Journal of Machine Learning Research (2018) Submitted 8/18; Published

Exact Clustering of Weighted Graphs via Semidefinite Programming

Aleksis Pirinen ALEKSIS@MATHS.LTH.SE
Centre for Mathematical Sciences
Lund University
Lund, Sweden

Brendan Ames BPAMES@UA.EDU
Department of Mathematics
The University of Alabama
Tuscaloosa, Alabama, AL 35487-0350, USA

Editor: TBA

Abstract

As a model problem for clustering, we consider the densest k-disjoint-clique problem of partitioning a weighted complete graph into k disjoint subgraphs such that the sum of the densities of these subgraphs is maximized. We establish that such subgraphs can be recovered from the solution of a particular semidefinite relaxation with high probability if the input graph is sampled from a distribution of clusterable graphs. Specifically, the semidefinite relaxation is exact if the graph consists of k large disjoint subgraphs, corresponding to clusters, with weight concentrated within these subgraphs, plus a moderate number of nodes not belonging to any cluster. Further, we establish that if noise is weakly obscuring these clusters, i.e, the between-cluster edges are assigned very small weights, then we can recover significantly smaller clusters. For example, we show that in approximately sparse graphs, where the between-cluster weights tend to zero as the size n of the graph tends to infinity, we can recover clusters of size polylogarithmic in n under certain conditions of the distribution of edge weights. Empirical evidence from numerical simulations is also provided to support these theoretical phase transitions to perfect recovery of the cluster structure.

Keywords: clustering, densest subgraph, stochastic block models, semidefinite programming, sparse graphs

1. Introduction

Clustering is a fundamental problem in machine learning and statistics, focusing on the identification and classification of groups, called *clusters*, of similar items in a given data set. Clustering is ubiquitous, playing a prominent role in varied fields such as computational biology, information retrieval, pattern recognition, image processing and computer vision, and network analysis. This problem is inherently ill-posed, as the partition or clustering of any given data set will depend heavily on how we quantify similarity between items in the data set and how we characterize clusters; it is not outside the realm of possibility to have two drastically different clusterings of the same data if two different similarity metrics are used in the clustering process. Regardless of the similarity metric used, clustering is a combinatorial optimization problem at its core: given data, identify a partition or labeling of the data (approximately) maximizing some measure of quality of the clustering. Due to the difficulties inherent with optimization over discrete sets, many popular approaches for clustering involve the approximate solution of an NP-hard combinatorial optimization problem; for example, the spectral clustering heuristic for the normalized cut problem (Dhillon et al., 2004; Ng et al., 2002), the convex relaxation approaches for the correlation clustering problem (Mathieu and Schudy, 2010), robust principal component analysis (Chen et al., 2014a; Oymak and Hassibi, 2011), and the densest k-disjoint-clique problem (Ames and Vavasis, 2014; Ames, 2014), among many others.

In spite of the inherent intractability of clustering, many recent analyses have established that if data is sampled from some distribution of clusterable data, then one can efficiently recover the underlying cluster structure using a variety of clustering algorithms; in particular, see Abbe et al. (2016); Ailon et al. (2013); Ames and Vavasis (2014); Ames (2014); Amini and Levina (2014); Cai et al. (2015); Chen et al. (2014a,b); Chen and Xu (2014); Guédon and Vershynin (2015); Hajek et al. (2015); Lei et al. (2015); Mathieu and Schudy (2010); Nellore and Ward (2013); Oymak and Hassibi (2011); Rohe et al. (2011); Qin and Rohe (2013); Vinayak et al. (2014). Most of these results assume that the similarity structure of the data can be modeled as a graph sampled from some generalization of the *stochastic block model* (Holland et al., 1983). In this model, the nodes of the graph, called the *similarity graph* of the data, are associated with the items in the data set. An edge is drawn between two items with fixed probability p if the corresponding items belong to the same cluster, and with fixed probability $q < p$ if the corresponding items belong to different clusters. Under this block model, the analyses cited above establish that the block structure of the data can be recovered in polynomial-time with high probability provided that the smallest cluster in the data is sufficiently large, typically larger than $\tilde{c}\sqrt{n}$, where n denotes the number of items in the data (and nodes in the similarity graph) and $\tilde{c}$ is a polylogarithmic factor in n depending on $p - q$.

Although valuable in establishing sufficient conditions for data to be clusterable, these results are not immediately applicable to data sets seen in many applications, particularly those arising from the analysis of social networks. For example, statistical analysis of social networks suggests that communities, playing the role of clusters, tend to be limited in size to several hundred users, while the networks themselves can contain thousands, if not millions or even billions, of users (Leskovec et al., 2008, 2009). However, several recent analyses (Chen et al., 2014a; Chen and Xu, 2014; Guédon and Vershynin, 2015; Jalali et al., 2015; Rohe et al., 2012) suggest that these clusterability results are overly conservative with respect to the size of clusters we can expect to recover in polynomial-time. Specifically, these analyses allow the edge probabilities p and q to vary with n, and investigate how the size of the smallest cluster that can be recovered depends on the relative scaling of p, q and n. In this case, the data is often assumed to be sampled from a *sparse* generalized stochastic block model where the parameters p and q governing edge formation are functions depending on the number of items n and one or both tends to 0 as $n \to \infty$. In the case where p tends to 0 much more slowly than q, the noise obscuring the block structure is significantly weaker than in the dense graph case (where p and q are assumed fixed). Here, sparsity refers to the fact that graphs generated according to the block model contain very few edges between clusters with high probability when n is large, and not that the graph itself is sparse in the sense that the nodes have small average degree. In this case, it has been shown that clusters significantly smaller than $\sqrt{n}$ can be recovered efficiently; specifically, several methods have been shown to recover clusters with size polylogarithmic in n under certain assumptions on the probability functions p and q (see Chen et al., 2014a; Chen and Xu, 2014; Guédon and Vershynin, 2015; Rohe et al., 2012). We should note that these results provide evidence of a computational limit for cluster recovery; that is, these results establish that clusters can be recovered in a computationally efficient way if the underlying data satisfies certain sufficient conditions. We should note further that the lower bounds on cluster size given by these sufficient conditions typically do not match information-theoretic limits; it is well-known that it is possible to identify clusters of size on the order of $\log n$ in certain settings, however, no polynomial-time algorithms are known to do so (see Chen and Xu (2014); Hajek et al. (2015) for further details).

The primary contribution of this paper is an analysis establishing similar clusterability results for a particular convex relaxation of the clustering problem. That is, we present an analysis establishing the following theorem, which provides conditions for perfect recovery of the underlying cluster structure from the solution of a particular semidefinite program. As an immediate corollary, the theorem establishes that one may identify clusters as small as $\Omega(\log n)$, i.e., there exists constant c such that the size of the smallest cluster recoverable cluster is bounded below by $c \log n$ for sufficiently large n, with high probability if the data

is sampled from the sparse block model described above for particular choices of p and q. Here, we say that an event occurs *with high probability (w.h.p.)* if the event occurs with probability tending polynomially to 1 as $n \to \infty$.

Theorem 1 *Suppose that the n-node graph $G = (V, E)$ is sampled from the generalized stochastic block model, with k disjoint blocks, in-cluster edge probability p, and between-cluster edge probability q. Let $\boldsymbol{A} \in \mathbf{R}^{n \times n}$ denote the adjacency matrix of G and let $\hat{r}$ and $\tilde{r}$ denote the cardinality of the smallest and largest clusters, respectively, in the block model for G. Then there exists constants $c_1, c_2, c_3 > 0$ such that the columns of the optimal solution $\boldsymbol{X}^*$ of the semidefinite program*

$$\max_{\boldsymbol{X} \in \Sigma_+^n} \{\mathrm{Tr}(\boldsymbol{A}\boldsymbol{X}) : \boldsymbol{X}\boldsymbol{e} \leq \boldsymbol{e}, \mathrm{Tr}(\boldsymbol{X}) = k, \boldsymbol{X} \geq \boldsymbol{0}\}$$

are scalar multiples of the characteristic vectors of the clusters in our underlying block model with high probability if

$$p - q \geq c_3 \max\left\{ \sqrt{\frac{\tilde{\sigma}^2 \log n}{\hat{r}}}, \frac{\log n}{\hat{r}} \right\},$$

where $\tilde{\sigma}^2 = \max\{p(1-p), q(1-q)\}$, and

$$(p-q)\hat{r} \geq c_1 \max\left\{ \sqrt{q(1-q)n}, \sqrt{\log n} \right\} + c_2 \max\left\{ \sqrt{p(1-p)\tilde{r}}, \sqrt{\log n} \right\}.$$

Moreover, in this case, every characteristic vector of a cluster in the block model is a scalar multiple of at least one column of $\boldsymbol{X}^*$.

Here, the characteristic vector of a set $S \subseteq \{1, 2, \ldots, n\}$ is the vector $\boldsymbol{x} \in \{0,1\}^n$ with ith element

$$x_i = \begin{cases} 1, & \text{if } i \in S \\ 0, & \text{otherwise.} \end{cases}$$

In Theorem 1, $\mathrm{Tr}(\boldsymbol{X})$ denotes the trace of the matrix $\boldsymbol{X}$, $\boldsymbol{e}$ denotes the all-ones vector of appropriate dimension, the notation $\boldsymbol{X} \geq \boldsymbol{0}$ indicates that the entries of $\boldsymbol{X}$ are nonnegative, and Σ_+^n denotes the cone of $n \times n$ symmetric positive semidefinite matrices.

Note that if G is sampled from the *dense* block model, i.e., p, q are independent of n, then Theorem 1 suggests that we have exact recovery if $\hat{r} \geq c\sqrt{n}$ with high probability, where c is a constant depending on p, q; this bound matches that established by Ames (2014) (among many others) up to constant terms. On the other hand, when G is sampled from the *sparse* block model, we see that Theorem 1 suggests that we may have perfect recovery of significantly smaller clusters. For example, suppose that $p = 1$ is fixed and $q = \log n/n$. Then we have exact recovery with high probability if the smallest cluster has size $\hat{r} = \Omega(\log n)$; see the discussion following Theorem 3.

We will show that analogous phenomena occur in what we will call *approximately* sparse graphs. In many practical applications, the expectation that we have a binary labeling indicating whether any pair of items in a given data set are similar or dissimilar is unrealistic. However, it is often possible to describe the *level* of similarity between any two items using some affinity function based on distance between the items in question. For example, we could consider the discrepancy in pixel intensity and geographic location in image segmentation applications or Euclidean distance between two items represented as vectors in a Euclidean space (or some other vector space with corresponding norm). In this case, we can summarize the pairwise similarity relationships within our data using a weighted graph, called a *weighted similarity graph*. Specifically, given a data set with affinity function f, the weighted similarity graph is the weighted complete graph with nodes corresponding to the items in the data set, and edge weight w_{ij} between nodes i and j given by the value of $f(i,j)$. Clearly, this contains the similarity graphs discussed earlier as a special case where $w_{ij} = 1$ if items i and j are known to be similar and $w_{ij} = 0$ otherwise; note that we assume that we have an undirected graph with symmetric adjacency matrix.

We can generalize the stochastic block model in an identical fashion. We assume that items in the same cluster are significantly more similar than pairs of items in different clusters. This corresponds to edge weights within clusters being larger, on average, than edge weights between clusters. This motivates the following random graph model, which we will call the *planted cluster model*. Let $G = (V, \boldsymbol{W})$ be the weighted complete graph whose node set represents the items in some data set containing k clusters and (potentially) some nodes that will not be assigned to a cluster. For each pair of nodes u, v in the same cluster C_ℓ, $\ell \in \{1, 2, \ldots, k\}$, we randomly sample edge weight $w_{uv} \geq 0$, and w_{vu} by symmetry, from some probability distribution Ω_ℓ with mean $\alpha_\ell \geq \alpha > 0$. If $u \in C_i$, $v \in C_j$, where $i \neq j$, i.e., u, v do not belong to the same cluster, we sample $w_{uv} = w_{vu} \geq 0$ from a different probability distribution Ω_{ij} with mean $\beta_{ij} \leq \beta \in [0, \alpha)$. Note that this model contains the generalized stochastic block model discussed earlier as a special case when Ω_ℓ and Ω_{ij} are Bernoulli distributions with probabilities of success $p_\ell = p$ and $q_{ij} = q$, respectively.

It was shown by Ames (2014) that if $G = (V, \boldsymbol{W})$ is sampled from the planted cluster model with minimum cluster size at least $c\sqrt{n}$ in the homogeneous case where all within-cluster edges are i.i.d. with mean α and all between-cluster edges are i.i.d. with mean β, where c is a constant depending on α and β, then we can recover the clusters from the optimal solution of the semidefinite program

$$\max_{\boldsymbol{X} \in \Sigma_+^n} \left\{ \operatorname{Tr}(\boldsymbol{W}\boldsymbol{X}) : \boldsymbol{X}\boldsymbol{e} \leq \boldsymbol{e}, \operatorname{Tr}(\boldsymbol{X}) = k, \boldsymbol{X} \geq \boldsymbol{0} \right\} \tag{1}$$

with high probability, where k is the number of clusters in the graph. We will show that these results can be strengthened to establish that much smaller clusters can be recovered in the

presence of *approximately* sparse noise. That is, we will see that if the between-cluster edge weights have expectation β and variance σ_2^2 approaching zero sufficiently quickly as $n \to \infty$, then we may recover clusters containing as few as $\Omega(\log n)$ nodes with high probability. We will derive the semidefinite program (1) as a relaxation of a particular model problem for clustering in Section 2.1 and formally state our recovery guarantees in Section 2.2; we will see that these results immediately specialize to those stated in Theorem 1 for the semidefinite program (1).

2. Semidefinite Relaxations of the Densest k-Disjoint Clique Problem

In this section, we derive a semidefinite relaxation for the densest k-disjoint clique problem and present an analysis illustrating a sufficient condition ensuring that this relaxation is exact. This problem will act as a model problem for clustering and we will see that we should expect to accurately recover the underlying cluster structure if the given data satisfies this sufficient condition.

2.1 The Densest k-disjoint Clique Problem

We begin by deriving a heuristic for the clustering problem based on semidefinite relaxation of the densest disjoint clique problem. A similar discussion motivating the relaxation was originally presented by Ames (2014); we repeat it here for completeness. Let $K_n = (V, \boldsymbol{W})$ be a weighted complete graph with vertex set $V = \{1, 2, \ldots, n\}$ and nonnegative edge weights $w_{ij} \in [0, 1]$ for all $i, j \in V$. Given a subgraph H of K_n, the *density* d_H of H is the average edge weight incident at a vertex in H:

$$d_H = \sum_{ij \in E(H)} \frac{w_{ij}}{|V(H)|}.$$

If we assume that K_n is the similarity graph of some data set consisting of k disjoint clusters and that weight is concentrated more heavily on within-cluster edges than between-cluster edges, then we may cluster this data set by finding the set of k disjoint subgraphs, corresponding to these clusters, with maximum density; we call this problem the *densest k-partition problem*. Unfortunately, the densest k-partition problem is NP-hard (see Peng and Wei, 2007). Moreover, this partition model excludes the case where some items in the data set do not naturally associate with any of the clusters in the data. To simultaneously motivate a convex relaxation of the densest k-partition problem and address the inclusion of nodes that do not naturally belong to clusters, we consider the *densest k-disjoint clique problem*.

Given a graph $G = (V, E)$, a *clique* of G is a pairwise adjacent subset of V. That is, $C \subseteq V$ is a clique of G if $ij \in E$ for every pair of nodes $i, j \in C$ or, equivalently, the subgraph $G(C)$ induced by C is complete. We say that H is a *k-disjoint-clique* subgraph of K_n if $V(H)$ consists of k disjoint cliques, i.e., H is the union of k disjoint complete subgraphs of K_n. The *densest k-disjoint-clique problem* seeks a k-disjoint-clique subgraph H^* maximizing the sum of the densities of the disjoint complete subgraphs comprising H^*. Note that if we add the additional constraint that each node in K_n belongs to exactly one k-disjoint-clique subgraph in K_n, then the densest k-disjoint-clique problem becomes the densest k-partition problem. However, in general, the densest k-disjoint-clique problem allows an assignment of nodes to clusters, represented by the disjoint cliques, which excludes some nodes. For example, if such nodes are present in the data, they would not be assigned to a cluster by the optimal k-disjoint-clique subgraph.

The complexity of the densest k-disjoint-clique problem is unknown; in particular, no polynomial-time algorithm for its solution is known. To address this potential intractability, we will attempt to approximately solve the k-disjoint-clique problem by convex relaxation. Suppose that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ are the characteristic vectors of a set of disjoint cliques $C_1, C_2, \ldots, C_k$ forming a k-disjoint-clique subgraph of K_n. Using this notation, the density of the complete subgraph induced by C_i is equal to

$$d_{G(C_i)} = \sum_{u,v \in C_i} \frac{w_{uv}}{|C_i|} = \frac{\boldsymbol{v}_i^T \boldsymbol{W} \boldsymbol{v}_i}{\boldsymbol{v}_i^T \boldsymbol{v}_i}.$$

If we let $\boldsymbol{P}$ be the $n \times k$ matrix with ith column equal to $\boldsymbol{v}_i / \|\boldsymbol{v}_i\|$, where $\|\cdot\| = \|\cdot\|_2$ denotes the standard Euclidean norm, then it is easy to see that

$$\sum_{i=1}^{k} d_{G(C_i)} = \mathrm{Tr}(\boldsymbol{P}^T \boldsymbol{W} \boldsymbol{P}).$$

We call such a matrix $\boldsymbol{P}$ a *normalized k-cluster matrix* and denote the set of normalized k-cluster matrices of the vertex set V by $ncm(V, k)$. It follows that the densest k-disjoint-clique problem may be formulated as

$$\max \left\{ \mathrm{Tr}(\boldsymbol{P}^T \boldsymbol{W} \boldsymbol{P}) : \boldsymbol{P} \in ncm(V, k) \right\}. \tag{2}$$

Again, the complexity of (2) is unknown, however, the maximization of quadratic functions subject to combinatorial constraints is known to be NP-hard.

A process for relaxation of (2) using matrix lifting is described by Ames (2014); a similar relaxation technique was applied by Ames and Vavasis (2011, 2014) and Ames (2015). In particular, each proposed cluster C_i, with characteristic vector $\boldsymbol{v}_i$, corresponds to the rank-one symmetric matrix

$$\boldsymbol{X}^{(i)} = \frac{\boldsymbol{v}_i \boldsymbol{v}_i^T}{\boldsymbol{v}_i^T \boldsymbol{v}_i}.$$

It is easy to see that the density of $G(C_i)$ is equal to

$$d_{G(C_i)} = \frac{\boldsymbol{v}_i^T \boldsymbol{W} \boldsymbol{v}_i}{\boldsymbol{v}_i^T \boldsymbol{v}_i} = \mathrm{Tr}(\boldsymbol{W}\boldsymbol{X}^{(i)}).$$

Moreover, each of the matrices $\boldsymbol{X}^{(i)}$ has row and column sums equal to either 0 or 1, and trace equal to 1. Finally, for each proposed clustering $C_1, \ldots, C_k$, the corresponding rank-one matrices are orthogonal in the trace inner product, due to the orthogonality of the characteristic vectors of the corresponding disjoint clusters. Thus, the matrix

$$\boldsymbol{X} = \sum_{i=1}^{k} \boldsymbol{X}^{(i)} = \sum_{i=1}^{k} \frac{\boldsymbol{v}_i \boldsymbol{v}_i^T}{\boldsymbol{v}_i^T \boldsymbol{v}_i} \tag{3}$$

has rank equal to k. This suggests that we may relax (2) as the rank-constrained semidefinite program

$$\max_{\boldsymbol{X} \in \Sigma_+^n} \left\{ \mathrm{Tr}(\boldsymbol{W}\boldsymbol{X}) : \boldsymbol{X}\boldsymbol{e} \leq \boldsymbol{e}, \mathrm{rank}\, \boldsymbol{X} = k, \mathrm{Tr}\, \boldsymbol{X} = k, \boldsymbol{X} \geq \boldsymbol{0} \right\}. \tag{4}$$

The relaxation (4) can be relaxed further to a semidefinite program by omitting the nonconvex rank constraint:

$$\max_{\boldsymbol{X} \in \Sigma_+^n} \left\{ \mathrm{Tr}(\boldsymbol{W}\boldsymbol{X}) : \boldsymbol{X}\boldsymbol{e} \leq \boldsymbol{e}, \mathrm{Tr}\, \boldsymbol{X} = k, \boldsymbol{X} \geq \boldsymbol{0} \right\}. \tag{5}$$

We should note that the semidefinite program (5) is remarkably similar to the semidefinite relaxation of the minimum sum of squared distance partition of Peng and Wei (2007) and the semidefinite relaxation of the maximum likelihood estimate of the stochastic block model considered by Amini and Levina (2014), among others, although our relaxation approach differs slightly from that used in these two papers.

2.2 Block Models and Recovery Guarantees

Given a set of clusterable data or, more accurately, a clusterable graph representation of data, Ames (2014) established that one can recover the underlying cluster structure from the optimal solution of the semidefinite program (5). Specifically, it is assumed that data with strong cluster structure should correspond to similarity graphs with heavy weight assigned to edges within clusters, relative to that between cluster edges. This corresponds to pairs of items within clusters being significantly more similar than pairs of items in different clusters. This motivates the following block model.

Let H^* be a k-disjoint-clique subgraph of $K_n = (V, \boldsymbol{W})$ with vertex set composed of the disjoint cliques $C_1, \ldots, C_k$ and let Σ^n denote the set of all $n \times n$ symmetric matrices. We consider weight matrices $\boldsymbol{W} = [w_{ij}] \in \Sigma^n$ with entries sampled independently from one of two probability distributions Ω_1, Ω_2 as follows.

- For each $i = 1, \ldots, k$ and each $u, v \in C_i$, we sample $w_{uv} = w_{vu}$ from a distribution Ω_1 such that
$$\mathbf{E}[w_{uv}] = \mathbf{E}[w_{vu}] = \alpha, \qquad 0 \leq w_{uv} \leq 1$$
for fixed $\alpha \in (0, 1]$.

- For each remaining edge uv, $u \in C_i, v \in C_j$, we sample the edge weight $w_{uv} = w_{vu}$ from a second distribution Ω_2 such that
$$\mathbf{E}[w_{uv}] = \mathbf{E}[w_{vu}] = \beta, \qquad 0 \leq w_{uv} \leq 1,$$
for fixed $\beta \in [0, \alpha)$ if $1 \leq i, j \leq k$ or $i = j = k + 1$ and, $\mathbf{E}[w_{uv}] = \beta/2$ otherwise.

We should note that the assumption that the entries of $\boldsymbol{W}$ are bounded between 0 and 1 is made for simplicity in the statement and proof of our main result; analogous recovery guarantees hold if we assume that random variables sampled according to Ω_1 and Ω_2 are bounded and nonnegative with high probability. We say that such random matrices $\boldsymbol{W}$ are sampled from the *planted cluster model*. Note that if $\boldsymbol{W}$ is sampled from the planted cluster model, then weight is concentrated on within-cluster edges (in expectation). This provides a natural generalization of the stochastic block model. Indeed, the stochastic block model corresponds to the planted cluster model in the special case that Ω_1 and Ω_2 are Bernoulli distributions with probabilities of success p and q, respectively. Ames (2014) established the following theorem, ensuring recovery of the planted cliques $C_1, \ldots, C_k$ from the optimal solution of (5) under the planted cluster model (see Ames, 2014, Theorem 2.1).

Theorem 2 *Suppose that the vertex sets $C_1, \ldots, C_k$ define a k-disjoint-clique subgraph H^* of the n-node weighted complete graph $K_n(V, \boldsymbol{W})$ and let $C_{k+1} := V \setminus \left(\cup_{i=1}^{t} C_i\right)$. Let $r_i := |C_i|$ for all $i = 1, \ldots, k+1$ and let $\hat{r} = \min_{i=1,\ldots,k} r_i$. Let $\boldsymbol{W} \in \Sigma^n$ be a random symmetric matrix sampled from the planted cluster model according to distributions Ω_1 and Ω_2 with means α and β, respectively, satisfying*
$$\gamma = \gamma(\alpha, \beta, r) := \alpha - \beta > 0.$$
Let $\boldsymbol{X}^*$ be the feasible solution of (5) corresponding to $C_1, \ldots, C_k$ defined by (3). Then there exist scalars $c_1, c_2, c_3 > 0$ such that if
$$c_1\sqrt{n} + c_2\sqrt{k r_{k+1}} + c_3 r_{k+1} \leq \gamma \hat{r}, \tag{6}$$
then $\boldsymbol{X}^*$ is the unique optimal solution of (5), and H^* is the unique maximum density k-disjoint-clique subgraph of K_n with probability tending exponentially to 1 as $\hat{r} \to \infty$.

In contrast to Theorem 1, the result of Theorem 2 implies that we can have perfect recovery if the graph contains a small number of nodes that shouldn't be assigned to any of the planted clusters. Each potential edge from each of these nodes to any other node is added independent to the graph with probability q, so that each node in C_{k+1} has roughly the same number of neighbours in each cluster block. This implies that such a node is not assigned to any of the planted clusters because it is weakly associated with all of the planted clusters. It is important to note that this edge assignment is performed randomly and not deterministically by an adversary attempting to obscure the cluster structure present in the graph. We present a new analysis that improves upon the recovery guarantee of Theorem 2 in two ways. First, the hypothesis of Theorem 2 assumes that between-cluster and within-cluster edge weights are i.i.d. We consider the more general heterogeneous case constructed as follows:

- For each $u \in C_i$, $v \in C_j$, we sample the edge weight $w_{uv} = w_{vu}$ from distribution Ω_{ij} with

$$\mathbf{E}[w_{uv}] = \mathbf{E}[w_{vu}] = \mu_{ij}, \qquad \mathrm{Var}[w_{uv}] = \mathrm{Var}[w_{vu}] = \sigma_{ij}^2, \qquad 0 \leq w_{uv} \leq 1.$$

This forces weights within the same block to be i.i.d., but weight may not be identically distributed in different blocks.

Second, the analysis leading to Theorem 2 assumes that the expectations of Ω_1, Ω_2 in the planted cluster model are fixed and that the variances are bounded by 1. We improve upon the recovery guarantee of Theorem 2 by considering the case where the parameters α and β depend on the number of nodes n in the graph. In particular, our recovery guarantees explicitly depend on the variances of the distributions Ω_{ij}, and their scaling with n, which will expand the set of graphs known to be clusterable by (5). We have the following theorem.

Theorem 3 *Suppose that the vertex sets $C_1, \ldots, C_k$ define a k-disjoint-clique subgraph K^* of the weighted complete graph $K_n = (V, \boldsymbol{W})$ on n vertices and let $C_{k+1} = V \setminus \left(\cup_{i=1}^k C_i\right)$. Let $r_i = |C_i|$ for all $i = 1, \ldots, k+1$ and let $\hat{r} = \min_{i=1,\ldots,k} r_i$. Let $\boldsymbol{W} \in \Sigma^n$ be a random symmetric matrix sampled from the heterogeneous planted cluster model according to distributions $\{\Omega_{ij}\}$ with expected values $\mu_{ij} = \mu_{ij}(n)$ and variances $\sigma_{ij}^2 = \sigma_{ij}^2(n)$. Let $\tilde{\sigma} := \max_{q,s} \sigma_{qs}$ and $\hat{\sigma} := \max_q \sigma_{qq}$. Let $\boldsymbol{X}^*$ be the feasible solution to (5) corresponding to $C_1, \ldots, C_k$ defined by (3). Let*

$$\gamma := \min_{\substack{q,s=1,2,\ldots,k \\ q \neq s}} \{\mu_{qq} - \mu_{qs}\}.$$

Then there exists scalar $c > 0$ such that if

$$\gamma \hat{r} \geq c \max\left\{\sqrt{\tilde{\sigma}^2 n},\ \sqrt{\tilde{\sigma}^2 \hat{r} \log n},\ \sqrt{\hat{\sigma}^2 k r_{k+1}},\ \sqrt{k r_{k+1} \log n / \hat{r}},\ \mu_{k+1,k+1} r_{k+1},\ \log n\right\} \quad (7)$$

then $\boldsymbol{X}^*$ *is the unique optimal solution for* (5)*, and* K^* *is the unique maximum density* k*-disjoint-clique subgraph of* K_n *with high probability.*

The *weak assortativity* condition (7) implies that we have perfect recovery provided that the gap between the cluster block expectation μ_{qq} and the largest between-cluster block expectation μ_{qs} is sufficiently large for all clusters C_q, $q = 1, \ldots, k$, relative to the minimum cluster size, number of unassigned nodes r_{k+1}, number of clusters, and edge weight variances. In the Bernoulli case, i.e., within-cluster and between-cluster edges are added independently with probabilities p and q, respectively, Theorem 3 and, in particular, (7) establish that we can recover the planted clusters provided that

$$\frac{(p-q)^2}{\tilde{\sigma}^2} = \frac{(p-q)^2}{\max\{p(1-p), q(1-q)\}} = \Omega\left(\frac{n}{\hat{r}^2}\right).$$

This result agrees with the Easy Regime for cluster recovery proposed by Chen and Xu (2014), where a polynomial-time algorithm exists for exact recovery of the planted clusters, in this case, the solution of the semidefinite relaxation (5). One distinct advantage of this result over similar recovery guarantees is that our model and phase transition are largely parameter free. For example, Amini and Levina (2014) present an analysis of three semidefinite relaxations that obtain nearly identical conditions on $\{\Omega_{ij}\}$ guaranteeing recovery but restrict their analysis to case where the clusters are identical in size or otherwise known and when $\{\Omega_{ij}\}$ are Bernoulli distributions; we should note that Amini and Levina (2014) consider heterogeneous Bernoulli distributions where the within-cluster and between-cluster probabilities of adding an edge vary across clusters. Similarly, Chen and Xu (2014) and Jalali et al. (2015) give identical conditions for recovery (up to constants and logarithmic terms) in the Bernoulli case to those in Theorem (3) for semidefinite relaxations that require the sizes of the clusters to be used as input parameters (or all clusters to have identical size), neither of which are realistic assumptions in practice. In contrast, our approach achieves this recovery guarantee using *only the desired number of clusters as a parameter*. Further, our guarantee extends to the general weighted case where the vast majority of existing recovery guarantees for stochastic block models are restricted to the Bernoulli case.

It is important to note that tighter recovery guarantees than those provided by Theorem 3 are known for specific problem settings. This is a natural consequence of the more general framework of our analysis. For example, Yan et al. (2017), studies a convex relaxation for cluster recovery in the Bernoulli (unweighted) case. The main theorem of this article establishes conditions for perfect recovery that allows larger clusters to have higher variance, although specialized for the unweighted case. Moreover, Yan et al. (2017) consider the use of a tuning parameter to allow recovery without knowledge of the number of clusters k. On the other hand, the results of (Amini and Levina, 2014; Jalali et al., 2015) also provide tigher

recovery guarantees but require knowledge of cluster sizes. The key contribution of this work is the presentation of a recovery guarantee that extends to the weighted case without strict assumptions regarding input parameters, as well as the first-order method for solution of (1) discussed in detail Section 4.

To further illustrate the consequences of this theorem, we consider several examples. In each, we assume that the graph is generated in the homogeneous setting where within-cluster weights are i.i.d. according to Ω_1 with mean α and variance σ_1^2, and between-cluster weights are i.i.d. according to Ω_2 with mean $\beta < \alpha$ and variance σ_2^2.

2.2.1 The Dense Case

When α, β are fixed, we obtain the same recovery guarantee as before, up to constants and logarithmic terms: we have exact recovery w.h.p. if $\hat{r} \geq \tilde{c}_1\sqrt{n}$ and $\hat{r} \geq \tilde{c}_2 r_{k+1}$ for some constants $\tilde{c}_1$ and $\tilde{c}_2$ depending on Ω_1 and Ω_2. Indeed, each of the pointwise maximums in the first three terms of (7) is bounded above by $O(\sqrt{n})$ since $\tilde{r} \leq n$, and $kr_{k+1} = O(n)$ if $r_{k+1} = O(\hat{r})$.

2.2.2 The Sparse Case

On the other hand, if noise in the form of between-cluster edge-weight is small, then we should expect to be able recover much smaller clusters. For example, suppose that Ω_2 is the Bernoulli distribution with probability of adding an edge q and that Ω_1 is the Bernoulli distribution with probability of adding an edge $p = 1$; the assumption that $p = 1$ is for the sake of simplicity in this example and we can expect analogous recovery guarantees for any p tending slowly enough to 0. Assume further that $q(1-q) \leq \log n/n$. Finally, again for simplicity, assume that we have k equally sized clusters of size $\hat{r} = n/k$ and $(r_{k+1} = 0)$. In this case, (7) holds if

$$\gamma\hat{r} \geq c\log n = c\max\left\{\sqrt{\log n},\ \sqrt{\frac{\hat{r}\log^2 n}{n}},\ \log n\right\} \geq c\max\left\{\sqrt{\tilde{\sigma}^2 n},\ \sqrt{\tilde{\sigma}^2 r\log n},\ \log n\right\}$$

since $\tilde{\sigma}^2 = \max\{p(1-p), q(1-q)\} = q(1-q) \leq \log n/n$ and the terms involving r_{k+1} and $1-p$ are equal to zero. This implies that we have exact recovery of the planted clusters w.h.p. provided $\hat{r} = \Omega(\log n)$. This exceeds the state of the art recovery bound of $\hat{r} = \Omega(\sqrt{\log n})$ established in Jalali et al. (2015) by a factor of $\sqrt{\log n}$. However, the convex relaxation proposed by Jalali et al. (2015) requires knowledge of $\sum_{i=1}^{k} r_i^2$, which is often an unrealistic expectation in practice; in contrast, our approach only requires knowledge of the number of clusters k present in the data. Further, the requirement $\hat{r} = \Omega(\log n)$ is enforced

by the gap inequality (7), which itself is a consequence of the use of the Bernstein inequality to establish certain dual variables are nonnegative in the proof of Theorem 3 (see Section 3.1 for more details); it may be possible to improve this bound to $\hat{r} = \Omega(\sqrt{\log n})$ with improved concentration inequalities but it is unclear what form these improvements may take.

2.2.3 The Planted Clique and Sparsest Subgraph

In the special case when $k = 1$ and Ω_1 and Ω_2 are Bernoulli distributions, the planted cluster model specializes to the planted clique model considered in Ames and Vavasis (2011) and Ames (2015). In this case, (7) suggests that we can recover a planted clique (in the dense case) of size $r_1 = \Omega\left(\max\left\{\sqrt{n}, n - r_1\right\}\right)$. This recovery guarantee is far more conservative than those provided in Ames and Vavasis (2011) and Ames (2015), among others, which establish that a planted clique of size $\Omega(\sqrt{n})$ can be recovered from the optimal solution of a particular nuclear norm relaxation of the maximum clique problem.

Unfortunately, it appears that this lower bound restricting the size of a recoverable planted clique to a constant multiple of the number of nonclique vertices is tight. For example, let p and q be the probabilities of adding an edge given by Ω_1 and Ω_2. Then the expected value of the proposed solution $\boldsymbol{X}^*$ in (5) is equal to

$$\mathbf{E}[\mathrm{Tr}(\boldsymbol{W}\boldsymbol{X}^*)] = \frac{1}{\hat{r}} \sum_{i \in C_1} \sum_{j \in C_1} \mathbf{E}[w_{ij}] = p\hat{r}.$$

On the other hand, the solution $\frac{1}{n}\boldsymbol{e}\boldsymbol{e}^T$ is also feasible for (5) with expected objective value

$$\mathbf{E}\left[\frac{1}{n}\,\mathrm{Tr}(\boldsymbol{W}\boldsymbol{e}\boldsymbol{e}^T)\right] \geq \hat{c}qn$$

for some constant $\hat{c}$. This implies that the proposed solution is suboptimal if $p\hat{r} < \hat{c}qn$, which holds unless $\hat{r} \geq \hat{c}(q/p)n$. We will see that the realized values of these sums are concentrated near their expectations and, thus, we cannot reasonably expect to recover planted clusters with unassigned nodes significantly outnumbering the smallest cluster. This implies that we cannot recover planted cliques of size $\omega(N)$ by maximizing density of a complete subgraph because the planted clique is not the index set of the densest such graph; in this case, the entire graph is a denser complete subgraph, as measured by average vertex degree, in expectation.

3. Derivation of the Recovery Guarantee

In this section, we show that if the hypothesis of Theorem 3 is satisfied then the solution $\boldsymbol{X}^*$ constructed according to (3) is optimal for (5) and the corresponding k-disjoint-clique

subgraph has maximum density. In particular, we will show that $\boldsymbol{X}^*$ satisfies the following sufficient condition for optimality of a feasible solution of (5) (see Ames, 2014, Theorem 4.1).

Theorem 4 *Let $\boldsymbol{X}$ be feasible for (5) and suppose that there exist some $\tau \in \mathbf{R}$, $\boldsymbol{\lambda} \in \mathbf{R}^n_+$, $\boldsymbol{\Xi} \in \mathbf{R}^{n\times n}_+$ and $\boldsymbol{S} \in \Sigma^n_+$ such that*

$$-\boldsymbol{W} + \boldsymbol{\lambda}\boldsymbol{e}^T + \boldsymbol{e}\boldsymbol{\lambda}^T - \boldsymbol{\Xi} + \tau\boldsymbol{I} = \boldsymbol{S} \tag{8}$$

$$\boldsymbol{\lambda}^T(\boldsymbol{X}\boldsymbol{e} - \boldsymbol{e}) = 0 \tag{9}$$

$$\mathrm{Tr}(\boldsymbol{X}\boldsymbol{\Xi}) = 0 \tag{10}$$

$$\mathrm{Tr}(\boldsymbol{X}\boldsymbol{S}) = 0. \tag{11}$$

Then $\boldsymbol{X}$ is optimal for (5).

Theorem 4 is a restriction of the Karush-Kuhn-Tucker optimality conditions to the semidefinite program (5) (see for example Boyd and Vandenberghe, 2009, Section 5.5.3). The goal of this section is to establish that we can construct dual variables $\tau \in \mathbf{R}$, $\boldsymbol{\lambda} \in \mathbf{R}^n_+$, $\boldsymbol{\Xi} \in \mathbf{R}^{n\times n}_+$ and $\boldsymbol{S} \in \Sigma^n_+$ which satisfy the hypothesis of Theorem 4 with high probability if the weight matrix $\boldsymbol{W}$ is sampled from a distribution of clusterable block models. To motivate our proposed choice of dual variables, we note that the complementary slackness condition $\mathrm{Tr}(\boldsymbol{X}\boldsymbol{S}) = 0$ holds if and only if $\boldsymbol{X}\boldsymbol{S} = \boldsymbol{0}$ under the assumption that both $\boldsymbol{X}$ and $\boldsymbol{S}$ are positive semidefinite. Therefore, the block structure of $\boldsymbol{X}$ implies that each block of $\boldsymbol{S}$ corresponding to a cluster block in $\boldsymbol{W}$ must sum to zero.

Before we continue with the construction of our dual variables, let us first remind ourselves of the notation of Theorem 3. Let K^* be a k-disjoint-clique subgraph of K_n with vertex set composed of the disjoint cliques $C_1, \ldots, C_k$ of sizes $r_1, \ldots, r_k$ and let $\boldsymbol{X}^*$ be the corresponding feasible solution of (5) defined by (3). Let $C_{k+1} := V \setminus (\cup_{i=1}^k C_i)$ and $r_{k+1} := n - \sum_{i=1}^k r_i$ be the size of of C_{k+1}. Moreover, let $\hat{r} := \min_{i=1,\ldots,k} r_i$ and $\tilde{r} := \max_{i=1,\ldots,k} r_i$ be the size of the smallest and largest clusters, respectively. Let $\boldsymbol{W} \in \Sigma^n$ be a random symmetric matrix sampled from the planted cluster model with planted clusters $C_1, \ldots, C_k$ and remaining nodes C_{k+1} according to the distributions $\{\Omega_{ij}\}$ with means $\{\mu_{ij}\}$ and variances $\{\sigma^2_{ij}\}$.

We now propose a choice of dual variables satisfying the complementary slackness condition $\boldsymbol{X}\boldsymbol{S} = \boldsymbol{0}$. Restricting this condition to the blocks $\boldsymbol{X}_{C_q,C_q}$ and $\boldsymbol{S}_{C_q,C_q}$ of $\boldsymbol{X}$ and $\boldsymbol{S}$ with rows and columns indexed by C_q, $q \in \{1, 2, \ldots, k\}$, we see that $\boldsymbol{X}^*\boldsymbol{S} = \boldsymbol{0}$ holds if and only if

$$\boldsymbol{0} = \boldsymbol{S}_{C_q,C_q}\boldsymbol{e} = \tau\boldsymbol{e} + r_q\boldsymbol{\lambda}_{C_q} + (\boldsymbol{\lambda}^T_{C_q}\boldsymbol{e})\boldsymbol{e} - \boldsymbol{W}_{C_q,C_q}\boldsymbol{e}$$

by the block structure of $\boldsymbol{X}^*$; note that $\boldsymbol{\Xi}_{C_q,C_q} = \mathbf{0}$ is chosen to satisfy the complementary slackness condition (10). Solving this linear system for $\boldsymbol{\lambda}_{C_q}$ using the Sherman-Morrison-Woodbury Formula (Golub and Van Loan, 2013, Equation (2.1.4)) gives

$$\boldsymbol{\lambda}_{C_q} = \frac{1}{r_q}\left(\boldsymbol{W}_{C_q,C_q}\boldsymbol{e} - \frac{1}{2}\left(\tau + \frac{\boldsymbol{e}^T\boldsymbol{W}_{C_q,C_q}\boldsymbol{e}}{r_q}\right)\boldsymbol{e}\right). \tag{12}$$

On the other hand, we choose $\boldsymbol{\lambda}_{C_{k+1}} = \mathbf{0}$ to satisfy the complementary slackness condition (9). Next, we use this choice of $\boldsymbol{\lambda}$ to construct the remaining dual variables.

Fix $q, s \in \{1, 2, \ldots, k+1\}$ such that $q \neq s$. We will choose $\boldsymbol{\Xi}_{C_q,C_s}$ so that $\boldsymbol{S}_{C_q,C_s}\boldsymbol{e} = \mathbf{0}$ and $\boldsymbol{S}_{C_s,C_q}\boldsymbol{e} = \mathbf{0}$. In particular, we choose

$$\boldsymbol{\Xi}_{C_q,C_s} = \left(\tfrac{1-\delta_{q,k+1}}{2}\left(\mu_{qq} - \tfrac{\tau}{r_q}\right) + \tfrac{1-\delta_{s,k+1}}{2}\left(\mu_{ss} - \tfrac{\tau}{r_s}\right) - \mu_{qs}\right)\boldsymbol{e}\boldsymbol{e}^T + \boldsymbol{y}^{q,s}\boldsymbol{e}^T + \boldsymbol{e}(\boldsymbol{z}^{q,s})^T, \tag{13}$$

where the vectors $\boldsymbol{y}^{q,s}$ and $\boldsymbol{z}^{q,s}$ are unknown vectors parametrizing the entries of $\boldsymbol{\Xi}_{C_q,C_s}$; here $\delta_{i,j}$ is the Kronecker delta function defined by $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise. That is, we choose $\boldsymbol{\Xi}_{C_q,C_s}$ to be the expected value of $\boldsymbol{\lambda}_{C_q}\boldsymbol{e}^T + \boldsymbol{e}\boldsymbol{\lambda}_{C_s}^T - \boldsymbol{W}_{C_q,C_s}$ plus the parametrizing term $\boldsymbol{y}^{q,s}\boldsymbol{e}^T + \boldsymbol{e}(\boldsymbol{z}^{q,s})^T$; the vectors $\boldsymbol{y}^{q,s}$ and $\boldsymbol{z}^{q,s}$ are chosen to be solutions of the systems of linear equations given by the complementary slackness conditions $\boldsymbol{S}_{C_q,C_s}\boldsymbol{e} = \mathbf{0}$ and $\boldsymbol{S}_{C_s,C_q}\boldsymbol{e} = \mathbf{0}$. It is reasonably straight-forward to show that we may choose

$$\boldsymbol{y}^{q,s} = \frac{1}{r_s}\left(\boldsymbol{b}_{q,s} - \frac{\boldsymbol{b}_{q,s}^T\boldsymbol{e}}{r_q + r_s}\boldsymbol{e}\right) \qquad \boldsymbol{z}^{q,s} = \frac{1}{r_q}\left(\boldsymbol{b}_{s,q} - \frac{\boldsymbol{b}_{s,q}^T\boldsymbol{e}}{r_q + r_s}\boldsymbol{e}\right), \tag{14}$$

where

$$\boldsymbol{b}_{q,s} = \left(\boldsymbol{\lambda}_{C_q}\boldsymbol{e}^T + \boldsymbol{e}\boldsymbol{\lambda}_{C_s}^T - \boldsymbol{W}_{C_q,C_s} - \mathbf{E}\left[\boldsymbol{\lambda}_{C_q}\boldsymbol{e}^T + \boldsymbol{e}\boldsymbol{\lambda}_{C_s}^T - \boldsymbol{W}_{C_q,C_s}\right]\right)\boldsymbol{e}. \tag{15}$$

Indeed, we must choose $\boldsymbol{y} = \boldsymbol{y}^{q,s}$ and $\boldsymbol{z} = \boldsymbol{z}^{q,s}$ to be solutions of the system

$$\begin{pmatrix} r_s\boldsymbol{I} + \boldsymbol{e}\boldsymbol{e}^T & 0 \\ 0 & r_q\boldsymbol{I} + \boldsymbol{e}\boldsymbol{e}^T \end{pmatrix}\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix} = \begin{pmatrix} \boldsymbol{b}^{q,s} \\ \boldsymbol{b}_{s,q} \end{pmatrix} \tag{16}$$

to ensure that the complementary slackness conditions are satisfied. Note that taking the inner product of each side of (16) with the vector $(\boldsymbol{e}; -\boldsymbol{e})$ yields

$$(r_q + r_s)(\boldsymbol{e}^T\boldsymbol{y} - \boldsymbol{e}^T\boldsymbol{z}) = \boldsymbol{e}^T\boldsymbol{b}^{q,s} - \boldsymbol{e}^T\boldsymbol{b}^{s,q} = 0$$

by the symmetry of $\boldsymbol{W}$. This establishes that the solution $(\boldsymbol{y}; \boldsymbol{z})$ of (16) is also a solution of the (singular) system of equations

$$\begin{pmatrix} r_s\boldsymbol{I} & \boldsymbol{e}\boldsymbol{e}^T \\ \boldsymbol{e}\boldsymbol{e}^T & r_q\boldsymbol{I} \end{pmatrix}\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix} = \begin{pmatrix} \boldsymbol{b}^{q,s} \\ \boldsymbol{b}_{s,q} \end{pmatrix}$$

imposed by the complementary slackness conditions $\boldsymbol{S}_{C_q,C_s}\boldsymbol{e} = \boldsymbol{0}$ and $\boldsymbol{S}_{C_s,C_q} = \boldsymbol{0}$. Solving (16) for $\boldsymbol{y}$ and $\boldsymbol{z}$ using the Sherman-Morrison-Woodbury Formula yields the formula for $\boldsymbol{y}$ and $\boldsymbol{z}$ given by (14). We set the remaining block $\boldsymbol{\Xi}_{C_{k+1},C_{k+1}} = \boldsymbol{0}$. See Ames (2014, Section 4.2) for more details.

Finally, we choose

$$\tau = \min_{\substack{q,s=1,\ldots,k \\ q\neq s}} \{\mu_{qq} - \mu_{qs}\}\, \epsilon\hat{r} =: \gamma\epsilon\hat{r}, \tag{17}$$

where $\epsilon > 0$ is a parameter to be chosen later. In particular, the analysis provided in Sections 3.1, 3.2, and 3.3 establishes that a suitable choice of ϵ exists if the hypothesis of Theorem 3 is satisfied.

The entries of $\boldsymbol{S}$ are chosen according to the stationarity condition (8), but we will also define an auxiliary variable $\tilde{\boldsymbol{S}} \in \Sigma^n$ as the following $(k+1)\times(k+1)$ block matrix:

$$\tilde{\boldsymbol{S}}_{C_q,C_s} = \begin{cases} \mu_{q,s}\boldsymbol{e}\boldsymbol{e}^T - \boldsymbol{W}_{C_q,C_s}, & \text{if } q,s \in \{1,\ldots,k\} \\ \mu_{q,k+1}\boldsymbol{e}\boldsymbol{e}^T - \boldsymbol{W}_{C_q,C_{k+1}} + (\boldsymbol{\lambda}_{C_q} - \mathbf{E}[\boldsymbol{\lambda}_{C_q}])\boldsymbol{e}^T, & \text{if } s = k+1 \\ \mu_{k+1,s}\boldsymbol{e}\boldsymbol{e}^T - \boldsymbol{W}_{C_{k+1},C_s} + \boldsymbol{e}(\boldsymbol{\lambda}_{C_s} - \mathbf{E}[\boldsymbol{\lambda}_{C_s}])^T, & \text{if } q = k+1. \end{cases} \tag{18}$$

We next provide the following theorem, first stated by Ames (2014, Theorem 4.2), which characterizes when the proposed dual variables satisfy the hypothesis of Theorem 4.

Theorem 5 *Suppose that the vertex sets $C_1,\ldots,C_k$ define a k-disjoint-clique subgraph K^* of the weighted complete graph $K_n = (V, \boldsymbol{W})$, where $\boldsymbol{W} \in \Sigma^n$ is a random symmetric matrix sampled from the planted cluster model according to the distributions $\{\Omega_{ij}\}$ with means $\{\mu_{ij}\}$ and variances $\{\sigma_{ij}^2\}$. Let $r_1,\ldots,r_{k+1}$, and $\hat{r}$ be defined as in Theorem 3. Let $\boldsymbol{X}^*$ be the feasible solution for (5) corresponding to $C_1,\ldots,C_k$ defined by (3). Let $\tau \in \mathbf{R}$, $\boldsymbol{\lambda} \in \mathbf{R}^n$, and $\boldsymbol{\Xi} \in \mathbf{R}^{n\times n}$ be chosen according to (12), (13), and (17), and let $\tilde{\boldsymbol{S}}$ be chosen according to (18). Suppose that the entries of $\boldsymbol{\lambda}$ and $\boldsymbol{\Xi}$ are nonnegative. Then $\boldsymbol{X}^*$ is optimal for (5), and K^* is the maximum density k-disjoint-clique subgraph of K_n corresponding to $\boldsymbol{W}$ if*

$$\|\tilde{\boldsymbol{S}}\| \leq \epsilon\gamma\hat{r}. \tag{19}$$

Moreover, if (19) is satisfied and

$$r_s\boldsymbol{e}^T\boldsymbol{W}_{C_q,C_q}\boldsymbol{e} > r_q\boldsymbol{e}^T\boldsymbol{W}_{C_q,C_s}\boldsymbol{e} \tag{20}$$

for all $q,s \in \{1,\ldots,k\}$ such that $q \neq s$, then $\boldsymbol{X}^*$ is the unique optimal solution of (5) and K^* is the unique maximum density k-disjoint-clique subgraph of K_n.

The proof of Theorem 5 is nearly identical to that by Ames (2014, Theorem 4.2), and is omitted. Theorem 5 provides a clear roadmap for the remainder of the proof; if we can

show that if $\boldsymbol{W}$ is sampled from the planted cluster model satisfying (7) then $\boldsymbol{\lambda}$ and $\boldsymbol{\Xi}$ are nonnegative and $\|\tilde{\boldsymbol{S}}\| \leq \epsilon\gamma\hat{r}$ with high probability, then we will have established that we can recover the underlying block structure with high probability in this case. We establish the necessary bounds on $\boldsymbol{\lambda}$, $\boldsymbol{\Xi}$, and $\|\tilde{\boldsymbol{S}}\|$ in the following sections.

3.1 Nonnegativity of $\boldsymbol{\lambda}$ and $\boldsymbol{\Xi}$

We first establish that the entries of $\boldsymbol{\Xi}$, as constructed according to (13), are nonnegative with high probability. To do so, we will make repeated use of the following specialization of the Bernstein inequality (see, for example, Lugosi (2009, Theorem 6)), which provides a bound on the tail of a sum of bounded independent random variables.

Theorem 6 *Let $x_1, \ldots, x_m$ be independent identically distributed (i.i.d.) variables with mean μ and variance σ^2. Let $S = x_1 + \cdots + x_m$. Then*

$$Pr\left(|S - \mu m| > 6\max\left\{\sqrt{\sigma^2 m \log T}, \log T\right\}\right) \leq 2T^{-6}, \tag{21}$$

for all $T > 1$.

The following bound on the parametrizing vectors $\boldsymbol{y}^{q,s}$ and $\boldsymbol{z}^{q,s}$ in the definition of the (C_q, C_s) block of $\boldsymbol{\Xi}$, c.f., (13), is an immediate consequence of Theorem 6.

Lemma 7 *There exists constant $c > 0$ such that*

$$\|\boldsymbol{y}^{q,s}\|_\infty + \|\boldsymbol{z}^{q,s}\|_\infty \leq c\max\left\{\sqrt{\frac{\tilde{\sigma}^2 \log n}{\hat{r}}}, \frac{\log n}{\hat{r}}\right\} \tag{22}$$

w.h.p., where $\tilde{\sigma} := \max\{\sigma_{ij} : i, j = 1, 2, \ldots, k+1\}$, for all $q, s \in \{1, \ldots, k+1\}$ such that $q \neq s$

For $q, s \in \{1, \ldots, k+1\}$ such that $q \neq s$, we define $\boldsymbol{y}^{q,s}$ and $\boldsymbol{z}^{q,s}$ as in (14). To bound the absolute values of the entries of $\boldsymbol{y}^{q,s}$ and $\boldsymbol{z}^{q,s}$, we must estimate the sums $\boldsymbol{e}^T\boldsymbol{W}_{C_q,C_q}\boldsymbol{e}$, $\boldsymbol{e}^T\boldsymbol{W}_{C_s,C_s}\boldsymbol{e}$ and $\boldsymbol{e}^T\boldsymbol{W}_{C_q,C_s}\boldsymbol{e}$; applying Theorem 6 to bound the tails of these sums yields Lemma 7. See Appendix A for the full argument.

We have the following bound on the entries of $\boldsymbol{\Xi}$ as an immediate consequence of Lemma 7.

Proposition 8 *Suppose that $\{\mu_{ij}\}$ satisfy (7). Then there exists constant $c > 0$ such that each entry of $\boldsymbol{\Xi}$ is nonnegative w.h.p. if ϵ satisfies*

$$0 < \epsilon \leq 1 - \frac{c}{\gamma}\max\left\{\sqrt{\frac{\tilde{\sigma}^2 \log n}{\hat{r}}}, \frac{\log n}{\hat{r}}\right\} \tag{23}$$

Proof Fix $q, s \in \{1, \dots, k\}$ such that $q \neq s$. By construction, we have

$$\begin{aligned}\boldsymbol{\Xi}_{C_q,C_s} &= \mathbf{E}\left[\boldsymbol{\lambda}_{C_q}\boldsymbol{e}^T + \boldsymbol{e}\boldsymbol{\lambda}_{C_s}^T - \boldsymbol{W}_{C_q,C_s}\right] + \boldsymbol{y}^{q,s}\boldsymbol{e}^T + \boldsymbol{e}\left(\boldsymbol{z}^{q,s}\right)^T \\ &= \left(\frac{1}{2}\left(\mu_{qq} - \frac{\tau}{r_q}\right) + \frac{1}{2}\left(\mu_{ss} - \frac{\tau}{r_s}\right) - \mu_{qs}\right)\boldsymbol{e}\boldsymbol{e}^T + \boldsymbol{y}^{q,s}\boldsymbol{e}^T + \boldsymbol{e}\left(\boldsymbol{z}^{q,s}\right)^T\end{aligned}$$

Using (17) and Lemma 7, we see that

$$\begin{aligned}\Xi_{ij} &\geq \frac{1}{2}\left(\mu_{qq} - \gamma\epsilon\right) + \frac{1}{2}\left(\mu_{ss} - \gamma\epsilon\right) - \mu_{sq} - \|\boldsymbol{y}^{q,s}\|_\infty - \|\boldsymbol{z}^{q,s}\|_\infty \\ &\geq (1-\epsilon)\gamma - c\max\left\{\sqrt{\frac{\tilde{\sigma}^2\log n}{\hat{r}}}, \frac{\log n}{\hat{r}}\right\},\end{aligned}$$

for all $i \in C_q$, $j \in C_s$ w.h.p., where c is the constant appearing in Lemma 7. Note that the right-hand side of this inequality is nonnegative if and only if

$$\epsilon \leq 1 - \frac{c}{\gamma}\max\left\{\sqrt{\frac{\tilde{\sigma}^2\log n}{\hat{r}}}, \frac{\log n}{\hat{r}}\right\}.$$

The argument for the case when one of q or s is equal to $k+1$ follows analogously. Applying the union bound over all blocks of $\boldsymbol{\Xi}$ shows that each entry of $\boldsymbol{\Xi}$ is nonnegative w.h.p. if ϵ satisfies (23). ■

We have an analogous result ensuring that the entries of $\boldsymbol{\lambda}$ are nonnegative with high probability; we present the proof of this result in Appendix B.

Proposition 9 *Suppose $\{\mu_{ij}\}$ satisfy (7). Then there exists constant $c' > 0$ such that each entry of $\boldsymbol{\lambda}$ is nonnegative w.h.p. if ϵ satisfies*

$$0 < \epsilon \leq \frac{1}{2\gamma}\left(\mu_{qq} - c'\max\left\{\sqrt{\frac{\sigma_{qq}^2\log n}{r_q}}, \frac{\log n}{r_q}\right\}\right) \tag{24}$$

for all $q \in \{1, \dots, k\}$.

We conclude this section with a result ensuring that the uniqueness condition (20) of Theorem 5 is satisfied for all $q, s \in \{1, \dots, k\}$ such that $q \neq s$; we provide a proof in Appendix C.

Proposition 10 *Suppose that*

$$\gamma \geq 12\max\left\{\sqrt{\frac{\tilde{\sigma}^2\log n}{\hat{r}^2}}, \frac{\log n}{\hat{r}^2}\right\}. \tag{25}$$

Then $r_s\boldsymbol{e}^T\boldsymbol{W}_{C_q,C_q}\boldsymbol{e} > r_q\boldsymbol{e}^T\boldsymbol{W}_{C_q,C_s}\boldsymbol{e}$ for all $q, s \in \{1, \dots, k\}$ such that $q \neq s$ with high probability.

3.2 A Bound on $\tilde{S}$

It remains to establish the following bound on the spectral norm of the matrix $\tilde{\boldsymbol{S}}$.

Proposition 11 *There exists scalars $C, C' > 0$ such that*

$$\|\tilde{\boldsymbol{S}}\| \leq C \max\left\{\tilde{\sigma}\sqrt{n}, \sqrt{\log n}\right\} + C'\left(\max\left\{\hat{\sigma}^2, \frac{\log n}{\hat{r}}\right\} k r_{k+1}\right)^{1/2} + \mu_{k+1,k+1} r_{k+1}, \quad (26)$$

where $\hat{\sigma}^2 = \max_{q=1,\ldots,k}\left\{\sigma_{qq}^2\right\}$, with high probability.

The proof of Proposition 11 follows the same structure as that of Ames (2014, Lemma 4.5). In particular, we decompose $\tilde{\boldsymbol{S}}$ as $\tilde{\boldsymbol{S}} = \tilde{\boldsymbol{S}}_1 + \tilde{\boldsymbol{S}}_2 + \tilde{\boldsymbol{S}}_3$, where

$$\tilde{\boldsymbol{S}}_1 = \mathbf{E}[\boldsymbol{W}] - \boldsymbol{W}, \quad (27)$$

$$[\tilde{\boldsymbol{S}}_2]_{C_q,C_s} = \begin{cases} \left(\boldsymbol{\lambda}_{C_q} - \mathbf{E}\left[\boldsymbol{\lambda}_{C_q}\right]\right)\boldsymbol{e}^T, & \text{if } s = k+1 \\ \boldsymbol{e}\left(\boldsymbol{\lambda}_{C_s} - \mathbf{E}\left[\boldsymbol{\lambda}_{C_s}\right]\right)^T, & \text{if } q = k+1 \\ \mathbf{0}, & \text{otherwise,} \end{cases} \quad (28)$$

$$[\tilde{\boldsymbol{S}}_3]_{C_q,C_s} = \begin{cases} -\mu_{k+1,k+1}\boldsymbol{e}\boldsymbol{e}^T, & \text{if } q = s = k+1 \\ \mathbf{0}, & \text{otherwise.} \end{cases} \quad (29)$$

Note that $\|\tilde{\boldsymbol{S}}_3\| = \mu_{k+1,k+1}\|\boldsymbol{e}\boldsymbol{e}^T\| = \mu_{k+1,k+1} r_{k+1}$. The following lemmas provide the necessary bounds on $\|\tilde{\boldsymbol{S}}_1\|$ and $\|\tilde{\boldsymbol{S}}_2\|$.

Lemma 12 *Suppose that $\tilde{\boldsymbol{S}}_1$ is constructed according to (27) for some $\boldsymbol{W} \in \Sigma^n$ sampled from the heterogeneous planted cluster model. Then there exists constant $C > 0$ such that*

$$\|\tilde{\boldsymbol{S}}_1\| \leq C \max\left\{\tilde{\sigma}\sqrt{n}, \sqrt{\log n}\right\} \quad (30)$$

with high probability.

Lemma 13 *Suppose that $\tilde{\boldsymbol{S}}_2$ is constructed according to (28) for some $\boldsymbol{W} \in \Sigma^n$ sampled from the heterogeneous planted cluster model. Then there exists constant $C' > 0$ such that*

$$\|\tilde{\boldsymbol{S}}_2\| \leq C'\left(\max\left\{\hat{\sigma}^2, \frac{\log n}{\hat{r}}\right\} k r_{k+1}\right)^{1/2} \quad (31)$$

with high probability, where $\hat{\sigma} := \max_{q=1,\ldots,k} \sigma_{qq}$.

We delay the proof of Lemmas 12 and 13 until Appendix D and Appendix E respectively. Combining the three bounds on $\|\tilde{\boldsymbol{S}}_1\|$, $\|\tilde{\boldsymbol{S}}_2\|$, and $\|\tilde{\boldsymbol{S}}_3\|$ and applying the triangle inequality one last time shows that (26) holds with high probability.

3.3 The Conclusion of the Proof

According to Theorem 5, it suffices to prove that $\|\tilde{\boldsymbol{S}}\| \leq \epsilon\gamma\hat{r}$ is satisfied with high probability in order to prove Theorem 3. According to Proposition 11, if

$$\gamma\epsilon\hat{r} \geq C\max\left\{\tilde{\sigma}\sqrt{n}, \sqrt{\log n}\right\} + C'\left(\max\left\{\hat{\sigma}^2, \frac{\log n}{\hat{r}}\right\} kr_{k+1}\right)^{1/2} + \mu_{k+1,k+1}r_{k+1}, \quad (32)$$

then $\|\tilde{\boldsymbol{S}}\| \leq \epsilon\gamma\hat{r}$ holds with high probability. Hence, we have three conditions, (23), (24) and (32), on $\epsilon > 0$ that need to be satisfied simultaneously; choosing any $\epsilon > 0$ satisfying all three establishes the desired recovery guarantee. We see that (23) and (32) can be simultaneously fulfilled if

$$1 - \tfrac{c}{\gamma}\max\left\{\sqrt{\tfrac{\tilde{\sigma}^2\log n}{\hat{r}}}, \tfrac{\log n}{\hat{r}}\right\} \geq \tfrac{1}{\gamma\hat{r}}\left(C\max\left\{\tilde{\sigma}\sqrt{n}, \sqrt{\log n}\right\} + C'\left(\max\left\{\hat{\sigma}^2, \tfrac{\log n}{\hat{r}}\right\} kr_{k+1}\right)^{1/2} + \mu_{k+1,k+1}r_{k+1}\right)$$

which holds if and only if

$$\hat{r}\left(\gamma - c\max\left\{\sqrt{\tfrac{\tilde{\sigma}^2\log n}{\hat{r}}}, \tfrac{\log n}{\hat{r}}\right\}\right) \geq C\max\left\{\tilde{\sigma}\sqrt{n}, \sqrt{\log n}\right\} + C'\left(\max\left\{\hat{\sigma}^2, \tfrac{\log n}{\hat{r}}\right\} kr_{k+1}\right)^{1/2} + \mu_{k+1,k+1}r_{k+1}. \quad (33)$$

Next, we see that (32) and (24) are simultaneously fulfilled if

$$\tfrac{1}{2\gamma}\left(\mu_{qq} - c'\max\left\{\sqrt{\tfrac{\sigma_{qq}^2\log n}{\hat{r}}}, \tfrac{\log n}{\hat{r}}\right\}\right) \geq \tfrac{1}{\gamma\hat{r}}\left(C\max\left\{\tilde{\sigma}\sqrt{n}, \sqrt{\log n}\right\} + C'\left(\max\left\{\hat{\sigma}^2, \tfrac{\log n}{\hat{r}}\right\} kr_{k+1}\right)^{1/2} + \mu_{k+1,k+1}r_{k+1}\right)$$

which holds if and only if

$$\hat{r}\left(\mu_{qq} - c'\max\left\{\sqrt{\tfrac{\sigma_{qq}^2\log n}{r_{qq}}}, \tfrac{\log n}{r_{qq}}\right\}\right) \geq 2\left(C\max\left\{\tilde{\sigma}\sqrt{n}, \sqrt{\log n}\right\} + C'\left(\max\left\{\hat{\sigma}^2, \tfrac{\log n}{\hat{r}}\right\} kr_{k+1}\right)^{1/2} + \mu_{k+1,k+1}r_{k+1}\right) \quad (34)$$

Finally, suppose that we choose the parameter $c_4 > \max\{c, c', 12\}$ so that gap condition (25) is satisfied and

$$\gamma > \max\{c, c', 12\}\max\left\{\sqrt{\frac{\tilde{\sigma}^2\log n}{\hat{r}}}, \frac{\log n}{\hat{r}}\right\}.$$

Then there exist constants c_1, c_2, c_3, depending on c_4, such that (33) and (34) are satisfied, i.e., there exists ϵ satisfying (23), (24) and (32) simultaneously, if

$$\gamma\hat{r} \geq c_1\max\left\{\tilde{\sigma}\sqrt{n}, \sqrt{\log n}\right\} + c_2\left(\max\left\{\sigma_1^2, \frac{\log n}{\hat{r}}\right\} kr_{k+1}\right)^{1/2} + c_3\mu_{k+1,k+1}r_{k+1}.$$

This concludes the proof of Theorem 3.

4. Numerical Methods and Simulations

We conclude with a discussion of an algorithm for solution of (5) based on the alternating direction method of multipliers (ADMM), and provide the results of a series of experiments that empirically verify the phase transitions predicted in Section 2.2. In particular, we randomly sample graphs $G = (V, \boldsymbol{W})$ from the planted cluster model and compare the optimal solution of (5) with the planted partition.

4.1 Alternating Direction Method of Multipliers for the Densest k-Disjoint Clique Problem

We solve (5) iteratively using the ADMM algorithm proposed by Ames (2014). Specifically, we split the decision variable $\boldsymbol{X}$ to obtain the equivalent formulation

$$\max\left\{\mathrm{Tr}(\boldsymbol{W}\boldsymbol{Y}) : \boldsymbol{X}-\boldsymbol{Y}=\boldsymbol{0}, \boldsymbol{X}\boldsymbol{e}\leq\boldsymbol{e}, \boldsymbol{X}\geq\boldsymbol{0},\ \mathrm{Tr}\,\boldsymbol{Y}=k, \boldsymbol{Y}\in\Sigma_+^V\right\}. \tag{35}$$

We then apply an approximate dual ascent scheme to maximize the augmented Lagrangian

$$L_\rho(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = \mathrm{Tr}(\boldsymbol{W}\boldsymbol{Y}) - \mathrm{Tr}(\boldsymbol{Z}(\boldsymbol{X}-\boldsymbol{Y})) + \frac{\rho}{2}\|\boldsymbol{X}-\boldsymbol{Y}\|_F^2,$$

where $\rho>0$ is a penalty parameter for violation of the linear equality constraint $\boldsymbol{X}-\boldsymbol{Y}=\boldsymbol{0}$. In particular, we minimize L_ρ with respect to $\boldsymbol{Y}$ and $\boldsymbol{X}$ successively, and then update $\boldsymbol{Z}=\boldsymbol{Z}-\rho(\boldsymbol{X}-\boldsymbol{Y})$ using approximate gradient ascent.

We update $\boldsymbol{Y}$ as the minimizer of the subproblem

$$\boldsymbol{Y}^{t+1} = \underset{\boldsymbol{Y}\in\Sigma_+^n}{\arg\min}\left\{\left\|\boldsymbol{Y}-\left(\boldsymbol{X}^t-\frac{\boldsymbol{W}+\boldsymbol{Z}^t}{\rho}\right)\right\|_F^2 : \mathrm{Tr}\,\boldsymbol{Y}=k\right\}, \tag{36}$$

where $(\boldsymbol{X}^t,\boldsymbol{Y}^t,\boldsymbol{Z}^t)$ is the current iterate after t iterations. That is, $\boldsymbol{Y}^{t+1}$ is the projection of the matrix $\boldsymbol{U}^t := \boldsymbol{X}^t-(\boldsymbol{W}+\boldsymbol{Z}^t)/\rho$ onto the intersection of the positive semidefinite cone and the set of matrices with trace equal to zero. Such a projection can be computed explicitly by projecting the vector of eigenvalues $\boldsymbol{\lambda}^t$ of $\boldsymbol{U}^t$ onto the nonnegative simplex $\{\boldsymbol{y}\in\mathbf{R}^n : \boldsymbol{e}^T\boldsymbol{y}=k,\ \boldsymbol{y}\geq\boldsymbol{0}\}$. Please see Zhang and Lu (2011, Proposition 2.6) and Van Den Berg and Friedlander (2008) for further details.

We update $\boldsymbol{X}^{t+1}$ as the optimal solution of

$$\boldsymbol{X}^{t+1} = \underset{\boldsymbol{X}\in\mathbf{R}^{n\times n}}{\arg\min}\left\{\left\|\boldsymbol{X}-\left(\boldsymbol{Y}^t+\boldsymbol{Z}^t/\rho\right)\right\|_F^2 : \boldsymbol{X}\geq\boldsymbol{0},\ \boldsymbol{X}\boldsymbol{e}\leq\boldsymbol{e}\right\}. \tag{37}$$

Applying strong duality, we know that the optimal solution of (37) is given by

$$\boldsymbol{X}^{t+1} = \left[\left(\boldsymbol{Y}^{t+1}+\boldsymbol{Z}^t/\rho\right)-\frac{\boldsymbol{z}^*\boldsymbol{e}+\boldsymbol{e}(\boldsymbol{z}^*)^T}{2}\right]_+, \tag{38}$$

where the operator $[\cdot]_+$ is the projection onto the symmetric nonnegative cone $\Sigma^V\cap\mathbf{R}_+^{V\times V}$ given by $[[\boldsymbol{Z}]_+]_{ij}=\max\{0,Z_{ij}\}$ for all $\boldsymbol{Z}\in\Sigma^V$, and $\boldsymbol{z}^*$ is the optimal solution of the dual problem of (37) given by

$$\min_{\boldsymbol{z}\geq\boldsymbol{0}}\frac{1}{2}\left\|\left[\left(\boldsymbol{Y}^{t+1}+\boldsymbol{Z}^t/\rho\right)-\frac{\boldsymbol{z}\boldsymbol{e}+\boldsymbol{e}\boldsymbol{z}^T}{2}\right]_+\right\|_F^2+\boldsymbol{z}^T\boldsymbol{e}-\frac{1}{2}\|\boldsymbol{Y}^{t+1}+\boldsymbol{Z}^t/\rho\|_F^2. \tag{39}$$

The objective function of the dual problem (39) is differentiable and coercive in $\boldsymbol{z}$, so it can be solved efficiently by applying the spectral projected gradient method of Birgin et al. (2000). We complete each iteration by performing an approximate dual ascent step to update the dual variable $\boldsymbol{Z}^{t+1}$. We stop the projected gradient method when the relative duality gap, given by $|v_p^{(t)} - v_d^{(t)}|/\max\{v_p^{(t)}, 1\}$ and primal constraint violation are both smaller than a desired error tolerance. We summarize the algorithm as Algorithm 1. Please see the work of Ames (2014, Section 6) for further implementation details.

4.2 Empirical Verification of Exact Recovery

We perform two sets of experiments, one to illustrate the recovery guarantee for dense graphs sampled from the heterogeneous planted cluster model and another to illustrate the guarantee when the noise is sparse. For the dense graph experiments, we fix $n = 1000$, and sample 10 graphs from the heterogeneous planted cluster model corresponding to the Bernoulli distributions $\Omega_{ij} = Bern(p_{ij})$ with probabilities of success p_{ij} given by

$$p_{ij} := \begin{cases} \left(1 - \left(\frac{0.35}{k+1}\right) i\right) p & \text{if } i = j \\ \left(1 - \left(\frac{0.35}{k+1}\right) \min\{i, j\}\right) q & \text{if } i \neq j, \end{cases}$$

for $q = 0.25$ and each $p = \{0.25, 0.275, 0.3, \ldots, 0.975, 1\}$ and $\hat{r} \in \{20, 40, \ldots, 500\}$. We choose the number of clusters $k = \lfloor n/\hat{r} \rfloor$ and distribute the remaining $n - k\hat{r}$ nodes evenly among $k - 1$ clusters to ensure that at least one cluster has minimum size. Under this choice of p_{ij} the smallest gap between the in-cluster and between-cluster means occurs when $i = 1$ and $j = k$; this implies that

$$\gamma = \left(1 - \frac{0.35k}{k+1}\right) p - q. \tag{40}$$

For each graph G, we call the ADMM algorithm sketched above to solve (5); in the algorithm, we use penalty parameter $\rho = \min\left\{\max\left\{5n/k, 80\right\}, 500\right\}/2$, stopping tolerance $\epsilon = 10^{-4}$, and maximum number of iterations 100. We declare the block structure of G to be recovered if $\|\boldsymbol{X}^* - \boldsymbol{X}_0\|_F^2 / \|\boldsymbol{X}_0\|_F^2 < 10^{-3}$, where $\boldsymbol{X}^*$ is the solution returned by the ADMM algorithm and $\boldsymbol{X}_0$ is the proposed solution given by (3). Note that Theorem 3 implies that we should expect exact recovery (w.h.p.) provided that $\gamma\hat{r} = \Omega\left(\sqrt{\tilde{\sigma}^2 n}\right)$. Figure 1(a) illustrates the empirical success rate for each choice of $\hat{r}$ and p, as well as the curve $p = (k+1)/(0.65k+1)(q + \frac{1}{2}\sqrt{n})$, where we use the upper bound $\tilde{\sigma}^2 \leq 1/4$ to estimate the constant term in (7).

We perform identical experiments for graphs sampled from the homogeneous planted cluster model with sparse noise. In particular, we fix $n = 1000$ and set $q = 1/\sqrt{n}$. We then sample 10 graphs from the planted cluster model corresponding to the Bernoulli distributions

Algorithm 1 ADMM for (1)

Input: Initial iterates $\boldsymbol{X}^0 = \boldsymbol{Y}^0 = \boldsymbol{Z}^0 = \boldsymbol{0}$, augmented Lagrangian parameter $\rho > 0$, and stopping tolerance $\epsilon > 0$.

Output: Approximate solution $(\boldsymbol{X}^*, \boldsymbol{Y}^*, \boldsymbol{Z}^*)$ of (1).

For $t = 0, 1, 2 \ldots$ until converged

Compute spectral decomposition $\boldsymbol{V}^t \text{Diag}\boldsymbol{\lambda}^t (\boldsymbol{V}^t)^T = \boldsymbol{U}^t = \boldsymbol{X}^t - (\boldsymbol{W} + \boldsymbol{Z}^t)/\rho$.

Project $\boldsymbol{\lambda}^t$ onto the nonnegative simplex $\{\boldsymbol{y} \in \mathbf{R}^n : \boldsymbol{e}^T \boldsymbol{y} = k,\ \boldsymbol{y} \geq \boldsymbol{0}\}$ to obtain $\bar{\boldsymbol{\lambda}}^t$.

Update $\boldsymbol{Y}^{t+1} = \boldsymbol{V}^t \text{Diag}\bar{\boldsymbol{\lambda}}^t (\boldsymbol{V}^t)^T$.

Compute approximate optimal solution $\boldsymbol{z}^*$ of the dual subproblem (39) using spectral projected gradient method of Birgin et al. (2000).

Update $\boldsymbol{X}^{t+1} = \left[\left(\boldsymbol{Y}^{ts+1} + \boldsymbol{Z}^t/\rho \right) - \frac{\boldsymbol{z}^* \boldsymbol{e} + \boldsymbol{e}(\boldsymbol{z}^*)^T}{2} \right]_+$.

Update $\boldsymbol{Z}^{t+1}$ using approximate dual ascent

$$\boldsymbol{Z}^{t+1} = \boldsymbol{Z}^t - \rho(\boldsymbol{X}^{t+1} - \boldsymbol{Y}^{t+1}).$$

Compute primal feasibilty gap

$$pfeas = \min \left\{ \min_{ij} Y_{ij}^t, \min \left(\boldsymbol{e} - \boldsymbol{Y}^t \boldsymbol{e} \right) \right\}.$$

Compute estimates of primal and dual objective values (note that $v_d^{(t+1)}$ is not necessarily a lower bound on the optimal dual value, but is asymptotically converging to the optimal dual value):

$$v_p^{(t+1)} = \text{Tr}(\boldsymbol{W}\boldsymbol{Y}^t) \qquad v_d^{(t+1)} = k\lambda_{min}(\boldsymbol{W} + \boldsymbol{Z}^{t+1}) - \text{Tr}(\boldsymbol{X}^{t+1}\boldsymbol{Z}^{t+1}).$$

.

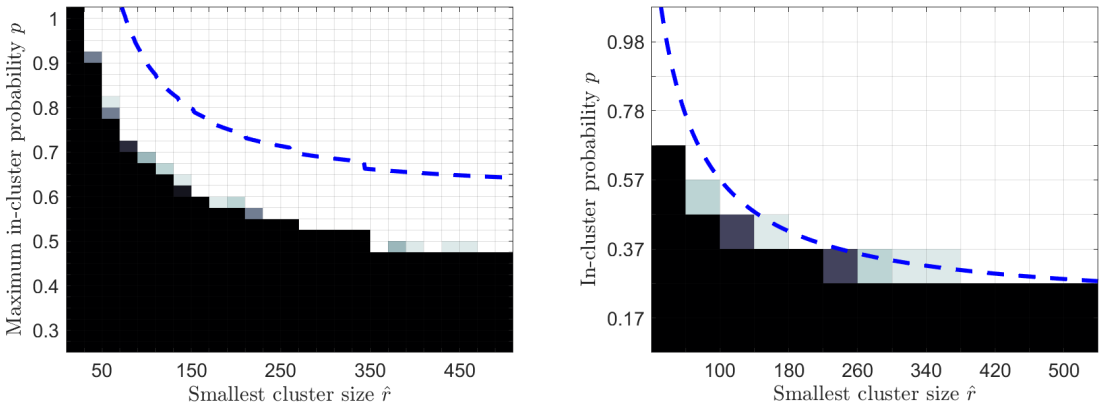
Calculate relative duality gap

$$relgap = \frac{|v_p^{(t+1)} = v_d^{(t+1)}|}{\max \left\{ |v_p^{(t+1)}|, 1 \right\}}.$$

Declare sequence of iterates to have converged if $relgap < \epsilon$ and $pfeas > -\epsilon$.

End For

$\Omega_{ij} = Bern(p)$ if $i = j$ and $\Omega_{ij} = Bern(q)$ if $i \neq j$ for each $\hat{r} \in \{20, 60, \ldots, 440, 500\}$ and $p = tq$ for 10 equally spaced scaling factors t between 2 and $\lfloor \sqrt{n} \rfloor$. As before, we set $k = \lfloor n/\hat{r} \rfloor$ and distribute the remaining nodes equally amongst the clusters so that the smallest has size $\hat{r}$ and $r_{k+1} = 0$. For each graph G, we call the ADMM algorithm to solve (5) (with the same parameters as before) and declare the block structure of G recovered if



(a) Dense noise.

(b) Sparse noise.

Figure 1: Empirical recovery rate for n-node graph with k planted cliques of size at least $\hat{r}$ and $\boldsymbol{W}$ generated according to the planted heterogeneous cluster model with distributions $\Omega_{ij} = Bern(p_{ij})$. Brighter colors indicate higher rates of recovery, with black corresponding to 0 recoveries and white corresponding to 10 recoveries (out of 10 trials). The dashed curves indicate the phase transition to perfect recovery predicted by Theorem 2.2.

$\|\boldsymbol{X}^* - \boldsymbol{X}_0\|_F^2 / \|\boldsymbol{X}_0\|_F^2 < 10^{-3}$. Theorem 3 suggests that we should expect recovery of the cluster structure in the case that

$$p > \Omega\left(\frac{1}{\sqrt{n}} + \frac{n^{1/4}}{\hat{r}}\right)$$

for this particular choice of p and q. Note that this implies that we have perfect recovery (w.h.p.) for $\hat{r} = \Omega(n^{1/4})$, rather than $\Omega(\sqrt{n})$ (as observed in the dense case). Figure 1(b) provides the empirical success rate for each choice of $\hat{r}$ and p, as well as the curve $p = 1/\sqrt{n} + n^{1/4}/\hat{r}$. It is clear that we are able to recover significantly smaller clusters under sparse noise than under dense noise, in accordance with (7).

5. Conclusions

We have established theoretical guarantees for graph clustering via a semidefinite relaxation of the densest k-disjoint problem. These results add to the growing corpus of evidence that clustering, while intractable in general, is possible if we seek to cluster clusterable data, i.e., data consisting of well-defined and well-separated groups of similar items. Moreover, our results provide further evidence that the $\omega(\sqrt{n})$ barrier can be broken for perfect cluster recovery in approximately sparse graphs and, specifically, that the size of recoverable clusters

scales logarithmically with n at worst in the special case that all clusters are roughly the same size. Finally, our semidefinite relaxation requires only an estimate of the number of clusters present in the data as input.

Our results suggest several areas of further research. The numerical simulations suggest that our theoretical guarantees may be overly conservative, especially in the dense noise case; further investigation is needed to determine if tighter estimates on the minimum size of clusters efficiently recoverable exist. Moreover, our model assumes clusters are disjoint. This is clearly not met in many practical applications; for example, returning to the social networking realm, users may belong to several overlapping communities. It would be worthwhile to see how our model and recovery guarantees can be modified to address overlapping clusters. Finally, our algorithm for graph clustering requires the solution of a semidefinite program, which may be impractical for even moderately large graphs. For example, the proposed algorithm, based on the ADMM, has per-iteration cost of $O(n^3)$ flops per iteration, primarily to compute the spectral decomposition needed to update $\boldsymbol{Y}$. Classical methods based on interior-point methods will scale even more poorly. Efficient, scalable methods for solving this semidefinite relaxation, and semidefinite programs in general, are needed.

Acknowledgments

We are grateful to John Bruer and Joel Tropp for their insights and helpful suggestions. Aleksis Pirinen was supported by a California Institute of Technology Summer Undergraduate Research Fellowship (SURF) using funds provided by Office of Naval Research (ONR) award N000014-11-1002. Brendan Ames was supported by University of Alabama Research Grant RG14678.

Appendix A. Proof of Lemma 7

In this appendix, we give the full proof of Lemma 7.

Proof We fix $q, s \in \{1, \ldots, k\}$ such that $q \neq s$ and assume without loss of generality that $r_q \leq r_s$. By the definition (14) of $\boldsymbol{y} := \boldsymbol{y}^{q,s}$ and the triangle inequality, we have

$$\|\boldsymbol{y}\|_\infty \leq \frac{1}{r_s}\left(\|\boldsymbol{b}_{q,s}\|_\infty + \frac{|\boldsymbol{b}_{q,s}^T \boldsymbol{e}|}{r_q + r_s}\right). \tag{41}$$

For simplicity, let $\boldsymbol{b}_1 := \boldsymbol{b}_{q,s}$ and $\boldsymbol{b}_2 := \boldsymbol{b}_{s,q}$. It follows from (15) and our choice of $\boldsymbol{\lambda}$ that the ith element of $\boldsymbol{b}_1$, denoted b_i^1, is given by

$$b_i^1 = r_s \left(\lambda_i - \frac{1}{2r_q}(\mu_{qq} r_q - \tau) \right) + \left(\boldsymbol{\lambda}_{C_s}^T \boldsymbol{e} - \frac{1}{2}(\mu_{ss} r_s - \tau) \right) - \left(\sum_{j \in C_s} w_{ij} - \mu_{qs} r_s \right).$$

It follows from the definition (12) of $\boldsymbol{\lambda}_{C_s}$ that

$$\boldsymbol{\lambda}_{C_s}^T \boldsymbol{e} = \frac{1}{2r_s} \left(\boldsymbol{e}^T \boldsymbol{W}_{C_s, C_s} \boldsymbol{e} - r_s \tau \right),$$

which implies that

$$\left| \boldsymbol{\lambda}_{C_s}^T \boldsymbol{e} - \frac{1}{2}(\mu_{ss} r_s - \tau) \right| = \frac{1}{2r_s} \left| \boldsymbol{e}^T \boldsymbol{W}_{C_s, C_s} \boldsymbol{e} - \mu_{ss} r_s^2 \right|.$$

Applying (21) with $T = n$ to the right-hand side in the equation above shows that

$$|\boldsymbol{e}^T \boldsymbol{W}_{C_s, C_s} \boldsymbol{e} - \mu_{ss} r_s^2| \leq 6 \max\{ \sqrt{\sigma_{ss}^2 r_s^2 \log n}, \log n \} \tag{42}$$

with high probability, which in turn implies that

$$\left| \boldsymbol{\lambda}_{C_s}^T \boldsymbol{e} - \frac{1}{2}(\mu_{ss} r_s - \tau) \right| \leq 3 \max \left\{ \sqrt{\sigma_{ss}^2 \log n}, \frac{\log n}{r_s} \right\} \tag{43}$$

with high probability. Similarly, applying (21) with $T = n$ to the sum $\sum_{j \in C_s} w_{ij}$ shows that

$$\left| \sum_{j \in C_s} w_{ij} - \mu_{qs} r_s \right| \leq 6 \max \left\{ \sqrt{\sigma_{qs}^2 r_s \log n}, \log n \right\} \tag{44}$$

for all $i \in C_q$ with high probability. Finally, we note that

$$\left| \lambda_i - \frac{1}{2r_q}(\mu_{qq} r_q - \tau) \right| \leq \frac{1}{r_q} \left| \sum_{j \in C_q} w_{ij} - \mu_{qq} r_q \right| + \frac{1}{2r_q^2} \left| \boldsymbol{e}^T \boldsymbol{W}_{C_q, C_q} \boldsymbol{e} - \mu_{qq} r_q^2 \right|. \tag{45}$$

We bound the first term in the sum using (21) with $T = n$, which establishes that

$$\left| \sum_{j \in C_q} w_{ij} - \mu_{qq} r_q \right| \leq 6 \max \left\{ \sqrt{\sigma_{qq}^2 r_q \log n}, \log n \right\}$$

w.h.p., and note that the second term has upper bound

$$\left| \boldsymbol{e}^T \boldsymbol{W}_{C_q, C_q} \boldsymbol{e} - \mu_{qq} r_q^2 \right| \leq 6 \max \left\{ \sqrt{\sigma_{qq}^2 r_q^2 \log n}, \log n \right\}$$

w.h.p. by a calculation identical to that used to obtain (42). Applying these bounds using the triangle inequality and the union bound over all $i \in C_q$, we conclude that

$$\begin{aligned}
\|\boldsymbol{b}_1\|_\infty &\le r_s \left|\lambda_i - \frac{1}{2r_q}(\mu_{qq} r_q - \tau)\right| + \left|\boldsymbol{\lambda}_{C_s}^T \boldsymbol{e} - \frac{1}{2}(\mu_{ss} r_s - \tau)\right| + \left|\sum_{j \in C_s} w_{ij} - \mu_{qs} r_s\right| \\
&\le\ r_s \left(\frac{6}{r_q} \max\left\{\sqrt{\sigma_{qq}^2 r_q \log n}, \log n\right\} + 3 \max\left\{\sqrt{\sigma_{qq}^2 \log n}, \frac{\log n}{r_q}\right\}\right) \\
&\qquad + 3 \max\left\{\sqrt{\sigma_{ss}^2 \log n}, \frac{\log n}{r_s}\right\} + 6 \max\left\{\sqrt{\sigma_{qs}^2 r_s \log n}, \log n\right\} \\
&= O\left(r_s \max\left\{\sqrt{\frac{\tilde{\sigma}^2 \log n}{\hat{r}}}, \frac{\log n}{\hat{r}}\right\}\right) \qquad (46)
\end{aligned}$$

with high probability.

We next bound $\left|\boldsymbol{b}_1^T \boldsymbol{e}\right|$. We have

$$\boldsymbol{b}_1^T \boldsymbol{e} = r_s \left(\boldsymbol{\lambda}_{C_q}^T \boldsymbol{e} - \frac{1}{2}\left(\mu_{qq} r_q - \tau\right)\right) + r_q \left(\boldsymbol{\lambda}_{C_s}^T \boldsymbol{e} - \frac{1}{2}\left(\mu_{ss} r_s - \tau\right)\right) + \left(\mu_{qs} r_s r_q - \boldsymbol{e}^T \boldsymbol{W}_{C_q, C_s} \boldsymbol{e}\right).$$

Applying (21) to bound the sum of the entries of $\boldsymbol{W}_{C_q,C_s}$ and the above concentration inequalities for $\boldsymbol{\lambda}_{C_q}^T \boldsymbol{e}$ and $\boldsymbol{\lambda}_{C_s}^T \boldsymbol{e}$ we have

$$\begin{aligned}
&\left|\boldsymbol{b}_1^T \boldsymbol{e}\right| \le r_s \left|\boldsymbol{\lambda}_{C_q}^T \boldsymbol{e} - \frac{1}{2}\left(\mu_{qq} r_q - \tau\right)\right| + r_q \left|\boldsymbol{\lambda}_{C_s}^T \boldsymbol{e} - \frac{1}{2}\left(\mu_{ss} r_s - \tau\right)\right| + \left|\mu_{qs} r_s r_q - \boldsymbol{e}^T \boldsymbol{W}_{C_q, C_s} \boldsymbol{e}\right| \\
&\le 3 r_s \max\left\{\sqrt{\sigma_{qq}^2 \log n}, \frac{\log n}{r_q}\right\} + 3 r_q \max\left\{\sqrt{\sigma_{ss}^2 \log n}, \frac{\log n}{r_s}\right\} + 6 \max\left\{\sqrt{\sigma_{qs}^2 r_q r_s \log n}, \log n\right\} \\
&= O\left((r_s + r_q) \max\left\{\sqrt{\tilde{\sigma}^2 \log n}, \frac{\log n}{\hat{r}}\right\}\right) \qquad (47)
\end{aligned}$$

w.h.p. Finally, we bound $\|\boldsymbol{y}\|_\infty$ using (46) and (47):

$$\|\boldsymbol{y}^{q,s}\|_\infty \le \frac{1}{r_s}\left(\|\boldsymbol{b}_1\|_\infty + \frac{|\boldsymbol{b}_1^T \boldsymbol{e}|}{r_q + r_s}\right) = O\left(\max\left\{\sqrt{\frac{\tilde{\sigma}^2 \log n}{\hat{r}}}, \frac{\log n}{\hat{r}}\right\}\right)$$

w.h.p. Since this holds for any $q, s \in \{1, \ldots, k\}$ such that $q \neq s$, we conclude that

$$\|\boldsymbol{y}\|_\infty = O\left(\max\left\{\sqrt{\frac{\tilde{\sigma}^2 \log n}{\hat{r}}}, \frac{\log n}{\hat{r}}\right\}\right) \qquad (48)$$

w.h.p. An identical argument shows that

$$\|\boldsymbol{z}^{q,s}\|_\infty = O\left(\max\left\{\sqrt{\frac{\tilde{\sigma}^2 \log n}{\hat{r}}}, \frac{\log n}{\hat{r}}\right\}\right) \qquad (49)$$

w.h.p. We conclude that

$$\|\boldsymbol{y}^{q,s}\|_\infty + \|\boldsymbol{z}^{q,s}\|_\infty = O\left(\max\left\{\sqrt{\frac{\tilde{\sigma}^2 \log n}{\hat{r}}}, \frac{\log n}{\hat{r}}\right\}\right) \tag{50}$$

w.h.p. ■

Appendix B. Proof of Proposition 9

We next prove Proposition 9.

Proof We follow the proof of Lemma 4.3 given by Ames (2014). Fix $q \in \{1, \dots, k\}$ and $i \in C_q$. It follows from (12) that

$$\lambda_i = \sum_{j \in C_q} w_{ij} - \frac{1}{2r_q} \boldsymbol{e}^T \boldsymbol{W}_{C_q, C_q} \boldsymbol{e} - \frac{\tau}{2}$$

for each $i \in C_q$. Applying (21) with $S = \sum_{j \in C_q} w_{ij}$ and $T = n$ yields

$$\sum_{j \in C_q} w_{ij} \geq \mu_{qq} r_q - 6 \max\left\{\sqrt{\sigma_{qq}^2 r_q \log n}, \log n\right\}$$

w.h.p. Moreover, by a similar argument, we have

$$\frac{1}{2r_q} \boldsymbol{e}^T \boldsymbol{W}_{C_q, C_q} \boldsymbol{e} \leq \frac{1}{2}\left(\mu_{qq} r_q + 6 \max\left\{\sqrt{\sigma_{qq}^2 \log n}, \frac{\log n}{r_q}\right\}\right)$$

w.h.p. Combining the above inequalities shows that

$$\lambda_i \geq r_q \left(\frac{\mu_{qq}}{2} - \epsilon\gamma - O\left(\max\left\{\sqrt{\frac{\sigma_{qq}^2 \log n}{r_q}}, \frac{\log n}{r_q}\right\}\right)\right)$$

w.h.p. Since $\gamma > 0$ by (7), this implies that there exists constant $c > 0$ such that if

$$\epsilon \leq \frac{1}{\gamma}\left(\mu_{qq} - c \max\left\{\sqrt{\frac{\sigma_{qq}^2 \log n}{r_q}}, \frac{\log n}{r_q}\right\}\right) \tag{51}$$

then $\lambda_i \geq 0$ w.h.p. Applying the union bound over all $q = 1, 2, \dots, k$ and $i \in C_q$ shows that each entry of $\boldsymbol{\lambda}_{C_q}$ is nonnegative w.h.p. if ϵ is chosen to satisfy (51) for all q. ■

Appendix C. Proof of Proposition 10

Our proof of Proposition 10 follows a similar structure to that of Ames (2014, Lemma 4.4).

Proof Fix $q \neq s$ with $q \in 1, \ldots, k$. Applying (42) and (21) with $S = \boldsymbol{e}^T \boldsymbol{W}_{C_q,C_s} \boldsymbol{e}$ and $T = n$, we have

$$\begin{aligned} & r_s \boldsymbol{e}^T \boldsymbol{W}_{C_q,C_q} \boldsymbol{e} - r_q \boldsymbol{e}^T \boldsymbol{W}_{C_q,C_s} \boldsymbol{e} \\ & \quad \geq (\mu_{qq} - \mu_{qs}) r_s r_q^2 - 6 r_s \max\left\{\sqrt{\sigma_{qq}^2 r_q^2 \log n}, \log n\right\} - 6 r_q \max\left\{\sqrt{\sigma_{qs}^2 r_s r_q \log n}, \log n\right\} \\ & \quad \geq r_q^2 r_s \left(\gamma - 12 \max\left\{\sqrt{\frac{\tilde{\sigma}^2 \log n}{\hat{r}^2}}, \frac{\log n}{\hat{r}^2}\right\}\right) \end{aligned}$$

w.h.p. This implies that $r_s \boldsymbol{e}^T \boldsymbol{W}_{C_q,C_q} \boldsymbol{e} \geq r_q \boldsymbol{e}^T \boldsymbol{W}_{C_q,C_s} \boldsymbol{e}$ w.h.p. if (25) is satisfied. ■

Appendix D. Proof of Lemma 12

In this appendix we prove Lemma 12.

Proof We will make repeated use of the following lemma, which specializes the concentration inequality on the spectral norm of a random symmetric matrix with i.i.d. mean zero entries given by Bandeira and van Handel (2016, Corollary 3.12).

Lemma 14 *Let $\boldsymbol{A} = [a_{ij}] \in \Sigma^n$ be a random symmetric matrix with i.i.d. mean zero entries a_{ij} having variance at most σ^2 and satisfying $|a_{ij}| \leq 1$. Then there exists constant $C > 0$ such that*

$$\Pr\left\{\|\boldsymbol{A}\| > C \max\left\{\sqrt{\sigma^2 n}, \sqrt{T}\right\}\right\} \leq n T^{-7} \tag{52}$$

for all $T > 0$.

Proof (of Lemma 14) Corollary 3.12 of Bandeira and van Handel (2016) establishes that for each $t > 0$ there exists $\tilde{c} = \tilde{c}(t) > 0$ such that

$$\Pr\left\{\|\boldsymbol{A}\| \geq 3\sqrt{\sigma^2 n} + t\right\} \leq n e^{-\tilde{c} t^2}. \tag{53}$$

Here, we have substituted the upper bound $\sigma^2 n \geq \tilde{\sigma}^2$, in place of $\tilde{\sigma} := \max_i \sum_j \mathbf{E}[X_{ij}^2]$ in the original statement of Corollary 3.12. Let $t = (C-3)\max\{\sqrt{\sigma^2 n}, \sqrt{\log T}\}$ where C is chosen large enough that $\tilde{c}(C-3)^2 > 7$. In this case, (53) specializes to

$$\Pr\left\{\|\boldsymbol{A}\| \geq C \max\{\sqrt{\sigma^2 n}, \sqrt{\log n}\}\right\} \leq n e^{-7 \max\{\sigma^2 n, \log n\}} \leq n e^{-7 \log n} = n T^{-7}.$$

This completes the proof. ■

Before we continue with the derivation of the desired bound on $\|\tilde{\boldsymbol{S}}_1\|$, we note that the entries $[\tilde{\boldsymbol{S}}_1]_{ij}$ of $\tilde{\boldsymbol{S}}_1$ all satisfy $|[\tilde{\boldsymbol{S}}_1]_{ij}| \leq 1$ if we assume that $w_{ij} \in [0,1]$ for all i,j; note that an identical argument establishes the result if we make the weaker assumption that the entries of $\boldsymbol{W}$ are bounded with high probability. On the other hand, note that the entries of $\tilde{\boldsymbol{S}}_1$ are not identically distributed (but are independent) since each w_{ij} is sampled according to Ω_{qs}, where $i \in C_q$, $j \in C_s$. However, we know that $\sigma_{qs}^2 \leq \tilde{\sigma}^2$ by our definition of $\tilde{\sigma}^2$. Moreover, $\mathbf{E}[(\tilde{\boldsymbol{S}}_1)_{ij}] = \mathbf{E}[\mu_{qs} - w_{ij}] = 0$. Thus, we can apply Lemma 14 to place a bound on $\|\tilde{\boldsymbol{S}}_1\|$. Doing so establishes that (30) holds w.h.p. ■

Appendix E. Proof of Lemma 13

We conclude with the following proof of Lemma 13.

Proof Note that $\|\tilde{\boldsymbol{S}}_2\| \leq \|\boldsymbol{\lambda} - \mathbf{E}[\boldsymbol{\lambda}]\|\sqrt{r_{k+1}}$. Thus, it remains to bound $\|\boldsymbol{\lambda} - \mathbf{E}[\boldsymbol{\lambda}]\|$.

To do so, fix $q \in \{1, 2, \ldots, k\}$. Recall that

$$\boldsymbol{\lambda}_{C_q} - \mathbf{E}[\boldsymbol{\lambda}_{C_q}] = \frac{1}{r_q}(\boldsymbol{W}_{C_q,C_q}\boldsymbol{e} - \mu_{qq}r_q\boldsymbol{e}) - \frac{1}{r_q^2}(\boldsymbol{e}^T\boldsymbol{W}_{C_q,C_q}\boldsymbol{e} - \mu_{qq}r_q^2)\boldsymbol{e}.$$

Applying (52) with $T = n$ establishes that

$$\begin{aligned}\|\boldsymbol{W}_{C_q,C_q}\boldsymbol{e} - \alpha r_q\boldsymbol{e}\| &\leq \|\boldsymbol{W}_{C_q,C_q} - \mu_{qq}\boldsymbol{e}\boldsymbol{e}^T\|\|\boldsymbol{e}\| \\ &\leq C\sqrt{r_q}\max\{\sigma_{qq}\sqrt{r_q}, \sqrt{\log n}\}\end{aligned}$$

w.h.p. On the other hand, Bernstein's inequality establishes that

$$|\boldsymbol{e}^T\boldsymbol{W}_{C_q,C_q}\boldsymbol{e} - \mu_{qq}r_q^2| \leq 6\max\left\{\sqrt{\sigma_{qq}^2 r_q \log n}, \log n\right\}$$

w.h.p. Combining these two inequalities using the triangle inequality establishes that

$$\begin{aligned}\|\boldsymbol{\lambda}_{C_q} - \mathbf{E}[\boldsymbol{\lambda}_{C_q}]\| \leq& C\max\left\{\sigma_{qq}, \sqrt{\frac{\log n}{r_q}}\right\} + 6\max\left\{\sqrt{\frac{\sigma_{qq}^2\log n}{r_q^2}}, \frac{\log n}{r_q^{3/2}}\right\} \\ &= O\left(\max\left\{\sigma_{qq}, \sqrt{\frac{\log n}{\hat{r}}}\right\}\right)\end{aligned}$$

w.h.p. Finally, applying the union bound over all choices of q shows that

$$\|\boldsymbol{\lambda} - \mathbf{E}[\boldsymbol{\lambda}]\|^2 = \sum_{q=1}^{t}\|\boldsymbol{\lambda}_{C_q} - \mathbf{E}[\boldsymbol{\lambda}_{C_q}]\|^2 = O\left(k\max\left\{\sigma_{qq}^2, \frac{\log n}{\hat{r}}\right\}\right)$$

w.h.p. This establishes that

$$\|\tilde{\boldsymbol{S}}_2\|^2 = O\left(kr_{k+1}\max\left\{\sigma_{qq}^2, \frac{\log n}{\hat{r}}\right\}\right)$$

w.h.p., as required. ■

References

Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *Information Theory, IEEE Transactions on*, 62(1):471–487, 2016.

Nir Ailon, Yudong Chen, and Xu Huan. Breaking the small cluster barrier of graph clustering. *arXiv preprint arXiv:1302.4549*, 2013.

Brendan PW Ames. Guaranteed clustering and biclustering via semidefinite programming. *Mathematical Programming*, 147(1-2):429–465, 2014.

Brendan PW Ames. Guaranteed recovery of planted cliques and dense subgraphs by convex relaxation. *Journal of Optimization Theory and Applications*, 167(2):653–675, 2015.

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

Brendan PW Ames and Stephen A Vavasis. Convex optimization for the planted k-disjoint-clique problem. *Mathematical Programming*, 143(1-2):299–337, 2014.

Arash A Amini and Elizaveta Levina. On semidefinite relaxations for the block model. *arXiv preprint arXiv:1406.5647*, 2014.

Afonso S Bandeira and Ramon van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability*, 44(4):2479–2506, 2016.

Ernesto G Birgin, José Mario Martínez, and Marcos Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 10(4): 1196–1211, 2000.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2009.

T Tony Cai, Xiaodong Li, et al. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *The Annals of Statistics*, 43(3):1027–1059, 2015.

Yudong Chen and Jiaming Xu. Statistical-computational phase transitions in planted models: The high-dimensional setting. In *ICML*, pages 244–252, 2014.

Yudong Chen, Ali Jalali, Sujay Sanghavi, and Huan Xu. Clustering partially observed graphs via convex optimization. *The Journal of Machine Learning Research*, 15(1): 2213–2238, 2014a.

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

Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–556. ACM, 2004.

GH Golub and C Van Loan. Matrix computations, 4th. *Johns Hopkins*, 2013.

Olivier Guédon and Roman Vershynin. Community detection in sparse networks via grothendieck's inequality. *Probability Theory and Related Fields*, pages 1–25, 2015.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming. In *Information Theory (ISIT), 2015 IEEE International Symposium on*, pages 1442–1446. IEEE, 2015.

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

Amin Jalali, Qiyang Han, Ioana Dumitriu, and Maryam Fazel. Relative density and exact recovery in heterogeneous stochastic block models. *arXiv preprint arXiv:1512.04937*, 2015.

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

Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th international conference on World Wide Web*, pages 695–704. ACM, 2008.

Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

G. Lugosi. Concentration-measure inequalities, 2009. Available from `http://www.econ.upf.edu/~lugosi/anu.pdf`.

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. Society for Industrial and Applied Mathematics, 2010.

Abhinav Nellore and Rachel Ward. Recovery guarantees for exemplar-based clustering. *arXiv preprint arXiv:1309.3256*, 2013.

Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

S. Oymak and B. Hassibi. Finding dense clusters via "low rank + sparse" decomposition. *Arxiv preprint arXiv:1104.5186*, 2011.

Jiming Peng and Yu Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM Journal on Optimization*, 18(1):186–205, 2007.

Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128, 2013.

K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Karl Rohe, Tai Qin, and Haoyang Fan. The highest dimensional stochastic blockmodel with a regularized estimator. *arXiv preprint arXiv:1206.2380*, 2012.

Ewout Van Den Berg and Michael P Friedlander. Probing the pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912, 2008.

Ramya Korlakai Vinayak, Samet Oymak, and Babak Hassibi. Sharp performance bounds for graph clustering via convex optimization. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 8297–8301. IEEE, 2014.

Bowei Yan, Purnamrita Sarkar, and Xiuyuan Cheng. Provable estimation of the number of blocks in block models. *arXiv preprint arXiv:1705.08580*, 2017.

Yong Zhang and Zhaosong Lu. Penalty decomposition methods for rank minimization. In *Advances in Neural Information Processing Systems*, pages 46–54, 2011.