SHARP PERFORMANCE BOUNDS FOR GRAPH CLUSTERING VIA CONVEX OPTIMIZATION

Ramya Korlakai Vinayak, Samet Oymak*, Babak Hassibi*

California Institute of Technology, Pasadena, CA, USA

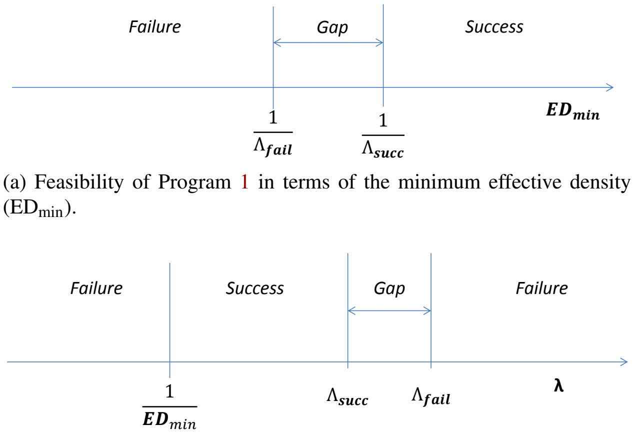
ABSTRACT

The problem of finding clusters in a graph arises in several applications such as social networks, data mining and computer networks. A typical, convex optimization approach, that is often adopted is to identify a sparse plus low-rank decomposition of the adjacency matrix of the graph, with the (dense) low-rank component representing the clusters. In this paper, we sharply characterize the conditions for successfully identifying clusters using this approach. In particular, we introduce the "effective density" of a cluster that measures its significance and we find explicit upper and lower bounds on the minimum effective density that demarcates regions of success or failure of this technique. Our conditions are in terms of (a) the size of the clusters, (b) the denseness of the graph, and (c) regularization parameter of the convex program. We also present extensive simulations that corroborate our theoretical findings.

Index Terms— Graph clustering, low rank plus sparse, convex optimization, thresholds.

1. INTRODUCTION

Given an unweighted graph, finding nodes that are well-connected with each other is a very useful problem with applications in social networks [1–3], data mining [4, 5], bioinformatics [6, 7], computer networks, sensor networks. Different versions of this problem have been studied as graph clustering [8–11], correlation clustering [12–15], graph partitioning on planted partition model [16–19]. Developments in convex optimization techniques to recover low-rank matrices [20–24] via nuclear norm minimization has recently led to the development of several convex algorithms to recover clusters in a graph [25–32].



(a) Feasibility of Program 1 in terms of the minimum effective density (ED_{min}).

(b) Feasibility of Program 1 in terms of the regularization parameter (λ).

Fig. 1: Characterization of the feasibility of Program (1) in terms of the minimum effective density and the value of the regularization parameter. The feasibility is determined by the values of these parameters in comparison with two constants Λ_{succ} and Λ_{fail}, derived in Theorem 1 and Theorem 2. The thresholds guaranteeing the success or failure of Program 1 derived in this paper are fairly close to each other.

Let us assume that a given graph has dense clusters; we can look at its adjacency matrix as a low-rank matrix with sparse noise. That is, the graph can be viewed as a union of cliques with some edges missing inside the cliques and extra edges between the cliques. Our aim is to recover the low-rank matrix since it is equivalent to finding clusters. In this paper, we will look at the following well known convex program which decomposes the adjacency matrix ($\mathbf{A}$) as the sum of a low-rank ($\mathbf{L}$) and a sparse ($\mathbf{S}$) component.

$$\underset{\mathbf{L},\mathbf{S}}{\text{minimize}} \quad \|\mathbf{L}\|_\star + \lambda \|\mathbf{S}\|_1 \tag{1}$$

$$\text{subject to}$$

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

$$\mathbf{L} + \mathbf{S} = \mathbf{A}$$

where $\lambda > 0$ is a regularization parameter. $\|\mathbf{X}\|_\star$ and $\|\mathbf{X}\|_1$ denote the nuclear norm (sum of the singular values) and the ℓ_1-norm (sum of the absolute values of all entries) of the matrix $\mathbf{X}$ respectively. This program is very intuitive and requires the knowledge of only the adjacency matrix. Program 1 has been proposed in several works [28–30].

We consider the popular *stochastic block model* (also called the planted partition model) for the graph. Under this model of generating random graphs, the existence of an

$*$ Authors contributed equally. 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. A shorter version of this work will appear on ICASSP 2014.

edge between any pair of vertices is independent of the other edges. The probability of the existence of an edge is identical within any individual cluster, but may vary across clusters. One may think of this as a heterogeneous form of the Erdös-Renyi model. We characterize the conditions under which Program 1 can successfully recover the correct clustering, and when it cannot. Our analysis reveals the dependence of its success on a metric that we term the *minimum effective density* of the graph. While defined more formally later in the paper, in a nutshell, the minimum effective density of a random graph tries to capture the density of edges in the sparsest cluster. We derive explicit upper and lower bounds on the value of this metric that determine the success or failure of Program 1 (as illustrated in Fig. 1a).

A second contribution of this paper is to explicitly characterize the efficacy of Program 1 with respect to the regularization parameter λ. We obtain bounds on the values of λ that permit the recovery of the clusters, or those that necessitate Program 1 to fail (as illustrated in Fig. 1b). Our results thus lead to a more principled approach towards the choice of the regularization parameter for the problem at hand.

Most of the convex algorithms proposed for graph clustering, for example, the recent works by Xu et al. [25], Ames and Vavasis [26, 27], Jalali et al. [28], Oymak and Hassibi [29], Chen et al. [30], Ames [31], Ailon et al. [32] are variants of Program 1. These results show that planted clusters can be identified via tractable convex programs as long as the cluster size is proportional to the square-root of the size of the adjacency matrix. However, the exact requirements on the cluster size are not known. In this work, we find sharp bounds for the identifiability as a function of cluster sizes, inter cluster density and intra cluster density. To the best of our knowledge, this is the first explicit characterization of the feasibility of the convex optimization based approach (1) towards this problem.

The rest of the paper is organized as follows. Section 2 formally introduces the model considered in this paper. Section 3 presents the main results of the paper: an analytical characterization of the feasibility of the low rank plus sparse based approximation for identifying clusters. Section 4 presents simulations that corroborate our theoretical results. Finally, the proofs of the technical results are deferred to Sections 7 and 8.

2. MODEL

For any positive integer m, let $[m]$ denote the set $\{1, 2, \ldots, m\}$. Let $\mathcal{G}$ be an unweighted graph on n nodes, $[n]$, with K disjoint (dense) clusters. Let $\mathcal{C}_i$ denote the set of nodes in the i^{th} cluster. Let n_i denote the size of the i^{th} cluster, i.e., the number of nodes in $\mathcal{C}_i$. We shall term the set of nodes that do not fall in any of these K clusters as *outliers* and denote them as $\mathcal{C}_{K+1} := [n] - \bigcup_{i=1}^{K} \mathcal{C}_i$. The number of outliers is thus $n_{K+1} := n - \sum_{i=1}^{K} n_i$. Since the clusters are assumed to be disjoint, we have $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ for all $i, j \in [n]$.

Let $\mathcal{R}$ be the region corresponding to the union of regions induced by the clusters, i.e., $\mathcal{R} = \bigcup_{i=1}^{K} \mathcal{C}_i \times \mathcal{C}_i \subseteq [n] \times [n]$. So, $\mathcal{R}^c = [n] \times [n] - \mathcal{R}$ is the region corresponding to out of cluster regions. 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 $n_{min} := \min\limits_{1 \le i \le K} n_i$.

Let $\mathbf{A} = \mathbf{A}^T$ denote the adjacency matrix of the graph $\mathcal{G}$. The diagonal entries of $\mathbf{A}$ are 1. The adjacency matrix will follow a probabilistic model, in particular, a more general version of the popular stochastic block model [16, 33].

Definition 1 (Stochastic Block Model)**.** *Let $\{p_i\}_{i=1}^{K}, q$ be constants between 0 and 1. Then, a random graph $\mathcal{G}$, generated according to stochastic block model, has the following adjacency matrix. Entries of $\mathbf{A}$ on the lower triangular part are independent random variables and for any $i > j$:*

$$\mathbf{A}_{i,j} = \begin{cases} \textit{Bernoulli}(p_l) & \textit{if both } \{i,j\} \in \mathcal{C}_l \textit{ for some } l \le K \\ \textit{Bernoulli}(q) & \textit{otherwise.} \end{cases}$$

So, an edge inside i^{th} cluster exists with probability p_i and an edge outside the clusters exists with probability q. Let $p_{min} := \min\limits_{1 \le i \le K} p_i$. We assume that the clusters are dense and the density of edges inside clusters is greater than outside, i.e., $p_{\min} > \frac{1}{2} > q > 0$. We note that the Program 1 does not require the knowledge of $\{p_i\}_{i=1}^{K}, q$ or K, and uses only the adjacency matrix $\mathbf{A}$ for its operation. However, the knowledge of $\{p_i\}_{i=1}^{K}, q$ will help us tune λ in a better way.

3. MAIN RESULTS

The desired solution to Program 1 is $(\mathbf{L}^0, \mathbf{S}^0)$ where $\mathbf{L}^0$ corresponds to the full cliques, when missing edges inside $\mathcal{R}$ are completed, and $\mathbf{S}^0$ corresponds to the missing edges and the extra edges between the clusters. In particular we want:

$$\mathbf{L}^0_{i,j} = \begin{cases} 1 & \text{if both } \{i,j\} \in \mathcal{C}_l \text{ for some } l \le K, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

$$\mathbf{S}^0_{i,j} = \begin{cases} -1 & \text{if both } \{i,j\} \in \mathcal{C}_l \text{ for some } l \le K, \text{ and } \mathbf{A}_{i,j} = 0, \\ 1 & \text{if } \{i,j\} \text{ are not in the same cluster and } \mathbf{A}_{i,j} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that the $(\mathbf{L}^0, \mathbf{S}^0)$ pair is feasible. We say that Program 1 *succeeds* when $(\mathbf{L}^0, \mathbf{S}^0)$ is the optimal solution to Program 1. In this section we present two theorems which give the conditions under which Program 1 succeeds or fails.

The following definitions are critical to our results.

- Define $\mathbf{ED}_i := n_i\,(2p_i - 1)$ as the effective density of cluster $\mathcal{C}_i$ and $\mathbf{ED}_{\min} = \min\limits_{1 \le i \le K} \mathbf{ED}_i$.

- Let $\gamma_{\text{succ}} := \max_{1\leq i\leq K} 4\sqrt{(q(1-q)+p_i(1-p_i))n_i}$, $\gamma_{\text{fail}} := \sum_{i=1}^{K} \frac{n_i^2}{n}$
- $\Lambda_{\text{fail}} := \frac{1}{\sqrt{q(n-\gamma_{\text{fail}})}}$ and $\Lambda_{\text{succ}} := \frac{1}{4\sqrt{q(1-q)n+\gamma_{\text{succ}}}}$.

Theorem 1. *Let $\mathcal{G}$ be a random graph generated according to the Stochastic Block Model 1 with K 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 δ, c_1, c_2 such that,*

1. *For any given $\lambda \geq 0$, if $\mathbf{ED}_{\min} \leq (1-\epsilon)\Lambda_{fail}^{-1}$ then Program 1 fails with probability $1 - c_1 \exp(-c_2|\mathcal{R}^c|)$.*
2. *Whenever $\mathbf{ED}_{\min} \geq (1+\epsilon)\Lambda_{succ}^{-1}$, for $\lambda = (1-\delta)\Lambda_{succ}$, Program 1 succeeds with probability $1 - c_1 n^2 \exp(-c_2 n_{\min})$.*

As it will be discussed in Sections 7 and 8, Theorem 1 is actually a special case of the following result, which characterizes success and failure as a function of λ.

Theorem 2. *Let $\mathcal{G}$ be a random graph generated according to the Stochastic Block Model 1 with K 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 $\lambda \geq (1+\epsilon)\Lambda_{fail}$, then Program 1 fails with probability $1 - c_1' \exp(-c_2'|\mathcal{R}^c|)$.*
2. *If $\lambda \leq (1-\epsilon)\Lambda_{succ}$ then,*
    - *If $\mathbf{ED}_{\min} \leq (1-\epsilon)\frac{1}{\lambda}$, then Program 1 fails with probability $1 - c_1' \exp(-c_2' n_{\min})$.*
    - *If $\mathbf{ED}_{\min} \geq (1+\epsilon)\frac{1}{\lambda}$, then Program 1 succeeds with probability $1 - c_1' n^2 \exp(-c_2' n_{\min})$.*

We see that the minimum effective density $\mathbf{ED}_{\min}$, Λ_{succ} and Λ_{fail} play a fundamental role in determining the success of Program 1. Theorem 1 gives a criteria for the inherent success of Program 1, whereas Theorem 2 characterizes the conditions for the success of Program 1 as a function of the regularization parameter λ. We illustrate these results in Figures 1a and 1b.

3.1. Sharp Performance Bounds

From our forward and converse results, we see that there is a gap between Λ_{fail} and Λ_{succ}. The gap is $\frac{\Lambda_{\text{fail}}}{\Lambda_{\text{succ}}} = \frac{4\sqrt{q(1-q)n+\gamma_{\text{succ}}}}{\sqrt{q(n-\gamma_{\text{fail}})}}$ times. In the small cluster regime where $\max_{1\leq i\leq K} n_i = o(n)$ and $\sum_{i=1}^K n_i^2 = o(n^2)$, the ratio $\frac{\Lambda_{\text{fail}}}{\Lambda_{\text{succ}}}$ takes an extremely simple form as we have $\gamma_{\text{fail}} \ll n$ and $\gamma_{\text{succ}} \ll \sqrt{n}$. In particular, $\frac{\Lambda_{\text{fail}}}{\Lambda_{\text{succ}}} = 4\sqrt{1-q} + o(1)$, which is at most 4 times in the worst case.

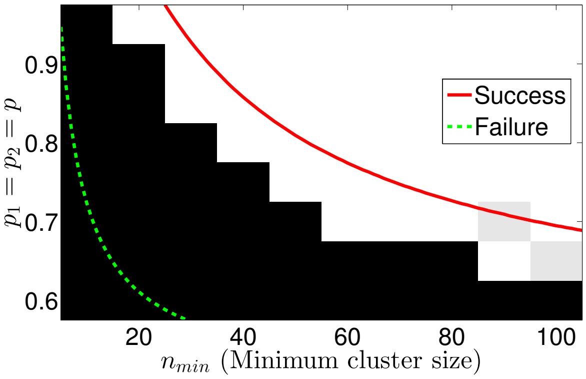


Fig. 2: Simulation results showing the region of success (white region) and failure (black region) of Program 1 with $\lambda = 0.99\Lambda_{\text{succ}}$. Also depicted are the thresholds for success (solid red curve on the top-right) and failure (dashed green curve on the bottom-left) predicted by Theorem 1.

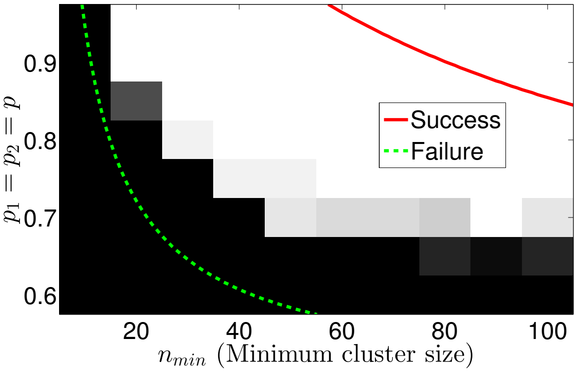


Fig. 3: Simulation results showing the region of success (white region) and failure (black region) of Program 1 with $\lambda = 2\mathbf{ED}_{\min}^{-1}$. Also depicted are the thresholds for success (solid red curve on the top-right) and failure (dashed green curve on the bottom-left) predicted by Theorem 2.

4. SIMULATIONS

We implement Program 1 using the inexact augmented Lagrangian multiplier method algorithm by Lin et al. [34]. We note that this algorithm solves the program approximately. Moreover, numerical imprecision prevents the output of the algorithm from being strictly 1 or 0. Hence we round each entry to 1 or 0 by comparing it with the mean of all entries of the output. In other words, if an entry is greater than the overall mean, we round it to 1 and to 0 otherwise. We declare success if the number of entries that are wrong in the rounded output compared to $\mathbf{L}^0$ (recall from (3)) is less than 0.1%.

We consider the set up with $n = 200$ nodes and two clusters of equal sizes, $n_1 = n_2$. We vary the cluster sizes from 10 to 100 in steps of 10. We fix $q = 0.1$ and vary the probability of edge inside clusters $p_1 = p_2 = p$ from 0.6 to 0.95 in steps of 0.05. We run the experiments 20 times and average over the outcomes. In the first set of experiments, we run the program with $\lambda = 0.99\Lambda_{succ}$ which ensures that $\lambda < \Lambda_{\text{succ}}$.

Figure 2 shows the region of success (white region) and failure (black region) for this experiment. From Theorem 1, we expect the program to succeed when $\mathbf{ED}_{\min} > \Lambda_{\text{succ}}^{-1}$, which is the region above the solid red curve in Figure 2, and fail when $\mathbf{ED}_{\min} < \Lambda_{\text{fail}}^{-1}$, which is the region below the dashed green curve in Figure 2.

In the second set of experiments, we run the program with $\lambda = \frac{2}{\mathbf{ED}_{\min}}$. This ensures that $\mathbf{ED}_{\min} > \frac{1}{\lambda}$. Figure 3 shows the region of success (white region) and failure (black region) for this experiment. From Theorem 2, we expect the program to succeed when $\lambda < \Lambda_{\text{succ}}$ which is the region above the solid red curve in Figure 3 and fail when $\lambda > \Lambda_{\text{fail}}$ which is the region below the dashed green curve in Figure 3.

We see that the transition indeed happens between the solid red curve and the dashed green curve in both Figure 2 and Figure 3 as predicted by Theorem 1 and Theorem 2 respectively.

5. DISCUSSION AND CONCLUSION

We provided sharp analysis of Program 1 which is commonly used to identify clusters in a graph and more generally, to decompose a matrix into low-rank and sparse components. We believe, our technique can be extended to tightly analyze variants of this approach. As a future work, we are looking at the extensions of Problem 1, where the adjacency matrix $\mathbf{A}$ is partially observed, and also modifying Program 1 for clustering weighted graphs, where the adjacency matrix $\mathbf{A}$ with $\{0, 1\}$-entries is replaced by a similarity matrix with real entries.

6. REFERENCES

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

[2] 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*, New York, NY, USA, 2001, KDD '01, pp. 57–66, ACM.

[3] Santo Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75 – 174, 2010.

[4] 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)*. August 1995, pp. 94–99, AAAI Press.

[5] Xiaowei Xu, Jochen Jäger, and Hans-Peter Kriegel, "A fast parallel clustering algorithm for large spatial databases," *Data Min. Knowl. Discov.*, vol. 3, no. 3, pp. 263–290, Sept. 1999.

[6] Ying Xu, Victor Olman, and Dong Xu, "Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees," *Bioinformatics*, vol. 18, no. 4, pp. 536–545, 2002.

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

[8] Satu Elisa Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27 – 64, 2007.

[9] Gary W. Flake, Robert E. Tarjan, and Kostas Tsioutsiouliklis, "Graph clustering and minimum cut trees," *Internet Mathematics*, vol. 1, no. 4, pp. 385–408, 2003.

[10] Moses Charikar, Venkatesan Guruswami, and Anthony Wirth, "Clustering with qualitative information.," *J. Comput. Syst. Sci.*, vol. 71, no. 3, pp. 360–383, 2005.

[11] Joachim Giesen and Dieter Mitsche, "Reconstructing many partitions using spectral techniques.," in *FCT*, Maciej Liskiewicz and Rdiger Reischuk, Eds. 2005, vol. 3623 of *Lecture Notes in Computer Science*, pp. 433–444, Springer.

[12] Dotan Emanuel and Amos Fiat, "Correlation clustering - minimizing disagreements on arbitrary weighted graphs.," in *ESA*, Giuseppe Di Battista and Uri Zwick, Eds. 2003, vol. 2832 of *Lecture Notes in Computer Science*, pp. 208–220, Springer.

[13] Nikhil Bansal, Avrim Blum, and Shuchi Chawla, "Correlation clustering," *Machine Learning*, vol. 56, no. 1-3, pp. 89–113, 2004.

[14] Ioannis Giotis and Venkatesan Guruswami, "Correlation clustering with a fixed number of clusters," *CoRR*, vol. abs/cs/0504023, 2005.

[15] Erik D. Demaine, Dotan Emanuel, Amos Fiat, and Nicole Immorlica, "Correlation clustering in general weighted graphs," *Theoretical Computer Science*, 2006.

[16] Anne Condon and Richard M. Karp, "Algorithms for graph partitioning on the planted partition model.," *Random Struct. Algorithms*, vol. 18, no. 2, pp. 116–140, 2001.

[17] Frank McSherry, "Spectral partitioning of random graphs.," in *FOCS*. 2001, pp. 529–537, IEEE Computer Society.

[18] B. Bollobás and A. D. Scott, "Max cut for random graphs with a planted partition," *Comb. Probab. Comput.*, vol. 13, no. 4-5, pp. 451–474, July 2004.

[19] R.R. Nadakuditi, "On hard limits of eigen-analysis based planted clique detection," in *Statistical Signal Processing Workshop (SSP), 2012 IEEE*, 2012, pp. 129–132.

[20] Emmanuel J. Candes and Justin Romberg, "Quantitative robust uncertainty principles and optimally sparse decompositions," *Found. Comput. Math.*, vol. 6, no. 2, pp. 227–254, Apr. 2006.

[21] Emmanuel J. Candes and Benjamin Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, Dec. 2009.

[22] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, June 2011.

[23] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky, "Rank-sparsity incoherence for matrix decomposition.," *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

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

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

[26] Brendan P. W. Ames and Stephen A. Vavasis, "Convex optimization for the planted k-disjoint-clique problem," *CoRR*, vol. abs/1008.2814, 2010.

[27] Brendan P. W. Ames and Stephen A. Vavasis, "Nuclear norm minimization for the planted clique and biclique problems," *Math. Program.*, vol. 129, no. 1, pp. 69–89, Sept. 2011.

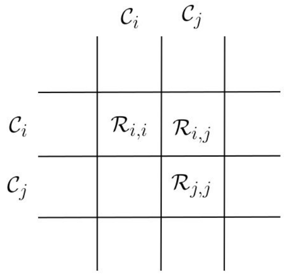


Fig. 4: Illustration of $\{\mathcal{R}_{i,j}\}$ dividing $[n] \times [n]$ into disjoint regions similar to a grid.

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

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

[30] Yudong Chen, Sujay Sanghavi, and Huan Xu, "Clustering sparse graphs.," in *NIPS*, Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Lon Bottou, and Kilian Q. Weinberger, Eds., 2012, pp. 2213–2221.

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

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

[33] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt, "Stochastic blockmodels: First steps," *Social Networks*, vol. 5, no. 2, pp. 109 – 137, 1983.

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

[35] Van H. Vu, "Spectral norm of random matrices," in *STOC*, Harold N. Gabow and Ronald Fagin, Eds. 2005, pp. 423–430, ACM.

7. PROOFS FOR SUCCESS

The theorems in Section 3 provide the conditions under which Program 1 succeeds or fails. In this section, we provide the proofs of the success results, i.e., the last statements of Theorems 1 and 2. The failure results will be the topic of Section 8.
Notation: Before we proceed, we need some additional notation. $\mathbb{1}^n$ will denote a vector in $\mathbb{R}^n$ with all ones. Complement of a set S will be denoted by S^c. 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 Figure 4. Thus, $\mathcal{R}_{i,i}$ is the region induced by i'th cluster for any $i \leq K$.

Let $\mathcal{A} \subseteq [n] \times [n]$ be the set of nonzero coordinates of $\mathbf{A}$. Then the sets,

1. $\mathcal{A} \cap \mathcal{R}$ corresponds to the edges inside the clusters.

2. $\mathcal{A}^c \cap \mathcal{R}$ corresponds to the missing edges inside the clusters.

3. $\mathcal{A} \cap \mathcal{R}^c$ corresponds to the set of edges outside the clusters, which should be ideally not present.

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, $\mathbb{1}_{\mathcal{A}}^{n \times n} = \mathbf{A}$. Given a matrix $\mathbf{A}$, sum($\mathbf{A}$) will denote the sum of all entries of $\mathbf{A}$. Finally, we introduce 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 $\mathcal{R}^c$. Given q, $\{p_i\}_{i=1}^K$, let,

$$\begin{aligned} \mathbf{D}_{\mathcal{A}} &= \frac{1}{2} \min\{1 - 2q, \{2p_i - 1 - \frac{1}{\lambda n_i}\}_{i=1}^{K}\} \\ &= \frac{1}{2} \min\{1 - 2q, \frac{\mathbf{ED}_i - \lambda^{-1}}{n_i}\} \end{aligned} \tag{4}$$

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, $f(n) \leq cn$.

Observe that the success condition of Theorem 1 is a special case of that of Theorem 2. Considering Theorem 1, suppose $\mathbf{ED}_{\min} \geq (1+\epsilon)\Lambda_{succ}^{-1}$ and $\lambda = (1-\delta)\Lambda_{succ}$ where $\delta > 0$ is to be determined. Choose δ so that $1 - \delta = (1+\epsilon)^{-1/2}$. Now, considering Theorem 2, we already have, $\lambda \leq (1-\delta)\Lambda_{succ}$ and we also satisfy the second requirement as we have $\mathbf{ED}_{\min} \geq (1+\epsilon)\Lambda_{succ}^{-1} = (1+\epsilon)(1-\delta)\lambda^{-1} = \sqrt{1+\epsilon}\lambda^{-1}$. Consequently, we will only prove Theorem 2 and we will assume that there exists a constant $\epsilon > 0$ such that,

$$\begin{aligned} \lambda &\leq (1-\epsilon)\Lambda_{succ} \\ \mathbf{ED}_{\min} &\geq (1+\epsilon)\lambda^{-1} \end{aligned} \tag{5}$$

This implies that $\mathbf{D}_{\mathcal{A}}$ is lower bounded by a positive constant. The reason is $p_{min} > 1/2$ hence $2p_i - 1 > 0$ and we additionally have that $2p_i - 1 \geq (1+\epsilon)\frac{1}{\lambda n_i}$. Together, these ensure, $2p_i - 1 - \frac{1}{\lambda n_i} \geq \frac{\epsilon}{1+\epsilon}(2p_i - 1)$.

7.1. Conditions for Success

In order to show that $(\mathbf{L}^0, \mathbf{S}^0)$ is the unique optimal solution to the program (1), we need to prove that the objective function strictly increases for any perturbation, i.e.,

$$(\|\mathbf{L}^0 + \mathbf{E}^L\|_\star + \lambda \, \|\mathbf{S}^0 + \mathbf{E}^S\|_1) - (\|\mathbf{L}^0\|_\star + \lambda \, \|\mathbf{S}^0\|_1) > 0, \tag{6}$$

for all feasible perturbations $(\mathbf{E}^L, \mathbf{E}^S)$.

For the following discussion, we will use a slightly abused notation where we denote a subgradient of a norm $\|\cdot\|_*$ at the point $\mathbf{x}$ by $\partial\|\mathbf{x}\|_*$. In the standard notation, $\partial\|\mathbf{x}\|_*$ denotes the set of all subgradients, i.e., the subdifferential.

We can lower bound the LHS of the equation (6) using the subgradients as follows,

$$\begin{aligned}(\|\mathbf{L}^0 + \mathbf{E}^L\|_\star + &\lambda \, \|\mathbf{S}^0 + \mathbf{E}^S\|_1) - \left(\|\mathbf{L}^0\|_\star + \lambda \, \|\mathbf{S}^0\|_1\right) \\ &\geq \langle \partial\|\mathbf{L}^0\|_\star, \mathbf{E}^L\rangle + \lambda\langle \partial\|\mathbf{S}^0\|_1, \mathbf{E}^S\rangle, \end{aligned} \tag{7}$$

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 $\left(\mathbf{L}^0, \mathbf{S}^0\right)$.

To make use of (7), it is crucial to choose good subgradients. Our efforts will now focus on construction of such subgradients.

7.1.1. Subgradient construction

Write $\mathbf{L}^0 = \mathbf{U}\Lambda\mathbf{U}^T$, 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

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

Then the subgradient $\partial\|\mathbf{L}^0\|_\star$ is of the form $\mathbf{U}\mathbf{U}^T + \mathbf{W}$ such that $\mathbf{W} \in \mathcal{M}_U := \{\mathbf{X} : \mathbf{X}\mathbf{U} = \mathbf{U}^T\mathbf{X} = 0, \|\mathbf{X}\| \leq 1\}$. 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 $\|\mathbf{Q}\|_\infty \leq 1$. We note that since $\mathbf{L} + \mathbf{S} = \mathbf{A}$, $\mathbf{E}^L = -\mathbf{E}^S$. Note that $\text{sign}(\mathbf{S}^0) = \mathbb{1}^{n\times n}_{\mathcal{A}\cap\mathcal{R}^c} - \mathbb{1}^{n\times n}_{\mathcal{A}^c\cap\mathcal{R}}$. Choosing $\mathbf{Q} = \mathbb{1}^{n\times n}_{\mathcal{A}\cap\mathcal{R}} - \mathbb{1}^{n\times n}_{\mathcal{A}^c\cap\mathcal{R}^c}$, we get,

$$\begin{aligned} \|\mathbf{L}^0 + &\mathbf{E}^L\|_\star + \lambda \, \|\mathbf{S}^0 + \mathbf{E}^S\|_1 - (\|\mathbf{L}^0\|_\star + \lambda \, \|\mathbf{S}^0\|_1) \\ &\geq \langle \partial\|\mathbf{L}^0\|_\star, \mathbf{E}^L\rangle + \lambda\langle \partial\|\mathbf{S}^0\|_1, \mathbf{E}^S\rangle \\ &= \langle \mathbf{U}\mathbf{U}^T + \mathbf{W}, \mathbf{E}^L\rangle + \lambda\langle \text{sign}(\mathbf{S}^0) + \mathbf{Q}, \mathbf{E}^S\rangle \\ &= \underbrace{\sum_{i=1}^K \frac{1}{n_i}\text{sum}(\mathbf{E}_{R_{i,i}}) + \lambda\left(\text{sum}(\mathbf{E}^L_{\mathcal{A}^c}) - \text{sum}(\mathbf{E}^L_{\mathcal{A}})\right)}_{:=g(\mathbf{E}^L)} \\ &\qquad + \left\langle \mathbf{W}, \mathbf{E}^L \right\rangle. \end{aligned} \tag{8}$$

Define,

$$g(\mathbf{E}^L) := \sum_{i=1}^K \frac{1}{n_i}\text{sum}(\mathbf{E}^L_{\mathcal{R}_{i,i}}) + \lambda\left(\text{sum}(\mathbf{E}^L_{\mathcal{A}^c}) - \text{sum}(\mathbf{E}^L_{\mathcal{A}})\right). \tag{9}$$

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

$$f\left(\mathbf{E}^L, \mathbf{W}\right) = g(\mathbf{E}^L) + \left\langle \mathbf{W}, \mathbf{E}^L\right\rangle > 0. \tag{10}$$

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

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

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

Proof. Let $c = 1 - \|\mathbf{W}\|$. Assume $\left\langle \mathbf{E}^L, \mathbf{W}'\right\rangle \neq 0$ for some $\mathbf{W}' \in \mathcal{M}_{\mathbf{U}}$. If $\left\langle \mathbf{E}^L, \mathbf{W}'\right\rangle > 0$, 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,

$$\begin{aligned} f(\mathbf{E}^L, \mathbf{W}^*) &= f(\mathbf{E}^L, \mathbf{W}) + |\left\langle \mathbf{E}^L, c\mathbf{W}'\right\rangle| \\ &> f(\mathbf{E}^L, \mathbf{W}) \geq 0 \end{aligned} \tag{11}$$

■

Notice that, for all $\mathbf{W} \in \mathcal{M}_{\mathbf{U}}$, $\left\langle \mathbf{E}^L, \mathbf{W}\right\rangle = 0$ is equivalent to $\mathbf{E}^L \in \mathcal{M}^\perp_{\mathbf{U}}$ 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:

$$\begin{aligned} \mathcal{M}^\perp_{\mathbf{U}} = \{\mathbf{X} \in \mathbb{R}^{n\times n} : \mathbf{X} = \mathbf{U}\mathbf{M}^T + \mathbf{N}\mathbf{U}^T \\ \text{for some } \mathbf{M}, \mathbf{N} \in \mathbb{R}^{n\times K}\}. \end{aligned} \tag{12}$$

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

As a first step, in Section 7.2, 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, $f(\mathbf{E}^L, \mathbf{W}) \geq 0$ for all feasible $\mathbf{E}^L$. This $\mathbf{W}$ is called the dual certificate. Secondly, in Section 7.3, we will show that, under certain conditions, for all $\mathbf{E}^L \in \mathcal{M}^\perp_{\mathbf{U}}$ with high probability, $g(\mathbf{E}^L) > 0$. Finally, combining these two arguments, and using Lemma 1 we will conclude that $(\mathbf{L}^0, \mathbf{S}^0)$ is the unique optimal with high probability.

7.2. Showing existence of the dual certificate

Recall that

$$\begin{aligned} f(\mathbf{E}^L, \mathbf{W}) \;\; = \;\; &\sum_{i=1}^K \frac{1}{n_i}\text{sum}(\mathbf{E}^L_{\mathcal{R}_{i,i}}) + \left\langle \mathbf{E}^L, \mathbf{W}\right\rangle \\ &+\lambda\left(\text{sum}\left(\mathbf{E}^L_{\mathcal{A}^c}\right) - \text{sum}\left(\mathbf{E}^L_{\mathcal{A}}\right)\right) \end{aligned}$$

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

7.2.1. Candidate $\mathbf{W}_0$

Based on Program 1, we propose the following,

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

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

We now have to find a bound on the spectral norm of $\mathbf{W}_0$. Note that $\mathbf{W}_0$ is a random matrix where randomness is due to $\mathcal{A}$. 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 Stochastic Block Model 1, 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 $c_i + \lambda(2p_i - 1)$ and $c + \lambda(2q - 1)$ respectively. Hence, we set,

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

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

$$\begin{aligned}\mathbf{W}_0 \;=\;& 2\lambda\left[\sum_{i=1}^{K}(1-p_i)\;\mathbb{1}_{\mathcal{R}_{i,i}\cap\mathcal{A}}^{n\times n} - p_i\;\mathbb{1}_{\mathcal{R}_{i,i}\cap\mathcal{A}^c}^{n\times n}\right] \\ &+2\lambda\left[(1-q)\;\mathbb{1}_{\mathcal{R}^c\cap\mathcal{A}}^{n\times n} - q\;\mathbb{1}_{\mathcal{R}^c\cap\mathcal{A}^c}^{n\times n}\right] \qquad (13)\end{aligned}$$

$$\begin{aligned}f(\mathbf{E}^L, \mathbf{W}_0) \;=\;& \lambda\left[(1-2q)\;\text{sum}(\mathbf{E}_{\mathcal{R}^c}^L)\right] \\ &-\lambda\left[\sum_{i=1}^{K}\left(2p_i - 1 - \frac{1}{\lambda n_i}\right)\;\text{sum}(\mathbf{E}_{\mathcal{R}_{i,i}}^L)\right] \qquad (14)\end{aligned}$$

From $\mathbf{L}^0$ and (2), it follows that,

$$\begin{aligned}&\mathbf{E}_{\mathcal{R}^c}^L \text{ is (entrywise) nonnegative.} \qquad (15)\\ &\mathbf{E}_{\mathcal{R}}^L \text{ is (entrywise) nonpositive.}\end{aligned}$$

Thus, $\text{sum}(\mathbf{E}_{\mathcal{R}^c}^L) \leq 0$ and $\text{sum}(\mathbf{E}_{\mathcal{R}_{i,i}}^L) \geq 0$. When $\lambda(2p_i - 1) - \frac{1}{n_i} \geq 0$ and $\lambda(2q-1) \leq 0$; we will have $f(\mathbf{E}^L, \mathbf{W}_0) \geq 0$ for all feasible $\mathbf{E}^L$. This indeed holds due to the assumptions of Theorem 1 (see (4)), as we assumed $2p_i - 1 > \frac{1}{\lambda n_i}$ for $i = 1, 2\cdots, K$ and $1 > 2q$.

We will now proceed to find a tight bound on the spectral norm of $\mathbf{W}^0$. Let us define the zero-mean Bernoulli distribution $\text{Bern}_0(\alpha)$ as follows. $X \sim \text{Bern}_0(\alpha)$ if,

$$X = \begin{cases} 1-\alpha & w.p. \;\; \alpha \\ -\alpha & w.p. \;\; 1-\alpha \end{cases}$$

Theorem 3. *Assume $\mathbf{A} \in \mathbb{R}^{n\times n}$ obeys the stochastic block model (1) and let $\mathbf{M} \in \mathbb{R}^{n\times n}$. Let entries of $\mathbf{M}$ be as follows.*

$$\mathbf{M}_{i,j} \sim \begin{cases} Bern_0(p_k) & if \;\; (i,j) \in \mathcal{R}_{k,k} \\ Bern_0(q) & if \;\; (i,j) \in \mathcal{R}^c \end{cases}$$

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

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

Proof. The entries of $\mathbf{M}$ are i.i.d. with maximum variance of $1/4$. Hence, the first statement follows directly from [35].

For the second statement, let,

$$\mathbf{M}_1(i,j) = \begin{cases} \mathbf{M}(i,j) & \text{if } i,j \in \mathbb{R}^c \\ \text{Bern}_0(q) \text{ else} \end{cases}$$

Also let $\mathbf{M}_2 = \mathbf{M} - \mathbf{M}_1$. Observe that, $\mathbf{M}_1$ has i.i.d. $\text{Bern}_0(q)$ entries. From standard results on random matrix theory, it follows that,

$$\|\mathbf{M}_1\| \leq (2\sqrt{q(1-q)} + \epsilon')\sqrt{n}$$

with the desired probability.

For $\mathbf{M}_2$, first observe that over $\mathcal{R}_{i,i}$ $\mathbf{M}_2$ has i.i.d. entries with variance $q(1-q) + p_i(1-p_i)$. This similarly gives,

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

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

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

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

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

Lemma 2. *Recall that, $\mathbf{W}_0$ is a random matrix; where randomness is on the stochastic block model $\mathcal{A}$ and it is given by,*

$$\begin{aligned}\mathbf{W}_0 = 2\lambda\sum_{i=1}^{K}&\left[(1-p_i)\mathbb{1}_{\mathcal{A}\cap\mathcal{R}_{i,i}}^{n\times n} - p_i\mathbb{1}_{\mathcal{A}^c\cap\mathcal{R}_{i,i}}^{n\times n}\right] \\ &+ 2\lambda\left[(1-q)\mathbb{1}_{\mathcal{A}\cap\mathcal{R}^c}^{n\times n} - q\mathbb{1}_{\mathcal{A}^c\cap\mathcal{R}^c}^{n\times n}\right] \qquad (16)\end{aligned}$$

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

$$\begin{aligned}\|\mathbf{W}_0\| &\leq 4\lambda\sqrt{q(1-q)}\sqrt{n} \\ &\quad + \max_{i\leq K}\; 4\lambda\sqrt{q(1-q) + p_i(1-p_i)}\sqrt{n_i} + \epsilon'\lambda\sqrt{n} \\ &\leq \lambda\Lambda_{succ}^{-1} + \epsilon'\lambda\sqrt{n}\end{aligned}$$

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

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

Proof. $\frac{1}{2\lambda}\mathbf{W}_0$ is a random matrix whose entries are i.i.d. and distributed as $\text{Bern}_0(p_i)$ on $\mathcal{R}_{i,i}$ and $\text{Bern}_0(q)$ on $\mathcal{R}^c$. Consequently, using Theorem 3 and recalling the definition of Λ_{succ} we obtain the result. ■

Lemma 2 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.

7.2.2. Correcting the candidate $\mathbf{W}_0$

Lemma 3. $\mathbf{W}_0$ *is as described previously in* (16)*. Let* $\mathbf{W}^H$ *be the projection of* $\mathbf{W}_0$ *on* $\mathcal{M}_{\mathbf{U}}$*. Then*

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

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

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 $\mathbf{U}_2 = [\mathbf{u}_{K+1} \ \dots \ \mathbf{u}_n]$ 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,

$$\begin{aligned}\mathbf{X} - \mathbf{P}_2\mathbf{X}\mathbf{P}_2 &= \mathbf{X} - (\mathbf{I}-\mathbf{P})\mathbf{X}(\mathbf{I}-\mathbf{P})\\ &= \mathbf{P}\mathbf{X} + \mathbf{X}\mathbf{P} - \mathbf{P}\mathbf{X}\mathbf{P} \in \mathcal{M}_{\mathbf{U}}^{\perp} \qquad (17)\end{aligned}$$

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

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

For analysis of $\|\mathbf{W}_0 - \mathbf{W}^H\|_\infty$ we can consider terms on the right hand side of (17) separately as we have:

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

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. $\text{Bern}_0(p_i)$ or $\text{Bern}_0(q)$ random variables scaled by n_i^{-1} for some $i \le K$ or 0. Hence any $c, d \in [n]$ and $\epsilon'' > 0$

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

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

Recall that,
Let $\gamma_{\text{succ}} := \max_{1\le i\le K} 4\sqrt{(q(1-q) + p_i(1-p_i))n_i}$, and
$\Lambda_{\text{succ}} := \frac{1}{4\sqrt{q(1-q)n} + \gamma_{\text{succ}}}$.

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

Lemma 4. $\mathbf{W}_0$ *is as described previously in* (13)*. Choose* $\mathbf{W}$ *to be projection of* $\mathbf{W}_0$ *on* $\mathcal{M}_{\mathbf{U}}$*. Also suppose* $\lambda \le (1-\delta)\Lambda_{succ}$*. Then, with probability* $1 - 6n^2\exp(-\Omega(n_{min})) - 4\exp(-\Omega(n))$ *we have,*

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

Proof. To begin with, observe that Λ_{succ}^{-1} is $\Omega(\sqrt{n})$. Since $\lambda \le \Lambda_{succ}$, $\lambda\sqrt{n} = \mathcal{O}(1)$. Consequently, using $\lambda\Lambda_{succ}^{-1} < 1$ and applying Lemma 2, 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 3 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 (14) for all $\mathbf{E}^L$, we have that,

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

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

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

Summary so far: Combining the last lemma with Lemma 1, with high probability, either there exists a dual vector $\mathbf{W}^*$ which ensures $f(\mathbf{E}^L, \mathbf{W}^*) > 0$ or $\mathbf{E}^L \in \mathcal{M}_{\mathbf{U}}^{\perp}$. If former, we are done. Hence, we need to focus on the latter case and show that for all perturbations $\mathbf{E}^L \in \mathcal{M}_{\mathbf{U}}^{\perp}$, the objective will strictly increase at $(\mathbf{L}^0, \mathbf{S}^0)$ with high probability.

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

Recall that,

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

Let us define,

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

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

so that, $g\left(\mathbf{X}\right) = g_1(\mathbf{X}) + \lambda g_2(\mathbf{X})$. 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 $\mathbf{E}^L \in \mathcal{M}_{\mathbf{U}}^{\perp}$. Then, by definition of $\mathcal{M}_{\mathbf{U}}^{\perp}$, we can write,

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

Let $\mathbf{m}_i, \mathbf{n}_i$ denote i'th columns of $\mathbf{M}, \mathbf{N}$ respectively. From $\mathbf{L}^0$ and (2) it follows that

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

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

$$\mathbf{E}^L = \sum_{i=1}^{K} (\mathbf{v}_i \mathbf{m}_i^T + \mathbf{n}_i \mathbf{v}_i^T) = \sum_{i=1}^{K+1} \sum_{j=1}^{K+1} \mathbf{E}^L_{\mathcal{R}_{i,j}} \tag{18}$$

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

Let $\{a_{i,j}\}_{j=1}^{n_i}$ be an (arbitrary) indexing of elements of $\mathcal{C}_i$ i.e. $\mathcal{C}_i = \{a_{i,1}, \dots, a_{i,n_i}\}$. For a vector $\mathbf{z} \in \mathbb{R}^n$, let $\mathbf{z}^i \in \mathbb{R}^{n_i}$ denote the vector induced by entries of $\mathbf{z}$ in $\mathcal{C}_i$. Basically, for any $1 \le j \le n_i$, $\mathbf{z}^i_j = \mathbf{z}_{a_{i,j}}$. Also, let $\mathbf{E}^{i,j} \in \mathbb{R}^{n_i \times n_j}$ which is $\mathbf{E}^L$ induced by entries on $\mathcal{R}_{i,j}$.

In other words,

$$\begin{aligned} \mathbf{E}^{i,j}_{c,d} = \mathbf{E}^L_{a_{i,c}, a_{j,d}} \quad &\text{for all } (i,j) \in \mathcal{C}_i \times \mathcal{C}_j \text{ and} \\ &\text{all } 1 \le c \le n_i,\ 1 \le d \le n_j \end{aligned}$$

Basically, $\mathbf{E}^{i,j}$ is same as $\mathbf{E}^L_{\mathcal{R}_{i,j}}$ when we get rid of trivial zero rows and zero columns. Then

$$\mathbf{E}^{i,j} = \mathbb{1}^{n_i} \mathbf{m}_i^{j\,T} + \mathbf{n}^i_j \mathbb{1}^{n_j\,T} \tag{19}$$

Clearly, given $\{\mathbf{E}^{i,j}\}_{1 \le i,j \le n}$, $\mathbf{E}^L$ 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* $\mathbf{E}^L$ subject to these constraints. Variables in such an optimization are $\mathbf{m}_i, \mathbf{n}_i$. Basically we are interested in,

$$\begin{aligned} &\min g(\mathbf{E}^L) \\ &\text{subject to} \\ &\quad \text{sum}(\mathbf{E}^{i,j}) = c_{i,j} \text{ for all } i, j \\ &\quad \mathbf{E}^{i,j} \begin{cases} \text{nonnegative if } i \neq j \\ \text{nonpositive if } i = j \end{cases} \end{aligned} \tag{20, 21}$$

where $\{c_{i,j}\}$ are constants. Constraint (21) follows from (15). **Remark:** For the special case of $i = j = K+1$, notice that $\mathbf{E}^{i,j} = 0$.

In (20), $g_1(\mathbf{E}^L)$ 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 $g_2(\mathbf{E}^L) = \text{sum}(\mathbf{E}^L_{\mathcal{A}^c}) - \text{sum}(\mathbf{E}^L_{\mathcal{A}})$.

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 $(i_1, j_1) \neq (i_2, j_2)$, $(\mathbf{m}^{j_1}_{i_1}, \mathbf{n}^{i_1}_{j_1})$ and $(\mathbf{m}^{j_2}_{i_2}, \mathbf{n}^{i_2}_{j_2})$ are independent variables. Consequently, due to (19), we can partition problem (20) into the following smaller disjoint problems.

$$\begin{aligned} &\min_{\mathbf{m}_i^j, \mathbf{n}_j^i} \ \text{sum}(\mathbf{E}^{i,j}_{\beta^c_{i,j}}) - \text{sum}(\mathbf{E}^{i,j}_{\beta_{i,j}}) \\ &\text{subject to} \\ &\quad \text{sum}(\mathbf{E}^{i,j}) = c_{i,j} \\ &\quad \mathbf{E}^{i,j} \text{ is } \begin{cases} \text{nonnegative if } i \neq j \\ \text{nonpositive if } i = j \end{cases} \end{aligned} \tag{22}$$

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 (20), where $*$ denotes the optimal solutions in problems (20) and (22). The following lemma will be useful for analysis of these local optimizations.

Lemma 5. *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* $C_0 \ge 0$ *be a constant. Also let* $\beta \subseteq [c] \times [d]$. *Consider the following optimization problem*

$$\begin{aligned} &\min_{\mathbf{a}, \mathbf{b}} \ \mathit{sum}(\mathbf{X}_{\beta^c}) - \mathit{sum}(\mathbf{X}_{\beta}) \\ &\textit{subject to} \\ &\quad \mathbf{X}_{i,j} \ge 0 \textit{ for all } i, j \\ &\quad \mathit{sum}(\mathbf{X}) = C_0 \end{aligned}$$

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:

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

which implies,

$$\begin{aligned}\text{sum}(\mathbf{X}^*_\beta) - \text{sum}(\mathbf{X}^*_{\beta^c}) &= \text{sum}(\mathbf{X}^0_\beta) - \text{sum}(\mathbf{X}^0_{\beta^c}) \\ &= \text{optimal value}\end{aligned}$$

■

Lemma 6. *A direct consequence of Lemma 5 is the fact that in the local optimizations (22), Without loss of generality, we can assume $(\mathbf{m}_i^j, \mathbf{n}_j^i)$ entrywise nonnegative whenever $i \neq j$ and entrywise nonpositive when $i = j$. This follows from the structure of $\mathbf{E}^{i,j}$ given in (19) and (15).*

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

Lemma 7. *Let β be a random set generated by choosing elements of $[c] \times [d]$ indecently with probability $0 \leq r \leq 1$. 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:*

$$\mathit{sum}(\mathbf{X}_\beta) > (r - \epsilon')\mathit{sum}(\mathbf{X}) \tag{23}$$

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

Proof. We'll only prove the first statement (23) 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(r) random variables. Then

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

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

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

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

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

■

Using Lemma 7, we can calculate a lower bound for $g(\mathbf{E}^L)$ with high probability as long as the cluster sizes are sufficiently large. Due to (18) and the linearity of $g(\mathbf{E}^L)$, 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 6. In particular, subvectors $\mathbf{m}_i^i$ and $\mathbf{n}_i^i$ of $\mathbf{m}_i, \mathbf{n}_i$ can be assumed to be nonpositive and rest of the entries are nonnegative.

Lemma 8. *Assume, $l \leq K$, $\mathbf{D}_\mathcal{A} > 0$. Then, with probability $1 - n\exp(-2\mathbf{D}^2_\mathcal{A}(n_l - 1))$, 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 $\mathbf{m}_l^i$ is nonpositive/nonnegative when $i = l/i \neq l$ for all i. Call $\mathbf{X}^i = \mathbb{1}^{n_l} {\mathbf{m}_l^i}^T$. 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 7 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,

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

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,

$$\begin{aligned}&\frac{1}{\lambda n_l}\text{sum}(\mathbf{X}^l) + h(\mathbf{X}^l, \beta^c_{l,l}) \geq \\ &\qquad \left(1 - p_l + \epsilon' + \frac{1}{\lambda n_l}\right)\text{sum}(\mathbf{X}^l) - (p_l - \epsilon')\,\text{sum}(\mathbf{X}^l) \\ &\qquad = -\left(2p_l - 1 - \frac{1}{\lambda n_l} - 2\epsilon'\right)\text{sum}(\mathbf{X}^l)\end{aligned} \tag{24}$$

Choosing $\epsilon' = \frac{\mathbf{D}_\mathcal{A}}{2}$ and using the facts that $1 - 2q - 2\mathbf{D}_\mathcal{A} \geq 0$, $2p_l - 1 - \frac{1}{\lambda n_l} - 2\mathbf{D}_\mathcal{A} \geq 0$ and using a union bound, with probability $1 - n\exp(-2\mathbf{D}^2_\mathcal{A}(n_l - 1))$, 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 8 and summarizes the main result of the section.

Lemma 9. *Let $\mathbf{D}_\mathcal{A}$ be as defined in (4) and assume $\mathbf{D}_\mathcal{A} > 0$. Then with probability $1 - 2nK\exp(-2\mathbf{D}^2_\mathcal{A}(n_{min} - 1))$ we have $g(\mathbf{E}^L) > 0$ for all nonzero feasible $\mathbf{E}^L \in \mathcal{M}_U^\perp$.*

7.4. The Final Step

Lemma 10. *Let $p_{min} > \frac{1}{2} > q$ and $\mathcal{G}$ be a random graph generated according to Model 1 with cluster sizes $\{n_i\}_{i=1}^K$. If $\lambda \leq (1 - \epsilon)\Lambda_{succ}$ and $\mathbf{ED}_{\min} = \min_{1\leq i\leq n}(2p_i - 1)\,n_i \geq (1 + \epsilon)\frac{1}{\lambda}$, then $(\mathbf{L}^0, \mathbf{S}^0)$ is the unique optimal solution to Program 1 with probability $1 - \exp(-\Omega(n)) - 6n^2\exp(-\Omega(n_{min}))$.*

Proof. Based on Lemma 4 and Lemma 9, with probability $1 - cn^2 \exp(-C\left(\min\{1-2q, 2p_{min}-1\}\right)^2 n_{min})$,

- 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 1, $(\mathbf{L}^0, \mathbf{S}^0)$ is the unique optimal of Problem 1. ∎

8. PROOFS FOR FAILURE

This section will provide the proofs of the failure results, i.e., the initial statements of Theorems 1 and 2. Let us start by arguing that, failure result of Theorem 2 implies failure result of Theorem 1. To see this, assume Theorem 2 holds and $\mathbf{ED}_{\min} \leq (1-\epsilon)\Lambda_{fail}^{-1}$. Let ϵ' be a constant to be determined. If $\lambda \geq (1+\epsilon')\Lambda_{fail}$ or $\mathbf{ED}_{\min} \leq (1-\epsilon')\lambda^{-1}$, due to Theorem 2, Program 1 would fail and we can conclude. Suppose, these are not the case, i.e., $\lambda \leq (1+\epsilon')\Lambda_{fail}$ and $\mathbf{ED}_{\min} \geq (1-\epsilon')\lambda^{-1}$. These would imply, $\mathbf{ED}_{\min} \geq \frac{1-\epsilon'}{1+\epsilon'}\Lambda_{fail}^{-1}$. We can end up with a contradiction by choosing ϵ' small enough to ensure $\frac{1-\epsilon'}{1+\epsilon'} > 1-\epsilon$. Consequntly, we will only prove Theorem 2.

Lemma 11. *Let $p_{min} > \frac{1}{2} > q$ and $\mathcal{G}$ be a random graph generated according to the Model 1 with cluster sizes $\{n_i\}_{i=1}^K$.*

1. *If $\min_i \{n_i(2p_i - 1)\} \leq (1-\epsilon)\frac{1}{\lambda}$, then $(\mathbf{L}^0, \mathbf{S}^0)$ is not an optimal solution to the Program 1 with probability at least $1 - K\exp\left(-\Omega(n_{\min}^2)\right)$.*
2. *If $\lambda \geq (1+\epsilon)\sqrt{\frac{n}{q\left(n^2 - \sum_{i=1}^K n_i^2\right)}}$, then $(\mathbf{L}^0, \mathbf{S}^0)$ is not an optimal solution to the Program 1 with high probability.*

Proof. **Proof of the first statement:** Choose ϵ' to be a constant satisfying $2p_i - 1 + \epsilon' < \frac{1}{\lambda n_i}$ for some $1 \leq i \leq K$. This is indeed possible if the assumption of the Statement 1 of Lemma 11 holds. Lagrange for the Problem 1 can be written as follows,

$$\mathscr{L}(\mathbf{L}, \mathbf{S}; \mathbf{M}, \mathbf{N}) = \|\mathbf{L}\|_\star + \lambda\|\mathbf{S}\|_1 + \text{trace}(\mathbf{M}(\mathbf{L} - \mathbb{1}^{n\times n})) - \text{trace}(\mathbf{NL}). \tag{25}$$

where $\mathbf{M}$ and $\mathbf{N}$ are dual variables corresponding to the inequality constraints (2).

For $\mathbf{L}^0$ to be an optimal solution to (1), it has to satisfy the KKT conditions. Therefore, the subgradient of (25) at $\mathbf{L}^0$ has to be 0, i.e.,

$$\partial\|\mathbf{L}^0\|_\star + \lambda\,\partial\|\mathbf{A} - \mathbf{L}^0\|_1 + \mathbf{M}^0 - \mathbf{N}^0 = 0. \tag{26}$$

where $\mathbf{M}^0$ and $\mathbf{N}^0$ are optimal dual variables.

Also, by complementary slackness,

$$\text{trace}(\mathbf{M}^0(\mathbf{L}^0 - \mathbb{1}^{n\times n})) = 0, \tag{27}$$

and

$$\text{trace}(\mathbf{N}^0\mathbf{L}^0) = 0. \tag{28}$$

From (3), (27), and (28), 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 $(\mathbf{M}^0 - \mathbf{N}^0)_{\mathcal{R}^c} \leq 0$.

Recall, $\mathbf{L}^0 = \mathbf{U}\Lambda\mathbf{U}^T$, where $\mathbf{U} = [\mathbf{u}_1 \ \ldots \ \mathbf{u}_K] \in \mathbb{R}^{n\times K}$,

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

Also, recall that the subgradient $\partial\|\mathbf{L}^0\|_\star$ is of the form $\mathbf{U}\mathbf{U}^T + \mathbf{W}$ such that $\mathbf{W} \in \{\mathbf{X} : \mathbf{XU} = \mathbf{U}^T\mathbf{X} = 0, \|\mathbf{X}\| \leq 1\}$. 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 $\|\mathbf{Q}\|_\infty \leq 1$.

From (26), we have,

$$\mathbf{U}\mathbf{U}^T + \mathbf{W} - \lambda\left(\text{sign}(\mathbf{S}^0) + \mathbf{Q}\right) + (\mathbf{M}^0 - \mathbf{N}^0) = 0. \tag{29}$$

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

$$\underbrace{\text{sum}\left(\mathbf{L}^0_{\mathcal{R}_{i,i}}\right)}_{n_i} - \text{sum}\left(\lambda\left(\text{sign}(\mathbf{S}^0) + \mathbf{Q}\right)_{\mathcal{R}_{i,i}}\right) + \underbrace{\text{sum}\left((\mathbf{M}^0 - \mathbf{N}^0)_{\mathcal{R}_{i,i}}\right)}_{\geq 0} = 0. \tag{30}$$

By Bernstein's inequality and using $\|\mathbf{Q}\|_\infty \leq 1$, with probability $1 - \exp\left(-\Omega(n_i^2)\right)$ we have,

$$\text{sum}\left(\text{sign}(\mathbf{S}^0)\right) \leq -n_i^2(1 - p_i - \frac{\epsilon'}{2}) \tag{31}$$

$$\text{sum}\left(\mathbf{Q}\right) \leq n_i^2(p_i + \frac{\epsilon'}{2}). \tag{32}$$

Thus, $-\text{sum}\left(\lambda\left(\text{sign}(\mathbf{S}^0) + \mathbf{Q}\right)_{\mathcal{R}_{i,i}}\right) \geq \lambda n_i^2(1 - 2p_i - \epsilon')$ and hence,

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

Now, choose $i = \arg\min_{1\leq j\leq K} n_i(2p_i - 1)$. From the initial choice of ϵ', we have that $n_i + \lambda n_i^2(1 - 2p_i - \epsilon') > 0$. Consequently, the equation (26) does not hold and hence $\mathbf{L}^0$ cannot be an optimal solution to the Program 1.

Proof of the second statement: Let ϵ' be a constant to be determined. 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 $\mathcal{R}^c \cap \mathcal{A}$ are nonpositive. Hence from (29),

$$\begin{aligned}\|\mathbf{W}\|_F^2 &\geq \| \left(\mathbf{U}\mathbf{U}^T+\mathbf{W}\right)_{\mathcal{R}^c\cap\mathcal{A}} \|_F^2 \\ &\geq \|\lambda \left(\text{sign}(\mathbf{S}^0)+\mathbf{Q}\right)_{\mathcal{R}^c\cap\mathcal{A}} \|_F^2. \end{aligned} \tag{33}$$

Recall that $\mathbf{S}^0_{\mathcal{R}^c\cap\mathcal{A}} \neq 0$ and hence $\mathbf{Q}_{\mathcal{R}^c\cap\mathcal{A}} = 0$. Further, recall that by Model 1, each entry of $\mathbf{A}$ over $\mathcal{R}^c$ is non-zero with probability q. Hence with probability at least $1-\exp\left(-\Omega(|\mathcal{R}^c|)\right)$, $|\mathcal{R}^c \cap \mathcal{A}| \geq (q-\epsilon')(n^2-\sum_{i=1}^K n_i^2)$. Thus from (33) we have,

$$\|\mathbf{W}\|_F^2 \geq \lambda^2(q-\epsilon')(n^2-\sum_{i=1}^K n_i^2), \tag{34}$$

Recall that $\|\mathbf{W}\| \leq 1$ should hold true for $\left(\mathbf{L}^0,\mathbf{S}^0\right)$ to be an optimal solution to the Program 1. Using the standard inequality $n\|\mathbf{W}\|^2 \geq \|\mathbf{W}\|_F^2$ and the equation (34), we find,

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

So, if $\lambda\sqrt{q(1-\epsilon')\left(n^2-\sum_{i=1}^K n_i^2\right)/n} > 1$ then, $\left(\mathbf{L}^0,\mathbf{S}^0\right)$ cannot be an optimal solution to Program 1. This is indeed the case with the choice $(1-\epsilon')^{-1/2} < (1+\epsilon)$. This gives us the Statement 2 of Lemma 11. ■