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 and so on. 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] etc. 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].

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.

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

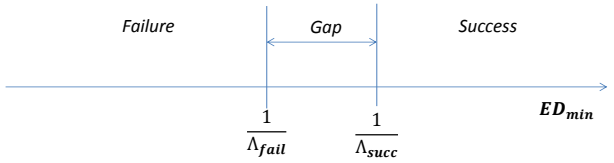
subject to

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

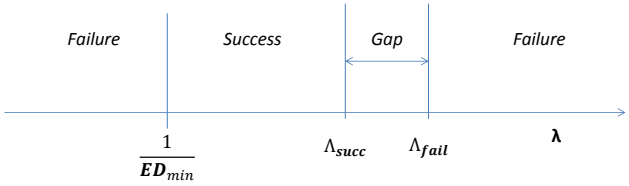
$$\mathbf{L} + \mathbf{S} = \mathbf{A} \tag{1.3}$$

where $\lambda > 0$ is a regularization parameter. $\|\mathbf{X}\|_\star$ and $\|\mathbf{X}\|_1$ denote the nuclear norm (maximum singular value) and the l_1-norm (sum of absolute values of all entries) respectively of the matrix $\mathbf{X}$. This program is very intuitive and requires the knowledge of only the adjacency matrix. Program 1.1 has been proposed in many works [28–30].

We consider the popular *stochastic block model* (also called the planted partition model) for the graph. Under



(a) Feasibility of Program 1.1 in terms of the minimum effective density (ED_{min}).



(b) Feasibility of Program 1.1 in terms of the regularization parameter (λ).

Fig. 1: Characterization of the feasibility of Program (1.1) 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 1.1 derived in this paper are fairly close to each other.

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.

this model of generating random graphs, the existence of an 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 Erdos-Renyi model. We characterize the conditions under which Program (1.1) 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 1.1 (as illustrated in Fig. 1a).

A second contribution of this paper is to explicitly characterize the efficacy of Program 1.1 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 1.1 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 1.1. 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 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 1.1 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 Appendix contains an outline of the proofs of the theoretical results presented in the paper.

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 $n_{min} := \min_{1 \le i \le K} n_i$.

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

Definition 2.1 (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 $i > j$:*

$$\mathbf{A}_{i,j} = \begin{cases} \textit{Bernoulli}(p_l) & \textit{if both } \{i,j\} \in \mathcal{C}_l \textit{ for some } l \le K \\ \textit{Bernoulli}(q) & \textit{otherwise.} \end{cases} \tag{2.1}$$

So, an edge inside i^{th} cluster exists with probability p_i and an edge outside clusters exists with probability q. Let $p_{min} := \min_{1 \le i \le K} p_i$. 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 1.1 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.

3. MAIN RESULTS

The desired solution to Program 1.1 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:

$$\mathbf{L}^0_{i,j} = \begin{cases} 1 & \text{if both } \{i,j\} \in \mathcal{C}_l \text{ for some } l \le K, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

$$\mathbf{S}^0_{i,j} = \begin{cases} -1 & \text{if both } \{i,j\} \in \mathcal{C}_l \text{ for some } l \le K, \text{ and } \mathbf{A}_{i,j} = 0, \\ 1 & \text{if } \{i,j\} \text{ are not in the same cluster and } \mathbf{A}_{i,j} = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

It is easy to see that the $(\mathbf{L}^0, \mathbf{S}^0)$ pair is feasible. We say that Program 1.1 *succeeds* when $(\mathbf{L}^0, \mathbf{S}^0)$ is the optimal solution to Program 1.1. In this section we present two theorems which give the conditions under which Program 1.1 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 $\mathbf{ED}_{\min} = \min_{1 \le i \le K} \mathbf{ED}_i$.

- Let $\gamma_{\text{succ}} := \max_{1\le i\le K} 4\sqrt{(q(1-q)+p_i(1-p_i))n_i}$, $\gamma_{\text{fail}} := \sum_{i=1}^{K} \frac{n_i^2}{n}$
- $\Lambda_{\text{fail}} := \frac{1}{\sqrt{q(n-\gamma_{\text{fail}})}}$ 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 2.1 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. *Whenever $\mathbf{ED}_{\min} \ge (1+\epsilon)\Lambda_{succ}^{-1}$, for $\lambda = (1-\delta)\Lambda_{succ}$, (1.1) succeeds with probability $1 - c_1 n^2 \exp(-c_2 n_{\min})$.*
2. *For any given $\lambda \ge 0$, if $\mathbf{ED}_{\min} \le (1-\epsilon)\Lambda_{fail}^{-1}$ then the (1.1) fails with probability $1 - c_1 \exp(-c_2|\mathcal{R}^c|)$.*

Theorem 2. *Let $\mathcal{G}$ be a random graph generated according to the stochastic block model 2.1 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 $\lambda \ge (1+\epsilon)\Lambda_{fail}$, then (1.1) fails with probability $1 - c_1' \exp(-c_2'|\mathcal{R}^c|)$.*
2. *If $\lambda \le (1-\epsilon)\Lambda_{succ}$ then,*
   - *If $\mathbf{ED}_{\min} \le (1-\epsilon)\frac{1}{\lambda}$, then (1.1) fails with probability $1 - c_1' \exp(-c_2' n_{\min})$.*
   - *If $\mathbf{ED}_{\min} \ge (1+\epsilon)\frac{1}{\lambda}$, then (1.1) succeeds with probability $1 - c_1' n^2 \exp(-c_2' n_{\min})$.*

We see that the minimum effective density $\mathbf{ED}_{\min}, \Lambda_{\text{succ}}$ and Λ_{fail} play a fundamental role in determining the success of Program 1.1. Theorem 1 gives criteria for the inherent success of Program 1.1 whereas theorem 2 characterizes the conditions for the success of Program 1.1 as a function of the regularization parameter λ. We illustrate these results in Figures 1a and 1b.

Sharp performance bounds: From the forward and converse, 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 $\max_{1\le i\le K} n_i = o(n)$ 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.

4. SIMULATIONS

We implement Program 1.1 using the inexact augmented Lagrangian multiplier method algorithm by Li et al. [34]. We note that this algorithm solves the program approximately.

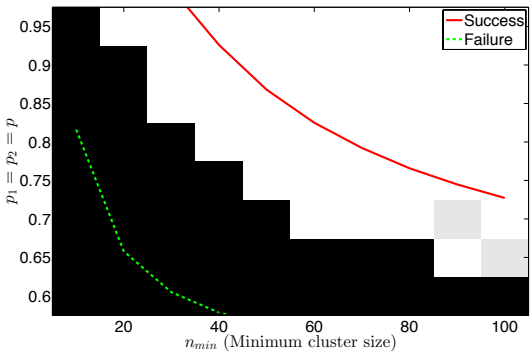


Fig. 2: Simulation results showing the region of success (white region) and failure (black region) of Program 1.1 with $\lambda = 0.99\Lambda_{\text{succ}}$. Also depicted are the thresholds for success (solid red curve on the top-right) and failure (dashed green curve on the bottom-left) predicted by Theorem 1.

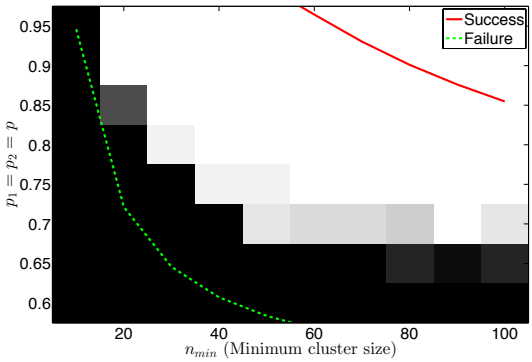


Fig. 3: Simulation results showing the region of success (white region) and failure (black region) of Program 1.1 with $\lambda = 2\mathbf{ED}_{\min}^{-1}$. Also depicted are the thresholds for success (solid red curve on the top-right) and failure (dashed green curve on the bottom-left) predicted by Theorem 2.

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 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 3.1) 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 0.5. We run the experiments 20 times and average the 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 1.1 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 1.1 where the adjacency matrix $\mathbf{A}$ is partially observed and also modifying Program 1.1 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.

[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*, Apr. 2011.

[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.

[36] R. Korlakai Vinayak, S. Oymak, and B. Hassibi, "Sharp performance bounds for graph clustering via convex optimizations," *http://www.its.caltech.edu/∼rkorlaka/KVOHSharpClustering.pdf*, 2013.

7. APPENDIX: PROOF OUTLINE

This section presents an outline of the proofs of the theorems stated in Section 3.

7.1. Additional Notation

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.,

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

Let $\mathcal{R}_{i,j} = \mathcal{C}_i \times \mathcal{C}_j$ for $1 \le i,j \le K+1$. One can see that $\{\mathcal{R}_{i,j}\}$ divides $[n] \times [n]$ into $(K+1)^2$ disjoint regions similar to a grid. Thus, $\mathcal{R}_{i,i}$ is the region induced by i'th cluster for any $i \le K$. Let $\mathcal{A} \subseteq [n] \times [n]$ be the set of nonzero coordinates of $\mathbf{A}$, i.e., $\mathbb{1}_{\mathcal{A}}^{n\times n} = \mathbf{A}$. The set $\mathcal{A}^c \cap \mathcal{R}$ corresponds to the missing edges inside the clusters and so on.

7.2. Sketch of proof

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 the following,

$$(\|\mathbf{L}^0 + \mathbf{E}^L\|_\star + \lambda \|\mathbf{S}^0 + \mathbf{E}^S\|_1) - (\|\mathbf{L}^0\|_\star + \lambda \|\mathbf{S}^0\|_1) > 0, \tag{7.1}$$

for all feasible perturbations $(\mathbf{E}^L, \mathbf{E}^S)$. Let $\mathbf{L}^0 = \mathbf{U}\Lambda\mathbf{U}^T$, where $\Lambda = \text{diag}\{n_1, n_2, \ldots, n_K\}$ and $\mathbf{U} = [\mathbf{u}_1 \ \ldots \ \mathbf{u}_K] \in \mathbb{R}^{n\times K}$,

$$\mathbf{u}_{l,i} = \begin{cases} \frac{1}{\sqrt{n_l}} & \text{if } i \in \mathcal{C}_l \\ 0 & \text{otherwise.} \end{cases} \tag{7.2}$$

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

We construct $\mathbf{W} \in \mathcal{M}_U$, from

$$\mathbf{W}_0 = \sum_{i=1}^{K} c_i \mathbb{1}_{\mathcal{R}_{i,i}}^{n\times n} + c\mathbb{1}_{\mathcal{R}^c}^{n\times n} + \lambda\left(\mathbb{1}_{\mathcal{A}}^{n\times n} - \mathbb{1}_{\mathcal{A}^c}^{n\times n}\right) \tag{7.3}$$

where $c_i = -\lambda(2p_i - 1), i = 1, 2, \ldots K$ and $c = -\lambda(2q - 1)$. Using results from [35] we compute upper bound on $\|\mathbf{W}_0\|$ as $\left(4\sqrt{(q(1-q)n} + \gamma_{\text{succ}} + \epsilon\sqrt{n}\right)\lambda$. Setting $\lambda < \left(4\sqrt{(q(1-q)n} + \gamma_{\text{succ}} + \epsilon\sqrt{n}\right)^{-1}$ we then show that equation 7.1 holds with high probability.

8. PROOF OF CONVERSE RESULTS

Lemma 8.1. *Let $p_{min} > \frac{1}{2} > q$ and $\mathcal{G}$ be a random graph generated according to the stochastic block model 2.1 with cluster sizes $\{n_i\}_{i=1}^K$.*

1. *If $\min_i \{n_i(2p_i - 1)\} < \frac{1}{\lambda}$, then $(\mathbf{L}^0, \mathbf{S}^0)$ is not an optimal solution to the program 1.1 with probability at least $1 - K\exp\left(-\Omega(k_{\min}^2)\right)$.*

2. *If $\lambda > \sqrt{\frac{n}{q\left(n^2 - \sum_{i=1}^K n_i^2\right)}}$, then $(\mathbf{L}^0, \mathbf{S}^0)$ is not an optimal solution to the program 1.1 with high probability.*

Proof. Lagrange for the problem 1.1 can be written as follows

$$\begin{aligned}\mathscr{L}(\mathbf{L}, \mathbf{S}; \mathbf{M}, \mathbf{N}) &= \|\mathbf{L}\|_\star + \lambda\|\mathbf{S}\|_1 + \text{trace}(\mathbf{M}(\mathbf{L} - \mathbb{1}\mathbb{1}^T)) \\ &\quad -\text{trace}(\mathbf{N}\mathbf{L}).\end{aligned} \tag{8.1}$$

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 8.1 at $\mathbf{L}^0$ has to be 0, i.e.,

$$\partial\|\mathbf{L}^0\|_\star + \lambda\,\partial\|\mathbf{A} - \mathbf{L}^0\|_1 + \mathbf{M}^0 - \mathbf{N}^0 = 0. \tag{8.2}$$

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

Also, by complementary slackness,

$$\text{trace}(\mathbf{M}^0(\mathbf{L}^0 - \mathbb{1}\mathbb{1}^T)) = 0, \tag{8.3}$$

and

$$\text{trace}(\mathbf{N}^0\mathbf{L}^0) = 0. \tag{8.4}$$

From 3.1 and 8.3, 8.4, we have $(\mathbf{M}^0)_\mathcal{R} \ge 0$, $(\mathbf{M}^0)_{\mathcal{R}^c} = 0$, $(\mathbf{N}^0)_\mathcal{R} = 0$ and $(\mathbf{N}^0)_{\mathcal{R}^c} \ge 0$. Hence $(\mathbf{M}^0 - \mathbf{N}^0)_\mathcal{R} \ge 0$ and $(\mathbf{M}^0 - \mathbf{N}^0)_{\mathcal{R}^c} = 0$.

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

$$\mathbf{u}_{l,i} = \begin{cases} \frac{1}{\sqrt{k_l}} & \text{if } i \in \mathcal{C}_l \\ 0 & \text{else.} \end{cases} \tag{8.5}$$

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

From 8.2, we have

$$\mathbf{U}\mathbf{U}^T + \mathbf{W} - \lambda\left(\text{sign}(\mathbf{S}^0) + \mathbf{Q}\right) + (\mathbf{M}^0 - \mathbf{N}^0) = 0. \tag{8.6}$$

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

$$\underbrace{\text{sum}\left(\mathbf{L}^0\right)_{\mathcal{R}_{i,i}}}_{k_i} - \text{sum}\left(\lambda\left(\text{sign}(\mathbf{S}^0) + \mathbf{Q}\right)_{\mathcal{R}_{i,i}}\right) + \underbrace{\text{sum}\left(\mathbf{M}^0 - \mathbf{N}^0\right)_{\mathcal{R}_{i,i}}}_{\geq 0} = 0 \tag{8.7}$$

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

$$\text{sum}\left(\text{sign}(\mathbf{S}^0)\right) = -k_i^2(1 - p_i)$$

and $\text{sum}\left(\mathbf{Q}\right) \leq k_i^2 p_i$.

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

$$\underbrace{\text{sum}\left(\mathbf{L}^0\right)_{\mathcal{R}_{i,i}}}_{k_i} - \text{sum}\left(\lambda\left(\text{sign}(\mathbf{S}^0) + \mathbf{Q}\right)_{\mathcal{R}_{i,i}}\right) + \underbrace{\text{sum}\left(\mathbf{M}^0 - \mathbf{N}^0\right)_{\mathcal{R}_{i,i}}}_{\geq 0} \geq k_i + \lambda k_i^2(1 - 2p_i). \tag{8.8}$$

If $k_i + \lambda k_i^2(1 - 2p_i) > 0$, then the equation 8.2 does not hold and hence $\mathbf{L}^0$ cannot be an optimal solution to the program 1.1. $k_i + \lambda k_i^2(1-2p_i) > 0$ would require $k_i\,(2p_i - 1) < \frac{1}{\lambda}$. (Note that, $p_i > \frac{1}{2}$ and hence $2p_i - 1 > 0$.) The same argument holds for each cluster $i = 1 \dots K$, and by union bound we get the result 1 in the Lemma 8.1.

Notice that $\left(\mathbf{U}\mathbf{U}^T\right)_{\mathcal{R}^c} = 0$ and entries of $\text{sign}(\mathbf{S}^0) + \mathbf{Q}$ and $\mathbf{M}^0 - \mathbf{N}^0$ over $\mathcal{R}^c \cap \mathcal{A}$ are negative. Hence from the equation 8.6,

$$\begin{aligned} \|\mathbf{W}\|_F^2 &\geq \|\left(\mathbf{U}\mathbf{U}^T + \mathbf{W}\right)_{(\mathcal{R}^c\cap\mathcal{A})}\|_F^2 \\ &\geq \|\lambda\left(\text{sign}(\mathbf{S}^0) + \mathbf{Q}\right)_{(\mathcal{R}^c\cap\mathcal{A})}\|_F^2. \end{aligned} \tag{8.9}$$

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 the stochastic block model 2.1, each entry of $\mathbf{A}$ over $\mathcal{R}_c$ is non-zero with probability q. Hence with probability at least $1 - \exp\left(-\Omega(|\mathcal{R}_c|)\right)$, $|\mathcal{R}_c \cap \mathcal{A}| = q(n^2 - \sum_{i=1}^K n_i^2)$. Thus from equation 8.9 we have,

$$\|\mathbf{W}\|_F^2 \geq \lambda^2 q(n^2 - \sum_{i=1}^K n_i^2), \tag{8.10}$$

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 1.1.

$$\|\mathbf{W}\| = |\sigma_{\max}(\mathbf{W})| \geq \frac{\|\mathbf{W}\|_F}{\sqrt{n}},$$

which on combining with equation 8.10 gives us,

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

So, if $\lambda\sqrt{q\left(n^2 - \sum_{i=1}^K n_i^2\right)/n} > 1$ then, $\left(\mathbf{L}^0, \mathbf{S}^0\right)$ cannot be an optimal solution to the program 1.1. This gives us result 2 in the Lemma 8.1.

■