Clustering Partially Observed Graphs via Convex Optimization

Ali Jalali, Yudong Chen, Sujay Sanghavi (ALIJ,YDCHEN,SANGHAVI)@MAIL.UTEXAS.EDU
UT Austin, 1 University Station C0806, Austin, TX 78712 USA

Huan Xu MPEXUH@NUS.EDU.SG
National University of Singapore , 9 Engineering Drive 1, Singapore 117575 SINGAPORE

Abstract

This paper considers the problem of clustering a partially observed unweighted graph – i.e. one where for some node pairs we know there is an edge between them, for some others we know there is no edge, and for the remaining we do not know whether or not there is an edge. We want to organize the nodes into disjoint clusters so that there is relatively dense (observed) connectivity within clusters, and sparse across clusters.

We take a novel yet natural approach to this problem, by focusing on finding the clustering that minimizes the number of "disagreements" - i.e. the sum of the number of (observed) missing edges within clusters, and (observed) present edges across clusters. Our algorithm uses convex optimization; its basis is a reduction of disagreement minimization to the problem of recovering an (unknown) low-rank matrix and an (unknown) sparse matrix from their partially observed sum. We show that our algorithm succeeds under certain natural assumptions on the optimal clustering and its disagreements. Our results significantly strengthen existing matrix splitting results for the special case of our clustering problem. Our results directly enhance solutions to the problem of Correlation Clustering (Bansal et al., 2002) with partial observations.

This work is supported in part by NSF CAREER grant 0954059, NSF grant EFRI-0735905, DTRA grant HDTRA1-08-0029 and NUS startup grant R-265-000-384-133.

1. Introduction

This paper is about the following task: given partial observation of an undirected unweighted graph, partition the nodes into disjoint clusters so that there are dense connections within clusters, and sparse connections across clusters. By partial observation, we mean that for some node pairs we know if there is an edge or not, and for other node pairs we do not know – these pairs are *unobserved*. This problem arises in several fields across science and engineering. For example, in sponsored search, each cluster is a submarket that represents a specific group of advertisers that do most of their spending on a group of query phrases – see e.g. (Inc, 2009) for such a project at Yahoo. In VLSI and design automation, it is useful in minimizing signaling between components, layout etc. – see e.g. (Kernighan & Lin, 1970) and references thereof. In social networks, clusters represent groups of mutual friends; finding clusters enables better recommendations, link prediction, etc (Mishra et al., 2007). In the analysis of document databases, clustering the citation graph is often an essential and informative first step (Ester et al., 1995). In this paper, we will focus not on specific application domains, but rather on the basic graph clustering problem itself.

As with any clustering problem, this needs a precise mathematical definition. We are not aware of any existing work with provable performance guarantees for partially observed graphs. Even most existing approaches to clustering fully observed graphs, which we review in section 1.1 below, either require an additional input (e.g. the number of clusters k required for spectral or k-means clustering approaches), or do not guarantee the performance of the clustering. Indeed, the specialization of our results to the fully observed case extends the known guarantees there.

Our Formulation: We focus on a natural formulation, one that *does not require any other extraneous input* besides the graph itself. It is based on minimizing *disagreements*, which we now define. Consider any candidate clustering; this will have *(a)* observed node pairs that are in different clusters, but have an edge between them, and *(b)* observed

node pairs that are in the same cluster, but do not have an edge between them. The total number of node pairs of types *(a)* and *(b)* is the number of disagreements between the clustering and the given graph. We focus on the problem of finding the *optimal clustering* – one that minimizes the number of disagreements. Note that we do *not* prespecify the number of clusters. For the special case of fully observed graphs, this formulation is exactly the same as the problem of "Correlation Clustering", first proposed by (Bansal et al., 2002). They showed that exact minimization of the above objective is NP-complete in the worst case – we survey and compare this and other related work in section 1.1. As we will see, our approach and results are very different.

Our Approach: We aim to achieve the combinatorial disagreement minimization objective using matrix splitting via convex optimization. In particular, as we show in section 2 below, one can represent the adjacency matrix of the given graph as the sum of an unknown low-rank matrix (corresponding to"ideal" clusters) and a sparse matrix (corresponding to disagreements from this "ideal" in the given graph). Our algorithm either returns a clustering, which is guaranteed to be disagreement minimizing, or returns a "failure" – it never returns a sub-optimal clustering. Our analysis provides both deterministic and probabilistic guarantees for when our algorithm succeeds. Our analysis uses the special structure of our problem to provide much stronger guarantees than are current results on general matrix splitting (Chandrasekaran et al., 2009; Candes et al., 2009; Hsu et al., 2010).

1.1. Related Work

Our problem can be interpreted in the general clustering context as one in which the presence of an edge between two points indicates a "similarity", and the lack of an edge means no similarity. The general field of clustering is of course vast, and a detailed survey of all methods therein is beyond our scope here. We focus instead on the two sets of papers most relevant to the problem here, namely the work on Correlation Clustering, and the other approaches to the specific problem of graph clustering.

Correlation Clustering: First formulated in (Bansal et al., 2002), correlation clustering looks at the following problem: given a complete graph where every edge is labelled "+" or "-", cluster the nodes to minimize the total of the number of "-" edges within clusters and "+" edges across clusters. As mentioned, for a completely observed graph, our problem is mathematically precisely the same as correlation clustering; in particular a "+" in correlation clustering corresponds to an edge in graph clustering, and a "-" to the lack of an edge. Disagreements are defined in the same way. Thus, this paper can equivalently be considered an algorithm, and guarantees, for *correlation clustering under partial observations*. (Bansal et al., 2002) show that exact minimization is NP-complete, and also provide *(a)* constant-factor approximation algorithm for the problem of minimizing the number of disagreements, and *(b)* a PTAS for maximizing agreements. Their algorithms are combinatorial in nature. Subsequently, there has been much work on devising alternative approximation algorithms for both the weighted and unweighted cases, and for both agreement and disagreement objectives (Emmanuel & Immorlica, 2003; Demaine et al., 2005; Swamy, 2004; Charikar et al., 2003; Emmanuel & Fiat, 2003; Becker, 2005). Approximations based on LP relaxation (Becker, 2005) and SDP relaxation (Swamy, 2004), followed by rounding, have also been developed. We emphasize that while we do convex relaxation as well, we do not do rounding; rather, our convex program itself yields an optimal clustering. We emphasize that ours is the *first* attempt at correlation clustering with partial observations.

Graph Clustering: The problem of graph clustering is well studied and very rich literature on the subject exists (see e.g. (Everitt, 1980; Jain & Dubes, 1988) and references thereof). One set of approaches seek to optimize criteria such as k-median, minimum sum or minimum diameter (Bern & Eppstein, 1996); typically these result in NP-hard problems with few global guarantees. Another option is a top-down hierarchical approach, i.e., recursively bisecting the graph into smaller and smaller clusters. Various algorithms in this category differ in the criterion used to determine where to split in each iteration. Notable examples of such criteria include small cut (Condon & Karp, 2001), maximal flow (Flake et al., 2004), low conductance (Shi & Malik, 2000), eigenvector of the Laplacian (aka spectral clustering) (Ng et al., 2002), and many others. Due to the iterative nature of these algorithms, global theoretical guarantees are hard to obtain.

As we mentioned before, we are not aware of any work on graph clustering with partial observations and provable guarantees.

2. Main Contributions

Our algorithm is based on convex optimization, and either *(a)* outputs a clustering that is guaranteed to be the one that minimizes the number of observed disagreements, or *(b)* declares "failure" – in which case one could potentially try some other approximate methods. In particular, it never produces a suboptimal clustering. We now briefly present the main idea, then describe the algorithm, and finally present our main results – analytical characterizations of when the algorithm succeeds.

Setup: We are given a partially observed graph, whose adjacency matrix is $\mathbf{A}$ – which has $a_{ij} = 1$ if there is an edge

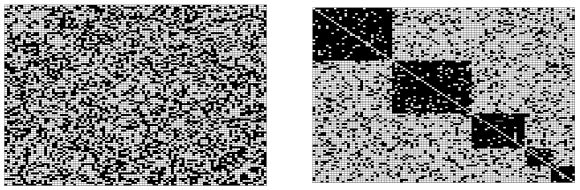

Figure 1. The adjacency matrix of a graph before (a) and after (b) proper reordering (i.e. clustering) of the nodes. The figure (b) is indicative of the matrix as a superposition of a sparse matrix and a low-rank one.

between nodes i and j, $a_{ij} = 0$ if there is no edge, and $a_{ij} =?$ if we do not know. Let Ω_{obs} be the set of observed entries, i.e. the set of elements of $\mathbf{A}$ that are known to be 0 or 1. We want to find the *optimal clustering*, i.e. the one that has the minimum number of disagreements in Ω_{obs}.

Idea: Consider first the fully observed case, i.e. every $a_{ij} = 0$ or 1. Suppose also that the graph is already ideally clustered – i.e. there is a partition of the nodes such that there are no edges between partitions, and each partition is a clique. In this case, the matrix $\mathbf{A} + \mathbf{I}$ is now a *low-rank* matrix, with the rank being equal to the number of clusters. This can be seen by noticing that if we re-ordered the rows and columns so that partitions appear together, the result would be a *block-diagonal* matrix, with each block being an all-ones sub-matrix – and thus rank one. Of course, this re-ordering does not change the rank of the matrix, and hence $\mathbf{A} + \mathbf{I}$ is (exactly) low-rank.
Consider now any given graph, still fully observed. In light of the above, we are looking for a decomposition of its $\mathbf{I} + \mathbf{A}$ into a low-rank part $\mathbf{K}$ (of block-diagonal all-ones, one block for each cluster) and a remaining $\mathbf{B}$ (the disagreements) – such that the number of entries in $\mathbf{B}$ is as small as possible; i.e. $\mathbf{B}$ is sparse. Finally, the problem we look at is recovery of the best $\mathbf{K}$ when we do not observe all entries. The idea is depicted in Figure 1.

Convex Optimization Formulation: We propose to do the matrix splitting using convex optimization, an approach recently taken in (Chandrasekaran et al., 2009; Candes et al., 2009) (however, we establish much stronger results for our special problem). Our approach consists of *dropping* any additional structural requirements, and just looking for a decomposition of the given $\mathbf{A} + \mathbf{I}$ as the sum of a sparse matrix $\mathbf{B}$ and a low-rank matrix $\mathbf{K}$. In particular, we use the following convex program

$$\min_{\mathbf{B},\mathbf{K}} \quad \eta\,||\mathbf{B}||_1 + (1-\eta)\,||\mathbf{K}||_* \qquad (1)$$
$$\text{s.t.} \quad \mathcal{P}_{\Omega_{obs}}(\mathbf{B}+\mathbf{K}) = \mathcal{P}_{\Omega_{obs}}(\mathbf{I}+\mathbf{A})$$

Here, for any matrix M, the term $\mathcal{P}_{\Omega_{obs}}(M)$ keeps all elements of M in Ω_{obs} unchanged, and sets all other elements to 0; the constraints thus state that the sparse and low-rank matrix should in sum be consistent with the observed entries. $||\mathbf{B}||_1 = \sum_{i,j} |b_{ij}|$ is the ℓ_1 norm of the entries of the matrix, which is well-known to be a convex surrogate for the number of non-zero entries $||\mathbf{B}||_0$. The second term is $||\mathbf{K}||_* = \sum_s \sigma_s(K)$ is "nuclear norm": the sum of singular values of $\mathbf{K}$. This has been shown recently to be the convex surrogate[1] for the rank function (Recht et al., 2009). Thus our objective function is a convex surrogate for the (natural) combinatorial objective $\eta\,||\mathbf{B}||_0 + (1-\eta)\text{rank}(\mathbf{K})$. (1) is, in fact, a semi-definite program SDP (Chandrasekaran et al., 2009).

Definitition: Validity: The convex program (1) is said to produce a *valid* output if the low-rank matrix part $\mathbf{K}$ of the optimum corresponds to a graph of disjoint cliques; i.e. its rows and columns can be re-ordered to yield a block-diagonal matrix with all-one matrices for each block.

Validity of a given $\mathbf{K}$ can easily be checked, either via elementary re-ordering operations, or via a singular value decomposition[2]. Our first simple, but crucial, insight is that whenever the convex program (1) yields a valid solution, it is the disagreement minimizer. This is true in-spite of the fact that we have clearly dropped several constraints of the original problem (e.g. we do not enforce the entries of $\mathbf{K}$ to be between 0 and 1, etc.).

Theorem 1 *For any $\eta > 0$, if the optimum of* (1) *is valid, then it is the clustering that minimizes the number of observed disagreements.*

Algorithm: Our algorithm takes the adjacency matrix of the network $\mathbf{A}$ and outputs either the optimal clustering or declares failure. Using the result of Theorem 1, if the clustering is valid, then we are guaranteed that the result is a disagreement minimizer clustering.

Algorithm 1 Optimal-Cluster(A)

```
for η ∈ (0, 1) do
  Solve (1)
  if Solution K is valid then
    Output the clustering w.r.t K and EXIT.
  end if
end for
Declare Failure.
```

We recommend using the fast implementation algorithms developed in (Lin et al., 2009), which is specially tailored

[1] In particular, it is the ℓ_1 norm of the singular value vector, while rank is the ℓ_0 norm of the same.

[2] An SVD of a valid $\mathbf{K}$ will yield singular vectors with disjoint supports. The supports correspond to the clusters.

for matrix splitting. Setting the parameter η can be done either via a simple line search from 0 to 1, binary search, or any other option. Whenever it results in a valid $\mathbf{K}$, we have found the optimal clustering.

Analysis: The main analytical contribution of this paper is conditions under which the above algorithm will find the clustering that minimizes the number of disagreements among the observed entries. We provide both deterministic/worst-case guarantees, and average case guarantees for a natural randomness assumption. Let $\mathbf{K}^*$ be the low-rank matrix corresponding to the optimal clustering (as described above). Let $\mathbf{B}^* = \mathcal{P}_{\Omega_{\text{obs}}}(\mathbf{A} + \mathbf{I} - \mathbf{K})$ be the matrix of observed disagreements for this clustering. Note that the support of $\mathbf{B}^*$ is contained in Ω_{obs}. Let $K_{\min}$ be the size of the smallest cluster in $\mathbf{K}^*$.

Deterministic guarantees: We first provide deterministic conditions under which (1) will find $\mathbf{K}^*$. For any node i, let $C(i)$ be the cluster in $\mathbf{K}^*$ that node i belongs to. For any cluster $c \neq C(i)$, define $d_{i,c} = |\{j \in c \,|\, a_{ij} = ? \text{ or } a_{ij} = 1\}|$ and for $c = C(i)$, define $d_{i,c} = |\{j \in c \,|\, a_{ij} = ? \text{ or } a_{ij} = 0\}|$. In words, for both cases, $d_{i,c}$ is the total number of disagreements and unobserved entries between i and c. We now define a quantity D_{max} as follows

$$D_{\max} \;=\; \max_{i,c} \frac{d_{i,c}}{\min\{|c|, C(i)\}}$$

Essentially, $D_{\max}$ is the largest *fraction* of "bad entries" (i.e. disagreements or unobserved) between a node and a cluster. Thus for the same $D_{\max}$, a node is allowed to have more bad entries to a larger cluster, but constrained to have a smaller to a smaller cluster. It is intuitively clear that a large D_{max} will cause problems, as a node will have so many disagreements (with respect to the corresponding cluster size) that it will be impossible to resolve. We now state our main theorem for the deterministic case.

Theorem 2 *If* $\frac{nD_{\max}}{K_{\min}} < \frac{1}{4}$*, then the optimal clustering* $(\mathbf{K}^*, \mathbf{B}^*)$ *is the unique solution of (1) for any*

$$\eta \in \left(\frac{1}{1 + \frac{1}{2}K_{\min}} \,,\, 1 - \frac{K_{\min}}{\left(1 + \frac{3}{4nD_{\max}}\right) K_{\min} - 1} \right).$$

Remarks on Theorem 2: Essentially, Theorem 2 allows for the number of disagreements and unobserved edges at a node to be as large as a third of the number of "good" edges (i.e. edges to its own cluster in the optimal clustering). This means that there is a lot of evidence "against" the optimal clustering, and missing evidence, making it that much harder to find. Theorem 2 allows a node to have many disagreements and unobserved edges overall; it just requires these to be distributed proportional to the cluster sizes.
In many applications, the size of the typical cluster may be much smaller than the size of the graph. Theorem 2 implies that the smallest cluster $K_{\min} > 4\sqrt{n}$ for any non-trivial problem (i.e. one where every cluster has at least one node with at least one disagreement or unobserved edge). Our method can thus handle as many as $\Theta(\sqrt{n})$ clusters; this can be compared to existing approaches to graph clustering, which often partition nodes into two or a constant number of clusters. The guarantees of this theorem are orderwise stronger than what would result from a direct application of the deterministic guarantees in (Chandrasekaran et al., 2009; Hsu et al., 2010). Indeed, the results in (Hsu et al., 2010) implies correct recovery as long as $D_{\max} \leq c\frac{K_{\min}^2}{n^2}$ for some constant c. (This result subsumes those in (Chandrasekaran et al., 2009).) Theorem 2 only requires $D_{\max} < \frac{K_{\min}}{4n}$, which is an order improvement if $K_{\min}$ grows slower than n.

Probabilistic Guarantees: We now provide much stronger guarantees for the case where both the locations of the observations, and the locations of the observed disagreements, are drawn uniformly at random. Specifically, consider a graph that is generated as follows: start with an initial "ideally clustered" graph with no disagreements – i.e. each cluster is completely connected (i.e. a full clique), and different clusters are completely disconnected (i.e. have no edges between them). Then for some $0 < \tau < 1$ and for each of the $\binom{n}{2}$ possible node pairs, flip the entry in this location with probability τ from 0 to 1 or 1 to 0, as the case may be – thus causing them to be disagreements. There are thus, on average, $\tau\binom{n}{2}$ disagreements in the resulting graph. The actual number is close to this with high probability, by standard concentration arguments. Further, this graph is observed at locations chosen uniformly at random. Specifically, for each node pair (i, j) there is a probability p_0 that $(i, j) \in \Omega_{\text{obs}}$, and this choice is made independently of any other node pair, or of the graph. Note that now it may be possible that the optimal clustering is not the original ideal clustering we started with; the following theorem says that we will still find the optimal clustering with high probability.

Theorem 3 *For any constant $c > 0$, there exist constants C_d, C_k, such that, with probability at least $1 - cn^{-10}$, the optimal clustering $(\mathbf{K}^*,\ \mathbf{B}^*)$ is the unique solution of (1) with $\eta = \frac{1}{1+\sqrt{np_0}}$ provided that*

$$\tau \leq C_d \qquad \textit{and} \qquad K_{\min} \geq C_k\sqrt{n(\log n)^4/p_0}.$$

Remarks on Theorem 3: This shows that our algorithm will succeed in the overwhelming majority of instances where as large as a constant fraction of all observations are disagreements. In particular the number of disagreements can be an order of magnitude larger than the number of "good" edges (i.e. those that agree with the clustering).

This remains true even if we observe a vanishingly small fraction of the total number of node pairs – p_0 above is allowed to be a function of n. Smaller p_0 however requires $K_{\min}$ to be correspondingly larger. The reason underlying these stronger results is that bounded matrices with random supports are very spectrally diffuse, and thus find it hard to "hide" a clique, which is highly structured. When p_0 is a constant, our theorem and the probabilistic guarantees in (Candes et al., 2009) can both handle the same value of corrupted fraction τ. However, our theorem goes beyond (Candes et al., 2009) as we allow p_0 to be a vanishing function of n.

Remarks on Outliers: Our algorithm has the capability to handle outliers (i.e., isolated nodes outside the true clusters with at most $D_{\max}|c|$ edges to each true cluster c) by classifying all their edges as disagreements - and hence automatically revealing each outlier as a single-node cluster. In the output of our algorithm, the low rank part $\mathbf{K}$ will have all zeroes in columns corresponding to outliers – all their edges will appear in the disagreement matrix $\mathbf{B}$.

3. Proof of Theorem 1

In this section, we prove Theorem 1; in particular, that if (1) produces a valid low-rank matrix, i.e. one that corresponds to a clustering of the nodes, then this is the disagreement minimizing clustering. Consider the following non-convex optimization problem

$$\begin{aligned} \min_{\mathbf{B},\mathbf{K}} \quad & \eta\,||\mathbf{B}||_1 + (1-\eta)\,||\mathbf{K}||_* \qquad (2) \\ \text{s.t.} \quad & P_{\Omega_{\text{obs}}}(\mathbf{B}+\mathbf{K}) = P_{\Omega_{\text{obs}}}(\mathbf{I}+\mathbf{A}) \\ & \mathbf{K} \text{ is valid} \end{aligned}$$

and let $(\mathbf{B},\mathbf{K})$ be any feasible solution. Since $\mathbf{K}$ represents a valid clustering, it is positive semidefinite and has all ones along its diagonal. Therefore, any valid $\mathbf{K}$ obeys $\|\mathbf{K}\|_* = \text{trace}(\mathbf{K}) = n$. On the other hand, because both $\mathbf{K}-\mathbf{I}$ and $\mathbf{A}$ are adjacency matrices, the entries of $\mathbf{B}=\mathbf{I}+\mathbf{A}-\mathbf{K}$ must be equal to -1, 1 or 0 (i.e. it is a disagreement matrix). Hence $\|\mathbf{B}\|_1 = \|\mathbf{B}\|_0$ when $\mathbf{K}$ is valid. We thus conclude that the above optimization problem is equivalent to minimizing $||\mathbf{B}||_0$ s.t. the constraints in (2) hold. This is exactly the minimization of the number of disagreements on the observed edges. Now notice that (1) is a relaxed version (2). Therefore, if the optimum of (1) is valid and feasible to (2), then it is also optimal to (2).

4. Proof Outline for Theorem 2 and 3

We now overview the main steps in the proof of Theorem 2 and 3; the following sections provide details. Recall that we would like to show that $\mathbf{K}^*$ and $\mathbf{B}^*$ corresponding to the optimal clustering is the unique optimum of our convex program (1). This involves the following steps:

Step 1: Write down sub-gradient based first-order sufficient conditions that need to be satisfied for $\mathbf{K}^*, \mathbf{B}^*$ to be the unique optimum of (1). In our case, this involves showing the existence of a matrix $\mathcal{Q}$ – the *dual certificate* – that satisfies certain properties. This step is technically involved – requiring us to delve into the intricacies of sub-gradients since our convex function is not smooth – but otherwise standard. Luckily for us, this has been done by (Chandrasekaran et al., 2009; Candes et al., 2009).

Step 2: Using the assumptions made on the optimal clustering and its disagreements $(\mathbf{K}^*, \mathbf{B}^*)$, construct a candidate dual certificate $\mathcal{Q}$ that meets the requirements – and thus certifies $\mathbf{K}^*, \mathbf{B}^*$ as being the unique optimum. This is where the "art" of the proof lies: different assumptions on the $\mathbf{K}^*, \mathbf{B}^*$ (e.g. we look at deterministic and random assumptions) and different ways to construct this $\mathcal{Q}$ will result in different performance guarantees.

The crucial second step is where we go beyond the existing literature on matrix splitting (Chandrasekaran et al., 2009; Candes et al., 2009). In particular, our sparse and low-rank matrices have a lot of additional structure, and we use some of this in new ways to generate dual certificates. This leads to much more powerful performance guarantees than those that could be obtained via a direct application of existing sparse and low-rank matrix splitting results.

4.1. Preliminaries

Definitions related to $\mathbf{K}^*$: By symmetry, the SVD of $\mathbf{K}^*$ is of the form $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$. We define the sub-space $\mathcal{T} = \left\{\mathbf{U}\mathbf{X}^T + \mathbf{Y}\mathbf{U}^T : \mathbf{X},\mathbf{Y} \in \mathbb{R}^{n\times p}\right\}$ to be the span of all matrices that share either the same column space or the same row space as $\mathbf{K}^*$. For any matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$, we can define its *orthogonal projection* to the space $\mathcal{T}$ as $\mathcal{P}_{\mathcal{T}}(\mathbf{M}) = \mathbf{U}\mathbf{U}^T\mathbf{M} + \mathbf{M}\mathbf{U}\mathbf{U}^T - \mathbf{U}\mathbf{U}^T\mathbf{M}\mathbf{U}\mathbf{U}^T$. We also define the projection onto $\mathcal{T}^\perp$, the complement orthogonal space of $\mathcal{T}$, as $\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{M}) = \mathbf{M} - P_{\mathcal{T}}(\mathbf{M})$.

Definitions related to $\mathbf{B}^*$: For any matrix $\mathbf{M}$ define its support set as $\text{supp}(\mathbf{M}) = \{(i,j) : m_{i,j} \neq 0\}$. Let $\Omega = \{\mathbf{B} \in \mathbb{R}^{n\times n} : \text{supp}(\mathbf{B}) \subseteq \text{supp}(\mathbf{B}^*)\}$ be the space of matrices with support sets that are a subset of the support set of $\mathbf{B}^*$. Let $\mathcal{P}_\Omega(\mathbf{N}) \in \mathbb{R}^{n\times n}$ be the orthogonal projection of the matrix $\mathbf{N}$ onto the space Ω, i.e., $\mathcal{P}_\Omega(\mathbf{N})$ is obtained from $\mathbf{N}$ by setting all entries not in the set $\text{supp}(\mathbf{B}^*)$ to zero. Let $\Omega^\perp$ be the orthogonal space to Ω – it is the space of all matrices whose entries in the set $\text{supp}(\mathbf{B}^*)$ are zero. The projection $\mathcal{P}_{\Omega^\perp}$ is defined accordingly. Finally, let $\text{sgn}(\mathbf{B}^*)$ be the matrix whose entries are +1 for every positive entry in $\mathbf{B}^*$, -1 for every negative entry, and 0 for all the zero entries.

Definitions related to partial observations: Let Ω_{obs} be the space of matrices with support sets that are a subset of

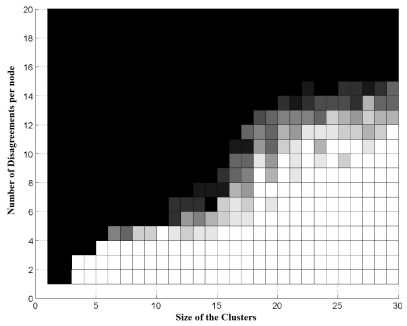

Figure 2. Simulation results for fully observed 1000-node graph with all clusters of the same size. For different cluster sizes $K_{\min}$ and different number of disagreements per node b, we plot the probability of success.

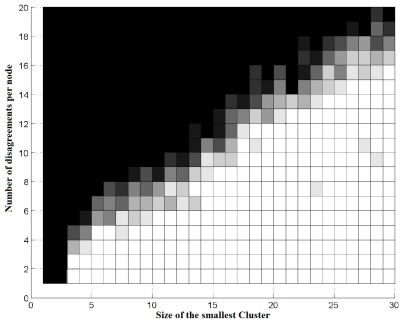

Figure 3. Simulation results for fully observed 1000-node graph with cluster of non-uniform sizes. The graph has clusters of at least size k. For different minimum cluster size $K_{\min}$ and number of disgreement per node b, we plot the probability of success.

the set of observed entries, and $\Gamma = \Omega^{\perp} \cap \Omega_{\text{obs}}$ is the set of matrices with support within the set of observed entries but outside the set of disgreements. Accordingly, define $\mathcal{P}_{\Omega_{\text{obs}}}$, $\mathcal{P}_{\Omega^{\perp}_{\text{obs}}}$, $\mathcal{P}_{\Gamma}$ and $\mathcal{P}_{\Gamma^{\perp}}$ similar to that of $\mathcal{P}_{\Omega}$ and $\mathcal{P}_{\Omega^{\perp}}$.

Norms: $\|\mathbf{M}\|$ and $\|\mathbf{M}\|_F$ represent the spectral and Frobenius norm of the matrix $\mathbf{M}$ respectively and $\|\mathbf{M}\|_{\infty} = \max_{i,j} |m_{i,j}|$.

5. Worst Case Analysis

In this section, we prove Theorem 2. We first state the determinisic first-order conditions required for $\mathbf{B}^*$ and $\mathbf{K}^*$ to be the unique optimum of our convex program (1).

Lemma 1 (Deterministic Sufficient Optimality) *(Chandrasekaran et al., 2009)* $\mathbf{B}^*$ *and* $\mathbf{K}^*$ *are unique solutions to* (1) *provided that* $\mathcal{T} \cap \Gamma^{\perp} = \{\mathbf{0}\}$ *and there exists a matrix* $\mathcal{Q}$ *such that*

(a). $\mathcal{P}_{\Omega^{\perp}_{obs}}(\mathcal{Q}) = 0$; (b). $\mathcal{P}_{\mathcal{T}}(\mathcal{Q}) = (1-\eta)\mathbf{U}\mathbf{U}^T$;
(c). $\mathcal{P}_{\Omega}(\mathcal{Q}) = \eta \operatorname{sgn}(\mathbf{B}^*)$; (d). $\|\mathcal{P}_{\mathcal{T}^{\perp}}(\mathcal{Q})\| < 1-\eta$;
(e). $\|\mathcal{P}_{\Omega^{\perp}}(\mathcal{Q})\|_{\infty} < \eta$.

The first condition, $\mathcal{T} \cap \Gamma^{\perp} = \{\mathbf{0}\}$, is satisfied under the assumption of the theorem; the proof follows from showing $\|\mathcal{P}_{\mathcal{T}}(\mathcal{P}_{\Gamma^{\perp}}(\mathbf{N}))\|_{\infty} < \|\mathbf{N}\|_{\infty}$. Next, we need to construct a suitable dual certificate $\mathcal{Q}$ that satisfies condition (a)-(e). We use the alternating projection method (see (Candes & Recht, 2009)) to construct $\mathcal{Q}$. The novelty of our analysis is that by taking advantage of the rich structures of the matrices $\mathbf{B}^*$ and $\mathbf{K}^*$, such as symmetricity, block-diagonal, etc, we improve the existing guarantees (Chandrasekaran et al., 2009; Candes et al., 2009) to a much larger class of matrices.

Dual Certificate Construction: For $\mathbf{M} \in \Gamma^{\perp}$ and $\mathbf{N} \in \mathcal{T}$, consider the infinite sums

$$\mathbf{S_M} = \mathbf{M} - \mathcal{P}_{\mathcal{T}}(\mathbf{M}) + \mathcal{P}_{\Gamma^{\perp}}\mathcal{P}_{\mathcal{T}}(\mathbf{M}) - \mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma^{\perp}}\mathcal{P}_{\mathcal{T}}(\mathbf{M}) + \cdots$$
$$\mathbf{V_N} = \mathbf{N} - \mathcal{P}_{\Gamma^{\perp}}(\mathbf{N}) + \mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma^{\perp}}(\mathbf{N}) - \mathcal{P}_{\Gamma^{\perp}}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma^{\perp}}(\mathbf{N}) + \cdots$$

Provided that these two sums converge, let $\mathcal{Q} = (1-\eta)\mathbf{V}_{\mathbf{U}\mathbf{U}^T} + \eta \mathbf{S}_{\operatorname{sgn}(\mathbf{B}^*)}$. It is easy to check that the equality conditions in Lemma 1 are satisfied. It remains to show that (i) the sums converge and (ii) the inequality conditions in Lemma 1 are satisfied. The proof again requires suitable bounds on $\|\mathcal{P}_{\mathcal{T}}(\mathcal{P}_{\Gamma^{\perp}}(\mathbf{N}))\|_{\infty}$, as well as on $\|\mathcal{P}_{\Gamma^{\perp}}\mathbf{M}\|$, which crucially depend on the assumptions imposed on $\mathbf{K}^*$ and $\mathbf{B}^*$; see supplementary materials. Combining the above discussion establishes the theorem.

6. Average Case Analysis

In this section, we prove Theorem 3. We first state the probabilistic first-order conditions required for $\mathbf{B}^*$ and $\mathbf{K}^*$ to be the unique optimum of (1) with high probability. By *with high probability* we mean with probability at least $1 - cn^{-10}$ for some constant $c > 0$.

Lemma 2 (Probabilistic Sufficient Optimality) *(Candes et al., 2009) Under the assumptions of Theorem 3,* $\mathbf{K}^*$ *and* $\mathbf{B}^*$ *are unique solutions to* (1) *with high proability provided that there exists* $\mathcal{Q} = \mathbf{W}^B + \mathbf{W}^K$ *such that*

(S1) $\left\|\mathcal{P}_{\mathcal{T}}(\mathbf{W}^B)\right\|_F \leq \frac{1}{2n^2}$. *(L1)* $\left\|\mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{W}^K)\right\| < \frac{1}{4}$.
(S2) $\left\|\mathcal{P}_{\mathcal{T}^{\perp}}(\mathbf{W}^B)\right\| < \frac{1}{4}$. *(L2)* $\left\|\mathcal{P}_{\mathcal{T}}(\mathbf{W}^K) - \mathbf{U}\mathbf{U}^T\right\|_F \leq \frac{1}{2n^2}$.
(S3) $\mathcal{P}_{\Omega}(\mathbf{W}^B) = \frac{\eta}{1-\eta}\operatorname{sgn}(\mathbf{B}^*)$
(S4) $\mathcal{P}_{\Omega^{\perp}_{obs}}(\mathbf{W}^B) = 0$. *(L3)* $\mathcal{P}_{\Gamma^{\perp}}(\mathbf{W}^K) = 0$.
(S5) $\left\|\mathcal{P}_{\Gamma}(\mathbf{W}^B)\right\|_{\infty} < \frac{1}{4}\frac{\eta}{1-\eta}$. *(L4)* $\left\|\mathcal{P}_{\Gamma}(\mathbf{W}^K)\right\|_{\infty} < \frac{1}{4}\frac{\eta}{1-\eta}$.

Dual Certificate construction: We used the so-called Golfing Scheme ((Candes et al., 2009; Gross, 2009)) to construct $(\mathbf{W}^{\mathrm{B}}, \mathbf{W}^{\mathrm{K}})$. Our application of Golfing Scheme

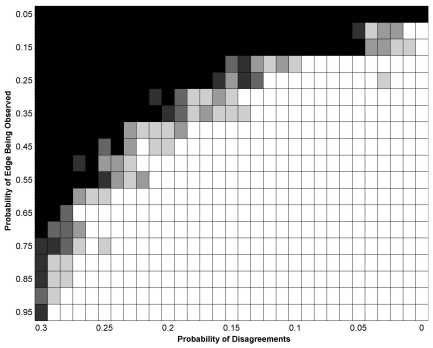

Figure 4. Simulation results for partially observed 400-node network with minimum cluster size fixed at $K_{\min} = 60$. Disagreements are placed on each (potential) edge with probability τ, and each edge is observed with probability p_0. The figure shows the probability of success in recovering the ideal cluster under different τ and p_0. Brighter colors show higher success.

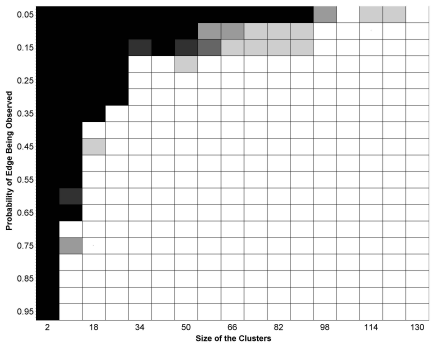

Figure 5. Simulation results for partially observed 400-node network with fixed probability $\tau = 0.04$ of placing a disgreement, and different $K_{\min}$ and p_0.

is different from (Candes et al., 2009), and the proof utilizes additional structure in our problem, which leads to stronger guarantees. In particular, we go beyond existing results by allowing the fraction of observed entries to be vanishing.

With slight abuse of notation, we use Ω_{obs}, Γ, and Ω to denote both the spaces of matrices, as well as the sets of indices these matrices are supported on. By definition, Γ (as a set of indices) contains each entry index with probability $p_0(1-\tau)$. Observe that Γ may be considered to be generated by $\cup_{1\le k\le k_0}\Gamma_i$, where each Γ_i contains each entry with probability q independent of all others, where q and k_0 are suitably chosen. For $1 \le k \le k_0$, define the operator $\mathcal{R}_{\Gamma_k}$ by

$$\mathcal{R}_{\Gamma_k}(\mathbf{M}) = \sum_{i=1}^{n} m_{i,i} e_i e_i^T + q^{-1} \sum_{1\le i<j\le n} \delta_{ij}^{(k)} m_{i,j} \left(e_i e_j^T + e_j e_i^T\right),$$

where, $\delta_{ij}^{(k)} = 1$ if $(i,j) \in \Gamma_k$ and 0 otherwise, and e_i is the i-th standard basis – i.e., the $n \times 1$ column vector with 1 in its i-th entry and 0 elsewhere. $\mathbf{W}^{\text{B}}$ and $\mathbf{W}^{\text{K}}$ are defined as

$$\mathbf{W}^{\text{B}} = \mathbf{W}_{k_0}^{\text{B}} + \frac{\eta}{1-\eta}\operatorname{sgn}(\mathbf{B}^*), \quad \mathbf{W}^{\text{K}} = \mathbf{W}_{k_0}^{\text{K}},$$

where, $(\mathbf{W}_{k_0}^{\text{K}}, \mathbf{W}_{k_0}^{\text{B}})$ is defined recursively by setting $\mathbf{W}_0^{\text{B}} = \mathbf{W}_0^{\text{K}} = 0$ and for all $k = 1, 2, \ldots, k_0$,

$$\mathbf{W}_k^{\text{B}} = \mathbf{W}_{k-1}^{\text{B}} - \mathcal{R}_{\Gamma_k}\mathcal{P}_{\mathcal{T}}\left(\frac{\eta}{1-\eta}\mathcal{P}_{\mathcal{T}}(\operatorname{sgn}(\mathbf{B}^*)) + \mathbf{W}_{k-1}^{\text{B}}\right)$$
$$\mathbf{W}_k^{\text{K}} = \mathbf{W}_{k-1}^{\text{K}} + \mathcal{R}_{\Gamma_k}\mathcal{P}_{\mathcal{T}}\left(\mathbf{U}\mathbf{U}^T - \mathbf{W}_{k-1}^{\text{K}}\right).$$

It is straightforward to verify that the equality constraints in Lemma 2 are satisfied. Moreover, $\mathbf{W}^{\text{K}}$ satisfies the inequality constraints. The proof is nearly identical to that of Y^L in section 7.3 in (Candes et al., 2009). It remains to prove that $\mathbf{W}^{\text{B}}$ also satisfies the corresponding inequalities in Lemma 2. As in the worst case analysis, the proof involves upper-bounding the norms of matrices after certain (random) linear tranformations, such as $\|\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma_k}\mathcal{P}_{\mathcal{T}}(\mathbf{M})\|$, $\|\mathcal{P}_{\Gamma_k}(\mathbf{M})\|$, $\|\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Gamma_k}\mathcal{P}_{\mathcal{T}}(\mathbf{M})\|_\infty$, and $\|\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}(\operatorname{sgn}(\mathbf{B}^*))\|_\infty$. These bounds are proven again using the assumptions imposed on $\mathbf{B}^*$, $\mathbf{K}^*$, and Ω_{obs}.

7. Experimental Results

We explore the performance of our algorithm as a various graph parameters of interest via simulation. We see that the performance matches well with the theory.

We first verify our deterministic guarantees for fully observed graphs and consider two cases: (1) all clusters have the same size equal to $K_{\min}$, and the number of disagreements involving each node is fixed at b across all nodes; (2) b is again fixed, but clusters may have different sizes no smaller than $K_{\min}$. For each pair $(b, K_{\min})$, a graph is picked randomly from all graphs with the desired property, and we use our algorithm to find $\mathbf{K}^*$ and $\mathbf{B}^*$. The optimization problem (1) is solved using the fast algorithm in (Lin et al., 2009) with η set via line search with step size 0.01. We check if the solution is a valid clustering and is equal to the underlying ideal cluster. The experiment is repeated for 10 times and we plot the probability of success in Fig. 2 and 3. One can see that the margin of the number of disagreements is higher in the second case, as these graphs have typically larger clusters than in the first case.

We next consider partially observed graphs. A test case is constructed by generating a 400-node graph with equal cluster size $K_{\min}$, and then placing a disagreement on each (potiential) edge with probability τ, independent of all others. Each edge is observed with probability p_0. In the first set of experiments, we fix $K_{\min} = 60$ and vary (p_0, τ). The

probability of success is ploted in Fig. 6. The second set of experiments have fixed $\tau = 0.04$ and different $(p_0, K_{\min})$, with results ploted in Fig. 6. One can see that our algorithm succeeds with p_0 as small as 10% and the average number of disagreements per node being on the same order of the cluster size. We expect that the fraction of observed entries can be even smaller for larger networks, where the concentration effect is more significant.

References

Bansal, N., Blum, A., and Chawla, S. Correlation clustering. In *Proceedings of the 43rd Symposium on Foundations of Computer Science*, 2002.

Becker, H. A survey of correlation clustering. Available online at http://www1.cs.columbia.edu/ hila/clustering.pdf, 2005.

Bern, M. and Eppstein, D. *Approximation Algorithms for Geometric Problems*. In Approximation Algorithms for NP-Hard Problems, edited by D. S. Hochbaum, Boston: PWS Publishing Company, 1996.

Candes, E. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 2009.

Candes, E., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? Technical report, Stanford University, CA, 2009.

Chandrasekaran, V., Sanghavi, S., Parrilo, S., and Willsky, A. Rank-sparsity incoherence for matrix decomposition. Available on arXiv:0906.2220v1, 2009.

Charikar, M., Guruswami, V., and Wirth, A. Culstering with qualitative information. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, 2003.

Condon, A. and Karp, R.M. Algorithms for graph partitioning on the planted partition model. *Random Structures & Algorithms*, 2001.

Demaine, E. D., Immorlica, N., Emmanuel, D., and Fiat, A. Correlation clustering in general weighted graphs. *SIAM special issue on approximation and online algorithms*, 2005.

Emmanuel, D. and Fiat, A. Correlation clustering minimizing disagreements on arbitrary weighted graphs. In *Proceedings of the 11th Annual European Symposium on Algorithms*, 2003.

Emmanuel, D. and Immorlica, N. Correlation clustering with partial information. In *Proceedings of the 6th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems*, 2003.

Ester, M., Kriegel, H., and Xu, X. A database interface for clustering in large spatial databases. 1995.

Everitt, B. *Cluster Analysis*. New York: Halsted Press, 1980.

Flake, G.W., Tarjan, R.E., and Tsioutsiouliklis, K. Graph clustering and minimum cut trees. *Internet Mathematics*, 2004.

Gross, D. Recovering low-rank matrices from few coefficients in any basis. Available on arXiv:0910.1879v4, 2009.

Hsu, Daniel, Kakade, Sham, and Zhang, Tong. Robust matrix decomposition with outliers. *Available at arXiv:1011.1518*, 2010.

Inc, Yahoo. Graph partitioning. Available at http://research.yahoo.com/project/2368, 2009.

Jain, A. K. and Dubes, R. C. *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

Kernighan, B.W. and Lin, S. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49(2):291–307, 1970.

Lin, Z., Chen, M., Wu, L., and Ma, Y. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. *UIUC Technical Report UILU-ENG-09-2215*, 2009.

Mishra, N., R. Schreiber, I. Stanton, and Tarjan, R. E. Clustering social networks. *Algorithms and Models for Web-Graph, Springer*, 2007.

Ng, A.Y., Jordan, M.I., and Weiss, Y. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2002.

Recht, B., Fazel, M., and Parillo, P. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. Available on arXiv:0706.4138v1, 2009.

Shi, J. and Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

Swamy, C. Correlation clustering: maximizing agreements via semidefinite programming. In *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2004.