet the all-ones recovery threshold and then remove them as specified in Algorithm 1. Taking a union bound over these $r \leq K \leq n$ rounds, we are guaranteed with probability at least $1 - O(n^{-2})$ that Algorithm 1 will recover all clusters of size at least $C'(p, q)\sqrt{n_\ell \log n}$ at each round ℓ .

Now to prove Theorem 10 we can simply analyze the sequence $Z_1, \dots, Z_r$. First, again the total number of rounds is $\leq K$ since we stop once no clusters are recovered in a given round.

Next, considering the round r where the algorithm terminates, since no clusters were recovered, this means that all clusters sizes in Z_r must be below $C'(p, q)\sqrt{n_r \log n}$. Since there are at most K total clusters in Z_r , this allows us to bound the total number of remaining nodes n_r as

$$\begin{aligned} n_r &\leq KC'(p, q)\sqrt{n_r \log n} \\ &\leq \frac{n^{\frac{1}{2}-\alpha}}{C'(p, q)\sqrt{\log n}} C'(p, q)\sqrt{n_r \log n} \\ &= n^{\frac{1}{2}-\alpha}\sqrt{n_r}. \end{aligned}$$

Rearranging, we obtain that $n_r \leq n^{1-2\alpha}$.

We prove the final part of the theorem by induction over rounds ℓ . First, after $\ell = 0$ rounds, there will be all n nodes remaining, matching the claim. Next, suppose inductively that after some fixed round ℓ we have $\leq n^{1-\alpha \sum_{i=0}^{\ell-1} 2^{-i}}$ unrecovered nodes remaining. Then the recovery threshold for the next round will be $\leq T = C'(p, q)\sqrt{\log n n^{\frac{1}{2}(1-\alpha \sum_{i=0}^{\ell-1} 2^{-i})}}$. The maximum number of nodes that

can be below this threshold during the next round is less than or equal to

$$\begin{aligned}
 KT &\leq \frac{n^{\frac{1}{2}-\alpha}}{C'(p, q)\sqrt{\log n}} C'(p, q) \sqrt{\log n} n^{\frac{1}{2}(1-\alpha \sum_{i=0}^{\ell-1} 2^{-i})} \\
 &= n^{\frac{1}{2}-\alpha} n^{\frac{1}{2}(1-\alpha \sum_{i=0}^{\ell-1} 2^{-i})} \\
 &= n^{\frac{1}{2}-\alpha} n^{\frac{1}{2}-\alpha \sum_{i=1}^{\ell} 2^{-i}} \\
 &= n^{1-\alpha \sum_{i=0}^{\ell} 2^{-i}}
 \end{aligned}$$

proving the desired claim. ■

Appendix E. Proofs for Clustering With a Faulty Oracle

In this section we prove Theorem 6 on faulty oracle clustering with a target recovery size of s , we prove our instance-adaptive query complexity guarantees in Theorem 7, and then finally we prove our Theorem 8 which achieves improved query complexity for the case that K is bounded.

First, we state several concentration inequalities which will be used in this section.

Theorem 42 (Hoeffding's Inequality (Hoeffding, 1994)) *Suppose $(X_i)_{i=1}^n$ are independent random variables such that $X_i \in [a_i, b_i]$ almost surely. Let $X = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}X$. Then for any $t \geq 0$,*

$$\mathbb{P}(|X - \mu| \geq t) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Theorem 43 (Chernoff's Inequality for Bernoulli random variables (Mitzenmacher and Upfal, 2005))

Suppose $(X_i)_{i=1}^n$ are independent random variables such that $X_i \in \{0, 1\}$ almost surely. Let $X = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}X$. Then

1. *For any $\delta \geq 0$, $\mathbb{P}(X \geq (1 + \delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{2 + \delta}\right)$.*
2. *For any $0 < \delta < 1$, $\mathbb{P}(X \leq (1 - \delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{2}\right)$.*
3. *For any $0 < \delta < 1$, $\mathbb{P}(|X - \mu| \geq \delta \mu) \leq 2 \exp\left(-\frac{\delta^2 \mu}{3}\right)$.*

Now before proving our theorems on faulty oracle clustering, we first start by fixing a subsample T and analyzing the majority voting procedure and the performance of the recovery SDP (1) on the subgraph on nodes T to prove two lemmas which will be useful for all subsequent proofs.

Lemma 44 *Let $T \subseteq [n]$ be fixed. There exists C_1 such that the following holds: With probability at least $1 - O(n^{-3})$, for each cluster $k \in [K]$ such that $|V_k \cap T| \geq \frac{C_1 \log n}{\delta^2}$, letting S'_k be the first $\frac{C_1 \log n}{\delta^2}$ nodes in $V_k \cap T$, for each node $i \notin T$, if we query (i, j) for all $j \in S'_k$, then:*

1. *If $i \in V_k$, then the majority of queries will be 1.*
2. *If $i \notin V_k$, then the majority of queries will be 0.*

Proof First we fix S'_k and i . For each $j \in S'_k$ let I_j be the outcome of the query (i, j) , and let $N = \sum_{j=1}^{|S'_k|} I_j$ be the sum of the queries. First, if $i \in V_k$, then I_j has distribution Bernoulli($\frac{1}{2} + \frac{\delta}{2}$), so then by Hoeffding's inequality

$$\begin{aligned} \mathbb{P}\left(N \leq \frac{|S'_k|}{2}\right) &= \mathbb{P}\left(N - \mathbb{E}N \leq -\frac{\delta|S'_k|}{2}\right) \\ &\leq \exp\left(\frac{-2\left(\frac{\delta|S'_k|}{2}\right)^2}{|S'_k|}\right) \\ &= \exp\left(-\frac{1}{2}\delta^2|S'_k|\right) \end{aligned}$$

which is $O(n^{-5})$ if $|S'_k| = \frac{C_1 \log n}{\delta^2}$ for sufficiently large C_1 (in fact $C_1 = 10$). For the other case that $i \notin V_k$, then I_j has distribution Bernoulli($\frac{1}{2} - \frac{\delta}{2}$), so again by Hoeffding's inequality

$$\begin{aligned} \mathbb{P}\left(N \geq \frac{|S'_k|}{2}\right) &= \mathbb{P}\left(N - \mathbb{E}N \geq \frac{\delta|S'_k|}{2}\right) \\ &\leq \exp\left(-\frac{1}{2}\delta^2|S'_k|\right) \end{aligned}$$

which is again $O(n^{-5})$.

Now taking a union bound over all $\leq n$ nodes $i \notin T$ and all $\leq n$ clusters k satisfying $|V_k \cap T| \geq \frac{C_1 \log n}{\delta^2}$, we obtain the desired conclusion with a failure probability of at most $O(n^{-3})$. $\blacksquare$

Lemma 45 *Let $T \subseteq [n]$ be fixed. There exists C such that with probability at least $1 - O(n^{-3})$, if we query all $i, j \in T$ to form an adjacency matrix A and apply the recovery SDP (1) with $m = n$, then*

1. *All clusters k such that $|T \cap V_k| \geq \frac{C\sqrt{|T| \log n}}{\delta}$ will be recovered.*
2. *No clusters k such that $|T \cap V_k| \leq \frac{C\sqrt{|T| \log n}}{3\delta}$ will be recovered (and the blocks of the recovery SDP solution will be zero).*

Proof This follows immediately from Theorem 2, once we note that by querying all $i, j \in T$, this is equivalent to an SBM instance with $p = \frac{1+\delta}{2}$ and $q = \frac{1-\delta}{2}$, and so $p - q = \delta$. $\blacksquare$

With Lemmas 44 and 45 we can now easily prove Theorem 6.

Proof of Theorem 6 Now we can simply unfix T and find a high-probability event under which Algorithm 2 succeeds.

Lemma 46 *There exists C_2 such that for any $s \geq C_2 \frac{\sqrt{n \log n}}{\delta}$, if we take $\gamma = C_2 \frac{n \log n}{s^2 \delta^2}$, then with probability at least $1 - O(n^{-3})$:*

1. $\frac{7}{8}\gamma n \leq |T| \leq \frac{9}{8}\gamma n$.
2. *For all clusters k such that $s_k \geq s$, $\frac{7}{8}\gamma s_k \leq |T \cap V_k| \leq \frac{9}{8}\gamma s_k$.*

3. For all clusters k such that $s_k \geq s$, $|T \cap V_k| \geq \frac{C\sqrt{|T|\log n}}{\delta}$.

Proof We let I_i be an indicator for the event that node i is included in T . Then I_i is distributed as $\text{Bernoulli}(\gamma)$ and $|T| = \sum_{i=1}^n I_i$. Then by Chernoff's inequality,

$$\mathbb{P}\left(|T| - \gamma n \geq \left(1 + \frac{1}{8}\right) \gamma n\right) \leq \exp\left(-\frac{\gamma n}{8^2 \cdot 3}\right).$$

Now since $\frac{\gamma n}{8^2 \cdot 3} \geq \frac{C_2}{8^2 \cdot 3} \frac{n^2 \log n}{s^2 \delta^2} \geq \frac{C_2}{8^2 \cdot 3} \frac{\log n}{\delta^2} \geq 5 \log n$ for sufficiently large C_2 , we have that $\frac{7}{8} \gamma n \leq |T| \leq \frac{9}{8} \gamma n$ with probability at least $1 - O(n^{-5})$. An identical argument holds for all clusters k such that $s_k \geq s$ (except we will use $\frac{s_k}{s} \geq 1$ in place of $\frac{n}{s} \geq 1$). Union bounding over all these events, we have a failure probability which is less than $O(n^{-3})$.

Lastly, under these events, for each k such that $s_k \geq s$ we have

$$|T \cap V_k| \geq \frac{7}{8} \gamma s_k \geq \frac{7}{8} \frac{s_k}{s} \frac{\sqrt{C_2 \gamma n \log n}}{\delta} \geq \frac{7}{8} \frac{\sqrt{C_2 \gamma n \log n}}{\delta} \geq \frac{C \sqrt{\frac{9}{8} \gamma n \log n}}{\delta} \geq \frac{C \sqrt{|T| \log n}}{\delta}$$

where the second-last inequality holds as long as C_2 is sufficiently large. $\blacksquare$

Now we can complete the proof of Theorem 6. For any fixed $T \subseteq [n]$ which satisfies the conclusions of Lemma 46, from Lemma 45 we have that all clusters k of size at least s will have $V_k \cap T$ correctly recovered by the recovery SDP with failure probability $O(n^{-3})$. Then by taking the first $\frac{C_1 \log n}{\delta^2}$ nodes from each recovered (sub)clusters, Lemma 44 guarantees that with failure probability $O(n^{-3})$ we will successfully recover each entire cluster. Note that Lemma 46 guarantees all clusters k with $s_k \geq s$ will have $|T \cap V_k| \geq \frac{C_1 \log n}{\delta^2}$, since we have

$$|T \cap V_k| \geq \frac{7}{8} \gamma n = \frac{7}{8} C_2 \frac{n^2 \log n}{s^2 \delta^2} \geq \frac{7}{8} C_2 \frac{\log n}{\delta^2} \geq C_1 \frac{\log n}{\delta^2}.$$

Now unfixing T , the event in Lemma 46 occurs with probability at least $1 - O(n^{-3})$, giving an overall failure probability of $O(n^{-3})$.

Finally, to check the sample complexity, we first note that by Lemma 46, we have

$$|T| \leq \frac{9}{8} \gamma n = \frac{9}{8} C_2 \frac{n^2 \log n}{s^2 \delta^2}$$

so querying all pairs in T requires $O(|T|^2) \leq O\left(\frac{n^4 \log^2 n}{s^4 \delta^4}\right)$ queries. Next, to bound the number of queries used for the majority voting phase, each subcluster of T which gets recovered will lead to at most $n \cdot \frac{C_1 \log n}{\delta^2} \leq O\left(\frac{n \log n}{\delta^2}\right)$ queries, so it remains to bound the number of subclusters contained in T that are recovered. From Lemma 45 we know that for the subcluster of cluster k to be recovered we need $|T \cap V_k| \geq \frac{C \sqrt{|T| \log n}}{3\delta}$. Therefore under the event from Lemma 46 the maximum number of such clusters is

$$\frac{|T|}{\frac{C \sqrt{|T| \log n}}{3\delta}} \leq O\left(\frac{\delta \sqrt{\gamma n}}{\sqrt{\log n}}\right) \leq O\left(\frac{n}{s}\right).$$

Therefore the majority voting procedure requires $O\left(\frac{n^2 \log n}{s\delta^2}\right)$ queries, leading to the claimed total query complexity. $\blacksquare$

Next we present the proof of the guarantees for our instance-adaptive clustering algorithm.

Proof of Theorem 7 We reuse the same value of C_1 as from Theorem 6, and we require C_2 to be at least as large as required in Theorem 6. Let $\underline{r}$ be the smallest r such that $\hat{s}_r = \frac{n}{2^{r-1}} \geq s_1$. Now we can outline our proof strategy. With this choice of $\underline{r}$, we are essentially guaranteed by Theorem 6 and the form of $\gamma_{\underline{r}}$ that cluster 1 would be recovered on round $\underline{r}$ (if that round is reached). This immediately implies the claimed query complexity since the algorithm will terminate no later than round $\underline{r}$. Then we will ensure that no clusters which are significantly smaller than s_1 will be recovered during any of the rounds before $\underline{r}$ (inclusive).

Our first step will be to analyze the subsampling procedure. We define T_r for all $r = 1, \dots, \underline{r}$ as the subsample of $[n]$ which would be sampled if we reached round r .

Lemma 47 *There exist absolute constants C_2, C_3 such that if we take $\gamma_r = C_2 \frac{n \log n}{\hat{s}_r^2 \delta^2}$ for each $r \in [\underline{r}]$, then with probability at least $1 - O(n^{-2})$:*

1. For all $r \in [\underline{r}]$, $\frac{7}{8}\gamma_r n \leq |T_r| \leq \frac{9}{8}\gamma_r n$.
2. For all $r \in [\underline{r}]$ and for all k such that $s_k \leq \frac{s_1}{C_3}$, we have $|T_r \cap V_k| \leq \frac{C\sqrt{|T_r| \log n}}{3\delta}$.
3. We have $|T_{\underline{r}} \cap V_1| \geq \frac{C\sqrt{|T_{\underline{r}}| \log n}}{\delta}$.

Proof For each fixed r , $\frac{7}{8}\gamma_r n \leq |T_r| \leq \frac{9}{8}\gamma_r n$ follows by applying Lemma 46, so then we can obtain the first item by applying a union bound over all $\underline{r} \leq n$ rounds.

The third item also follows from the third item of Lemma 46 (the conditions of which hold for round $\underline{r}$ and for cluster 1 by the definition of $\underline{r}$).

Now we check the second item. First we fix a cluster k such that $s_k \leq \frac{s_1}{C_3}$ and a round $r \leq \underline{r}$. We know from the first item that $|T_r| \geq \frac{7}{8}\gamma_r n$, so it suffices to show that $|T_r \cap V_k| \leq \sqrt{\frac{7}{8} \frac{C\sqrt{\gamma_r n \log n}}{3\delta}}$. Now for each $i \in V_k$ we let I_i be an indicator for the event that node i is included in T_r and note that the I_i are independent, they have distribution Bernoulli(γ_r), and their sum is $|T_r \cap V_k|$. We calculate that

$$\begin{aligned} \mathbb{P}\left(|T_r \cap V_k| \geq \sqrt{\frac{7}{8} \frac{C\sqrt{\gamma_r n \log n}}{3\delta}}\right) &= \mathbb{P}\left(|T_r \cap V_k| - \mathbb{E}|T_r \cap V_k| \geq \sqrt{\frac{7}{8} \frac{C\sqrt{\gamma_r n \log n}}{3\delta}} - \gamma_r s_k\right) \\ &= \mathbb{P}\left(|T_r \cap V_k| - \mathbb{E}|T_r \cap V_k| \geq \left(\sqrt{\frac{7}{8} \frac{C\sqrt{n \log n}}{3\delta\sqrt{\gamma_r s_k}}} - 1\right) \gamma_r s_k\right) \\ &= \mathbb{P}\left(|T_r \cap V_k| - \mathbb{E}|T_r \cap V_k| \geq \left(C' \frac{\hat{s}_r}{s_k} - 1\right) \gamma_r s_k\right) \end{aligned}$$

where we use the fact that by definition of γ_r ,

$$\sqrt{\frac{7}{8} \frac{C\sqrt{n \log n}}{3\delta\sqrt{\gamma_r s_k}}} = \sqrt{\frac{7}{8} \frac{C\sqrt{n \log n}}{3\delta s_k}} \frac{\hat{s}_r \delta}{\sqrt{C_2 n \log n}} = \frac{\sqrt{\frac{7}{8} C}}{3\sqrt{C_2}} \frac{\hat{s}_r}{s_k} = C' \frac{\hat{s}_r}{s_k}$$

and the last equality is a definition. Therefore by Chernoff's inequality

$$\mathbb{P} \left(|T_r \cap V_k| \geq \sqrt{\frac{7}{8}} \frac{C \sqrt{\gamma_r n \log n}}{3\delta} \right) \leq \exp \left(- \frac{\left(C' \frac{\hat{s}_r}{s_k} - 1 \right)^2 \gamma_r s_k}{1 + C' \frac{\hat{s}_r}{s_k}} \right). \quad (51)$$

Note that since $r \leq \underline{r}$ and $\frac{s_1}{C_3} \geq s_k$, by definition of $\underline{r}$ we have that

$$\hat{s}_r \geq \hat{s}_{\underline{r}} = \frac{1}{2} \hat{s}_{\underline{r}-1} > \frac{1}{2} s_1 \geq \frac{C_3}{2} s_k.$$

Also if x is sufficiently large ($x \geq 5$ works) then we will have

$$\frac{(x-1)^2}{1+x} = \frac{x^2 - 2x + 1}{1+x} \geq \frac{x}{2}.$$

Using this in (51) by requiring C_3 to be sufficiently large so that $C' \frac{\hat{s}_r}{s_k} \geq 5$, we have

$$\begin{aligned} \mathbb{P} \left(|T_r \cap V_k| \geq \sqrt{\frac{7}{8}} \frac{C \sqrt{\gamma_r n \log n}}{3\delta} \right) &\leq \exp \left(- \frac{C' \frac{\hat{s}_r}{s_k}}{2} \gamma_r s_k \right) \\ &= \exp \left(- \frac{C'}{2} \hat{s}_r \gamma_r \right) \\ &= \exp \left(- \frac{C'}{2} C_2 \frac{n \log n}{\hat{s}_r \delta^2} \right) \\ &\leq \exp \left(- \frac{C'}{2} C_2 \log n \right) \\ &= \exp \left(- \frac{1}{2} \frac{\sqrt{\frac{7}{8}} C}{3 \sqrt{C_2}} C_2 \log n \right) \\ &\leq O(n^{-5}) \end{aligned}$$

where we obtain the final inequality by requiring C_2 sufficiently large. Now we can conclude by taking a union bound over all $\leq n$ clusters k such that $\frac{s_1}{C_3} \geq s_k$ and all $\leq n$ rounds $r \leq \underline{r}$. $\blacksquare$

Now, for each round $r \in [\underline{r}]$, we can apply Lemma 45 to ensure the success of the recovery SDP when applied to the subsample T_r , and we can apply Lemma 44 to ensure the success of the majority voting procedure (if applied to a recovered subcluster). Each of these lemmas has failure probability $O(n^{-3})$ and we apply each $\underline{r} \leq n$ times, giving the claimed failure probability by a union bound.

The third item of Lemma 47 combined with Lemma 45 guarantees that if the algorithm has not terminated by round $\underline{r}$, then a subcluster from cluster 1 will be recovered on this round (as well as possibly other subclusters) so the algorithm will terminate by round $\underline{r}$. Therefore the maximum query complexity due to querying all pairs within subsamples T_r (up until the first round a subcluster is

recovered) is

$$\begin{aligned}
O(|T_1|^2 + \dots + |T_{\underline{r}}|^2) &\leq O\left(\left(\frac{9}{8}\gamma_1 n\right)^2 + \dots + \left(\frac{9}{8}\gamma_{\underline{r}} n\right)^2\right) \\
&\leq O\left(\left(\frac{9}{8}(2^{1-1})^2 \frac{\log n}{\delta^2}\right)^2 + \dots + \left(\frac{9}{8}(2^{r-1})^2 \frac{\log n}{\delta^2}\right)^2\right) \\
&\leq O\left(\frac{n^4 \log^2 n}{s_1^4 \delta^4}\right)
\end{aligned}$$

using the facts that $\gamma_r n = \frac{n^2 \log n}{\hat{s}_r^2 \delta^2} = (2^{r-1})^2 \frac{\log n}{\delta^2}$, that $2^{r-1} \leq \frac{n}{s_1}$, and that a geometric series is bounded by a multiple of its largest term. Adding in the $O\left(\frac{n \log n}{\delta^2}\right)$ sample complexity for the majority voting procedure (since we only apply it to find one cluster), we obtain the claimed query complexity.

Lastly, the second item of Lemma 47 combined with Lemma 45 guarantees that no subcluster from a cluster k such that $s_k \leq \frac{s_1}{C_3}$ will be recovered at any round up to and including round $\underline{r}$, so we are guaranteed that whenever the first subcluster is recovered, it must be from a cluster k such that $s_k \geq \frac{s_1}{C_3}$. Furthermore from our earlier application of Lemma 44 the entirety of this cluster k will be correctly recovered. $\blacksquare$

Finally we present the proof of the guarantees for our algorithm for faulty oracle clustering when K is bounded.

Proof of Theorem 8 We face the same technical issue as in the proof of Theorem 10 on recursive clustering, which concisely is that we would like to apply Theorem 2 to guarantee success of the recovery SDP (1) when given each constructed adjacency matrix A_r as input, for each round r . This would simply follow from a union bound over all rounds if we knew which clusters would be present at each round. However, since there are mid-size clusters whose recovery is random, we face a combinatorial explosion of lists of clusters which could still be remaining at each round, and therefore we cannot guarantee the success of the recovery SDP for each option at each round via a union bound. Again, we believe that this could ideally be overcome by very carefully tracking each random event required for the success of the recovery SDP for each possible instance of remaining clusters and noting that the same events should guarantee success of the recovery SDP for many different instances.

Instead, we devise the following technical workaround which involves modifying Algorithm 4: Each round r , for any subcluster recovered by the recovery SDP, the majority voting procedure will (w.h.p.) correctly identify all nodes in the cluster, making the size of the cluster known. If the size of the cluster is at least $\frac{n_r}{K_r}$, we remove it as usual. However, if we recover a cluster smaller than $\frac{n_r}{K_r}$, we do not remove its nodes, but we remember the set S of nodes. This way, we are guaranteed that each round r , only the clusters of size at least $\frac{n_r}{K_r}$ are removed, making the list of clusters which remain at each round deterministic (under the good event). If, during a later round r' , we again recover a subcluster of S , we do not apply the majority voting procedure again, and we will remove the nodes S if and only if we now have $|S| \geq \frac{n_{r'}}{K_{r'}}$. This ensures that each cluster will only have the majority voting procedure applied to it once, ensuring that our workaround does not increase the query complexity.

Now we analyze this modified algorithm. First, similarly to the proof of Theorem 10, we let $F_1 = [K]$ and given F_r , we recursively construct $F_{r+1} = \{k \in F_r : s_k < \frac{\sum_{i \in F_r} s_i}{|F_r|}\}$. This choice is so that (under the good event) F_r will represent all clusters which have not been removed at the start of round r , and then F_{r+1} is obtained by removing all clusters which have size at least $(\# \text{ nodes remaining in } F_r) / (\# \text{ clusters remaining in } F_r)$.

Now we define a good event for the subsamples $T_1, \dots, T_K$ (which may not all be reached by the algorithm). First, for convenience, we let $\hat{n}_r = \sum_{i \in F_r} s_i$ denote the number of nodes remaining in the clusters represented by F_r and we let $\hat{K}_r = |F_r|$ denote the number of clusters represented by F_r . For each r , we sample $\hat{T}_r$ from the nodes which are members of the clusters included in F_r . We differentiate these notationally from the n_r, K_r, T_r which are defined over the course of the algorithm, but under the good event below, $\hat{n}_r, \hat{K}_r$ will be identical to n_r, K_r and $\hat{T}_r$ will have the same distribution as T_r .

Lemma 48 *There exists an absolute constant C_2 such that, taking $\gamma_r = C_2 \frac{\hat{K}_r^2 \log n}{\hat{n}_r \delta^2}$, with probability at least $1 - O(n^{-2})$:*

1. For each $r \in [K]$, we have $\frac{7}{8} \gamma_r \hat{n}_r \leq \hat{T}_r \leq \frac{9}{8} \gamma_r \hat{n}_r$.
2. For each cluster k , letting r be the first round such that $s_k \geq \frac{\hat{n}_r}{\hat{K}_r}$ (if such a round exists), we have $\frac{7}{8} \gamma_r s_k \leq |\hat{T}_r \cap V_k| \leq \frac{9}{8} \gamma_r s_k$.
3. For each cluster k , letting r be the first round such that $s_k \geq \frac{\hat{n}_r}{\hat{K}_r}$ (if such a round exists), we have $|\hat{T}_r \cap V_k| \geq \frac{C \sqrt{|\hat{T}_r| \log n}}{\delta}$.

Proof The first item is very similar to the first item of Lemma 46 except we use a different choice subsampling ratio γ_r . Fix a round r . For each node i which is in a cluster $k \in F_r$, let I_i be an indicator for the event that node i is included in $\hat{T}_r$. Then I_i is distributed as Bernoulli(γ_r) and $|\hat{T}_r| = \sum_{i=1}^{\hat{n}_r} I_i$. Then by Chernoff's inequality,

$$\begin{aligned} \mathbb{P} \left(\left| |\hat{T}_r| - \gamma_r \hat{n}_r \right| \geq \left(1 + \frac{1}{8} \right) \gamma_r \hat{n}_r \right) &\leq \exp \left(-\frac{\gamma_r \hat{n}_r}{8^2 \cdot 3} \right) \\ &= \exp \left(-\frac{C_2}{8^2 \cdot 3} \frac{\hat{K}_r^2 \log n}{\delta^2} \right) \\ &\leq \exp \left(-\frac{C_2}{8^2 \cdot 3} \log n \right) \\ &\leq O(n^{-3}) \end{aligned}$$

where the final inequality is for sufficiently large C_2 . Now unfixing r and applying a union bound over all $\leq K \leq n$ rounds, we obtain the first item.

For the second item, fix a cluster k and let r be the first round such that $s_k \geq \frac{\hat{n}_r}{\hat{K}_r}$. By an identical argument to the previous step we have

$$\begin{aligned}
\mathbb{P}\left(\left|\hat{T}_r \cap V_k\right| - \gamma_r s_k\right| &\geq \left(1 + \frac{1}{8}\right) \gamma_r s_k\right) \leq \exp\left(-\frac{\gamma_r s_k}{8^2 \cdot 3}\right) \\
&= \exp\left(-\frac{C_2}{8^2 \cdot 3} \frac{\hat{K}_r^2 \log n}{\hat{n}_r \delta^2} s_k\right) \\
&\leq \exp\left(-\frac{C_2}{8^2 \cdot 3} \frac{\hat{K}_r \log n}{\delta^2}\right) \\
&\leq \exp\left(-\frac{C_2}{8^2 \cdot 3} \log n\right) \\
&\leq O(n^{-3})
\end{aligned}$$

using the fact that $s_k \geq \frac{\hat{n}_r}{\hat{K}_r}$. Now we obtain the second item by taking a union bound over all $K \leq n$ clusters.

The third item follows from the previous two. For each cluster k , letting r be the first round such that $s_k \geq \frac{\hat{n}_r}{\hat{K}_r}$ (if such a round exists), we have

$$\begin{aligned}
|\hat{T}_r \cap V_k| &\geq \frac{7}{8} \gamma_r s_k \geq \frac{7\sqrt{C_2}}{8} \frac{\hat{K}_r \sqrt{\log n}}{\delta \sqrt{\hat{n}_r}} \sqrt{\gamma_r s_k} \geq \frac{7\sqrt{C_2}}{8} \frac{\sqrt{\hat{n}_r \log n}}{\delta} \sqrt{\gamma_r} \\
&\geq \frac{C \sqrt{\frac{9}{8} \gamma_r \hat{n}_r \log n}}{\delta} \geq \frac{C \sqrt{|\hat{T}_r| \log n}}{\delta}
\end{aligned}$$

where the second line is for sufficiently large C_2 . ■

Now, for each r , we apply Lemma 45 to guarantee the success of the recovery SDP when applied to the adjacency matrix obtained from the subsample $\hat{T}_r$ with probability at least $1 - O(n^{-3})$, and combining this with the third item of the previous Lemma 48, this guarantees that when the nodes are those belonging to the clusters of F_r , we will recover subclusters from each cluster k such that $s_k \geq \frac{\hat{n}_r}{\hat{K}_r}$. Now applying Lemma 44 we are guaranteed that the majority voting procedure will correctly recover the remaining nodes from each such subcluster, because Lemma 48 guarantees that we will have

$$|\hat{T}_r \cap V_k| \geq \frac{7}{8} \gamma_r s_k = \frac{7}{8} C_2 \frac{\hat{K}_r^2 \log n}{\hat{n}_r \delta^2} s_k \geq \frac{7}{8} C_2 \frac{\hat{K}_r \log n}{\delta^2} \geq \frac{C_1 \log n}{\delta^2}.$$

The (modified) algorithm therefore ensures that all nodes belonging to clusters such that $s_k \geq \frac{\hat{n}_r}{\hat{K}_r}$ would be recovered and removed.

Now since this holds for any round r , and we start with $\hat{n}_1 = n_1 = n$ and $F_1 = [K]$, this guarantees that T_1 has the same distribution as $\hat{T}_1$, and then the above argument shows that with probability at least $1 - O(n^{-2})$, for each round r , we have $n_r = \hat{n}_r$, $K_r = \hat{K}_r$, and the clusters k such that $s_k \geq \frac{n_r}{K_r} = \frac{\hat{n}_r}{\hat{K}_r}$ are recovered and removed (which is at least one cluster), leaving exactly the clusters remaining in F_{r+1} .

Now it remains to check the query complexity and establish the recovery threshold. The majority voting procedure requires $O(\frac{n \log n}{\delta^2})$ queries per subcluster and gets applied at most K times, giving a term of $O(\frac{nK \log n}{\delta^2})$. Under the aforementioned events, the subsampling procedure for the r th round makes

$$O(|T_r|^2) = O(\gamma_r^2 n_r^2) = O\left(\frac{K_r^4 \log^2 n}{\delta^4}\right) \leq O\left(\frac{K^4 \log^2 n}{\delta^4}\right)$$

queries, and there are obviously at most K rounds, leading to a total query complexity of

$$O\left(\frac{nK \log n}{\delta^2} + \frac{K^5 \log^2 n}{\delta^4}\right).$$

To establish the recovery threshold, note that the algorithm terminates at the first round r such that $n_r \leq C_2 \frac{K_r^2 \log n}{\delta^2}$, implying that each remaining cluster is smaller than this amount and thus also smaller than $C_2 \frac{K^2 \log n}{\delta^2}$. $\blacksquare$

Appendix F. Semirandom Recovery Proof

Proof of Theorem 4 Recall that A' is the adjacency matrix sampled from the standard SBM, and the observed A is A' subject to the semirandom perturbations. For purposes of analysis, we can apply our Theorem 2 to the SBM problem with adjacency matrix A' . Let E be the event that the conclusions of Theorem 2 hold for the solution $\hat{Y}'$ of the analogous recovery SDP (1) but with A' in place of A (it is shown in Theorem 2 that $\mathbb{P}(E) \geq 1 - O(n^3)$). Under the event E , we have that $\hat{Y}'$ is the unique solution to the recovery SDP, meaning for all feasible $Y \neq \hat{Y}'$ we have

$$\left\langle \hat{Y}', A' - \frac{p+q}{2} J \right\rangle - \lambda \text{Tr}(\hat{Y}') > \left\langle Y, A' - \frac{p+q}{2} J \right\rangle - \lambda \text{Tr}(Y). \quad (52)$$

Next we argue that also for all feasible Y , we have

$$\left\langle \hat{Y}', A - A' \right\rangle \geq \left\langle Y, A - A' \right\rangle. \quad (53)$$

We show this in an elementwise fashion by using the form of $\hat{Y}'$ and the definition of the large cluster semirandom model. If $A_{ij} > A'_{ij}$, then i, j must be two nodes within a cluster k such that $\frac{p-q}{2} s_k \geq \frac{3}{2} B \sqrt{pn \log n} \geq \left(1 + \frac{1}{B}\right) \sqrt{pn \log n}$, using the fact that $\kappa = B$ and $\bar{B} > 2$ so $\left(1 + \frac{1}{B}\right) \leq \frac{3}{2}$. Therefore the conclusion of Theorem 2, which hold under the event E , guarantees that $\hat{Y}'_{ij} = 1$. By feasibility of Y , we have $Y_{ij} \leq 1$, and therefore $\hat{Y}'_{ij}(A_{ij} - A'_{ij}) \geq Y_{ij}(A_{ij} - A'_{ij})$. If $A_{ij} < A'_{ij}$, then i, j belong to two different clusters, and under the event E we have $\hat{Y}'_{ij} = 0$. Since $Y_{ij} \geq 0$ by feasibility of Y , we again have that $\hat{Y}'_{ij}(A_{ij} - A'_{ij}) \geq Y_{ij}(A_{ij} - A'_{ij})$. Finally, this same inequality obviously holds if $A_{ij} = A'_{ij}$. Summing over i, j we obtain the inequality (53).

Finally, by adding inequalities (52) and (53) and cancelling A' , we obtain that

$$\left\langle \hat{Y}', A - \frac{p+q}{2} J \right\rangle - \lambda \text{Tr}(\hat{Y}') > \left\langle Y, A - \frac{p+q}{2} J \right\rangle - \lambda \text{Tr}(Y).$$

Therefore we have shown that $\widehat{Y}'$ is also the unique optimal solution to the recovery SDP (1) (which uses A , not A'), meaning that $\widehat{Y} = \widehat{Y}'$ under the event E . Therefore we have shown that all conclusions of Theorem 2 hold under the event E , and thus Theorem 2 still holds under the large cluster semirandom model, as desired. $\blacksquare$

Appendix G. Eigenvalue Perturbation Bound Proofs

Proof of Theorem 12 We begin in the same way as the argument from (Eldridge et al., 2018, Theorem 6). Define $S_{t:T}$ as the set of all vectors in $\text{Col}(U_{t:T})$ with norm 1, and likewise for $S_{T+1:n}$ and $S_{t:n}$. Then by the Courant-Fischer min-max principle we have that

$$\lambda_t(M + H) \leq \max_{x \in S_{t:n}} x^\top (M + H) x.$$

Now we can write any $x \in S_{t:n}$ as $x = \alpha u + \beta u_\perp$, where $u \in S_{t:T}$, $u_\perp \in S_{T+1:n}$ and $\alpha^2 + \beta^2 = 1$. Therefore the above is equivalent to

$$\lambda_t(M + H) \leq \max_{\alpha, \beta, \alpha^2 + \beta^2 = 1} \max_{u \in S_{t:T}} \max_{u_\perp \in S_{T+1:n}} \left\{ \begin{aligned} &\alpha^2 u^\top M u + \alpha^2 u^\top H u + 2\alpha\beta u^\top M u_\perp + 2\alpha\beta u^\top H u_\perp + \beta^2 u_\perp^\top M u_\perp + \beta^2 u_\perp^\top H u_\perp \end{aligned} \right\}.$$

Now we begin to deviate slightly from (Eldridge et al., 2018, Theorem 6). We have $u^\top M u \leq \lambda_t$, $u^\top H u \leq h$ by assumption, $u^\top M u_\perp = 0$, $u^\top H u_\perp \leq \|H U_{t:T}\|_{\text{op}}$, $u_\perp^\top M u_\perp \leq \lambda_{T+1}$, and $u_\perp^\top H u_\perp \leq \|H\|_{\text{op}}$. Thus

$$\begin{aligned} \lambda_t(M + H) &\leq \max_{\alpha, \beta, \alpha^2 + \beta^2 = 1} \alpha^2 \lambda_t + \alpha^2 h + \alpha\beta \|H U_{t:T}\|_{\text{op}} + \beta^2 \lambda_{T+1} + \beta^2 \|H\|_{\text{op}} \\ &= \max_{\beta \in [0, 1]} (1 - \beta^2)(\lambda_t + h) + 2(1 - \beta)\beta \|H U_{t:T}\|_{\text{op}} + \beta^2(\lambda_{T+1} + \|H\|_{\text{op}}) \\ &= \max_{\beta \in [0, 1]} \lambda_t + h + \beta^2(\lambda_{T+1} - \lambda_t + \|H\|_{\text{op}} - h - 2\|H U_{t:T}\|_{\text{op}}) + 2\beta\|H U_{t:T}\|_{\text{op}}. \end{aligned}$$

Now by our assumption the leading coefficient $(\lambda_{T+1} - \lambda_t + \|H\|_{\text{op}} - h - 2\|H U_{t:T}\|_{\text{op}})$ is < 0 , so the quadratic has unique maximum over $\mathbb{R}$.

Using the elementary formula for the maximum of a quadratic, we have that

$$\begin{aligned} \lambda_t(M + H) &\leq \lambda_t + h + \frac{-4\|H U_{t:T}\|_{\text{op}}^2}{4(\lambda_{T+1} - \lambda_t + \|H\|_{\text{op}} - h - 2\|H U_{t:T}\|_{\text{op}})} \\ &= \lambda_t + h + \frac{\|H U_{t:T}\|_{\text{op}}^2}{\lambda_t(M) - \lambda_{T+1}(M) - \|H\|_{\text{op}} + h + 2\|H U_{t:T}\|_{\text{op}}}. \end{aligned}$$

$\blacksquare$