An ADMM Algorithm for Clustering Partially Observed Networks

N. S. Aybat* S. Zarmehri† S. Kumara‡

Abstract

Community detection has attracted increasing attention during the past decade, and many algorithms have been proposed to find the underlying community structure in a given network. Many of these algorithms are based on modularity maximization, and these methods suffer from the resolution limit. In order to detect the underlying cluster structure, we propose a new convex formulation to decompose a partially observed adjacency matrix of a network into low-rank and sparse components. In such decomposition, the low-rank component encodes the cluster structure under certain assumptions. We also devise an alternating direction method of multipliers with increasing penalty sequence to solve this problem; and compare it with Louvain method, which maximizes the modularity, on some synthetic randomly generated networks. Numerical results show that our method outperforms Louvain method on the randomly generated networks when variance among cluster sizes increases. Moreover, empirical results also demonstrate that our formulation is indeed tighter than the robust PCA formulation, and is able to find the true clustering when the robust PCA formulation fails.

1 Introduction

Community detection or clustering is one of the most important topics in network science [18]. A cluster is defined loosely as a group of nodes which are more densely connected with each other than with nodes in the other groups of the network. Many clustering algorithms have been proposed to identify the underlying community structure in a given network. The goodness of the identified communities can be evaluated by a quality function [9]. Modularity is the most popular quality function [9] which was introduced by Girvan and Newman [19]. It is assumed that a higher modularity value indicates a better community structure. Although this is not always true, it has formed the motivation for developing many algorithms based on modularity maximization [9]. Modularity maximization is an NP-complete problem [6] and the algorithms are only able to find a good approximation to the global solution. Recently it has been shown that modularity maximization has some problems for large networks [15]. One major problem is due to resolution limit [10]. Many modularity based algorithms tend to merge smaller clusters with bigger ones even when the small size cluster is a clique, and it is connected to a larger cluster by a single edge [15, 10]. This problem arises from the definition of modularity and particularly from the assumption of its null model that each node can interact with any other node in the network [9]. If there are two communities with sufficiently small sizes (and hence small degrees), the expected number of edges between them for the null model is small. In this case, even the existence of an edge between the two communities can merge them together [9]. Given a partition of nodes, modularity is the sum of values, each corresponding to a group in the partition. Hence, in modularity maximization, one searches for the best partition, which is equivalent to looking for the best trade-off between the number of groups in the partition and their corresponding values. However, as discussed in [10], the partition corresponding to the highest modularity may not be correlated with the underlying unknown community structure. Indeed, there are some instances of real networks [10] and benchmark graphs [15] such that the modularity maximization fails to properly identify the community structure. Recently, it has been shown that the modularity maximization problem can have different local maxima which are structurally different but have high modularity values [11]. These solutions may disagree on many community structure properties such as the distribution of cluster sizes. This kind of disagreement may have serious impact on real world networks such as metabolic networks [11]. The need to find an accurate community structure motivates us to develop a new method for community detection which depends more on the network structure and less on the quality function.

Our method is based on convex optimization, and is inspired by the work in [7]. Suppose we have a data matrix \( D \in \mathbb{R}^{m \times n} \) which is a summation of a low rank matrix \( L \) and a sparse matrix \( S \), i.e. \( D = L + S \). Consider the following convex optimization problem:

\[
(1.1) \quad (L^*_\rho, S^*_\rho) \in \arg\min_{L, S \in \mathbb{R}^{m \times n}} \{ \|L\|_* + \rho \|S\|_1 : D = L + S \},
\]
where $\|Z\|_* := \sum_{i=1}^{\text{rank}(Z)} \sigma_i(Z)$ denotes the nuclear norm of $Z \in \mathbb{R}^{m \times n}$; i.e., sum of singular values of its argument, $\|Z\|_1 = \sum_{i=1}^m \sum_{j=1}^n |Z_{ij}|$, and $\rho = 1/\sqrt{\max\{m, n\}}$. It has been shown in [7] that under some technical conditions on $L$ and $\bar{S}$, problem in (1.1) with very high probability has a unique solution $(L^*_\rho, S^*_\rho)$ such that $(L^*_\rho, S^*_\rho) = (\bar{L}, \bar{S})$.

Suppose we are given an undirected network $G = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N} = \{1, \ldots, n\}$ and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ denote the set of nodes and edges, respectively. Suppose there are $r \ll n$ communities in $\mathcal{G}$, and $\mathcal{N}_\ell \subseteq \mathcal{N}$ denotes the subset of nodes in community-$\ell$ for $1 \leq \ell \leq r$. We assume that every node belongs to exactly one community, i.e., $\bigcup_{\ell=1}^r \mathcal{N}_\ell = \mathcal{N}$ and $\mathcal{N}_\ell \cap \mathcal{N}_\ell' = \emptyset$ for all $\ell_1 \neq \ell_2$. Let $D \in \mathbb{R}^{n \times n}$ denote the node-node incidence matrix of $\mathcal{G}$ such that $D_{ii} = 1$ for all $i \in \mathcal{N}$, $D_{ij} = 1$ if either $(i, j) \in \mathcal{E}$ or $(j, i) \in \mathcal{E}$, and $D_{ij} = 0$ otherwise. Our idea is to decompose $D$ into a low rank matrix and a sparse matrix to recover the underlying community structure in $\mathcal{G}$. Here we discuss that such decomposition is feasible under the assumption that the number of node pairs not connected by an edge in each cluster and the number of edges connecting two different clusters are both small for the underlying community structure in $\mathcal{G}$. To motivate the upcoming discussion, first consider the scenario where the subgraph of $\mathcal{G}$ restricted to $\mathcal{N}_\ell$ is a clique for all $\ell = 1, \ldots, r$, and there is no inter-community edge in $\mathcal{E}$, i.e., $\mathcal{N}_\ell \times \mathcal{N}_{\ell'} \subseteq \mathcal{E}^c$ for all $\ell_1 \neq \ell_2$ and $\mathcal{E}^c$ denotes the complement of $\mathcal{E}$. Clearly, $D$ is a block diagonal matrix with each block on the diagonal consisting of all ones and the off-diagonal blocks consisting of all zeros – from now on we refer to such matrices as block diagonal matrices of ones (BDO). Note that $D$ is a low-rank matrix such that $\text{rank}(D) = r$ and $\lambda_1(D) = |\mathcal{N}_\ell|$ for all $1 \leq \ell \leq r$, where $\{\lambda_\ell(D)\}_{\ell=1}^r$ denotes the non-zero eigenvalues of $D$. Hence, $D = L + \bar{S}$ such that low-rank component $L = D$ and sparse component $\bar{S} = 0_n$, where $0_n \in \mathbb{R}^{n \times n}$ is the matrix of zeros.

Now consider a more realistic scenario where for any $\ell \in \{1, \ldots, r\}$, a small number of node pairs from $\mathcal{N}_\ell$ may not be connected by an edge, and for any $\ell_1 \neq \ell_2$ there may be a small number of edges with one end in $\mathcal{N}_{\ell_1}$ and the other in $\mathcal{N}_{\ell_2}$, i.e., the clusters may not be cliques, and there can be inter-cluster edges. Given $\mathcal{G}$ with a non-overlapping community structure $\{\mathcal{N}_\ell\}_{\ell=1}^r$, we define $L = D - \bar{S}$ and $\bar{S}$ such that $\bar{S}_{ij}$ defined in (1.2) is sparse, due to our assumption on the underlying community structure in $\mathcal{G}$, and $\bar{L} = D - \bar{S}$ is low-rank. Indeed, $\bar{L}$ is a BDO obtained by completing the clusters into cliques and deleting the inter-cluster edges; therefore, $\bar{L}$ is low rank because each block has rank one and $\text{rank}(\bar{L})$ is equal to the number of diagonal blocks, i.e., $\text{rank}(\bar{L}) = r$.

In this paper, we propose a convex model similar to (1.1) of which optimal solution is equal to $(\bar{L}, \bar{S})$, defined as in (1.2), with very high probability. In particular, by adding constraints $L \succeq 0$, $\text{diag}(L) = 1$, $L \succeq 0$, and $|S_{ij}| \leq 1$ for $1 \leq i \neq j \leq n$ to (1.1), we will obtain a tighter convex model. Another important property of our model is its ability to handle cases where $D$ is partially observed. We will discuss this property in more detail in the next section, and develop an alternating direction method of multipliers (ADMM) algorithm for the proposed model. Finally, in Section 3, we first discuss how to generate a random family of networks for which modularity maximization fails, then compare our results with those generated by Louvain method [4], which is a greedy algorithm to solve the modularity maximization problem. While we were working on this idea independently, we found out a similar work by Chen et al. [3], which is also based on decomposition of the adjacency matrix into low-rank and sparse components by solving (1.1) when $D$ is partially observed. In the next section, we will discuss the similarities among the two methods, and emphasize the advantages of our method over the one in [3]; and compare both methods in Section 3.

Numerical results show that our convex formulation is indeed tighter than the robust PCA formulation, and when the variance among cluster sizes increases our method is able to find the true clustering while both Louvain method and the one in [3] fail.

2 Methodology

2.1 Theoretical results Model (1.1) was proposed by Candès et al. in [7], and it has been shown under some technical conditions on the components of $D$ that solving this model recovers the low rank and sparse components of the data matrix exactly with high probability - see [7] for more details. When all the data is not available, to decompose partially observed $D$, Tao and Yuan [21] proposed the following model:

$$\min_{L, S \in \mathbb{R}^{n \times n}} \{\|L\|_* + \rho \|S\|_1 : \pi_\Omega(L + S) = \pi_\Omega(D),\}$$

where $\Omega \subset \{(i, j) : 1 \leq i, j \leq n\}$ is the set of observable indices and $\pi_\Omega$ is the projection operator. For an arbitrary matrix $Z$, $(\pi_\Omega(Z))_{ij} = Z_{ij}$ when $(i, j) \in \Omega$ and $(\pi_\Omega(Z))_{ij} = 0$ otherwise. This model inspired us to develop a new method for network clustering based on convex optimization.
In order to compute a clustering for a partially observed network \( \mathcal{G} \), Chen et al. [S] proposed to solve (2.3) repeatedly for different values of \( \rho \), where \( D \) is the node-node incidence matrix of \( \mathcal{G} \) with a diagonal of ones. In particular, the authors of [S] proposed doing bisection on \( \rho \) until \( L^* \) is a BDO, which they called a valid result. The main advantage of the proposed method in this paper over the one in [8] is that we solve a tighter convex problem one time that has the same complexity with solving (1.1) one time, while in [S] the authors propose to solve (2.3) repeatedly for different values of \( \rho \). Moreover, our numerical results show that our method is not only better in computation time, but also in clustering quality. Indeed, when the network size increases and/or the variance among cluster sizes increases, the method in [S] fails to cluster correctly while ours always succeeds.

**Assumption 1.** Let \( \mathcal{G} = (\mathcal{N}, \mathcal{E}) \) be an undirected network with \( \mathcal{N} = \{1, \ldots, n\} \), and \( \mathcal{E} \subset \mathcal{N} \times \mathcal{N} \). Suppose that \( \{\mathcal{N}_i\}_{i=1}^K \) is a partition of \( \mathcal{N} \) representing the non-overlapping communities in \( \mathcal{G} \). Let \( \Omega := \Omega \cup \mathcal{D} \) denote the set of observable entries of the adjacency matrix \( D \) such that \( \text{diag}(D) = 1 \), where \( \mathcal{D} = \{(i, i) : i \in \mathcal{N}\} \), and \( \Omega \subset \mathcal{N} \times \mathcal{N} \setminus \mathcal{D} \). Assume that an edge exists between two nodes in the same cluster with probability \( p \), and it exists between two nodes from different clusters with probability \( q \) such that \( p > q \); and for any two nodes whether there is an edge between them or not is known with probability \( p_0 \).

In this paper, we propose an efficient method that can recover the underlying community structure of \( \mathcal{G} \) by decomposing \( D \) into \((\bar{L}, \bar{S})\) defined as in (1.2) with very high probability under Assumption 1. In such decomposition, the community structure is encoded in \( \bar{L} \), and the off-diagonal zero entries of \( \bar{S} \) correspond to the pair nodes that are in the same cluster but not connected by any edge, or to the edges connecting two different clusters. Following [S], the total number of off-diagonal non-zeros in \( \bar{S} \), denoted by \( \|\bar{S}\|_0 \), will be called total number of disagreements. Before we introduce our model, we investigate some properties of \((\bar{L}, \bar{S})\).

**Lemma 2.1.** Given \( \mathcal{G} = (\mathcal{N}, \mathcal{E}) \), and \( \Omega \subset \mathcal{N} \times \mathcal{N} \), define \( \chi := \{(\bar{L}, \bar{S}) \in S_n \times S_n : \text{diag}(L + S) = \pi_0(D), |S_{ij}| \leq 1 \ \forall \ 1 \leq i \neq j \leq n, \text{diag}(D) = 1, L \geq 0_n, L \geq 0_n\} \), where \( S_n \) denotes the subspace of \( n \times n \) symmetric real matrices. Under Assumption 1 we have \((\bar{L}, \bar{S}) \in \chi \).

**Proof.** Directly from the definition of \( \bar{S} \) in (1.2), and from the facts: \( \bar{L} := D - \bar{S}, \text{diag}(D) = 1 \), it follows that the first four constraints are satisfied at \((\bar{L}, \bar{S})\). Therefore it is enough to show that \( \bar{L} \) is positive semidefinite. Under Assumption 1 \( \bar{L} \) is BDO with \( r \) blocks since \( \{\mathcal{N}_i\}_{i=1}^\ell \) is a partition of \( \mathcal{N} \). Now let \( v^i \in \mathbb{R}^n \) be such that \( v^i_j = 1 \) if \( j \in \mathcal{N}_i \), and \( v^i_j = 0 \) otherwise. It is easy to show that \( \bar{L} = \sum_{i=1}^\ell v^i(v^i)^T \).

Therefore, \( \bar{L} \geq 0_n \) such that \( \text{rank} \bar{L} = r \).

Based on Lemma 2.1 we define our new model as:

\[
(L^*, S^*) \in \arg \min_{L, S \in S_n} \text{Tr}(L) + \rho \|S\|_1
\]

subject to:

\[
\text{diag}(S) = 0, |S_{ij}| \leq 1 \ \forall \ i \neq j,
\]

\[
L \geq 0_n, L \geq 0_n.
\]

Note that in (2.1), we replaced \( \|L\|_1 \) in (2.3) with \( \text{Tr}(L) \), and also replaced \( \text{diag}(L) = 1 \) constraint in \( \chi \) set with \( \text{diag}(S) = 0 \) constraint. The first follows from the fact that \( \bar{L} \geq 0_n \) implies \( \|\bar{L}\|_1 = \text{Tr}(\bar{L}) \). Indeed, for a positive semidefinite matrix \( L \), its non-zero singular values \( \{\sigma_i\}_{i=1}^{\text{rank}(L)} \) are equal to its non-zero eigenvalues \( \{\lambda_i\}_{i=1}^{\text{rank}(L)} \). Therefore, we have \( \|L\|_1 = \sum_{i=1}^{\text{rank}(L)} \lambda_i = \text{Tr}(L) \). Moreover, since \( \bar{S} = D - \bar{L} \) and \( \text{diag}(D) = 1 \), we have \( \text{diag}(S) = 0 \). It is important to note that replacing \( \text{diag}(L) = 1 \) constraint in \( \chi \) set with \( \text{diag}(S) = 0 \) is the key point for developing an ADMM algorithm with subproblems that can be solved efficiently, which will be discussed in the next section.

The following two theorems show the importance and special properties of our formulation. Theorem 2.1 and Theorem 2.2 were originally proved in [S] for model (2.3). Since the feasible region of our model in (2.4) is a subset of that in (2.3), these important results can be trivially extended for our formulation.

**Theorem 2.1.** For any \( \rho > 0 \), if \( L^* \) is a BDO, then it provides the optimal clustering which means the total number of observed disagreements, i.e. \( \|\pi_0(S^*)\|_0 \), is minimized.

**Proof.** See proof of Theorem 2 in [S].

**Theorem 2.2.** Given \( \mathcal{G} = (\mathcal{N}, \mathcal{E}) \) and \( \Omega \subset \mathcal{N} \times \mathcal{N} \) satisfying Assumption 1. Let \( \{\mathcal{N}_i\}_{i=1}^\ell \) represent the true underlying community structure of \( \mathcal{G} \), and \( \bar{L} = D - \bar{S} \), where \( \bar{S} \) is defined in (1.2). Then for all \( c > 0 \), there exists \( C > 0 \) such that with probability of at least \( 1 - cn^{-10} \), \((\bar{L}, \bar{S})\) is the unique optimal solution of (2.4) when \( \rho = \frac{1}{n^{\sqrt{0.02}}} \), provided that

\[
(2.5) \quad n \log^2 n \leq CK_{\text{min}}^2 p_0(1 - 2\gamma)^2,
\]

where \( \gamma = \max\{1 - p, q\} \) and \( K_{\text{min}} := \min\{|\mathcal{N}_\ell| : 1 \leq \ell \leq r\} \) is the size of the smallest cluster.

**Proof.** Given \( c > 0 \), Theorem 2 in [S] shows that there exists \( C > 0 \) such that \((\bar{L}, \bar{S})\) is the unique optimal solution to (2.3) with probability at least \( 1 - cn^{-10} \).
provided that \((2.5)\) holds. Lemma 2.1 implies that \((L, S)\) is feasible to \((2.4)\); hence, it must be an optimal solution to the more tighter problem in \((2.4)\) as well. Moreover, under the assumptions of Theorem 2.2 \((2.5)\) has a unique solution with high probability, which implies that \((L, S)\) is the unique optimal solution to \((2.4)\) w.p. at least \(1 - cn^{-10}\).

2.2 Algorithm

In this section, we develop an ADMM algorithm to solve \((2.4)\). Define \(\phi \in S_n \times S_n\) as

\[
(2.6) \quad \phi := \left\{ (X, S) : X \geq 0_n, \text{diag}(S) = 0, \quad |S_{ij}| \leq 1, \quad 1 \leq i \neq j \leq n \right\}.
\]

By using partial variable splitting as in \([11, 12, 2.4]\) can be written equivalently as follows:

\[
(2.7) \quad (L^*, L^*, S^*) = \arg\min_{L, X, S \in S_n} \text{Tr}(L) + \rho \|S\|_1 \text{ s.t. } X = L, \quad L \geq 0_n, \quad (X, S) \in \phi.
\]

Given a penalty parameter \(\mu > 0\), the partial augmented Lagrangian \([3]\) of \((2.7)\) is defined for any \(L \geq 0_n, \quad (X, S) \in \phi, \quad Y \in S_n\) as follows

\[
L_\mu(L, X, S; Y) = \text{Tr}(L) + \rho \|S\|_1 + (Y, X - L) - \frac{\mu}{2} \|X - L\|_F^2.
\]

Given \(Y \in S_n\), since it is not easy to minimize \(L_\mu(L, X, S; Y)\) jointly in \((L, X, S)\), method of multipliers is not a practical approach to solve \((2.7)\). On the other hand, given \(Y\), alternating minimization of \(L_\mu(L, X, S; Y)\) in \((X, S)\) for fixed \(L\), and in \(L\) for fixed \((X, S)\) can be done efficiently. Therefore, we propose ADMPIC, which is an ADMM algorithm with increasing penalty sequence, to solve \((2.7)\). Each step of ADMPIC is displayed in Figure 1. The subproblems in Step 3 and Step 4 are the computational bottlenecks, and they can be solved efficiently as explained in Lemma 2.2 and Lemma 2.3. The initialization part will be discussed in Section 3.2.

The convergence of ADMPIC directly follows from [14]. Indeed, the variable penalty ADMM algorithms in \([12, 13, 14]\) are proposed to solve variational inequalities (VI) of the form:

\[
(x - x^*)^T F(x^*) + (y - y^*)^T G(y^*) \geq 0, \quad \forall (x, y) \in \Omega,
\]

where \(A \in \mathbb{R}^{m \times n_1}, B \in \mathbb{R}^{m \times n_2}\), and \(b \in \mathbb{R}^m\). The convergence proofs in \([12, 13, 14]\) require that both \(F : \mathcal{X} \rightarrow \mathbb{R}^{n_1}\) and \(G : \mathcal{Y} \rightarrow \mathbb{R}^{n_2}\) are continuous point-to-point maps that are monotone with respect to the non-empty closed convex sets \(\mathcal{X} \subset \mathbb{R}^{n_1}\) and \(\mathcal{Y} \subset \mathbb{R}^{n_2}\), respectively. When these variable penalty ADMM methods for VI are applied to the VI reformulation of convex optimization problems of the form \(\min \{f(x) + g(y) : (x, y) \in \Omega\}\), the requirement that \(F\) and \(G\) be continuous point-to-point maps implies that \(F(x) = \nabla f(x)\), and \(G(y) = \nabla g(y)\). On the other hand, if \(f\) (similarly \(g\)) is a non-smooth convex function, then \(F\) (similarly \(G\)) should be a point-to-set map, i.e., multi-function; therefore, the convergence proofs for variable penalty ADMM algorithms in \([12, 13, 14]\) do not extend to non-smooth convex optimization problems – see Assumption A and the following discussion on page 107 in [14]. However, even though the objective in \((2.7)\) is non-smooth, the following result establishes that the convergence of ADMPIC follows directly from [14].

**Theorem 2.3.** Let \(Z_k = (L_k, X_k, S_k, Y_k)\) denote the iterates generated by ADMPIC in Figure 1 and \(Z^*\) denote the set of optimal primal-dual pairs to \((2.7)\), i.e., \((L^*, X^*, S^*, Y^*) \in Z^*\) if and only if

\[
\{L_n - Y^*, L - L^*\} \geq 0, \quad \forall L \geq 0_n, 
\]

\[
\rho(G, S - S^*) + (Y^*, X - X^*) \geq 0, \quad \forall (X, S) \in \phi, 
\]

\[
X^* = L^*, 
\]

\[
G \in \partial \|S^*\|_1.
\]

Then \(\min \{\|Z_k - Z\|_F : Z \in Z^*\} \rightarrow 0\). Moreover, \(\{Z_k\}\) is bounded.

**Proof.**

Using the change of variables \(S := S^+ - S^-\) for \(S^+, S^- \geq 0_n, \quad \|S\|_1\) can be equivalently written as \(\langle E_n, S^+ + S^-\rangle\), where \(E_n \in \mathbb{R}^{n \times n}\) is a matrix of ones. Similarly, let \(\phi' \subset \prod_{i=1}^n S_n\) such that

\[
\phi' := \left\{ (X, S^+, S^-) : \begin{aligned} 
\pi_\Omega \left( X + S^+ - S^- \right) = \pi_\Omega(D) ; 
X \geq 0_n, \quad S^+ \geq 0_n, \quad S^- \geq 0_n, 
\text{diag}(S^+) = \text{diag}(S^-) = 0, 
S^+_{ij} + S^-_{ij} \leq 1, \quad \forall 1 \neq j 
\end{aligned} \right\}.
\]

Note that \((2.8)\) is a smooth convex optimization problem equivalent to \((2.7)\), and it satisfies all the assumption in [14]. Given a nondecreasing penalty sequence \(\{\mu_k\}\) such that \(\sup_{k \leq k} \mu_k < \infty\), let \(\{L_k, X_k, S_k^+, S_k^-, Y_k\}\) be the iterate sequence generated by the variable penalty ADMM in [14] on

\[
(2.8) \quad \min \left\{ \text{Tr}(L) + \rho \langle E_n, S^+ + S^- \rangle : X = L, \quad L \geq 0_n, \quad (X, S^+, S^-) \in \phi' \right\}
\]

alternatingly minimizing in \((X, S^+, S^-) \in \phi'\), and in \(L \geq 0_n\). Define \(\tilde{Z}_k := (L_k, X_k, S_k, Y_k)\), where \(S_k := S_k^+ - S_k^-\). It is easy to see that \(\{\tilde{Z}_k\}\) would be the same with the one generated by ADMPIC in Figure 1 i.e., \(\tilde{Z}_k = Z_k\) for all \(k \geq 1\). The result follows from Theorem 4 in [14], which shows that \(\{\tilde{Z}_k\}\) is bounded and \(\min \{\|Z_k - Z\|_F : Z \in Z^*\} \rightarrow 0\).
Lemma 2.2. shows that the subproblem in Step 6 can be solved efficiently by computing a partial eigenvalue decomposition of an $n \times n$ matrix.

**Lemma 2.2.** The solution to the subproblem in Step 6 can be written in closed form:

$$L_{k+1} = W \text{ diag } \left( \max \{ \lambda_k - \mu_k^{-1} \} \right) W^T,$$

where $W \text{ diag } (\lambda_k) W^T$ is the eigenvalue decomposition of $Q_k^X := X_k + \frac{Y_k}{\mu_k}$.

**Proof.** Since $\text{Tr}(L) = (I_n, L)$, the subproblem in Step 6 can be equivalently written as

$$L_{k+1} = \arg\min_{L \succeq 0} \| L - (Q_k^X - \mu_k^{-1} I_n) \|_F.$$

Since $Q_k^X - \mu_k^{-1} I_n = W \text{ diag } (\lambda_k - \mu_k^{-1} 1) W^T$, (2.9) follows from the properties of Euclidean projection onto the positive semidefinite cone of symmetric matrices.

The main reason for using an increasing sequence of penalties $\{\mu_k\}$ in ADMIPC is because the work required for eigenvalue decomposition in Step 6 reduces significantly as fewer leading eigenvalues are needed for small values of $\mu_k$. Note that according to (2.9), we do not need to compute eigenvalues of $Q_k^X$ that are smaller than $\mu_k^{-1}$. Indeed, $Q_k^X$ may not be low rank, and many of its eigenvalues may be large during the initial iterations in the transient phase of the algorithm. This could make Step 6 an expensive operation for a constant penalty ADMM method with $\mu_k = \mu$, as there may be many leading eigenvalues that will be larger than $\mu$. However, based on (2.9), by choosing small values for $\mu_k$ during the initial iterations and then gradually increasing it, we can avoid computing all the eigenvalues of $Q_k^X$. Refer to [14] for more details about this concept.

Moreover, it is shown in [14] that results of Theorem 2.3 are still true if Step 6 is computed inexactly. Since with very high probability the optimal solution $L^*$ is unique and equal to $L$, which has low rank, Step 6 can be computed approximately by calculating only a small number of leading eigenvalues, and the approximation error will be small for all sufficiently large $k$. Indeed, Theorem 2.3 implies that with high probability $L_k \to L^*$ (due to uniqueness of $L^*$), and that for any $\delta > 0$, there exists $\{\mu_k\}$ such that $\|Q_k^X - L_k\|_F \leq \delta$ due to boundedness of $\{Y_k\}$. Hence, $\|Q_k^X - \tilde{L}\|_F \leq \delta$. Moreover, since eigenvalues of a matrix are a continuous function of its entries, previous discussion implies that for all $\epsilon > 0$, there exists $\{\mu_k\}$ such that $\|\lambda_k - \lambda\|_\infty \leq \epsilon$, where $\lambda \in \mathbb{R}^n$ denotes the vector of eigenvalues of $L$. Since the number of nonzeros in $\lambda = r \ll n$, $n - r$ components of $\lambda_k$ is between $-\epsilon$ and $\epsilon$.

In order to compute the eigenvalue decomposition of $Q_k^X$ in Step 6, we used LANSVD routine in PROPACK package. LANSVD routine is based on the Lanczos bidiagonalization algorithm with partial reorthogonalization for computing partial singular value decomposition (SVD). Let $\lambda := \lambda_k - \mu_k^{-1} 1$, and $U \text{ diag } (\sigma)^T$ denote SVD of $Q_k^X - \mu_k^{-1} I_n$, where $U$, $V$ denote left and right singular vectors corresponding to $i$-th singular value $\sigma_i$. It is clear that if $U_i^T V_i = 1$, then $\lambda_i = \sigma_i > 0$; and if $U_i^T V_i = -1$, then $\lambda_i = -\sigma_i < 0$. Hence, (2.9) can be computed efficiently using a partial SVD.

**Lemma 2.3.** Let $\Omega := \Omega \cup \mathcal{D}$ denote the set of observable entries of the adjacency matrix $D$ such that $\text{diag}(D) = 1$, where $\mathcal{D} = \{(i, i) : i \in N\}$, and $\Omega \subset N \times N \setminus \mathcal{D}$. The solution to the subproblem in Step 6 can be written in closed form:

$$\begin{align*}
C_1 &= \text{sgn} \left( \pi_0(D - Q_k^L) \right), \\
C_2 &= \max \{ |\pi_0(D - Q_k^L)| - \mu_k^{-1} E_n, 0_n \}, \\
S_{k+1} &= \min \{ \pi_0(D), \max \{ -E_n, C_1 \odot C_2 \} \}, \\
X_{k+1} &= \pi_0(D - S_{k+1}) + \max \{ \pi_0(-Q_k^L), 0_n \},
\end{align*}$$

where $Q_k^L := L_{k+1} - \frac{Y_k}{\mu_k} E_n \in \mathbb{R}^{n \times n}$ is a matrix of ones, and $\odot$ represents the component-wise multiplication operator.

**Proof.** The subproblem in step 5 can be written as

$$(X_{k+1}, S_{k+1}) = \arg\min_{(X, S) \in \phi} \rho \|S\|_1 + \frac{\mu_k}{2} \|X - Q_k^L\|_F^2.$$ 

For $(X, S) \in \phi$, we have

$$
X - Q_k^L = \pi_0(D - S - Q_k^L) + \pi_D(D - Q_k^L) + \pi_0(X - Q_k^L).
$$
Moreover, from the optimality conditions for (2.11), it is clear that $(S_{k+1})_{ij} = 0$ for all $(i, j) \in D \cup \Omega^*$. Therefore, the following problem is equivalent to (2.11).

$$\min_{(X, S) \in S_n \times S} \rho |\pi_1(S)| + h(X, S)$$

(2.13) 

s.t. 

$$0_n \leq \pi_0(X),$$

$$-\pi_0(E_n) \leq \pi_1(S) \leq \pi_0(D),$$

where $h(X, S) := \frac{\rho}{2} \|\pi_1(S) - \pi_0(D - Q_k)\| - \pi_0(X - Q_k)\|^2_f$. For the sake of notational simplicity, let $S_{ij} := (D - Q_k)_{ij}$ for all $(i, j) \in \Omega$. Note that (2.13) is separable over $(i, j)$. Therefore, for all $(i, j) \in \Omega$,

$$(S_{k+1})_{ij} = \arg \min_{S_{ij} \in \mathbb{R}} \rho |S_{ij}| + \frac{\mu_k}{2} (S_{ij} - \bar{S}_{ij})^2$$

(2.14) 

s.t. 

$$-1 \leq S_{ij} \leq D_{ij}.$$ 

Given $\bar{t} \in \mathbb{R}$ and $\mu > 0$, define $f : \mathbb{R} \to \mathbb{R}$ such that $f(t) = \rho |t| + \frac{\mu_k}{2}(t - \bar{t})$. Let $t^*_n := \arg \min_{t \in \mathbb{R}} f(t)$ and $t^*_c := \arg \min_{t \in \mathbb{R}} f(t) : \alpha \leq t \leq b$. Since $f$ is convex on $\mathbb{R}$, it is easy to show that

$$t^*_n = \text{sgn}(\bar{t}) \max \left\{ |\bar{t}| - \frac{\rho}{\mu_k}, 0 \right\},$$

$$t^*_c = \min\{b, \max\{a, t^*_n\}\}.$$

Thus, (2.15) implies that for all $(i, j) \in \Omega$, we have

$$(S_{k+1})_{ij} = \min \{ D_{ij}, \max \{ -1, c_{ij} \} \},$$

(2.16) 

$$c_{ij} = \text{sgn}(\bar{S}_{ij}) \max \left\{ |\bar{S}_{ij}| - \frac{\rho}{\mu_k}, 0 \right\}.$$ 

The structure of $S_{k+1}$ follows from (2.16) and the fact that $(S_{k+1})_{ij} = 0$ for all $(i, j) \in D \cup \Omega^*$.

Since $(X_{k+1}, S_{k+1}) \in \phi$, clearly for all $(i, j) \in \Omega$, we have $(X_{k+1})_{ij} = (D - S_{k+1})_{ij}$. Moreover, (2.12) implies that for all $(i, j) \in \Omega^*$, we have

$$(X_{k+1})_{ij} = \arg \min_{X_{ij} \geq 0} (X_{ij} - (Q_k)_{ij})^2,$$

(2.17) 

$$= \max\{(Q_k)_{ij}, 0\}.$$ 

3 Numerical results

In Section 3.3.1, we compared our formulation (2.4) with the robust PCA formulation in (2.3), which is adopted by Chen et al. in [8]. Numerical results show that our formulation is more tighter, and is able to recover many clusterings which cannot be detected using the methodology given in [8]. Next, in Section 3.3.2, we compared ADMIPC with Louvain method, which is based on modularity maximization, on randomly generated test problems. The results show that as the number of nodes in the network increases, Louvain method starts merging small clusters; and this phenomena becomes more apparent when the variation among cluster sizes gets bigger. The empirical results presented in Section 3.3.2 indeed confirm that resolution limit [10] becomes a major drawback for modularity maximization.

3.1 Random network generation. In this section we describe the random network generation used in our experiments. Let $G = (\mathcal{N}, E)$ be a random undirected network, and $\{\mathcal{N}_\ell\}_{\ell=1}^r$, a partition of $\mathcal{N}$, be the underlying clustering in $G$ chosen such that

$$\mathcal{N}_\ell := \left\{ \sum_{i=1}^{\ell-1} n_i + 1, \ldots, \sum_{i=1}^{\ell} n_i \right\}, \quad \forall \ell \in \{1, \ldots, r\},$$

where $n_\ell := |\mathcal{N}_\ell|$ denote the size of $\ell$-th cluster. Let $0 < \alpha \leq 1$ be a parameter that will control the variation among cluster sizes $\{n_\ell\}_{\ell=1}^r$. Note that $n = \sum_{\ell=1}^r \frac{1}{\alpha^{\ell-1}}. \alpha$. Given $x > 0$, let $\lceil x \rceil$ denote the nearest integer to $x$. The cluster sizes are chosen as

(3.18) $n_\ell = \left[ \frac{1 - \alpha}{1 - \alpha^\ell} \right] \alpha^{\ell-1}, \quad \forall \ell \in \{1, \ldots, r\}.$

In our experiments, we choose $r = \lceil 0.05n \rceil$. For instance, suppose $n = 100$, then the total number of clusters is $r = 5$; and in Table 3.1 we present the size of each cluster for different values of $\alpha$.

| $\alpha$ | $n_1$ | $n_2$ | $n_3$ | $n_4$ | $n_5$ |
|---|---|---|---|---|---|
| 0.1 | 20 | 20 | 20 | 20 | 20 |
| 0.9 | 24 | 22 | 20 | 18 | 16 |
| 0.8 | 30 | 24 | 19 | 15 | 12 |
| 0.7 | 36 | 25 | 18 | 12 | 9 |
| 0.6 | 43 | 26 | 16 | 9 | 6 |
| 0.5 | 52 | 26 | 13 | 6 | 3 |

Table 1: Cluster sizes for different values of $\alpha$ when $n = 100$ and $r = 5$.

In the next step, after we choose the underlying clustering $\{\mathcal{N}_\ell\}_{\ell=1}^r$ as above, we generated the edges in $G$ as follows. Let $D = \{(i, j) : i \in \mathcal{N}\}$, and $\mathcal{E} \subset \mathcal{N} \times \mathcal{N} \setminus \mathcal{D}$ be such that $|\mathcal{E}| = \lceil 0.05mn \rceil$ and the elements are randomly chosen with equal probability; define $\mathcal{E} := \{(i, j) \in \mathcal{N} \times \mathcal{N} : \exists \ell \in \{1, \ldots, r\} \text{ s.t. } i \in \mathcal{N}_\ell, j \in \mathcal{N}_\ell\}$. Then we set $\mathcal{E}$ as the symmetric difference of $\mathcal{E}$ and $\mathcal{E}^\ell$, i.e. $\mathcal{E} := \mathcal{E} \Delta \mathcal{E}$. Note that $G = (\mathcal{N}, \mathcal{E})$ generated this way will satisfy Assumption [1].

Let $D \in \mathfrak{S}_n$ be the node-node incidence matrix corresponding to $G$ such that $\text{diag}(D) = 1$, i.e. $D_{ij} = 1$ if $(i, j) \in \mathcal{E}$ or $i = j$, and $D_{ij} = 0$ otherwise. Let $\Omega \subset \mathcal{N} \times \mathcal{N}$ be the set of indices corresponding to the observable entries of $D$. Note that according to Assumption [1] we know whether an edge is in $\mathcal{E}$ or not with probability $p_0$. Hence, to generate $\Omega$, let $\Omega^U$ be the set of $N_0 := \left\lfloor \frac{\rho(n^2 - n)}{2} \right\rfloor$ indices of the upper triangular entries of $D$ chosen uniformly at random, i.e. $\Omega^U \subset \mathcal{N} \times \mathcal{N} \setminus \{(i, j) : i \leq j\}$ such that $|\Omega^U| = N_0$. Since $D$ is symmetric, the lower triangular elements symmetric to those in $\Omega^U$ should be in $\Omega$; and we also know that the diagonal entries are all ones. Therefore, we set $\Omega := \Omega^U \cup \Omega^L \cup \Omega^D$, where $\Omega^L := \{(i, j) : (j, i) \in \Omega^U\}$, and $\Omega^D := \{(i, i) : i \in \mathcal{N}\}$.
3.2 Initialization and stopping criterion. In all the experiments, we set $\rho = \frac{1}{\sqrt{n}}$, $L_0 = 0_n$, $\mu_0 = \frac{5}{\|\sigma(D)\|_1}$, and $Y_0 = \max\{\|\sigma(D)\|_1, r^{-2}\|\sigma(D)\|_1\}$ in AD-MIPC. The penalty multiplier sequence $\{\mu_k\}$ is chosen such that $\mu_{k+1} = \min\{\kappa \mu_k, \bar{\mu}\}$ for $k \geq 1$, where $\bar{\mu} = 10^7$ and $\kappa = 1.2$. We terminate the algorithm when either the following two conditions hold, or the maximum number of iterations (maxIter = 1000) is reached:

\begin{equation}
\|L_{k+1} - X_{k+1}\|_F \leq \text{tol}_p, \quad \frac{\|L_{k+1} - X_k\|_F}{\|\sigma(D)\|_F} \leq \text{tol}_d,
\end{equation}

where $\text{tol}_p = \epsilon_r \max\{\|L_{k+1}\|_F, \|X_{k+1}\|_F\}$, $\text{tol}_d = \epsilon_r \|Y_{k+1}\|_F$, and $\epsilon_r = 6 \times 10^{-4}$. The first equation gives the primal stopping criterion and the second one gives the dual stopping criterion. For more details about the stopping criteria refer to [5].

3.3 Results All experiments are conducted on HP ProLiant SL390s G7 1U server with Intel Xeon X5675 Six-Core 3.06 GHz processor using MATLAB 2013a (64 bit). We consider two different cases. In the first case, we assume that $E$ is perfectly known, i.e. all the entries of $D$ are observed. In the second case, we assume that $E$ is partially observable, i.e. we only know the entries of $D$ corresponding to indices in $\Omega$. For both cases, we compared ADMIPC with the method proposed in [8] and with Louvain method for different values of $(n, \alpha)$.

3.3.1 ADMIPC vs ADMM on RPCA First, for $\rho = \frac{1}{\sqrt{n}}$, we compare our formulation, given in (2.4), with the robust PCA (RPCA) formulation in (2.3) to check whether the proposed formulation (2.3) is tighter than (2.4). In particular, given randomly generated networks as described in Section 3.1 we solve (2.4) using ADIPC and compare the results with those obtained by solving (2.3) using a modified version of IALM in [16]. IALM is nothing but an increasing penalty ADMM method customized for (1.1). Note that IALM [16] works when $D$ is fully observed, and it does not work on (2.3). However, Theorem 1.1 in [2] shows that (2.3) is equivalent to

\begin{equation}
\min_{L, S \in \mathbb{R}^{n \times n}} \{\|L\|_* + \rho\|\sigma(S)\|_1 : L + S = \sigma(D)\};
\end{equation}

and one can easily modify IALM [16] to solve (3.20). We call the modified version as $M -$ IALM($\rho$). The results presented in this section show that, for $\rho = 1/\sqrt{n}$, our formulation (2.4) is indeed tighter than the RPCA formulation (2.3).

Next, we compared ADMIPC with the method developed in [8], which is based on RPCA formulation (2.3). We call the method in [8] as RPCA with bisection (RPCAB). RPCAB calls M-IALM on (2.3) for changing values of $\rho$. In particular, for a given $\rho > 0$, RPCAB calls M-IALM to compute $L^*_\rho$, the optimal low-rank component to (2.3). Next, if $\text{Tr}(L^*_\rho) \neq n$, then RPCAB updates $\rho$ as follows: when $\text{Tr}(L^*_\rho) > n$, then $\rho \leftarrow \rho/2$; otherwise, $\rho \leftarrow 2\rho$. After $\rho$ is updated, RPCAB calls M-IALM on (2.3) with the new $\rho$ value. As mentioned in [8], the initial value of $\rho$ is not crucial for RPCAB. In all the numerical tests we set the initial value of $\rho = \frac{1}{\sqrt{n}}$. Based on the discussion in [3] on stopping criteria for ADMM, the dual stopping criterion for M-IALM is chosen as in (3.19) such that $\text{tol}_d = \epsilon_r \|Y_{k+1}\|_F$ and the primal stopping criterion for M-IALM is chosen as $\|\pi(D) - (L_{k+1} + S_{k+1})\|_F \leq \text{tol}_p$, where $\text{tol}_p = \epsilon_r \max\{\|L_{k+1}\|_F, \|S_{k+1}\|_F, \|\sigma(D)\|_F\}$, and $\epsilon_r = 4 \times 10^{-4}$.

The following two cases are considered when we compare the low-rank component output by AD-MIPC with those generated by $M -$ IALM(1/\sqrt{n}) and by RPCAB. For each $n \in \{100, 200, 300, 400, 500\}$, random networks are generated as described in Section 3.3 for $\alpha \in \{0.5, 0.6, \ldots, 1\}$. For each $(n, \alpha)$ setting, we generated 20 random graphs, and corresponding $D$. In Case 1, all the entries of $D$ are observed, i.e., $\rho_0 = 1$, and in Case 2, we assume that $E$ is partially observable; hence, there are missing entries in $D$, i.e. $\rho_0 < 1$.

Let $\tilde{L}$ represent the underlying clustering in $G$, i.e. $\tilde{L} = D - \tilde{S}$ for $\tilde{S}$ defined in (1.2), and $L^*$ is the optimal low-rank component computed by one of the algorithms mentioned above. By definition, $\tilde{L}$ is BDO with $r$ diagonal blocks, each of size $n_\ell \times n_\ell$ for $\ell = 1, \ldots, r$. For $\ell_1, \ell_2 \in \{1, \ldots, r\}$, define matrices $\tilde{B}_{\ell_1, \ell_2} \in \mathbb{R}^{n_{\ell_1} \times n_{\ell_2}}$, and $\tilde{B}_{\ell_1, \ell_2} = (L_{ij})_{i \in N_{\ell_1}, j \in N_{\ell_2}}$. Clearly, $\tilde{B}_{\ell, \ell} = 1_{n_\ell}$ for $\ell \in \{1, \ldots, r\}$, and $\tilde{B}_{\ell, \ell_2} = 0_{n_{\ell_1} \times n_{\ell_2}}$ if $\ell_1 \neq \ell_2$. For all $1 \leq \ell_1, \ell_2 \leq r$, define $R_{\ell_1, \ell_2} := \|\tilde{B}_{\ell_1, \ell_2} - B^*_{\ell_1, \ell_2}\|_F$.

Given a random graph corresponding to $(n, \alpha)$, for each algorithm ADMIPC, M - IALM(1/\sqrt{n}) and RPCAB, we compute five different statistics. The first three statistics are the maximum, minimum and average of $\{R_{\ell_1, \ell_2}: \ell_1, \ell_2 \in \ell_1, \ell_2 \in \{1, \ldots, r\}\}$, and are denoted by $s_{\text{max}}$, $s_{\text{min}}$, and $s_{\text{avg}}$, respectively. The fourth statistic is $s_{\text{off}} := \sum_{\ell_1, \ell_2} R_{\ell_1, \ell_2}^2 / \sum_{\ell_1, \ell_2} \tilde{B}_{\ell_1, \ell_2}^2$. These first four statistics show how close $L^*$ to the true clustering encoded by the BDO matrix $\tilde{L}$. The fifth statistic $s_f$ is about the fraction of clusters recovered correctly. For $\ell \in \{1, \ldots, r\}$, define

\begin{equation}
E_{\ell} := \frac{R_{\ell, \ell}}{n_{\ell}}, \quad E_{\ell}^* := \frac{\sqrt{\sum_{\ell_1 \neq \ell} R_{\ell_1, \ell_1}^2}}{\sqrt{\sum_{\ell_1 \neq \ell} n_{\ell_1}^2}}.
\end{equation}

We call cluster $\ell$ “recovered” if $E_{\ell} < \tau_1$ and $E_{\ell}^* < \tau_2$. We set $\tau_1 = 0.4$ and $\tau_2 = 0.1$ for all three algorithms. Let $\tilde{r}$ denote the number of recovered clusters, i.e.
\(\bar{r} := \{|(\ell : E_\ell < \tau_1, E_{\ell'}^c < \tau_2)|\}\). The fifth statistic reported is \(s_f := \frac{\bar{r}}{p}\).

Given \((n, \alpha)\) and \(p_0 \in \{0.9, 0.8\}\), the underlying clustering of each 20 random graphs are estimated using ADMIPC, M-IALM, and RPCAB. Table 1, Table 2, and Table 3 report the averages of 5 statistics: \(s_{\text{max}}, s_{\text{min}}, s_{\text{av}}, s_{\text{off}}, s_f\), over the 20 instances for \(p_0 = 1\), \(p_0 = 0.9\) and \(p_0 = 0.8\), respectively. Numerical results show that increasing \(n\), and/or decreasing \(\alpha\) adversely affect the performance of all three methods. However, the negative impact is more serious for M-IALM and RPCAB. Indeed, the results corresponding to \(s_f\) statistic show that while ADMIPC can detect more than 80% of clusters all the time, \(s_f\) values for M-IALM, and RPCAB decreases significantly (there are many instances for which \(s_f\) is 0) when \(n\), \(\alpha\), and \(p_0\) changes as discussed above. By investigating the results carefully, we see that usually the large values of \(E_r\) cause the failure of ADMIPC and M-IALM. But for RPCAB, the failure is mainly due to \(E_r > \tau_2\). Although \(E_r\) are not reported, one can drive this result by comparing \(s_{\text{off}}\) values corresponding to different scenarios.

### 3.3.2 Clustering with Convex Optimization vs Modularity Maximization

In this section, we compare ADMIPC with Louvain method \[\text{(1)}\]. Let \(\tau_d = 0.05\) and \(L^*\) be the optimal low-rank component computed by ADMIPC. If \(|L_{ii}^* - 1| > \tau_d\) for some \(i = 1, \ldots, n\), we declare failure. Otherwise, a new matrix \(T^* \in \mathbb{S}_n\) is constructed as follows: \(T_{ij}^* = 1\) if \(L_{ij}^* > \bar{r}\), and \(T_{ij}^* = 0\) otherwise, where \(\bar{r} = 0.55\). Based on the new matrix, \(T^*\), we put nodes \(i\) and \(j\) in the same cluster if \(T_{ij}^* = 1\). We compare the clusterings generated by ADMIPC, and Louvain method with the ground truth. The three different measures of similarity are used to assess the results.

Let \(\mathcal{G} = (\mathcal{N}, \mathcal{E})\) denote the network, \(\mathcal{C} = \{\mathcal{N}_i\}_{i=1}^r\), which is a partition of \(\mathcal{N}\), represent the ground truth, and \(\mathcal{C}' = \{\mathcal{N}_j\}_{j=1}^{r'}\) represent the clustering computed by an algorithm. The following measures are used:

1. **Jaccard’s index:** Let \(a\) be the number of node pairs that belong to the same clusters in both \(\mathcal{C}\) and \(\mathcal{C}'\), \(b\) be the number of pairs that are in the same cluster in \(\mathcal{C}\) but not in different clusters in \(\mathcal{C}'\), and \(c\) be the number of pairs that are in the same cluster in \(\mathcal{C}'\) but not in different clusters in \(\mathcal{C}\). The Jaccard’s index is defined as \(\frac{a}{a+b+c}\). It has many applications in geology and ecology \[23\]; but it is a sensitive measure \[17\]. The Jaccard’s index is in \([0, 1]\) interval. It is 1 when \(\mathcal{C}\) and \(\mathcal{C}'\) are exactly the same, and equal to 0 when there is no common pair classified in the same cluster in both \(\mathcal{C}\) and \(\mathcal{C}'\).

2. **Normalized Mutual Information (NMI):** This measure is based on information theory, and it quantifies the reduction in our uncertainty about one cluster if we know the other one \[22\]. Define \(n_i := |\mathcal{N}_i|\) for \(i = 1, \ldots, r\), \(n_j := |\mathcal{N}_j|\) for \(j = 1, \ldots, r'\), and \(m_{ij} := |\mathcal{N}_i \cap \mathcal{N}_j|\) for all \(i, j\). Then the mutual information \(I(\mathcal{C}, \mathcal{C}') := \sum_{i=1}^{r} \sum_{j=1}^{r'} \frac{m_{ij}}{n} \log_2 \left( \frac{m_{ij}/n}{n_i/n_j} \right) \). By normalizing the mutual information, we force it to be between fixed ranges as well as improving its sensitivity \[22\] \[24\]. Let \(H(\mathcal{C}) := -\sum_{i=1}^{r} \frac{n_i}{n} \log_2 \left( \frac{n_i}{n} \right)\) represent the entropy associated with clustering \(\mathcal{C}\) and \(H(\mathcal{C}') := -\sum_{j=1}^{r'} \frac{n_j}{n} \log_2 \left( \frac{n_j}{n} \right)\) represent the entropy associated with clustering \(\mathcal{C}'\). There are different ways of normalizing but we use the method introduced by Strehl and Ghosh \[20\]. In this method, \(\mathcal{N} I_{\mathcal{SG}} = \frac{I(\mathcal{C}, \mathcal{C}')}{\sqrt{H(\mathcal{C}) H(\mathcal{C}')}}\) is between 0 and 1.

3. **Portion of Exactly Recovered Clusters (PERC):** This measure is a secondary measure and we introduced it to make the comparison in case of tightness in the other measures. Let \(\bar{r}\) represent the total number of clusters that are both in \(\mathcal{C}\) and \(\mathcal{C}'\), i.e. the number of clusters identified by the algorithm correctly. Then PERC is equal to \(\frac{\bar{r}}{\hat{r}}\). For each \(n \in \{100, 200, 300, 400, 500\}\) random networks are generated as described in Section 3.1 for \(\alpha \in \{0.5, 0.6, \ldots, 1\}\). For each \((n, \alpha)\) setting, we generated 20 random graphs, and the corresponding \(D\). Then we compute two clusterings using ADMIPC and Louvain method. When \(p_0 < 1\), if an edge is not observable, then we set the corresponding entry to 0 in the data matrix for Louvain method.

For a fixed \((n, \alpha)\) setting and \(p_0 \in \{0.9, 0.8\}\), each of the three measures are evaluated on the 20 clusterings generated by ADMIPC corresponding to 20 random instances. The mean of these values are reported in Table 4, Table 5, and Table 6 for \(p_0 = 1\), \(p_0 = 0.9\), and \(p_0 = 0.8\), respectively. Note that the output of Louvain method depends on the initial ordering of the nodes chosen. Hence, for the same graph, this method can generate different clusterings for different ordering of nodes. Therefore, for each 20 random graphs corresponding to fixed \((n, \alpha)\), we run Louvain method for 200 different ordering of nodes (generated randomly such that each ordering is equally likely). The mean values for each three measures are reported similarly in Table 4, Table 5, and Table 6. Numerical results show that our method outperforms Louvain method almost every time for all the three measures. It is important to note that increasing \(n\) adversely affects the performance of both methods. Moreover, for Louvain method, the clustering quality decreases significantly as \(\alpha\) decreases, i.e. the variation among the cluster sizes increases, while it is not the case for our method, and the clustering quality
is not impacted significantly. Analyzing the results we find that for a fixed $n$, Louvain method tends to merge small clusters when $\alpha$ is small. By increasing $\alpha$, the variation among the cardinality of clusters decreases in the ground truth and the clusters with small sizes get bigger. Therefore, the quality of Louvain method’s output increases. On the other hand, ADMIPC does not show any trend for the first two measures for changing $\alpha$. When $\alpha$ is fixed and $n$ increases, the clustering performance of Louvain method decreases, which agrees with the discussion on resolution limit. As $n$ increases, the number of small size clusters increases in the ground truth as well, and more clusters are merged together by Louvain method. On the other hand, by increasing $n$, the number of isolated nodes which originally belong to different clusters increases for our algorithm, and hence the performance decreases. From the above discussion, there are two points which makes our algorithm more reliable than Louvain method. First, our algorithm works well even for small values of $\alpha$. Second, by increasing $n$, the performance of both algorithms decrease: Louvain method tends to merge smaller clusters, while our algorithm generates some isolated nodes. But as discussed in [17], generating some isolated nodes is less severe than merging some clusters, which makes our algorithm more reliable.

References

[1] N. S. Aybat and G. Iyengar, An alternating direction method with increasing penalty for stable principal component pursuit, arXiv:1309.6553v2, 2013.
[2] N. S. Aybat, S. Ma, and D. Goldfarb, Efficient algorithms for robust and stable principal component pursuit problems, Computational Optimization and Applications, (2014), pp. 1–29.
[3] D. P. Bertsekas, Constrained Optimization and Lagrange Multiplier Methods, Academic Press, 1982.
[4] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E.L.J.S. Mech, Fast unfolding of communities in large networks, J. Stat. Mech, (2008), p. P10008.
[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn., 3 (2011), pp. 1–122.
[6] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner, On modularity clustering, IEEE Transactions on Knowledge and Data Engineering, 20 (2008), pp. 172–188.
[7] E. J. Candès, X. Li, Y. Ma, and Wright J., Robust principle component analysis?, Journal of ACM, 58 (2011), pp. 1–37.
[8] Y. Chen, A. Jalali, S. Sanghavi, and H. Xu, Clustering partially observed graphs via convex optimization, Journal of Machine Learning Research, 15 (2014), pp. 2213–2238.
[9] S. Fortunato, Community detection in graphs, Physics Reports, 486 (2010), pp. 75–174.
[10] S. Fortunato and M. Barthélemy, Resolution limit in community detection, Proceedings of the National Academy of Sciences, 104 (2007), pp. 36–41.
[11] B. H. Good, Y. de Montjoye, and A. Clauset, Performance of modularity maximization in practical contexts, Phys. Rev. E, 81 (2010), p. 046106.
[12] B. He and H. Yang, Some convergence properties of a method of multipliers for linearly constrained monotone variational inequalities, Operations Research Letters, 23 (1998), pp. 151–161.
[13] B.S. He, H. Yang, and S.L. Wang, Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities, Journal of Optimization Theory and Applications, 106 (2000), pp. 337–356.
[14] B. S. He, L. Z. Liao, D. R. Han, and H. Yang, A new inexact alternating directions method for monotone variational inequalities, Mathematical Programming, Series A, 92 (2002), pp. 103–118.
[15] A. Lancichinetti and S. Fortunato, Limits of modularity maximization in community detection, Phys. Rev. E, 84 (2011), p. 066122.
[16] Z. Lin, M. Chen, and Y. Ma, The augmented la-grange multiplier method for exact recovery of corrupted low-rank matrices. arXiv:1009.5055v3, 2013.
[17] G. W. Milligan and D. A. Schilling, Asymptotic and finite sample characteristics of four external criterion measures, Multivariate Behavioral Research, 20 (1985), pp. 97–109.
[18] M. E. J. Newman, Detecting community structure in networks, EPJB, 38 (2004), pp. 321–330.
[19] M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E, 69 (2004), p. 026113.
[20] A. Strehl and J. Ghosh, Cluster ensembles - a knowledge reuse framework for combining multiple partitions, J. Mach. Learn. Res., 3 (2003), pp. 583–617.
[21] M. Tao and X. Yuan, Recovering low-rank and sparse components of matrices from incomplete and noisy observations, SIAM Journal on Optimization, 21 (2011), pp. 57–81.
[22] N. X. Vinh, J. Epps, and J. Bailey, Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance, J. Mach. Learn. Res., 11 (2010), pp. 2837–2854.
[23] S. Wagner and D. Wagner, Comparing clusterings-an overview, Tech. Report 2006-04, Universität Karlsruhe (TH), 2007.
[24] J. Wu, H. Xiong, and J. Chen, Adapting the right measures for k-means clustering, in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’09, ACM, 2009, pp. 877–886.
| n   | α   | s_{max} | s_{min} | s_{av} | s_{off} | s_f | s_{max} | s_{min} | s_{av} | s_{off} | s_f | s_{max} | s_{min} | s_{av} | s_{off} | s_f |
|-----|-----|---------|---------|--------|---------|-----|---------|---------|--------|---------|-----|---------|---------|--------|---------|-----|
| 100 | 1   | 0.8     | 0.1     | 100    | 0       | 0   | 0       | 0       | 0      | 0       | 0   | 0       | 0       | 0      | 0       | 0   |
|     | 0.9 | 0.7     | 0.1     | 0.1    | 0       | 0   | 0       | 0       | 0      | 0       | 0   | 0       | 0       | 0      | 0       | 0   |
|     | 0.8 | 0.6     | 0.1     | 100    | 0       | 0   | 0       | 0       | 0      | 0       | 0   | 0       | 0       | 0      | 0       | 0   |
|     | 0.7 | 0.5     | 0.1     | 0.1    | 100    | 0   | 0       | 0       | 0      | 0       | 0   | 0       | 0       | 0      | 0       | 0   |
|     | 0.6 | 0.4     | 0.1     | 0.2    | 99.9   | 0   | 0       | 0       | 0      | 0       | 0   | 0       | 0       | 0      | 0       | 0   |
|     | 0.5 | 0.3     | 0.1     | 0.4    | 99.9   | 0   | 0       | 0       | 0      | 0       | 0   | 0       | 0       | 0      | 0       | 0   |

Table 2: The mean values for 5 statistics in % when \( p_0 = 1 \).
Table 4: The mean values for 5 statistics in % when $p_0 = 0.8$.

| $n$ | $\alpha$ | $s_{\max}$ | $s_{\min}$ | $s_{av}$ | $s_{off}$ | $s_f$ | $s_{\max}$ | $s_{\min}$ | $s_{av}$ | $s_{off}$ | $s_f$ |
|-----|-----------|-------------|------------|----------|-----------|-------|-------------|------------|----------|-----------|-------|
| 1   | 0.1       | 0           | 0          | 0        | 100       |       | 9.6         | 0.1        | 3.1      | 0.1       | 100   |
|     | 0.9       | 0           | 0          | 0        | 100       |       | 14.9        | 0.1        | 5        | 0.1       | 100   |
|     | 0.8       | 0.7         | 0.1        | 0.1      | 100       |       | 95.4        | 0.2        | 23.3     | 0.8       | 100   |
|     | 0.7       | 0.4        | 0.9        | 0.1      | 100       |       | 99.9        | 0.0        | 39.6     | 0.1      | 60    |
|     | 0.6       | 0.2        | 2.7        | 0.4      | 97.1      |       | 99.9        | 0.0        | 42.4     | 0.1      | 59    |
|     | 0.5       | 0.6        | 5.7        | 0.6      | 97       |       | 99.9        | 0.0        | 53.6     | 0.1      | 46    |
|     |           | 0.6        | 8.5        | 0.6      | 93.7     |       | 99.9        | 0.0        | 62.9     | 0.1      | 37.5  |
| 200 | 1         | 0.4        | 0.1        | 0.1      | 100      |       | 86.6        | 22.5       | 53.7     | 0.3       | 30    |
|     | 0.9       | 0.3        | 1.9        | 0.1      | 100      |       | 99.9        | 0.0        | 56.7     | 0.2       | 42.5  |
|     | 0.8       | 0.4        | 5.4        | 0.4      | 96.5     |       | 99.9        | 0.0        | 60.3     | 0.2       | 40    |
|     | 0.7       | 0.2        | 10.8       | 0.5      | 92      |       | 99.9        | 0.0        | 61.9     | 0.2       | 39    |
|     | 0.6       | 0.2        | 10.3       | 0.6      | 93      |       | 99.9        | 0.0        | 69.8     | 0.2       | 30    |
|     | 0.5       | 0.5        | 8.5        | 0.6      | 93.7     |       | 99.9        | 0.0        | 62.9     | 0.1      | 37.5  |

Table 5: The mean values for 3 measures in % when $p_0 = 1$.

| $n$ | $\alpha$ | Louvain | ADMIPC |
|-----|-----------|---------|--------|
| 1   | 0.1       | 99.9    | 99.9   |
|     | 0.9       | 99.8    | 94.9   |
|     | 0.8       | 99.5    | 83.9   |
|     | 0.7       | 98.2    | 82.0   |
|     | 0.6       | 94.3    | 86.1   |
|     | 0.5       | 94.3    | 88.7   |
| 200 | 0.9       | 99.9    | 99.9   |
|     | 0.8       | 99.9    | 99.9   |
|     | 0.7       | 99.9    | 99.9   |
|     | 0.6       | 99.9    | 99.9   |
|     | 0.5       | 99.9    | 99.9   |

| $n$ | $\alpha$ | Louvain | ADMIPC |
|-----|-----------|---------|--------|
| 1   | 0.1       | 99.9    | 99.9   |
|     | 0.9       | 99.8    | 94.9   |
|     | 0.8       | 99.5    | 83.9   |
|     | 0.7       | 98.2    | 82.0   |
|     | 0.6       | 94.3    | 86.1   |
|     | 0.5       | 94.3    | 88.7   |
| 200 | 0.9       | 99.9    | 99.9   |
|     | 0.8       | 99.9    | 99.9   |
|     | 0.7       | 99.9    | 99.9   |
|     | 0.6       | 99.9    | 99.9   |
|     | 0.5       | 99.9    | 99.9   |
### Table 6: The mean values for 3 measures in % when $p_0 = 0.9$. 

| $\alpha$ | Louvain | ADMIPC |
|---|---|---|
| | 100 | 200 | 300 | 400 | 500 | 100 | 200 | 300 | 400 | 500 |
| 1 | 99.9 | 99.4 | 97.4 | 93.7 | 86.5 | 100 | 100 | 100 | 100 | 100 | 99.9 |
| 0.9 | 99.9 | 95.9 | 78.9 | 66.4 | 60.2 | 100 | 100 | 99.9 | 99.7 | 99.4 |
| 0.8 | 99.7 | 85.1 | 67.5 | 67.4 | 67.3 | 100 | 99.9 | 99.8 | 99.8 | 99.8 |
| 0.7 | 97.9 | 82.3 | 79.3 | 75.7 | 75.4 | 99.9 | 99.8 | 99.9 | 99.9 | 99.9 |
| 0.6 | 93.4 | 87.1 | 87.5 | 87.4 | 87.1 | 99.9 | 99.9 | 99.9 | 99.9 | 99.9 |
| 0.5 | 94.3 | 88.5 | 89.2 | 88.8 | 88.5 | 99.9 | 99.9 | 99.9 | 99.9 | 99.9 |

### Table 7: The mean values for 3 measures in % when $p_0 = 0.8$. 

| $\alpha$ | Louvain | ADMIPC |
|---|---|---|
| | 100 | 200 | 300 | 400 | 500 | 100 | 200 | 300 | 400 | 500 |
| 1 | 100 | 99.9 | 99.6 | 99.1 | 98.3 | 100 | 100 | 100 | 100 | 100 | 99.9 |
| 0.9 | 99.9 | 99.0 | 94.6 | 90.7 | 87.9 | 100 | 100 | 99.9 | 99.7 | 99.2 |
| 0.8 | 99.9 | 93.7 | 88.4 | 86.1 | 85.3 | 100 | 99.8 | 99.6 | 99.4 | 99.4 |
| 0.7 | 98.8 | 89.3 | 87.1 | 85.7 | 85.6 | 99.9 | 99.7 | 99.5 | 99.5 | 99.6 |
| 0.6 | 95.1 | 88.3 | 88.2 | 88.1 | 88.7 | 99.8 | 99.6 | 99.5 | 99.7 | 99.7 |
| 0.5 | 93.1 | 87.2 | 87.5 | 86.5 | 86.3 | 99.8 | 99.7 | 99.8 | 99.8 | 99.8 |

| | Louvain | ADMIPC |
|---|---|---|
| | 100 | 200 | 300 | 400 | 500 | 100 | 200 | 300 | 400 | 500 |
| 1 | 99.9 | 99.3 | 97.2 | 93.2 | 85.1 | 100 | 100 | 100 | 100 | 100 | 99.8 |
| 0.9 | 99.9 | 91.8 | 52.3 | 29.9 | 14.4 | 100 | 100 | 99 | 92.7 | 80.6 |
| 0.8 | 99.8 | 45.4 | 16.9 | 8.88 | 10.0 | 100 | 96.5 | 89.3 | 78.1 | 78.6 |
| 0.7 | 91.4 | 24.6 | 8.17 | 6.61 | 6.73 | 99 | 93 | 84 | 77.6 | 81.1 |
| 0.6 | 64.8 | 14.7 | 9.61 | 8.58 | 10.08 | 98 | 89 | 80.4 | 80.2 | 84.6 |
| 0.5 | 42.7 | 14.3 | 12.0 | 10.7 | 12.1 | 97 | 89.3 | 88.8 | 87.7 | 90 |