Math. Program., Ser. A (2014) 147:429–465
DOI 10.1007/s10107-013-0729-x

FULL LENGTH PAPER

Guaranteed clustering and biclustering via semidefinite programming

Brendan P. W. Ames

Received: 16 February 2012 / Accepted: 9 November 2013 / Published online: 30 November 2013

Abstract Identifying clusters of similar objects in data plays a significant role in a wide range of applications. As a model problem for clustering, we consider the densest k-disjoint-clique problem, whose goal is to identify the collection of k disjoint cliques of a given weighted complete graph maximizing the sum of the densities of the complete subgraphs induced by these cliques. In this paper, we establish conditions ensuring exact recovery of the densest k cliques of a given graph from the optimal solution of a particular semidefinite program. In particular, the semidefinite relaxation is exact for input graphs corresponding to data consisting of k large, distinct clusters and a smaller number of outliers. This approach also yields a semidefinite relaxation with similar recovery guarantees for the biclustering problem. Given a set of objects and a set of features exhibited by these objects, biclustering seeks to simultaneously group the objects and features according to their expression levels. This problem may be posed as that of partitioning the nodes of a weighted bipartite complete graph such that the sum of the densities of the resulting bipartite complete subgraphs is maximized. As in our analysis of the densest k-disjoint-clique problem, we show that the correct partition of the objects and features can be recovered from the optimal solution of a semidefinite program in the case that the given data consists of several disjoint sets of objects exhibiting similar features. Empirical evidence from numerical experiments supporting these theoretical guarantees is also provided.

Keywords Semidefinite programming · Clustering · Planted clique · Biclustering · Co-clustering

Mathematics Subject Classification 90C22 · 62H30 · 68Q25 · 91C20

B. P. W. Ames (✉)
Department of Computing and Mathematical Sciences, California Institute of Technology, 1200 E. California Blvd., Mail Code 305-16, Pasadena, CA 91125, USA
e-mail: bpames@caltech.edu

1 Introduction

The goal of *clustering* is to partition a given data set into groups of similar objects, called *clusters*. Clustering is a fundamental problem in statistics and machine learning and plays a significant role in a wide range of applications, including information retrieval, pattern recognition, computational biology, and image processing. The complexity of finding an optimal clustering depends significantly on the measure of fitness of a proposed partition, but most interesting models for clustering are posed as an intractable combinatorial problem. For this reason, heuristics are used to cluster data in most practical applications. Unfortunately, although much empirical evidence exists for the usefulness of these heuristics, few theoretical guarantees ensuring the quality of the obtained partition are known, even for data containing well separated clusters. For a recent survey of clustering techniques and heuristics, see Berkhin [7]. In this paper, we establish conditions ensuring that the optimal solution of a particular convex optimization problem yields a correct clustering under certain assumptions on the input data set.

Our approach to clustering is based on partitioning the similarity graph of a given set of data. Given a data set S and measure of similarity between any two objects, the *similarity graph* G_S is the weighted complete graph with nodes corresponding to the objects in the data set and each edge ij having weight equal to the level of similarity between objects i and j. For this representation of data, clustering the data set S is equivalent to partitioning the nodes of G_S into disjoint cliques such that edges connecting any two nodes in the same clique have significantly higher weight than those between different cliques. Therefore, a clustering of the data may be obtained by identifying dense, in the sense of having large average edge weight, subgraphs of G_S.

We consider the densest k-partition problem as a model problem for clustering. Given a weighted complete graph $K = (V, E, W)$ and integer $k \in \{1, \ldots, |V|\}$, the *densest k-partition problem* aims to identify the partition of V into k disjoint sets such that the sum of the average edge weights of the complete subgraphs induced by these cliques is maximized. Unfortunately, the densest k-partition problem is NP-hard, since it contains the minimum sum of squared Euclidean distance problem, known to be NP-hard [2], as a special case. In Sect. 2, we consider the related problem of finding the set of k disjoint complete subgraphs maximizing the sum of their densities. We model this problem as a quadratic program with combinatorial constraints and relax to a semidefinite program using matrix lifting. This relaxation approach is similar to that employed in several recent papers [19,35,43], although we consider a different model problem for clustering and establish stronger recovery properties. We show that the optimal solution of this semidefinite relaxation coincides with that of the original combinatorial problem for certain program inputs. In particular, we show that the set of input graphs for which the relaxation is exact includes the set of graphs with edge weights concentrated on a particular collection of disjoint subgraphs, and provide a general formula for the clique sizes and number of cliques that may be recovered.

In Sect. 3, we establish similar results for the biclustering problem. Given a set of objects and features, *biclustering*, also known as *co-clustering*, aims to simultaneously group the objects and features according to their expression levels. That is, we would like to partition the objects and features into groups of objects and features, called

biclusters, such that objects strongly exhibit features within their bicluster relative to the features within the other biclusters. Hence, biclustering differs from clustering in the sense that it does not aim to obtain groups of similar objects, but instead seeks groups of objects similar with respect to a particular subset of features. Applications of biclustering include identifying subsets of genes exhibiting similar expression patterns across subsets of experimental conditions in analysis of gene expression data, grouping documents by topics in document clustering, and grouping customers according to their preferences in collaborative filtering and recommender systems. For an overview of the biclustering problem, see Busygin et al. [11], Fan et al. [18].

As a model problem for biclustering, we consider the problem of partitioning a bipartite graph into dense disjoint subgraphs. If the given bipartite graph has vertex sets corresponding to sets of objects and features with edges indicating expression level of each feature by each object, each dense subgraph will correspond to a bicluster of objects strongly exhibiting the contained features. Given a weighted bipartite complete graph $K = ((U, V), E, W)$ and integer $k \in \{1, \ldots, \min\{|U|, |V|\}\}$, we seek the set of k disjoint bipartite complete subgraphs with sum of their densities maximized. We establish that this problem may be relaxed as a semidefinite program and show that, for certain program instances, the correct partition of K can be recovered from the optimal solution of this relaxation. In particular, this relaxation is exact in the special case that the edge weights of the input graph are concentrated on some set of disjoint bipartite subgraphs. When the input graph arises from a given data set, the relaxation is exact when the underlying data set consists of several disjoint sets strongly exhibiting nonoverlapping sets of features.

Our results build upon those of recent papers regarding clusterability of data. These papers generally contain results of the following form: if a data set is randomly sampled from a distribution of "clusterable" data, then the correct partition of the data can be obtained efficiently using some heuristic, such as the k-means algorithm or other iterative partitioning heuristics [1,6,32,42], spectral clustering [5,27,31,40], or convex optimization [3,26,30,33]. Recent papers by Kolar et al. [28], Rohe and Yu [41], and Flynn and Perry [20] establish analogous recovery guarantees for biclustering; the latter two of these papers appeared shortly after the initial preprint release of this paper. Our results are of a similar form. If the underlying data set consists of several sufficiently distinct clusters or biclusters, then the correct partition of the data can be recovered from the optimal solution of our relaxations. We model this ideal case for clustering using random edge weight matrices constructed so that weight is, in expectation, concentrated heavily on the edges of a few disjoint subgraphs. We will establish that this random model for clustered data contains those previously considered in the literature and, in this sense, our results are a generalization of these earlier theoretical guarantees.

More generally, our results follow in the spirit of, and borrow techniques from, recent work regarding sparse optimization and, in particular, the nuclear norm relaxation for rank minimization. The goal of *matrix rank minimization* is to find a solution of minimum rank of a given linear system, i.e., to find the optimal solution $X^* \in \mathbf{R}^{m\times n}$ of the optimization problem $\min\{\text{rank}\, X : \mathcal{A}(X) = \mathbf{b}\}$ for given linear operator $\mathcal{A} : \mathbf{R}^{m\times n} \to \mathbf{R}^p$ and vector $\mathbf{b} \in \mathbf{R}^p$. Although this problem is well-known to be NP-hard, several recent papers ([4,12,13,24,34,36–38], among others) have established

that, under certain assumptions on $\mathcal{A}$ and $\mathbf{b}$, the minimum rank solution is equal to the optimal solution of the convex relaxation obtained by replacing rank X with the sum of the singular values of X, the nuclear norm $\|X\|_*$. This relaxation may be thought of as a matrix analogue of the ℓ_1 norm relaxation for the cardinality minimization problem, and these results generalize similar recovery guarantees for compressed sensing (see [14–16]). For example, the nuclear norm relaxation is exact with high probability if $\mathcal{A}$ is a random linear transform with matrix representation having i.i.d. Gaussian or Bernoulli entries and $\mathbf{b} = \mathcal{A}(X_0)$ is the image of a sufficiently low rank matrix X_0 under $\mathcal{A}$. We prove analogous results for an instance of rank constrained optimization. To identify the densest k complete subgraphs of a given graph, we seek a rank-k matrix X maximizing some linear function of X, depending only on the edge weights W of the input graph, subject to linear constraints. We will see that the optimal rank-k solution is equal to that obtained by relaxing the rank constraint to the corresponding nuclear norm constraint if the matrix W is randomly sampled from a probability distribution satisfying certain assumptions.

2 A semidefinite relaxation of 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 $i, j \in C$. Let $K_N = (V, E, W)$ be a complete graph with vertex set $V = \{1, 2, \ldots, N\}$ and nonnegative edge weights $W_{ij} \in [0, 1]$ for all $i, j \in V$. A *k-disjoint-clique subgraph* of K_N is a subgraph of K_N consisting of k disjoint complete subgraphs, i.e., the vertex sets of each of these subgraphs is a *clique*. For any subgraph H of K_N, the density of H, denoted d_H, is the average edge weight incident at a vertex in H:

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

The *densest k-disjoint-clique problem* concerns choosing a k-disjoint-clique subgraph of K_N such that the sum of the densities of the subgraphs induced by the cliques is maximized. Given a k-disjoint-clique subgraph with vertex set composed of cliques $C_1, \ldots, C_k$, the sum of the densities of the subgraphs induced by the cliques is equal to

$$\sum_{i=1}^{k} d_{G(C_i)} = \sum_{i=1}^{k} \frac{\mathbf{v}_i^T W \mathbf{v}_i}{\mathbf{v}_i^T \mathbf{v}_i}, \tag{2.1}$$

where $\mathbf{v}_i$ is the characteristic vector of C_i. In the special case that $C_1, \ldots, C_k$ defines a partition of V and $W_{ij} = 1 - \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^2$ for a given set of N vectors $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ in $\mathbf{R}^n$ with maximum distance between any two points at most one, we have

$$\sum_{\ell=1}^{k} d_{G(C_\ell)} = \sum_{\ell=1}^{k} \frac{1}{|C_\ell|} \left(\sum_{i \in C_\ell} \sum_{j \in C_\ell} (1 - (\mathbf{x}^{(i)} - \mathbf{x}^{(j)})^T (\mathbf{x}^{(i)} - \mathbf{x}^{(j)})) \right)$$

$$= \sum_{\ell=1}^{k} \left(|C_\ell| - 2 \left(\sum_{i \in C_\ell} \|\mathbf{x}^{(i)}\|^2 - \sum_{i \in C_\ell} \sum_{j \in C_\ell} (\mathbf{x}^{(i)})^T \mathbf{x}^{(j)} \right) \right)$$

$$= N - 2 \sum_{\ell=1}^{k} \sum_{i \in C_\ell} \|\mathbf{x}^{(i)} - \mathbf{c}^{(\ell)}\|^2,$$

where $\mathbf{c}^{(\ell)} = \sum_{i \in C_\ell} \mathbf{x}^{(i)} / |C_\ell|$ is the center of the vectors assigned to C_ℓ for all $\ell = 1, \ldots, k$, since $\sum_{\ell=1}^{k} |C_\ell| = N$ for this choice of W. Here, and in the rest of the note, $\|\mathbf{x}\| = \sqrt{\mathbf{x}^T \mathbf{x}}$ denotes the ℓ_2 norm in $\mathbf{R}^n$. For this choice of W, the *densest k-partition problem*, i.e., finding a partition $C_1, \ldots, C_k$ of V such that the sum of densities of the subgraphs induced by $C_1, \ldots, C_k$ is maximized, is equivalent to finding the partition of V such that the sum of the squared Euclidean distances

$$f(\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}, \{C_1, \ldots, C_k\}) = \sum_{\ell=1}^{k} \sum_{i \in C_\ell} \|\mathbf{x}^{(i)} - \mathbf{c}^{(\ell)}\|^2 \tag{2.2}$$

from each vector $\mathbf{x}^{(i)}$ to its assigned cluster center is minimized. Unfortunately, minimizing f over all potential partitions of V is NP-hard and, thus, so is the densest k-partition problem (see Peng and Wei [35]). It should be noted that the complexity of the densest k-disjoint-clique subgraph problem is unknown, although the problem of minimizing f over all k-disjoint-clique subgraphs has the trivial solution of assigning exactly one point to each cluster and setting all other points to be outliers.

If we let X be the $N \times k$ matrix with ith column equal to $\mathbf{v}_i / \|\mathbf{v}_i\|$, we have $\sum_{i=1}^{k} d_{G(C_i)} = \mathrm{Tr}\,(X^T W X)$. We call such a matrix X a normalized k-partition matrix. That is, X is a *normalized k-partition matrix* if the columns of X are the normalized characteristic vectors of k disjoint subsets of V. We denote by $npm(V, k)$ the set of all normalized k-partition matrices of V. We should note that the term normalized k-partition matrix is a slight misnomer; the columns of $X \in npm(V, k)$ do not necessarily define a partition of V into k disjoint sets but do define a partition of V into the $k + 1$ disjoint sets given by the columns $X(:, 1), \ldots, X(:, k)$ of X and their complement. Using this notation, the densest k-disjoint-clique problem may be formulated as the quadratic program

$$\max\{\mathrm{Tr}\,(X^T W X) : X \in npm(V, k)\}. \tag{2.3}$$

Unfortunately, quadratic programs with combinatorial constraints are NP-hard in general.

The quadratic program (2.3) may be relaxed to a rank constrained semidefinite program using matrix lifting. We replace each column $\mathbf{x}_i$ of X with a rank-one semidefinite variable $\mathbf{x}_i \mathbf{x}_i^T$ to obtain the new decision variable $\tilde{X} = \sum_{i=1}^{k} \mathbf{x}_i \mathbf{x}_i^T$. The new variable $\tilde{X}$ has both rank and trace exactly equal to k since the summands $\mathbf{x}_i \mathbf{x}_i^T$ are orthogonal rank-one matrices, each with nonzero eigenvalue equal to 1. Moreover, since $[\mathbf{x}_i]_j = 1/\sqrt{r_i}$ for all $j \in C_i$ and all remaining components equal to 0 where r_i is equal to the number of nonzero entries of $\mathbf{x}_i$, we have

$$\tilde{X}\mathbf{e} = \sum_{i=1}^{k} \mathbf{x}_i (\mathbf{x}_i^T \mathbf{e}) = \sum_{i=1}^{k} \sqrt{r_i}\mathbf{x}_i.$$

Thus, the matrix $\tilde{X}$ has row sum equal to one for each vertex in the subgraph of K_N defined by the columns of X and zero otherwise. Therefore, we may relax (2.3) as the rank constrained semidefinite program

$$\max \left\{ \mathrm{Tr}\,(WX) : X\mathbf{e} \leq \mathbf{e}, \mathrm{rank}\, X = k, \mathrm{Tr}\, X = k, X \geq 0, X \in \Sigma_+^N \right\} \tag{2.4}$$

Here Σ_+^N denotes the cone of $N \times N$ symmetric positive semidefinite matrices and $\mathbf{e}$ denotes the all-ones vector in R^N. The nonconvex program (2.4) may be relaxed further to a semidefinite program by ignoring the nonconvex constraint $\mathrm{rank}\,(X) = k$:

$$\max \left\{ \mathrm{Tr}\,(WX) : X\mathbf{e} \leq \mathbf{e}, \mathrm{Tr}\, X = k, X \geq 0, X \in \Sigma_+^N \right\}. \tag{2.5}$$

Note that a k-disjoint-clique subgraph with vertex set composed of disjoint cliques $C_1, \ldots, C_k$ defines a feasible solution of (2.5) with rank exactly equal to k and objective value equal to (2.1) by

$$X^* = \sum_{i=1}^{k} \frac{\mathbf{v}_i \mathbf{v}_i^T}{\mathbf{v}_i^T \mathbf{v}_i}, \tag{2.6}$$

where $\mathbf{v}_i$ is the characteristic vector of C_i for all $i = 1, \ldots, k$. This feasible solution is exactly the lifted solution corresponding to the cliques $\{C_1, \ldots, C_k\}$. This relaxation approach mirrors that for the planted k-disjoint-clique problem considered in Ames and Vavasis [3]. In Ames and Vavasis [3], entrywise nonnegativity constraints can be ignored for the sake of computational efficiency due to explicit constraints forcing all entries of a feasible solution corresponding to unadjacent nodes to be equal to 0. Due to the lack of such constraints in (2.5), the nonnegativity constraints are required to ensure that the optimal solution of (2.5) is unique if the input data is sufficiently clusterable. Indeed, suppose that

$$W = \begin{pmatrix} \mathbf{e}\mathbf{e}^T & 0 \\ 0 & \mathbf{e}\mathbf{e}^T \end{pmatrix},$$

where $\mathbf{e}$ is the all-ones vector in $\mathbf{R}^n$. Then both

$$\frac{1}{n}\begin{pmatrix} \mathbf{e}\mathbf{e}^T & 0 \\ 0 & \mathbf{e}\mathbf{e}^T \end{pmatrix} \quad \text{and} \quad \frac{1}{n}\begin{pmatrix} \mathbf{e}\mathbf{e}^T & -\mathbf{e}\mathbf{e}^T \\ -\mathbf{e}\mathbf{e}^T & \mathbf{e}\mathbf{e}^T \end{pmatrix}$$

are positive semidefinite, have trace equal to 2, row sums bounded above by 1, and have objective value equal to $2n$. Therefore, the nonnegativity constraints are necessary to distinguish between these two solutions. We should also point out that the constraints of (2.5) are similar to those of the semidefinite relaxation used to approximate the

minimum sum of squared Euclidean distance partition by Peng and Wei [35], although with different derivation.

The relaxation (2.5) may be thought of as a nuclear norm relaxation of (2.4). Indeed, since the eigenvalues and singular values of a positive semidefinite matrix are identical, every feasible solution X satisfies $\mathrm{Tr}\,(X) = \sum_{i=1}^{N} \sigma_i(X) = \|X\|_*$. Moreover, since every feasible solution X is symmetric and has row sums at most 1, we have $\|X\|_1 = \|X\|_\infty \leq 1$ for every feasible X. Here $\|\cdot\|_1$, $\|\cdot\|$, and $\|\cdot\|_\infty$ denote the matrix norms on $\mathbf{R}^{N\times N}$ induced by the ℓ_1, ℓ_2, and ℓ_∞ norms on $\mathbf{R}^N$ respectively. This implies that every feasible X satisfies $\|X\| \leq 1$ since $\|X\| \leq \sqrt{\|X\|_1\|X\|_\infty}$ (see [23, Corollary 2.3.2]). Since $\|X\|_*$ is the convex envelope of rank (X) on the set $\{X : \|X\| \leq 1\}$ (see, for example, [37, Theorem 2.2]), the semidefinite program (2.5) is exactly the relaxation of (2.4) obtained by ignoring the rank constraint and only constraining the nuclear norm of a feasible solution. Many recent results have shown that the minimum rank solution of a set of linear equations $\mathcal{A}(X) = \mathbf{b}$ is equal to the minimum nuclear norm solution, under certain assumption on the linear operator $\mathcal{A}$. We would like to prove analogous results for the relaxation (2.5). That is, we would like to identify conditions on the input graph that guarantee recovery of the densest k-disjoint-clique subgraph by solving (2.5).

Ideally, a clustering heuristic should be able to correctly identify the clusters in data that is known a priori to be clusterable. In our graph theoretic model, this case corresponds to a graph $G_S = (V, E, W)$ admitting a k-disjoint-clique subgraph with very high weights on edges connecting nodes within the cliques and relatively low weights on edges between different cliques. We focus our attention on input instances for the densest k-disjoint-clique problem that are constructed to possess this structure. Let K^* be a k-disjoint-clique subgraph of K_N with vertex set composed of disjoint cliques $C_1, C_2, \ldots, C_k$. We consider random symmetric matrices $W \in \Sigma^N$ with entries sampled independently from one of two distributions Ω_1, Ω_2 as follows:

- For each $q = 1, \ldots, k$, the entries of each diagonal block W_{C_q,C_q} are independently sampled from a probability distribution Ω_1 satisfying $\mathbf{E}[W_{ij}] = \mathbf{E}[W_{ji}] = \alpha$ and $W_{ij} \in [0, 1]$ for all $i, j \in C_q$.
- All remaining entries of W are independently sampled from a probability distribution Ω_2 satisfying $\mathbf{E}[W_{ij}] = \mathbf{E}[W_{ji}] = \beta$ and $W_{ij} \in [0, 1]$ for all $(i, j) \in (V \times V) \backslash \cup_{q=1}^{k} (C_q \times C_q)$.

That is, we sample the random variable W_{ij} from the probability distribution Ω_1 with mean α if the nodes i, j are in the same planted clique; otherwise, we sample W_{ij} from the distribution Ω_2 with mean β. We say that such random matrices W are sampled from the *planted cluster model*. We should note the planted cluster model is a generalization of the planted k-disjoint-clique subgraph model considered in Ames and Vavasis [3], as well as the stochastic block/probabilistic cluster model considered in Jalali et al. [26], Oymak and Hassibi [33], Rohe et al. [40]. Indeed, the stochastic block model is generated by independently adding edges within planted dense subgraphs with probability p and independently adding edges between cliques with probability q for some $p > q$. The planted k-disjoint-clique subgraph model is simply the stochastic block model in the special case that $p = 1$. Therefore, choosing Ω_1 and Ω_2 to be

Bernoulli distributions with probabilities of success p and q, respectively, yields W sampled from the stochastic block model.

The following theorem describes which partitions $\{C_1, C_2, \ldots, C_{k+1}\}$ of V yield random symmetric matrices W drawn from the planted cluster model such that the corresponding planted k-disjoint-clique subgraph K is the densest k-disjoint-clique subgraph and can be found with high probability by solving (2.5).

Theorem 2.1 *Suppose that the vertex sets $C_1, \ldots, C_k$ define a k-disjoint-clique subgraph K^* of the complete graph $K_N = (V, E)$ on N vertices and let $C_{k+1} := V \backslash (\cup_{i=1}^{k} C_i)$. Let $r_i := |C_i|$ for all $i = 1, \ldots, k+1$, and let $\hat{r} = \min_{i=1,\ldots,k} r_i$. Let $W \in \Sigma^N$ 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*

$$\gamma = \gamma(\alpha, \beta, r) := \alpha(1 + \delta_{0,r_{k+1}}) - 2\beta > 0, \tag{2.7}$$

where $\delta_{i,j}$ is the Kronecker delta function defined by $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise. Let X^* be the feasible solution for (2.5) corresponding to $C_1, \ldots, C_k$ defined by (2.6). Then there exist scalars $c_1, c_2, c_3 > 0$ such that if

$$c_1\sqrt{N} + c_2\sqrt{kr_{k+1}} + c_3 r_{k+1} \leq \gamma \hat{r} \tag{2.8}$$

then X^* is the unique optimal solution for (2.5), and K^* is the unique maximum density k-disjoint-clique subgraph of K_N corresponding to W with probability tending exponentially to 1 as $\hat{r} \to \infty$.

Note that the condition (2.7) implies that $\alpha > \beta$ if $r_{k+1} = 0$ and $\alpha > 2\beta$ otherwise. That is, if $\{C_1, \ldots, C_k\}$ defines a partition of V then the restriction that $\alpha > 2\beta$ can be relaxed to $\alpha > \beta$. On the other hand, the condition (2.8) cannot be satisfied unless $N = O(\hat{r}^2)$ and $r_{k+1} = O(\hat{r})$. We now provide a few examples of $r_1, \ldots, r_k$ satisfying the hypothesis of Theorem 2.1.

- Suppose that we have k cliques $C_1, \ldots, C_k$ of size $r_1 = r_2 = \cdots = r_k = N^\epsilon$. Then (2.8) implies that we may recover the k-disjoint-clique subgraph corresponding to $C_1, \ldots, C_k$ if $N^\epsilon \geq \Omega(N^{1/2})$. Since the cliques $C_1, \ldots, C_k$ are disjoint and contain $\Omega(N)$ nodes, we must have $\epsilon \geq 1/2$. Therefore, our heuristic may recover $O(N^{1/2})$ planted cliques of size $N^{1/2}$.
- On the other hand, we may have cliques of different sizes. For example, suppose that we wish to recover k_1 cliques of size $N^{3/4}$ and k_2 smaller cliques of size $N^{1/2}$. Then the right-hand side of (2.8) must be at least $\Omega\big(\sqrt{N} + \sqrt{(k_1 + k_2)r_{k+1}} + r_{k+1}\big)$. Therefore, we may recover the planted cliques provided that $k_1 = O(N^{1/4})$, $k_2 = O(N^{1/2})$, and $r_{k+1} = O(N^{1/2})$.

Although we consider a more general model for clustered data, our recovery guarantee agrees (up to constants) with those existing in the literature. In particular, the bound on the minimum size of the planted clique recoverable by the relaxation (2.5), $\hat{r} = \Omega(N^{1/2})$, provided by Theorem 2.1 matches that given in Jalali et al. [26], Oymak and Hassibi [33]. However, among the existing recovery guarantees in the literature,

few consider noise in the form of diversionary nodes. As a consequence of our more general model, the relaxation (2.5) is exact for input graphs containing up to $O(\hat{r})$ noise nodes, fewer than the bound, $O(\hat{r}^2)$, provided by [3, Theorem 4.5].

3 A semidefinite relaxation of the densest k-disjoint-biclique problem

Given a bipartite graph $G = ((U, V), E)$, a pair of disjoint independent subsets $U' \subseteq U,\ V' \subseteq V$ is a *biclique* of G if the subgraph of G induced by (U', V') is complete bipartite. That is, (U', V') is a biclique of G if $uv \in E$ for all $u \in U', v \in V'$. A *k-disjoint-biclique subgraph* of G is a subgraph of G with vertex set composed of k disjoint bicliques of G. Let $K_{M,N} = ((U, V), E, W)$ be a weighted complete bipartite graph with vertex sets $U = \{1, 2, \ldots, M\}$, $V = \{1, \ldots, N\}$ with matrix of edge weights $W \in [0, 1]^{U \times V}$. We are interested in identifying the densest k-disjoint-biclique subgraph of $K_{M,N}$ with respect to W. We define the density of a subgraph $H = (U', V', E')$ of $K_{M,N}$ to be the total edge weight incident at each vertex divided by the square root of the number of edges from U' to V':

$$d_H = \frac{1}{\sqrt{|E'|}} \sum_{u \in U', v \in V'} W_{uv}. \tag{3.1}$$

Note that the density of H, as defined by (3.1), is not necessarily equal to the average edge weight incident at a vertex of H, since the square root of the number of edges is not equal to the total number of vertices if $|U'| \neq |V'|$ or H is not complete. The goal of the *densest k-disjoint-biclique problem* is to identify a set of k disjoint bicliques of $K_{M,N}$ such that the sum of the densities of the complete subgraphs induced by these bicliques is maximized. That is, we want to find a set of k disjoint bicliques, with characteristic vectors $(\mathbf{u}_1, \mathbf{v}_1), \ldots, (\mathbf{u}_k, \mathbf{v}_k)$, maximizing the sum

$$\sum_{i=1}^{k} \frac{\mathbf{u}_i^T W \mathbf{v}_i}{\|\mathbf{u}_i\| \|\mathbf{v}_i\|}. \tag{3.2}$$

As in our analysis of the densest k-disjoint-clique problem, this problem may be posed as the nonconvex quadratic program

$$\max\{\mathrm{Tr}\,(X^T W Y) : X \in npm(U),\ Y \in npm(V)\}. \tag{3.3}$$

By letting $Z = (X^T,\ Y^T)^T (X^T,\ Y^T)$, we have $\mathrm{Tr}\,(X^T W Y) = \frac{1}{2}\mathrm{Tr}\,(\tilde{W} Z)$, where

$$\tilde{W} = \begin{pmatrix} 0 & W \\ W^T & 0 \end{pmatrix}.$$

Using this change of variables, we relax to the rank constrained semidefinite program

$$\begin{array}{ll} \max & \frac{1}{2}\mathrm{Tr}\,(\tilde{W}Z) \\ \text{s.t.} & Z_{U,U}\mathbf{e} \leq \mathbf{e}, \quad Z_{V,V}\mathbf{e} \leq \mathbf{e}, \\ & \mathrm{Tr}\,(Z_{U,U}) = k, \quad \mathrm{Tr}\,(Z_{V,V}) = k, \\ & \mathrm{rank}\,(Z_{U,U}) = k, \quad \mathrm{rank}\,(Z_{V,V}) = k, \\ & Z \geq 0, \quad Z \in \Sigma_+^{M+N}, \end{array} \tag{3.4}$$

where $Z_{U,U}$ and $Z_{V,V}$ are the blocks of Z with rows and columns indexed by U and V respectively. Ignoring the nonconvex rank constraints yields the semidefinite relaxation

$$\begin{array}{ll} \max & \frac{1}{2}\mathrm{Tr}\,(\tilde{W}Z) \\ \text{s.t.} & Z_{U,U}\mathbf{e} \leq \mathbf{e}, \quad Z_{V,V}\mathbf{e} \leq \mathbf{e}, \\ & \mathrm{Tr}\,(Z_{U,U}) = k, \quad \mathrm{Tr}\,(Z_{V,V}) = k, \\ & Z \geq 0, \quad Z \in \Sigma_+^{M+N}. \end{array} \tag{3.5}$$

As in our analysis of the densest k-disjoint-clique problem, we would like to identify sets of program instances of the k-disjoint-biclique problem that may be solved using the semidefinite relaxation (3.5). As before, we consider input graphs where it is known a priori that a k-disjoint-biclique subgraph with large edge weights, relative to the edges of its complement, exists. We consider random program instances generated as follows. Let G^* be a k-disjoint-biclique subgraph of $K_{M,N}$ with vertex set composed of the disjoint bicliques $(U_1, V_1), \ldots, (U_k, V_k)$. We construct a random matrix $W \in \mathbf{R}_+^{M\times N}$ with entries sampled independently from one of two distributions Ω_1, Ω_2 as follows.

- If $u \in U_i, v \in V_i$ for some $i \in \{1, \ldots, k\}$, then we sample W_{uv} from the distribution Ω_1, with mean α. If u and v are in different bicliques of K^*, then we sample W_{uv} according to the probability distribution Ω_2, with mean $\beta < \alpha$.
- The probability distributions Ω_1, Ω_2 are chosen such that $u \in U, v \in V$, $0 \leq W_{uv} \leq 1$.

We say that such W are sampled from the *planted bicluster model*. Note that G^* defines a feasible solution for (3.5) by

$$Z^* = \sum_{i=1}^{k} \begin{pmatrix} \mathbf{u}_i/\|\mathbf{u}_i\| \\ \mathbf{v}_i/\|\mathbf{v}_i\| \end{pmatrix} \begin{pmatrix} \mathbf{u}_i/\|\mathbf{u}_i\| \\ \mathbf{v}_i/\|\mathbf{v}_i\| \end{pmatrix}^T, \tag{3.6}$$

where $\mathbf{u}_i, \mathbf{v}_i$ are the characteristic vectors of U_i and V_i, respectively, for all $i = 1, \ldots, k$. Moreover, Z^* has objective value equal to (3.2). The following theorem describes which partitions $\{U_1, \ldots, U_k\}$ and $\{V_1, \ldots, V_k\}$ of U and V yield random matrices W drawn from the planted bicluster model such that Z^* is the unique optimal solution of the semidefinite relaxation (3.5) and G^* is the unique densest k-disjoint-biclique subgraph.

Theorem 3.1 *Suppose that the vertex sets $(U_1, V_1), \ldots, (U_k, V_k)$ define a k-disjoint-biclique subgraph K^* of the complete bipartite graph $K_{M,N} = ((U, V), E)$. Let*

$U_{k+1} := U \backslash (\cup_{i=1}^{k} U_i)$ and $V_{k+1} := V \backslash (\cup_{i=1}^{k} V_i)$. Let $m_i = |U_i|$ and $n_i = |V_i|$ for all $i = 1, \ldots, k+1$ and $\hat{n} := \min_{i=1,\ldots,k} n_i$. Let Z^* be the feasible solution for (3.5) corresponding to K^* given by (3.6). Let $W \in \mathbf{R}_+^{M \times N}$ be a random matrix sampled from the planted bicluster model according to distributions Ω_1 and Ω_2 with means α, β satisfying

$$\gamma = \gamma(\alpha, \beta, m, n) := \alpha(1 + \delta_{0,m_{k+1}} \delta_{0,n_{k+1}}) - 2\beta > 0. \tag{3.7}$$

Suppose that there exist scalars $\{\tau_1, \ldots, \tau_{k+1}\}$ such that $m_i = \tau_i^2 n_i$ for all $i \in \{1, \ldots, k+1\}$ and

$$\alpha \tau_i > \beta \tau_j \tag{3.8}$$

for all $i, j \in \{1, \ldots, k+1\}$. Then there exist scalars $c_1, c_2 > 0$ depending only on α, β, and $\{\tau_1, \ldots, \tau_{k+1}\}$ such that if

$$c_1 \left(\sqrt{k} + \sqrt{n_{k+1}} + 1 \right) \sqrt{N} + \beta \tau_{k+1} n_{k+1} \leq c_2 \gamma \hat{n} \tag{3.9}$$

then Z^* is the unique optimal solution of (3.5) and G^* is the unique maximum density k-disjoint-biclique subgraph with respect to W with probability tending exponentially to 1 as $\hat{n}$ tends to ∞.

For example, Theorem 3.1 implies that $O(N^{1/3})$ bicliques of size $\hat{m} = \hat{n} = N^{2/3}$ can be recovered from a graph sampled from the planted bicluster model with up to $O(N^{1/3})$ diversionary nodes by solving (3.5).

4 Proof of Theorem 2.1

This section comprises a proof of Theorem 2.1. The proof of Theorem 3.1 is essentially identical to that of Theorem 2.1, although with some modifications made to accommodate the different relaxation and lack of symmetry of the weight matrix W; an outline of the proof of Theorem 3.1 is given in Sect. 5.

4.1 Optimality conditions

We begin with the following sufficient condition for the optimality of a feasible solution of (2.5).

Theorem 4.1 *Let X be feasible for (2.5) and suppose that there exist some $\mu \in \mathbf{R}$, $\lambda \in \mathbf{R}_+^N$, $\eta \in \mathbf{R}_+^{N \times N}$ and $S \in \Sigma_+^N$ such that*

$$-W + \lambda \mathbf{e}^T + \mathbf{e} \lambda^T - \eta + \mu I = S \tag{4.1}$$

$$\lambda^T (X\mathbf{e} - \mathbf{e}) = 0 \tag{4.2}$$

$$\mathrm{Tr}\,(X\eta) = 0 \tag{4.3}$$

$$\mathrm{Tr}\,(XS) = 0. \tag{4.4}$$

Then X is optimal for (2.5).

Note that $X = (k/N - \epsilon)I + \epsilon \mathbf{e}\mathbf{e}^T$ is a strictly feasible solution of (2.5) for sufficiently small $\epsilon > 0$. Thus, Slater's constraint qualification holds for (2.5). Therefore, a feasible solution X is optimal for (2.5) if and only if it satisfies the Karush-Kuhn-Tucker conditions. Theorem 4.1 provides the necessary specialization to (2.5) of these necessary and sufficient conditions (see, for example, [10, Section 5.5.3] or [39, Theorem 28.3]).

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 X^* be the corresponding feasible solution of (2.5) defined by (2.6). Let $C_{k+1} := V \backslash (\cup_{i=1}^k C_i)$ and $r_{k+1} := N - \sum_{i=1}^k r_i$. Let $\hat{r} := \min_{i=1,\ldots,k} r_i$. Let $W \in \Sigma^N$ be a random symmetric matrix sampled from the planted cluster model according to Ω_1 and Ω_2 with means α and β. To show that X^* is optimal for (2.5), we will construct multipliers $\mu \in \mathbf{R}$, $\lambda \in \mathbf{R}_+^N$, $\eta \in \mathbf{R}_+^{N \times N}$, and $S \in \Sigma_+^N$ satisfying (4.1), (4.2), (4.3), and (4.4). Note that the gradient Eq. (4.1) provides an explicit formula for the multiplier S for any choice of multipliers μ, λ, and η.

The proof of Theorem 2.1 uses techniques similar to those used in Ames and Vavasis [3]. Specifically, the proof of Theorem 2.1 relies on constructing multipliers satisfying Theorem 4.1. The multipliers λ and η will be constructed in blocks inherited from the block structure of the proposed solution X^*. Again, once the multipliers μ, λ, and η are chosen, (4.1) provides an explicit formula for the multiplier S.

The dual variables must be chosen so that the complementary slackness condition (4.4) is satisfied. The condition $\mathrm{Tr}\,(X^*S) = 0$ is satisfied if and only if $X^*S = 0$, since both X^* and S are desired to be positive semidefinite (see [44, Proposition 1.19]). Therefore, the multipliers must be chosen so that the left-hand side of (4.1) is orthogonal to the columns of X^*. That is, we must choose the multipliers μ, λ, and η such that S, as defined by (4.1), has nullspace containing the columns of X^*. By the special block structure of X^*, this is equivalent to requiring S_{v,C_q} to sum to 0 for all $q \in \{1, \ldots, k\}$ and $v \in V$.

The gradient condition Eq. (4.1), coupled with the requirement that the columns of X^* reside in the nullspace of S, provides an explicit formula for the multiplier λ. Moreover, the complementary slackness condition (4.3) implies that all diagonal blocks η_{C_q,C_q}, $q = 1, \ldots, k$, are equal to 0. To construct the remaining multipliers, we parametrize the remaining blocks of S using the vectors $\mathbf{y}^{q,s}$ and $\mathbf{z}^{q,s}$ for all $q \neq s$. These vectors are chosen to be the solutions of the system of linear equations defined by $SX^* = X^*S = 0$. As in Ames and Vavasis [3], we will show that this system is a perturbation of a linear system with known solution and will use this known solution to obtain estimates of $\mathbf{y}^{q,s}$ and $\mathbf{z}^{q,s}$.

Once the multipliers are chosen, we must establish dual feasibility to prove that X^* is optimal for (2.5). In particular, we must show that λ and η are nonnegative and S is positive semidefinite. To establish nonnegativity of λ and η, we will show that these multipliers are strictly positive in expectation and close to their respective means with extremely high probability. To establish that S is positive semidefinite, we will show that the diagonal blocks of S dominate the off diagonal blocks with high probability.

4.2 Choice of the multipliers and a sufficient condition for uniqueness and optimality

We construct the multipliers λ, η, and S in blocks indexed by the vertex sets $C_1, \ldots, C_{k+1}$. The complementary slackness condition (4.4) implies that the columns of X are in the nullspace of S since $\mathrm{Tr}\,(XS) = 0$ if and only if $XS = 0$ for all positive semidefinite X, S. Since $X^*_{C_q,C_q}$ is a multiple of the all-ones matrix $\mathbf{e}\mathbf{e}^T$ for each $q = 1, \ldots, k$, and all other entries of X^* are equal to 0, (4.4) implies that the block S_{C_q,C_s} must have row and column sums equal to 0 for all $q, s \in \{1, \ldots, k\}$. Moreover, since all entries of $X^*_{C_q,C_q}$ are nonzero, $\eta_{C_q,C_q} = 0$ for all $q = 1, \ldots, k$ by (4.3).

To compute an explicit formula for λ, note that the condition $S_{C_q,C_q}\mathbf{e} = 0$ is satisfied if

$$0 = S_{C_q,C_q}\mathbf{e} = \mu\mathbf{e} + r_q\lambda_{C_q} + (\lambda^T_{C_q}\mathbf{e})\mathbf{e} - W_{C_q,C_q}\mathbf{e} \tag{4.5}$$

for all $q = 1, \ldots, k$. Rearranging (4.5) shows that λ_{C_q} is the solution to the system

$$(r_q I + \mathbf{e}\mathbf{e}^T)\lambda_{C_q} = W_{C_q,C_q}\mathbf{e} - \mu\mathbf{e} \tag{4.6}$$

for all $q = 1, \ldots, k$. We will use the Sherman-Morrision-Woodbury formula (see, for example, [23, Equation (2.1.4)]), stated in the following lemma, to obtain the desired formula for λ.

Lemma 4.1 *Let $A \in \Sigma^{n\times n}$ be nonsingular and $\mathbf{u}, \mathbf{v} \in \mathbf{R}^n$ be such that $1 + \mathbf{v}^T A^{-1}\mathbf{u} \neq 0$. Then*

$$(A + UV^T)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^T A^{-1}}{1 + \mathbf{v}^T A^{-1}\mathbf{u}}. \tag{4.7}$$

Applying (4.7) with $A = r_q I$, $\mathbf{u} = \mathbf{v} = \mathbf{e}$ shows that choosing

$$\lambda_{C_q} = \frac{1}{r_q}\left(W_{C_q,C_q}\mathbf{e} - \frac{1}{2}\left(\mu + \frac{\mathbf{e}^T W_{C_q,C_q}\mathbf{e}}{r_q}\right)\mathbf{e}\right) \tag{4.8}$$

ensures that $\mathrm{Tr}\,(S_{C_q,C_q}X^*_{C_q,C_q}) = 0$ for all $q = 1, \ldots, k$.

We next construct η. Fix $q, s \in \{1, \ldots, k+1\}$ such that $q \neq s$. To ensure that $S_{C_q,C_s}\mathbf{e} = 0$ and $S_{C_s,C_q}\mathbf{e} = 0$, we parametrize the entries of η_{C_q,C_s} using the vectors $\mathbf{y}^{q,s}$ and $\mathbf{z}^{q,s}$. In particular, we take

$$\eta_{C_q,C_s} = \left(\frac{\bar{\delta}_{q,k+1}}{2}\left(\alpha - \frac{\mu}{r_q}\right) + \frac{\bar{\delta}_{s,k+1}}{2}\left(\alpha - \frac{\mu}{r_s}\right) - \beta\right)\mathbf{e}\mathbf{e}^T + \mathbf{y}^{q,s}\mathbf{e}^T + \mathbf{e}(\mathbf{z}^{q,s})^T. \tag{4.9}$$

Here $\bar{\delta}_{ij} := 1 - \delta_{ij}$, where δ_{ij} is the Kronecker delta function defined by $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. That is, we take η_{C_q,C_s} to be the expected value of $\lambda_{C_q}\mathbf{e}^T + \mathbf{e}\lambda^T_{C_s} - W_{C_q,C_s}$, plus the parametrizing terms $\mathbf{y}^{q,s}\mathbf{e}^T$ and $\mathbf{e}(\mathbf{z}^{q,s})^T$. The vectors

$\mathbf{y}^{q,s}$ and $\mathbf{z}^{q,s}$ are chosen to be the solutions to the systems of linear equations imposed by the requirement that $X^*S = SX^* = 0$. As we will see, this system of linear equations is a perturbation of a linear system with known solution. Using the solution of the perturbed system we obtain bounds on $\mathbf{y}^{q,s}$ and $\mathbf{z}^{q,s}$, which are used to establish that η is nonnegative and S is positive semidefinite.

Let

$$\tilde{\eta}_{C_q,C_s} := \lambda_{C_q}\mathbf{e}^T + \mathbf{e}\lambda_{C_s}^T - W_{C_q,C_s}. \tag{4.10}$$

Note that the symmetry of W implies that $\tilde{\eta}_{C_s,C_q} = \tilde{\eta}_{C_q,C_s}^T$. Let $\mathbf{b} = \mathbf{b}^{q,s} \in \mathbf{R}^{C_q \cup C_s}$ be defined by $\mathbf{b}_{C_q} := \tilde{\eta}_{C_q,C_s}\mathbf{e} - \mathbf{E}[\tilde{\eta}_{C_q,C_s}]\mathbf{e}$ and $\mathbf{b}_{C_s} = \tilde{\eta}_{C_s,C_q}\mathbf{e} - \mathbf{E}[\tilde{\eta}_{C_s,C_q}]\mathbf{e}$. We choose $\mathbf{y} = \mathbf{y}^{q,s}$ and $\mathbf{z} = \mathbf{z}^{q,s}$ to be solutions of the system

$$\begin{pmatrix} r_s I + \theta \mathbf{e}\mathbf{e}^T & (1-\theta)\mathbf{e}\mathbf{e}^T \\ (1-\theta)\mathbf{e}\mathbf{e}^T & r_q I + \theta \mathbf{e}\mathbf{e}^T \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \mathbf{b} \tag{4.11}$$

for some scalar $\theta > 0$ to be defined later. The requirement that the row sums of S_{C_q,C_s} are equal to zero is equivalent to $\mathbf{y}$ and $\mathbf{z}$ satisfying the system of linear equations

$$\begin{aligned} 0 = &-r_s\mathbf{y}_i - \mathbf{z}^T\mathbf{e} + r_s\left(\lambda_i - \frac{\bar{\delta}_{q,k+1}}{2r_q}(\alpha r_q - \mu)\right) + \left(\lambda_{C_s}^T\mathbf{e} - \frac{\bar{\delta}_{s,k+1}}{2}(\alpha r_s - \mu)\right) \\ &- ([W_{C_q,C_s}\mathbf{e}]_i - r_s\beta) \end{aligned} \tag{4.12}$$

for all $i \in C_q$. Similarly, the column sums of S_{C_q,C_s} are equal to zero if and only if $\mathbf{y}$ and $\mathbf{z}$ satisfy

$$\begin{aligned} 0 = &-r_q\mathbf{z}_i - \mathbf{y}^T\mathbf{e} + r_q\left(\lambda_i - \frac{\bar{\delta}_{s,k+1}}{2r_s}(\alpha r_s - \mu)\right) + \left(\lambda_{C_q}^T\mathbf{e} - \frac{\bar{\delta}_{q,k+1}}{2}(\alpha r_q - \mu)\right) \\ &- ([W_{C_s,C_q}\mathbf{e}]_i - r_q\beta) \end{aligned} \tag{4.13}$$

for all $i \in C_s$. Note that the system of equations defined by (4.12) and (4.13) is equivalent to (4.11) in the special case that $\theta = 0$. However, when $\theta = 0$, the system of equations in (4.11) is singular, with nullspace spanned by the vector $(\mathbf{e}; -\mathbf{e})$. When θ is nonzero, each row of the system (4.11) has an additional term of the form $\theta(\mathbf{e}^T\mathbf{y} - \mathbf{e}^T\mathbf{z})$. However, any solution $(\mathbf{y}; \mathbf{z})$ of (4.11) for $\theta > 0$ is also a solution in the special case that $\theta = 0$. Indeed, since $(\mathbf{e}; -\mathbf{e})$ is in the nullspace of the matrix

$$\begin{pmatrix} r_s I & \mathbf{e}\mathbf{e}^T \\ \mathbf{e}\mathbf{e}^T & r_q I \end{pmatrix}$$

and $\mathbf{b}_{C_q}^T\mathbf{e} = \mathbf{b}_{C_s}^T\mathbf{e}$, taking the inner product of each side of (4.11) with $(\mathbf{e}; -\mathbf{e})$ yields

$$\theta(r_q + r_s)(\mathbf{e}^T\mathbf{y} - \mathbf{e}^T\mathbf{z}) = \mathbf{b}_{C_q}^T\mathbf{e} - \mathbf{b}_{C_s}^T\mathbf{e} = 0.$$

Therefore, the unique solution $(\mathbf{y}; \mathbf{z})$ of (4.11) also satisfies (4.12) and(4.13) for any $\theta > 0$ such that (4.11) is nonsingular. In particular, note that (4.11) is nonsingular for $\theta = 1$. For this choice of θ, $\mathbf{y}$ and $\mathbf{z}$ are the unique solutions of the systems $(r_s I + \mathbf{e}\mathbf{e}^T)\mathbf{y} = \mathbf{b}_1$ and $(r_q I + \mathbf{e}\mathbf{e}^T)\mathbf{z} = \mathbf{b}_2$, where $\mathbf{b}_1 := \mathbf{b}_{C_q}$ and $\mathbf{b}_2 := \mathbf{b}_{C_s}$. Applying (4.7) with $A = r_s I$, $\mathbf{u} = \mathbf{v} = \mathbf{e}$ and $A = r_q I$, $\mathbf{u} = \mathbf{v} = \mathbf{e}$ yields

$$\mathbf{y} = \frac{1}{r_s}\left(\mathbf{b}_1 - \frac{(\mathbf{b}_1^T\mathbf{e})}{r_q + r_s}\mathbf{e}\right) \quad \text{and} \quad \mathbf{z} = \frac{1}{r_q}\left(\mathbf{b}_2 - \frac{(\mathbf{b}_2^T\mathbf{e})}{r_q + r_s}\mathbf{e}\right) \tag{4.14}$$

respectively. Finally, we choose $\mu = \epsilon\gamma\hat{r}$, where $\gamma = \gamma(\alpha, \beta, r) = \alpha(1 + \delta_{0,r_{k+1}}) - 2\beta$, and $\epsilon > 0$ is a scalar to be chosen later.

In summary, we choose the multipliers $\mu \in \mathbf{R}$, $\lambda \in \mathbf{R}^N$, $\eta \in \mathbf{R}^{N\times N}$ as follows:

$$\mu = \epsilon\gamma\hat{r} \tag{4.15}$$

$$\lambda_{C_q} = \begin{cases} \dfrac{1}{r_q}\left(W_{C_q,C_q}\mathbf{e} - \dfrac{1}{2}\left(\mu + \dfrac{\mathbf{e}^T W_{C_q,C_q}\mathbf{e}}{r_q}\right)\mathbf{e}\right), & \text{if } q \in \{1,\ldots,k\} \\ 0, & \text{if } q = k+1 \end{cases} \tag{4.16}$$

$$\eta_{C_q,C_s} = \begin{cases} \mathbf{E}[\tilde{\eta}_{C_q,C_s}] + \mathbf{y}^{q,s}\mathbf{e}^T + \mathbf{e}(\mathbf{z}^{q,s})^T, & \text{if } q,s \in \{1,\ldots,k+1\}, q \neq s \\ 0, & \text{otherwise} \end{cases} \tag{4.17}$$

where $\epsilon > 0$ is a scalar to be defined later, $\tilde{\eta}_{C_q,C_s}$ is defined as in (4.10), and $\mathbf{y}^{q,s}$, $\mathbf{z}^{q,s}$ are given by (4.14) for all $q, s \in \{1,\ldots,k+1\}$ such that $q \neq s$. We choose S according to (4.1). Finally, we define the $(k+1)\times(k+1)$ block matrix $\tilde{S}$ in Σ^N by

$$\tilde{S}_{C_q,C_s} = \begin{cases} \alpha\mathbf{e}\mathbf{e}^T - W_{C_q,C_s}, & \text{if } q = s,\ q,s \in \{1,\ldots,k\} \\ \beta\mathbf{e}\mathbf{e}^T - W_{C_q,C_s}, & \text{if } q \neq s,\ q,s \in \{1,\ldots,k\} \\ \beta\mathbf{e}\mathbf{e}^T - W_{C_q,C_s} + (\lambda_{C_q} - \mathbf{E}[\lambda_{C_q}])\mathbf{e}^T, & \text{if } s = k+1 \\ \beta\mathbf{e}\mathbf{e}^T - W_{C_q,C_s} + \mathbf{e}(\lambda_{C_s} - \mathbf{E}[\lambda_{C_s}])^T, & \text{if } q = k+1. \end{cases} \tag{4.18}$$

We conclude with the following theorem, which provides a sufficient condition ensuring that the proposed solution X^* is the unique optimal solution for (2.5) and K^* is the unique maximum density k-disjoint-clique subgraph of K_N corresponding to W.

Theorem 4.2 *Suppose that the vertex sets $C_1, \ldots, C_k$ define a k-disjoint-clique subgraph K^* of the complete graph $K_N = (V, E)$ on N vertices and let $C_{k+1} := V\backslash(\cup_{i=1}^k C_i)$. Let $r_i := |C_i|$ for all $i = 1, \ldots, k+1$, and let $\hat{r} = \min_{i=1,\ldots,k} r_i$. Let $W \in \Sigma^N$ be a random symmetric matrix sampled from the planted cluster model*

according to distributions Ω_1, Ω_2 *with means* α, β *satisfying* (2.7). *Let* X^* *be the feasible solution for* (2.5) *corresponding to* $C_1, \ldots, C_k$ *defined by* (2.6). *Let* $\mu \in \mathbf{R}$, $\lambda \in \mathbf{R}^N$, *and* $\eta \in \mathbf{R}^{N \times N}$ *be chosen according to* (4.15), (4.16), *and* (4.17), *and let* S *be chosen according to* (4.1). *Suppose that the entries of* λ *and* η *are nonnegative. Then there exists scalar* $c > 0$ *such that if* $\|\tilde{S}\| \leq c\gamma\hat{r}$ *then* X^* *is optimal for* (2.5), *and* K^* *is the maximum density* k*-disjoint-clique subgraph of* K_N *corresponding to* W. *Moreover, if*

$$r_s \mathbf{e}^T W_{C_q, C_q} \mathbf{e} > r_q \mathbf{e}^T W_{C_q, C_s} \mathbf{e} \tag{4.19}$$

for all $q, s \in \{1, \ldots, k\}$ *such that* $q \neq s$, *then* X^* *is the unique optimal solution of* (2.5) *and* K^* *is the unique maximum density* k*-disjoint-clique subgraph of* K_N.

Proof By construction, μ, λ, η, and S satisfy (4.1), (4.2), (4.3), and (4.4). Moreover, λ and η are nonnegative by assumption. Therefore, it suffices to show that S is positive semidefinite to prove that X^* is optimal for (2.5). To do so, we fix $\mathbf{x} \in \mathbf{R}^N$ and decompose $\mathbf{x}$ as $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$ where

$$\mathbf{x}_1(C_i) = \begin{cases} \phi_i \mathbf{e}, & \text{if } i \in \{1, \ldots, k\} \\ 0, & \text{if } i = k+1 \end{cases}$$

for some $\phi \in \mathbf{R}^k$ chosen such that $\mathbf{x}_2(C_i)$ is orthogonal to $\mathbf{e}$ for all $i = 1, \ldots, k$, and $\mathbf{x}_2(C_{k+1}) = \mathbf{x}(C_{k+1})$. Here, and in the rest of the note, the notation $\mathbf{v}(A)$ denotes the vector in $\mathbf{R}^{|A|}$ with entries equal to those of $\mathbf{v}$ indexed by A. Similarly, the notation $M(A, B)$ denotes the $|A| \times |B|$ matrix with entries equal those of M indexed by A and B respectively. We have

$$\mathbf{x}^T S \mathbf{x} = \mathbf{x}_2^T S \mathbf{x}_2 = \mathbf{x}_2^T (\tilde{S} + \mu I) \mathbf{x}_2 \geq \left(\mu - \|\tilde{S}\|\right) \|\mathbf{x}_2\| \geq \left(\epsilon\gamma\hat{r} - \|\tilde{S}\|\right) \|\mathbf{x}_2\|$$

since $\mathbf{x}_2(C_i)$ is orthogonal to $\mathbf{e}$ for all $i = 1, \ldots, k$ and, hence, $\mathbf{x}_2^T (S - \tilde{S} - \mu I) \mathbf{x}_2 = 0$. Therefore, if $\|\tilde{S}\| \leq \epsilon\gamma\hat{r}$, then $\mathbf{x}^T S \mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbf{R}^N$ with equality if and only if $\mathbf{x}_2 = 0$. In this case, X^* is optimal for (2.5). Moreover, $\mathbf{v}_i$ is in the nullspace of S for all $i = 1, \ldots, k$ by (4.4) and the fact that $X^* = \sum_{i=1}^k \mathbf{v}_i \mathbf{v}_i^T / r_i$. Since $\mathbf{x}^T S \mathbf{x} = 0$ if and only if $\mathbf{x}_2 = 0$, the nullspace of S is exactly equal to the span of $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ and S has rank equal to $N - k$.

To see that X^* is the unique optimal solution for (2.5) if Assumption (4.19) holds, suppose, on the contrary, that $\tilde{X}$ is also optimal for (2.5). By (4.4), we have $\mathrm{Tr}\,(\tilde{X} S) = 0$, which holds if and only if $\tilde{X} S = 0$. Therefore, the row and column spaces of $\tilde{X}$ lie in the nullspace of S. Since $\tilde{X} \succeq 0$ and $\tilde{X} \geq 0$, we may write $\tilde{X}$ as

$$\tilde{X} = \sum_{i=1}^k \sum_{j=1}^k \sigma_{ij} \mathbf{v}_i \mathbf{v}_j^T \tag{4.20}$$

for some $\sigma \in \mathbf{R}_+^{k\times k}$. The fact that $\tilde{X}$ satisfies $\tilde{X}\mathbf{e} \leq \mathbf{e}$ implies that

$$\sigma_{qq} r_q + \sum_{\substack{s=1 \\ s\neq q}}^{k} \sigma_{qs} r_s \leq 1 \tag{4.21}$$

for all $q =, 1 \ldots, k$. Moreover, since $\mathrm{Tr}\,(W\tilde{X}) = \mathrm{Tr}\,(WX^*)$, there exists some $q \in \{1, \ldots, k\}$ such that

$$\sigma_{qq} \mathbf{v}_q^T W \mathbf{v}_q + \sum_{\substack{s=1 \\ s\neq q}}^{k} \sigma_{qs} \mathbf{v}_q^T W \mathbf{v}_s \geq \frac{\mathbf{v}_q^T W \mathbf{v}_q}{r_q}. \tag{4.22}$$

Combining (4.21) and (4.22) shows that

$$\begin{aligned} 0 &\leq \mathbf{v}_q^T W \mathbf{v}_q \left(\frac{1}{r_q} - \sum_{\substack{s=1 \\ s\neq q}}^{k} \frac{\sigma_{qs} r_s}{r_q} \right) + \sum_{\substack{s=1 \\ s\neq q}}^{k} \sigma_{qs} \mathbf{v}_q^T W \mathbf{v}_s - \frac{\mathbf{v}_q^T W \mathbf{v}_q}{r_q} \\ &= \sum_{\substack{s=1 \\ s\neq q}}^{k} \frac{\sigma_{qs}}{r_q} (r_q \mathbf{v}_q^T W \mathbf{v}_s - r_s \mathbf{v}_q^T W \mathbf{v}_q), \end{aligned}$$

contradicting Assumption (4.19). Therefore, X^* is the unique optimal solution of (2.5) as required. □

4.3 Nonnegativity of λ and η in the planted case

We now establish that the entries of λ and η are nonnegative with probability tending exponentially to 1 as $\hat{r}$ approaches ∞ for sufficiently small choice of ϵ in (4.15).

We begin by deriving lower bounds on the entries of η. To do so, we will repeatedly apply the following theorem of Hoeffding (see [25, Theorem 1]), which provides a bound on the tail distribution of a sum of bounded, independent random variables.

Theorem 4.3 (Hoeffding's Inequality) *Let $X_1, \ldots, X_m$ be independent identically distributed (i.i.d.) variables sampled from a distribution satisfying $0 \leq X_i \leq 1$ for all $i = 1, \ldots, m$. Let $S = X_1 + \cdots + X_m$. Then*

$$Pr(|S - \mathbf{E}[S]| > t) \leq 2\exp\left(\frac{-2t^2}{m}\right) \tag{4.23}$$

for all $t > 0$.

To show that $\eta_{ij} \geq 0$ for all $i, j \in V$ with high probability, we will use the following lemma, which provides an upper bound on $\|\mathbf{y}^{q,s}\|_\infty$ and $\|\mathbf{z}^{q,s}\|_\infty$ for all

$q, s \in \{1, \ldots, k+1\}$ such that $q \neq s$, holding with probability tending to 1 as $\hat{r}$ tends to ∞.

Lemma 4.2 *There exists scalar $\tilde{c} > 0$ such that $\|\mathbf{y}^{q,s}\|_\infty + \|\mathbf{z}^{q,s}\|_\infty \leq \tilde{c}\hat{r}^{-1/4}$ for all $q, s \in \{1, \ldots, k+1\}$ such that $q \neq s$ with probability tending exponentially to 1 as $\hat{r} \to \infty$.*

Proof Fix $q, s \in \{1, \ldots, k\}$ such that $q \neq s$. Without loss of generality, we assume that $r_q \leq r_s$. The proof for the case when either q or s is equal to $k+1$ is analogous. We first obtain an upper bound on $\|\mathbf{y}\|_\infty = \|\mathbf{y}^{q,s}\|_\infty$. By the triangle inequality, we have

$$\|\mathbf{y}\|_\infty \leq \frac{1}{r_s}\left\|\mathbf{b}_1 + \frac{|\mathbf{b}_1^T\mathbf{e}|}{r_q + r_s}\mathbf{e}\right\|_\infty \leq \frac{1}{r_s}\left(\|\mathbf{b}_1\|_\infty + \frac{|\mathbf{b}_1^T\mathbf{e}|}{r_q + r_s}\right). \tag{4.24}$$

Hence, to obtain an upper bound on $\|\mathbf{y}\|_\infty$, it suffices to obtain bounds on $\|\mathbf{b}_1\|_\infty$ and $|\mathbf{b}_1^T\mathbf{e}|$. We begin with $\|\mathbf{b}_1\|_\infty$. Recall that we have

$$\mathbf{b}_i = r_s\left(\lambda_i - \frac{1}{2r_q}(\alpha r_q - \mu)\right) + \left(\lambda_{C_s}^T\mathbf{e} - \frac{1}{2}(\alpha r_s - \mu)\right) - \left(\sum_{j\in C_s} W_{ij} - \beta r_s\right). \tag{4.25}$$

for each $i \in C_q$. Note that

$$\lambda_{C_s}^T\mathbf{e} = \frac{1}{r_s}\left(\mathbf{e}^T W_{C_s,C_s}\mathbf{e} - \frac{1}{2}r_s\mu - \frac{1}{2}\mathbf{e}^T W_{C_s,C_s}\mathbf{e}\right) = \frac{1}{2r_s}(\mathbf{e}^T W_{C_s,C_s}\mathbf{e} - r_s\mu).$$

Applying (4.23) with $S = \mathrm{Tr}\,(W_{C_s,C_s})$, $t = r_s^{3/2}$ and $S = \sum_{i\in C_s}\sum_{j\in C_s, j>i} W_{ij}$, $t = r_s^{3/2}/2$ shows that

$$\begin{aligned}\left|\lambda_{C_s}^T\mathbf{e} - \frac{1}{2}(\alpha r_s - \mu)\right| &= \frac{1}{2r_s}|\mathbf{e}^T W_{C_s,C_s}\mathbf{e} - \alpha r_s^2| \\ &\leq \frac{1}{2r_s}\left(|\mathrm{Tr}\,(W_{C_s,C_s}) - \alpha r_s| + 2\left|\sum_{i\in C_s}\sum_{\substack{j\in C_s\\ j>i}} W_{ij} - \frac{\alpha r_s(r_s-1)}{2}\right|\right) \leq \sqrt{r_s}\end{aligned} \tag{4.26}$$

with probability at least $1 - 2\exp(-2r_s^2) - 2\exp(-r_s^2/(r_s-1)) \geq 1 - \tilde{p}_1$, where $\tilde{p}_1 := 2\exp(-2\hat{r}^2) + 2\exp(-\hat{r})$. Next, applying (4.23) with $S = \sum_{\ell\in C_s} W_{i\ell}$ and $t = r_s^{3/4}$ shows that

$$\left|\sum_{\ell\in C_s} W_{i\ell} - \beta r_s\right| \leq r_s^{3/4} \tag{4.27}$$

with probability at least $1 - \tilde{p}_2$ where $\tilde{p}_2 := 2\exp(-2\hat{r}^{1/2})$. Finally, applying (4.23) with $S = \sum_{\ell \in C_q} W_{i\ell}$, $t = r_q^{3/4}$ and (4.26) shows that

$$\begin{aligned}\left|\lambda_i - \frac{1}{2r_q}(\alpha r_q - \mu)\right| &\leq \frac{1}{r_q}\left|\sum_{\ell \in C_q} W_{i\ell} - r_q\alpha\right| + \frac{1}{2r_q^2}\left|\mathbf{e}^T W_{C_q,C_q}\mathbf{e} - \alpha r_q^2\right| \\ &\leq r_q^{-1/4} + r_q^{-1/2} \leq 2r_q^{-1/4} \end{aligned} \tag{4.28}$$

with probability at least $1 - \tilde{p}_1 - \tilde{p}_2$. Combining (4.26), (4.27) and (4.28) and applying the union bound shows that

$$\|\mathbf{b}_1\|_\infty \leq 4r_q^{-1/4} r_s \tag{4.29}$$

with probability at least $1 - \tilde{p}_1 - 2r_q\tilde{p}_2$. By a similar argument, $\|\mathbf{b}_2\|_\infty \leq 4r_q^{3/4}$ with probability at least $1 - \tilde{p}_1 - 2r_s\tilde{p}_2$.

We next obtain an upper bound on $|\mathbf{b}_1^T\mathbf{e}|$ and $|\mathbf{b}_2^T\mathbf{e}|$. We have

$$\begin{aligned}\mathbf{b}_1^T\mathbf{e} = r_s\left(\lambda_{C_q}^T\mathbf{e} - \frac{1}{2}(\alpha r_q - \mu)\right) + r_q\left(\lambda_{C_s}^T\mathbf{e} - \frac{1}{2}(\alpha r_s - \mu)\right) \\ +(\beta r_s r_q - \mathbf{e}^T W_{C_q,C_s}\mathbf{e}).\end{aligned} \tag{4.30}$$

By (4.26) and the union bound, we have

$$\left|\lambda_{C_s}^T\mathbf{e} - \frac{1}{2}(\alpha r_s - \mu)\right| \leq \sqrt{r_s} \tag{4.31}$$

$$\left|\lambda_{C_q}^T\mathbf{e} - \frac{1}{2}(\alpha r_q - \mu)\right| \leq \sqrt{r_q} \tag{4.32}$$

with probability at least $1 - 2\tilde{p}_1$. Moreover, applying (4.23) with $S = \mathbf{e}^T W_{C_q,C_s}\mathbf{e}$ and $t = r_q\sqrt{r_s}$ shows that

$$|\mathbf{e}^T W_{C_q,C_s}\mathbf{e} - \beta r_s r_q| \leq r_q\sqrt{r_s} \tag{4.33}$$

with probability at least $1 - \tilde{p}_3$, where $\tilde{p}_3 := 2\exp(-2\hat{r})$. Substituting (4.31) and (4.33) into (4.30), we have

$$|\mathbf{b}_1^T\mathbf{e}| \leq 3r_s\sqrt{r_q} \tag{4.34}$$

for some scalar $c_3 > 0$ with probability at least $1 - 2\tilde{p}_1 - \tilde{p}_3$ by the union bound. Similarly,

$$|\mathbf{b}_2^T\mathbf{e}| \leq 3r_s\sqrt{r_q} \tag{4.35}$$

with probability at least $1 - 2\tilde{p}_1 - \tilde{p}_3$. Substituting (4.29) and (4.34) in (4.24) yields

$$\|\mathbf{y}\|_\infty \leq \tilde{c}_1 r_q^{-1/4}, \tag{4.36}$$

for some scalar $\tilde{c}_1 > 0$, with probability at least $1 - (3\tilde{p}_1 + 2r_q\tilde{p}_2 + \tilde{p}_3)$. Similarly, there exists scalar $\tilde{c}_2 > 0$ such that

$$\|\mathbf{z}\|_\infty \leq \tilde{c}_2 r_q^{-1/4} \tag{4.37}$$

with probability at least $1 - (3\tilde{p}_1 + 2r_s\tilde{p}_2 + \tilde{p}_3)$. Combining (4.36) and (4.37) and applying the union bound over all q, s completes the proof. □

As an immediate consequence of Lemma 4.2, η is nonnegative with probability tending exponentially to 1 for sufficiently large values of $\hat{r}$.

Corollary 4.1 *Suppose that α and β satisfy (2.7). Then the entries of the matrix η are nonnegative with probability tending exponentially to 1 as $\hat{r}$ approaches ∞.*

Proof Fix $i \in C_q$, $j \in C_s$ for some $q, s \in \{1, \ldots, k+1\}$ such that $q \neq s$. Recall that

$$\eta_{C_q,C_s} = \left(\frac{\bar{\delta}_{q,k+1}}{2}\left(\alpha - \frac{\mu}{r_q}\right) + \frac{\bar{\delta}_{s,k+1}}{2}\left(\alpha - \frac{\mu}{r_s}\right) - \beta\right)\mathbf{e}\mathbf{e}^T + \mathbf{y}^{q,s}\mathbf{e}^T + \mathbf{e}(\mathbf{z}^{q,s})^T.$$

Therefore, since $\gamma > 0$ by (2.7), Lemma 4.2 implies that

$$\begin{aligned}\eta_{ij} &\geq \frac{(1-\delta_{q,k+1})}{2}\left(\alpha - \frac{\mu}{r_q}\right) + \frac{(1-\delta_{s,k+1})}{2}\left(\alpha - \frac{\mu}{r_s}\right) - \beta - \|\mathbf{y}^{q,s}\|_\infty - \|\mathbf{z}^{q,s}\|_\infty \\ &\geq \left(\frac{1}{2} - \epsilon\right)\gamma - \tilde{c}\hat{r}^{-1/4} \geq 0,\end{aligned}$$

for all sufficiently small $\epsilon > 0$ and sufficiently large $\hat{r}$ with probability tending exponentially to 1 as $\hat{r} \to \infty$, since at most one of q and s is equal to $k+1$. □

The following lemma provides a similar lower bound on the entries of λ.

Lemma 4.3 *There exist scalars $\bar{c}_1, \bar{c}_2 > 0$ such that $\lambda_i \geq \hat{r}(\bar{c}_1 - \bar{c}_2\hat{r}^{-1/4})$ for all $i \in V \backslash C_{k+1}$ with probability tending exponentially to 1 as $\hat{r} \to \infty$.*

Proof Fix $q \in \{1, \ldots, k\}$ and $i \in C_q$. Recall that

$$\lambda_i = \sum_{j \in C_q} W_{ij} - \frac{1}{2r_q}\mathbf{e}^T W_{C_q,C_q}\mathbf{e} - \frac{\mu}{2}.$$

Applying (4.23) with $S = \sum_{j \in C_q} W_{ij}$ and $t = r_q^{3/4}$ yields

$$\sum_{j \in C_q} W_{ij} \geq \alpha r_q - r_q^{3/4} \tag{4.38}$$

with probability at least $1 - \tilde{p}_2$. Moreover, (4.32) implies that

$$\frac{1}{2r_q}\mathbf{e}^T W_{C_q,C_q}\mathbf{e} \leq \frac{1}{2}(\alpha r_q + \sqrt{r_q}) \tag{4.39}$$

with probability at least $1 - \tilde{p}_1$. Combining (4.38) and (4.39) and applying the union bound shows that there exist scalars $\bar{c}_1, \bar{c}_2 > 0$ such that

$$\lambda_i \geq \alpha r_q - r_q^{3/4} - \frac{1}{2}(\alpha r_q + \sqrt{r_q}) - \frac{\mu}{2} \geq r_q(\bar{c}_1 - \bar{c}_2 r_q^{-1/4})$$

with probability at least $1 - \tilde{p}_1 - \tilde{p}_2$ for sufficiently small choice of $\epsilon > 0$ in (4.15). Applying the union bound over all $i \in V \backslash C_{k+1}$ completes the proof. □

Note that Lemma 4.3 implies that $\lambda \geq 0$ with probability tending exponentially to 1 as $\hat{r}$ tends to ∞. Therefore, μ, λ, η constructed according to (4.15), (4.16), and (4.17) are dual feasible for (2.5) with probability tending exponentially to 1 as $\hat{r} \to \infty$ if the left-hand side of (4.1) is positive semidefinite. The following lemma states the uniqueness condition given by (4.19) is also satisfied with high probability for sufficiently large $\hat{r}$.

Lemma 4.4 *If $\hat{r} > 9/(\alpha - \beta)^2$ then $r_s \mathbf{e}^T W_{C_q,C_q}\mathbf{e} > r_q \mathbf{e}^T W_{C_q,C_s}\mathbf{e}$ for all $q, s \in \{1, \ldots, k\}$ such that $q \neq s$ with probability tending exponentially to 1 as $\hat{r} \to \infty$.*

Proof Fix $q \neq s$ such that $r_q \leq r_s$. Combining (4.26) and (4.33) shows that

$$r_s \mathbf{e}^T W_{C_q,C_q}\mathbf{e} - r_q \mathbf{e}^T W_{C_q,C_s}\mathbf{e} \geq r_s r_q^2(\alpha - \beta - 2r_q^{-1/2} - r_s^{-1/2}) \geq r_s r_q^2(\alpha - \beta - 3\hat{r}^{-1/2})$$

with probability at least $1 - \tilde{p}_1 - \tilde{p}_3$. Noting that this lower bound is positive if $\hat{r} > 9/(\alpha - \beta)^2$ and applying the union bound over all choices of q and s completes the proof. □

We have shown that μ, λ, η constructed according to (4.15), (4.16), and (4.17) are dual feasible for (2.5) and the uniqueness condition (4.19) is satisfied with probability tending exponentially to 1 as $\hat{r} \to \infty$. In the next subsection, we derive an upper bound on the norm of $\tilde{S}$ and use this bound to obtain conditions ensuring dual feasibility of S and, hence, optimality of X^* for (2.5).

4.4 An upper bound on $\|\tilde{S}\|$

In this section, we derive an upper bound on $\|\tilde{S}\|$, which will be used to verify that the conditions on the partition $C_1, \ldots, C_{k+1}$ imposed by (2.8) ensure that the k-disjoint-clique subgraph of K_N composed of the cliques $C_1, \ldots, C_k$ is the unique maximum density k-disjoint-clique of K_N with respect to W and can be recovered by solving (2.5) with probability tending exponentially to 1 as $\hat{r} \to \infty$. In particular, we will prove the following lemma.

Lemma 4.5 *There exist scalars $\rho_1, \rho_2 > 0$ such that*

$$\|\tilde{S}\| \leq \rho_1 \sqrt{N} + \rho_2 \sqrt{k r_{k+1}} + \beta r_{k+1} \tag{4.40}$$

with probability tending exponentially to 1 *as $\hat{r}$ approaches ∞.*

This lemma, along with Theorem 4.2, Lemma 4.3, and Corollary 4.1, establishes Theorem 2.1. Indeed, if the right-hand side of (4.40) is bounded above by $O(\gamma \hat{r})$ then Theorem 4.2, Lemma 4.3, and Corollary 4.1 imply that the k-disjoint-clique subgraph given by $C_1, \ldots, C_k$ is the densest k-disjoint-clique subgraph corresponding to W and can be recovered by solving (2.5).

The remainder of this section consists of a proof of Lemma 4.5. We decompose $\tilde{S}$ as $\tilde{S} = \tilde{S}_1 + \tilde{S}_2 + \tilde{S}_3$ where $\tilde{S}_i \in \Sigma^N$, $i = 1, \ldots, 3$, are $(k+1)$ by $(k+1)$ block matrices such that

$$\tilde{S}_1(C_q, C_s) = \mathbf{E}[W] - W$$

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

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

To bound the norm of each matrix in this decomposition, we will make repeated use of the following bound on the norm of a random symmetric matrix (see [21], [4, Theorem 1]).

Theorem 4.4 *Let $A \in \Sigma^n$ be a random symmetric matrix with i.i.d. entries sampled from a distribution with mean μ and variance σ^2 such that $A_{ij} \in [0, 1]$ for all $i, j \in \{1, \ldots, n\}$. Then $\|A - \mu \mathbf{e}\mathbf{e}^T\| \leq 3\sigma\sqrt{n}$ with probability at least $1 - \exp(-cn^{1/6})$ where c depends only on σ.*

We are now ready to compute the desired bound on $\|\tilde{S}\|$. By Theorem 4.4, there exist $\rho_1 > 0$ such that $\|\tilde{S}_1\| \leq \rho_1 \sqrt{N}$ with probability tending exponentially to 1 as $N \to \infty$. Morever, we have $\|\tilde{S}_3\| = \beta \|\mathbf{e}\mathbf{e}^T\| = \beta r_{k+1}$. It remains to obtain an upper bound on $\|\tilde{S}_2\|$.

Note that $\|\tilde{S}_2\| \leq 2\|\lambda - \mathbf{E}[\lambda]\|\sqrt{r_{k+1}}$ by the triangle inequality. Recall that

$$\lambda_{C_q} - \mathbf{E}[\lambda_{C_q}] = \frac{1}{r_q}\left(\left(W_{C_q, C_q}\mathbf{e} - \alpha r_q \mathbf{e}\right) - \frac{1}{r_q}\left(\mathbf{e}^T W_{C_q, C_q}\mathbf{e} - \alpha r_q^2\right)\mathbf{e}\right)$$

for all $q = 1, \ldots, k$. Applying Theorem 4.4, there exists $\varphi > 0$ such that

$$\|W_{C_q, C_q}\mathbf{e} - \alpha r_q \mathbf{e}\| \leq \|W_{C_q, C_q} - \alpha \mathbf{e}\mathbf{e}^T\|\|\mathbf{e}\| \leq \varphi r_q$$

with probability tending exponentially to 1 as $\hat{r} \to \infty$. On the other hand, (4.26) implies that $|\mathbf{e}^T W_{C_q, C_q}\mathbf{e} - \alpha r_q^2| \leq 2r_q^{3/2}$ with probability at least $1 - \tilde{p}_1$. It follows

that there exists scalar $\rho_2 > 0$ such that $\|\lambda_{C_q} - \mathbf{E}[\lambda_{C_q}]\|^2 \leq \rho_2^2/4$ for all $q = 1, \ldots, k$ with probability tending exponentially to 1 as $\hat{r} \to \infty$. Therefore, $\|\lambda - \mathbf{E}[\lambda]\|^2 = \sum_{q=1}^{k} \|\lambda_{C_q} - \mathbf{E}[\lambda_{C_q}]\|^2 \leq k\rho_2^2/4$ with high probability, as required. This completes the proof of Lemma 4.5.

5 Proof of Theorem 3.1

5.1 Optimality conditions and choice of multipliers

We provide of a sketch of the proof of Theorem 3.1 here; many of the technical details are identical to those in the proof of Theorem 2.1 and are omitted. As before, we will establish that a proposed solution satisfies a set of sufficient conditions for optimality for (3.5), given by the following theorem, with high probability if the input graph satisfies the assumptions of Theorem 3.1.

Theorem 5.1 *Let Z be feasible for (3.5) and suppose that there exist some $\mu_1, \mu_2 \in \mathbf{R}$, $\lambda \in \mathbf{R}_+^M$, $\phi \in \mathbf{R}_+^N$, $\eta \in \mathbf{R}_+^{(M+N)\times(M+N)}$ and $S \in \Sigma_+^{M+N}$ such that*

$$\begin{pmatrix} \mu_1 I + \lambda \mathbf{e}^T + \mathbf{e}\lambda^T & -W \\ -W^T & \mu_2 I + \phi \mathbf{e}^T + \mathbf{e}\phi^T \end{pmatrix} - \eta = S \tag{5.1}$$

$$\lambda^T (Z_{U,U}\mathbf{e} - \mathbf{e}) = 0 \tag{5.2}$$

$$\phi^T (Z_{V,V} - \mathbf{e}) = 0 \tag{5.3}$$

$$\mathrm{Tr}\,(Z\eta) = 0 \tag{5.4}$$

$$\mathrm{Tr}\,(ZS) = 0. \tag{5.5}$$

Then Z is optimal for (3.5).

Let $(U_1, V_1), \ldots, (U_k, V_k)$ denote the vertex sets of the k-disjoint-biclique subgraph G^* of the bipartite complete graph $K_{M,N} = ((U, V), E)$ with vertex sets U and V of size M and N respectively. Let $U_{k+1} := U \backslash (\cup_{i=1}^{k} U_i)$ and $V_{k+1} := V \backslash (\cup_{i=1}^{k} V_i)$. Let $W \in \mathbf{R}^{M\times N}$ be a random nonnegative matrix sampled from the planted bicluster model according to distributions Ω_1, Ω_2 with means α, β. Let $m_i := |U_i|, n_i := |V_i|$ for all $i = 1, \ldots, k+1$, and let $\hat{m} = \min_{i=1,\ldots,k} m_i$, $\hat{n} = \min_{i=1,\ldots,k} n_i$. Let $C_i := U_i \cup V_i$ and let $r_i := |C_i| = m_i + n_i$ for all $i = 1, \ldots, k+1$. We assume that m_i is equal to a scalar multiple τ_i^2 of n_i for all $i \in \{1, \ldots, k+1\}$. That is, $m_i = \tau_i^2 n_i$ for some $\tau_i > 0$ for all $i = 1, \ldots, k+1$.

As before, we establish optimality of Z^* by constructing dual multipliers satisfying the assumptions of Theorem 5.1. The matrix S and, hence, λ, ϕ, and η will be constructed in blocks indexed by the vertex sets $U_1, \ldots, U_{k+1}$ and $V_1, \ldots, V_{k+1}$. Note that the diagonal blocks of $Z^*_{U,U}$ indexed by $U_1, \ldots, U_k$ consist of multiples of the all-ones matrix and the remaining blocks are equal to 0. Therefore, $\lambda_{U_{k+1}} = 0$ by (5.2). Similarly, the block structure of Z^* implies that $\phi_{V_{k+1}} = 0$ by (5.3) and $\eta_{C_q,C_q} = 0$ for all $q = 1, \ldots, k$ by (5.4).

Since both S and Z^* are assumed to be positive semidefinite matrices, the complementary slackness condition, $\mathrm{Tr}\,(Z^*S) = 0$, is equivalent to requiring the columns of

Z^* to reside in the nullspace of S. For each $q \in \{1, \dots, k\}$, we choose λ_{U_q} so that S_{U_q,C_q} is orthogonal to $Z^*_{U_q,C_q}$. In particular, it suffices to choose λ such that

$$0 = S_{U_q,U_q}\mathbf{e} + \tau_q S_{U_q,V_q}\mathbf{e} = \mu_1\mathbf{e} + m_q\lambda_{U_q} + (\lambda_{U_q}^T\mathbf{e})\mathbf{e} - \tau_q W_{U_q,V_q}\mathbf{e} \tag{5.6}$$

for all $q = 1, \dots, k$. Rearranging (5.6) shows that λ_{U_q} is the solution to the system

$$(m_q I + \mathbf{e}\mathbf{e}^T)\lambda_{U_q} = \tau_q W_{U_q,V_q}\mathbf{e} - \mu_1\mathbf{e} \tag{5.7}$$

for all $q = 1, \dots, k$. As before, the Sherman-Morrision-Woodbury formula yields an explicit formula for λ; for each $q \in \{1, \dots, k\}$, applying (4.7) with $A = m_q I$, $\mathbf{u} = \mathbf{v} = \mathbf{e}$ shows that

$$\lambda_{U_q} = \frac{1}{m_q}\left(\tau_q W_{U_q,V_q}\mathbf{e} - \frac{1}{2}\left(\mu_1 + \frac{\mathbf{e}^T W_{U_q,V_q}\mathbf{e}}{\tau_q n_q}\right)\mathbf{e}\right). \tag{5.8}$$

Similarly, choosing

$$\phi_{V_q} = \frac{1}{n_q}\left(\frac{W^T_{U_q,V_q}\mathbf{e}}{\tau_q} - \frac{1}{2}\left(\mu_2 + \frac{\mathbf{e}^T W_{U_q,V_q}\mathbf{e}}{\tau_q n_q}\right)\mathbf{e}\right) \tag{5.9}$$

forces the rows of S_{V_q,C_q} to be orthogonal to the columns of $Z^*_{C_q,C_q}$ for all $q \in \{1, \dots, k\}$. Note that $\mathbf{E}[\lambda_{U_q}] = (\alpha/(2\tau_q) - \mu_1/(2m_q))\mathbf{e}$ for all $q \in \{1, \dots, k\}$. We choose $\mu_1 = \epsilon\gamma\hat{m}$ for some scalar $\epsilon > 0$ to be defined later to ensure that λ is nonnegative in expectation. Similarly, $\mathbf{E}[\phi_{V_q}] = \left(\alpha\tau_q/2 - \mu_2/(2n_q)\right)\mathbf{e}$ for all $q = 1, \dots, k$. Again, we choose $\mu_2 = \epsilon\gamma\hat{n}$ for small enough $\epsilon > 0$ to ensure that ϕ is nonnegative in expectation.

We next construct the multiplier η. We set $\eta_{C_{k+1},C_{k+1}} = 0$ and parametrize η_{C_q,C_s} using the vectors $\mathbf{y}^{q,s}$ and $\mathbf{z}^{q,s}$ for each $q \neq s$. For each $q = 1, \dots, k+1$, let $\mathbf{w}_q$ be the vector in $\mathbf{R}^{C_q}$ such that $\mathbf{w}_q(U_q) = \mathbf{e}$ and $\mathbf{w}_q(V_q) = \tau_q\mathbf{e}$. We choose $\eta_{C_q,C_s} = \Pi^{q,s} + \mathbf{y}^{q,s}\mathbf{w}_s^T + \mathbf{w}_q(\mathbf{z}^{q,s})^T$, where

$$\Pi^{q,s} = \begin{pmatrix} \pi_{U_q,U_s}\mathbf{e}\mathbf{e}^T & \tau_s\pi_{U_q,V_s}\mathbf{e}\mathbf{e}^T \\ \tau_q\pi_{V_q,U_s}\mathbf{e}\mathbf{e}^T & \tau_q\tau_s\pi_{V_q,V_s}\mathbf{e}\mathbf{e}^T \end{pmatrix}$$

for some scalars $\pi_{U_q,U_s}, \pi_{U_q,V_s}, \pi_{V_q,U_s}, \pi_{V_q,V_s} > 0$ to be defined later. As before, we choose $\mathbf{y}^{q,s}$ and $\mathbf{z}^{q,s}$ to be solutions of the system of equations given by $S_{C_q,C_s}Z^*_{C_s,C_s} = 0$ and $S_{C_s,C_q}Z^*_{C_q,C_q} = 0$. By the symmetry of S and Z^*, $\mathbf{y}^{q,s} = \mathbf{z}^{s,q}$ for all $q \neq s$.

For all $q, s \in \{1, \dots, k+1\}$ such that $q \neq s$, let

$$\bar{S}_{C_q,C_s} := \begin{pmatrix} \lambda_{U_q}\mathbf{e}^T + \mathbf{e}\lambda_{U_s}^T & -W_{U_q,V_s} \\ -W^T_{U_s,V_q} & \phi_{V_q}\mathbf{e}^T + \mathbf{e}\phi_{V_s}^T \end{pmatrix}, \tag{5.10}$$

and let $\mathbf{b} = \mathbf{b}^{q,s} \in \mathbf{R}^{C_q \cup C_s}$ be the vector defined by $\mathbf{b}_{C_q} = \big(\bar{S}_{C_q,C_s} - \mathbf{E}[\bar{S}_{C_q,C_s}]\big)\mathbf{w}_s$ and $\mathbf{b}_{C_s} = \big(\bar{S}_{C_s,C_q} - \mathbf{E}[\bar{S}_{C_s,C_q}]\big)\mathbf{w}_q$. The parameters $\pi_{U_q,U_s}, \pi_{U_q,V_s}, \pi_{V_q,U_s}, \pi_{V_q,V_s} > 0$ will be chosen so that

$$\big(\mathbf{E}[\bar{S}_{C_q,C_s}] - \Pi^{q,s}\big)\mathbf{w}_s = 0 \quad \text{and} \quad \big(\mathbf{E}[\bar{S}_{C_s,C_q}] - \Pi^{s,q}\big)\mathbf{w}_q = 0. \tag{5.11}$$

We will establish that such a choice of $\Pi^{q,s}$ exists in Lemma 5.1.

Fix $q, s \in \{1, \ldots, k\}$ such that $q \neq s$. It is easy to see that the requirement that the rows of S_{C_q,C_s} be orthogonal to the columns of $Z^*_{C_s,C_s}$ is satisfied if $\mathbf{y} = \mathbf{y}^{q,s}$ and $\mathbf{z} = \mathbf{z}^{q,s}$ are chosen to be be the unique solutions of the system

$$\begin{pmatrix} 2m_s I + \mathbf{w}_q\mathbf{w}_q^T & 0 \\ 0 & 2m_q I + \mathbf{w}_s\mathbf{w}_s^T \end{pmatrix}\begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \mathbf{b}. \tag{5.12}$$

Applying (4.7) with $A = 2m_s$, $\mathbf{u} = \mathbf{v} = \mathbf{w}_q$ and $A = 2m_q$, $\mathbf{u} = \mathbf{v} = \mathbf{w}_s$ yields

$$\mathbf{y} = \frac{1}{2m_s}\left(I - \frac{\mathbf{w}_q\mathbf{w}_q^T}{2(m_q + m_s)}\right)\mathbf{b}_{C_q} \quad \text{and} \quad \mathbf{z} = \frac{1}{2m_q}\left(I - \frac{\mathbf{w}_s\mathbf{w}_s^T}{2(m_q + m_s)}\right)\mathbf{b}_{C_s}$$

respectively.

For $q \in \{1, \ldots, k\}$, we set $\mathbf{z}^{k+1,q} = 0$ and choose $\mathbf{y} = \mathbf{y}^{k+1,q}$ so that the rows of S_{C_{k+1},C_q} are orthogonal to $\mathbf{w}_q$. By our choice of $\Pi^{k+1,q}$, $\mathbf{y}$ must satisfy

$$2m_q\mathbf{y} = \begin{pmatrix} \mathbf{e}(\lambda_{U_q} - \mathbf{E}[\lambda_{U_q}])^T & -W_{U_{k+1},V_q} + \beta\mathbf{e}\mathbf{e}^T \\ \beta\mathbf{e}\mathbf{e}^T - W^T_{U_q,V_{k+1}} & \mathbf{e}(\phi_{V_q} - \mathbf{E}[\phi_{V_q}]) \end{pmatrix}\mathbf{w}_q = \mathbf{b}^{k+1,q}$$

Therefore, we choose $\mathbf{y}^{k+1,q} = (1/(2mq))\mathbf{b}^{k+1,q}$. We choose the remaining blocks of η symmetrically. That is, we choose $\mathbf{y}^{q,k+1} = 0$ and set $\mathbf{z}^{q,k+1} = \mathbf{y}^{k+1,q}$ for all $q = 1, \ldots, k$.

To establish that S is positive semidefinite with high probability, we decompose S as the sum $S = S_1 + S_2 + S_3 + S_4$ where

$$S_1(C_q, C_s) := \begin{cases} S_{C_{k+1},C_{k+1}}, & \text{if } q = s = k+1 \\ \bar{S}_{C_q,C_s} - \mathbf{E}[\bar{S}_{C_q,C_s}], & \text{otherwise,} \end{cases} \tag{5.13}$$

$$S_2(C_q, C_s) := \begin{cases} \mathbf{E}[\bar{S}_{C_q,C_s}] - \Pi^{q,s}, & \text{if } q \neq s \\ \mathbf{E}[\bar{S}_{C_q,C_q}], & \text{if } q = s, q \in \{1, \ldots, k\}, \\ 0, & \text{otherwise,} \end{cases} \tag{5.14}$$

$$S_3(C_q, C_s) := \big\{\mathbf{y}^{q,s}\mathbf{w}_s^T + \mathbf{w}_q(\mathbf{z}^{q,s})^T, \quad \text{for all } q, s \in \{1, \ldots, k+1\} \tag{5.15}$$

and

$$S_4 := \begin{pmatrix} \mu_1 I & 0 \\ 0 & \mu_2 I \end{pmatrix}. \tag{5.16}$$

We conclude with the following theorem, which provides a sufficient condition for optimality and uniqueness of the proposed solution Z^* for (3.5).

Theorem 5.2 *Let Z^* be the feasible solution for (3.5) corresponding to G^* defined by (3.6). Then there exist scalars $\xi_1, \xi_2 > 0$ such that if*

$$\|S_1\| + \xi_1 (n_{k+1} N)^{1/2} \leq \xi_2 \gamma \hat{n} \tag{5.17}$$

then Z^* is the unique optimal solution of (3.5) and G^* is the unique densest k-disjoint-biclique subgraph of $K_{M,N}$ with probability tending exponentially to 1 as $\hat{n} \to \infty$.

The remainder of this section consists of a proof of Theorem 5.2. We first establish that Z^* is optimal for (3.5) and G^* is the unique densest k-disjoint-biclique subgraph of $K_{M,N}$ with probability tending exponentially to 1 as $\hat{n} \to \infty$ if (5.17) is satisfied. By construction, μ, λ, ϕ, η and S satisfy (5.1), (5.2), (5.3), (5.4), and (5.5). Moreover, a series of arguments similar to those in Sect. 4.3 establish that λ, ϕ, and η are non-negative with probability tending exponentially to 1 as $\hat{n} \to \infty$. Therefore, it suffices to show that S is positive semidefinite with probability tending exponentially to 1 as $\hat{n} \to \infty$ if (5.17) is satisfied. To do so, we will establish that $\mathbf{x}^T S \mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbf{R}^{M+N}$ in this case.

Fix $\mathbf{x} \in \mathbf{R}^{M+N}$. We decompose $\mathbf{x}$ as $\mathbf{x} = \sum_{i=1}^{k} \varphi_i \mathbf{x}_i + \bar{\mathbf{x}}$ for some $\varphi_1, \ldots, \varphi_k$, where $\mathbf{x}_i(C_i) = \mathbf{w}_i$ and all remaining entries of $\mathbf{x}_i$ are equal to 0, and $\bar{\mathbf{x}}$ is orthogonal to the span of $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. Note that span $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \subseteq$ Null S since $\mathbf{x}_i$ is a scalar multiple of a column of Z^* for all $i = 1, \ldots, k$. It follows that

$$\mathbf{x}^T S \mathbf{x} = \bar{\mathbf{x}}^T S \bar{\mathbf{x}} = \sum_{i=1}^{4} \bar{\mathbf{x}}^T S_i \bar{\mathbf{x}}. \tag{5.18}$$

Note that $\bar{\mathbf{x}}^T S_3 \bar{\mathbf{x}} = 0$ since $\bar{\mathbf{x}}(C_q)$ is orthogonal to $\mathbf{w}_q$ for all $q = 1, \ldots, k$ and $\bar{\mathbf{x}}^T S_4 \bar{\mathbf{x}} \geq \min\{\mu_1, \mu_2\} \|\bar{\mathbf{x}}\|^2 = \epsilon\gamma \min\{\hat{m}, \hat{n}\} \|\bar{\mathbf{x}}\|^2$. by our choice of μ_1 and μ_2. The following lemma provides a similar lower bound on $\bar{\mathbf{x}}^T S_2 \bar{\mathbf{x}}$.

Lemma 5.1 *Suppose that $\alpha, \beta, \tau_1, \ldots, \tau_{k+1}$ satisfy (3.7) and (3.8). Then, for all $q, s \in \{1, \ldots, k+1\}$ such that $q \neq s$, there exist scalars $\pi_{U_q,U_s}, \pi_{U_q,V_s}, \pi_{V_q,U_s}, \pi_{V_q,V_s} > 0$ and $\hat{c} > 0$, depending only on $\alpha, \beta, \tau_1, \ldots, \tau_{k+1}$ such that $\bar{\mathbf{x}}^T S_2 \bar{\mathbf{x}} \geq -\hat{c} \|\bar{\mathbf{x}}\|^2 \sqrt{n_{k+1} N}$ and (5.11) is satisfied.*

Proof Fix $q, s \in \{1, \ldots, k\}$ such that $q \neq s$. Let $\pi_1 := \pi_{U_q,U_s}$, $\pi_2 := \pi_{U_q,V_s}$, $\pi_3 := \pi_{V_q,U_s}$, and $\pi_4 := \pi_{V_q,V_s}$. Then the system of equations defined by (5.11) is equivalent to

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix} = \begin{pmatrix} \bar{\lambda} - \beta/\tau_s \\ \bar{\phi} - \tau_s \beta \\ \bar{\lambda} - \beta/\tau_q \\ \bar{\phi} - \tau_q \beta \end{pmatrix}, \tag{5.19}$$

where

$$\bar{\lambda} := \frac{\alpha}{2}\left(\frac{1}{\tau_q}+\frac{1}{\tau_s}\right)-\frac{\mu_1}{2}\left(\frac{1}{m_q}+\frac{1}{m_s}\right),$$
$$\bar{\phi} := \frac{\alpha}{2}(\tau_q+\tau_s)-\frac{\mu_2}{2}\left(\frac{1}{n_q}+\frac{1}{n_s}\right). \tag{5.20}$$

The system (5.19) is singular with solutions of the form

$$\pi_1 = \bar{\lambda}-\frac{\bar{\phi}}{\tau_q\tau_s}+\pi_4, \quad \pi_2 = \frac{\bar{\phi}}{\tau_q\tau_s}-\frac{\beta}{\tau_s}-\pi_4,$$
$$\pi_3 = \frac{\bar{\phi}}{\tau_q\tau_s}-\frac{\beta}{\tau_q}-\pi_4. \tag{5.21}$$

We next show that there exists some choice of $\pi_4 > 0$, independent of $\hat{n}$, such that the desired bound on $\bar{\mathbf{x}}^T S_2\bar{\mathbf{x}}$ holds and π_1, π_2, π_3 are bounded below by a positive scalar whenever (3.7) and (3.8) are satisfied.

Suppose that $\alpha, \beta, \tau_1, \ldots, \tau_{k+1}$ satisfy (3.7) and (3.8). Let $\pi_4 := (\rho_1\bar{\phi}-\rho_2\beta)/\tau_q\tau_s$ for some $\rho_1, \rho_2 > 0$ to be chosen later. For π_4 to be strictly positive, we need $\rho_2\beta < \rho_1\bar{\phi}$. Substituting our choice of π_4 into the formulas for π_2 and π_3 given by (5.21) and rearranging shows that ρ_1 and ρ_2 must satisfy

$$\rho_2\beta > \beta\max\{\tau_q, \tau_s\}+(\rho_1-1)\bar{\phi} \tag{5.22}$$

for π_2, π_3 to be positive. When (3.8) is satisfied

$$\bar{\phi}-\beta\max\{\tau_q, \tau_s\} \geq \frac{\alpha}{2}(\tau_q+\tau_s)-\beta\max\{\tau_q, \tau_s\}-\epsilon\gamma > 0$$

for sufficiently small $\epsilon > 0$ in our choice of μ_1 and μ_2. Therefore, we choose ρ_2 such that

$$\rho_2 = \rho_1\bar{\phi}-\kappa(\beta\max\{\tau_q, \tau_s\}-\bar{\phi})$$

for some $\kappa \in (0, 1)$. Then $\pi_4 = \kappa(\bar{\phi}-\beta\max\{\tau_q, \tau_s\})$ is bounded below by a positive scalar, depending only on α, β, τ_q, and τ_s by our choice of μ_2. Since our choice of ρ_1, ρ_2 satisfies (5.22), π_2, π_3 are also bounded below by a positive scalar. Finally, since π_4 is at least a positive scalar, we can always take $\epsilon > 0$ small enough that π_1 is also bounded below by a positive scalar depending only on α, β, τ_q and τ_s. The case when $q \in \{1, \ldots, k\}$ and $s = k+1$ follows by an identical argument.

It remains to show that this particular choice of Π yields the desired lower bound on $\bar{\mathbf{x}}^T S_2\bar{\mathbf{x}}$. Let $\mathbf{u}_q := \bar{\mathbf{x}}(U_q)$ and $\mathbf{v}_q := \bar{\mathbf{x}}(V_q)$ denote the entries of $\bar{\mathbf{x}}$ indexed by U_q and V_q respectively, for all $q = 1, \ldots, k+1$. For all $q = 1, \ldots, k$, we have

$\mathbf{u}_q^T\mathbf{e} = -\tau_q\mathbf{v}_q^T\mathbf{e}$ since $\bar{\mathbf{x}}$ is orthogonal to span $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. Fix $s \in \{1, \ldots, k\}$. By our choice of $\pi_1^{k+1,s}$, $\pi_2^{k+1,s}$, $\pi_3^{k+1,s}$, and $\pi_4^{k+1,s}$ we have

$$S_2(C_{k+1}, C_s) = S_2(C_s, C_{k+1})^T = \frac{1}{2}\begin{pmatrix} (\bar{\lambda}_{k+1,s} + \beta/\tau_s)\mathbf{e}\mathbf{e}^T & -(\tau_s\bar{\lambda}_{k+1,s} + \beta)\mathbf{e}\mathbf{e}^T \\ -(\bar{\phi}_{k+1,s}/\tau_s + \beta)\mathbf{e}\mathbf{e}^T & (\bar{\phi}_{k+1,s} + \tau_s\beta)\mathbf{e}\mathbf{e}^T \end{pmatrix}$$
$$= \frac{1}{2}\begin{pmatrix} (\bar{\lambda}_{k+1,s} + \beta/\tau_s)\mathbf{e} \\ -(\bar{\phi}_{k+1,s}/\tau_s + \beta)\mathbf{e} \end{pmatrix}(\mathbf{e}^T \ -\tau_s\mathbf{e}^T).$$

It follows that

$$\begin{aligned} &\sum_{s=1}^{k} \bar{\mathbf{x}}(C_{k+1})^T S_2(C_{k+1}, C_s)\bar{\mathbf{x}}(C_s) \\ &\quad \geq -\frac{1}{2}\sum_{s=1}^{k} \max\left\{\bar{\lambda} + \frac{\beta}{\tau_s}, \frac{\bar{\phi}}{\tau_s} + \beta\right\} \|\bar{\mathbf{x}}(C_{k+1})\|_1 \left(\|\mathbf{u}_s\|_1 + \tau_s\|\mathbf{v}_s\|_1\right) \\ &\quad \geq= -\hat{c}\|\bar{\mathbf{x}}(C_{k+1})\|_1 \left(\|\bar{\mathbf{x}}\|_1 - \|\bar{\mathbf{x}}(C_{k+1})\|_1\right), \end{aligned}$$

where $\tau_{\min} := \min_{i=1,\ldots,k} \tau_i$, $\tau_{\max} := \max_{i=1,\ldots,k} \tau_i$, and

$$\hat{c} := \frac{1}{2}\left(\frac{\alpha}{2} + \beta\right)\left(\frac{\max\{\tau_{\max}, 1\}}{\min\{\tau_{\min}, 1\}}\right).$$

The optimization problem

$$\max_{\mathbf{w}_1 \in \mathbf{R}^{\ell_1}, \mathbf{w}_2 \in \mathbf{R}^{\ell_2}} \left\{\|\mathbf{w}_1\|_1\|\mathbf{w}_2\|_1 : \|\mathbf{w}_1\|^2 + \|\mathbf{w}_2\|^2 = \Psi^2\right\}$$

has optimal solution $\mathbf{w}_1^* = (\Psi/\sqrt{2\ell_1})\mathbf{e}$, $\mathbf{w}_2^* = (\Psi/\sqrt{2\ell_2})\mathbf{e}$, with optimal value $\Psi^2\sqrt{\ell_1\ell_2}/2$. Taking $\mathbf{w}_1 := \bar{\mathbf{x}}(C_{k+1})$ and $\mathbf{w}_2 = (\bar{\mathbf{x}}(C_1); \ldots; \bar{\mathbf{x}}(C_k))$ and $\Psi = \|\bar{\mathbf{x}}\|$, shows that

$$\|\bar{\mathbf{x}}(C_{k+1})\|_1\left(\|\bar{\mathbf{x}}\|_1 - \|\bar{\mathbf{x}}(C_{k+1})\|_1\right) \leq \frac{\|\bar{\mathbf{x}}\|^2}{2}\sqrt{r_{k+1}(N - r_{k+1})}$$

and, consequently,

$$\sum_{s=1}^{k} \bar{\mathbf{x}}(C_{k+1})^T S_2(C_{k+1}, C_s)\bar{\mathbf{x}}(C_s) \geq -\frac{\hat{c}\|\bar{\mathbf{x}}\|^2}{2}\sqrt{r_{k+1}N}. \tag{5.23}$$

Similarly,

$$\bar{\mathbf{x}}(C_q)^T S_2(C_q, C_q)\bar{\mathbf{x}}(C_q) = (\mathbf{v}_q^T\mathbf{e})^2\left(4\tau_q\alpha - \frac{\mu_1 + \mu_2}{n_q}\right) \tag{5.24}$$

for all $q = 1, \ldots, k$. Finally, for $q, s \in \{1, \ldots, k\}$ such that $q \neq s$, we have

$$\begin{aligned}
&\bar{\mathbf{x}}(C_q)^T S_2(C_q, C_s)\bar{\mathbf{x}}(C_s) \\
&= (\mathbf{v}_q^T \mathbf{e})(\mathbf{v}_s^T \mathbf{e})\Big(\tau_q \tau_s \bar{\lambda}^{q,s} + \beta(\tau_q + \tau_s) + \bar{\phi}^{q,s} - \tau_q \tau_s(\pi_1^{q,s} - \pi_2^{q,s} - \pi_3^{q,s} + \pi_4^{q,s})\Big) \qquad (5.25) \\
&= 4(\mathbf{v}_q^T \mathbf{e})(\mathbf{v}_s^T \mathbf{e})(\bar{\phi}^{q,s} - \tau_q \tau_s \pi_4^{q,s}). \qquad (5.26)
\end{aligned}$$

Here (5.26) is obtained by substituting (5.21), into (5.25). Let $\bar{v}_q := \mathbf{v}_q^T \mathbf{e}$ for all $q = 1, \ldots, k$. Combining (5.23), (5.24) and (5.26) shows that

$$\begin{aligned}
&\bar{\mathbf{x}}^T S_2 \bar{\mathbf{x}} \\
&\geq -\hat{c}\|\bar{\mathbf{x}}\|^2 \sqrt{r_{k+1} N} + \sum_{q=1}^{k} \bar{v}_q^2 \left(\tau_q \alpha - \frac{\mu_1 + \mu_2}{n_q}\right) \\
&\quad + 2\sum_{q=1}^{k} \sum_{s=q+1}^{k} 4\bar{v}_q \bar{v}_s((1-\kappa)\bar{\phi}^{q,s} + \kappa\beta\tilde{\tau}_{qs}) \\
&\geq -\hat{c}\|\bar{\mathbf{x}}\|^2 \sqrt{r_{k+1} N} + 8\sum_{q=1}^{k} \sum_{s=q+1}^{k} |\bar{v}_q \bar{v}_s| \\
&\quad \times \left(\alpha\tau_{\min} - \frac{\mu_1 + \mu_2}{4\hat{n}} - (1-\kappa)\bar{\phi}^{q,s} - \kappa\beta\tilde{\tau}_{qs}\right),
\end{aligned}$$

where $\tilde{\tau}_{qs} := \max\{\tau_q, \tau_s\}$, since $\sum_{q=1}^{k} \bar{v}_q^2 \geq \sum_{q \neq s} |\bar{v}_q \bar{v}_s|$. If $\alpha\tau_{\min} > \beta\tau_i$ for all $i = 1, \ldots, k$ then, for all $\epsilon > 0$ sufficiently small and κ sufficiently close to 1, we have

$$\begin{aligned}
&\alpha\tau_{\min} - \frac{\mu_1 + \mu_2}{4\hat{n}} - (1-\kappa)\bar{\phi}^{q,s} - \kappa\beta \max\{\tau_q, \tau_s\} \\
&\quad \geq \alpha\tau_{\min} - \beta \max\{\tau_q, \tau_s\} - (1-\kappa)(\alpha - \beta)\max\{\tau_q, \tau_s\} \\
&\quad - \frac{\epsilon\gamma}{4}\left(\frac{\hat{m}}{\hat{n}} + 1\right) \geq 0
\end{aligned}$$

for all $q \neq s$. It follows immediately that $\bar{\mathbf{x}}^T S_2 \bar{\mathbf{x}} \geq -\hat{c}\|\bar{\mathbf{x}}\|^2 \sqrt{r_{k+1} N}$. □

Substituting the respective bounds on $\bar{\mathbf{x}}^T S_i \bar{\mathbf{x}}$ into (5.18) shows that

$$\bar{\mathbf{x}}^T S \bar{\mathbf{x}} \geq \left(\min\{\mu_1, \mu_2\} - \hat{c}\sqrt{r_{k+1} N} - \|S_1\|\right)\|\bar{\mathbf{x}}\|^2. \qquad (5.27)$$

Since μ_1, μ_2 are both scalar multiples of $\hat{n}$, where the scalar depends only on $\alpha, \beta, \tau_1, \ldots, \tau_{k+1}$, there exists scalar $\xi > 0$, also depending only $\alpha, \beta, \tau_1, \ldots, \tau_{k+1}$, such that the right-hand side of (5.27) is nonnegative if $\|S_1\| + \hat{c}\sqrt{r_{k+1} N} \leq \xi\gamma\hat{n}$.

It remains to show that Z^* is the unique optimal solution with high probability if (5.17) holds. An argument similar to that in the proof of Theorem 4.2 show that Z^* is the unique optimal solution of (3.5) if $n_s \mathbf{e}^T W_{U_q,V_q} \mathbf{e} > n_q \mathbf{e}^T W_{U_q,V_S} \mathbf{e}$ for all $q, s \in \{1, \ldots, k\}$. Moreover, an argument identical to that of the proof of Lemma 4.4, establishes that this uniqueness condition holds with high probability for sufficiently large $\hat{n}$. This completes the proof.

5.2 Positive semidefiniteness of S

It remains to show that S, as defined by (5.1), satisfies (5.17) to prove that Z^* is the unique optimal solution of (3.5). In particular, we will derive the following upper bound on the spectral norm of S_1.

Lemma 5.2 *There exist scalars $c_1, c_2 > 0$ such that*

$$\|S_1\| \leq c_1\sqrt{kN} + c_2\sqrt{N} + \beta\tau_{k+1}n_{k+1} \tag{5.28}$$

with probability tending exponentially to 1 *as $\hat{n}$ approaches ∞.*

To establish Lemma 5.2, we decompose S_1 as $S_1 = \tilde{S}_1 + \tilde{S}_2 + \tilde{S}_3 + \tilde{S}_4$, where $\tilde{S}_i \in \Sigma^{M+N}$, $i = 1, \ldots, 4$, are defined as follows. We take

$$\begin{aligned}\tilde{S}_1(U_q, U_s) &= (\lambda_{U_q} - \mathbf{E}[\lambda_{U_q}])\mathbf{e}^T + \mathbf{e}(\lambda_{U_s} - \mathbf{E}[\lambda_{U_s}])^T,\\ \tilde{S}_1(V_q, V_s) &= (\phi_{V_q} - \mathbf{E}[\phi_{V_q}])\mathbf{e}^T + \mathbf{e}(\phi_{V_s} - \mathbf{E}[\phi_{V_s}])^T,\end{aligned}$$

for all $q, s \in \{1, \ldots, k+1\}$ and set all remaining entries of $\tilde{S}_1$ to be 0. Next, let

$$\tilde{S}_2(U_q, V_s) = \begin{cases} \beta\mathbf{e}\mathbf{e}^T - R^{q,q}, & \text{if } q = s,\ q \in \{1, \ldots, k\}\\ \beta\mathbf{e}\mathbf{e}^T - W_{U_q,V_s}, & \text{otherwise,}\end{cases}$$

where $R^{q,q}$ is a $m_q \times n_q$ random matrix with independent identically distributed (i.i.d.) entries sampled according to Ω_2, the distribution of the off-diagonal blocks of W. We choose $\tilde{S}_2(V_q, U_s) = \tilde{S}_2(U_s, V_q)^T$ and set all other entries of $\tilde{S}_2$ equal to 0. Next, we set $\tilde{S}_3(U_q, V_q) = \alpha\mathbf{e}\mathbf{e}^T - W_{U_q,V_q}$ and $\tilde{S}_3(V_q, U_q) = \tilde{S}_3(U_q, V_q)^T$ for all $q = 1, \ldots, k$, and set all remaining entries of $\tilde{S}_3$ equal to 0. Finally, $\tilde{S}_4$ is the correction matrix for the diagonal blocks of $\tilde{S}_2$. That is, we take $\tilde{S}_4(U_q, V_q) = R^{q,q} - \beta\mathbf{e}\mathbf{e}^T$, and $\tilde{S}_4(V_q, U_q) = \tilde{S}_4(U_q, V_q)^T$, for all $q = 1, \ldots, k$, we take $\tilde{S}_4(U_{k+1}, V_{k+1}) = \tilde{S}_4(V_{k+1}, U_{k+1})^T = -\beta\mathbf{e}\mathbf{e}^T$, and all remaining entries of $\tilde{S}_4$ are 0. To obtain the desired bound on $\|S_1\|$, we bound each of $\|\tilde{S}_1\|$, $\|\tilde{S}_2\|$, $\|\tilde{S}_3\|$, and $\|\tilde{S}_4\|$ individually. To do so, we will repeatedly invoke the following bound on the norm of a random rectangular matrix (see Geman [22] and [4, Theorem 2]).

Theorem 5.3 *Let A be a $\lceil yn \rceil \times n$ random matrix with independent identically distributed (i.i.d.) entries sampled from a distribution with mean μ and variance σ^2 such that $A_{ij} \in [0, 1]$ for all $i \in \{1, \ldots, \lceil yn \rceil\}$, $j \in \{1, \ldots, n\}$ for fixed $y \in \mathbf{R}_+$. Then there*

exist $c_1, c_2, c_3, c_4 > 0$ *depending only on* σ *and* y *such that* $\|A - \mu \mathbf{e}\mathbf{e}^T\| \leq c_4 \sigma \sqrt{n}$ *with probability at least* $1 - c_1 \exp(-c_2 n^{c_3})$.

Applying Theorem 5.3 shows that $\|\tilde{S}_2\| = \|\tilde{S}_2(U, V)\| \leq \tilde{c}_2 \sqrt{N}$ for some scalar $\tilde{c}_2$ with probability tending exponentially to 1 as $\hat{n} \to \infty$ by the block structure of $\tilde{S}_2$. A similar argument shows that there exists scalar $\tilde{c}_3 > 0$ such that $\|\tilde{S}_3\| \leq \tilde{c}_3 \max_{q=1,\ldots,k} \sqrt{n_q} \leq \tilde{c}_3 \sqrt{N}$ with probability tending exponentially to 1 as $\hat{n} \to \infty$. Next,

$$\|\tilde{S}_4\| = \max \left\{ \tilde{c}_4 \max_{q=1,\ldots,k} \sqrt{n_q}, \beta \sqrt{m_{k+1} n_{k+1}} \right\} \leq \tilde{c}_4 \sqrt{N} + \beta \tau_{k+1} n_{k+1}$$

for some scalar $\tilde{c}_4 > 0$ with probability tending exponentially to 1 as $\hat{n} \to \infty$, again by Theorem 5.3. Finally, a calculation similar to the derivation of the bound on $\|\tilde{S}_2\|$ in the proof of Lemma 4.5 shows that $\|\tilde{S}_1\| = O(\sqrt{kN})$ with probability tending exponentially to 1 as $\hat{n} \to \infty$. Applying the triangle inequality and the union bound completes the proof.

6 Numerical experiments

In this section, we empirically verify the performance of our heuristics for a variety of program inputs. Specifically, we randomly generate symmetric matrices W according to the planted cluster model, for a number of distributions on the entries of W and partitions $\{C_1, \ldots, C_{k+1}\}$ of the rows and columns of W, and compare the optimal solution of (2.5) to that corresponding to the planted partition. Similarly, we also compare the optimal solution of (3.5) to the matrix representation of the planted partition for W sampled from the planted bicluster model.

In each experiment, we solve either (2.5) or (3.5) using the Alternating Direction Method of Multipliers (ADMM). A comprehensive description of ADMM and similar algorithms is well beyond the scope of this manuscript; we direct the reader to the recent survey [9] for more details. To solve (2.5), we represent the feasible region as the intersection of two sets and apply ADMM to solve the resulting equivalent formulation. In particular, let $\Xi := \{X \in \Sigma^V : X\mathbf{e} \leq \mathbf{e}, \ X \geq 0\}$, and $\Omega := \{X \in \Sigma^V : \mathrm{Tr}(X) = k, \ X \in \Sigma_+^V\}$. Then we may rewrite (2.5) as $\max\{\mathrm{Tr}(WY) : X - Y = 0, \ X \in \Xi, \ Y \in \Omega\}$. We solve this problem iteratively as follows. In each iteration, we approximately minimize the augmented Lagrangian $L_\beta(X, Y, U) = \mathrm{Tr}(WY) - \mathrm{Tr}(U(X - Y)) + \frac{\beta}{2}\|X - Y\|_F^2$ with respect to Y and X successively, and then update the dual multiplier U as $U = U - \beta(X - Y)$.[1] Here $\|\cdot\|_F$ denotes the Frobenius norm on Σ^V defined by $\|X\|_F^2 = \mathrm{Tr}(X^2)$. As we will see, the resulting subproblems are convex and can be solved efficiently; therefore, this algorithm will converge to the optimal solution of (2.5) (see [17, Theorem 8]).

[1] The penalty parameter $\beta = \min\{\max\{5n/k, 80\}, 500\}/2$ was chosen via simulation, and seems to work well for most problem instances.

Let (X^k, Y^k, U^k) be the current iterate after k iterations. To update in the Y direction, we minimize L_β with respect to Y. That is, Y^{k+1} is a minimizer of the subproblem

$$\min_{Y \in \Omega} \frac{1}{2} \left\| Y - \left(X^k - \frac{W + U^k}{\beta} \right) \right\|_F^2. \tag{6.1}$$

Let $X^k - (W + U^k)/\beta$ have eigenvalue decomposition $V \mathrm{Diag}\,(\mathbf{v}^k) V^T$. Then, by the fact that both the Frobenius norm and the set Ω are invariant under unitary similarity transformations, we have $Y^{k+1} = V \mathrm{Diag}\,(\mathbf{y}^*) V^T$, where $\mathbf{y}^*$ is the optimal solution of $\min\{\|\mathbf{y} - \mathbf{v}^k\|^2 : \mathbf{e}^T \mathbf{y} = k, \ \mathbf{y} \geq 0\}$, by [29, Proposition 2.6]. This latter subproblem admits an analytic solution, which can be computed efficiently; see Berg and Friedlander [45].

Next, we take X^{k+1} to be the optimal solution of

$$\min_{X \in \Xi} \frac{1}{2} \left\| X - \left(Y^k + \frac{U^k}{\beta} \right) \right\|_F^2. \tag{6.2}$$

Unfortunately, this subproblem does not admit a closed-form solution. Instead, we approximately solve (6.2) by applying the spectral projected gradient method of Birgin et al. [8] to the dual of (6.2). Taking the dual of (6.2) shows that

$$X^{k+1} = \left[\left(Y^{k+1} + \frac{U^k}{\beta} \right) - \frac{\mathbf{z}^* \mathbf{e}^T + \mathbf{e} (\mathbf{z}^*)^T}{2} \right]_+,$$

where $\mathbf{z}^*$ is the optimal solution of the dual problem

$$\min_{\mathbf{z} \geq 0} \frac{1}{2} \left\| \left[\left(Y^{k+1} + \frac{U^k}{\beta} \right) - \frac{\mathbf{z}\mathbf{e}^T + \mathbf{e}\mathbf{z}^T}{2} \right]_+ \right\|_F^2 + \mathbf{z}^T \mathbf{e} - \frac{1}{2} \left\| Y^{k+1} + \frac{U^k}{\beta} \right\|_F^2; \tag{6.3}$$

Here, the operator $[\cdot]_+ : \Sigma^V \to \Sigma^V \cap \mathbf{R}_+^{V \times V}$ maps each $Z \in \Sigma^V$ to the matrix with (i, j)th entry equal to $[[Z]_+]_{ij} = \max\{0, Z_{ij}\}$ for all $i, j \in V$. Moreover, the objective function of the dual problem (6.3) is both differentiable and coercive in $\mathbf{z}$, and, therefore, the dual can be solved efficiently [8]. The algorithm is stopped when the relative duality gap $|v_p^{(k)} - v_d^{(k)}| / \max\{|v_p^{(k)}, 1\}$ and primal constraint violation are smaller than a desired error tolerance.

We solve (3.5) in a similar manner. In particular, we apply ADMM to minimize the augmented Lagrangian of the convex program $\max\{\mathrm{Tr}\,(WY) : X - Y = 0, \ X \in \Xi_B, \ Y \in \Omega_B\}$, where $\Xi_B = \{X \in \Sigma^{U \cup V} : X_{U,U} \mathbf{e} \leq \mathbf{e}, \ X_{V,V} \mathbf{e} \leq \mathbf{e}, \ X \geq 0\}$ and $\Omega_B = \{Y \in \Sigma_+^{U \cup V} : \mathrm{Tr}\,(Y) = 2k\}$. It is important to note that Ω_B is a relaxation of the set $\{Y \in \Sigma^{U \cup V} : \mathrm{Tr}\,(Y_{UU}) = k, \ \mathrm{Tr}\,(Y_{VV}) = k\}$ and, therefore, we are actually applying ADMM to solve a relaxation of (3.5). Here, the penalty parameter $\beta = \min\{\max\{5n/k, 80\}, 500\}$ is used in the augmented Lagrangian. As before, the

subproblem to update Y admits a closed-form solution using simplex projection, and we update X by applying the spectral projected gradient method to the dual subproblem

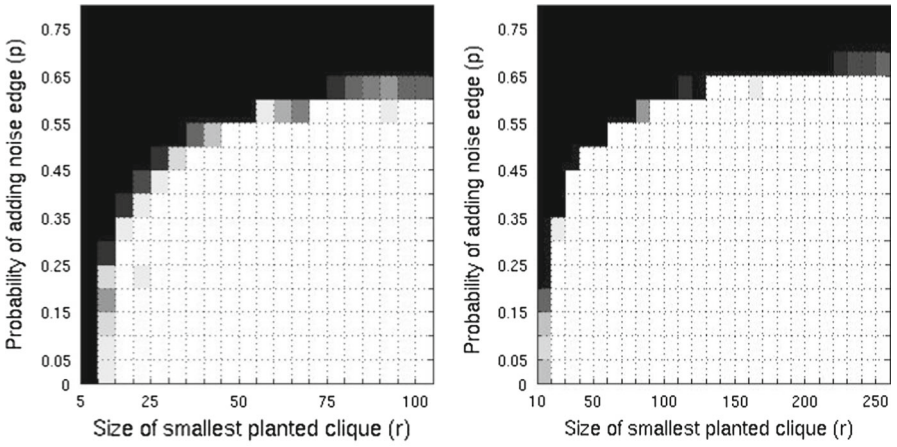
$$\min_{\lambda\geq 0,\phi\geq 0} \left\| \left[\left(Y^{k+1} + \frac{U^k}{\beta} \right) - \Lambda \right]_+ \right\|_F^2 + \lambda^T \mathbf{e} + \phi^T \mathbf{e} - \frac{1}{2} \left\| Y^{k+1} + \frac{U^k}{\beta} \right\|_F^2,$$

where

$$\Lambda := \begin{pmatrix} \frac{1}{2}(\lambda \mathbf{e}^T + \mathbf{e}\lambda^T) & 0 \\ 0 & \frac{1}{2}(\phi \mathbf{e}^T + \mathbf{e}\phi^T) \end{pmatrix}.$$

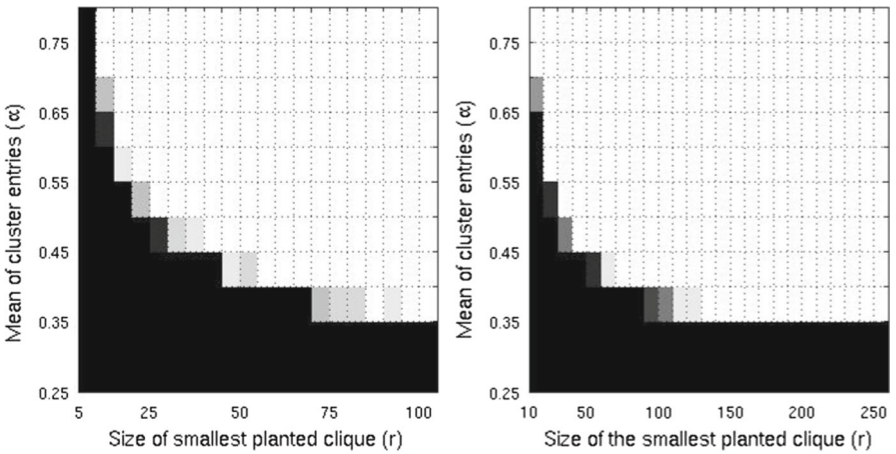
Again, we stop the algorithm when both the relative duality gap and primal constraint violation are within a desired error tolerance.

For $N = 200$ and $N = 500$ and a variety of choices of $\hat{r} \in \{1, \ldots, N\}$, the following procedure was repeated 10 times. We first partition the indices $\{1, \ldots, N\}$ into $k = \lfloor N/\hat{r} \rfloor$ subsets $\{C_1, C_2, \ldots, C_k\}$ of size at least $\hat{r}$. We then generate a random symmetric matrix $W \in \Sigma^n$ according to the planted cluster model with respect to $\{C_1, C_2, \ldots, C_k\}$ and one of two sets of probability distributions. In the first, W_{ij} is a Bernoulli random variable with probability of success 0.75 if i, j both belong to C_ℓ for some $\ell \in \{1, 2, \ldots, k\}$ and is a Bernoulli random variable with probability of success p otherwise, for some fixed probability p. In the second, each W_{ij} is Gaussian with $\mu = \alpha$, $\sigma = 0.25$ for some $\alpha \in [0.25, 1]$ if i and j belong to the same block, and $(\mu, \sigma^2) = (0.25, 0.25)$ otherwise. For each choice of p and α, the ADMM procedure described above is called to approximately solve (2.5). In each experiment, the algorithm is terminated if the stopping criteria is achieved with error tolerance $\epsilon = 10^{-5}$ or after 500 iterations, and the subproblem (6.2) is solved to



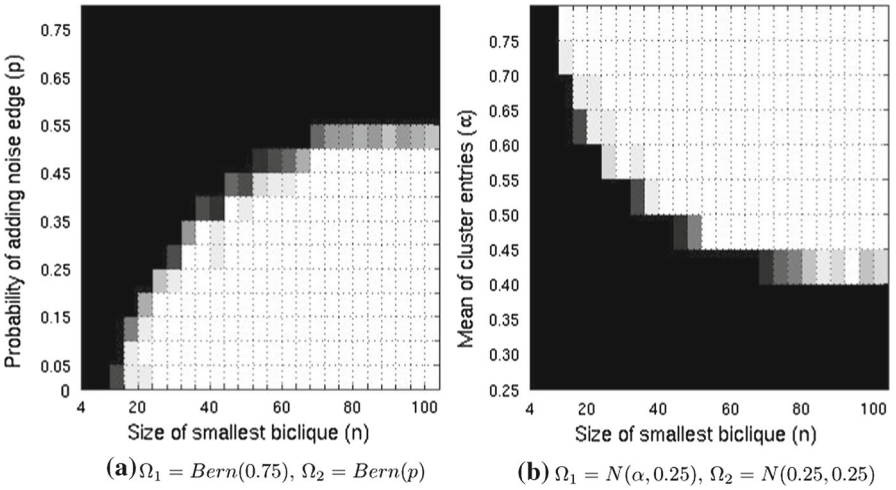
(a) $N = 200$ **(b)** $N = 500$

Fig. 1 Number of recoveries for N-node graph with k planted cliques of size at least $\hat{r}$ and W generated according to the distributions $\Omega_1 = Bern(0.75)$, $\Omega_2 = Bern(p)$. *Brighter colours* indicate a higher rate of recovery



(a) $N = 200$ **(b)** $N = 500$

Fig. 2 Number of recoveries for N-node graph with k planted cliques of size at least $\hat{r}$ and W generated according to the distributions $\Omega_1 = N(\alpha, 0.25)$, $\Omega_2 = N(0.25, 0.25)$



(a) $\Omega_1 = Bern(0.75)$, $\Omega_2 = Bern(p)$ **(b)** $\Omega_1 = N(\alpha, 0.25)$, $\Omega_2 = N(0.25, 0.25)$

Fig. 3 Number of recoveries for (250, 200)-node bipartite graph with k planted bicliques of size at least $(1.25\hat{n}, \hat{n})$ and W generated according to the distributions Ω_1 and Ω_2

within error tolerance $\epsilon/10$ during each iteration. Let X^* denote the optimal solution for (2.5) returned by the ADMM algorithm. We declare the block structure of W to be successfully recovered if $\|X^* - X_0\|_F^2 / \|X_0\|_F^2 < 10^{-3}$, where X_0 is the proposed solution constructed according to (2.6).

Figures 1 and 2 display the average number of successes for each choice of $\hat{r}$ for W sampled from the planted cluster model according to the Bernoulli and Gaussian distributions, respectively. The empirical performance of our heuristic appears to match that predicted by Theorem 2.1. For each choice of p or α, there is a sharp phase transition between zero and perfect recovery as $\hat{r}$ increases past some threshhold. It should

be noted that the recovery guarantees given by Theorem 2.1 appear to be conservative compared to those observed empirically; we have perfect recovery for values of $\hat{r}$ smaller than the left-hand side of (2.8) for many trials. Moreover, W generated according to the Gaussian model do not necessarily satisfy the assumption $0 \leq W_{ij} \leq 1$ for $i, j \in \{1, \ldots, N\}$.

We repeated the experiment for bipartite graphs drawn from the planted bicluster model. For $(M, N) = (250, 200)$, and various minimum bicluster sizes $(\hat{m}, \hat{n})$, we randomly sample weight matrices W from the planted bicluster model according to some partition of $\{1, \ldots, M\} \times \{1, \ldots, N\}$ into $k = \lfloor N/\hat{n} \rfloor$ bicliques with left and right vertex sets of size at least $\hat{m}$ and $\hat{n}$ respectively. For each W, we solve (3.5) using the ADMM algorithm described above and declare the block structure to be recovered if the returned optimal solution Z^* satisfies $\|Z^* - Z_0\|_F^2 / \|Z_0\|_F^2 < 10^{-3}$, where Z_0 is the proposed solution constructed according to (3.6). We plot the number of successful recoveries for each $(\hat{m}, \hat{n})$ and generating distribution in Fig. 3. As before, the empirical behaviour of our heuristic reflects that predicted by Theorem 3.1, although there is some evidence that this theoretical recovery guarantee may be overly pessimistic.

Acknowledgments This research was supported in part by the Institute for Mathematics and its Applications with funds provided by the National Science Foundation, and by a Postgraduate Scholarship from NSERC (Natural Science and Engineering Research Council of Canada). I am grateful to Stephen Vavasis, Henry Wolkowicz, Levent Tunçel, Shai Ben-David, Inderjit Dhillon, Ben Recht, and Ting Kei Pong for their helpful comments and suggestions. I am especially grateful to Ting Kei for his help implementing the ADMM algorithms used to perform the numerical trials. I would also like to thank Warren Schudy for suggesting some relevant references that were omitted in an earlier version. Finally, I thank the two anonymous reviewers, whose suggestions vastly improved the presentation and organization of this paper.

References

1. Ackerman, M., Ben-David, S.: Clusterability: a theoretical study. In: Proceedings of the AISTATS-09. JMLR: W&CP **5**, 1–8 (2009)
2. Aloise, D., Deshpande, A., Hansen, P., Popat, P.: Np-hardness of euclidean sum-of-squares clustering. Mach. Learn. **75**(2), 245–248 (2009)
3. Ames, B., Vavasis, S.: Convex optimization for the planted k-disjoint-clique problem. Arxiv, preprint arXiv:1008.2814, 2010
4. Ames, B., Vavasis, S.: Nuclear norm minimization for the planted clique and biclique problems. Math. Program. **129**(1), 1–21 (2011)
5. Balakrishnan, S., Xu, M., Krishnamurthy, A., Singh, A.: Noise thresholds for spectral clustering. Adv. Neural Inf. Process. Syst. **25**(3), 954–962 (2011)
6. Bansal, N.A., Blum, A., Chawla, S.: Correlation clustering. Mach. Learn. **56**(1), 89–113 (2004)
7. Berkhin, P.: A survey of clustering data mining techniques. In: Kogan, J., Nicholas, C., Teboulle, M. (eds.) Grouping Multidimensional Data: Recent Advances in Clustering, pp. 25–71. Springer, Berlin, Heidelberg (2006)
8. Birgin, E., Martínez, J., Raydan, M.: Nonmonotone spectral projected gradient methods on convex sets. SIAM J. Optimiz. **10**(4), 1196–1211 (2000)
9. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends. Mach. Learn. **3**(1), 1–122 (2011)
10. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
11. Busygin, S., Prokopyev, O., Pardalos, P.: Biclustering in data mining. Comput. Oper. Res. **35**(9), 2964–2987 (2008)
12. Candès, E., Plan, Y.: Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. Arxiv, preprint arXiv:1001.0339 (2010)

13. Candès, E., Recht, B.: Exact matrix completion via convex optimization. Found. Comput. Math. **9**(6), 717–772 (2009)
14. Candès, E., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
15. Candès, E., Tao, T.: Decoding by linear programming. IEEE Trans. Inf. Theory **51**(12), 4203–4215 (2005)
16. Donoho, D.: Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
17. Eckstein, J., Bertsekas, D.P.: On the douglasrachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**(1–3), 293–318 (1992)
18. Fan, N., Boyko, N., Pardalos, P.: Recent advances of data biclustering with application in computational neuroscience. Computat. Neurosci. pp. 105–132 (2010)
19. Fan, N., Pardalos, P.: Multi-way clustering and biclustering by the ratio cut and normalized cut in graphs. J. Comb. Optimiz. **23**(2), 224–251 (2012)
20. Flynn, C., Perry, P.: Consistent biclustering. Arxiv, preprint arXiv:1206.6927 (2012)
21. Füredi, Z., Komlós, J.: The eigenvalues of random symmetric matrices. Combinatorica **1**(3), 233–241 (1981)
22. Geman, S.: A limit theorem for the norm of random matrices. Ann. Prob. **8**(2), 252–261 (1980)
23. Golub, G.Van, Loan, C.: Matrix Computations. Johns Hopkins University Press, Baltimore (1996)
24. Gross, D.: Recovering low-rank matrices from few coefficients in any basis. IEEE Trans. Inf. Theory **57**(3), 1548–1566 (2011)
25. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**, 13–30 (1962)
26. Jalali, A., Chen, Y., Sanghavi, S., Xu, H.: Clustering partially observed graphs via convex optimization. Arxiv, preprint arXiv:1104.4803 (2011)
27. Kannan, R., Vempala, S., Vetta, A.: On clusterings: good, bad and spectral. J. ACM **51**(3), 497–515 (2004)
28. Kolar, M., Balakrishnan, S., Rinaldo, A., Singh, A.: Minimax localization of structural information in large noisy matrices. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P. (eds.) Advances in Neural Information Processing Systems, pp. 909–917. FCN, Weinberger (2011)
29. Lu, Z., Zhang, Y.: Penalty decomposition methods for rank minimization. Technical report. Department of Mathematics, Simon Fraser University. Arxiv, preprint (2010)
30. Mathieu, C., Schudy, W.: Correlation clustering with noisy input. In: Proceedings of the twenty-first annual ACM-SIAM symposium on discrete algorithms, pp. 712–728. Society for Industrial and Applied Mathematics (2010)
31. Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: analysis and an algorithm. Adv. Neural Inf. Process. Syst. **2**, 849–856 (2002)
32. Ostrovsky, R., Rabani, Y., Schulman, L., Swamy, C.: The effectiveness of Lloyd-type methods for the k-means problem. In: Proceedings of 47st Annual IEEE Symposium on the Foundations of Computer Science (2006)
33. Oymak, S., Hassibi, B.: Finding dense clusters via "low rank + sparse" decomposition. Arxiv, preprint arXiv:1104.5186 (2011)
34. Oymak, S., Mohan, K., Fazel, M., Hassibi, B.: A simplified approach to recovery conditions for low rank matrices. In: Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on IEEE, pp. 2318–2322 (2011)
35. Peng, J., Wei, Y.: Approximating k-means-type clustering via semidefinite programming. SIAM J. Optimiz. **18**(1), 186–205 (2007)
36. Recht, B.: A simpler approach to matrix completion. J. Mach. Learn. Res. **12**, 3413–3430 (2011)
37. Recht, B., Fazel, M., Parrilo, P.: Guaranteed minimum-rank solutions of linear matrix equations via muclear norm minimization. SIAM Rev. **52**, 471 (2010)
38. Recht, B., Xu, W., Hassibi, B.: Null space conditions and thresholds for rank minimization. Mathematical Programming, pp. 1–28 (2010)
39. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1997)
40. Rohe, K., Chatterjee, S., Yu, B.: Spectral clustering and the high-dimensional stochastic blockmodel. Ann. Stat. **39**(4), 1878–1915 (2011)
41. Rohe, K., Yu, B.: Co-clustering for directed graphs; the stochastic co-blockmodel and a spectral algorithm. Arxiv, preprint arXiv:1204.2296 (2012)

42. Shamir, R., Tsur, D.: Improved algorithms for the random cluster graph model. Alg. Theory SWAT **2002**, 230–239 (2002)
43. Singh, V., Mukherjee, L., Peng, J., Xu, J.: Ensemble clustering using semidefinite programming with applications. Mach. Learn. **79**(1), 177–200 (2010)
44. Tunçel, L.: Polyhedral and Semidefinite Programming Methods in Combinatorial Optimization. American Mathematical Society, Providence (2010)
45. Van Den Berg, E., Friedlander, M.: Probing the pareto frontier for basis pursuit solutions. SIAM J. Sci. Comput. **31**(2), 890–912 (2008)