Journal of Machine Learning Research 16 (2015) 455-490 Submitted 3/14; Revised 8/14; Published 3/15

Iterative and Active Graph Clustering Using Trace Norm Minimization Without Cluster Size Constraints*

Nir Ailon NAILON@CS.TECHNION.AC.IL
Department of Computer Science
Technion IIT Haifa, Israel

Yudong Chen YUDONG.CHEN@EECS.BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA 94720, USA

Huan Xu MPEXUH@NUS.EDU.SG
Department of Mechanical Engineering
National University of Singapore
Singapore 117575

Editor: Tong Zhang

Abstract

This paper investigates graph clustering under the planted partition model in the presence of *small clusters*. Traditional results dictate that for an algorithm to provably correctly recover the underlying clusters, all clusters must be sufficiently large—in particular, the cluster sizes need to be $\tilde{\Omega}(\sqrt{n})$, where n is the number of nodes of the graph. We show that this is not really a restriction: by a refined analysis of a convex-optimization-based recovery approach, we prove that small clusters, under certain mild assumptions, do not hinder recovery of large ones. Based on this result, we further devise an iterative algorithm to provably recover *almost all clusters* via a "peeling strategy": we recover large clusters first, leading to a reduced problem, and repeat this procedure. These results are extended to the partial observation setting, in which only a (chosen) part of the graph is observed. The peeling strategy gives rise to an *active* learning algorithm, in which edges adjacent to smaller clusters are queried more often after large clusters are learned (and removed). We expect that the idea of iterative peeling—that is, sequentially identifying a subset of the clusters and reducing the problem to a smaller one—is useful more broadly beyond the specific implementations (based on convex optimization) used in this paper.

Keywords: graph clustering, community detection, active clustering, convex optimization, planted partition model, stochastic block model

1. Introduction

This paper considers the following classic graph clustering problem: given an undirected unweighted graph, partition the nodes into disjoint clusters so that the density of edges within each cluster is higher than those across clusters. Graph clustering arises naturally in many applications across science and engineering; prominent examples include community

*. This work extends and improves a preliminary conference version Ailon et al. (2013).

detection in social networks (Mishra et al., 2007; Zhao et al., 2011), submarket identification in E-commerce and sponsored search (Yahoo!-Inc, 2009), and co-authorship analysis in document database (Ester et al., 1995), among others. From a purely binary classification theoretical point of view, the edges of the graph are (noisy) labels of "similarity" or "affinity" between pairs of objects, and the concept class consists of clusterings of the objects (encoded graphically by identifying clusters with cliques).

Many theoretical results in graph clustering consider the *Planted Partition Model* (Condon and Karp, 2001), in which the edges are generated randomly based on an unknown set of underlying clusters; see Section 1.1 for more details. While numerous different methods have been proposed, their performance guarantees under the planted partition model generally have the following form: under certain conditions of the density of edges (within clusters and across clusters), the method succeeds to recover the correct clusters exactly *if all clusters are larger than a threshold size*, typically $\tilde{\Omega}(\sqrt{n})$;[1] see e.g., McSherry (2001); Bollobás and Scott (2004); Ames and Vavasis (2011); Chen et al. (2012); Chaudhuri et al. (2012); Anandkumar et al. (2014).

In this paper, we aim to relax this cluster size constraint of graph clustering under the planted partition model. Identifying extremely small clusters is inherently hard as they are easily confused with "fake" clusters generated by noisy edges,[2] and is not the focus of this paper. Instead, in this paper we investigate a question that has not been addressed before: Can we still recover large clusters in the presence of small clusters? Intuitively, this should be doable. To illustrate, consider an extreme example where the given graph G consists of two subgraphs G_1 and G_2 with disjoint node sets. Suppose G_1, if presented alone, can be correctly clustered using some existing methods, G_2 is a very small clique, and there are relatively few edges connecting G_1 and G_2. The graph G certainly violates the minimum cluster size requirement of previous results, but why should G_2 spoil our ability to correctly cluster G_1?

Our main result confirms this intuition. We show that the cluster size barrier arising in previous work is not really a restriction, but rather an artifact of the attempt to solve the problem in a single shot and recover large and small clusters simultaneously. Using a more careful analysis, we prove that a mixed trace-norm and ℓ_1-norm based convex formulation can recover clusters of size $\tilde{\Omega}(\sqrt{n})$ even in the presence of smaller clusters. That is, small clusters do not interfere with recovery of the large clusters.

The main implication of this result is that one can apply an *iterative* "peeling" strategy to recover smaller and smaller clusters. The intuition is simple: suppose the *number* of clusters is limited, then either all clusters are large, or the sizes of the clusters vary significantly. The first case is obviously easy. But the second is also tractable, for a different reason: using the aforementioned convex formulation, the larger clusters can be correctly identified; if we remove all nodes from these larger clusters, the remaining subgraph contains significantly fewer nodes than the original graph, which leads to a much lower threshold on the size of the cluster for correct recovery, making it possible for correctly identify some

1. The notations $\tilde{\Omega}(\cdot)$ and $\tilde{O}(\cdot)$ ignore logarithmic factors.
2. Indeed, even in a more lenient setup where one clique (i.e., a perfect cluster) of size K is embedded in an Erdos-Renyi graph of n nodes and 0.5 probability of forming an edge, the best known polynomial-time method requires $K = \Omega(\sqrt{n})$ in order to recover the hidden clique, and it has been a long standing open problem to relax this requirement.

smaller clusters. By repeating this procedure, indeed, we can recover the cluster structure for almost all nodes *with no lower bound on the minimal cluster size*. Below we summarize our main contributions and techniques:

1. We provide a refined analysis (Theorem 2) of the mixed trace-norm and ℓ_1-norm convex relaxation approach for exact cluster recovery proposed in Chen et al. (2014a, 2012), focusing on the case where small clusters exist. We show that in the planted partition setting, if each cluster is either large (more precisely, of size at least $\sigma \approx \sqrt{n}$) or small (of size at most σ/C for some global constant $C > 1$), then with high probability, this convex relaxation approach correctly identifies all large clusters while "ignoring" the small ones. In fact, it is possible to arbitrarily increase the tuning parameter σ in quest of an interval $(\sigma/C, \sigma)$ that is disjoint from the set of cluster sizes. The analysis is done by identifying a certain feasible solution to the convex program and proving its almost sure optimality. This solution easily identifies the large clusters. Previous analysis is performed only in the case where all clusters are of size greater than $\sqrt{n}$.

2. We provide a converse (Theorem 5) of the result just described. More precisely, we show that if for some value of the tuning parameter σ, an optimal solution to the convex relaxation program is an exact representation of a collection of large clusters (a partial clustering), then these clusters are actual ground truth clusters, even if the particular interval corresponding to σ isn't really free of cluster sizes. This allows the practitioner to be certain that the optimal solution is useful. Moreover, this has important algorithmic implications for an iterative recovery procedure which we describe below.

3. The last two points imply that if some interval of the form $(\sigma/C, \sigma)$ is free of cluster sizes, then an exhaustive search of this interval will constructively find large clusters, though not necessarily for that particular interval (Theorem 6). Removing the recovered large clusters leads to a reduced problem with a smaller graph. Repeating this procedure gives rise to an iterative algorithm (Algorithm 2), using a "peeling strategy", to recover smaller and smaller clusters that are otherwise impossible to recover. Using this iterative algorithm, we prove that as long as the *number* of clusters is bounded by $O(\log n)$, regardless of the cluster sizes, we can correctly recover the cluster structure for an overwhelming fraction of nodes (Theorem 7). To the best of our knowledge, this is the first result of provably correct graph clustering assuming only an upper bound on the number of clusters, but otherwise no assumption on the cluster sizes.

4. We extend the result to the partial observation setting, where only a fraction of similarity labels (i.e., edge/no edge) are queried. As expected, large clusters can be identified using small observation rates, and a higher rate is needed to find smaller clusters. Hence, the observation rate serves as the tuning parameter. This gives rise to an *active learning algorithm* (Algorithm 4) based on adaptively increasing the rate of sampling in order to hit an interval free of cluster sizes, and spending more queries on smaller subgraphs after we identify large clusters and peel them off. Performance

guarantees are given for this algorithm (Corollary 8–Theorem 11). This active learning scheme requires significantly fewer samples than uniform sampling .

Beside these technical contributions, this paper suggests a new strategy that is potentially useful for general low-rank matrix recovery and other high-dimensional statistical problems, where the data are typically assumed to have certain low-dimensional structures. Many methods have been developed to exploit this *a priori* structural information so that consistent estimation is possible even when the dimensionality of the problem is larger than the number of samples. Our result shows that one may combine these methods with a "peeling strategy" to further push the envelope of learning structured data: by iteratively recovering the easier structural components and reducing the problem complexity, it may be possible to learn complicated structures that are otherwise difficult to recover using existing one-shot approaches.

1.1 Related Work

The literature of graph clustering is too vast for a detailed survey here; we concentrate on the most related work, and in particular those provide provable guarantees on exact cluster recovery.

1.1.1 Planted Partition Model

Also known as the *stochastic block model* (Holland et al., 1983; Condon and Karp, 2001), this classical model assumes that n nodes are partitioned into subsets, referred to as the "true clusters", and a graph is randomly generated as follows: for each pair of nodes, depending on whether or not they belong to the same subset, an edge connecting them is generated with a probability p or q respectively. The goal is to correctly recover the clusters given the random graph. The planted partition model has a large body of literature. Earlier work focused on the setting where the minimal cluster size is $\Theta(n)$ (Boppana, 1987; Condon and Karp, 2001; Carson and Impagliazzo, 2001; Bollobás and Scott, 2004). Subsequently, a number of methods have been proposed methods to handle sublinear cluster sizes, including randomized algorithms (Shamir and Tsur, 2007), spectral clustering (McSherry, 2001; Chaudhuri et al., 2012; Rohe et al., 2011; Kumar and Kannan, 2010), convex optimization based approaches (Jalali et al., 2011; Chen et al., 2014a, 2012; Ames and Vavasis, 2011; Oymak and Hassibi, 2011) and tensor decomposition methods (Anandkumar et al., 2014). See Chen et al. (2014b) for a survey of existing theoretical guarantees for the planted partition model. While the methodology differs, all the work above requires, sometimes implicitly, a constraint on the minimum size of the true clusters; in particular, the size must be $\Omega(\sqrt{n})$. Our analysis is carried under the planted partition model, and our approach requires no constraint on the cluster sizes. We also mention the work of Zhao et al. (2011) for community detection in social networks, which works under a type of planted partition model. Like ours, their algorithm extracts clusters in an iterative manner and is also amenable to outliers. However, their theoretical guarantees are only shown to hold when $n \to \infty$ and the cluster sizes grow linearly with n.

1.1.2 Low-rank and Sparse Matrix Decomposition via Trace Norm

Motivated by robustifying principal component analysis (PCA), several authors (Chandrasekaran et al., 2011; Candès et al., 2011) show that it is possible to recover a low-rank matrix from sparse errors of arbitrary magnitude, where the key ingredient is using the trace norm (also known as the nuclear norm) as a convex surrogate of the rank. Similar results are obtained when the low rank matrix is corrupted by other types of noise (Xu et al., 2012). Of particular relevance to this paper is the work by Jalali et al. (2011), Oymak and Hassibi (2011) and Chen et al. (2012, 2014a), where they apply this approach to graph clustering, and specifically to the planted partition model. These works require the $\widetilde{\Omega}(\sqrt{n})$ bound on the minimal cluster size. Our approach uses the trace norm relaxation, combined with a more refined analysis and an iterative/active peeling strategy.

1.1.3 Active Learning/Active Clustering

Another line of work that motivates this paper is the study of active learning (a setting in which labeled instances are chosen by the learner, rather than by nature), and in particular active learning algorithms for clustering. The most related work is Ailon et al. (2014), who investigated active learning for the *correlation clustering* problem (Bansal et al., 2004), where the goal is to find a set of clusters whose Hamming distance from the graph is minimized. Ailon et al. (2014) obtain a $(1+\varepsilon)$-approximate solution with respect to the optimum, while (actively) querying no more than $O(n\,\mathrm{poly}(\log n, k, \varepsilon^{-1}))$ edges, where k is the number of clusters. Their result imposed no restriction on cluster sizes and hence inspired this work, but differs in at least two major ways. First, Ailon et al. (2014) did not consider *exact* cluster recovery as we do. Second, their guarantees fall in the Empirical Risk Minimization (ERM) framework, with *no* running time guarantees. Our work uses a convex relaxation algorithm, and is hence computationally efficient. The problem of active learning has also been investigated in other setups including clustering based on distance matrix (Voevodski et al., 2012; Shamir and Tishby, 2011), hierarchical clustering (Eriksson et al., 2011; Krishnamurthy et al., 2012) and low-rank matrix/tensor recovery (Krishnamurthy and Singh, 2013). These setups differ significantly from ours..

Remark 1 (A note on a preliminary version of this paper) *The authors published a weaker version of the results in this paper in a preliminary conference paper (Ailon et al., 2013). An exact comparison is stated after each theorem in the text.*

2. Notation and Setup

In this paper the following notations are used. We use $X(i,j)$ to denote the (i,j)-the entry of a matrix X. For a matrix $X \in \mathbb{R}^{n\times n}$ and a subset $S \subseteq [n]$ of size m, the matrix $X[S] \in \mathbb{R}^{m\times m}$ is the principal minor of X corresponding to the set of indexes S. For a matrix M, $\mathrm{s}(M)$ denotes the support of M, namely, the set of index pairs (i,j) such that $M(i,j) \neq 0$. For any subset Φ of $[n]\times[n]$, $\mathcal{P}_\Phi M$ is the matrix that satisfies

$$(\mathcal{P}_\Phi M)(i,j) = \begin{cases} M(i,j), & (i,j) \in \Phi \\ 0, & \text{otherwise.} \end{cases}$$

We now describe the problem setup. Throughout the paper, V denotes a ground set of elements, which we identify with the set $[n] = \{1, \ldots, n\}$. We assume a ground truth clustering of V given by a pairwise disjoint covering $V_1, \ldots, V_k$, where k is the number of clusters. We say $i \sim j$ if $i, j \in V_a$ for some $a \in [k]$, otherwise $i \not\sim j$. We let $n_a := |V_a|$ be the size of the a-th cluster for each $a \in [k]$. For each $i \in [n]$, $\langle i \rangle$ is index of the cluster that contains i, the unique index satisfying $i \in V_{\langle i \rangle}$.

The ground truth clustering matrix, denoted as K^*, is defined as the $n \times n$ matrix so that $K^*(i, j) = 1$ if $i \sim j$, otherwise 0. This is a block diagonal matrix, each block consisting of 1's only, and its rank is k. The input is a symmetric $n \times n$ matrix A, which is a noisy version of K^*. It is generated according to the *planted partition model* with parameters p and q as follows.

> We think of A as the adjacency matrix of an undirected random graph, where the edge (i, j) is in the graph for $i > j$ with probability p_{ij} if $i \sim j$, otherwise with probability q_{ij}, independent of other choices, where we only assume the edge probabilities satisfy $(\min p_{ij}) =: p > q := (\max q_{ij})$.

We use the convention that the diagonal entries of A are all 1. The matrix $B^* := A - K^*$ can be viewed as the noise matrix. Given A, the task is to find the ground truth clusters.

We remark that the setup above is more flexible than the standard planted partition model: we allow the clusters to have different sizes, and the edges probabilities (p_{ij} and q_{ij}) need not be uniform across node pairs (i, j). One consequence is that the node degrees may not be uniform or correlated with the sizes of the associated clusters. Non-uniformity makes some simple heuristics, such as degree counting and single linkage clustering, vulnerable. For example, we cannot distinguish between large and small clusters simply by looking at the node degrees, since nodes in a small cluster may also have high expected degrees. The single linkage clustering approach also fails in the presence of non-uniformity. We illustrate this with an example. Suppose there are $\sqrt{n}$ clusters of equal size, $p = 1$ and $q = 0.1$. We use the number of common neighbors as the distance function in single linkage clustering. If all q_{ij} are equal to q, then it is easy to see that single linkage clustering will succeed, since with high probability node pairs in the same cluster will have more common neighbors than those in different clusters. Yet, this is not true for non-uniform q_{ij}'s. Consider three nodes 1, 2 and 3, where nodes 1 and 2 are in the same cluster, and node 3 belongs to a different cluster. Suppose for all $i > 3$, $q_{1i} = 0$, $q_{2i} = q_{3i} = 0.1$. The expected number of common neighbors between nodes 1 and 2 is $\sqrt{n}$, whereas the expected number of common neighbors between nodes 2 and 3 is $0.2\sqrt{n} + 0.01(n - 2\sqrt{n})$, which is larger than $\sqrt{n}$ for large n and hence single linkage clustering fails. In contrast, the proposed convex-optimization based method can handle such non-uniform settings, as we show in what follows.

3. Main Results

We remind the reader that the trace norm of a matrix is the sum of its singular values, and the (entry-wise) ℓ_1 norm of a matrix M is $\|M\|_1 := \sum_{i,j} |M(i, j)|$. Consider the following convex program, combining the trace norm of a matrix variable K with the ℓ_1 norm of

another matrix variable B using two parameters c_1, c_2 that will be determined later:

$$\text{(CP)} \qquad \min_{K,B\in\mathbb{R}^{n\times n}} \quad \|K\|_* + c_1 \left\|\mathcal{P}_{\mathrm{s}(A)}B\right\|_1 + c_2 \left\|\mathcal{P}_{\mathrm{s}(A)^c}B\right\|_1$$
$$\text{s.t.} \quad K + B = A,$$
$$0 \le K_{ij} \le 1, \forall (i,j).$$

Here the trace norm term in the objective promotes low-rank solutions and thus encourages the matrix K to have the zero-one block-diagonal structure of a clustering matrix. The matrix $\mathcal{P}_{\mathrm{s}(A)}B = \mathcal{P}_{\mathrm{s}(A)}(A-K)$ is non-zero only on the pairs (i,j) between which there is an edge in the graph ($A_{ij}=1$) but the candidate solution has $K_{ij}=0$, and thus $\mathcal{P}_{\mathrm{s}(A)}B$ corresponds to the "cross-cluster disagreements" between A and K. Similarly, the matrix $\mathcal{P}_{\mathrm{s}(A)^c}B$ corresponds to the "in-cluster disagreements". Hence, the last two terms in the objective is the weighted sum of these two types of disagreements. The formulation (CP) can therefore be considered as a convex relaxation of the so-called *weighted correlation clustering* approach (Bansal et al., 2004), whose objective is to find a clustering that minimizes the weighted disagreements. See Oymak and Hassibi (2011); Mathieu and Schudy (2010); Chen et al. (2014a) for related formulations.

Important to subsequent development is the following new theoretical guarantee for the formulation (CP). We show that (CP) identifies the large clusters whose sizes are above a threshold (chosen by the user) even when small clusters are present. The proof is given in Section 5.1.

Theorem 2 *There exist universal constants $b_3 > 1 > b_4 > 0$ such that the following is true. For any (user-specified) parameters $\kappa \ge 1$ and $t \in [\frac{1}{4}p + \frac{3}{4}q, \frac{3}{4}p + \frac{1}{4}q]$, define*

$$\ell_\sharp := b_3 \frac{\kappa\sqrt{p(1-q)n}}{p-q} \max\left\{1, \frac{\sqrt{p(1-q)}\log^4 n}{\kappa(p-q)\sqrt{n}}\right\}, \quad \ell_\flat := b_4 \frac{\kappa\sqrt{p(1-q)n}}{p-q}\ , \tag{1}$$

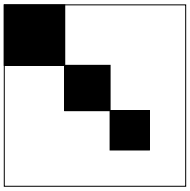
and set

$$c_1 := \frac{1}{100\kappa\sqrt{n}}\sqrt{\frac{1-t}{t}}, \quad c_2 := \frac{1}{100\kappa\sqrt{n}}\sqrt{\frac{t}{1-t}}\ . \tag{2}$$

If (i) $n \ge \ell_\sharp$ and $n \ge 700$, and (ii) for each $a \in [k]$, either $n_a \ge \ell_\sharp$ or $n_a \le \ell_\flat$, then with probability at least $1 - n^{-3}$, the optimal solution to (CP) with c_1, c_2 given above is unique and equal to $(\hat{K}, \hat{B}) = (\mathcal{P}_\sharp K^*, A - \hat{K})$, where for a matrix M, $\mathcal{P}_\sharp M$ is the matrix defined by

$$(\mathcal{P}_\sharp M)(i,j) = \begin{cases} M(i,j), & \max\{n_{\langle i\rangle}, n_{\langle j\rangle}\} \ge \ell_\sharp \\ 0, & \textit{otherwise.} \end{cases}$$

The theorem improves on a weaker version in Ailon et al. 2013, where the ratio $\ell_\sharp/\ell_\flat$ was larger by a factor of $\log^2 n$ than here. The theorem says that the solution to (CP) identifies clusters of size larger than $\ell_\sharp = \Omega(\kappa\sqrt{n})$ and ignores other clusters smaller than $\ell_\flat$. Setting $\kappa = 1$ we recover the usual $\sqrt{n}$ scaling in previous theoretical results. The main novelty here is the treatment of small clusters, whereas in previous work only large clusters were allowed, and there was no guarantee for recovery when small clusters are present.



Black represents 1, white represents 0. Here $\sigma_{\min}(K)$ is the side length of the smallest black square.

Figure 1: Illustration of a partial clustering matrix K.

Note that by the theorem's premise, $\hat{K}$ is the matrix obtained from K^* after zeroing out blocks corresponding to clusters of size at most $\ell_\flat$. Also note that under the assumption

$$p - q \geq \sqrt{p(1-q)} \log^4 n / \sqrt{n} \ , \tag{3}$$

we get the following simpler expression for $\ell_\sharp$ in the theorem, replacing its definition in (1):

$$\ell_\sharp = b_3 \frac{\kappa \sqrt{p(1-q)n}}{p-q} \ . \tag{4}$$

In this case, $\ell_\sharp$ and $\ell_\flat$ differ by only a multiplicative absolute constant b_3/b_4. We will make the assumption (3) in what follows for simplicity, although it is not generally necessary.

Remark 3 *The requirement of having a multiplicative constant gap b_3/b_4 between the sizes $\ell_\sharp$ and $\ell_\flat$ of the large and small clusters, is not an artifact of our analysis; cf. the discussion at the end of Section 4.*

For the convenience of subsequent discussion, we use the following definition.

Definition 4 (Partial Clustering Matrix) *An $n \times n$ matrix K is said to be a partial clustering matrix if there exists a collection of pairwise disjoint sets $U_1, \ldots, U_r \subseteq V$ (called the* induced clusters*) such that $K(i,j) = 1$ if and only if $i, j \in U_a$ for some $a \in [r]$, otherwise 0. If K is a partial clustering matrix then $\sigma_{\min}(K)$ is defined as $\min_{a \in [r]} |U_a|$.*

The definition is depicted in Figure 1. The key message in Theorem 2 is that by choosing κ properly such that no cluster size falls in the interval $(\ell_\flat, \ell_\sharp)$, the unique optimal solution $(\hat{K}, \hat{B})$ to the convex program (CP) is such that $\hat{K}$ is a partial clustering corresponding to large ground truth clusters.

But how can we choose a proper κ? Moreover, given that we chose a κ (say, by exhaustive search), how can we certify that it was indeed chosen properly? In order to develop an algorithm, we would need a type of converse of Theorem 2: There exists an event with high probability (in the random process generating the input graph), such that conditioned on this event, for all values of κ, if an optimal solution to the corresponding (CP) is a partial clustering matrix with the structure illustrated in Figure 1, then the blocks of $\hat{K}$ correspond to ground truth clusters.

Theorem 5 *There exist absolute constants $C_1, C_2 > 0$ such that with probability at least $1 - n^{-3}$, the following holds. For all $\kappa \geq 1$ and $t \in [\frac{3}{4}q + \frac{1}{4}p, \frac{1}{4}q + \frac{3}{4}p]$, if (K, B) is*

an optimal solution to (CP) with c_1, c_2 as defined in Theorem 2, and additionally K is a partial clustering corresponding to $U_1, \ldots, U_r \subseteq V$, with

$$\sigma_{\min}(K) \geq \max\left\{ \frac{C_1 k \log n}{(p-q)^2}, \frac{C_2 \kappa \sqrt{p(1-q)n \log n}}{p-q} \right\}, \tag{5}$$

then $U_1, \ldots, U_r$ are actual ground truth clusters, namely, there exists an injection $\phi : [r] \mapsto [k]$ such that $U_a = V_{\phi(a)}$ for all $a \in [r]$.

Algorithm 1 RecoverBigFullObs(V, A, p, q)

require: ground set V, graph $A \in \mathbb{R}^{V \times V}$, probabilities p, q
$n \leftarrow |V|$
$t \leftarrow \frac{1}{4}p + \frac{3}{4}q$ (or anything in $[\frac{1}{4}p + \frac{3}{4}q, \frac{3}{4}p + \frac{1}{4}q]$)
$\ell_\sharp \leftarrow n$, $g \leftarrow \frac{b_3}{b_4}$
// (If have prior bound k_0 on the number of clusters, take $\ell_\sharp \leftarrow n/k_0$)
while $\ell_\sharp \geq \max\left\{ \frac{C_1 k \log n}{(p-q)^2}, \frac{C_2\sqrt{p(1-q)n \log n}}{p-q} \right\}$ **do**
  solve for κ using (1), set c_1, c_2 as in (2)
  $(K, B) \leftarrow$ optimal solution to (CP) with c_1, c_2
  if K is a partial clustering matrix with $\sigma_{\min}(K) \geq \ell_\sharp$ **then**
    return induced clusters $\{U_1, \ldots, U_r\}$ of K
  end if
  $\ell_\sharp \leftarrow \ell_\sharp / g$
end while
return $\emptyset$

The proof is given in Section 5.2. The combination of Theorems 2 and 5 implies the following, which we state in rough terms for simplicity. Let $g := b_3/b_4$. Assume that we iteratively solve (CP) for κ taking values in some decreasing geometric progression of common ratio g (starting at roughly $\kappa = \sqrt{n}$), and halt if the optimal solution is a partial clustering with clusters of size at least $\ell_\sharp = \ell_\sharp(\kappa)$ (see Algorithm 1). Then these clusters are (extremely likely to be) ground truth clusters. Moreover, if for some κ in the sequence, (i) the interval $(\ell_\flat = \ell_\flat(\kappa), \ell_\sharp = \ell_\sharp(\kappa))$ intersects no cluster size, and (ii) there is at least one cluster at least of size $\ell_\sharp$, then such a halt will (be extremely likely to) occur.

The next question is, when are (i) and (ii) guaranteed? If the number of clusters k is a priori bounded by some k_0, then there is at least one cluster of size at least n/k_0 (alluding to (ii)), and by the pigeonhole principle, any set of $k_0 + 1$ pairwise disjoint intervals of the form $(\alpha, g\alpha)$ contains at least one interval that intersects no clusters size (alluding to (i)). For simplicity, we make an exact quantification of this principle for the case in which p, q are assumed to be fixed and independent of n.[3] As the following theorem shows, it turns out that in this regime, k_0 can be assumed to be asymptotically logarithmic in n to ensure recovery of at least one cluster.[4] In what follows, notation such as $C(p, q), C_3(p, q)$ denotes positive functions that depend on p, q only.

3. In fact, we need only fix $(p - q)$, but we wish to keep this exposition simple.
4. In comparison, Ailon et al. (2014) require k_0 to be constant for their guarantees, as do the Correlation Clustering PTAS in Giotis and Guruswami (2006).

Algorithm 2 RecoverFullObs(V, A, p, q)

require: ground set V, matrix $A \in \mathbb{R}^{V \times V}$, probabilities p, q
$\{U_1, \ldots, U_r\} \leftarrow$ RecoverBigFullObs(V, A, p, q)
$V' \leftarrow [n] \setminus (U_1 \cup \cdots \cup U_r)$
if $r = 0$ **then**
  return $\emptyset$
else
  return RecoverFullObs$(V', A[V'], p, q) \cup \{U_1, \ldots, U_r\}$
end if

Theorem 6 *There exist $C_3(p,q), C_4(p,q), C_5 > 0$ such that the following holds. Assume that $n > C_4(p,q)$, and that we are guaranteed that $k \le k_0$, where $k_0 = C_3(p,q)\log n$. Then with probability at least $1 - 2n^{-3}$, Algorithm 1 will recover at least one cluster in at most $C_5 k_0$ iterations.*

The theorem improves on a counterpart in the preliminary paper (Ailon et al., 2013), where k_0 was smaller by a factor of $\log\log n$ than here.

Proof Consider the set of intervals

$$\left(n/(gk_0), n/k_0\right), \left(n/(g^2k_0), n/(gk_0)\right), \ldots, \left(n/(g^{k_0+1}k_0), n/(g^{k_0}k_0)\right) .$$

By the pigeonhole principle, one of these intervals must not intersect the set of cluster sizes. Assume this interval is $(n/(g^{i_0+1}k_0), n/(g^{i_0}k_0))$, for some $0 \le i_0 \le k_0$. By setting $C_3(p,q)$ small enough so that n/k_0 is at least $\Omega(\sqrt{n\log n})$, and $C_4(p,q)$ large enough so that $n/g^{k_0+1}k_0$ is at least $\Omega(\sqrt{n\log n})$, one easily checks that both the requirements of Theorems 2 and 5 are fulfilled. ■

Theorem 6 ensures that by trying at most a logarithmic number of values of κ, we can recover at least one large cluster, assuming the number of clusters is logarithmic in n. After recovering and removing such a cluster, we are left with an input of size $n' < n$, together with an updated upper bound $k_0' < k_0$ on the number of clusters. As long as k_0' is logarithmic in n', we can continue identifying another large cluster (with respect to the smaller problem) using the same procedure. Clearly, as long as the input size is of size at most $\exp\{C_3(p,q)k_0\}$, we can iteratively continue this process. The following has been proved:

Theorem 7 *Assume an upper bound k_0 on the number k of clusters, and also that n, k_0 satisfy the requirements of Theorem 6. Then with probability at least $1 - 2n^{-2}$, Algorithm 2 recovers clusters covering all but at most $\max\{\exp\{C_3(p,q)k_0\}, C_4(p,q)\}$ elements, without any restriction on the minimal cluster size.*

The theorem improves on a counterpart in the preliminary paper (Ailon et al., 2013). The consequence is, for example, that if $k_0 \le \frac{1}{2C_3(p,q)}\log n$, then the algorithm recovers with high probability clusters covering all but at most $O(n^{1/2})$ elements, without any restriction on the minimal cluster size.

3.1 Partial Observations and Active Sampling

We now consider the case where the input matrix A is not given to us in entirety, but rather that we have oracle access to $A(i,j)$ for (i,j) of our choice. Unobserved values are formally marked as $A(i,j) = *$.

Consider a more particular setting in which the edge probabilities are p' and q', and the probability of sampling an observation is ρ. More precisely: For $i \sim j$ we have $A(i,j) = 1$ with probability $\rho p'$, 0 with probability $\rho(1-p')$ and $*$ with remaining probability, independently of other pairs. For $i \not\sim j$ we have $A(i,j) = 1$ with probability $\rho q'$, 0 with probability $\rho(1-q')$ and $*$ with remaining probability, independently of other pairs. Clearly, by pretending that the values $*$ in A are 0, we emulate the full observation case of the planted partition model with parameters $p = \rho p'$, $q = \rho q'$.

Of particular interest is the case in which p', q' are held fixed and ρ tends to zero as n grows. In this regime, by varying ρ and fixing $\kappa = 1$, Theorem 2 implies the following:

Corollary 8 *There exist constants $b_3(p',q') > b_4(p',q') > 0$ and $b_5(p',q') > 0$ such that the following is true. For any sampling probability parameter $0 < \rho \leq 1$, define*

$$\ell_\sharp = b_3(p',q')\frac{\sqrt{n}}{\sqrt{\rho}} \max\left\{1, \frac{\log^4 n}{\sqrt{\rho n}}\right\}, \qquad \ell_\flat = b_4(p',q')\frac{\sqrt{n}}{\sqrt{\rho}} \ . \tag{6}$$

If for each $a \in [k]$, either $n_a \geq \ell_\sharp$ or $n_a \leq \ell_\flat$, then, with probability at least $1 - n^{-3}$, the program (CP) (after setting $*$ in A to 0) with

$$c_1 = c_1(p',q') = \frac{1}{100\sqrt{n}}\sqrt{\frac{1 - b_5(p',q')\rho}{b_5(p',q')\rho}}$$
$$c_2 = c_2(p',q') = \frac{1}{100\sqrt{n}}\sqrt{\frac{b_5(p',q')}{1 - b_5(p',q')\rho}} \ ,$$

has a unique optimal solution equal to $(\hat{K}, \hat{B}) = (\mathcal{P}_\sharp K^*, A - \hat{K})$, where $\mathcal{P}_\sharp$ is as defined in Theorem 2.

Note that we have slightly abused notation by reusing previously defined global constants (e.g., b_1) with global *functions* of p', q' (e.g., $b_1(p',q')$). Notice now that the sampling probability ρ can be used as a tuning parameter for controlling the sizes of the clusters we try to recover, instead of κ. In what follows, we will always assume the following bound on the observation rate:

$$\rho \geq \frac{\log^8 n}{n} \ , \tag{7}$$

so that the definition of $\ell_\sharp$ in (6) can be replaced by the simpler:

$$\ell_\sharp = b_3(p',q')\frac{\sqrt{n}}{\sqrt{\rho}} \ . \tag{8}$$

This assumption is made for simplicity of the exposition, and a more elaborate (though tedious) derivation can be done without it.

We now present an analogue of the converse result in Theorem 5 for the partial observation setting. Our main focus is to understand the asymptotics as $\rho \to 0$.

Theorem 9 *There exist constants $C_1(p',q'), C_2(p',q') > 0$ such that the following holds with probability at least $1 - n^{-3}$. For all observation rate parameters $\rho \leq 1$, if (K, B) is an optimal solution to (CP) with c_1, c_2 as defined in Corollary 8, and additionally K is a partial clustering corresponding to $U_1, \ldots, U_r \subseteq V$, and also*

$$\sigma_{\min}(K) \geq \max\left\{ \frac{C_1(p',q')k\log n}{\rho}, \frac{C_2(p',q')\sqrt{n\log n}}{\sqrt{\rho}} \right\} , \tag{9}$$

then $U_1, \ldots, U_r$ are actual ground truth clusters, namely, there exists an injection $\phi : [r] \mapsto [k]$ such that $U_a = V_{\phi(a)}$ for each $a \in [r]$.

The proof is similar to that of Theorem 5. The necessary changes are outlined in Section 5.3. Using the same reasoning as before, we derive the following:

Theorem 10 *Let $g = (b_3(p',q')/b_4(p',q'))^2$ (with $b_3(p',q'), b_4(p',q')$ defined in Corollary 8). There exist constants $C_3(p',q')$ and $C_4(p',q')$ such that the following holds. Assume $n \geq C_3(p',q')$ and the number of clusters k is bounded by some known number $k_0 \leq C_4(p',q')\log n$. Let $\rho_0 = \frac{b_3(p',q')^2 k_0^2 \log n}{n}$. Then there exists ρ in the set $\{\rho_0, \rho_0 g, \ldots, \rho_0 g^{k_0}\}$ for which, if A is obtained with sampling rate ρ (zeroing $*$'s), then with probability at least $1 - 2n^{-3}$, any optimal solution (K, B) to (CP) with $c_1(p',q'), c_2(p',q')$ from Corollary 8 satisfies that K is a partial clustering with the property in (9).*

Note that the upper bound on k_0 ensures that ρg^{k_0} is a probability. The theorem improves on a counterpart in the preliminary paper (Ailon et al., 2013), where k_0 was smaller by a factor of $\log\log n$ compared to here. The theorem is proven, again, using a simple pigeonhole principle, noting that one of the intervals $(\ell_\flat(\rho), \ell_\sharp(\rho))$ must be disjoint from the set of cluster sizes, and there is at least one cluster of size at least n/k_0. The value of ρ_0 is chosen so that n/k_0 is larger than the RHS of (9). This theorem motivates the iterative procedure in Algorithm 3: we start with a low sampling rate ρ, which is then increased geometrically until the program (CP) returns a partial clustering.

Theorem 10 together with Corollary 8 and Theorem 9 ensures the following. On one end of the spectrum, if k_0 is a constant (and n is large enough), then with high probability Algorithm 3 recovers at least one large cluster (of size at least n/k_0) after querying no more than

$$O\left(nk_0^2(\log n)\left(\frac{b_3(p',q')}{b_4(p',q')}\right)^{2k_0}\right) \tag{10}$$

values of $A(i,j)$. On the other end of the spectrum, if $k_0 \leq \delta\log n$ and n is large enough (exponential in $1/\delta$), then Algorithm 3 recovers at least one large cluster after querying no more than $n^{1+O(\delta)}$ values of $A(i,j)$. Iteratively recovering and removing large clusters leads to Algorithm 4 with the following guarantees.

Theorem 11 *Assume an upper bound k_0 on the number of clusters k. As long as n is larger than some function of k_0, p', q', Algorithm 4 will recover, with probability at least $1 - n^{-2}$, at least one cluster of size at least n/k_0, regardless of the size of other (small) clusters. Moreover, if k_0 is a constant, then clusters covering all but a constant number of elements will be recovered with probability at least $1 - 2n^{-2}$, and the total number of observation queries is given by (10), hence almost linear.*

Algorithm 3 RecoverBigPartialObs(V, k_0) (Assume p', q' known, fixed)

require: ground set V, oracle access to $A \in \mathbb{R}^{V \times V}$, upper bound k_0 on number of clusters

$n \leftarrow |V|$
$\rho_0 \leftarrow \frac{b_3(p',q')^2 k_0^2 \log n}{n}$
$g \leftarrow b_3(p', q')^2 / b_4(p', q')^2$
for $s \in \{0, \ldots, k_0\}$ **do**
  $\rho \leftarrow \rho_0 g^s$
  obtain matrix $A \in \{0, 1, *\}^{V \times V}$ by sampling oracle at rate ρ, then zero $*$ values in A
  // (can reuse observations from previous iterations)
  $c_1(p', q'), c_2(p', q') \leftarrow$ as in Corollary 8
  $(K, B) \leftarrow$ an optimal solution to (CP)
  if K is a partial clustering matrix satisfying (9) **then**
    return induced clusters $\{U_1, \ldots, U_r\}$
  end if
end for
return $\emptyset$

Algorithm 4 RecoverPartialObs(V, k_0) (Assume p', q' known, fixed)

require: ground set V, oracle access to $A \in \mathbb{R}^{V \times V}$, upper bound k_0 on number of clusters

$\{U_1, \ldots, U_r\} \leftarrow$ RecoverBigPartialObs(V, k_0)
$V' \leftarrow [n] \setminus (U_1 \cup \cdots \cup U_r)$
if $r = 0$ **then**
  return $\emptyset$
else
  return RecoverFullObs$(V', k_0 - r) \cup \{U_1, \ldots, U_r\}$
end if

The theorem improves on a counterpart in the preliminary paper (Ailon et al., 2013), where the recovery covers all but a super-constant (in n) number of elements. Unlike previous convex relaxation based approaches for this problem, which require all cluster sizes to be of size at least roughly $\sqrt{n}$ to succeed, there is no constraint on the cluster sizes for our algorithm.

Also note that our algorithm is an *active learning* one, because more observations fall in smaller clusters which survive deeper in the recursion of Algorithm 4. This feature can lead to a significant saving in the number of queries. When small clusters of size $\tilde{\Theta}(\sqrt{n})$ are present, previous one-shot algorithms for graph clustering with partial observations (e.g., Jalali et al., 2011; Oymak and Hassibi, 2011; Chen et al., 2014a) only guarantee recovery using $O(n^2)$ queries, which is much larger than the almost linear requirement $\tilde{O}(n)$ of our active algorithm.

4. Experiments

We test our main Algorithms 2 and 4 (with subroutines Algorithms 1 and 3) on synthetic data. In all experiment reports below, we use a variant of the Alternating Direction Method of Multipliers (ADMM) to solve the semidefinite program (CP); see Lin et al. (2011); Chen et al. (2012). The main cost of ADMM is the computation of the Singular Value Decomposition (SVD) of an $n \times n$ matrix in each round. Note that one can take advantage of the sparsity of the observations to speed up the SVD (cf. Lin et al. 2011). As is discussed in previous work, and also observed empirically by us, ADMM converges linearly, so the number of SVD needed is usually small. See the references above for further discussion of the optimization issues. The overall computation time also depends on the number of recursive calls in Algorithm 2 and 4, as well as the number of iterations used in Algorithm 1 and 3 in search for suitable values for κ and ρ (using a multiplicative update rule). These two numbers are at most $O(\max(k, \log n))$ (k is the number of clusters) under the conditions of the theorems, and in our experiments they are both quite small.

In the experiments we consider simplified versions of the algorithms: we did not make an effort to compute the constants $\ell_\sharp/\ell_\flat$ defining the algorithms, creating a difficulty in exact implementation. Instead, for Algorithm 1, we start with $\kappa = 1$ and increase κ by a multiplicative factor of 1.1 in each iteration until a partial clustering matrix is found. Similarly, in Algorithm 3, the sampling rate ρ has an initial value of 0 and is increased by an additive factor of 0.025. Still, it is obvious that our experiments support our theoretical findings. A more practical "user's guide" for this method with actual constants is subject to future work.

Whenever we say that "clusters $\{V_{i_1}, V_{i_2}, \dots\}$ were recovered", we mean that a corresponding instantiation of (CP) resulted in an optimal solution (K, B) for which K was a partial clustering matrix induced by $\{V_{i_1}, V_{i_2}, \dots\}$.

4.1 Experiment 1 (Full Observation)

Consider $n = 1100$ nodes partitioned into 4 clusters $V_1, \dots, V_4$, of sizes $800, 200, 80, 20$, respectively. The graph is generated according to the planted partition model with $p = 0.5$ and $q = 0.2$, and we assume the full observation setting. We apply the simplified version of Algorithm 2 described previously, which terminates in 4 iterations using 44 seconds. The recovered clusters at each iteration are detailed in Table 1. The table also shows the values of κ adaptively chosen by the algorithm at each iteration (which happens to equal 1 throughout). We note that the first iteration of the algorithm is similar to existing convex optimization based approaches to graph clustering (Jalali et al., 2011; Oymak and Hassibi, 2011; Chen et al., 2012); the experiment shows that these approaches by itself fail to recover all the clusters in one shot, thus necessitating the iterative procedure proposed in this paper.

4.2 Experiment 2 (Partial Observation, Fixed Sample Rate)

We have $n = 1100$ with clusters $V_1, \dots, V_4$ of sizes $800, 200, 50, 50$. The observed graph is generated with $p' = 0.7$, $q' = 0.1$, and observation rate $\rho = 0.3$. We repeatedly solve (CP) with c_1, c_2 given in Corollary 8. At each iteration, we see that at least one large

Iteration	κ	# nodes left	Clusters recovered
1	1	1100	V_1
2	1	300	V_2
3	1	100	V_3
4	1	20	V_4

Table 1: Results for experiment 1: $n = 1100$, $\{|V_a|\} = \{800, 200, 80, 20\}$, $p = 0.5$, $q = 0.2$, fixed $\rho = 1$.

Iteration	κ	# nodes left	Clusters recovered
1	1	1100	V_1
2	1	300	V_2
3	1	100	V_3, V_4

Table 2: Results for experiment 2: $n = 1100$, $\{|V_a|\} = \{800, 200, 50, 50\}$, $p' = 0.7$, $q' = 0.1$, fixed $\rho = 0.3$.

cluster (compared to the input size at that iteration) is recovered exactly and removed. The experiment terminates in 3 iterations using 18 seconds. Results are shown in Table 2.

4.3 Experiment 3 (Partial Observation, Adaptive Sampling Rate)

We use the simplified version of Algorithm 4 described previously. We have $n = 1100$ with clusters $V_1, \ldots, V_4$ of sizes $800, 200, 50, 50$. The graph is generated with $p' = 0.8$ and $q' = 0.2$, and then adaptively sampled by the algorithm. The algorithm terminates in 3 iterations using 148 seconds. Table 3 shows the recovery result and the sampling rates used in each iteration. From the table we can see that the expected total number of observed entries used by the algorithm is

$$1100^2 \cdot 0.125 + 300^2 \cdot 0.25 + 100^2 \cdot 0.55 = 179250,$$

which is 14.8% of all possible node pairs (the actual number of observations is very close to this expected value). In comparison, we perform another experiment using a non-adaptive sampling rate, for which we need $\rho = 97.5\%$ in order to recover all the clusters in one shot. Therefore, our adaptive algorithm achieves a significant saving in the number of queries.

4.4 Experiment 3A

We repeat the above experiment with a larger instance: $n = 4500$ with clusters $V_1, \ldots, V_6$ of sizes $3200, 800, 200, 200, 50, 50$, and $p' = 0.8, q' = 0.2$. The algorithm terminates in 182 seconds, with results shown in Table 4. Note that we recover the smallest clusters, whose sizes are below $\sqrt{n}$. The expected total number of observations used by the algorithm is 3388000, which is 16.7% of all possible node pairs. Using a non-adaptive sampling rate

Iteration	ρ	# nodes left	Clusters recovered
1	0.125	1100	V_1
2	0.25	300	V_2
3	0.55	100	V_3, V_4

Table 3: Results for experiment 3: $n = 1100$, $\{|V_a|\} = \{800, 200, 50, 50\}$, $p' = 0.8$, $q' = 0.2$.

Iteration	ρ	# nodes left	Clusters recovered
1	0.15	4500	V_1
2	0.175	1300	V_2
3	0.2	500	V_3, V_4
4	0.475	100	V_5, V_6

Table 4: Results for experiment 3A: $n = 4500$, $\{|V_a|\} = \{3200, 800, 200, 200, 50, 50\}$, $p' = 0.8$, $q' = 0.2$.

$\rho = 35.0\%$ only recovers the 4 largest clusters, and we are unable to recover all 6 clusters in one shot even with $\rho = 1$.

4.5 Experiment 4 (Mid-Size Clusters)

Our current theoretical results do not say anything about the mid-size clusters—those with sizes between $\ell_\flat$ and $\ell_\sharp$. It is interesting to investigate the behavior of (CP) in the presence of mid-size clusters. We generate an instance with $n = 750$ nodes partitioned into four clusters of sizes $\{500, 150, 70, 30\}$, edge probabilities $p = 0.8, q = 0.2$ and a sampling rate $\rho = 0.12$. We then solve (CP) with a fixed $\kappa = 1$. The low-rank part K of the solution is shown in Figure 2. The large cluster of size 500 is completely recovered in K, while the two small clusters of sizes 70 and 30 are entirely ignored. The medium cluster of size 150, however, exhibits a pattern we find difficult to characterize. This shows that the constant gap between $\ell_\sharp$ and $\ell_\flat$ in our theorems is a real phenomenon and not an artifact of our proof techniques. Nevertheless, the mid-size cluster appears clean, and might allow recovery using a simple combinatorial procedure. If this is true in general, it might not be necessary to search for a gap free of cluster sizes. In particular, perhaps for any κ, (CP) identifies all large clusters above $\ell_\sharp$ after a simple mid-size cleanup procedure, and ignores all other clusters. Understanding this phenomenon and its algorithmic implications is of much interest.

5. Proofs

We use the following notation and conventions throughout the proofs. *With high probability* or *w.h.p.* means with probability at least $1 - n^{-6}$. The expressions $a \vee b$ and $a \wedge b$ mean $\max\{a, b\}$ and $\min\{a, b\}$, respectively. For a real $n \times n$ matrix M, we use the unadorned norm $\|M\|$ to denote its spectral norm. The notation $\|M\|_F$ refers to the Frobenius norm,

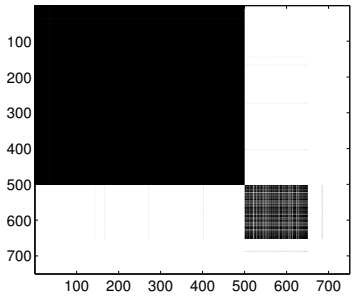


Figure 2: The solution to (CP) with mid-size clusters.

$\|M\|_1$ is $\sum_{i,j} |M(i,j)|$, and $\|M\|_\infty$ is $\max_{ij} |M(i,j)|$. We shall use the standard inner product $\langle X, Y\rangle := \sum_{i,j=1}^{n} X(i,j)Y(i,j)$.

We will also study operators on the space of matrices, and denote them using a calligraphic font, e.g., $\mathcal{P}$. The norm $\|\mathcal{P}\|$ of an operator is defined as

$$\|\mathcal{P}\| := \sup_{M \in \mathbb{R}^{n\times n}: \|M\|_F = 1} \|\mathcal{P}M\|_F \ .$$

For a fixed real $n \times n$ matrix M, we define the matrix linear subspace $T(M)$ as follows:

$$T(M) := \{YM + MX : X, Y \in \mathbb{R}^{n\times n}\} \ .$$

In words, this subspace is the set of matrices spanned by matrices each row of which is in the row space of M, and matrices each column of which is in the column space of M. We let $T(M)^\perp$ denote the orthogonal subspace to $T(M)$ with respect to $\langle \cdot, \cdot \rangle$, which is given by

$$T(M)^\perp := \{X \in \mathbb{R}^{n\times n} : \langle X, Y\rangle = 0, \forall Y \in T(M)\} \ .$$

It is a well known fact that the projection $\mathcal{P}_{T(X)}$ onto $T(X)$ w.r.t. $\langle \cdot, \cdot \rangle$ is given by

$$\mathcal{P}_{T(X)}M := \mathcal{P}_{C(X)}M + \mathcal{P}_{R(X)}M - \mathcal{P}_{C(X)}\mathcal{P}_{R(X)}M \ ,$$

where $\mathcal{P}_{C(X)}$ is projection (of each column of a matrix) onto the column space of X, and $\mathcal{P}_{R(X)}$ is projection onto the row space of X. The projection onto $T(M)^\perp$ is $\mathcal{P}_{T(X)^\perp}M = M - \mathcal{P}_{T(X)}M$.

Finally, we recall that $\mathrm{s}(M)$ is the support of M, $\mathcal{P}_{\mathrm{s}(M)}X$ is the matrix obtained from X by setting its entries outside $\mathrm{s}(M)$ to zero, and $\mathcal{P}_{\mathrm{s}(M)^c}X := X - \mathcal{P}_{\mathrm{s}(M)}X$.

5.1 Proof of Theorem 2

The proof builds on the analysis in Chen et al. (2012). We need some additional notation:

1. We let $V_\flat \subseteq V$ denote the set of of elements i such that $n_{\langle i\rangle} \le \ell_\flat$. (We remind the reader that $n_{\langle i\rangle} = |V_{\langle i\rangle}|$.)

2. We remind the reader that the projection $\mathcal{P}_\sharp$ is defined as follows:

$$(\mathcal{P}_\sharp M)(i,j) = \begin{cases} M(i,j), & \max\{n_{\langle i\rangle}, n_{\langle j\rangle}\} \ge \ell_\sharp \\ 0, & \text{otherwise.} \end{cases}$$

3. The projection $\mathcal{P}_\flat$ is defined as follows:

$$(\mathcal{P}_\flat M)(i,j) = \begin{cases} M(i,j), & \max\{n_{\langle i\rangle}, n_{\langle j\rangle}\} \le \ell_\flat \\ 0, & \text{otherwise.} \end{cases}$$

In words, $\mathcal{P}_\flat$ projects onto the set of matrices supported on $V_\flat \times V_\flat$. Note that by the theorem assumption, $\mathcal{P}_\sharp + \mathcal{P}_\flat = \mathcal{I}d$ (equivalently, $\mathcal{P}_\sharp$ projects onto the set of matrices supported on $(V \times V) \setminus (V_\flat \times V_\flat)$).

4. We use $U\Sigma U^\top$ to denote the rank-k' Singular Value Decomposition (SVD) of the symmetric matrix $\hat{K}$, where $k' = \mathrm{rank}(\hat{K})$ and equals the number of clusters with size at least $\ell_\sharp$.

5. Define the set

$$\mathfrak{D} := \left\{\Delta \in \mathbb{R}^{n\times n} | \Delta_{ij} \le 0, \forall i \sim j, (i,j) \notin V_\flat \times V_\flat; 0 \le \Delta_{ij}, \forall i \not\sim j, (i,j) \notin V_\flat \times V_\flat\right\},$$

which strictly contains all feasible deviation from $\hat{K}$.

6. For simplicity we write $T := T(\hat{K})$.

We will make use of the following facts:

1. $\mathcal{I}d = \mathcal{P}_{\mathrm{s}(\hat{B})} + \mathcal{P}_{\mathrm{s}(\hat{B})^c} = \mathcal{P}_{\mathrm{s}(A)} + \mathcal{P}_{\mathrm{s}(A)^c}$.
2. $\mathcal{P}_\sharp, \mathcal{P}_\flat, \mathcal{P}_{\mathrm{s}(\hat{B})}, \mathcal{P}_{\mathrm{s}(\hat{B})^c}, \mathcal{P}_{\mathrm{s}(A)}$, and $\mathcal{P}_{\mathrm{s}(A)^c}$ commute with each other.

5.1.1 Approximate Dual Certificate Condition

We begin by giving a deterministic sufficient condition for $(\hat{K}, \hat{B})$ to be the unique optimal solution to the program (CP).

Proposition 12 *$(\hat{K}, \hat{B})$ is the unique optimal solution to (CP) if there exists a matrix $Q \in \mathbb{R}^{n\times n}$ and a number $0 < \epsilon < 1$ satisfying:*

1. $\|Q\| < 1$;
2. $\|\mathcal{P}_T(Q)\|_\infty \le \frac{\epsilon}{2}\min\{c_1, c_2\}$;
3. $\forall \Delta \in \mathfrak{D}$:
   (a) $\left\langle UU^\top + Q, \mathcal{P}_{\mathrm{s}(A)}\mathcal{P}_{\mathrm{s}(\hat{B})}\mathcal{P}_\sharp\Delta\right\rangle = (1+\epsilon)c_1\left\|\mathcal{P}_{\mathrm{s}(A)}P_{\mathrm{s}(\hat{B})}\mathcal{P}_\sharp\Delta\right\|_1$,
   (b) $\left\langle UU^\top + Q, \mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}(\hat{B})}\mathcal{P}_\sharp\Delta\right\rangle = (1+\epsilon)c_2\left\|\mathcal{P}_{\mathrm{s}(A)^c}P_{\mathrm{s}(\hat{B})}\mathcal{P}_\sharp\Delta\right\|_1$;
4. $\forall \Delta \in \mathfrak{D}$:
   (a) $\left\langle UU^\top + Q, \mathcal{P}_{\mathrm{s}(A)}\mathcal{P}_{\mathrm{s}(\hat{B})^c}\mathcal{P}_\sharp\Delta\right\rangle \ge -(1-\epsilon)c_1\left\|\mathcal{P}_{\mathrm{s}(A)}P_{\mathrm{s}(\hat{B})^c}\mathcal{P}_\sharp\Delta\right\|_1$,
   (b) $\left\langle UU^\top + Q, \mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}(\hat{B})^c}\mathcal{P}_\sharp\Delta\right\rangle \ge -(1-\epsilon)c_2\left\|\mathcal{P}_{\mathrm{s}(A)^c}P_{\mathrm{s}(\hat{B})^c}\mathcal{P}_\sharp\Delta\right\|_1$;

5. $\mathcal{P}_{\mathrm{s}(\hat{B})}\mathcal{P}_\flat(UU^\top + Q) = c_1\mathcal{P}_\flat\hat{B}$;

6. $\left\|\mathcal{P}_{\mathrm{s}(\hat{B})^c}\mathcal{P}_\flat(UU^\top + Q)\right\|_\infty \le c_2$.

Proof Consider any feasible solution $(\hat{K}+\Delta, \hat{B}-\Delta)$ to (CP); we know $\Delta \in \mathfrak{D}$ due to the inequality constraints in (CP). We will show that this solution will have strictly higher objective value than $(\hat{K}, \hat{B})$ if $\Delta \neq 0$.

For this Δ, let G_Δ be a matrix in $T^\perp \cap \mathrm{Range}(\mathcal{P}_\flat)$ satisfying $\|G_\Delta\| = 1$ and $\langle G_\Delta, \Delta\rangle = \|\mathcal{P}_{T^\perp}\mathcal{P}_\flat\Delta\|_*$; such a matrix always exists because $\mathrm{Range}\mathcal{P}_\flat \subseteq T^\perp$. Suppose $\|Q\| = b$. Clearly, $\mathcal{P}_{T^\perp}Q + (1-b)G_\Delta \in T^\perp$ and, due to Property 1 in the proposition, we have $b < 1$ and $\|\mathcal{P}_{T^\perp}Q + (1-b)G_\Delta\| \le \|Q\| + (1-b)\|G_\Delta\| = b + (1-b) = 1$. Therefore, $UU^\top + \mathcal{P}_{T^\perp}Q + (1-b)G_\Delta$ is a subgradient of $f(K) = \|K\|_*$ at $K = \hat{K}$. On the other hand, define the matrix $F_\Delta = -\mathcal{P}_{\mathrm{s}(\hat{B})^c}\mathrm{sgn}(\Delta)$. We have $F_\Delta \in \mathrm{s}(\hat{B})^c$ and $\|F_\Delta\|_\infty \le 1$. Therefore, $\mathcal{P}_{\mathrm{s}(A)}(\hat{B} + F_\Delta)$ is a subgradient of $g_1(B) = \left\|\mathcal{P}_{\mathrm{s}(A)}B\right\|_1$ at $B = \hat{B}$, and $\mathcal{P}_{\mathrm{s}(A)^c}(\hat{B} + F_\Delta)$ is a subgradient of $g_2(B) = \left\|\mathcal{P}_{\mathrm{s}(A)^c}B\right\|_1$ at $B = \hat{B}$. Using these three subgradients, the difference in the objective value can be bounded as follows:

$$
\begin{aligned}
&d(\Delta)\\
\triangleq&\left\|\hat{K}+\Delta\right\|_* + c_1\left\|\mathcal{P}_{\mathrm{s}(A)}(\hat{B}-\Delta)\right\|_1 + c_2\left\|\mathcal{P}_{\mathrm{s}(A)^c}(\hat{B}-\Delta)\right\|_1 - \left\|\hat{K}\right\|_* - c_1\left\|\mathcal{P}_{\mathrm{s}(A)}\hat{B}\right\|_1\\
&- c_2\left\|\mathcal{P}_{\mathrm{s}(A)^c}\hat{B}\right\|_1\\
\ge&\left\langle UU^\top + \mathcal{P}_{T^\perp}Q + (1-b)G_\Delta, \Delta\right\rangle + c_1\left\langle\mathcal{P}_{\mathrm{s}(A)}(\hat{B}+F_\Delta), -\Delta\right\rangle + c_2\left\langle\mathcal{P}_{\mathrm{s}(A)^c}(\hat{B}+F_\Delta), -\Delta\right\rangle\\
=&(1-b)\|\mathcal{P}_{T^\perp}\mathcal{P}_\flat\Delta\|_* + \left\langle UU^\top + \mathcal{P}_{T^\perp}Q, \Delta\right\rangle + c_1\left\langle\mathcal{P}_{\mathrm{s}(A)}\hat{B}, -\Delta\right\rangle + c_2\left\langle\mathcal{P}_{\mathrm{s}(A)^c}\hat{B}, -\Delta\right\rangle\\
&+ c_1\left\langle\mathcal{P}_{\mathrm{s}(A)}F_\Delta, -\Delta\right\rangle + c_2\left\langle\mathcal{P}_{\mathrm{s}(A)^c}F_\Delta, -\Delta\right\rangle\\
=&(1-b)\|\mathcal{P}_{T^\perp}\mathcal{P}_\flat\Delta\|_* + \left\langle UU^\top + \mathcal{P}_{T^\perp}Q, \Delta\right\rangle + c_1\left\langle\mathcal{P}_\flat\mathcal{P}_{\mathrm{s}(A)}\hat{B}, -\Delta\right\rangle + c_2\left\langle\mathcal{P}_\flat\mathcal{P}_{\mathrm{s}(A)^c}\hat{B}, -\Delta\right\rangle\\
&+ c_1\left\langle\mathcal{P}_\sharp\mathcal{P}_{\mathrm{s}(A)}\hat{B}, -\Delta\right\rangle + c_2\left\langle\mathcal{P}_\sharp\mathcal{P}_{\mathrm{s}(A)^c}\hat{B}, -\Delta\right\rangle + c_1\left\langle\mathcal{P}_{\mathrm{s}(A)}F_\Delta, -\Delta\right\rangle + c_2\left\langle\mathcal{P}_{\mathrm{s}(A)^c}F_\Delta, -\Delta\right\rangle.
\end{aligned}
$$

The last six terms of the last RHS satisfy:

1. $c_1\left\langle\mathcal{P}_\flat\mathcal{P}_{\mathrm{s}(A)}\hat{B}, -\Delta\right\rangle + c_2\left\langle\mathcal{P}_\flat\mathcal{P}_{\mathrm{s}(A)^c}\hat{B}, -\Delta\right\rangle = c_1\left\langle\mathcal{P}_\flat\hat{B}, -\Delta\right\rangle$, because $\mathcal{P}_\flat\hat{B} \in \mathrm{s}(A)$.

2. $\left\langle\mathcal{P}_\sharp\mathcal{P}_{\mathrm{s}(A)}\hat{B}, -\Delta\right\rangle \ge -\left\|\mathcal{P}_\sharp\mathcal{P}_{\mathrm{s}(A)}\mathcal{P}_{\mathrm{s}(\hat{B})}\Delta\right\|_1$ and $\left\langle\mathcal{P}_\sharp\mathcal{P}_{\mathrm{s}(A)^c}\hat{B}, \Delta\right\rangle \ge -\left\|\mathcal{P}_\sharp\mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}(\hat{B})}\Delta\right\|_1$, because $\hat{B} \in \mathrm{s}(\hat{B})$ and $\left\|\hat{B}\right\|_\infty \le 1$.

3. $\left\langle\mathcal{P}_{\mathrm{s}(A)}F_\Delta, -\Delta\right\rangle = \left\|\mathcal{P}_{\mathrm{s}(A)}\mathcal{P}_{\mathrm{s}(\hat{B})^c}\Delta\right\|_1$ and $\left\langle\mathcal{P}_{\mathrm{s}(A)^c}F_\Delta, -\Delta\right\rangle = \left\|\mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}(\hat{B})^c}\Delta\right\|_1$, due to the definition of F.

It follows that

$$
\begin{aligned}
d(\Delta) \ge&(1-b)\|\mathcal{P}_{T^\perp}\mathcal{P}_\flat\Delta\|_* + \left\langle UU^\top + \mathcal{P}_{T^\perp}Q, \Delta\right\rangle + c_1\left\langle\mathcal{P}_\flat\hat{B}, -\Delta\right\rangle - c_1\left\|\mathcal{P}_\sharp\mathcal{P}_{\mathrm{s}(A)}\mathcal{P}_{\mathrm{s}(\hat{B})}\Delta\right\|_1\\
&- c_2\left\|\mathcal{P}_\sharp\mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}(\hat{B})}\Delta\right\|_1 + c_1\left\|\mathcal{P}_{\mathrm{s}(A)}\mathcal{P}_{\mathrm{s}^c(\hat{B})}\Delta\right\|_1 + c_2\left\|\mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}^c}\Delta\right\|_1. \qquad (11)
\end{aligned}
$$

Consider the second term in the last RHS, which equals

$$\left\langle UU^\top + \mathcal{P}_{T^\perp} Q, \Delta \right\rangle = \left\langle UU^\top + Q, \mathcal{P}_\sharp \Delta \right\rangle + \left\langle UU^\top + Q, \mathcal{P}_\flat \Delta \right\rangle - \langle \mathcal{P}_T Q, \Delta \rangle .$$

We bound these three terms separately. For the first term, we have

$$\begin{aligned}
&\left\langle UU^\top + Q, \mathcal{P}_\sharp \Delta \right\rangle \\
=&\left\langle UU^\top + Q, \left(\mathcal{P}_{\mathrm{s}(A)} \mathcal{P}_{\mathrm{s}(\hat{B})} \mathcal{P}_\sharp + \mathcal{P}_{\mathrm{s}(A)^c} \mathcal{P}_{\mathrm{s}(\hat{B})} \mathcal{P}_\sharp + \mathcal{P}_{\mathrm{s}(A)} \mathcal{P}_{\mathrm{s}(\hat{B})^c} \mathcal{P}_\sharp + \mathcal{P}_{\mathrm{s}(A)^c} \mathcal{P}_{\mathrm{s}^c} \mathcal{P}_\sharp \right) \Delta \right\rangle \\
\geq&(1+\epsilon) c_1 \left\| \mathcal{P}_{\mathrm{s}(A)} P_{\mathrm{s}(\hat{B})} \mathcal{P}_\sharp \Delta \right\|_1 + (1+\epsilon) c_2 \left\| \mathcal{P}_{\mathrm{s}(A)^c} P_{\mathrm{s}(\hat{B})} \mathcal{P}_\sharp \Delta \right\|_1 - (1-\epsilon) c_1 \left\| \mathcal{P}_{\mathrm{s}(A)} P_{\mathrm{s}(\hat{B})^c} \mathcal{P}_\sharp \Delta \right\|_1 \\
&- (1-\epsilon) c_2 \left\| \mathcal{P}_{\mathrm{s}(A)^c} P_{\mathrm{s}(\hat{B})^c} \mathcal{P}_\sharp \Delta \right\|_1 . \quad \text{(Using Properties 3 and 4.)}
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
&\left\langle UU^\top + Q, \mathcal{P}_\flat \Delta \right\rangle \\
=&\left\langle \mathcal{P}_{\mathrm{s}(\hat{B})} \mathcal{P}_\flat (UU^\top + Q), \Delta \right\rangle + \left\langle \mathcal{P}_{\mathrm{s}(\hat{B})^c} \mathcal{P}_\flat (UU^\top + Q), \Delta \right\rangle \\
\geq&c_1 \left\langle \mathcal{P}_\flat \hat{B}, \Delta \right\rangle - c_2 \left\| \mathcal{P}_{\mathrm{s}(\hat{B})^c} \mathcal{P}_\flat \Delta \right\|_1 \quad \text{(using Properties 5 and 6)} \\
=&c_1 \left\langle \mathcal{P}_\flat \hat{B}, \Delta \right\rangle - c_2 \left\| \mathcal{P}_{\mathrm{s}(A)^c} \mathcal{P}_{\mathrm{s}(\hat{B})^c} \mathcal{P}_\flat \Delta \right\|_1 . \quad \text{(Because } \mathcal{P}_{\mathrm{s}(A)^c} \mathcal{P}_{\mathrm{s}(\hat{B})^c} \mathcal{P}_\flat = \mathcal{P}_{\mathrm{s}(\hat{B})^c} \mathcal{P}_\flat .)
\end{aligned}$$

Finally, for the third term, Due to the block diagonal structure of the elements of T, we have $\mathcal{P}_T = \mathcal{P}_\sharp \mathcal{P}_T$ and therefore

$$\langle -\mathcal{P}_T Q, \Delta \rangle = - \langle \mathcal{P}_T Q, \mathcal{P}_\sharp \Delta \rangle \geq - \left\| \mathcal{P}_T Q \right\|_\infty \left\| \mathcal{P}_\sharp \Delta \right\|_1 \geq - \frac{\epsilon}{2} \min \{c_1, c_2\} \left\| \mathcal{P}_\sharp \Delta \right\|_1 .$$

Combining the above three bounds with Eq. (11), we obtain

$$\begin{aligned}
&d(\Delta) \\
\geq&(1-b) \left\| \mathcal{P}_{T^\perp} \mathcal{P}_\flat \Delta \right\|_* + \epsilon c_1 \left\| \mathcal{P}_\sharp \mathcal{P}_{\mathrm{s}(A)} \mathcal{P}_{\mathrm{s}(\hat{B})} \Delta \right\|_1 + \epsilon c_2 \left\| \mathcal{P}_\sharp \mathcal{P}_{\mathrm{s}(A)^c} \mathcal{P}_{\mathrm{s}(\hat{B})} \Delta \right\|_1 + \epsilon c_1 \left\| \mathcal{P}_{\mathrm{s}(A)} \mathcal{P}_{\mathrm{s}(\hat{B})^c} \mathcal{P}_\sharp \Delta \right\|_1 \\
&+ \epsilon c_2 \left\| \mathcal{P}_{\mathrm{s}(A)^c} \mathcal{P}_{\mathrm{s}(\hat{B})^c} \mathcal{P}_\sharp \Delta \right\|_1 + c_1 \left\| \mathcal{P}_{\mathrm{s}(A)} \mathcal{P}_{\mathrm{s}(\hat{B})^c} \mathcal{P}_\flat \Delta \right\|_1 - \frac{\epsilon}{2} \min \{c_1, c_2\} \left\| \mathcal{P}_\sharp \Delta \right\|_1 \\
=&(1-b) \left\| \mathcal{P}_{T^\perp} \mathcal{P}_\flat \Delta \right\|_* + \epsilon c_1 \left\| \mathcal{P}_\sharp \mathcal{P}_{\mathrm{s}(A)} \Delta \right\|_1 + \epsilon c_2 \left\| \mathcal{P}_\sharp \mathcal{P}_{\mathrm{s}(A)^c} \Delta \right\|_1 - \frac{\epsilon}{2} \min \{c_1, c_2\} \left\| \mathcal{P}_\sharp \Delta \right\|_1 \\
&\text{(note that } \mathcal{P}_{\mathrm{s}(A)} \mathcal{P}_{\mathrm{s}(\hat{B})^c} \mathcal{P}_\flat \Delta = 0) \\
\geq&(1-b) \left\| \mathcal{P}_\flat \Delta \right\|_* + \frac{\epsilon}{2} \min \{c_1, c_2\} \left\| \mathcal{P}_\sharp \Delta \right\|_1 ,
\end{aligned}$$

which is strictly greater than zero for $\Delta \neq 0$. ■

5.1.2 Constructing Q

To prove the theorem, it suffices to show that with probability at least $1 - n^{-3}$, there exists a matrix Q with the properties required by Proposition 12. We do this by explicitly

constructing Q. Suppose we take

$$\epsilon := \frac{100}{\sqrt{t(1-t)}} \max\left\{ \frac{\kappa\sqrt{n}}{\ell_\sharp}, \sqrt{\frac{\log^4 n}{\ell_\sharp}} \right\},$$

and use the weights c_1 and c_2 given in Theorem 2. We specify $\mathcal{P}_\sharp Q$ and $\mathcal{P}_\flat Q$ separately.

The matrix $\mathcal{P}_\sharp Q$ is given by $\mathcal{P}_\sharp Q = \mathcal{P}_\sharp Q_1 + \mathcal{P}_\sharp Q_2 + \mathcal{P}_\sharp Q_3$, where for $(i,j) \notin V_\flat \times V_\flat$,

$$\mathcal{P}_\sharp Q_1(i,j) = \begin{cases} -\frac{1}{n_{\langle i \rangle}}, & i \sim j, (i,j) \in \mathrm{s}(\hat{B}) \\ \frac{1}{n_{\langle i \rangle}} \cdot \frac{1-p_{ij}}{p_{ij}}, & i \sim j, (i,j) \in \mathrm{s}(\hat{B})^c \\ 0, & i \not\sim j \end{cases}$$

$$\mathcal{P}_\sharp Q_2(i,j) = \begin{cases} -(1+\epsilon)c_2, & i \sim j, (i,j) \in \mathrm{s}(\hat{B}) \\ (1+\epsilon)c_2 \frac{1-p_{ij}}{p_{ij}}, & i \sim j, (i,j) \in \mathrm{s}(\hat{B})^c \\ 0, & i \not\sim j \end{cases}$$

$$\mathcal{P}_\sharp Q_3(i,j) = \begin{cases} (1+\epsilon)c_1, & i \not\sim j, (i,j) \in \mathrm{s}(\hat{B}) \\ -(1+\epsilon)c_1 \frac{q_{ij}}{1-q_{ij}}, & i \not\sim j, (i,j) \in \mathrm{s}(\hat{B})^c \\ 0, & i \sim j. \end{cases}$$

Note that these matrices have zero-mean entries. (Recall that $\mathrm{s}(\hat{B}) = \mathrm{s}(A - \hat{K})$ is a random set since the graph A is random.)

$\mathcal{P}_\flat Q$ is given as follows. For $(i,j) \in V_\flat \times V_\flat$,

$$\mathcal{P}_\flat Q(i,j) = \begin{cases} c_1, & i \sim j, (i,j) \in \mathrm{s}(A) \\ -c_2, & i \sim j, (i,j) \in \mathrm{s}(A)^c \\ c_1, & i \not\sim j, (i,j) \in \mathrm{s}(A) \\ c_2 W(i,j), & i \not\sim j, (i,j) \in \mathrm{s}(A)^c, \end{cases}$$

where W is a symmetric matrix whose upper-triangle entries are independent and obey

$$W(i,j) = \begin{cases} +1, & \text{with probability } \frac{t-q}{2t(1-q)}, \\ -1, & \text{with remaining probability.} \end{cases}$$

Note that we introduced additional randomness in W.

5.1.3 Validating Q

Under the choice of t in Theorem 2, we have $\frac{1}{4}p \le t \le p$ and $\frac{1}{4}(1-q) \le 1-t \le 1-q$. Also under the assumption (1) in the theorem and since $p - q \le p(1-q)$, $\ell_\sharp \le n$, we have $p(1-q) \ge \frac{b_3^2 \kappa^2 n}{\ell_\sharp^2} \vee \frac{b_3 \log^4 n}{\ell_\sharp} \ge \frac{b_3 \log^4 n}{n}$. Using these inequalities, it is easy to check that $\epsilon < \frac{1}{2}$ provided that the constant b_3 is sufficiently large. We will make use of these facts frequently in the proof.

We now verify that the Q constructed above satisfy the six properties in Proposition 12 with probability at least $1 - n^{-3}$.

Property 1:

Suppose the matrix $Q_\sim$ is obtained from Q by setting all $Q(i,j)$ with $i \not\sim j$ to zero, and $Q_{\not\sim} = Q - Q_\sim$. Note that $\|Q\| \le \|\mathcal{P}_\sharp Q_\sim\| + \|\mathcal{P}_\sharp Q_{\not\sim}\| + \|\mathcal{P}_\flat Q_\sim\| + \|\mathcal{P}_\flat Q_{\not\sim}\|$. Below we show that with high probability, the first term is upper-bounded by $\frac{7}{32}$ and the other threes terms are upper-bounded by $\frac{1}{4}$, which establishes that $\|Q\| \le \frac{31}{32}$.

(a) $\mathcal{P}_\flat Q_\sim$ is a block diagonal matrix support on $V_\flat \times V_\flat$, where the size of each block is at most $\ell_\flat$. Note that $\mathcal{P}_\flat Q_\sim = \mathbb{E}[\mathcal{P}_\flat Q_\sim] + (\mathcal{P}_\flat Q_\sim - \mathbb{E}[\mathcal{P}_\flat Q_\sim])$. Here $\mathbb{E}[\mathcal{P}_\flat Q_\sim]$ is a deterministic matrix with all non-zero entries equal to $\frac{1}{100\kappa\sqrt{n}}\frac{p-t}{\sqrt{t(1-t)}}$. We thus have

$$\|\mathbb{E}[\mathcal{P}_\flat Q_\sim]\| \le \ell_\flat \frac{1}{100\kappa\sqrt{n}}\frac{p-t}{\sqrt{t(1-t)}} \le \frac{1}{32},$$

where the last inequality holds under the definition of $\ell_\flat$ in Theorem 2. On the other hand, the matrix $\mathcal{P}_\flat Q_\sim - \mathbb{E}[\mathcal{P}_\flat Q_\sim]$ is a random matrix whose entries are independent, bounded almost surely by $B := \max\{c_1, c_2\}$ and have zero mean with variance bounded by $\frac{1}{100^2\kappa^2 n} \cdot \frac{p(1-p)}{t(1-t)}$. If $\ell_\flat \le n^{2/3}$, we apply part 1 of Lemma 17 to obtain

$$\begin{aligned}\|\mathcal{P}_\flat Q_\sim - \mathbb{E}[\mathcal{P}_\flat Q_\sim]\| &\le 10\max\left\{\frac{1}{100\kappa\sqrt{n}}\sqrt{\frac{p(1-p)}{t(1-t)}\ell_\flat \log n}, (c_1 \vee c_2)\log n\right\} \\ &\le \max\left\{\frac{1}{10\kappa}\sqrt{\frac{p(1-p)}{t(1-t)}\frac{\log n}{n^{1/3}}}, \frac{1}{10\kappa\sqrt{n}}\left(\sqrt{\frac{1-t}{t}} \vee \sqrt{\frac{t}{1-t}}\right)\log n\right\} \le \frac{3}{16},\end{aligned}$$

where the last inequality follows from $t(1-t) \ge \frac{p(1-q)}{16} \gtrsim \frac{\log^4 n}{n}$. If $\ell_\flat \ge n^{2/3} \ge 76$, then the variance of the entries is bounded by $\sigma^2 := \frac{1}{100^2\kappa^2 nt(1-t)}\left(p(1-p) \vee \frac{t^2\log^4 n}{\ell_\flat} \vee \frac{(1-t)^2\log^4 n}{\ell_\flat}\right)$, and $\sigma \gtrsim \frac{B\log^2 n}{\sqrt{\ell_\flat}}$. Hence we can apply part 2 of Lemma 17 to get

$$\|\mathcal{P}_\flat Q_\sim - \mathbb{E}[\mathcal{P}_\flat Q_\sim]\| \le 10\sigma\sqrt{\ell_\flat} \le \frac{3}{16}, \text{w.h.p.},$$

where in the last inequality we use $n \ge \ell_\flat$ and $t(1-t) \ge \frac{1}{16}p(1-q) \gtrsim \frac{\log^4 n}{n}$. We conclude that $\|\mathcal{P}_\flat Q_\sim\| \le \|\mathbb{E}[\mathcal{P}_\flat Q_\sim]\| + \|\mathcal{P}_\flat Q_\sim - \mathbb{E}[\mathcal{P}_\flat Q_\sim]\| \le \frac{1}{32} + \frac{3}{16} = \frac{7}{32}$ w.h.p.

(b) $\mathcal{P}_\flat Q_{\not\sim}$ is a random matrix supported on $V_\flat \times V_\flat$, whose entries are independent, zero mean, bounded almost surely by $B' := \max\{c_1, c_2\}$, and have variance $\frac{1}{100^2\kappa^2 n} \cdot \frac{t^2+q-2tq}{(1-t)t}$. If $\ell_\flat \le n^{2/3}$, we apply part 1 of Lemma 17 to obtain

$$\begin{aligned}\|\mathcal{P}_\flat Q_{\not\sim}\| &\le 10\max\left\{\frac{1}{100\kappa\sqrt{n}}\sqrt{\frac{t^2+q-2tq}{t(1-t)}\ell_\flat \log n}, (c_1 \vee c_2)\log n\right\} \\ &\le \max\left\{\frac{1}{10\kappa}\sqrt{\frac{t^2+q-2tq}{t(1-t)}\frac{\log n}{n^{1/3}}}, \frac{1}{10\kappa\sqrt{n}}\left(\sqrt{\frac{1-t}{t}} \vee \sqrt{\frac{t}{1-t}}\right)\log n\right\} \le \frac{1}{4},\end{aligned}$$

where the last inequality follows from $t(1-t) \ge \frac{p(1-q)}{16} \gtrsim \frac{\log^4 n}{n}$. If $\ell_\flat \ge n^{2/3} \ge 76$, one verifies that the variance of the entries is bounded by $(\sigma')^2 := \frac{1}{100^2\kappa^2 n}\cdot\left(\frac{t^2+q-2tq}{(1-t)t} \vee \frac{t\log^4 n}{(1-t)\ell_\flat} \vee \frac{(1-t)\log^4 n}{t\ell_\flat}\right)$,

and $\sigma' \gtrsim \frac{B' \log^2 n}{\sqrt{\ell_\flat}}$. Hence we can apply part 2 of Lemma 17 to obtain

$$\|\mathcal{P}_\flat Q_\not\sim\| \le 10\sigma'\sqrt{\ell_\flat} \le \frac{1}{4}, \text{w.h.p.},$$

where in the last inequality we use $n \ge \ell_\flat$ and $t(1-t) \ge \frac{1}{16}p(1-q) \gtrsim \frac{\log^4 n}{n}$.

(c) Note that $\mathcal{P}_\sharp Q_\sim = \mathcal{P}_\sharp Q_1 + \mathcal{P}_\sharp Q_2$. By construction these two matrices are both block-diagonal, have independent zero-mean entries which are bounded almost surely by $B_{\sim,1} := \frac{1}{\ell_\sharp p}$ and $B_{\sim,2} := \frac{2c_2}{p}$ respectively, and and have variance bounded by $\sigma^2_{\sim_1} := \frac{1}{p\ell_\sharp^2}$ and $\sigma^2_{\sim_2} := \frac{4(1-t)}{p}c_2^2$ respectively. One verifies that $\sigma_{\sim,i} \gtrsim \frac{B_{\sim,i}\log^2 n}{\sqrt{n}}$ for $i = 1, 2$. We can then apply part 2 of Lemma 17 to obtain $\|\mathcal{P}_\sharp Q_\sim\| \le 10(\sigma_{\sim,1} + \sigma_{\sim,2})\sqrt{n} \le \frac{1}{4}$ w.h.p.

(d) Note that $\mathcal{P}_\sharp Q_\not\sim = \mathcal{P}_\sharp Q_3$ is a random matrix with independent zero-mean entries which are bounded almost surely by $B_\not\sim := \frac{2c_1}{1-q}$ and have variance bounded by $\sigma^2_\not\sim := \frac{4t}{1-q}c_1^2$. One verifies that $\sigma_\not\sim \ge \frac{B_\not\sim \log^2 n}{\sqrt{n}}$. We can then apply part 2 of Lemma 17 to obtain $\|\mathcal{P}_\sharp Q_\not\sim\| \le 4\sigma_\not\sim\sqrt{n} \le \frac{1}{4}$ w.h.p.

Property 2:

Due to the structure of T, we have

$$\begin{aligned}\|\mathcal{P}_T Q\|_\infty = \|\mathcal{P}_T\mathcal{P}_\sharp Q\|_\infty &= \left\|UU^\top(\mathcal{P}_\sharp Q) + (\mathcal{P}_\sharp Q)UU^\top + UU^\top(\mathcal{P}_\sharp Q)UU^\top\right\|_\infty \\ &\le 3\left\|UU^\top\mathcal{P}_\sharp Q\right\|_\infty \le 3\sum_{m=1}^{3}\left\|UU^\top\mathcal{P}_\sharp Q_m\right\|_\infty.\end{aligned}$$

Now observe that $(UU^\top\mathcal{P}_\sharp Q_m)(i,j) = \sum_{l\in V_{\langle i\rangle}}\frac{1}{n_{\langle i\rangle}}\mathcal{P}_\sharp Q_m(l,j)$ is the sum of independent zero-mean random variables with bounded magnitude and variance. Using the Bernstein inequality in Lemma 19, we obtain that for each (i,j) and with probability at least $1-n^{-8}$,

$$\left|(UU^\top\mathcal{P}_\sharp Q_1)(i,j)\right| \le \frac{10}{n_{\langle i\rangle}\ell_\sharp}\left(\sqrt{\frac{1-p}{p}}\cdot\sqrt{n_{\langle i\rangle}\log n} + \frac{\log n}{p}\right) \le \frac{1}{24\kappa}\sqrt{\frac{\log^2 n}{n\ell_\sharp}}, \quad \text{w.h.p.},$$

where in the last inequality we use $p \gtrsim \frac{\kappa^2 n}{\ell_\sharp^2}$. For $i \in V_\flat$, clearly $(UU^\top\mathcal{P}_\sharp Q_1)(i,j) = 0$. By union bound we conclude that $\left\|UU^\top\mathcal{P}_\sharp Q_1\right\|_\infty \le \frac{1}{24\kappa}\sqrt{\frac{\log^2 n}{n\ell_\sharp}}$ w.h.p. We can bound $\left\|UU^\top\mathcal{P}_\sharp Q_2\right\|_\infty$ and $\left\|UU^\top\mathcal{P}_\sharp Q_3\right\|_\infty$ in a similar fashion: for each (i,j) and with probability at least $1-n^{-8}$:

$$\begin{aligned}\left|(UU^\top\mathcal{P}_\sharp Q_2)(i,j)\right| &\le 10\frac{(1+\epsilon)c_2}{n_{\langle i\rangle}}\left(\sqrt{\frac{1-p}{p}}\cdot\sqrt{n_{\langle i\rangle}\log n} + \frac{\log n}{p}\right) \\ &\le \frac{15}{100\kappa}\sqrt{\frac{t}{(1-t)n}}\cdot\left(\sqrt{\frac{(1-p)\log n}{p\ell_\sharp}} + \frac{\log n}{\ell_\sharp p}\right) \le \frac{1}{6\kappa}\sqrt{\frac{\log^2 n}{n\ell_\sharp}},\end{aligned}$$

where the last inequality follows from $p(1-t) \gtrsim \frac{\log n}{\ell_\sharp}$, and

$$\begin{aligned}\left|(UU^\top \mathcal{P}_\sharp Q_3)(i,j)\right| &\le 10\frac{(1+\epsilon)c_1}{n_{\langle i\rangle}}\left(\sqrt{\frac{q}{1-q}}\cdot\sqrt{n_{\langle i\rangle}\log n}+\frac{\log n}{1-q}\right)\\ &\le \frac{15}{100\kappa}\sqrt{\frac{1-t}{tn}}\cdot\left(\sqrt{\frac{q\log n}{(1-q)\ell_\sharp}}+\frac{\log n}{\ell_\sharp(1-q)}\right)\le \frac{1}{6\kappa}\sqrt{\frac{\log^2 n}{n\ell_\sharp}},\end{aligned}$$

where the last inequality follows from $t(1-q) \gtrsim \frac{\log n}{\ell_\sharp}$. On the other hand, under the definition of c_1, c_2 and ϵ, we have

$$c_1\epsilon \ge \frac{1}{100\kappa}\sqrt{\frac{1-t}{tn}}\cdot 100\sqrt{\frac{\log^4 n}{t(1-t)\ell_\sharp}} = \frac{1}{\kappa t}\cdot\sqrt{\frac{\log^4 n}{n\ell_\sharp}} \ge \frac{3}{\kappa}\sqrt{\frac{\log^2 n}{n\ell_\sharp}},$$

and similarly

$$c_2\epsilon \ge \frac{1}{100\kappa}\sqrt{\frac{t}{(1-t)n}}\cdot 100\sqrt{\frac{\log^4 n}{t(1-t)\ell_\sharp}} \ge \frac{3}{\kappa}\sqrt{\frac{\log^2 n}{n\ell_\sharp}}.$$

It follows that $\|\mathcal{P}_T Q\|_\infty \le 3\cdot\left(\frac{1}{24}+\frac{1}{6}+\frac{1}{6}\right)\cdot\frac{\epsilon}{3}(c_1\wedge c_2) \le \frac{\epsilon}{2}(c_1\wedge c_2)$ w.h.p., proving Property 2).

Properties 3(a) and 3(b):

For 3(a), by construction of Q we have

$$\begin{aligned}\left\langle UU^\top + Q, \mathcal{P}_{\mathrm{s}(A)}\mathcal{P}_{\mathrm{s}(\hat{B})}\mathcal{P}_\sharp\Delta\right\rangle &= \left\langle \mathcal{P}_{\mathrm{s}(A)}\mathcal{P}_{\mathrm{s}(\hat{B})}\mathcal{P}_\sharp Q_3, \mathcal{P}_{\mathrm{s}(A)}\mathcal{P}_{\mathrm{s}(\hat{B})}\mathcal{P}_\sharp\Delta\right\rangle\\ &= (1+\epsilon)c_1\cdot\sum_{(i,j)\in\mathrm{s}(\hat{B})\cap\mathrm{s}(A)}\mathcal{P}_\sharp\Delta(i,j)\\ &= (1+\epsilon)c_1\left\|\mathcal{P}_{\mathrm{s}(A)}\mathcal{P}_{\mathrm{s}(\hat{B})}\mathcal{P}_\sharp\Delta\right\|_1,\end{aligned}$$

where the last equality follows from $\Delta \in \mathfrak{D}$. Similarly, since

$$\mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}(\hat{B})}\mathcal{P}_\sharp Q_1 = \mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}(\hat{B})}\mathcal{P}_\sharp(-UU^\top),$$

we have

$$\begin{aligned}\left\langle UU^\top + Q, \mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}(\hat{B})}\mathcal{P}_\sharp\Delta\right\rangle &= \left\langle \mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}(\hat{B})}\mathcal{P}_\sharp Q_2, \mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}(\hat{B})}\mathcal{P}_\sharp\Delta\right\rangle\\ &= -(1+\epsilon)c_2\cdot\sum_{(i,j)\in\mathrm{s}(\hat{B})\cap\mathrm{s}(A)^c}\mathcal{P}_\sharp\Delta(i,j)\\ &= (1+\epsilon)c_2\left\|\mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}(\hat{B})}\mathcal{P}_\sharp\Delta\right\|_1,\end{aligned}$$

where the last equality again follows from $\Delta \in \mathfrak{D}$; this proves Property 3(b).

Properties 4(a) and 4(b):

For 4(a), we have

$$\begin{aligned}
&\left\langle UU^\top + Q, \mathcal{P}_{\mathrm{s}(A)}\mathcal{P}_{\mathrm{s}(\hat{B})^c}\mathcal{P}_\sharp\Delta\right\rangle \\
=&\left\langle \mathcal{P}_{\mathrm{s}(A)}\mathcal{P}_{\mathrm{s}(\hat{B})^c}\mathcal{P}_\sharp\left(UU^\top + \mathcal{P}_\sharp Q_1 + \mathcal{P}_\sharp Q_2\right), \mathcal{P}_{\mathrm{s}(A)}\mathcal{P}_{\mathrm{s}(\hat{B})^c}\mathcal{P}_\sharp\Delta\right\rangle \\
=&\sum_{(i,j)\in \mathrm{s}(\hat{B})^c\cap \mathrm{s}(A)} \left(\frac{1}{n_{\langle i\rangle}} + \frac{1}{n_{\langle i\rangle}}\frac{1-p_{ij}}{p_{ij}} + (1+\epsilon)c_2\frac{1-p_{ij}}{p_{ij}}\right)\mathcal{P}_\sharp\Delta(i,j) \\
\geq& -\left(\frac{1}{p\ell_\sharp} + (1+\epsilon)c_2\frac{1-p}{p}\right)\left\|\mathcal{P}_{\mathrm{s}(A)}P_{\mathrm{s}(\hat{B})^c}\mathcal{P}_\sharp\Delta\right\|_1, \qquad (12)
\end{aligned}$$

where the last inequality follows from $\Delta \in \mathfrak{D}, p_{ij} \geq p$ and $n_{\langle i\rangle} \geq \ell_\sharp, \forall i \in V_\sharp$. Consider the two terms in the parenthesis in (12). For the first term, we have

$$\frac{1}{p\ell_\sharp} = \frac{100\kappa}{\ell_\sharp}\sqrt{\frac{n}{t(1-t)}}\cdot\sqrt{\frac{t(1-t)}{100^2\kappa^2p^2n}} \leq \frac{100\kappa}{\ell_\sharp}\sqrt{\frac{n}{t(1-t)}}\cdot\frac{1}{100\kappa}\sqrt{\frac{1-t}{tn}} \leq \epsilon c_1.$$

For the second term in (12), we have the following:

$$\begin{aligned}
p - t \geq \frac{p-q}{4} \geq& \frac{1}{4}\max\left\{\frac{\kappa\sqrt{b_3p(1-q)n}}{\ell_\sharp}, \sqrt{\frac{b_3p(1-q)\log^4 n}{\ell_\sharp}}\right\} \\
=&\frac{\sqrt{b_3}}{4}\cdot p(1-t)\cdot\frac{\sqrt{t(1-q)}}{\sqrt{p(1-t)}}\cdot\max\left\{\frac{\kappa\sqrt{n}}{\ell_\sharp\sqrt{t(1-t)}}, \sqrt{\frac{\log^4 n}{t(1-t)\ell_\sharp}}\right\} \\
\geq& 8p(1-t)\cdot 100\max\left\{\frac{\kappa\sqrt{n}}{\ell_\sharp\sqrt{t(1-t)}}, \sqrt{\frac{\log^4 n}{t(1-t)\ell_\sharp}}\right\} = 8p(1-t)\epsilon.
\end{aligned}$$

A little algebra shows that this implies $(1+\epsilon)\sqrt{\frac{t}{1-t}}\frac{1-p}{p} \leq (1-\epsilon)\sqrt{\frac{1-t}{t}}$, or equivalently $(1+\epsilon)c_2\frac{1-p}{p} \leq (1-2\epsilon)c_1$. Substituting back to (12), we conclude that

$$\left\langle UU^\top + Q, \mathcal{P}_{\mathrm{s}(A)}\mathcal{P}_{\mathrm{s}(\hat{B})^c}\mathcal{P}_\sharp\Delta\right\rangle \geq -\left(\epsilon c_1 + (1-2\epsilon)c_1\right)\left\|\mathcal{P}_{\mathrm{s}(A)}P_{\mathrm{s}(\hat{B})^c}\mathcal{P}_\sharp\Delta\right\|_1,$$

proving Property 4(a).

For 4(b), we have

$$\begin{aligned}
\left\langle UU^\top + Q, \mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}(\hat{B})^c}\mathcal{P}_\sharp\Delta\right\rangle &= \left\langle \mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}(\hat{B})^c}\mathcal{P}_\sharp Q_3, \mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}(\hat{B})^c}\mathcal{P}_\sharp\Delta\right\rangle \\
&= \sum_{(i,j)\in \mathrm{s}(A)^c\cap \mathrm{s}(\hat{B})^c} -(1+\epsilon)\frac{c_1q_{ij}}{1-q_{ij}}\mathcal{P}_\sharp\Delta(i,j) \\
&\geq -(1+\epsilon)\frac{c_1q}{1-q}\left\|\mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}(\hat{B})^c}\mathcal{P}_\sharp\Delta\right\|_1, \qquad (13)
\end{aligned}$$

where the last inequality follows from $q_{ij} \le q$. Consider the factor before the norm in (13). Similarly as before, we have

$$t - q \ge \frac{p-q}{4} \ge \frac{1}{4}\max\left\{\frac{\kappa\sqrt{b_3 p(1-q)n}}{\ell_\sharp}, \sqrt{\frac{b_3 p(1-q)\log^4 n}{\ell_\sharp}}\right\}$$
$$\ge 2t(1-q)\cdot 100\max\left\{\frac{\kappa\sqrt{n}}{\ell_\sharp\sqrt{t(1-t)}}, \sqrt{\frac{\log^4 n}{t(1-t)\ell_\sharp}}\right\} = 2t(1-q)\epsilon.$$

A little algebra shows that this implies $(1+\epsilon)\sqrt{\frac{1-t}{t}\frac{q}{1-q}} \le (1-\epsilon)\sqrt{\frac{t}{1-t}}$, or equivalently $(1+\epsilon)c_1\frac{q}{1-q} \le (1-\epsilon)c_2$. Substituting back to (13), we conclude that

$$\left\langle UU^\top + Q, \mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}(\hat{B})^c}\mathcal{P}_\sharp\Delta\right\rangle \ge -(1-\epsilon)c_2\left\|\mathcal{P}_{\mathrm{s}(A)^c}\mathcal{P}_{\mathrm{s}(\hat{B})^c}\mathcal{P}_\sharp\Delta\right\|_1,$$

proving Property 4(b).

Properties 5 and 6:

Note that $\mathcal{P}_\flat UU^\top = 0$ and $\mathcal{P}_{\mathrm{s}(\hat{B})}\mathcal{P}_\flat = \mathcal{P}_{\mathrm{s}(A)}\mathcal{P}_\flat$. These two properties hold by construction of Q.

We note that Properties (3)-(6) hold deterministically.

Combining the above results and applying the union bound, we conclude that with probability at least $1 - n^{-3}$, there exists a matrix Q (which is the one constructed and verified above) that satisfies the properties in Proposition 12, where the probability is with respect to the randomness in the graph A and the matrix W. Since W is independent of A, integrating out the randomness in W proves the theorem.

5.2 Proof of Theorem 5

To ease notation, throughout the proof, C denotes a general universal positive constant that can take different values at different locations. We let $\Omega := \mathrm{s}(B^*)$ denote the *noise locations*.

Fix $\kappa \ge 1$ and t in the allowed range, let (K, B) be an optimal solution to (CP), and assume K is a partial clustering induced by $U_1, \ldots, U_r$ for some integer r, and also assume $\sigma_{\min}(K) = \min_{i\in[r]}|U_i|$ satisfies (5). Let $M = \sigma_{\min}(K)$. We need a few helpful facts. Note that from the definition of t, c_1, c_2,

$$q + \frac{1}{4}(p-q) \le \frac{c_2}{c_1+c_2} = t \le p - \frac{1}{4}(p-q)\ . \tag{14}$$

We say that a pair of sets $Y \subseteq V, Z \subseteq V$ is *cluster separated* if there is no pair $(y,z) \in Y \times Z$ satisfying $y \sim z$.

Assumption 13 *For all pairs of cluster-separated sets Y, Z of size at least $m := \frac{C\log n}{(p-q)^2}$ each,*

$$|\hat{d}_{Y,Z} - q| < \frac{1}{4}(p-q)\ , \tag{15}$$

where $\hat{d}_{Y,Z} := \frac{|(Y\times Z)\cap\Omega|}{|Y|\cdot|Z|}$.

This is proven by a Hoeffding tail bound and a union bound to hold with probability at least $1 - n^{-4}$. To see why, fix the sizes m_Y, m_Z of $|Y|, |Z|$, assume $m_Y \leq m_Z$ w.l.o.g. For each such choice, there are at most $\exp\{C(m_Y + m_Z)\log n\} \leq \exp\{2Cm_Z \log n\}$ possibilities for the choice of sets Y, Z. For each such choice, the probability that (15) does not hold is

$$\exp\{-Cm_Y m_Z (p-q)^2\} \tag{16}$$

using Hoeffding inequality. Hence, as long as $m_Y \geq m$ as defined above, using union bound (over all possibilities of m_Y, m_Z and of Y, Z) we obtain (15) uniformly. If we also assume that

$$M \geq 3m \ , \tag{17}$$

the implication of Assumption 13 is that it cannot be the case that some U_i contains a subset U_i' of size in the range $[m, |U_i| - m]$ such that $U_i' = V_g \cap U_i$ for some g. Otherwise, if such a set existed, then we would find a strictly better solution to (CP), call it (K', B'), which is defined so that K' is obtained from K by splitting the block corresponding to U_i into two blocks, one corresponding to U_i' and the other to $U_i \setminus U_i'$. The difference Δ between the cost of (K, B) and (K', B') is (renaming $Y := U_i'$ and $Z := U \setminus U_i'$)

$$\Delta = c_1|(Y \times Z) \cap \Omega| - c_2|(Y \times Z) \cap \Omega^c| = (c_1 + c_2)\hat{d}_{Y,Z}|Y|\,|Z| - c_2|Y|\,|Z| \ .$$

But the sign of Δ is exactly the sign of $\hat{d}_{Y,Z} - \frac{c_2}{c_1+c_2}$ which is strictly negative by (15) and (14). (We also used the fact that the trace norm part of the utility function is equal for both solutions: $\|K'\|_* = \|K\|_*$).

The conclusion is that for each i, the sets $(U_i \cap V_1), \ldots, (U_i \cap V_k)$ must all be of size at most m, except maybe for at most one set of size at least $|U_i| - m$. But note that by the theorem's assumption,

$$M > km = (kC \log n)/(p-q)^2 \ , \tag{18}$$

so we conclude that not all the sets $(U_i \cap V_1), \ldots, (U_i \cap V_k)$ can be of size at most m. Hence exactly one of these sets must have size at least $|U_i| - m$. From this we conclude that there is a function $\phi : [r] \mapsto [k]$ such that for all $i \in [r]$,

$$|U_i \cap V_{\phi(i)}| \geq |U_i| - m \ .$$

We now claim that this function is an injection. We will need the following assumption:

Assumption 14 *For any 4 pairwise disjoint subsets (Y, Y', Z, Z') such that $(Y \cup Y') \subseteq V_i$ for some i, $(Z \cup Z') \subseteq [n] \setminus V_i$, $\max\{|Z|, |Z'|\} \leq m$, $\min\{|Y|, |Y'|\} \geq M - m$:*

$$\begin{aligned} |Y| \cdot |Y'|\,\hat{d}_{Y,Y'} - |Y| \cdot |Z|\,\hat{d}_{Y,Z} - |Y'| \cdot |Z'|\,\hat{d}_{Y',Z'} > \\ \frac{c_2}{c_1 + c_2}(|Y| \cdot |Y'| - |Y| \cdot |Z| - |Y'| \cdot |Z'|) \end{aligned} \tag{19}$$

The assumption holds with probability at least $1 - n^{-4}$ by using Hoeffding inequality, union bounding over all possible sets Y, Y', Z, Z' as above. Indeed, notice that for fixed $m_Y, m_{Y'}, m_Z, m_{Z'}$ (with, say, $m_Y \geq m_{Y'}$), and for each tuple Y, Y', Z, Z' such that $|Y| = m_Y, |Y'| = m_{Y'}, |Z| = m_Z, |Z'| = m_{Z'}$, the probability that (19) is violated is at most

$$\exp\{-C(p-q)^2(m_Y m_{Y'} + m_Y m_Z + m_{Y'} m_{Z'})\} \ . \tag{20}$$

Using (17), this is at most

$$\exp\{-C(p-q)^2(m_Y m_{Y'})\} \ . \tag{21}$$

Now notice that the number of possibilities to choose such a 4 tuple of sets is bounded above by $\exp\{Cm_Y \log n\}$. Assuming

$$M \geq \frac{C\log n}{(p-q)^2} \ , \tag{22}$$

and applying a union bound over all possible combinations Y, Y', Z, Z' of sizes $m_Y, m_{Y'}, m_Z, m_{Z'}$ respectively, of which there are at most $\exp\{Cm_Y \log n\}$, we conclude that (19) is violated for some combination with probability at most

$$\exp\{-C(p-q)^2 m_Y m_{Y'}/2\} \tag{23}$$

which is at most $\exp\{-C\log n\}$ if

$$M \geq \frac{C\log n}{(p-q)^2} \ . \tag{24}$$

Apply a union bound now over the possible combinations of the tuple $(m_Y, m_{Y'}, m_Z, m_{Z'})$, of which there are at most $\exp\{C\log n\}$ to conclude that (19) holds uniformly for all possibilities of Y, Y', Z, Z' with probability at least $1 - n^{-4}$.

Now assume by contradiction that ϕ is not an injection, so $\phi(i) = \phi(i') =: j$ for some distinct $i, i' \in [r]$. Set $Y = U_i \cap V_j, Y' = U_{i'} \cap V_j$, $Z = U_i \setminus Y, Z' = U_{i'} \setminus Y'$. Note that $\max\{|Z|, |Z'|\} \leq m$ and $\min\{|Y|, |Y'|\} \geq M - m$ by the derivations to this point. Consider the solution (K', B') where K' is obtained from K by replacing the two blocks corresponding to $U_i, U_{i'}$ with four blocks: Y, Y', Z, Z'. Inequality (19) guarantees that the cost of (K', B') is strictly lower than that of (K, B), contradicting optimality of the latter. (Note that we used the fact that the corresponding contributions $\|K\|_*$ and $\|K'\|_*$ to the trace-norm part of the utility function are equal.)

We can now also conclude that $r \leq k$. Fix $i \in [r]$. We show that not too many elements of $V_{\phi(i)}$ can be contained in $V \setminus \{U_1 \cup \cdots \cup U_r\}$. We need the following assumption.

Assumption 15 *For all pairwise disjoint sets $Y, X, Z \subseteq V$ such that $|Y| \geq M - m$, $|X| \geq m$, $(Y \cup X) \subseteq V_j$ for some $j \in [k]$, $|Z| \leq m$, $Z \cap V_j = \emptyset$:*

$$\begin{aligned} |X|\cdot|Y|\hat{d}_{X,Y} + \binom{|X|}{2}\hat{d}_{x,x} - |Y|\cdot|Z|\hat{d}_{Y,Z} > \\ \frac{c_2}{c_1+c_2}(|X|\cdot|Y| + \binom{|X|}{2} - |Y|\cdot|Z|) + \frac{|X|}{c_1+c_2} \ . \end{aligned} \tag{25}$$

The assumption holds with probability at least $1 - n^{-4}$. To see why, first notice that $|X|/(c_1+c_2) \leq \frac{1}{8}(p-q)|X|\cdot|Y|$ by (5), as long as C_2 is large enough. This implies that the RHS of (25) is upper bounded by

$$\left(p - \frac{1}{8}(p-q)\right)|X|\cdot|Y| + \frac{c_2}{c_1+c}\left(\binom{|X|}{2} - |Y|\cdot|Z|\right) \tag{26}$$

Proving that the LHS of (25) (denoted $f(X,Y,Z)$) is larger than (26) (denoted $g(X,Y,Z)$) uniformly w.h.p. can now be easily done as follows. By fixing $m_Y = |Y|, m_X = |X|$, the number of combinations for Y, X, Z is at most $\exp\{C(m_Y + m_X)\log n\}$ for some global $C > 0$. On the other hand, the probability that $f(X,Y,Z) \leq g(X,Y,Z)$ for any such option is at most

$$\exp\{-C(p-q)^2 m_Y m_X\} \ . \tag{27}$$

Hence, by union bounding, the probability that some tuple Y, X, Z of sizes m_Y, m_X, m_Z respectively satisfies $f(X,Y,Z) \leq g(X,Y,Z)$ is at most

$$\exp\{-C(p-q)^2 m_Y/2\} \ , \tag{28}$$

which is at most $\exp\{-C\log n\}$ assuming

$$M \geq C(\log n)/(p-q)^2 \ . \tag{29}$$

Another union bound over the possible choices of m_Y, m_X, m_Z proves that (25) holds uniformly with probability at least $1 - n^{-4}$.

Now assume, by way of contradiction, that for some $i \in [r]$, the set $X := V_{\phi(i)} \cap ([n] \setminus \{U_1 \cup \cdots \cup U_r\})$ is of size greater than m. Set $Y := V_{\phi(i)} \cap U_i$ and $Z = U_i \setminus V_{\phi(i)}$. Define the solution (K', B') where K' is obtained from K by replacing the block corresponding to $U_i = Y \cup Z$ in K with two blocks: $Y \cup X$ and Z. Assumption 15 tells us that the cost of (K', B') is strictly lower than that of (K, B). Note that the expression $\frac{|X|}{c_1+c_2}$ in the RHS of (25) accounts for the trace norm difference $\|K'\|_* - \|K\|_* = |X|$.

We are prepared to perform the final "cleanup" step. At this point we know that for each $i \in [r]$, the set $T_i = U_i \cap V_{\phi(i)}$ satisfies

$$t_i := |T_i| \geq \max\{|U_i| - m, |V_{\phi(i)}| - rm\} \ . \tag{30}$$

(To see why $t_i \geq |V_{\phi(i)}| - rm$, note that at most m elements of $V_{\phi(i)}$ may be contained in $U_{i'}$ for $i' \neq i$, and another at most m elements in $V \setminus (U_1 \cup \cdots \cup U_r)$.) We are now going to conclude from this that $U_i = V_{\phi(i)}$ for all i. To that end, let (K', B') be the feasible solution to (CP) defined so that K' is a partial clustering induced by $V_{\phi(1)}, \ldots, V_{\phi(r)}$. We would like to argue that if $K \neq K'$ then the cost of (K', B') is strictly smaller than that of (K, B). Fix the value of the collection

$$\begin{aligned}\mathcal{Y} := ((&r, \phi(1), \ldots, \phi(r),\\ &\big(m_{ij} := |V_{\phi(i)} \cap U_j|\big)\big)_{i,j\in[r],i\neq j} \ ,\\ &\big(m'_i := |V_{\phi(i)} \cap (V \setminus (U_1 \cup \cdots \cup U_r))|\big)_{i\in[r]}).\end{aligned}$$

Let $\beta(\mathcal{Y})$ denote the number of $i \neq j$ such that $m_{ij} > 0$ plus the number of $i \in [r]$ such that $m'_i > 0$. We can assume $\beta(\mathcal{Y}) > 0$, otherwise $U_i = V_{\phi(i)}$ for all $i \in [r]$. The number of possibilities for K giving rise to $\mathcal{Y}$ is $\exp\{C(\sum_{i\neq j} m_{ij} + \sum_i m_i)\log n\}$. Fix such a possibility, and let

$$D_{ij} = V_{\phi(i)} \cap U_j, \qquad D'_i = V_{\phi(i)} \cap (V \setminus (U_1 \cup \cdots \cup U_r)) \ .$$

The difference $\delta(K, K')$ between the (CP) costs of solutions K and K' is given by the following expression:

$$\begin{aligned}\delta =& c_1 \sum_i \sum_{j\neq i} |(D_{ij} \times U_i) \cap \Omega| + c_1 \sum_i \sum_{\substack{j_1<j_2 \\ j_1,j_2\neq i}} |(D_{ij_1} \times D_{ij_2}) \cap \Omega| \\ &+ c_1 \sum_i |((V_{\phi(i)} \setminus D'_i) \times D'_i) \cap \Omega| + c_2 \sum_i \sum_{j\neq i} |(D_{ij} \times U_j) \cap \Omega^c| \\ &- c_2 \sum_i \sum_{j\neq i} |(D_{ij} \times U_i) \cap \Omega^c| - c_2 \sum_i \sum_{\substack{j_1<j_2 \\ j_1,j_2\neq i}} |(D_{ij_1} \times D_{ij_2}) \cap \Omega^c| \\ &- c_2 \sum_i |((V_{\phi(i)} \setminus D'_i) \times D'_i) \cap \Omega^c| - c_1 \sum_i \sum_{j\neq i} |(D_{ij} \times U_j) \cap \Omega| - \sum m'_i \ ,\end{aligned}$$

where the expression $\sum m'_i$ comes from the trace norm contribution. If the quantity $\delta(K, K')$ is non-positive, then at least one of the following must be true:

(i) $\sum_i \sum_{j\neq i} |(D_{ij} \times U_i) \cap \Omega| + \sum_i \sum_{\substack{j_1<j_2 \\ j_1,j_2\neq i}} |(D_{ij_1} \times D_{ij_2}) \cap \Omega|$
$< \frac{c_2}{c_1+c_2} \sum_i \sum_{j\neq i} |(D_{ij} \times U_i)| + \frac{c_2}{c_1+c_2} \sum_i \sum_{\substack{j_1<j_2 \\ j_1,j_2\neq i}} |(D_{ij_1} \times D_{ij_2})|$

(ii) $\sum_i |((V_{\phi(i)} \setminus D'_i) \times D'_i) \cap \Omega| < \frac{c_1}{c_1+c_2} \sum_i |((V_{\phi(i)} \setminus D'_i) \times D'_i)| + \frac{1}{c_1+c_2} \sum_i m'_i$.

(iii) $\sum_i \sum_{j\neq i} |(D_{ij} \times U_j) \cap \Omega| > \frac{c_2}{c_1+c_2} \sum_i \sum_{j\neq i} |(D_{ij} \times U_j)|$.

Inequality (i) occurs with probability at most

$$\exp\left\{-C(p-q)^2 \sum_i M \sum_{j\neq i} m_{ij}\right\} \tag{31}$$

using Hoeffding bound; we also used (30). Inequality (ii) occurs with probability at most

$$\exp\left\{-C(p-q)^2 \sum_i M m'_i\right\} \tag{32}$$

using Hoeffding inequalities. (We also used the fact that the rightmost expression of (ii), $\frac{1}{c_1+c_2} \sum_i m'_i$, is bounded above by $\frac{1}{4}(p-q) \sum_i |((V_{\phi(i)} \setminus D'_i) \times D'_i)|$ due to the theorem assumptions.) Inequality (iii) occurs occurs with probability at most (31), using Hoeffding bounds again.

Now notice that the number of choices of K' giving rise to our fixed $\mathcal{Y}$ and $\beta(\mathcal{Y})$ is, by a gross estimation, at most $\exp\{(\sum_{j\neq i} m_{ij} + \sum_i m'_i) \log n\}$. The assumptions of the theorem ensure that, using a union bound over all such possibilities K, and then over all options for $\beta(\mathcal{Y})$ and $\mathcal{Y}$, with probability at least $1 - n^{-4}$ the difference $\delta(K, K')$ is positive. This means that the (CP) cost of K' is simultaneously strictly lower than that of K for all K we have enumerated over.

Taking the theorem's C_1, C_2 large enough to satisfy the requirements above concludes the proof.

5.3 Proof of Theorem 9

The proof of Theorem 5 in the previous section made repeated use of Hoeffding tail inequalities, for uniformly bounding the size of the intersection of the noise support Ω with various submatrices (with high probability). This is tight for p, q which are bounded away from 0 and 1. However, if $p = \rho p', q = \rho q'$, the noise probabilities p', q' are fixed and ρ tends to 0, a sharper bound is obtained using Bernstein tail bound (Lemma 18 in see Appendix A.2). Using Bernstein inequality instead of Hoeffding inequality, gives the required result. To see how this is done, the counterpart of Assumption 13 above is as follows:

Assumption 16 *For all pairs of cluster-separated sets Y, Z of size at least $m := \frac{C \log n}{\rho}$ each,*

$$|\hat{d}_{Y,Z} - q| < \frac{1}{4}(\rho p' - \rho q') \ , \tag{33}$$

where $\hat{d}_{Y,Z} := \frac{|(Y \times Z) \cap \Omega|}{|Y| \cdot |Z|}$.

Note: In this section, C (and hence also m) depends on p', q' only, which are assumed fixed. Defining henceforth m as in Assumption 9, Assumption 14 holds with probability at least $1 - n^{-4}$. This can be seen by replacing the Hoeffding bound in (20) with a Chernoff bound:

$$\exp\{-C(p', q')\rho(m_Y m_{Y'} + m_Y m_Z + m_{Y'} m_{Z'})\} \ . \tag{34}$$

The rest of the proof is obtained by a similar step by step technical alteration of the proof in Section 5.2.

6. Discussion

An immediate future research is to better understand the "mid-size crisis". Our current results say nothing about clusters that are neither large nor small, falling in the interval $(\ell_\flat, \ell_\sharp)$. Our numerical experiments confirm that the mid-size phenomenon is real: they are neither completely recovered nor entirely ignored by the optimal $\hat{K}$. The part of $\hat{K}$ restricted to these clusters does not seem to have an obvious pattern. Proving whether we can still efficiently recover large clusters in the presence of mid-size clusters is an interesting open problem.

Our study was mainly theoretical, focusing on the planted partition model. As such, our experiments focused on confirming the theoretical findings with data generated exactly according to the distribution we could provide provable guarantees for. It would be interesting to apply the presented methodology to real applications, particularly large data sets merged from web application and social networks.

Another interesting direction is extending the "peeling strategy" to other settings. Our algorithms use the convex program (CP) as a subroutine, taking advantage of the fact that the recovery of large clusters via (CP) is not hindered by the presence of small clusters, and that (CP) has a tunable parameter that controls the sizes of the clusters that are considered large. It is possible that other clustering routines also have these properties and thus can be used as a subroutine in our iterative and active algorithms. More generally, our problem concerns the inference of an unknown structure, and our high-level strategy is to sequentially infer and remove the "easy" (or low-resolution) part of the problem and

zoom into the "hard" (or high-resolution) part. It is interesting to explore this strategy in a broader context, and to understand for what problems and under what conditions this strategy may work.

Acknowledgments

The authors are grateful to the anonymous reviewers for their thorough reviews of this work and valuable suggestions on improving the manuscript. N. Ailon acknowledges the support of a Marie Curie International Reintegration Grant PIRG07-GA-2010-268403, and a grant from Technion-Cornell Innovation Institute (TCII). Y. Chen was supported by NSF grant CIF-31712-23800 and ONR MURI grant N00014-11-1-0688. The work of H. Xu was partially supported by the Ministry of Education of Singapore through AcRF Tier Two grant R265-000-443-112.

Appendix A. Technical Lemmas

In this section we state several lemmas needed in the proofs of our main results.

A.1 The Spectral Norm of Random Matrices

Lemma 17 *Suppose $A \in \mathbb{R}^{N\times N}$ is a symmetric matrix, where A_{ij}, $1 \le i \le j \le m$ are independent random variables, each of which has mean 0 and variance at most σ^2 and is bounded in absolute value by B a.s.*

1. *If $n \ge N$, then with probability at least $1 - n^{-6}$, the first singular value A satisfies*

$$\lambda_1(A) \le 10 \max\left\{\sigma\sqrt{N \log n}, B \log n\right\}.$$

2. *If further $n \ge N \ge 76$, $N \ge n^{2/3}$ and $\sigma \ge c_1 \frac{B \log^2 n}{\sqrt{N}}$ for some absolute constant $c_1 > 0$, then with probability at least $1 - n^{-6}$, we have*

$$\lambda_1(A) \le 10\sigma\sqrt{N}.$$

Proof We first prove part 1 of the lemma. Let e_i be the i-th standard basis in $\mathbb{R}^N$. Define $Z_{ij} = A_{ij}e_ie_j^\top + A_{ji}e_je_i^\top$ for $1 \le i < j \le N$, and $Z_{ii} = A_{ii}e_ie_i^\top$ for $i \in [N]$. Then the Z_{ij}'s are zero-mean random matrices independent of each other, and $A = \sum_{1\le i\le j\le N} Z_{ij}$. We have $\|Z_{ij}\| \le B$ almost surely. We also have

$$\left\|\sum_{1\le i\le j\le N} \mathbb{E}(Z_{ij}Z_{ij}^\top)\right\| = \left\|\sum_{1\le i\le N} \mathbb{E}(A_{ii}^2)e_ie_i^\top + \sum_{1\le i\le N} e_ie_i^\top \sum_{j:j\ne i} \mathbb{E}(A_{ij}^2)\right\| \le N\sigma^2.$$

Similarly, we have $\|\sum_{1\le i\le j\le N} \mathbb{E}(Z_{ij}^\top Z_{ij})\| \le N\sigma^2$. Applying the Matrix Bernstein Inequality (Theorem 1.6 in Tropp 2012) with $t = 10\max\left\{\sigma\sqrt{N\log n}, B\log n\right\}$ yields the desired bound.

We turn to part 2 of the lemma. Let A' be an independent copy of A, and define

$$\bar{A} := \begin{bmatrix} 0 & A \\ A' & 0 \end{bmatrix}.$$

Note that $\bar{A}$ is an $2N \times 2N$ random matrix with i.i.d. entries. If $\sigma \geq c_1 \frac{B \log^2 n}{\sqrt{N}}$ for some sufficiently large absolute constant $c_1 > 0$, then by Theorem 3.1 in Achlioptas and Mcsherry (2007) we know that with probability at least $1 - n^{-6}$, $\lambda_1(\bar{A}) \leq 10\sigma\sqrt{N}$. The lemma follows from noting that $\lambda_1(A) \leq \lambda_1(\bar{A})$. ■

A.2 Standard Bernstein Inequality for the Sum of Independent Variables

Lemma 18 *(*Bernstein inequality*) Let $Y_1, \ldots, Y_N$ be independent random variables, each of which has variance bounded by σ^2 and is bounded in absolute value by B a.s.. Then we have that*

$$\Pr\left[\left|\sum_{i=1}^{N} Y_i - \mathbb{E}\left[\sum_{i=1}^{N} Y_i\right]\right| > t\right] \leq 2\exp\left\{\frac{t^2/2}{N\sigma^2 + Bt/3}\right\}.$$

The following lemma is an immediate consequence of Lemma 18.

Lemma 19 *Let $Y_1, \ldots, Y_N$ be independent random variables, each of which has variance bounded by σ^2 and is bounded in absolute value by B a.s. Then we have*

$$\left|\sum_{i=1}^{N} Y_i - \mathbb{E}\left[\sum_{i=1}^{N} Y_i\right]\right| \leq 10\left(\sigma\sqrt{N\log n} + B\log n\right)$$

with probability at least $1 - 2n^{-8}$.

References

Dimitris Achlioptas and Frank Mcsherry. Fast computation of low-rank matrix approximations. *Journal of the ACM*, 54(2):9, 2007.

Nir Ailon, Yudong Chen, and Huan Xu. Breaking the small cluster barrier of graph clustering. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 995–1003, 2013.

Nir Ailon, Ron Begleiter, and Esther Ezra. Active learning using smooth relative regret approximations with applications. *Journal of Machine Learning Research*, 15:885–920, 2014.

Brendan P. W. Ames and Stephen A. Vavasis. Nuclear norm minimization for the planted clique and biclique problems. *Mathematical Programming*, 129(1):69–89, 2011.

Anima Anandkumar, Rong Ge, Daniel Hsu, and Sham M. Kakade. A tensor spectral approach to learning mixed membership community models. *Journal of Machine Learning Research*, 15:2239–2312, June 2014.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine Learning*, 56:89–113, 2004.

Béla Bollobás and Alex D. Scott. Max cut for random graphs with a planted partition. *Combinatorics, Probability and Computing*, 13(4-5):451–474, 2004.

Ravi B. Boppana. Eigenvalues and graph bisection: an average-case analysis. In *Proceedings of Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 280–285, 1987.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58:1–37, 2011.

Ted Carson and Russell Impagliazzo. Hill-climbing finds random planted bisections. In *Proceedings of the 12th Annual Symposium on Discrete Algorithms*, pages 903–909, 2001.

Venkat Chandrasekaran, Sujay Sanghavi, Pablo Parrilo, and Alan Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

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 (COLT)*, pages 35.1–35.23, 2012.

Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering sparse graphs. In *Advances in Neural Information Processing Systems 25*, pages 2204–2212. Curran Associates, Inc., 2012.

Yudong Chen, Ali Jalali, Sujay Sanghavi, and Huan Xu. Clustering partially observed graphs via convex optimization. *Journal of Machine Learning Research*, 15:2213–2238, June 2014a.

Yudong Chen, Sujay Sanghavi, and Huan Xu. Improved graph clustering. *IEEE Transactions on Information Theory*, 60(10):6440–6455, 2014b.

Anne Condon and Richard M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Structures and Algorithms*, 18(2):116–140, 2001.

Brian Eriksson, Gautam Dasarathy, Aarti Singh, and Robert Nowak. Active clustering: robust and efficient hierarchical clustering using adaptively selected similarities. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, pages 260–268, 2011.

Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. A database interface for clustering in large spatial databases. In *Proceedings of 1st International Conference on Knowledge Discovery and Data Mining (KDD)*, 1995.

Ioannis Giotis and Venkatesan Guruswami. Correlation clustering with a fixed number of clusters. *Theory of Computing*, 2(1):249–266, 2006.

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

Ali Jalali, Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering partially observed graphs via convex optimization. In *Proceedigns of the 28th International Conference on Machine Learning*, pages 1001–1008, 2011.

Akshay Krishnamurthy and Aarti Singh. Low-rank matrix and tensor completion via adaptive sampling. In *Advances in Neural Information Processing Systems 26*, pages 836–844, 2013.

Akshay Krishnamurthy, Sivaraman Balakrishnan, Min Xu, and Aarti Singh. Efficient active algorithms for hierarchical clustering. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 887–894, 2012.

Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *Proceedings of 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 299–308, 2010.

Zhouchen Lin, Risheng Liu, and Zhixun Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *Advances in Neural Information Processing Systems 24*, pages 612–620, 2011.

Claire Mathieu and Warren Schudy. Correlation clustering with noisy input. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 712–728. SIAM, 2010.

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

Nina Mishra, Robert Schreiber, Isabelle Stanton, and Robert E. Tarjan. Clustering social networks. In *Algorithms and Models for the Web-Graph*, pages 56–67. Springer, 2007.

Samet Oymak and Babak Hassibi. Finding dense clusters via low rank + sparse decomposition. arXiv:1104.5186v1, 2011.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic block model. *Annals of Statistics*, 39:1878–1915, 2011.

Ohad Shamir and Naftali Tishby. Spectral clustering on a budget. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 661–669, 2011.

Ron Shamir and Dekel Tsur. Improved algorithms for the random cluster graph model. *Random Structure and Algorithm*, 31(4):418–449, 2007.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Konstantin Voevodski, Maria-Florina Balcan, Heiko Röglin, Shang-Hua Teng, and Yu Xia. Active clustering of biological sequences. *Journal of Machine Learning Research*, 13: 203–225, 2012.

Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust PCA via outlier pursuit. *IEEE Transactions on Information Theory*, 58(5):3047–3064, 2012.

Yahoo!-Inc. Graph partitioning. `http://research.yahoo.com/project/2368`, 2009.

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Community extraction for social networks. *Proceedings of the National Academy of Sciences*, 108(18):7321–7326, 2011.