Graph Clustering With Missing Data : Convex Algorithms and Analysis

Ramya Korlakai Vinayak, Samet Oymak, Babak Hassibi
Department of Electrical Engineering
California Institute of Technology, Pasadena, CA 91125
{ramya, soymak}@caltech.edu, hassibi@systems.caltech.edu

Abstract

We consider the problem of finding clusters in an unweighted graph, when the graph is partially observed. We analyze two programs, one which works for dense graphs and one which works for both sparse and dense graphs, but requires some a priori knowledge of the total cluster size, that are based on the convex optimization approach for low-rank matrix recovery using nuclear norm minimization. For the commonly used Stochastic Block Model, we obtain *explicit* bounds on the parameters of the problem (size and sparsity of clusters, the amount of observed data) and the regularization parameter characterize the success and failure of the programs. We corroborate our theoretical findings through extensive simulations. We also run our algorithm on a real data set obtained from crowdsourcing an image classification task on the Amazon Mechanical Turk, and observe significant performance improvement over traditional methods such as k-means.

1 Introduction

Clustering [1] broadly refers to the problem of identifying data points that are similar to each other. It has applications in various problems in machine learning, data mining [2, 3], social networks [4–6], bioinformatics [7, 8], etc. In this paper we focus on graph clustering [9] problems where the data is in the form of an unweighted graph. Clearly, to observe the entire graph on n nodes requires $\binom{n}{2}$ measurements. In most practical scenarios this is infeasible and we can only expect to have *partial observations*. That is, for some node pairs we know whether there exists an edge between them or not, whereas for the rest of the node pairs we do not have this knowledge. This leads us to the problem of clustering graphs with *missing data*.

Given the adjacency matrix of an *unweighted graph*, a cluster is defined as a set of nodes that are densely connected to each other when compared to the rest of the nodes. We consider the problem of identifying such clusters when the input is a partially observed adjacency matrix. We use the popular *Stochastic Block Model* (SBM) [10] or *Planted Partition Model* [11] to analyze the performance of the proposed algorithms. SBM is a random graph model where the edge probability depends on whether the pair of nodes being considered belong to the same cluster or not. More specifically, the edge probability is higher when both nodes belong to the same cluster. Further, we assume that each entry of the adjacency matrix of the graph is observed independently with probability r. We will define the model in detail in Section 2.1.

1.1 Clustering by Low-Rank Matrix Recovery and Completion

The idea of using convex optimization for clustering has been proposed in [12–21]. While each of these works differ in certain ways, and we will comment on their relation to the current paper in Section 1.3, the common approach they use for clustering is inspired by recent work on low-rank matrix recovery and completion via regularized nuclear norm (trace norm) minimization [22–26].

In the case of unweighted graphs, an ideal clustered graph is a union of disjoint cliques. Given the adjacency matrix of an unweighted graph with clusters (denser connectivity inside the clusters compared to outside), we can interpret it as an ideal clustered graph with missing edges inside the clusters and erroneous edges in between clusters. Recovering the low-rank matrix corresponding to the disjoint cliques is equivalent to finding the clusters.

We will look at the following well known convex program which aims to recover and complete the low-rank matrix ($\mathbf{L}$) from the partially observed adjacency matrix ($\mathbf{A}^{obs}$):

Simple Convex Program:

$$\underset{\mathbf{L},\mathbf{S}}{\text{minimize}} \ \|\mathbf{L}\|_\star + \lambda\|\mathbf{S}\|_1 \tag{1.1}$$

subject to

$$1 \geq \mathbf{L}_{i,j} \geq 0 \ \text{ for all } i,j \in \{1,2,\dots n\} \tag{1.2}$$

$$\mathbf{L}^{obs} + \mathbf{S}^{obs} = \mathbf{A}^{obs} \tag{1.3}$$

where $\lambda \geq 0$ is the regularization parameter, $\|.\|_\star$ is the nuclear norm (sum of the singular values of the matrix), and $\|.\|_1$ is the l_1-norm (sum of absolute values of the entries of the matrix). $\mathbf{S}$ is the sparse error matrix that accounts for the missing edges inside the clusters and erroneous edges outside the clusters on the observed entries. $\mathbf{L}^{obs}$ and $\mathbf{S}^{obs}$ denote entries of $\mathbf{L}$ and $\mathbf{S}$ that correspond to the observed part of the adjacency matrix.

Program 1.1 is very simple and intuitive. Further, it does not require any information other than the observed part of the adjacency matrix. In [13], the authors analyze Program 1.1 without the constraint (1.2). While dropping (1.2) makes the convex program less effective, it does allow [13] to make use of low-rank matrix completion results for its analysis. In [16] and [21], the authors analyze Program 1.1 when the entire adjacency matrix is observed. In [17], the authors study a slightly more general program, where the regularization parameter is different for the extra edges and the missing edges. However, the adjacency matrix is completely observed.

It is not difficult to see that, when the edge probability inside the cluster is $p < 1/2$, that (as $n \to \infty$) Program 1.1 will return $\mathbf{L}^0 = 0$ as the optimal solution (since if the cluster is not dense enough it is more costly to complete the missing edges). As a result our analysis of Program 1.1, and the main result of Theorem 1, assumes $p > 1/2$. Clearly, there are many instances of graphs we would like to cluster where $p < 1/2$. If the total size of the cluster region (i.e, the total number of edges in the cluster, denoted by $|\mathcal{R}|$) is known, then the following convex program can be used, and can be shown to work for $p < 1/2$ (see Theorem 2).

Improved Convex Program:

$$\underset{\mathbf{L},\mathbf{S}}{\text{minimize}} \ \|\mathbf{L}\|_\star + \lambda\|\mathbf{S}\|_1 \tag{1.4}$$

subject to

$$1 \geq \mathbf{L}_{i,j} \geq \mathbf{S}_{i,j} \geq 0 \ \text{ for all } i,j \in \{1,2,\dots n\} \tag{1.5}$$

$$\mathbf{L}_{i,j} = \mathbf{S}_{i,j} \text{ whenever } \mathbf{A}^{obs}_{i,j} = 0 \tag{1.6}$$

$$\text{sum}(\mathbf{L}) \geq |\mathcal{R}| \tag{1.7}$$

As before, $\mathbf{L}$ is the low-rank matrix corresponding to the ideal cluster structure and $\lambda \geq 0$ is the regularization parameter. However, $\mathbf{S}$ is now the sparse error matrix that accounts only for the missing edges inside the clusters on the observed part of adjacency matrix. [16] and [19] study programs similar to Program 1.4 for the case of a completely observed adjacency matrix. In [19], the constraint 1.7 is a strict equality. In [15] the authors analyze a program close to Program 1.4 but without the l_1 penalty.

If $\mathcal{R}$ is not known, it is possible to solve Problem 1.4 for several values of $\mathcal{R}$ until the desired performance is obtained. Our empirical results reported in Section 3, suggest that the solution is not very sensitive to the choice of $\mathcal{R}$.

1.2 Our Contributions

- We analyze the Simple Convex Program 1.1 for the SBM with partial observations. We provide *explicit bounds* on the regularization parameter as a function of the parameters of the SBM, that

characterizes the success and failure conditions of Program 1.1 (see results in Section 2.2). We show that clusters that are either too small or too sparse constitute the bottleneck. Our analysis is helpful in understanding the **phase transition** from failure to success for the simple approach.
- We also analyze the Improved Convex Program 1.4. We explicitly characterize the conditions on the parameters of the SBM and the regularization parameter for successfully recovering clusters using this approach (see results in Section 2.3).
- Apart from providing theoretical guarantees and corroborating them with simulation results (Section 3), we also apply Programs 1.1 and 1.4 on a real data set (Section 3.3) obtained by crowdsourcing an image labeling task on Amazon Mechanical Turk.

1.3 Related Work

In [13], the authors consider the problem of identifying clusters from partially observed unweighted graphs. For the SBM with partial observations, they analyze Program 1.1 without constraint (1.2), and show that under certain conditions, the minimum cluster size must be at least $\mathcal{O}(\sqrt{n(\log(n))^4/r})$ for successful recovery of the clusters. Unlike our analysis, the exact requirement on the cluster size is not known (since the constant of proportionality is not known). Also they do not provide conditions under which the approach fails to identify the clusters. Finding the explicit bounds on the constant of proportionality is critical to understanding the phase transition from failure to successfully identifying clusters.

In [14–19], analyze convex programs similar to the Programs 1.1 and 1.4 for the SBM and show that the minimum cluster size should be at least $\mathcal{O}(\sqrt{n})$ for successfully recovering the clusters. However, the exact requirement on the cluster size is not known. Also, they do not provide explicit conditions for failure, and except for [16] they do not address the case when the data is missing.

In contrast, we consider the problem of clustering with missing data. We explicitly characterize the constants by providing bounds on the model parameters that decide if Programs 1.1 and 1.4 can successfully identify clusters. Furthermore, for Program 1.1, we also explicitly characterize the conditions under which the program fails.

In [16], the authors extend their results to partial observations by scaling the edge probabilities by r (observation probability), which will *not* work for $r < 1/2$ or $1/2 < p < 1/2r$ in Program 1.1 . [21] analyzes Program 1.1 for the SBM and provides conditions for success and failure of the program when the entire adjacency matrix is observed. The dependence on the number of observed entries emerges non-trivially in our analysis. Further, [21] does not address the drawback of Program 1.1, which is $p > 1/2$, whereas in our work we analyze Program 1.4 that overcomes this drawback.

2 Partially Observed Unweighted Graph

2.1 Model

Definition 2.1 (Stochastic Block Model)**.** *Let $\mathbf{A} = \mathbf{A}^T$ be the adjacency matrix of a graph on n nodes with K disjoint clusters of size n_i each, $i = 1, 2, \cdots, K$. Let $1 \geq p_i \geq 0,\ i = 1, \cdots, K$ and $1 \geq q \geq 0$. For $l > m$,*

$$\mathbf{A}_{l,m} = \begin{cases} 1 \text{ w.p. } p_i, & \text{if both nodes } l, m \text{ are in the same cluster } i. \\ 1 \text{ w.p. } q, & \text{if nodes } l, m \text{ are not in the same cluster.} \end{cases} \tag{2.1}$$

If $p_i > q$ for each i, then we expect the density of edges to be higher inside the clusters compared to outside. We will say the random variable Y has a $\Phi(r, \delta)$ distribution, for $0 \leq \delta, r \leq 1$, written as $Y \sim \Phi(r, \delta)$, if

$$Y = \begin{cases} 1, \text{ w.p. } r\delta \\ 0, \text{ w.p. } r(1-\delta) \\ *, \text{ w.p. } (1-r) \end{cases}$$

where $*$ denotes unknown.

Definition 2.2 (Partial Observation Model)**.** *Let $\mathbf{A}$ be the adjacency matrix of a random graph generated according to the Stochastic Block Model of Definition 2.1. Let $0 < r \leq 1$. Each entry of*

the adjacency matrix $\mathbf{A}$ *is observed independently with probability* r*. Let* $\mathbf{A}^{obs}$ *denote the observed adjacency matrix. Then for* $l > m$*:* $(\mathbf{A}^{obs})_{l,m} \sim \Phi(r, p_i)$ *if both the nodes* l *and* m *belong to the same cluster* i*. Otherwise,* $(\mathbf{A}^{obs})_{l,m} \sim \Phi(r, q)$*.*

2.2 Results : Simple Convex Program

Let $[n] = \{1, 2, \cdots, n\}$. Let $\mathcal{R}$ be the union of regions induced by the clusters and $\mathcal{R}^c = [n] \times [n] - \mathcal{R}$ its complement. Note that $|\mathcal{R}| = \sum_{i=1}^{K} n_i^2$ and $|\mathcal{R}^c| = n^2 - \sum_{i=1}^{K} n_i^2$. Let $n_{\min} := \min_{1 \le i \le K} n_i$, $p_{min} := \min_{1 \le i \le K} p_i$ and $n_{\max} := \max_{1 \le i \le K} n_i$.

The following definitions are important to describe our results.

- Define $\mathbf{D}_i := n_i \, r \, (2p_i - 1)$ as the **effective density** of cluster i and $\mathbf{D}_{\min} = \min_{1 \le i \le K} \mathbf{D}_i$.
- $\gamma_{\text{succ}} := \max_{1 \le i \le K} 2r\sqrt{n_i}\sqrt{2(\frac{1}{r} - 1) + 4\left(q(1-q) + p_i(1-p_i)\right)}$ and $\gamma_{\text{fail}} := \sum_{i=1}^{K} \frac{n_i^2}{n}$
- $\mathbf{\Lambda}_{\text{succ}}^{-1} := 2r\sqrt{n}\sqrt{\frac{1}{r} - 1 + 4q(1-q)} + \gamma_{\text{succ}}$ and $\mathbf{\Lambda}_{\text{fail}}^{-1} := \sqrt{rq(n - \gamma_{\text{fail}})}$.

We note that the thresholds, $\mathbf{\Lambda}_{\text{succ}}$ and $\mathbf{\Lambda}_{\text{fail}}$ depend only the parameters of the model. Some simple algebra shows that $\mathbf{\Lambda}_{\text{succ}} < \mathbf{\Lambda}_{\text{fail}}$.

Theorem 1 (Simple Program)**.** *Consider a random graph generated according to the Partial Observation Model of Definition (2.2) with K disjoint clusters of sizes* $\{n_i\}_{i=1}^{K}$*, and probabilities* $\{p_i\}_{i=1}^{K}$ *and* q*, such that* $p_{min} > \frac{1}{2} > q > 0$*. Given* $\epsilon > 0$*, there exists positive constants* c_1', c_2' *such that,*

1. *If* $\lambda \ge (1 + \epsilon)\mathbf{\Lambda}_{fail}$*, then Program 1.1 fails to correctly recover the clusters with probability* $1 - c_1' \exp(-c_2'|\mathcal{R}^c|)$*.*
2. *If* $0 < \lambda \le (1 - \epsilon)\mathbf{\Lambda}_{succ}$*,*
   - *If* $\mathbf{D}_{\min} \ge (1 + \epsilon)\frac{1}{\lambda}$*, then Program 1.1 succeeds in correctly recovering the clusters with probability* $1 - c_1' n^2 \exp(-c_2' n_{\min})$*.*
   - *If* $\mathbf{D}_{\min} \le (1 - \epsilon)\frac{1}{\lambda}$*, then Program 1.1 fails to correctly recover the clusters with probability* $1 - c_1' \exp(-c_2' n_{\min})$*.*

Discussion:

1. Theorem 1 characterizes the success and failure of Program 1.1 as a function of the regularization parameter λ. In particular, if $\lambda > \mathbf{\Lambda}_{\text{fail}}$, Program 1.1 fails with high probability. If $\lambda < \mathbf{\Lambda}_{\text{succ}}$, Program 1.1 succeeds with high probability *if and only if* $\mathbf{D}_{\min} > \frac{1}{\lambda}$. However, Theorem 1 has nothing to say about $\mathbf{\Lambda}_{\text{succ}} < \lambda < \mathbf{\Lambda}_{\text{fail}}$.
2. **Small Cluster Regime:** When $n_{\max} = o(n)$, we have $\mathbf{\Lambda}_{\text{succ}}^{-1} = 2r\sqrt{n}\sqrt{\left(\frac{1}{r} - 1 + 4q(1-q)\right)}$. For simplicity let $p_i = p$, $\forall\, i$, which yields $\mathbf{D}_{\min} = n_{\min} r(2p-1)$. Then $\mathbf{D}_{\min} > \mathbf{\Lambda}_{\text{succ}}^{-1}$ implies,

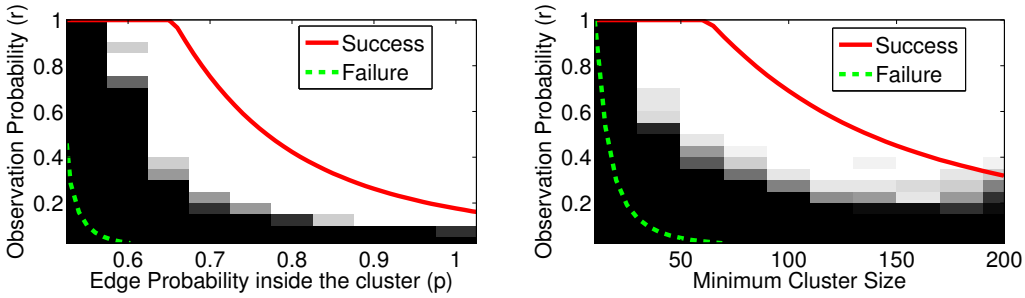
$$n_{\min} > \frac{2\sqrt{n}}{2p-1}\sqrt{\left(\frac{1}{r} - 1 + 4q(1-q)\right)}, \tag{2.2}$$

   giving a lower bound on the minimum cluster size that is sufficient for success.

2.3 Results: Improved Convex Program

The following definitions are critical to describe our results.

- Define $\tilde{\mathbf{D}}_i := n_i \, r \, (p_i - q)$ as the effective density of cluster i and $\tilde{\mathbf{D}}_{\min} = \min_{1 \le i \le K} \tilde{\mathbf{D}}_i$.
- $\tilde{\gamma}_{\text{succ}} := 2 \max_{1 \le i \le K} r\sqrt{n_i}\sqrt{(1-p_i)(\frac{1}{r} - 1 + p_i) + (1-q)(\frac{1}{r} - 1 + q)}$



(a) (b)

Figure 1: Region of success (white region) and failure (black region) of Program 1.1 with $\lambda = 1.01\mathbf{D}_{\min}^{-1}$. The solid red curve is the threshold for success ($\lambda < \Lambda_{\text{succ}}$) and the dashed green line which is the threshold for failure ($\lambda > \Lambda_{\text{fail}}$) as predicted by Theorem 1.

- $\tilde{\mathbf{\Lambda}}_{\text{succ}}^{-1} := 2r\sqrt{n}\sqrt{(\frac{1}{r} - 1 + q)(1-q)} + \tilde{\gamma}_{\text{succ}}$.

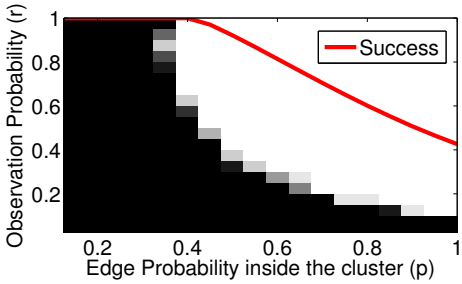
We note that the threshold, $\tilde{\mathbf{\Lambda}}_{\text{succ}}$ depends only on the parameters of the model.

Theorem 2 (Improved Program)**.** *Consider a random graph generated according to the Partial Observation Model of Definition 2.2, with K disjoint clusters of sizes $\{n_i\}_{i=1}^K$, and probabilities $\{p_i\}_{i=1}^K$ and q, such that $p_{min} > q > 0$. Given $\epsilon > 0$, there exists positive constants c_1', c_2' such that: If $0 < \lambda \leq (1-\epsilon)\tilde{\mathbf{\Lambda}}_{succ}$ and $\tilde{\mathbf{D}}_{\min} \geq (1+\epsilon)\frac{1}{\lambda}$, then Program 1.4 succeeds in recovering the clusters with probability $1 - c_1' n^2 \exp(-c_2' n_{\min})$.*

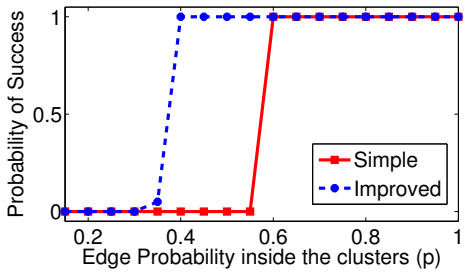
Discussion:[1]

1. Theorem 2 gives a sufficient condition for the success of Program 1.4 as a function of λ. In particular, for any $\lambda > 0$, we succeed if $\tilde{\mathbf{D}}_{\min}^{-1} < \lambda < \tilde{\mathbf{\Lambda}}_{\text{succ}}$.
2. **Small Cluster Regime:** When $n_{\max} = o(n)$, we have $\tilde{\mathbf{\Lambda}}_{\text{succ}}^{-1} = 2r\sqrt{n}\sqrt{\left(\frac{1}{r} - 1 + q\right)(1-q)}$. For simplicity let $p_i = p$, $\forall\, i$, which yields $\tilde{\mathbf{D}}_{\min} = n_{\min} r(p-q)$. Then $\tilde{\mathbf{D}}_{\min} > \tilde{\mathbf{\Lambda}}_{\text{succ}}^{-1}$ implies,
$$n_{\min} > \frac{2\sqrt{n}}{p-q}\sqrt{\left(\frac{1}{r} - 1 + q\right)(1-q)}, \tag{2.3}$$
which gives a lower bound on the minimum cluster size that is sufficient for success.
3. (p, q) **as a function of** n**:** We now briefly discuss the regime in which cluster sizes are large (i.e. $\mathcal{O}(n)$) and we are interested in the parameters (p, q) as a function of n that allows proposed approaches to be successful. Critical to Program 1.4 is the constraint (1.6): $\mathbf{L}_{i,j} = \mathbf{S}_{i,j}$ when $\mathbf{A}_{i,j}^{obs} = 0$ (which is the only constraint involving the adjacency $\mathbf{A}^{obs}$). With missing data, $\mathbf{A}_{i,j}^{obs} = 0$ with probability $r(1-p)$ inside the clusters and $r(1-q)$ outside the clusters. Defining $\hat{p} = rp + 1 - r$ and $\hat{q} = rq + 1 - r$, the number of constraints in (1.6) becomes statistically equivalent to those of a *fully observed* graph where p and q are replaced by $\hat{p}$ and $\hat{q}$. Consequently, for a fixed $r > 0$, from (2.3), we require $p \geq p - q \gtrsim \mathcal{O}(\frac{1}{\sqrt{n}})$ for success. However, setting the unobserved entries to 0, yields $\mathbf{A}_{i,j} = 0$ with probability $1 - rp$ inside the clusters and $1 - rq$ outside the clusters. This is equivalent to a fully observed graph where p and q are replaced by rp and rq. In this case, we can allow $p \approx \mathcal{O}(\frac{polylog(n)}{n})$ for success which is order-wise better, and matches closely to the results in McSherry [27]. Intuitively, clustering a fully observed graph with parameters $\hat{p} = rp + 1 - r$ and $\hat{q} = rq + 1 - r$ is much more difficult than one with rp and rq, since the links are *more noisy* in the former case. Hence, while it is beneficial to leave the unobserved entries blank in Program 1.1, for Program 1.4 it is in fact beneficial to set the unobserved entries to 0.

[1]The proofs for Theorems 1 and 2 are provided in the supplementary material.



(a) Region of success (white region) and failure (black region) of Program 1.4 with $\lambda = 0.49\tilde{\mathbf{\Lambda}}_{\text{succ}}$. The solid red curve is the threshold for success ($\tilde{\mathbf{D}}_{\min} > \lambda^{-1}$) as predicted by Theorem 2.



(b) Comparison range of edge probability p for Simple Program 1.1 and Improved Program 1.4.

Figure 2: Simulation results for Improved Program.

3 Experimental Results

We implement Program 1.1 and 1.4 using the inexact augmented Lagrange method of multipliers [28]. Note that this method solves the Program 1.1 and 1.4 approximately. Further, the numerical imprecisions will prevent the entries of the output of the algorithms from being strictly equal to 0 or 1. We use the mean of all the entries of the output as a hard threshold to round each entry. That is, if an entry is less than the threshold, it is rounded to 0 and to 1 otherwise. We compare the output of the algorithm after rounding to the optimal solution ($\mathbf{L}^0$), and declare success if the number of wrong entries is less than 0.1%.

Set Up: We consider at an unweighted graph on $n = 600$ nodes with 3 disjoint clusters. For simplicity the clusters are of equal size $n_1 = n_2 = n_3$, and the edge probability inside the clusters are same $p_1 = p_2 = p_3 = p$. The edge probability outside the clusters is fixed, $q = 0.1$. We generate the adjacency matrix randomly according to the Stochastic Block Model 2.1 and Partial Observation Model 2.2. All the results are an average over 20 experiments.

3.1 Simulations for Simple Convex Program

Dependence between r and p: In the first set of experiments we keep $n_1 = n_2 = n_3 = 200$, and vary p from 0.55 to 1 and r from 0.05 to 1 in steps of 0.05.

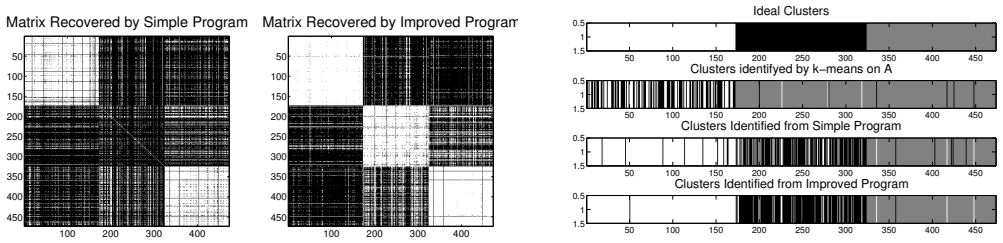
Dependence between $n_{\min}$ and r: In the second set of experiments we keep the edge probability inside the clusters fixed, $p = 0.85$. The cluster size is varied from $n_{\min} = 20$ to $n_{\min} = 200$ in steps of 20 and r is varied from 0.05 to 1 in steps of 0.05.

In both the experiments, we set the regularization parameter $\lambda = 1.01\mathbf{D}_{\min}^{-1}$, ensuring that $\mathbf{D}_{\min} > 1/\lambda$, enabling us to focus on observing the transition around $\mathbf{\Lambda}_{\text{succ}}$ and $\mathbf{\Lambda}_{\text{fail}}$. The outcome of the experiments are shown in the Figures 1a and 1b. The experimental region of success is shown in white and the region of failure is shown in black. The theoretical region of success is about the solid red curve ($\lambda < \mathbf{\Lambda}_{\text{succ}}$) and the region of failure is below dashed green curve ($\lambda > \mathbf{\Lambda}_{\text{fail}}$). As we can see the transition indeed occurs between the two thresholds $\mathbf{\Lambda}_{\text{succ}}$ and $\mathbf{\Lambda}_{\text{fail}}$.

3.2 Simulations for Improved Convex Program

We keep the cluster size, $n_1 = n_2 = n_3 = 200$ and vary p from 0.15 to 1 and r from 0.05 to 1 in steps of 0.05. We set the regularization parameter, $\lambda = 0.49\tilde{\mathbf{\Lambda}}_{\text{succ}}$, ensuring that $\lambda < \tilde{\mathbf{\Lambda}}_{\text{succ}}$, enabling us to focus on observing the condition of success around $\tilde{\mathbf{D}}_{\min}$. The outcome of this experiment is shown in the Figure 2a. The experimental region of success is shown in white and region of failure is shown in black. The theoretical region of success is above solid red curve.

Comparison with the Simple Convex Program: In this experiment, we are interested in observing the range of p for which the Programs 1.1 and 1.4 work. Keeping the cluster size $n_1 = n_2 = n_3 =$



(a) (b) (c) Comparing with k-means clustering.

Figure 3: Result of using (a) Program 1.1 (Simple) and (b) Program 1.4 (Improved) on the real data set. (c) Comparing the clustering output after running Program 1.1 and Program 1.4 with the output of applying k-means clustering directly on A (with unknown entries set to 0).

200 and $r = 1$, we vary the edge probability inside the clusters from $p = 0.15$ to $p = 1$ in steps of 0.05. For each instance of the adjacency matrix, we run both Program 1.1 and 1.4. We plot the probability of success of both the algorithms in Figure 2b. As we can observe, Program 1.1 starts succeeding only after $p > 1/2$, whereas for Program 1.4 it starts at $p \approx 0.35$.

3.3 Labeling Images: Amazon MTurk Experiment

Creating a training dataset by labeling images is a tedious task. It would be useful to crowdsource this task instead. Consider a specific example of a set of images of dogs of different breeds. We want to cluster them such that the images of dogs of the same breed are in the same cluster. One could show a set of images to each worker, and ask him/her to identify the breed of dog in each of those images. But such a task would require the workers to be experts in identifying the dog breeds. A relatively reasonable task is to ask the workers to compare pairs of images, and for each pair, answer whether they think the dogs in the images are of the same breed or not. If we have n images, then there are $\binom{n}{2}$ distinct pairs of images, and it will pretty quickly become unreasonable to compare all possible pairs. This is an example where we could obtain a subset of the data and try to cluster the images based on the partial observations.

Image Data Set: We used images of 3 different breeds of dogs : Norfolk Terrier (172 images), Toy Poodle (151 images) and Bouvier des Flandres (150 images) from the Standford Dogs Dataset [29]. We uploaded all the 473 images of dogs on an image hosting server (we used imgur.com).

MTurk Task: We used Amazon Mechanical Turk [30] as the platform for crowdsourcing. For each worker, we showed 30 pairs of images chosen randomly from the $\binom{n}{2}$ possible pairs. The task assigned to the worker was to compare each pair of images, and answer whether they think the dogs belong to the same breed or not. If the worker's response is a "yes", then there we fill the entry of the adjacency matrix corresponding to the pair as 1, and 0 if the answer is a "no".

Collected Data: We recorded around 608 responses. We were able to fill $16,750$ out of $111,628$ entries in $\mathbf{A}$. That is, we observed 15% of the total number of entries. Compared with true answers (which we know a priori), the answers given by the workers had around 23.53% errors (3941 out of 16750). The empirical parameters for the partially observed graph thus obtained is shown Table 1.

We ran Program 1.1 and Program 1.4 with regularization parameter, $\lambda = 1/\sqrt{n}$. Further, for Program 1.4, we set the size of the cluster region, $\mathcal{R}$ to 0.125 times $\binom{n}{2}$. Figure 3a shows the recovered matrices. Entries with value 1 are depicted by white and 0 is depicted by black. In Figure 3c we compare the clusters output by running the k-means algorithm directly on the adjacency matrix $\mathbf{A}$ (with unknown entries set to 0) to that obtained by running k-means algorithm on the matrices recovered after running Program 1.1 (Simple Program) and Program 1.4 (Improved Program) respectively. The overall error with k-means was 40.8% whereas the error significantly reduced to 15.86% and 7.19% respectively when we used the matrices recoverd from Programs 1.1 and 1.4 respectively (see Table 2). Further, note that for running the k-means algorithm we need to know the exact number of clusters. A common heuristic is to identify the top K eigenvalues that are much

Table 1: Empirical Parameters from the real data.

Params	Value	Params	Value
n	473	r	0.1500
K	3	q	0.1929
n_1	172	p_1	0.7587
n_2	151	p_2	0.6444
n_3	150	p_3	0.7687

Table 2: Number of miss-classified images

Clusters→	1	2	3	Total
K-means	39	150	4	193
Simple	9	57	8	74
Improved	1	29	4	34

larger than the rest. In Figure 4 we plot the sorted eigenvalues for the adjacency matrix $\mathbf{A}$ and the recovered matrices. We can see that the top 3 eigen values are very easily distinguished from the rest for the matrix recovered after running Program 1.4.

A sample of the data is shown in Figure 5. We observe that factors such as color, grooming, posture, face visibility etc. can result in confusion while comparing image pairs. Also, note that the ability of the workers to distinguish the dog breeds is neither guaranteed nor uniform. Thus, the edge probability inside and outside clusters are not uniform. Nonetheless, Programs 1.1 and Program 1.4, especially Program 1.4, are quite successful in clustering the data with only 15% observations.

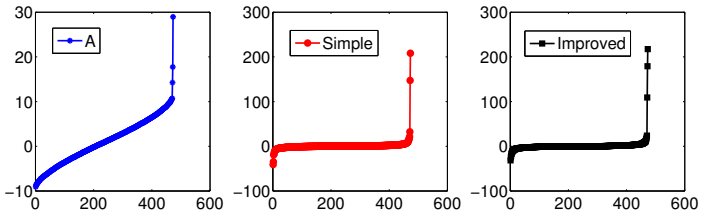


Figure 4: Plot of sorted eigen values for (1) Adjacency matrix with unknown entries filled by 0, (2) Recovered adjacency matrix from Program 1.1, (3) Recovered adjacency matrix from Program 1.4

Figure 5: Sample images of three breeds of dogs that were used in the MTurk experiment.

The authors thank the anonymous reviewers for their insightful comments. This work was supported in part by the National Science Foundation under grants CCF-0729203, CNS-0932428 and CIF-1018927, by the Office of Naval Research under the MURI grant N00014-08-1-0747, and by a grant from Qualcomm Inc. The first author is also supported by the Schlumberger Foundation Faculty for the Future Program Grant.

References

[1] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, September 1999.

[2] M. Ester, H.-P. Kriegel, and X. Xu. A database interface for clustering in large spatial databases. In *Proceedings of the 1st international conference on Knowledge Discovery and Data mining (KDD'95)*, pages 94–99. AAAI Press, August 1995.

[3] Xiaowei Xu, Jochen Jäger, and Hans-Peter Kriegel. A fast parallel clustering algorithm for large spatial databases. *Data Min. Knowl. Discov.*, 3(3):263–290, September 1999.

[4] Nina Mishra, Robert Schreiber, Isabelle Stanton, and Robert Tarjan. Clustering Social Networks. In Anthony Bonato and Fan R. K. Chung, editors, *Algorithms and Models for the Web-Graph*, volume 4863 of *Lecture Notes in Computer Science*, chapter 5, pages 56–67. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[5] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 57–66, New York, NY, USA, 2001. ACM.

[6] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.

[7] Ying Xu, Victor Olman, and Dong Xu. Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics*, 18(4):536–545, 2002.

[8] Qiaofeng Yang and Stefano Lonardi. A parallel algorithm for clustering protein-protein interaction networks. In *CSB Workshops*, pages 174–177. IEEE Computer Society, 2005.

[9] Satu Elisa Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27 – 64, 2007.

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

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

[12] Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. In John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta, editors, *NIPS*, pages 2496–2504. Curran Associates, Inc., 2010.

[13] Ali Jalali, Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering partially observed graphs via convex optimization. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 1001–1008, New York, NY, USA, June 2011. ACM.

[14] Brendan P. W. Ames and Stephen A. Vavasis. Convex optimization for the planted k-disjoint-clique problem. *Math. Program.*, 143(1-2):299–337, 2014.

[15] Brendan P. W. Ames and Stephen A. Vavasis. Nuclear norm minimization for the planted clique and biclique problems. *Math. Program.*, 129(1):69–89, September 2011.

[16] S. Oymak and B. Hassibi. Finding Dense Clusters via "Low Rank + Sparse" Decomposition. *arXiv:1104.5186*, April 2011.

[17] Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering sparse graphs. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Lon Bottou, and Kilian Q. Weinberger, editors, *NIPS*, pages 2213–2221, 2012.

[18] Yudong Chen, Ali Jalali, Sujay Sanghavi, and Constantine Caramanis. Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory*, 59(7):4324–4337, 2013.

[19] Brendan P. W. Ames. Robust convex relaxation for the planted clique and densest k-subgraph problems. 2013.

[20] Nir Ailon, Yudong Chen, and Huan Xu. Breaking the small cluster barrier of graph clustering. *CoRR*, abs/1302.4549, 2013.

[21] Ramya Korlakai Vinayak, Samet Oymak, and Babak Hassibi. Sharp performance bounds for graph clustering via convex optimizations. In *Proceedings of the 39th International Conference on Acoustics, Speech and Signal Processing*, ICASSP '14, 2014.

[22] Emmanuel J. Candes and Justin Romberg. Quantitative robust uncertainty principles and optimally sparse decompositions. *Found. Comput. Math.*, 6(2):227–254, April 2006.

[23] Emmanuel J. Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, December 2009.

[24] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011.

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

[26] Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky. Rejoinder: Latent variable graphical model selection via convex optimization. *CoRR*, abs/1211.0835, 2012.

[27] Frank McSherry. Spectral partitioning of random graphs. In *FOCS*, pages 529–537. IEEE Computer Society, 2001.

[28] Zhouchen Lin, Minming Chen, and Yi Ma. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. *Mathematical Programming*, 2010.

[29] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

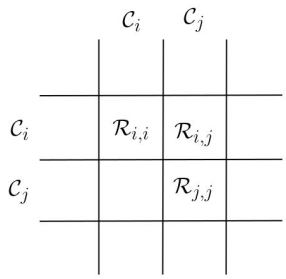


Figure 6: Illustration of $\{\mathcal{R}_{i,j}\}$ dividing $[n] \times [n]$ into disjoint regions similar to a grid

[30] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, January 2011.

[31] Van H. Vu. Spectral norm of random matrices. In Harold N. Gabow and Ronald Fagin, editors, *STOC*, pages 423–430. ACM, 2005.

4 Proof of Results for Simple Convex Program

Let $1 \geq p_{min} > \frac{1}{2} > q > 0$ and $0 \leq r \leq 1$. $\mathcal{G}$ be a random graph generated according to the stochastic block model 2.1 with cluster sizes $\{n_i\}_{i=1}^K$. Let the observation model be as defined in (Defn 2.2). Theorem 1 is based on the following lemmas:

Lemma 4.1. *If $\lambda > \Lambda_{fail}$, then $\left(\mathbf{L}^0, \mathbf{S}^0\right)$ is not an optimal solution to the Program 1.1 with high probability.*

Lemma 4.2. *If $\lambda < \Lambda_{succ}$ and $\mathbf{D}_{\min} > \frac{1}{\lambda}$, then $\left(\mathbf{L}^0, \mathbf{S}^0\right)$ is the unique optimal solution to Program 1.1 with high probability.*

Before we proceed, we need some additional notations. Let $\mathcal{R}_{i,j} = \mathcal{C}_i \times \mathcal{C}_j$ for $1 \leq i, j \leq K+1$. One can see that $\{\mathcal{R}_{i,j}\}$ divides $[n] \times [n]$ into $(K+1)^2$ disjoint regions similar to a grid which is illustrated in the Figure 6. Thus, $\mathcal{R}_{i,i}$ is the region induced by i'th cluster for any $1 \leq i \leq K$.

Let Γ^{out} be the set of entries of adjacency matrix that are *not* observed. Let $\mathcal{A} \subseteq [n] \times [n]$ be the set of observed coordinates of $\mathbf{A_{obs}}$. Let $\mathcal{A}_1 \subseteq [n] \times [n]$ be the set of nonzero coordinates of $\mathbf{A_{obs}}$, and $\mathcal{A}_0 \subseteq [n] \times [n]$ be the set of coordinates of $\mathbf{A_{obs}}$ that are zero. Then the sets,

1. $\mathcal{A}_1 \cap \mathcal{R}$ corresponds to the edges inside the clusters that are observed.
2. $\mathcal{A}_1 \cap \mathcal{R}^c$ corresponds to the set of edges outside the clusters that are observed.
3. $\mathcal{A}_0 \cap \mathcal{R}$ corresponds to the missing edges inside the clusters, that are observed (that is, we know that the edge does not exist).

Let c and d be positive integers. Consider a matrix, $\mathbf{X} \in \mathbb{R}^{c \times d}$. Let β be a subset of $[c] \times [d]$. Then, let $\mathbf{X}_\beta$ denote the matrix induced by the entries of $\mathbf{X}$ on β i.e.,

$$(\mathbf{X}_\beta)_{i,j} = \begin{cases} \mathbf{X}_{i,j} & \text{if } (i,j) \in \beta \\ 0 & \text{otherwise .} \end{cases}$$

In other words, $\mathbf{X}_\beta$ is a matrix whose entries match those of $\mathbf{X}$ in the positions $(i,j) \in \beta$ and zero otherwise. For example, $\mathbb{1}_{\mathcal{A}^{obs}}^{n \times n} = \mathbf{A}^{obs}$. Given a matrix $\mathbf{X}$, sum$(\mathbf{X})$ will denote the sum of all entries of $\mathbf{X}$. Finally, we introduce the following parameter which will be useful for the subsequent analysis. Given $q, \{p_i\}_{i=1}^K$, let,

$$\begin{aligned} \mathbf{D}_{\mathcal{A}} &= \frac{1}{2} \min \left\{ r(1-2q), \left\{ r(2p_i - 1) - \frac{1}{\lambda n_i} \right\}_{i=1}^K \right\} \\ &= \frac{1}{2} \min \left\{ r(1-2q), \frac{\mathbf{D}_i - \lambda^{-1}}{n_i} \right\} \end{aligned} \tag{4.1}$$

For our proofs, we will make use of the following Big O notation. $f(n) = \Omega(n)$ will mean there exists a positive constant c such that for sufficiently large n, $f(n) \geq cn$. $f(n) = O(n)$ will mean there exists a positive constant c such that for sufficiently large n, $f(n) \leq cn$.

4.1 Proof of Lemma 4.1

Lagrange for the problem (1.1) can be written as follows

$$\mathscr{L}(\mathbf{L}, \mathbf{S}; \mathbf{M}, \mathbf{N}) \quad = \quad \|\mathbf{L}\|_\star + \lambda \|\mathbf{S_{obs}}\|_1 + \text{trace}(\mathbf{M}(\mathbf{L} - \mathbb{1}\mathbb{1}^T)) - \text{trace}(\mathbf{NL}). \tag{4.2}$$

where $\mathbf{M}$ and $\mathbf{N}$ are dual variables corresponding to the inequality constraints (1.2).

For $\mathbf{L}^0$ to be an optimal solution to (1.1), it has to satisfy the KKT conditions. Therefore, the subgradient of (4.2) at $\mathbf{L}^0$ has to be 0, i.e.,

$$\partial \|\mathbf{L}^0\|_\star + \lambda \, \partial \|\mathbf{A_{obs}} - \mathbf{L_{obs}}^0\|_1 + \mathbf{M}^0 - \mathbf{N}^0 = 0. \tag{4.3}$$

where $\mathbf{M}^0$ and $\mathbf{N}^0$ are optimal dual variables, and $\partial \|\mathbf{L}^0\|_\star$ and $\partial \|\mathbf{S}^0\|_1$ are subgradients of nuclear norm and ℓ_1-norm respectively at the points $\left(\mathbf{L}^0, \mathbf{S}^0\right)$. Note that in the standard notation, $\partial \|\mathbf{x}\|_*$ denotes the set of all subgradients, i.e., the subdifferential. We have slightly abused the notation by denoting a subgradient of a norm $\|\cdot\|_*$ at the point $\mathbf{x}$ by $\partial \|\mathbf{x}\|_*$.

Also, by complementary slackness,

$$\text{trace}(\mathbf{M}^0(\mathbf{L}^0 - \mathbb{1}\mathbb{1}^T)) = 0, \tag{4.4}$$

and

$$\text{trace}(\mathbf{N}^0\mathbf{L}^0) = 0. \tag{4.5}$$

From (5.1) and (4.4), (4.5), we have $(\mathbf{M}^0)_{\mathcal{R}} \geq 0$, $(\mathbf{M}^0)_{\mathcal{R}^c} = 0$, $(\mathbf{N}^0)_{\mathcal{R}} = 0$ and $(\mathbf{N}^0)_{\mathcal{R}^c} \geq 0$. Hence $(\mathbf{M}^0 - \mathbf{N}^0)_{\mathcal{R}} \geq 0$ and $(\mathbf{M}^0 - \mathbf{N}^0)_{\mathcal{R}^c} \leq 0$.

Let $\mathbf{L}^0 = \mathbf{U}\Lambda\mathbf{U}^T$, where $\Lambda = \text{diag}\{n_1, n_2, \ldots, n_K\}\mathbf{U} = [\mathbf{u}_1 \ \ldots \ \mathbf{u}_K] \in \mathbb{R}^{n \times K}$,

$$\mathbf{u}_{l,i} = \begin{cases} \frac{1}{\sqrt{n_l}} & \text{if } i \in \mathcal{C}_l \\ 0 & \text{else.} \end{cases} \tag{4.6}$$

Then the subgradient $\partial \|\mathbf{L}^0\|_\star$ is of the form $\mathbf{U}\mathbf{U}^T + \mathbf{W}$ such that $\mathbf{W} \in \{\mathbf{X} : \mathbf{X}\mathbf{U} = \mathbf{U}^T\mathbf{X} = 0, \|\mathbf{X}\| \leq 1\}$. The subgradient $\partial \|\mathbf{S}^0\|_1$ is of the form $\text{sign}(\mathbf{S}^0) + \mathbf{Q}$ where $\mathbf{Q}_{i,j} = 0$ if $\mathbf{S}_{i,j} \neq 0$ and $\|\mathbf{Q}\|_\infty \leq 1$. Further, note that the subgradient of $\mathbf{S}^0$ over unobserved entries is zero. That is, $\partial \|\mathbf{S}^0_{unobs}\|_1 = 0$ since $\mathbf{S}^0_{unobs} = 0$.

From (4.3), we have

$$\mathbf{U}\mathbf{U}^T + \mathbf{W} - \lambda \left(\text{sign}(\mathbf{S}^0) + \mathbf{Q}\right) + (\mathbf{M}^0 - \mathbf{N}^0) = 0. \tag{4.7}$$

Consider the sum of the entires corresponding to the cluster i ($\mathcal{R}_{i,i}$), i.e.,

$$\begin{aligned} &\underbrace{\text{sum}\left(\mathbf{L}^0\right)_{\mathcal{R}_{i,i}}}_{n_i} - \text{sum}\left(\lambda \left(\text{sign}(\mathbf{S}^0) + \mathbf{Q}\right)_{\mathcal{R}_{i,i}}\right) \\ &+ \underbrace{\text{sum}\left(\mathbf{M}^0 - \mathbf{N}^0\right)_{\mathcal{R}_{i,i}}}_{\geq 0} = 0 \end{aligned} \tag{4.8}$$

Since each entry of the adjacency matrix is observed with probability r, and the probability of missing edge inside cluster i is $1 - p_i$, we note that $(\mathbf{S}^0_{\mathcal{R}_{i,i}})_{l,m} \neq 0$ with probability $r(1 - p_i)$. Recall that $\mathbf{Q}_{l,m} = 0$ if $\mathbf{S}^0_{l,m} \neq 0$.

Then by Bernstein's inequality and using $\|\mathbf{Q}\|_\infty \leq 1$, with probability $1 - \exp\left(-\Omega(n_i^2)\right)$ we have $\text{sum}\left(\text{sign}(\mathbf{S}^0)\right) = -n_i^2 r(1 - p_i)$ and $\text{sum}\left(\mathbf{Q}\right) \leq n_i^2 r p_i$.

Thus,

$$\begin{aligned} -\text{sum}\left(\lambda \left(\text{sign}(\mathbf{S}^0) + \mathbf{Q}\right)_{\mathcal{R}_{i,i}}\right) \quad &\geq \quad \lambda n_i^2 r(1 - p_i) - \lambda n_i^2 r p_i \\ &= \quad \lambda n_i^2 r \left(1 - 2p_i\right). \end{aligned} \tag{4.9}$$

and hence LHS of equation (4.8) can be lower bounded as ,

$$n_i - \text{sum}\left(\lambda\left(\text{sign}(\mathbf{S}^0) + \mathbf{Q}\right)_{\mathcal{R}_{i,i}}\right) + \underbrace{\text{sum}\left(\mathbf{M}^0 - \mathbf{N}^0\right)_{\mathcal{R}_{i,i}}}_{\geq 0} \geq n_i + \lambda n_i^2 r\left(1 - 2p_i\right). \tag{4.10}$$

We see that $n_i r\left(2p_i - 1\right) < \frac{1}{\lambda}$ would imply $n_i + \lambda n_i^2 r\left(1 - 2p_i\right) > 0$, in which case, the equation (4.3) does not hold. Hence $\mathbf{L}^0$ cannot be an optimal solution to the Program 1.1. (Note that, $p_i > \frac{1}{2}$ and hence $2p_i - 1 > 0$.)

Notice that $\left(\mathbf{U}\mathbf{U}^T\right)_{\mathcal{R}^c} = 0$ and the entries of $-\left(\text{sign}(\mathbf{S}^0) + \mathbf{Q}\right)$ and $\mathbf{M}^0 - \mathbf{N}^0$ over $\mathcal{R}^c \cap \mathcal{A}_1$ are negative. Hence from the equation (4.7),

$$\|\mathbf{W}\|_F^2 \quad \geq \quad \|\left(\mathbf{U}\mathbf{U}^T + \mathbf{W}\right)_{(\mathcal{R}^c \cap \mathcal{A}_1)}\|_F^2 \geq \|\lambda\left(\text{sign}(\mathbf{S}^0) + \mathbf{Q}\right)_{(\mathcal{R}^c \cap \mathcal{A}_1)}\|_F^2. \tag{4.11}$$

Recall that $\mathbf{S}^0_{(\mathcal{R}_c \cap \mathcal{A}_1)} \neq 0$ and hence $\mathbf{Q}_{(\mathcal{R}_c \cap \mathcal{A}_1)} = 0$. Further, recall that by the Stochastic Block Model, each entry of $\mathbf{A}$ over $\mathcal{R}^c$ is non-zero with probability q and by observation model (Defn 2.2), each entry of $\mathbf{A}$ is observed with probability r. Hence with probability at least $1 - \exp\left(-\Omega(|\mathcal{R}_c|)\right)$, $|\mathcal{R}_c \cap \mathcal{A}_1| = rq(n^2 - \sum_{i=1}^K n_i^2)$. Thus from equation (4.11) we have,

$$\|\mathbf{W}\|_F^2 \geq \lambda^2 rq(n^2 - \sum_{i=1}^K n_i^2), \tag{4.12}$$

Recall that $\|\mathbf{W}\| \leq 1$ should hold true for $\left(\mathbf{L}^0, \mathbf{S}^0\right)$ to be an optimal solution to Program 1.1. $\|\mathbf{W}\| = |\sigma_{\max}(\mathbf{W})| \geq \frac{\|\mathbf{W}\|_F}{\sqrt{n}}$, which on combining with equation (4.12) gives us,

$$\|\mathbf{W}\| \geq \lambda\sqrt{\frac{rq\left(n^2 - \sum_{i=1}^K n_i^2\right)}{n}}.$$

So, if $\lambda\sqrt{rq\left(n^2 - \sum_{i=1}^K n_i^2\right)/n} > 1$ then, $\left(\mathbf{L}^0, \mathbf{S}^0\right)$ cannot be an optimal solution to Program 1.1. This gives us the result in Lemma 4.1.

4.2 Proof of Lemma 4.2

In order to show that $(\mathbf{L}^0, \mathbf{S}^0)$ is the unique optimal solution to the Program 1.1, we need to prove that for all feasible perturbations $(\mathbf{E}^L, \mathbf{E}^S)$,

$$(\|\mathbf{L}^0 + \mathbf{E}^L\|_\star + \lambda\|\mathbf{S}^0 + \mathbf{E}^S\|_1) - (\|\mathbf{L}^0\|_\star + \lambda\|\mathbf{S}^0\|_1) > 0. \tag{4.13}$$

We note that $\mathbf{S}$ can be split as $\mathbf{S} = \mathbf{S_{obs}} + \mathbf{S_{rest}}$, where $\mathbf{S_{rest}}$ denotes the entries of $\mathbf{S}$ other than those corresponding to the observed entries of $\mathbf{A}$. Furthermore, we claim that at the optimal, $\mathbf{S_{rest}} = 0$, since if otherwise, the objective can be strictly decreased by setting $\mathbf{S_{rest}} = 0$. Hence, $\mathbf{S} = \mathbf{S_{obs}}$.

We can lower bound the LHS of the equation (4.13) using the subgradients as follows,

$$(\|\mathbf{L}^0 + \mathbf{E}^L\|_\star + \lambda\|\mathbf{S}^0 + \mathbf{E}^S\|_1) - \left(\|\mathbf{L}^0\|_\star + \lambda\|\mathbf{S}^0\|_1\right) \geq \langle \partial\|\mathbf{L}^0\|_\star, \mathbf{E}^L\rangle + \lambda\langle \partial\|\mathbf{S}^0\|_1, \mathbf{E}^S\rangle, \tag{4.14}$$

where $\partial\|\mathbf{L}^0\|_\star$ and $\partial\|\mathbf{S}^0\|_1$ are subgradients of nuclear norm and ℓ_1-norm respectively at the points $\left(\mathbf{L}^0, \mathbf{S}^0\right)$. Note that in the standard notation, $\partial\|\mathbf{x}\|_*$ denotes the set of all subgradients, i.e., the subdifferential. We have slightly abused the notation by denoting a subgradient of a norm $\|\cdot\|_*$ at the point $\mathbf{x}$ by $\partial\|\mathbf{x}\|_*$.

To make use of (4.14), it is very important to choose good subgradients. In the following section we will focus on construction of such subgradients.

4.2.1 Subgradient construction

Recall that, $\mathbf{L}^0 = \mathbf{U}\Lambda\mathbf{U}^T$, where $\Lambda = \text{diag}\{n_1, n_2, \ldots, n_K\}$ and $\mathbf{U} = [\mathbf{u}_1 \ \ldots \ \mathbf{u}_K] \in \mathbb{R}^{n\times K}$, with $\mathbf{u}_l$ as defined before. Then the subgradient $\partial\|\mathbf{L}^0\|_\star$ is of the form $\mathbf{U}\mathbf{U}^T + \mathbf{W}$ such that $\mathbf{W} \in \mathcal{M}_U := \{\mathbf{X} : \mathbf{X}\mathbf{U} = \mathbf{U}^T\mathbf{X} = 0, \|\mathbf{X}\| \leq 1\}$. $\|.\|$ is spectral norm (maximum singular value). The subgradient $\partial\|\mathbf{S}^0\|_1$ is of the form $\text{sign}(\mathbf{S}^0) + \mathbf{Q}$ where $\mathbf{Q}_{i,j} = 0$ if $\mathbf{S}^0_{i,j} \neq 0$ and $\|\mathbf{Q}\|_\infty \leq 1$.

$$\begin{aligned}\|\mathbf{L}^0 + \mathbf{E}^L\|_\star + \lambda\|\mathbf{S}^0 + \mathbf{E}^S\|_1 - (\|\mathbf{L}^0\|_\star + \lambda\|\mathbf{S}^0\|_1) &\geq \langle\partial\|\mathbf{L}^0\|_\star, \mathbf{E}^L\rangle + \lambda\langle\partial\|\mathbf{S}^0\|_1, \mathbf{E}^S\rangle \\ &= \langle\mathbf{U}\mathbf{U}^T + \mathbf{W}, \mathbf{E}^L\rangle + \lambda\langle\text{sign}(\mathbf{S}^0) + \mathbf{Q}, \mathbf{E}^S\rangle\end{aligned}$$

Note that, due to the condition $\mathbf{L_{obs}} + \mathbf{S_{obs}} = \mathbf{A_{obs}}$, we have $\mathbf{E}^S = {\mathbf{E_{obs}}}^L$. Further, note that $\text{sign}(\mathbf{S}^0) = \mathbb{1}^{n\times n}_{\mathcal{A}_1\cap\mathcal{R}^c} - \mathbb{1}^{n\times n}_{\mathcal{A}_0\cap\mathcal{R}}$. Choosing $\mathbf{Q} = \mathbb{1}^{n\times n}_{\mathcal{A}_1\cap\mathcal{R}} - \mathbb{1}^{n\times n}_{\mathcal{A}_0\cap\mathcal{R}^c}$, we get,

$$\begin{aligned}\|\mathbf{L}^0 + \mathbf{E}^L\|_\star + \lambda\|\mathbf{S}^0 + \mathbf{E}^S\|_1 - (\|\mathbf{L}^0\|_\star + \lambda\|\mathbf{S}^0\|_1) \geq \big\langle\mathbf{W}, \mathbf{E}^L\big\rangle \\ + \underbrace{\sum_{i=1}^{K}\frac{1}{n_i}\text{sum}(\mathbf{E}_{R_{i,i}}) + \lambda\left(\text{sum}(\mathbf{E}^L_{\mathcal{A}_0}) - \text{sum}(\mathbf{E}^L_{\mathcal{A}_1})\right)}_{:=g(\mathbf{E}^L)}\end{aligned} \tag{4.15}$$

From this point onward, for simplicity we will ignore the superscript L on $\mathbf{E}^L$ and just use $\mathbf{E}$.

Define,

$$g(\mathbf{E}) := \sum_{i=1}^{K}\frac{1}{n_i}\text{sum}(\mathbf{E}_{\mathcal{R}_{i,i}}) + \lambda\left(\text{sum}(\mathbf{E}_{\mathcal{A}_0}) - \text{sum}(\mathbf{E}_{\mathcal{A}_1})\right). \tag{4.16}$$

Also, define $f(\mathbf{E}, \mathbf{W}) := g(\mathbf{E}) + \langle\mathbf{W}, \mathbf{E}\rangle$. Our aim is to show that for all feasible perturbations $\mathbf{E}$, there exists $\mathbf{W}$ such that,

$$f(\mathbf{E}, \mathbf{W}) = g(\mathbf{E}) + \langle\mathbf{W}, \mathbf{E}\rangle > 0. \tag{4.17}$$

Note that $g(\mathbf{E})$ does not depend on $\mathbf{W}$.

Lemma 4.3. *Given* $\mathbf{E}$*, assume there exists* $\mathbf{W} \in \mathcal{M}_\mathbf{U}$ *with* $\|\mathbf{W}\| < 1$ *such that* $f(\mathbf{E}, \mathbf{W}) \geq 0$*. Then at least one of the followings holds:*

- *There exists* $\mathbf{W}^* \in \mathcal{M}_\mathbf{U}$ *with* $\|\mathbf{W}^*\| \leq 1$ *and* $f(\mathbf{E}, \mathbf{W}^*) > 0$*.*
- *For all* $\mathbf{W} \in \mathcal{M}_\mathbf{U}$*,* $\langle\mathbf{E}, \mathbf{W}\rangle = 0$*.*

Proof. Let $c = 1 - \|\mathbf{W}\|$. Assume $\langle\mathbf{E}, \mathbf{W}'\rangle \neq 0$ for some $\mathbf{W}' \in \mathcal{M}_\mathbf{U}$. If $\langle\mathbf{E}, \mathbf{W}'\rangle > 0$, choose $\mathbf{W}^* = \mathbf{W} + c\mathbf{W}'$. Otherwise, choose $\mathbf{W}^* = \mathbf{W} - c\mathbf{W}'$. Since $\|\mathbf{W}'\| \leq 1$, we have, $\|\mathbf{W}^*\| \leq 1$ and $\mathbf{W}^* \in \mathcal{M}_\mathbf{U}$. Consequently,

$$f(\mathbf{E}, \mathbf{W}^*) = f(\mathbf{E}, \mathbf{W}) + \langle\mathbf{E}, c\mathbf{W}'\rangle > f(\mathbf{E}, \mathbf{W}) \geq 0 \tag{4.18}$$

■

Notice that, for all $\mathbf{W} \in \mathcal{M}_\mathbf{U}$, $\langle\mathbf{E}, \mathbf{W}\rangle = 0$ is equivalent to $\mathbf{E} \in \mathcal{M}^\perp_\mathbf{U}$ which is the orthogonal complement of $\mathcal{M}_\mathbf{U}$ in $\mathbb{R}^{n\times n}$. $\mathcal{M}^\perp_\mathbf{U}$ has the following characterization:

$$\mathcal{M}^\perp_\mathbf{U} = \{\mathbf{X} \in \mathbb{R}^{n\times n} : \mathbf{X} = \mathbf{U}\mathbf{M}^T + \mathbf{N}\mathbf{U}^T \ \text{ for some } \mathbf{M}, \mathbf{N} \in \mathbb{R}^{n\times K}\}. \tag{4.19}$$

Now we have broken down our aim into two steps.

1. Construct $\mathbf{W} \in \mathcal{M}_\mathbf{U}$ with $\|\mathbf{W}\| < 1$, such that $f(\mathbf{E}, \mathbf{W}) \geq 0$ for all feasible perturbations $\mathbf{E}$.
2. For all non-zero feasible $\mathbf{E} \in \mathcal{M}^\perp_\mathbf{U}$, show that $g(\mathbf{E}) > 0$.

As a first step, in Section 4.3, we will argue that, under certain conditions, there exists a $\mathbf{W} \in \mathcal{M}_\mathbf{U}$ with $\|\mathbf{W}\| < 1$ such that with high probability, $f(\mathbf{E}, \mathbf{W}) \geq 0$ for all feasible $\mathbf{E}$. This $\mathbf{W}$ is called the dual certificate. Secondly, in Section 4.4, we will show that, under certain conditions, for all $\mathbf{E} \in \mathcal{M}^\perp_\mathbf{U}$ with high probability, $g(\mathbf{E}) > 0$. Finally, combining these two arguments, and using Lemma 4.3 we will conclude that $(\mathbf{L}^0, \mathbf{S}^0)$ is the unique optimal with high probability.

4.3 Showing existence of the dual certificate

Recall that

$$f(\mathbf{E}, \mathbf{W}) = \sum_{i=1}^{K} \frac{1}{n_i} \text{sum}(\mathbf{E}_{\mathcal{R}_{i,i}}) + \langle \mathbf{E}, \mathbf{W} \rangle + \lambda \left(\text{sum}\left(\mathbf{E}_{\mathcal{A}_0}\right) - \text{sum}\left(\mathbf{E}_{\mathcal{A}_1}\right) \right)$$

$\mathbf{W}$ will be constructed from the candidate $\mathbf{W}_0$, which is given as follows.

4.3.1 Candidate $\mathbf{W}_0$

Based on Program 1.1, we propose the following,

$$\mathbf{W}_0 = \sum_{i=1}^{K} c_i \mathbb{1}^{n\times n}_{\mathcal{R}_{i,i}} + c\mathbb{1}^{n\times n}_{\mathcal{R}^c} + \lambda \left(\mathbb{1}^{n\times n}_{\mathcal{A}_1} - \mathbb{1}^{n\times n}_{\mathcal{A}_0} \right), \tag{4.20}$$

where $\{c_i\}_{i=1}^{K}, c$ are real numbers to be determined.

$$f(\mathbf{E}, \mathbf{W}^0) = \sum_{i=1}^{K} (\frac{1}{n_i} + c_i) \text{ sum}(\mathbf{E}_{\mathcal{R}_{i,i}}) + c \text{ sum}(\mathbf{E}_{\mathcal{R}^c})$$

Note that $\mathbf{W}_0$ is a random matrix where randomness is due to $\mathbf{A_{obs}}$. In order to ensure a small spectral norm, we will set its expectation to 0, i.e., we will choose $c, \{c_i\}'s$ to ensure that $\mathbb{E}[\mathbf{W}_0] = 0$.

Following from the partially observed Stochastic Block Model (Defn 2.1 and 2.2), the expectation of an entry of $\mathbf{W}_0$ on $\mathcal{R}_{i,i}$ (region corresponding to cluster i) and $\mathcal{R}^c$ (region outside the clusters) is $c_i + \lambda r(2p_i - 1)$ and $c + \lambda r(2q - 1)$ respectively. Hence, we set,

$$c_i = -\lambda r(2p_i - 1) \quad \text{and} \quad c = -\lambda r(2q - 1),$$

With these, choices, the candidate $\mathbf{W}_0$ and $f(\mathbf{E}, \mathbf{W}_0)$ take the following forms,

$$\begin{aligned} \mathbf{W}_0 \quad = \quad & \lambda \left[\sum_{i=1}^{K} (1 + r(1 - 2p_i)) \ \mathbb{1}^{n\times n}_{\mathcal{R}_{i,i} \cap \mathcal{A}_1} + (-1 + r(1 - 2p_i)) \ \mathbb{1}^{n\times n}_{\mathcal{R}_{i,i} \cap \mathcal{A}_0} + r(1 - 2p_i) \mathbb{1}^{n\times n}_{\mathcal{R}_{i,i} \cap \Gamma^{out}} \right] \\ & + \lambda \left[(1 + r(1 - 2q)) \ \mathbb{1}^{n\times n}_{\mathcal{R}^c \cap \mathcal{A}_1} + (-1 + r(1 - 2q)) \ \mathbb{1}^{n\times n}_{\mathcal{R}^c \cap \mathcal{A}_0} + r(1 - 2q) \mathbb{1}^{n\times n}_{\mathcal{R}^c \cap \Gamma^{out}} \right] \end{aligned} \tag{4.21}$$

$$f(\mathbf{E}, \mathbf{W}_0) = \lambda \left[r(1 - 2q) \text{ sum}(\mathbf{E}_{\mathcal{R}^c}) \right] - \lambda \left[\sum_{i=1}^{K} \left(r(2p_i - 1) - \frac{1}{\lambda n_i} \right) \text{ sum}(\mathbf{E}_{\mathcal{R}_{i,i}}) \right]$$

From $\mathbf{L}^0$ and the constraint $1 \geq \mathbf{L}_{i,j} \geq 0$, it follows that,

$$\begin{aligned} & \mathbf{E}_{\mathcal{R}^c} \text{ is (entrywise) nonnegative.} \\ & \mathbf{E}_{\mathcal{R}} \text{ is (entrywise) nonpositive.} \end{aligned} \tag{4.22}$$

Thus, $\text{sum}(\mathbf{E}_{\mathcal{R}^c}) \leq 0$ and $\text{sum}(\mathbf{E}_{\mathcal{R}_{i,i}}) \geq 0$. When $\lambda(2p_i - 1) - \frac{1}{n_i} \geq 0$ and $\lambda(2q - 1) \leq 0$; we will have $f(\mathbf{E}, \mathbf{W}_0) \geq 0$ for all feasible $\mathbf{E}$. This indeed holds due to the assumptions of Theorem 1 (see (4.1)), as we assumed $r(2p_i - 1) > \frac{1}{\lambda n_i}$ for $i = 1, 2 \cdots, K$ and $1 > 2q$.

We will now proceed to find a tight bound on the spectral norm of $\mathbf{W}^0$. We will say that random variable X has a $\Delta(\zeta, \delta)$ distribution for $0 \leq \zeta, \delta \leq 1$, written as $X \sim \Delta(\zeta, \delta)$ if,

$$X = \begin{cases} 1 + \zeta(1 - 2\delta) \text{ w.p. } \zeta\delta \\ -1 + \zeta(1 - 2\delta) \text{ w.p. } \zeta(1 - \delta) \\ \zeta(1 - 2\delta) \text{ w.p. } 1 - \zeta \end{cases}$$

Variance of the above distribution is

$$\text{Var}(X) = \zeta(1 - \zeta + 4\,\zeta\,\delta\,(1 - \delta)). \tag{4.23}$$

Theorem 3. *Assume* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *obeys the Stochastic Block Model* (2.1) *and let* $\mathbf{M} \in \mathbb{R}^{n\times n}$. *Let entries of* $\mathbf{M}$ *be as follows.*

$$\mathbf{M}_{i,j} \sim \begin{cases} \Delta(r, p_k) & \text{if } (i,j) \in \mathcal{R}_{k,k} \\ \Delta(r,q) & \text{if } (i,j) \in \mathcal{R}^c \end{cases}$$

Then, for a constant ϵ' *(to be determined) each of the following holds with probability* $1 - \exp(-\Omega(n))$.

- $\|\mathbf{M}\| \le 2\sqrt{nr}\sqrt{1-r+4rq(1-q)}$
  $+ \max_{1\le i\le K} 2\sqrt{n_i r}\sqrt{2(1-r)+4r(q(1-q)+p_i(1-p_i))} + \epsilon'\sqrt{n}.$
- *Assume* $\max_{1\le i\le K} n_i = o(n)$. *Then, for sufficiently large* n,

$$\|\mathbf{M}\| \le (2\sqrt{r(1-r+4rq(1-q))} + \epsilon')\sqrt{n}.$$

Proof. For the first statement, let $\mathbf{M}_1$ be a random matrix with independent entries distributed as:

$$\mathbf{M}_1(i,j) \sim \Delta(r,q).$$

From standard results on random matrix theory [31], it follows that,

$$\|\mathbf{M}_1\| \le (2\sqrt{r(1-r+4rq(1-q))} + \epsilon')\sqrt{n}$$

with the desired probability.

Also let $\mathbf{M}_2 = \mathbf{M} - \mathbf{M}_1$. We note that $\mathbf{M}_2$ is a block diagonal random matrix. Observe that $\mathbf{M}_2$ over $\mathcal{R}_{i,i}$, $\mathbf{M}_{2,\mathcal{R}_{i,i}}$ is sum of two independent random variables $\mathbf{M}_{\mathcal{R}_{i,i}} \sim \Delta(r,p_i)$ and $-\mathbf{M}_{1,\mathcal{R}_{i,i}} \sim \Delta(r,q)$. So, the variance is $2r(1-r) + 4r^2(q(1-q)+p_i(1-p_i))$. This similarly gives,

$$\|\mathbf{M}_{2,\mathcal{R}_{i,i}}\| \le 2\sqrt{2r(1-r)+4r^2(q(1-q)+p_i(1-p_i))}\sqrt{n_i} + \epsilon'\sqrt{n}$$

Now, observing, $\|\mathbf{M}_2\| = \sup_{1\le i\le K} \|\mathbf{M}_{2,\mathcal{R}_{i,i}}\|$ and using a union bound over $i \le K$ we have,

$$\|\mathbf{M}_2\| \le \max_{1\le i\le K} 2\sqrt{2r(1-r)+4r^2(q(1-q)+p_i(1-p_i))}\sqrt{n_i} + \epsilon'\sqrt{n}$$

Finally, we use the triangle inequality $\|\mathbf{M}\| \le \|\mathbf{M}_1\| + \|\mathbf{M}_2\|$ to conclude. ■

The following lemma gives a bound on $\|\mathbf{W}_0\|$.

Lemma 4.4. *Recall that,* $\mathbf{W}_0$ *is a random matrix; where randomness is on the partially observed stochastic block model* $\mathbf{A}_{\mathbf{obs}}$ *and it is given by,*

$$\begin{aligned}\mathbf{W}_0 \;=\; & \lambda\left[\sum_{i=1}^{K}(1+r(1-2p_i))\;\mathbb{1}^{n\times n}_{\mathcal{R}_{i,i}\cap\mathcal{A}_1} + (1-r(1-2p_i))\;\mathbb{1}^{n\times n}_{\mathcal{R}_{i,i}\cap\mathcal{A}_0} + r(1-2p_i)\mathbb{1}^{n\times n}_{\mathcal{R}_{i,i}\cap\Gamma^{out}}\right] \\ & +\lambda\left[(1+r(1-2q))\;\mathbb{1}^{n\times n}_{\mathcal{R}^c\cap\mathcal{A}_1} + (-1+r(1-2q))\;\mathbb{1}^{n\times n}_{\mathcal{R}^c\cap\mathcal{A}_0} + r(1-2q)\mathbb{1}^{n\times n}_{\mathcal{R}^c\cap\Gamma^{out}}\right]\end{aligned}$$

Then, for any $\epsilon' > 0$, *with probability* $1 - \exp(-\Omega(n))$, *we have*

$$\|\frac{1}{\lambda}\mathbf{W}_0\| \le 2\sqrt{nr}\sqrt{1-r+4rq(1-q)} + \max_{1\le i\le K} 2\sqrt{n_i r}\sqrt{2(1-r)+4r(q(1-q)+p_i(1-p_i))} + \epsilon'\sqrt{n}$$

Further, if $\max_{1\le i\le K} n_i = o(n)$. *Then, for sufficiently large* n, *with the same probability,*

$$\|\mathbf{W}_0\| \le 2\lambda\sqrt{nr}\sqrt{1-r+4rq(1-q)} + \epsilon'\lambda\sqrt{n}.$$

Proof. $\frac{1}{\lambda}\mathbf{W}_0$ is a random matrix whose entries are i.i.d. and distributed as $\Delta(r,p_i)$ on $\mathcal{R}_{i,i}$ and $\Delta(r,q)$ on $\mathcal{R}^c$. Consequently, using Theorem 3 we obtain the result. ■

Lemma 4.4 verifies that asymptotically with high probability we can make $\|\mathbf{W}_0\| < 1$ as long as λ is sufficiently small. However, $\mathbf{W}_0$ itself is not sufficient for construction of the desired $\mathbf{W}$, since we do not have any guarantee that $\mathbf{W}_0 \in \mathcal{M}_{\mathbf{U}}$. In order to achieve this, we will *correct* $\mathbf{W}_0$ by projecting it onto $\mathcal{M}_{\mathbf{U}}$. Following lemma suggests that $\mathbf{W}_0$ does not change much by such a correction.

4.3.2 Correcting the candidate $\mathbf{W}_0$

Lemma 4.5. *$\mathbf{W}_0$ is as described previously in (4.21). Let $\mathbf{W}^H$ be the projection of $\mathbf{W}_0$ on $\mathcal{M}_{\mathbf{U}}$. Then*

- $\|\mathbf{W}^H\| \leq \|\mathbf{W}_0\|$
- *For any $\epsilon'' > 0$ (constant to be determined), with probability $1 - 6n^2\exp(-2\epsilon''^2 n_{min})$ we have*

$$\|\mathbf{W}_0 - \mathbf{W}^H\|_\infty \leq 3\lambda\epsilon''$$

Proof. Choose arbitrary vectors $\{\mathbf{u}_i\}_{i=K+1}^n$ to make $\{\mathbf{u}_i\}_{i=1}^n$ an orthonormal basis in $\mathbb{R}^n$. Call $\mathbf{U}_2 = [\mathbf{u}_{K+1}\ \ldots\ \mathbf{u}_n]$ and $\mathbf{P} = \mathbf{U}\mathbf{U}^T$, $\mathbf{P}_2 = \mathbf{U}_2\mathbf{U}_2^T$. Now notice that for any matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$, $\mathbf{P}_2\mathbf{X}\mathbf{P}_2$ is in $\mathcal{M}_{\mathbf{U}}$ since $\mathbf{U}^T\mathbf{U}_2 = 0$. Let $\mathbf{I}$ denote the identity matrix. Then,

$$\begin{aligned}\mathbf{X} - \mathbf{P}_2\mathbf{X}\mathbf{P}_2 &= \mathbf{X} - (\mathbf{I}-\mathbf{P})\mathbf{X}(\mathbf{I}-\mathbf{P}) \\ &= \mathbf{P}\mathbf{X} + \mathbf{X}\mathbf{P} - \mathbf{P}\mathbf{X}\mathbf{P} \in \mathcal{M}_{\mathbf{U}}^{\perp}\end{aligned} \tag{4.24}$$

Hence, $\mathbf{P}_2\mathbf{X}\mathbf{P}_2$ is the orthogonal projection on $\mathcal{M}_{\mathbf{U}}$. Clearly,

$$\|\mathbf{W}^H\| = \|\mathbf{P}_2\mathbf{W}_0\mathbf{P}_2\| \leq \|\mathbf{P}_2\|^2\|\mathbf{W}_0\| \leq \|\mathbf{W}_0\|$$

For analysis of $\|\mathbf{W}_0 - \mathbf{W}^H\|_\infty$ we can consider terms on the right hand side of (4.24) separately as we have:

$$\|\mathbf{W}_0 - \mathbf{W}^H\|_\infty \leq \|\mathbf{P}\mathbf{W}_0\|_\infty + \|\mathbf{W}_0\mathbf{P}\|_\infty + \|\mathbf{P}\mathbf{W}_0\mathbf{P}\|_\infty$$

Clearly $\mathbf{P} = \sum_{i=1}^K \frac{1}{n_i}\mathbb{1}_{\mathbb{R}_{i,i}}^{n\times n}$. Then, each entry of $\frac{1}{\lambda}\mathbf{P}\mathbf{W}_0$ is either a summation of n_i i.i.d. $\Delta(r, p_i)$ or $\Delta(r, q)$ mean zero random variables scaled by n_i^{-1} for some $i \leq K$ or 0. Hence any $c, d \in [n]$ and $\epsilon'' > 0$

$$\mathbb{P}[|(\mathbf{P}\mathbf{W}_0)_{c,d}| \geq \lambda\epsilon''] \leq 2\exp(-2\epsilon''^2 n_{min})$$

Same (or better) bounds holds for entries of $\mathbf{W}_0\mathbf{P}$ and $\mathbf{P}\mathbf{W}_0\mathbf{P}$. Then a union bound over all entries of the three matrices will give with probability $1-6n^2\exp(-2\epsilon''^2 n_{min})$, we have $\|\mathbf{W}_0 - \mathbf{W}^H\|_\infty \leq 3\lambda\epsilon''$. ■

Recall that, $\gamma_{\text{succ}} := \max_{1\leq i\leq K} 2r\sqrt{n_i}\sqrt{2(\frac{1}{r}-1) + 4\left(q(1-q) + p_i(1-p_i)\right)}$, and

$\mathbf{\Lambda}_{\text{succ}}^{-1} := 2r\sqrt{n}\sqrt{\frac{1}{r} - 1 + 4q(1-q)} + \gamma_{\text{succ}}$.

We can summarize our discussion so far in the following lemma,

Lemma 4.6. *$\mathbf{W}_0$ is as described previously in (4.21). Choose $\mathbf{W}$ to be projection of $\mathbf{W}_0$ on $\mathcal{M}_{\mathbf{U}}$. Also suppose $\lambda \leq (1-\delta)\mathbf{\Lambda}_{succ}$. Then, with probability $1 - 6n^2\exp(-\Omega(n_{min})) - 4\exp(-\Omega(n))$ we have,*

- $\|\mathbf{W}\| < 1$
- *For all feasible $\mathbf{E}$, $f(\mathbf{E}, \mathbf{W}) \geq 0$.*

Proof. To begin with, observe that $\mathbf{\Lambda}_{\text{succ}}^{-1}$ is $\Omega(\sqrt{n})$. Since $\lambda \leq \Lambda_{\text{succ}}$, $\lambda\sqrt{n} = \mathcal{O}(1)$. Consequently, using $\lambda\Lambda_{\text{succ}}^{-1} < 1$ and applying Lemma 4.4, and choosing a sufficiently small $\epsilon' > 0$, we conclude with,

$$\|\mathbf{W}\| \leq \|\mathbf{W}_0\| < 1$$

with probability $1-\exp(-\Omega(n))$ where the constant in the exponent depends on the constant $\epsilon' > 0$.

Next, from Lemma 4.5 with probability $1 - 6n^2 \exp(-\frac{2}{9}\epsilon''^2 n_{min})$ we have $\|\mathbf{W}_0 - \mathbf{W}\|_\infty \leq \lambda\epsilon''$. Then based on (5.10) for all $\mathbf{E}$, we have that,

$$\begin{aligned} f(\mathbf{E}, \mathbf{W}) &= f(\mathbf{E}, \mathbf{W}_0) - \langle \mathbf{W}_0 - \mathbf{W}, \mathbf{E} \rangle \\ &\geq f(\mathbf{E}, \mathbf{W}_0) - \lambda\epsilon'' \left(\text{sum}(\mathbf{E}_{\mathcal{R}}) - \text{sum}(\mathbf{E}_{\mathcal{R}^c})\right) \\ &= \lambda \left[(r(1-2q) - \epsilon'')\text{sum}(\mathbf{E}_{\mathcal{R}^c})\right] \\ &\quad -\lambda \sum_{i=1}^{K} \left[\left(r(2p_i - 1) - \frac{1}{\lambda n_i} - \epsilon'' \right) \text{sum}(\mathbf{E}_{\mathcal{R}_{i,i}}) \right] \\ &\geq 0 \end{aligned}$$

where we chose ϵ'' to be a sufficiently small constant. In particular, we set $\epsilon'' < \mathbf{D}_{\mathcal{A}}$, i.e., set $\epsilon'' < r(1-2q)$ and $\epsilon'' < r(2p_i - 1) - \frac{1}{\lambda n_i}$ for all $1 \leq i \leq K$.

Hence, by using a union bound $\mathbf{W}$ satisfies both of the desired conditions. ■

Summary so far: Combining the last lemma with Lemma 4.3, with high probability, either there exists a dual vector $\mathbf{W}^*$ which ensures $f(\mathbf{E}, \mathbf{W}^*) > 0$ or $\mathbf{E} \in \mathcal{M}_{\mathbf{U}}^{\perp}$. If former, we are done. Hence, we need to focus on the latter case and show that for all perturbations $\mathbf{E} \in \mathcal{M}_{\mathbf{U}}^{\perp}$, the objective will strictly increase at $(\mathbf{L}^0, \mathbf{S}^0)$ with high probability.

4.4 Solving for $\mathbf{E}^L \in \mathcal{M}_{\mathbf{U}}^{\perp}$ case

Recall that,

$$g(\mathbf{E}) = \sum_{i=1}^{K} \frac{1}{n_i} \text{sum}(\mathbf{E}_{R_{i,i}}) + \lambda \left(\text{sum}(\mathbf{E}_{\mathcal{A}_0}) - \text{sum}(\mathbf{E}_{\mathcal{A}_1})\right)$$

Let us define,

$$g_1(\mathbf{X}) := \sum_{i=1}^{K} \frac{1}{n_i} \text{sum}(\mathbf{X}_{\mathcal{R}_{i,i}}),$$

$$g_2(\mathbf{X}) := \text{sum}(\mathbf{X}_{\mathcal{A}_0}) - \text{sum}(\mathbf{X}_{\mathcal{A}_1}),$$

so that, $g(\mathbf{X}) = g_1(\mathbf{X}) + \lambda g_2(\mathbf{X})$. Also let $\mathbf{V} = [\mathbf{v}_1 \ \ldots \ \mathbf{v}_K]$ where $\mathbf{v}_i = \sqrt{n_i}\mathbf{u}_i$. Thus, $\mathbf{V}$ is basically obtained by, normalizing columns of $\mathbf{U}$ to make its nonzero entries 1. Assume $\mathbf{E} \in \mathcal{M}_{\mathbf{U}}^{\perp}$. Then, by definition of $\mathcal{M}_{\mathbf{U}}^{\perp}$, we can write,

$$\mathbf{E} = \mathbf{V}\mathbf{M}^T + \mathbf{N}\mathbf{V}^T.$$

Let $\mathbf{m}_i, \mathbf{n}_i$ denote i'th columns of $\mathbf{M}, \mathbf{N}$ respectively. From $\mathbf{L}^0$ and (1.3) it follows that

$\mathbf{E}_{\mathcal{R}^c}$ is (entrywise) nonnegative
$\mathbf{E}_{\mathcal{R}}$ is (entrywise) nonpositive

Now, we list some simple observations regarding structure of $\mathbf{E}$. We can write

$$\mathbf{E} = \sum_{i=1}^{K} (\mathbf{v}_i \mathbf{m}_i^T + \mathbf{n}_i \mathbf{v}_i^T) = \sum_{i=1}^{K+1} \sum_{j=1}^{K+1} \mathbf{E}_{\mathcal{R}_{i,j}} \tag{4.25}$$

Notice that only two components : $\mathbf{v}_i \mathbf{m}_i^T$ and $\mathbf{n}_j \mathbf{v}_j^T$, contribute to the term $\mathbf{E}_{\mathcal{R}_{i,j}}$.

Let $\mathbf{E}^{i,j} \in \mathbb{R}^{n_i \times n_j}$ which is $\mathbf{E}$ induced by entries on $\mathcal{R}_{i,j}$. Basically, $\mathbf{E}^{i,j}$ is same as $\mathbf{E}_{\mathcal{R}_{i,j}}$ when we get rid of trivial zero rows and zero columns. Then

$$\mathbf{E}^{i,j} = \mathbb{1}^{n_i} (\mathbf{m}_i^{\mathcal{C}_j})^T + \mathbf{n}_j^{\mathcal{C}_i} \mathbb{1}^{n_j T} \tag{4.26}$$

where $\mathbf{m}_i^{\mathcal{C}_j}$ is the vector corresponding to the entries of $\mathcal{C}_j$ in $\mathbf{m}_i$. Similarly, $\mathbf{n}_j^{\mathcal{C}_i}$ is the vector corresponding to the entries of $\mathcal{C}_i$ in $\mathbf{n}_j$.

Clearly, given $\{\mathbf{E}^{i,j}\}_{1\le i,j\le n}$, $\mathbf{E}$ is uniquely determined. Now, assume we fix $\text{sum}(\mathbf{E}^{i,j})$ for all i,j and we would like to find the *worst* $\mathbf{E}$ subject to these constraints. Variables in such an optimization are $\mathbf{m}_i, \mathbf{n}_i$. Basically we are interested in,

$$\min g(\mathbf{E}) \tag{4.27}$$
$$\text{subject to}$$
$$\text{sum}(\mathbf{E}^{i,j}) = c_{i,j} \text{ for all } i,j$$
$$\mathbf{E}^{i,j} \begin{cases} \text{nonnegative if } i \neq j \\ \text{nonpositive if } i = j \end{cases} \tag{4.28}$$

where $\{c_{i,j}\}$ are constants. Constraint (4.28) follows from (5.10).
Remark: For the special case of $i = j = K+1$, notice that $\mathbf{E}^{i,j} = 0$.

In (4.27), $g_1(\mathbf{E})$ is fixed and is equal to $\sum_{i=1}^{K} \frac{1}{n_i} c_{i,i}$. Consequently, we just need to do the optimization with the objective $g_2(\mathbf{E}) = \text{sum}(\mathbf{E}_{\mathcal{A}_0}) - \text{sum}(\mathbf{E}_{\mathcal{A}_1})$.

Let $\beta_{i,j} \subseteq [n_i] \times [n_j]$ be a set of coordinates defined as follows. For any $(c,d) \in [n_i] \times [n_j]$

$$(c,d) \in \beta_{i,j} \text{ iff } (a_{i,c}, a_{j,d}) \in \mathcal{A}$$

For $(i_1,j_1) \neq (i_2,j_2)$, $(\mathbf{m}_{i_1}^{\mathcal{C}_{j_1}}, \mathbf{n}_{j_1}^{\mathcal{C}_{i_1}})$ and $(\mathbf{m}_{i_2}^{\mathcal{C}_{j_2}}, \mathbf{n}_{j_2}^{\mathcal{C}_{i_2}})$ are independent variables. Consequently, due to (4.26), we can partition problem (4.27) into the following smaller disjoint problems.

$$\min_{\mathbf{m}_i^j, \mathbf{n}_j^i} \ \text{sum}(\mathbf{E}^{i,j}_{\beta^c_{i,j}}) - \text{sum}(\mathbf{E}^{i,j}_{\beta_{i,j}}) \tag{4.29}$$
$$\text{subject to}$$
$$\text{sum}(\mathbf{E}^{i,j}) = c_{i,j}$$
$$\mathbf{E}^{i,j} \text{ is } \begin{cases} \text{nonnegative if } i \neq j \\ \text{nonpositive if } i = j \end{cases}$$

Then, we can solve these problems locally (for each i,j) to finally obtain,

$$g_2(\mathbf{E}^{L,*}) = \sum_{i,j} \text{sum}(\mathbf{E}^{i,j,*}_{\beta^c_{i,j}}) - \sum_{i,j} \text{sum}(\mathbf{E}^{i,j,*}_{\beta_{i,j}})$$

to find the overall result of problem (4.27), where $*$ denotes the optimal solutions in problems (4.27) and (4.29). The following lemma will be useful for analysis of these local optimizations.

Lemma 4.7. *Let $\mathbf{a} \in \mathbb{R}^c$, $\mathbf{b} \in \mathbb{R}^d$ and $X = \mathbb{1}^c \mathbf{b}^T + \mathbf{a} \mathbb{1}^{d^T}$ be variables and $C_0 \geq 0$ be a constant. Also let $\beta \subseteq [c] \times [d]$. Consider the following optimization problem*

$$\min_{\mathbf{a},\mathbf{b}} \ \textit{sum}(\mathbf{X}_\beta)$$
$$\textit{subject to}$$
$$\mathbf{X}_{i,j} \geq 0 \ \textit{for all } i,j$$
$$\textit{sum}(\mathbf{X}) = C_0$$

For this problem there exists a (entrywise) nonnegative minimizer $(\mathbf{a}^0, \mathbf{b}^0)$.

Proof. Let x_i denotes i'th entry of vector $\mathbf{x}$. Assume $(\mathbf{a}^*, \mathbf{b}^*)$ is a minimizer. Without loss of generality assume $b_1^* = \min_{i,j}\{\mathbf{a}_i^*, \mathbf{b}_j^*\}$. If $b_1^* \geq 0$ we are done. Otherwise, since $\mathbf{X}_{i,j} \geq 0$ we have $a_i^* \geq -b_1^*$ for all $i \leq c$. Then set $\mathbf{a}^0 = \mathbf{a}^* + \mathbb{1}^c b_1^*$ and $\mathbf{b}^0 = \mathbf{b}^* - \mathbb{1}^d b_1^*$. Clearly, $(\mathbf{a}^0, \mathbf{b}^0)$ is nonnegative. On the other hand, we have:

$$\mathbf{X}^* = \mathbb{1}^c \mathbf{b}^{*T} + \mathbf{a}^* \mathbb{1}^{d^T} = \mathbb{1}^c \mathbf{b}^{0T} + \mathbf{a}^0 \mathbb{1}^{d^T} = \mathbf{X}^0,$$

which implies,

$$\text{sum}(\mathbf{X}^*_\beta) = \text{sum}(\mathbf{X}^0_\beta) = \text{optimal value}$$

■

Lemma 4.8. *A direct consequence of Lemma 4.7 is the fact that in the local optimizations (4.29), Without loss of generality, we can assume $(\mathbf{m}_i^{\mathcal{C}_j}, \mathbf{n}_j^{\mathcal{C}_i})$ entrywise nonnegative whenever $i \neq j$ and entrywise nonpositive when $i = j$. This follows from the structure of $\mathbf{E}^{i,j}$ given in (4.26) and (5.10).*

The following lemma will help us characterize the relationship between $\text{sum}(\mathbf{E}^{i,j})$ and $\text{sum}(\mathbf{E}^{i,j}_{\beta^c_{i,j}})$.

Lemma 4.9. *Let β be a random set generated by choosing elements of $[c] \times [d]$ independently with probability $0 \leq \eta \leq 1$. Then for any $\epsilon' > 0$ with probability $1 - d\exp(-2\epsilon'^2 c)$ for all nonzero and entrywise nonnegative $\mathbf{a} \in \mathbb{R}^d$ we'll have:*

$$\textit{sum}(\mathbf{X}_\beta) > (\eta - \epsilon')\textit{sum}(\mathbf{X}) \tag{4.30}$$

where $\mathbf{X} = \mathbb{1}^c \mathbf{a}^T$. Similarly, with the same probability, for all such $\mathbf{a}$, we'll have $\textit{sum}(\mathbf{X}_\beta) < (\eta + \epsilon')\textit{sum}(\mathbf{X})$

Proof. We'll only prove the first statement (4.30) as the proofs are identical. For each $i \leq d$, a_i occurs exactly c times in $\mathbf{X}$ as i'th column of X is $\mathbb{1}^c a_i$. By using a Chernoff bound, we can estimate the number of coordinates of i'th column which are element of β (call this number C_i) as we can view this number as a sum of c i.i.d. Bernoulli(η) random variables. Then

$$\mathbb{P}(C_i \leq c(\eta - \epsilon')) \leq \exp(-2\epsilon'^2 c)$$

Now, we can use a union bound over all columns to make sure for all i, $C_i > c(\eta - \epsilon')$

$$\mathbb{P}(C_i > c(r - \epsilon') \text{ for all } i \leq d) \geq 1 - d\exp(-2\epsilon'^2 c)$$

On the other hand if each $C_i > c(\eta - \epsilon')$ then for any nonnegative $\mathbf{a} \neq 0$,

$$\text{sum}(\mathbf{X}_\beta) = \sum_{(i,j)\in\beta} \mathbf{X}_{i,j} = \sum_{i=}^{d} C_i a_i > c(\eta - \epsilon') \sum_{i=1}^{d} a_i = (\eta - \epsilon')\text{sum}(\mathbf{X})$$

■

Using Lemma 4.9, we can calculate a lower bound for $g(\mathbf{E})$ with high probability as long as the cluster sizes are sufficiently large. Due to (4.25) and the linearity of $g(\mathbf{E})$, we can focus on contributions due to specific clusters i.e. $\mathbf{v}_i \mathbf{m}_i^T + \mathbf{n}_i \mathbf{v}_i^T$ for the i'th cluster. We additionally know the simple structure of $\mathbf{m}_i, \mathbf{n}_i$ from Lemma 4.8. In particular, subvectors $\mathbf{m}_i^{\mathcal{C}_i}$ and $\mathbf{n}_i^{\mathcal{C}_i}$ of $\mathbf{m}_i, \mathbf{n}_i$ can be assumed to be nonpositive and rest of the entries are nonnegative.

Lemma 4.10. *Assume, $1 \leq l \leq K$, $\mathbf{D}_{\mathcal{A}} > 0$. Then, with probability $1 - n\exp(-2\mathbf{D}_{\mathcal{A}}^2(n_l - 1))$, we have $g(\mathbf{v}_l \mathbf{m}_l^T) \geq 0$ for all $\mathbf{m}_l$. Also, if $\mathbf{m}_l \neq 0$ then inequality is strict.*

Proof. Recall that $\mathbf{m}_l$ satisfies $\mathbf{m}_l^{\mathcal{C}_i}$ is nonpositive/nonnegative when $i = l/i \neq l$ for all i. Define $\mathbf{X}^i := \mathbb{1}^{n_l} \mathbf{m}_l^{\mathcal{C}_i T}$. We can write

$$g(\mathbf{v}_l \mathbf{m}_l^T) = \frac{1}{n_l}\text{sum}(\mathbf{X}^l) + \sum_{i=1}^{K} \lambda h(\mathbf{X}^i, \beta^c_{l,i})$$

where $h(\mathbf{X}^i, \beta^c_{l,i}) = \text{sum}(\mathbf{X}^i_{\beta^c_{l,i}}) - \text{sum}(\mathbf{X}^i_{\beta_{l,i}})$. Now assume $i \neq l$. Using Lemma 4.9 and the fact that $\beta_{l,i}$ is a randomly generated subset (with parameter q), with probability $1 - n_i \exp(-2\epsilon'^2 n_l)$, for all $\mathbf{X}^i$, we have,

$$\begin{aligned} h(\mathbf{X}^i, \beta^c_{l,i}) &\geq (r(1-q) - \epsilon')\text{sum}(\mathbf{X}^i) - (rq + \epsilon')\text{sum}(\mathbf{X}^i) \\ &= (r(1-2q) - 2\epsilon')\text{sum}(\mathbf{X}^i) \end{aligned}$$

where inequality is strict if $X^i \neq 0$. Similarly, when $i = l$ with probability at least $1 - n_l \exp(-2\epsilon'^2(n_l - 1))$, we have,

$$\begin{aligned} \frac{1}{\lambda n_l}\text{sum}(\mathbf{X}^l) + h(\mathbf{X}^l, \beta^c_{l,l}) &\geq \left(r(1-p_l) + \epsilon' + \frac{1}{\lambda n_l}\right)\text{sum}(\mathbf{X}^l) - (rp_l - \epsilon')\,\text{sum}(\mathbf{X}^l) \\ &= -\left(r(2p_l - 1) - \frac{1}{\lambda n_l} - 2\epsilon'\right)\text{sum}(\mathbf{X}^l) \end{aligned}$$

Choosing $\epsilon' = \frac{\mathbf{D}_{\mathcal{A}}}{2}$ and using the facts that $r(1-2q) - 2\mathbf{D}_{\mathcal{A}} \geq 0$, $r(2p_l - 1) - \frac{1}{\lambda n_l} - 2\mathbf{D}_{\mathcal{A}} \geq 0$ and using a union bound, with probability $1 - n\exp(-2\mathbf{D}_{\mathcal{A}}^2(n_l - 1))$, we have $g(\mathbf{v}_l\mathbf{m}_l^T) \geq 0$ and the inequality is strict when $\mathbf{m}_l \neq 0$ as at least one of the $\mathbf{X}^i$'s will be nonzero. ■

The following lemma immediately follows from Lemma 4.10 and summarizes the main result of the section.

Lemma 4.11. *Let $\mathbf{D}_{\mathcal{A}}$ be as defined in (4.1) and assume $\mathbf{D}_{\mathcal{A}} > 0$. Then with probability $1 - 2nK\exp(-2\mathbf{D}_{\mathcal{A}}^2(n_{min} - 1))$ we have $g(\mathbf{E}^L) > 0$ for all nonzero feasible $\mathbf{E}^L \in \mathcal{M}_U^{\perp}$.*

4.5 The Final Step

Lemma 4.12. *Let $p_{min} > \frac{1}{2} > q$ and $\mathcal{G}$ be a random graph generated according to Model 2.1 and 2.2 with cluster sizes $\{n_i\}_{i=1}^K$. If $\lambda \leq (1-\epsilon)\mathbf{\Lambda}_{succ}$ and $\mathbf{D}_{\min} = \min_{1\leq i\leq n} r\,(2p_i - 1)\,n_i \geq (1+\epsilon)\frac{1}{\lambda}$, then $\left(\mathbf{L}^0, \mathbf{S}^0\right)$ is the unique optimal solution to Program 1.1 with probability $1 - \exp(-\Omega(n)) - 6n^2\exp(-\Omega(n_{min}))$.*

Proof. Based on Lemma 4.6 and Lemma 4.11,
with probability $1 - cn^2\exp(-C\left(\min\{r(1-2q), r(2p_{min}-1)\}\right)^2 n_{min})$,

- There exists $\mathbf{W} \in \mathcal{M}_{\mathbf{U}}$ with $\|\mathbf{W}\| < 1$ such that for all feasible $\mathbf{E}^L$, $f(\mathbf{E}^L, \mathbf{W}) \geq 0$.
- For all nonzero $\mathbf{E}^L \in \mathcal{M}_{\mathbf{U}}^{\perp}$ we have $g(\mathbf{E}^L) > 0$.

Consequently based on Lemma 4.3, $(\mathbf{L}^0, \mathbf{S}^0)$ is the unique optimal of Problem 1.1. ■

5 Proof of Results for Improved Convex Program

This section will show that, the optimal solution of Problem 1.4 is the pair $(\mathbf{L}^0, \mathbf{S}^0)$ under reasonable conditions, where,

$$\mathbf{L}^0 = \mathbb{1}_{\mathcal{R}}^{n\times n},\ \mathbf{S}^0 = \mathbf{S}_{\mathbf{obs}}^0 = \mathbb{1}_{\mathcal{R}\cap\mathcal{A}_0}^{n\times n} \tag{5.1}$$

Also denote the true optimal pair by $(\mathbf{L}^*, \mathbf{S}^*)$. Let $1 \geq p_{min} > q > 0$ and $0 \leq r \leq 1$. $\mathcal{G}$ be a random graph generated according to the stochastic block model 2.1 with cluster sizes $\{n_i\}_{i=1}^K$. Let the observation model be as defined in (2.2). Theorem 2 is based on the following lemma:

Lemma 5.1. *If $\lambda < \tilde{\mathbf{\Lambda}}_{succ}$ and $\tilde{\mathbf{D}}_{\min} > \frac{1}{\lambda}$, then $\left(\mathbf{L}^0, \mathbf{S}^0\right)$ is the unique optimal solution to Program 1.4 with high probability.*

Given $q, \{p_i\}_{i=1}^K$, define the following parameter which will be useful for the subsequent analysis. This parameter can be seen as a measure of distinctness of the "worst" cluster from the "background noise". Here, by background noise we mean the edges over $\mathcal{R}^c$.

$$\begin{aligned}\tilde{\mathbf{D}}_{\mathcal{A}} &= \frac{1}{2}\min\left\{r(1-q), \left\{r(p_i - q) - \frac{1}{\lambda n_i}\right\}_{i=1}^K\right\} \\ &= \frac{1}{2}\min\left\{r(1-q), \frac{\tilde{\mathbf{D}}_i - \lambda^{-1}}{n_i}\right\}\end{aligned} \tag{5.2}$$

5.1 Perturbation Analysis

Our aim is to show that $\left(\mathbf{L}^0, \mathbf{S}^0\right)$ defined in (5.1) is unique optimal solution to Problem 1.4.

Lemma 5.2. *Let $(\mathbf{E}^L, \mathbf{E}^S)$ be a feasible perturbation. Then, the objective will increase by at least,*

$$f(\mathbf{E}^L, \mathbf{W}) = \sum_{i=1}^K \frac{1}{n_i} sum(\mathbf{E}_{\mathcal{R}_{i,i}}^L) + \langle \mathbf{E}^L, \mathbf{W}\rangle + \lambda sum(\mathbf{E}_{\mathcal{A}_0}^L) \tag{5.3}$$

for any $\mathbf{W} \in \mathcal{M}$, $\|\mathbf{W}\| \leq 1$.

Proof. From constraint (1.6), we have $\mathbf{L}_{i,j} = \mathbf{S}_{i,j}$ whenever $\mathbf{A}^{obs}_{i,j} = 0$. So, $\mathbf{L}^*_{\mathcal{A}_0} = \mathbf{S}^*_{\mathcal{A}_0}$. Further, recall that $\mathbf{S}$ can be split as $\mathbf{S} = \mathbf{S_{obs}} + \mathbf{S_{rest}}$, where $\mathbf{S_{rest}}$ denotes the entries of $\mathbf{S}$ other than those corresponding to the observed entries of $\mathbf{A}$. Furthermore, we claim that at the optimal, $\mathbf{S_{rest}} = 0$, since if otherwise, the objective can be strictly decreased by setting $\mathbf{S_{rest}} = 0$. So, without loss of generality,

$$\mathbf{S}^* = \mathbf{L}^*_{\mathcal{A}_0}. \tag{5.4}$$

Recall that,

$$\begin{aligned}\|\mathbf{L}^0 + \mathbf{E}^L\|_\star + \lambda \|\mathbf{S}^0 + \mathbf{E}^S\|_1 - (\|\mathbf{L}^0\|_\star + \lambda \|\mathbf{S}^0\|_1) &\geq \langle \partial\|\mathbf{L}^0\|_\star, \mathbf{E}^L\rangle + \lambda\langle \partial\|\mathbf{S}^0\|_1, \mathbf{E}^S\rangle \\ &= \langle \mathbf{U}\mathbf{U}^T + \mathbf{W}, \mathbf{E}^L\rangle + \lambda\langle \text{sign}(\mathbf{S}^0) + \mathbf{Q}, \mathbf{E}^S\rangle\end{aligned}$$

Using $\text{sign}(\mathbf{S}^0) = \mathbb{1}^{n\times n}_{\mathcal{A}_0\cap\mathcal{R}}$, and choosing $\mathbf{Q} = \mathbb{1}^{n\times n}_{\mathcal{A}_0-(\mathcal{A}_0\cap\mathcal{R})}$, we get,

$$\begin{aligned}\|\mathbf{L}^0 + \mathbf{E}^L\|_\star + \lambda \|\mathbf{S}^0 + \mathbf{E}^S\|_1 - (\|\mathbf{L}^0\|_\star + \lambda \|\mathbf{S}^0\|_1) \geq \big\langle \mathbf{W}, \mathbf{E}^L\big\rangle \\ + \underbrace{\sum_{i=1}^{K} \frac{1}{n_i}\text{sum}(\mathbf{E}^L_{R_{i,i}}) + \lambda\left(\text{sum}(\mathbf{E}^L_{\mathcal{A}_0})\right)}_{:=g(\mathbf{E}^L)}\end{aligned} \tag{5.5}$$

for any $\mathbf{W} \in \mathcal{M}$. ■

From this point onward, for simplicity we will ignore the superscript L on $\mathbf{E}^L$ and just use $\mathbf{E}$.

Define,

$$g(\mathbf{E}) := \sum_{i=1}^{K} \frac{1}{n_i}\text{sum}(\mathbf{E}_{\mathcal{R}_{i,i}}) + \lambda\text{sum}(\mathbf{E}_{\mathcal{A}_0})). \tag{5.6}$$

Also, define $f(\mathbf{E}, \mathbf{W}) := g(\mathbf{E}) + \langle \mathbf{W}, \mathbf{E}\rangle$. Our aim is to show that for all feasible perturbations $\mathbf{E}$, there exists $\mathbf{W}$ such that,

$$f(\mathbf{E}, \mathbf{W}) = g(\mathbf{E}) + \langle \mathbf{W}, \mathbf{E}\rangle > 0. \tag{5.7}$$

Note that $g(\mathbf{E})$ does not depend on $\mathbf{W}$.

We can directly use Lemma 4.3. So, as in the previous section, we have broken down our aim into two steps.

1. Construct $\mathbf{W} \in \mathcal{M}_\mathbf{U}$ with $\|\mathbf{W}\| < 1$, such that $f(\mathbf{E}, \mathbf{W}) \geq 0$ for all feasible perturbations $\mathbf{E}$.
2. For all non-zero feasible $\mathbf{E} \in \mathcal{M}^\perp_\mathbf{U}$, show that $g(\mathbf{E}) > 0$.

As a first step, in Section 5.2, we will argue that, under certain conditions, there exists a $\mathbf{W} \in \mathcal{M}_\mathbf{U}$ with $\|\mathbf{W}\| < 1$ such that with high probability, $f(\mathbf{E}, \mathbf{W}) \geq 0$ for all feasible $\mathbf{E}$. Recall that such a $\mathbf{W}$ is called the dual certificate. Secondly, in Section 5.3, we will show that, under certain conditions, for all $\mathbf{E} \in \mathcal{M}^\perp_\mathbf{U}$ with high probability, $g(\mathbf{E}) > 0$. Finally, combining these two arguments, and using Lemma 4.3 we will conclude that $(\mathbf{L}^0, \mathbf{S}^0)$ is the unique optimal with high probability.

5.2 Showing existence of the dual certificate

Recall that

$$f(\mathbf{E}, \mathbf{W}) = \sum_{i=1}^{K} \frac{1}{n_i}\text{sum}(\mathbf{E}_{\mathcal{R}_{i,i}}) + \langle \mathbf{E}, \mathbf{W}\rangle + \lambda\text{sum}\left(\mathbf{E}_{\mathcal{A}_0}\right)$$

$\mathbf{W}$ will be constructed from the candidate $\mathbf{W}_0$, which is given as follows.

5.2.1 Candidate $\mathbf{W}_0$

Based on Program 1.4, we propose the following,

$$\mathbf{W}_0 = \sum_{i=1}^{K} c_i \mathbb{1}^{n\times n}_{\mathcal{R}_{i,i}} + c\mathbb{1}^{n\times n} - \lambda \mathbb{1}^{n\times n}_{\mathcal{A}_0}, \tag{5.8}$$

where $\{c_i\}_{i=1}^K, c$ are real numbers to be determined.

$$f(\mathbf{E}, \mathbf{W}^0) = \sum_{i=1}^{K} (\frac{1}{n_i} + c_i)\ \text{sum}(\mathbf{E}_{\mathcal{R}_{i,i}}) + c\ \text{sum}(\mathbf{E})$$

Note that $\mathbf{W}_0$ is a random matrix where randomness is due to $\mathbf{A_{obs}}$. In order to ensure a small spectral norm, we will set its expectation to 0, i.e., we will choose $c, \{c_i\}'s$ to ensure that $\mathbb{E}[\mathbf{W}_0] = 0$.

Following from the partially observed Stochastic Block Model (Definition 2.1 and Definition 2.2), the expectation of an entry of $\mathbf{W}_0$ on $\mathcal{R}_{i,i}$ (region corresponding to cluster i) and $\mathcal{R}^c$ (region outside the clusters) is $c_i + \lambda r(p_i - q)$ and $c + \lambda r(q-1)$ respectively. Hence, we set,

$$c_i = -\lambda r(p_i - q) \quad \text{and} \quad c = \lambda r(1-q),$$

With these, choices, the candidate $\mathbf{W}_0$ and $f(\mathbf{E}, \mathbf{W}_0)$ take the following forms,

$$\begin{aligned}\mathbf{W}_0 \quad = \quad & \lambda \left[\sum_{i=1}^{K} -(1 - r(1-p_i))\ \mathbb{1}^{n\times n}_{\mathcal{R}_{i,i}\cap\mathcal{A}_0} + r(1-p)\ \left(\mathbb{1}^{n\times n}_{\mathcal{R}_{i,i}\cap\mathcal{A}_1} + \mathbb{1}^{n\times n}_{\mathcal{R}_{i,i}\cap\Gamma^{out}} \right) \right] \\ & + \lambda \left[-(1 - r(1-q)) \mathbb{1}^{n\times n}_{\mathcal{R}^c\cap\mathcal{A}_0} + r(1-q)\ \left(\mathbb{1}^{n\times n}_{\mathcal{R}^c\cap\mathcal{A}_1} + \mathbb{1}^{n\times n}_{\mathcal{R}^c\cap\Gamma^{out}} \right) \right] \end{aligned} \tag{5.9}$$

$$f(\mathbf{E}, \mathbf{W}_0) = \lambda \left[r(1-q)\ \text{sum}(\mathbf{E}) \right] - \lambda \left[\sum_{i=1}^{K} \left(r(p_i - q) - \frac{1}{\lambda n_i} \right)\ \text{sum}(\mathbf{E}_{\mathcal{R}_{i,i}}) \right]$$

From $\mathbf{L}^0$ and the constraint $1 \geq \mathbf{L}_{i,j} \geq 0$, it follows that,

$$\begin{aligned} &\mathbf{E}_{\mathcal{R}^c} \text{ is (entrywise) nonnegative.} \\ &\mathbf{E}_{\mathcal{R}} \text{ is (entrywise) nonpositive.} \end{aligned} \tag{5.10}$$

Thus, $\text{sum}(\mathbf{E}_{\mathcal{R}^c}) \leq 0$ and $\text{sum}(\mathbf{E}_{\mathcal{R}_{i,i}}) \geq 0$. When $\lambda r(p_i - q) - \frac{1}{n_i} \geq 0$ and $\lambda(1-q) \geq 0$; we will have $f(\mathbf{E}, \mathbf{W}_0) \geq 0$ for all feasible $\mathbf{E}$. This indeed holds due to the assumptions of Theorem 2 (see (5.2)), as we assumed $r(p_i - q) > \frac{1}{\lambda n_i}$ for $i = 1, 2 \cdots, K$ and $1 > q$.

Using the same technique as in Theorem 3, we can bound the spectral norm of $\mathbf{W}^0$ as follows

Lemma 5.3. *Recall that, $\mathbf{W}_0$ is a random matrix; where randomness is on the partially observed stochastic block model $\mathbf{A_{obs}}$ and it is given by,*

$$\begin{aligned}\mathbf{W}_0 \quad = \quad & \lambda \left[\sum_{i=1}^{K} -(1 - r(1-p_i))\ \mathbb{1}^{n\times n}_{\mathcal{R}_{i,i}\cap\mathcal{A}_0} + r(1-p)\ \left(\mathbb{1}^{n\times n}_{\mathcal{R}_{i,i}\cap\mathcal{A}_1} + \mathbb{1}^{n\times n}_{\mathcal{R}_{i,i}\cap\Gamma^{out}} \right) \right] \\ & + \lambda \left[-(1 - r(1-q)) \mathbb{1}^{n\times n}_{\mathcal{R}^c\cap\mathcal{A}_0} + r(1-q)\ \left(\mathbb{1}^{n\times n}_{\mathcal{R}^c\cap\mathcal{A}_1} + \mathbb{1}^{n\times n}_{\mathcal{R}^c\cap\Gamma^{out}} \right) \right] \end{aligned}$$

Then, for any $\epsilon' > 0$, with probability $1 - \exp(-\Omega(n))$, we have

$$\|\frac{1}{\lambda}\mathbf{W}_0\| \leq 2\sqrt{nr}\sqrt{(1-q)(1-r+rq)} + \max_{1\leq i\leq K} 2\sqrt{n_i r}\sqrt{(1-p_i)(1-r+rp_i) + (1-q)(1-r+rq)} + \epsilon'\sqrt{n}$$

Further, if $\max_{1\leq i\leq K} n_i = o(n)$. Then, for sufficiently large n, with the same probability,

$$\|\mathbf{W}_0\| \leq 2\lambda\sqrt{nr}\sqrt{(1-q)(1-r+rq)} + \epsilon'\lambda\sqrt{n}.$$

Lemma 5.3 verifies that asymptotically with high probability we can make $\|\mathbf{W}_0\| < 1$ as long as λ is sufficiently small. However, $\mathbf{W}_0$ itself is not sufficient for construction of the desired $\mathbf{W}$, since we do not have any guarantee that $\mathbf{W}_0 \in \mathcal{M}_{\mathbf{U}}$. In order to achieve this, we will *correct* $\mathbf{W}_0$ by projecting it onto $\mathcal{M}_{\mathbf{U}}$. Lemma 4.5 can be used to here.

Recall that, $\tilde{\gamma}_{\text{succ}} := 2 \max_{1 \le i \le K} r\sqrt{n_i}\sqrt{(1-p_i)(\frac{1}{r}-1+p_i)+(1-q)(\frac{1}{r}-1+q)}$ and $\tilde{\mathbf{\Lambda}}^{-1}_{\text{succ}} := 2r\sqrt{n}\sqrt{(\frac{1}{r}-1+q)(1-q)}+\tilde{\gamma}_{\text{succ}}$.

We can summarize our discussion so far in the following lemma,

Lemma 5.4. *$\mathbf{W}_0$ is as described previously in (5.9). Choose $\mathbf{W}$ to be projection of $\mathbf{W}_0$ on $\mathcal{M}_{\mathbf{U}}$. Also suppose $\lambda \le (1-\delta)\tilde{\mathbf{\Lambda}}_{succ}$. Then, with probability $1-6n^2\exp(-\Omega(n_{min}))-4\exp(-\Omega(n))$ we have,*

- $\|\mathbf{W}\| < 1$
- *For all feasible $\mathbf{E}$, $f(\mathbf{E},\mathbf{W}) \ge 0$.*

Proof. To begin with, observe that $\tilde{\mathbf{\Lambda}}^{-1}_{\text{succ}}$ is $\Omega(\sqrt{n})$. Since $\lambda \le \tilde{\mathbf{\Lambda}}_{\text{succ}}$, $\lambda\sqrt{n} = \mathcal{O}(1)$. Consequently, using $\lambda\tilde{\mathbf{\Lambda}}^{-1}_{\text{succ}} < 1$ and applying Lemma 5.3, and choosing a sufficiently small $\epsilon' > 0$, we conclude with,

$$\|\mathbf{W}\| \le \|\mathbf{W}_0\| < 1$$

with probability $1-\exp(-\Omega(n))$ where the constant in the exponent depends on the constant $\epsilon' > 0$.

Next, from Lemma 4.5 with probability $1-6n^2\exp(-\frac{2}{9}\epsilon''^2 n_{min})$ we have $\|\mathbf{W}_0-\mathbf{W}\|_\infty \le \lambda\epsilon''$. Then based on (5.10) for all $\mathbf{E}$, we have that,

$$\begin{aligned} f(\mathbf{E},\mathbf{W}) &= f(\mathbf{E},\mathbf{W}_0) - \langle \mathbf{W}_0-\mathbf{W},\mathbf{E}\rangle \\ &\ge f(\mathbf{E},\mathbf{W}_0) - \lambda\epsilon''\left(\text{sum}(\mathbf{E}_{\mathcal{R}}) - \text{sum}(\mathbf{E}_{\mathcal{R}^c})\right) \\ &= \lambda\left[(r(1-q)-\epsilon'')\text{sum}(\mathbf{E}_{\mathcal{R}^c})\right] \\ &\quad -\lambda\sum_{i=1}^{K}\left[\left(r(p_i-q)-\frac{1}{\lambda n_i}-\epsilon''\right)\text{sum}(\mathbf{E}_{\mathcal{R}_{i,i}})\right] \\ &\ge 0 \end{aligned}$$

where we chose ϵ'' to be a sufficiently small constant. In particular, we set $\epsilon'' < \tilde{\mathbf{D}}_{\mathcal{A}}$, i.e., set $\epsilon'' < r(1-q)$ and $\epsilon'' < r(p_i-q)-\frac{1}{\lambda n_i}$ for all $1 \le i \le K$.

Hence, by using a union bound $\mathbf{W}$ satisfies both of the desired conditions. ■

Summary so far: Combining the last lemma with Lemma 4.3, with high probability, either there exists a dual vector $\mathbf{W}^*$ which ensures $f(\mathbf{E},\mathbf{W}^*) > 0$ or $\mathbf{E} \in \mathcal{M}^{\perp}_{\mathbf{U}}$. If former, we are done. Hence, we need to focus on the latter case and show that for all perturbations $\mathbf{E} \in \mathcal{M}^{\perp}_{\mathbf{U}}$, the objective will strictly increase at $(\mathbf{L}^0,\mathbf{S}^0)$ with high probability.

5.3 Solving for $\mathbf{E}^L \in \mathcal{M}^{\perp}_{\mathbf{U}}$ case

Recall that,

$$g(\mathbf{E}) = \sum_{i=1}^{K}\frac{1}{n_i}\text{sum}(\mathbf{E}_{R_{i,i}}) + \lambda\text{sum}(\mathbf{E}_{\mathcal{A}_0})$$

Let us define,

$$g_1(\mathbf{X}) := \sum_{i=1}^{K}\frac{1}{n_i}\text{sum}(\mathbf{X}_{\mathcal{R}_{i,i}}),$$

$$g_2(\mathbf{X}) := \text{sum}(\mathbf{X}_{\mathcal{A}_0}),$$

so that, $g(\mathbf{X}) = g_1(\mathbf{X}) + \lambda g_2(\mathbf{X})$. Also let $\mathbf{V} = [\mathbf{v}_1 \ \dots \ \mathbf{v}_K]$ where $\mathbf{v}_i = \sqrt{n_i}\mathbf{u}_i$. Thus, $\mathbf{V}$ is basically obtained by, normalizing columns of $\mathbf{U}$ to make its nonzero entries 1. Assume $\mathbf{E} \in \mathcal{M}_{\mathbf{U}}^{\perp}$. Then, by definition of $\mathcal{M}_{\mathbf{U}}^{\perp}$, we can write,

$$\mathbf{E} = \mathbf{V}\mathbf{M}^T + \mathbf{N}\mathbf{V}^T.$$

Let $\mathbf{m}_i, \mathbf{n}_i$ denote i'th columns of $\mathbf{M}, \mathbf{N}$ respectively.

Again as in previous section 4.4, we consider optimization problem 4.27. Since $g_1(\mathbf{E})$ is fixed, we just need to optimize over $g_2(\mathbf{E})$. This optimization can be reduced to local optimizations 4.29. Since $\mathbf{L}^0 = \mathbb{1}_{\mathcal{R}}^{n\times n}$ and the condition (1.3),

$$\mathbf{E}_{\mathcal{R}^c} \text{ is (entrywise) nonnegative}$$
$$\mathbf{E}_{\mathcal{R}} \text{ is (entrywise) nonpositive}$$

We can make use of Lemma 4.8 and assume $\mathbf{m}_l^{\mathcal{C}_i}$ is nonpositive/nonnegative when $i = l/i \neq l$ for all i. Hence using Lemma 4.30 we lower bound $g(\mathbf{v}_l\mathbf{m}_l^T)$ as described in the following section.

5.3.1 Lower bounding $g(\mathbf{E})$

Lemma 5.5. *Assume, $1 \leq l \leq K$, $\tilde{\mathbf{D}}_{\mathcal{A}} > 0$. Then, with probability $1 - n\exp(-2\tilde{\mathbf{D}}_{\mathcal{A}}^2(n_l - 1))$, we have $g(\mathbf{v}_l\mathbf{m}_l^T) \geq \lambda(1 - q - \tilde{\mathbf{D}}_{\mathcal{A}})sum(\mathbf{v}_l\mathbf{m}_l^T)$ for all $\mathbf{m}_l$. Also, if $\mathbf{m}_l \neq 0$ then inequality is strict.*

Proof. Recall that $\mathbf{m}_l$ satisfies $\mathbf{m}_l^{\mathcal{C}_i}$ is nonpositive/nonnegative when $i = l/i \neq l$ for all i. Define $\mathbf{X}^i := \mathbb{1}^{n_l}{\mathbf{m}_l^{\mathcal{C}_i}}^T$. We can write

$$g(\mathbf{v}_l\mathbf{m}_l^T) = \frac{1}{n_l}\text{sum}(\mathbf{X}^l) + \sum_{i=1}^{K}\lambda\text{sum}(\mathbf{X}^i_{\beta^c_{l,i}})$$

Now assume $i \neq l$. Using Lemma 4.9 and the fact that $\beta_{l,i}$ is a randomly generated subset (with parameter q), with probability $1 - n_i\exp(-2\epsilon'^2 n_l)$, for all $\mathbf{X}^i$, we have,

$$\text{sum}(\mathbf{X}^i_{\beta^c_{l,i}}) \geq (r(1-q) - \epsilon')\text{sum}(\mathbf{X}^i) \tag{5.11}$$

where inequality is strict if $X^i \neq 0$. Similarly, when $i = l$ with probability at least $1 - n_l\exp(-2\epsilon'^2(n_l - 1))$, we have,

$$\frac{1}{\lambda n_l}\text{sum}(\mathbf{X}^l) + \text{sum}(\mathbf{X}^l_{\beta^c_{l,l}}) \geq \left(\frac{1}{\lambda n_l} + r(1-p_l) + \epsilon'\right)\text{sum}(\mathbf{X}^l)$$

Together,

$$\begin{aligned} g(\mathbf{v}_l\mathbf{m}_l^T) &\geq \lambda\sum_{i\neq l}(r(1-q) - \epsilon')\text{sum}(\mathbf{X}^i) + \left(\frac{1}{\lambda n_l} + r(1-p_l) + \epsilon'\right)\text{sum}(\mathbf{X}^l) \\ &\geq \lambda(r(1-q) - \epsilon')\sum_{i=1}^{K}\text{sum}(\mathbf{X}^i) = \lambda(r(1-q) - \epsilon')\text{sum}(\mathbf{v}_l\mathbf{m}_l^T) \end{aligned} \tag{5.12}$$

Choosing $\epsilon' = \frac{\tilde{\mathbf{D}}_{\mathcal{A}}}{2}$ and using the facts that $r(1-q) - 2\tilde{\mathbf{D}}_{\mathcal{A}} \geq 0$, $r(p_l - q) - \frac{1}{\lambda n_l} - 2\tilde{\mathbf{D}}_{\mathcal{A}} \geq 0$ and using a union bound, with probability $1 - n\exp(-2\tilde{\mathbf{D}}_{\mathcal{A}}^2(n_l - 1))$, we have $g(\mathbf{v}_l\mathbf{m}_l^T) \geq 0$ and the inequality is strict when $\mathbf{m}_l \neq 0$ as at least one of the $\mathbf{X}^i$'s will be nonzero. ■

The following lemma immediately follows from Lemma 5.5 and summarizes the main result of the section.

Lemma 5.6. *Let $\tilde{\mathbf{D}}_{\mathcal{A}}$ be as defined in (4.1) and assume $\tilde{\mathbf{D}}_{\mathcal{A}} > 0$. Then with probability $1 - 2nK\exp(-2\tilde{\mathbf{D}}_{\mathcal{A}}^2(n_{min} - 1))$ we have $g(\mathbf{E}^L) > 0$ for all nonzero feasible $\mathbf{E}^L \in \mathcal{M}_U^{\perp}$.*

5.4 The Final Step

Lemma 5.7. *Let $p_{min} > q$ and $\mathcal{G}$ be a random graph generated according to Model 2.1 and 2.2 with cluster sizes $\{n_i\}_{i=1}^K$. If $\lambda \leq (1-\epsilon)\tilde{\mathbf{\Lambda}}_{succ}$ and $\mathbf{D}_{\min} = \min_{1\leq i\leq n} r\,(p_i - q)\,n_i \geq (1+\epsilon)\frac{1}{\lambda}$, then $\left(\mathbf{L}^0, \mathbf{S}^0\right)$ is the unique optimal solution to Program 1.1 with probability $1 - \exp(-\Omega(n)) - 6n^2\exp(-\Omega(n_{min}))$.*

Proof. Based on Lemma 5.4 and Lemma 5.6,
with probability $1 - cn^2\exp(-C(r(p_{min}-q))^2 n_{min})$,

- There exists $\mathbf{W} \in \mathcal{M}_{\mathbf{U}}$ with $\|\mathbf{W}\| < 1$ such that for all feasible $\mathbf{E}^L$, $f(\mathbf{E}^L, \mathbf{W}) \geq 0$.
- For all nonzero $\mathbf{E}^L \in \mathcal{M}_{\mathbf{U}}^{\perp}$ we have $g(\mathbf{E}^L) > 0$.

Consequently based on Lemma 4.3, $(\mathbf{L}^0, \mathbf{S}^0)$ is the unique optimal of Problem 1.4. ■