Similarity Clustering in the Presence of Outliers: Exact Recovery via Convex Program

Ramya Korlakai Vinayak and Babak Hassibi
Dept. of Electrical Engineering, California Institute of Technology

***Abstract*—We study the problem of clustering a set of data points based on their similarity matrix, each entry of which represents the similarity between the corresponding pair of points. We propose a convex-optimization-based algorithm for clustering using the similarity matrix, which has provable recovery guarantees. It needs no prior knowledge of the number of clusters and it behaves in a robust way in the presence of outliers and noise. Using a generative stochastic model for the similarity matrix (which can be thought of as a generalization of the classical Stochastic Block Model) we obtain *precise bounds* (not orderwise) on the sizes of the clusters, the number of outliers, the noise variance, separation between the mean similarities inside and outside the clusters and the values of the regularization parameter that guarantee the exact recovery of the clusters with high probability. The theoretical findings are corroborated with extensive evidence from simulations.**

I. Introduction

Big data sets are collected by companies, governments and research institutions with the aim of extracting useful and relevant information. Clustering [1] is a widely used pattern recognition tool that broadly refers to the problem of grouping together data points that are similar to each other. In certain instances, the data points can be embedded in Euclidean space; in others, they can be categorical data, which do not readily lend themselves to such an embedding, or a combination of both. For example, in the case of census data, each individual person has different attributes such as age and income which which are numerical and race, religion, address etc., that are categorical [2]. Simple encoding schemes, such as using a $D-$dimensional vector for a categorical field of size D, not only artificially inflate the dimension of the data, but might also give poor results when used with a numerical algorithm when compared to learning an embedding based on similarity or kernel methods [3], [4]. Depending on the data and application domain, it is often possible to construct a similarity map between pairs of data points that assigns a numerical value to how similar (or dissimilar) two data points are. This in turn leads to a *similarity matrix* (also referred to as an affinity matrix in the literature).

If we have noiseless data and an ideal similarity map, then all pairs of points in the same cluster would be mapped to the same similarity value, say 1, and 0 otherwise. However, in reality, the data will be noisy and it is difficult to design a perfectly ideal similarity map. We assume a simple but reasonable probabilistic generative model for the similarity matrix where the average similarity between two data points is higher if they are in the same cluster and lower otherwise. This model can be seen as a natural extension of the popular Stochastic Block Model [5] which is a random unweighted graph model where the probability of the existence of an edge between nodes in the graph that are in the same cluster is higher than those that are not.

The data, apart from being *noisy* might also contain *outliers*, that is data points that do not belong to any clusters. Thus, given a noisy similarity matrix, denoted by $\mathbf{A}$, with outliers, and no other side information, we want to reliably find the clusters. In this regard, we seek to identify a matrix $\mathbf{X}$ whose rank is equal to the number of clusters and in the regions corresponding to the same cluster has entries that are non-zero and equal, and zero elsewhere - thus reflecting the cluster regions (defined formally in Section III). In most practical scenarios, the number of clusters is much lesser than the total number of data points, which makes the matrix $\mathbf{X}$ low-rank.

Convex programs for clustering have drawn attention recently as they are robust to noise and lend themselves to analysis. A general convex approach of using low-rank plus sparse matrix decomposition via trace-norm minimization with a regularized l_1-norm penalty (robust PCA) for finding clusters in *unweighted graphs* has been well-studied [6]–[15]. While the sparse noise model and hence the l_1 penalty works very well for unweighted graphs, it does not fit the similarity model.

Inspired by the robust PCA-based clustering algorithms for unweighted graphs, we propose the following convex program to find the low-rank matrix $\mathbf{X}$ for similarity clustering:

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{A}-\mathbf{X}\|_F^2 + \lambda \ \text{trace}(\mathbf{X}) \\ \text{subject to} \quad & \\ & \mathbf{X} \succcurlyeq 0, \ \mathbf{X}_{i,j} \geq 0 \quad \text{for all } i,j \in [n] \\ & \sum_j \mathbf{X}_{i,j} \leq 1, \quad \text{for all } i \in [n] \end{aligned} \tag{I.1}$$

where $\|.\|_F$ is the Frobenius norm (square root of the sum of the squares of the entries of the matrix), $[n]$ denotes the set $\{1,2,\ldots,n\}$ and $\lambda > 0$ is a regularization parameter (we will later comment on how to set this). Also, by $\mathbf{X} \succcurlyeq 0$, we mean that $\mathbf{X}$ is *symmetric* and has non-negative eigenvalues. The constraints $\sum_j \mathbf{X}_{i,j} \leq 1$ along with $\mathbf{X} \geq 0$ helps in forcing the entries outside the cluster to zero and those corresponding to the same cluster to be equal.

Program I.1 is very simple and intuitive. Furthermore, it does not require any information other than the similarity matrix itself. The goal of this work is to understand the *fundamental limits* of the simple Program I.1, that is, the conditions under which it successfully recovers clusters. In particular, we aim to address the following questions (Section III-D):

1) How noisy can the similarity matrix be? At a given noise level, how separated should the average similarity inside and outside the clusters be?
2) How small can the clusters be? How much can their relative sizes vary?
3) How many outliers can be tolerated? How is the performance affected as the number of outliers becomes large, say larger than the size of the clusters?

Our contributions are multifold:

1) We analyze Program I.1 on a generative model (defined further below in Section II), that is a natural extension of the Stochastic Block Model, and obtain *precise thresholds* (not orderwise) as a function of the problem parameters sufficient for the exact recovery of the underlying cluster structure (Section III). Though our analysis uses the problem parameters, the program itself is agnostic to them.
2) We provide insights into the behavior of the solution in the presence of outliers (Sections III-B and III-C), which is important from a practical standpoint. In the presence of a large number of outliers, Program I.1 exhibits an interesting difference from what occurs in robust PCA-based convex algorithms for unweighted graphs.
3) Our analysis also gives insights into the effect of the noise variance in the similarity matrix on the successful recovery of clusters.

II. Model

We consider the following random generative model for similarity matrices:

Definition 2.1 (Similarity Block Model): Let n be the number of data points comprised of K disjoint clusters and a set of outliers (points that do not belong to any clusters). Let $\mathbf{A} = \mathbf{A}^T$ be the similarity matrix with entries $\mathbf{A}_{l,m} \in [0,1]$. The entries $\mathbf{A}_{l,m}$ with $l \geq m$ are random, independent of each other given the cluster assignment, with variance σ^2 and the means given by:

$$\mathrm{E}(\mathbf{A}_{l,m}) = \begin{cases} \mu_i, & \text{if } l,m \text{ are in the same cluster } i. \\ \mu_{out}, & \text{if } l,m \text{ are not in the same cluster.} \end{cases}$$

III. Main Results

Let n_i, where $i \in [K]$ denote the number of nodes in cluster i, which we will refer to as the *size* of cluster i. If there are outliers, that is nodes that do not belong to any cluster, we denote the number of outliers by n_{K+1} (or n_{out}). Assume that the similarity matrix is generated from the model in Definition 2.1. In this section we present the conditions to guarantee the exact recovery of the underlying cluster structure in the cases when there are (a) no outliers, (b) a small number of outliers and (c) a large number of outliers. The results presented hold with probability at least $1 - n^2 \exp\{-\Omega(n_{min})\}$, where $n_{min} = \min_{i \leq K} n_i$ is the size of the smallest cluster.

A. No Outliers

In the case where there are no outliers, we aim to recover the following matrix via Program I.1,

$$\mathbf{X}^* = \sum_{i=1}^{K} \mathbf{x}_i \mathbf{x}_i^T, \quad \mathbf{x}_i = \frac{1}{\sqrt{n_i}} \mathbf{c}_i \tag{III.1}$$

where $\mathbf{c}_i \in \mathrm{R}^n$ is the indicator vector for cluster i, with ones in the entries corresponding to the data points that belong to cluster i and zeros everywhere else. $\mathbf{x}_i$ is the normalized indicator vector for cluster i. So, the entries of $\mathbf{X}^*$ are:

$$\mathbf{X}^*_{l,m} = \begin{cases} \frac{1}{n_i}, & \text{if both nodes } l,m \text{ are in the same cluster } i \\ 0, & \text{if nodes } l,m \text{ are not in the same cluster} \end{cases}$$

The following quantities are important for our results:

- *Cluster Density:* For each cluster $i \in [K]$, define $\rho_i := n_i \mu_i > 0$, requiring $\mu_i > 0$.
- *Minimum Cluster Density* is defined as $\rho_{min} := \min_{i \leq K} \rho_i > 0$.
- *Cross Cluster Density:* For each pair of clusters $i \neq j \in [K]$, define the *cross cluster density* as $\gamma_{ij} := 2 \left(\frac{\mu_i + \mu_j}{2} - \mu_{out} \right) \left(\frac{1}{n_i} + \frac{1}{n_j} \right)^{-1} > 0$, requiring $\frac{\mu_i + \mu_j}{2} > \mu_{out}$. That is, the average of the mean similarity of any two clusters clusters i and j must be at least as big as the mean similarity between them.
- *Minimum Cross Cluster Density:* is defined as $\gamma_{min} := \min_{i \neq j \leq K} \gamma_{ij} > 0$.
- *Noise threshold*, $\Lambda := 2\, \sigma \sqrt{n}$ which depends only on the noise variance and number of data points.

Theorem 1: [No Outliers] When there are no outliers, if the regularizer λ is within the following range,

$$\Lambda < \lambda < \min \{\rho_{min}, \gamma_{min}\} - 1 \tag{III.2}$$

then, $\mathbf{X}^*$ is the unique optimal solution to Program I.1 with high probability. If $\lambda > \min \{\rho_{min}, \gamma_{min}\} - 1$, then Program I.1 fails to recover $\mathbf{X}^*$ with high probability.

B. Small Number of Outliers

In the presence of outliers, the solution to Program I.1 depends on the number of outliers compared to the size of the smallest cluster. When $n_{out} \leq \mathcal{O}(n_{min})$, we refer to it as a small number of outliers. In addition to the cross cluster density γ defined before, define the following:

- *Effective cluster density* in presence of outliers for each cluster i as $\eta_i := (\mu_i - 2\mu_{out})\, n_i > 0$, implying $\mu_i > 2\mu_{out}$, required only in the case of small number of outliers.
- *Minimum effective density* be $\eta_{min} := \min_{i \leq K} \eta_i$.

Theorem 2: [Small Number of Outliers] If the regularizer λ is within the following range,

$$\Lambda + \mu_{out} n_{K+1} < \lambda < \min \{\eta_{min}, \gamma_{min}\} - 1 \tag{III.3}$$

then $\mathbf{X}^*$ is the unique optimal solution to Program I.1 with high probability. If $\lambda > \min \{\eta_{min}, \gamma_{min}\} - 1$ then, Program I.1 fails to recover $\mathbf{X}^*$ with high probability.

C. Large Number of Outliers

When the number of outliers is large (at least $\Omega(\sqrt{n})$) and is comparable or larger than the size of clusters, we cannot hope to recover $\mathbf{X}^*$ which requires the entries corresponding to the outlier region to be all zeros. Instead Program I.1 groups all the outliers together to give an extra cluster and hence recovers $\tilde{\mathbf{X}} := \sum_{i=1}^{K+1} \mathbf{x}_i \mathbf{x}_i^T$ where $\mathbf{x}_{K+1}$ is the normalized indicator vector for the cluster of outliers. So, the entries of $\tilde{\mathbf{X}}$ are:

$$\tilde{\mathbf{X}}_{l,m} = \begin{cases} \frac{1}{n_i}\,, & \text{if nodes } l,m \text{ are in the same cluster } i. \\ 0\,, & \text{if nodes } l,m \text{ are in different clusters.} \\ \frac{1}{n_{K+1}} & \text{if both nodes } l,m \text{ are outliers.} \end{cases}$$

Note that this is not a bad scenario. Rather, it is good that outliers get separated out as a cluster and do not get merged with other clusters. Once the cluster structure is revealed, one can compare the average similarity inside the clusters obtained and the average similarity outside to decide if any of the clusters obtained has average similarity very close to that of outside cluster region, and hence discard it.

In addition to the cluster densities ρ and the cross cluster denisities γ, define the following:

- *Outlier Density:* $\rho_{K+1} := \mu_{out} n_{K+1}$.
- *Cross Cluster-Outlier Density:* For each $i \in [K]$, define cross density of cluster i with outliers as, $\gamma_{i,K+1} := (\mu_i - \mu_{out}) \left(\frac{1}{n_i} + \frac{1}{n_{K+1}}\right)^{-1} > 0$.
- *Minimum cluster density* in the presence of *outliers* as $\rho_{min}^{out} := \min_{i \le K+1} \rho_i$.
- *Minimum cross cluster density* in the presence of *outliers* $\gamma_{min}^{out} := \min_{i \neq j \le K+1} \gamma_{ij}$.

Theorem 3: [Large Number of Outliers] If the regularizer λ is within the following range,

$$\Lambda < \lambda < \min\{\rho_{min}^{out}, \gamma_{min}^{out}\} - 1, \tag{III.4}$$

then $\tilde{\mathbf{X}}$ is the unique optimal solution to Program I.1 with high probability. If $\lambda > \min\{\rho_{min}^{out}, \gamma_{\min}^{out}\} - 1$, then, the Program I.1 fails to recover $\tilde{\mathbf{X}}$ with high probability.

Note that Theorems 2 and 3 are not in contradiction, since if the conditions on mean and cluster sizes are satisfied in both cases, setting $\lambda > 2\sigma\sqrt{n} + \mu_{out} n_{K+1}$ (Equation III.3 in Theorem 2) would violate $\mu_{out} n_{K+1} > \lambda + 1$ (Equation III.4 in Theorem 3).

D. Discussion:

1) **Size of the Smallest Cluster:** All three theorems stated in this section imply $\rho_i = \mu_i n_i > \Lambda + 1 = 2\sigma\sqrt{n} + 1 \ \forall i$, and hence we require $n_{min} \geq \Omega(\sqrt{n})$ to guarantee success, which matches the earlier known results. The results cannot guarantee success when $\Lambda + 1 < \rho_{min}$, that is, when $n_{min} < \mathcal{O}(\sqrt{n})$. In this regime it is not known whether the clustering problem can be efficiently solved.
2) **Relative Size of Clusters:** Note that the results do not place any restrictions on the relative size of clusters. So, we can have clusters of varying sizes: for example, some clusters of size $\Theta(n)$ and some of size $\Theta(\sqrt{n})$.
3) **Size of Outliers**: In the presence of outliers, Theorem 2 implies $\rho_i > \eta_i > \Lambda + \mu_{out} n_{K+1} + 1$. So to guarantee the exact recovery of $\mathbf{X}^*$ as the optimal solution to Program I.1 we require $n_{min} \geq \max\{\Omega(\sqrt{n}), \Omega(n_{K+1})\}$. Note that this requirement is automatically satisfied if the number of outliers is $o(\sqrt{n})$ as we need $n_{min} \geq \Omega(\sqrt{n})$ in all cases. If the number of outliers is large in comparison to the size of the smallest cluster, we cannot guarantee the recovery of a solution with all zero entries in the region corresponding to the outliers.
4) **Large Number of Outliers:** Theorem 3 implies $\rho_{out} = \mu_{out} n_{out} \geq \Lambda + 1 = 2\sigma\sqrt{n} + 1$. So, if the number of outliers is at least $\Omega(\sqrt{n})$, i.e, the number of outliers is *large*, then we can guarantee that they form their own cluster under the conditions in Theorem 3.
5) **Separation Between the Means Compared to the Noise Variance:** For simplicity, assume all the clusters are of equal size, $n_i = m$ and $\mu_i = \mu_{in}\ \forall i$.
   a) In the case of no outliers, from $\gamma_{ij} > 2\sigma\sqrt{n} + 1$ (Equation III.2) we get the following sufficient condition:
   $$\frac{\mu_{in} - \mu_{out}}{\sigma} > \frac{1}{m}\left(2\sqrt{n} + \frac{1}{\sigma}\right).$$
   If $m = \Theta(\sqrt{n})$, then as $n \to \infty$, we require $\frac{\mu_{in} - \mu_{out}}{\sigma} > \Omega(1)$, whereas if $m = \Theta(n)$ then $\mu_{in} - \mu_{out} > 0$ is sufficient to guarantee exact recovery.
   b) In the case of large number of outliers, from Equation III.4 we get:
   $$\frac{\mu_{in} - \mu_{out}}{\sigma} > \left(\frac{1}{m} + \frac{1}{n_{K+1}}\right)\left(2\sqrt{n} + \frac{1}{\sigma}\right).$$
   c) In the case of small number of outliers, from Equation III.3 we get:
   $$\frac{\mu_{in} - 2\mu_{out}}{\sigma} > \frac{1}{m}\left(2\sqrt{n} + \frac{\mu_{out} n_{K+1} + 1}{\sigma}\right).$$
   So the average similarity inside the clusters will have to be higher than twice the average similarity outside to recover $\mathbf{X}^*$ in the presence of small number of outliers.
6) **Regularization Parameter:** If the noise variance is known, then the regularizer can be set to $\lambda = 2\sigma\sqrt{n}$. In case there is no information about σ, then we suggest using the empirical variance of $\mathbf{A}$ or setting $\lambda = 2\sqrt{n}$. The value of λ provides a bound on how much noise can be tolerated. For e.g, if we set $\lambda = 2\sqrt{n}$, then $\sigma \leq 1$ can be tolerated.

E. Brief Proof Outline

Due to reasons of space, we are only able to provide a brief outline of the proof. However, the interested reader may consult the supplementary material posted at this url [1]. Define dual variables for the constraints of Program I.1,

1) $\mathbf{Y} \in \mathbb{R}^{n \times n}$, $\mathbf{Y} \succcurlyeq 0$ for constraint $\mathbf{X} \succcurlyeq 0$.
2) $\nu \in \mathbb{R}^n$, $\nu \geq 0$ for constraints $\sum_j \mathbf{X}_{i,j} \leq 1$, $\forall i$.
3) $\mathbf{Z} \in \mathbb{R}^{n \times n}$, $\mathbf{Z} \geq 0$ for constraints $\mathbf{X} \geq 0$.

[1] http://www.its.caltech.edu/~rkorlaka/SimilarityClusteringSupplementary.pdf

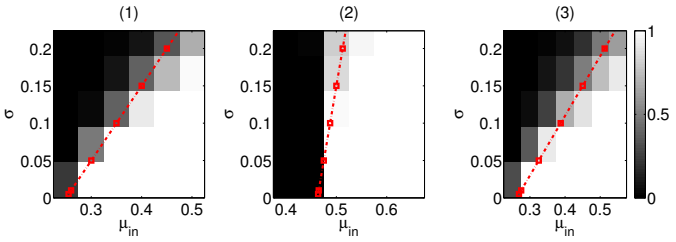

Fig. 1: Fraction of correct entries in the solution obtained by running Program I.1 with $n = 100$, $\mu_{out} = 0.2$ and varying μ_{in} and σ, for three cases: (1) no outliers, (2) small number of outliers and (3) large number of outliers. Dotted red line is the threshold for μ_{in} predicted by theory for the corresponding σ. The white and black regions represent the empirical regions of success and failure respectively.

If a feasible $\hat{\mathbf{X}}$ is an optimal solution to Program (I.1), then the following conditions have to hold (from KKT conditions and complementary slackness):

$$\mathbf{Z} + \mathbf{Y} = \lambda \mathbf{I} + \hat{\mathbf{X}} + \mathbb{1}\nu^T - \mathbf{A} + \nu \mathbb{1}^T$$

$$\text{trace}(\hat{\mathbf{X}}\mathbf{Y}) = 0,\ \text{trace}(\hat{\mathbf{X}}\mathbf{Z}) = 0,\ \nu^T(\hat{\mathbf{X}}\mathbb{1} - \mathbb{1}) = 0.$$

We first construct dual variables that satisfy the conditions above. The dual variables $\mathbf{Z}, \mathbf{Y}, \nu$ thus obtained are functions of the problem parameters $\{\{\mu_i\}_{i\in[K]}, \mu_{out}, \sigma, \{n_i\}_{i\in[K]}, n_{out}\}$. The condition $\mathbf{Y} \succcurlyeq 0$ will give the lower bound on λ of the form Λ or $\Lambda + \mu_{out} n_{out}$ depending on the case. The conditions $\nu \geq 0$ gives the lower bounds on the cluster densities ρ. The conditions $\mathbf{Z} \geq 0$ gives the lower bounds on cross-cluster densities γ and the effective cluster densities η.

IV. Simulations

We consider an example of $n = 100$ data points to illustrate the sharp transitions predicted in the main results in Section III via numerical simulations. The similarity matrix is generated as follows: For $l \geq m$, $\mathbf{A}_{lm}$ is sampled independently from $\mathcal{N}(\mu_i, \sigma^2)$ if both the points l, m belong to the same cluster i, else it is sampled independently from $\mathcal{N}(\mu_{out}, \sigma^2)$. Note that this does not necessarily satisfy $\mathbf{A}_{l,m} \in [0, 1]$. However, we will choose the μ_{in}, μ_{out} and range of σ such that we will not be violating the condition of the average similarity inside the clusters being positive. All the results presented are an average over 5 experiments unless otherwise stated. We use the CVX package for Matlab [16] to run Program I.1. We set the mean similarity between nodes that are not in the same cluster to $\mu_{out} = 0.2$. The standard deviation σ starts with 0.01 and is then varied from 0.05 to 0.20 in steps of 0.05. Note that for the clusters sizes of 20 and 80, non-zero entries of the ideal solution are 0.05 and 0.0125 respectively. We declare an entry of the solution matrix $\mathbf{X}_{l,m}$ to be in error if $|\mathbf{X}_{l,m} - \mathbf{X}^{ideal}_{l,m}| > 10^{-3}$. For Theorems 1 and 2, $\mathbf{X}^{ideal}$ is $\mathbf{X}^*$ and for Theorem 3, $\mathbf{X}^{ideal}$ is $\tilde{\mathbf{X}}$.

1) **No Outliers:** We consider five clusters of size 20 each, and no outliers. We set the regularization parameter $\lambda = 1.001\Lambda = 1.001 \times 2\sigma\sqrt{n}$ (lower bound on λ in Equation III.2). We vary the mean similarity inside the clusters μ_{in} from 0.25 to 0.5 in steps of 0.05. From Theorem 1, we expect Program I.1 to succeed (to obtain solution $\mathbf{X}^*$) with high probability when, $\mu_{in} > \mu_{out} + (\lambda + 1)/n_1$.
2) **Small Number of Outliers:** For this case, we consider one cluster of size $n_1 = 80$ and the rest $n_{out} = 20$ are outliers. We set the regularization parameter to $\lambda = 1.001(\Lambda + \mu_{out} n_{out})$ (lower bound on λ in Equation III.4) and vary μ_{in} from 0.4 to 0.65 in steps of 0.05. From Theorem 2, we expect Program I.1 to succeed (obtain solution $\mathbf{X}^*$) with high probability when, $\mu_{in} > 2\mu_{out} + (\lambda + 1)/n_1$.
3) **Large Number of Outliers:** In this case, we consider one cluster of size $n_1 = 20$ and the rest $n_{out} = 80$ are outliers. We set the regularization parameter to $\lambda = 1.001\Lambda$ (lower bound on λ in Equation III.4) and vary μ_{in} from 0.25 to 0.55 in steps of 0.05. From Theorem 3, we expect Program I.1 to succeed (obtain solution $\tilde{\mathbf{X}}$) with high probability when, $\mu_{in} > \mu_{out} + (\lambda + 1)\,(1/n_1 + 1/n_{out})$.

Figure 1 shows the fraction of correct entries of the output of Program I.1 for each of the three cases described above. The white and black regions corresponds to the empirical regions of success and failure respectively. The dashed-red line is the threshold for μ_{in} as predicted by our results in Section III. We observe that the transition occurs around μ_{in} predicted from our results. Our theoretical thresholds are *sharp* even for $n = 100$ with cluster sizes as small as 20.

V. Related Works

In this section we will discuss some related works.

Convex Penalties: [17]–[20] have introduced regularized convex relaxations for *hierarchical clustering* and show that as the regularizer is varied in a certain range there is a coalescing of clusters that gives rise to a hierarchical tree. However, they do not give guarantees on the problem parameters that give rise to a particular clustering at any point in the tree. Also, they do not provide theoretical guarantees on clustering in the presence of outliers.

Spectral Clustering: [21] analyzes the spectral partitioning of graphs under the Stochastic Block Model and [22] studies the asymptotic correctness of spectral clustering for these models. [23], [24] study the stability of the eigenvectors of the graph Laplacian under noisy perturbations. [24] provides guarantees on the exact recovery of clusters for spectral clustering under noisy perturbations to the similarity matrix. However, they require the clusters be balanced (i.e., the size of the clusters are constant fractions of each other). Moreover, these results do not hold when there are outliers.

Convex Programs for Graph Clustering: [6]–[15] consider clustering unweighted graphs via convex optimization based on a low-rank + sparse decomposition of the *unweighted* adjacency matrix of the graph via nuclear norm minimization with l_1 regularization. In the case of unweighted graphs, if the edge density inside the clusters is bigger than that outside, under mild conditions, convex programs can recover clusters of size $\Omega(\sqrt{n})$, regardless of the size of the outliers. Whereas, in the case of similarity clustering (Section III-C), if the number of outliers is larger than the smallest cluster, Program I.1 gives an extra cluster that contains all the outliers.

Submatrix Localization: [25] considers the special case of submatrix localization (bi-clustering) when the number of clusters, $K = 1$ and provides guarantees for exact recovery via message passing algorithm when the size of the cluster is known. Section 3 in [26] considers the problem of submatrix localization when the clusters are homogeneous (same mean inside all the clusters) and have same size. They provide order-wise bounds on signal strength required for exact recovery of clusters from a convex program which requires the knowledge of cluster sizes and number of clusters. We can recover the results for the case of symmetric submatrix localization by setting $\mu_i = \mu, n_i = m, \ \forall i \in [K]$. Note that the quantities *signal* and *SNR* defined in [25] and [26] are related to the separation between means (discussed in 5 in Section III-D) that arises via cross-cluster density.

Convex Program for Similarity Clustering: The work that is closest to ours in terms of the approach and analysis is [27], which considers the problem of clustering a similarity matrix when the number of clusters are known. [27] analyzes a convex program and provides guarantees for recovering the clusters as long as the number of number of outliers is not too large (less than the size of the smallest cluster). While the results in [27] are interesting, it does not comment on the quality of the solution when the number of outliers is large. Further, the convex program in [27] requires the knowledge of the number of clusters, which can be problematic in the presence of outliers. In contrast, Program I.1 is oblivious to the exact number of clusters and naturally figures it out as a function of the regularization parameter. This is helpful in understanding the behavior of the program when there are large number of outliers. Our analysis shows that Program I.1 can recover the clusters as long as the regularization parameter is within a *range* rather than a specific number, which is robust to error when it is heuristically set. The model and analysis in [27] does not capture the effect of the noise variance in the similarity matrix on the performance of the program. Though our analysis technique is inspired by the work [27], we extend the analysis to understand the behavior of the solution in the presence of large number of outliers as well as to capture the effect of the noise variance. Also, our analysis gives *precise thresholds* for the successful recovery conditions, whereas the results in [27] are orderwise.

VI. Conclusions

In this work we focus on understanding the performance of convex-optimization-based clustering in the presence of outliers when we only have the similarity matrix given to us (with no additional information). We analyze a simple and intuitive convex program (I.1), and for the stochastic similarity model, we provide guarantees on its performance by deriving precise thresholds (not orderwise) on the cluster sizes, the strength of similarity compared to noise, the number of outliers, and the regularization parameter. We corroborate our results through simulations. One of the drawbacks of convex approach is that it is computationally intensive. In the future, we want to focus on scaling the convex approach to work with large datasets.

References

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, 1999.

[2] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.

[3] K. Zhang, Q. Wang, Z. Chen, I. Marsic, V. Kumar, G. Jiang, and J. Zhang, *From Categorical to Numerical: Multiple Transitive Distance Learning and Embedding*, ch. 6, pp. 46–54.

[4] J. Couto, "Kernel k-means for categorical data," in *Advances in Intelligent Data Analysis VI*, ser. Lecture Notes in Computer Science, A. Famili, J. Kok, J. Pea, A. Siebes, and A. Feelders, Eds. Springer Berlin Heidelberg, 2005, vol. 3646, pp. 46–56.

[5] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Networks*, vol. 5, no. 2, pp. 109 – 137, 1983.

[6] H. Xu, C. Caramanis, and S. Sanghavi, "Robust pca via outlier pursuit." in *NIPS*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 2496–2504.

[7] A. Jalali, Y. Chen, S. Sanghavi, and H. Xu, "Clustering partially observed graphs via convex optimization," in *Proceedings of the 28th ICML*. ACM, June 2011, pp. 1001–1008.

[8] B. P. W. Ames and S. A. Vavasis, "Convex optimization for the planted k-disjoint-clique problem," *Math. Program.*, vol. 143, no. 1-2, pp. 299–337, 2014.

[9] ——, "Nuclear norm minimization for the planted clique and biclique problems," *Math. Program.*, vol. 129, no. 1, pp. 69–89, Sep. 2011.

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

[11] Y. Chen, S. Sanghavi, and H. Xu, "Clustering sparse graphs." in *NIPS*, 2012, pp. 2213–2221.

[12] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis, "Low-rank matrix recovery from errors and erasures," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4324–4337, 2013.

[13] B. P. W. Ames, "Robust convex relaxation for the planted clique and densest k-subgraph problems," *ArXiv e-prints*, 2013.

[14] R. K. Vinayak, S. Oymak, and B. Hassibi, "Sharp performance bounds for graph clustering via convex optimizations," in *Proceedings of the 39th*, ser. ICASSP '14, 2014.

[15] ——, "Graph clustering with missing data: Convex algorithms and analysis," in *NIPS 2014, Montreal, Canada*, 2014, pp. 2996–3004.

[16] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," Mar. 2014.

[17] F. Lindsten, H. Ohlsson, and L. Ljung, "Clustering using sum-of-norms regularization : With application to particle filter output computation," in *Proceedings of the 2011 IEEE Statistical Signal Processing Workshop*, 2011, pp. 201–204.

[18] T. D. Hocking, A. Joulin, F. Bach, and J.-P. Vert, "Clusterpath An Algorithm for Clustering using Convex Fusion Penalties," in *28th international conference on machine learning*, United States, Jun. 2011.

[19] J. R. K. L. Gary K. Chen, Eric Chi, "Convex clustering: An attractive alternative to hierarchical clustering," *PLOS Computational Biology*, 2015, in press.

[20] G. I. A. Eric C. Chi and R. G. Baraniuk, "Convex biclustering," arXiv:1408.0856 [stat.ME], August 2014.

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

[22] K. Rohe, S. Chatterjee, and B. Yu, "Spectral clustering and the high-dimensional stochastic blockmodel," *Annals of Statistics*, vol. 39, no. 4, pp. 1878–1915, Aug. 2011.

[23] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*. MIT Press, 2001, pp. 849–856.

[24] S. Balakrishnan, M. Xu, A. Krishnamurthy, and A. Singh, "Noise thresholds for spectral clustering," in *In Advances in Neural Information Processing Systems 25*, 2011.

[25] B. Hajek, Y. Wu, and J. Xu, "Submatrix localization via message passing," *arXiv:1510.09219*, Oct. 2015.

[26] Y. Chen and J. Xu, "Statistical-Computational Tradeoffs in Planted Problems and Submatrix Localization with a Growing Number of Clusters and Submatrices," *arXiv:1402.1267*, 2014.

[27] B. P. W. Ames, "Guaranteed clustering and biclustering via semidefinite programming," *Math. Program.*, vol. 147, no. 1-2, pp. 429–465, 2014.