Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

# Covariate Regularized Community Detection in Sparse Graphs

Bowei Yan & Purnamrita Sarkar

**To cite this article:** Bowei Yan & Purnamrita Sarkar (2021) Covariate Regularized Community Detection in Sparse Graphs, Journal of the American Statistical Association, 116:534, 734-745, DOI: 10.1080/01621459.2019.1706541

**To link to this article:** https://doi.org/10.1080/01621459.2019.1706541

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 116, NO. 534, 734–745: Theory and Methods
https://doi.org/10.1080/01621459.2019.1706541



# Covariate Regularized Community Detection in Sparse Graphs

Bowei Yan and Purnamrita Sarkar

Department of Statistics and Data Sciences, University of Texas at Austin, Austin, TX

**ABSTRACT**

In this article, we investigate community detection in networks in the presence of node covariates. In many instances, covariates and networks individually only give a partial view of the cluster structure. One needs to jointly infer the full cluster structure by considering both. In statistics, an emerging body of work has been focused on combining information from both the edges in the network and the node covariates to infer community memberships. However, so far the theoretical guarantees have been established in the dense regime, where the network can lead to perfect clustering under a broad parameter regime, and hence the role of covariates is often not clear. In this article, we examine sparse networks in conjunction with finite dimensional sub-Gaussian mixtures as covariates under moderate separation conditions. In this setting each individual source can only cluster a nonvanishing fraction of nodes correctly. We propose a simple optimization framework which improves clustering accuracy when the two sources carry partial information about the cluster memberships, and hence perform poorly on their own. Our optimization problem can be solved by scalable convex optimization algorithms. With a variety of simulated and real data examples, we show that the proposed method outperforms other existing methodology. Supplementary materials for this article are available online.

**ARTICLE HISTORY**
Received June 2017
Accepted November 2019

**KEYWORDS**
Asymptotic analysis; Kernel method; Semidefinite programming; Stochastic block models; Sub-Gaussian mixture.

## 1. Introduction

Community detection in networks is a fundamental problem in machine learning and statistics. A variety of important practical problems such as analyzing socio-political ties among politicians (Gil-Mendieta and Schmidt 1996), understanding brain graphs arising from diffusion MRI data (Binkiewicz, Vogelstein, and Rohe 2017), investigating ecological relationships between different tiers of the food chain (Jacob et al. 2011) can be framed as community detection problems. Much attention has been focused on developing models and methodology to recover latent community memberships. Among generative models, the stochastic block model (SBM) (Holland, Laskey, and Leinhardt 1983) and its variants (Airoldi et al. 2008), etc. have attracted a lot of attention, since their simplicity facilitates efficient algorithms and asymptotic analysis (Rohe, Chatterjee, and Yu 2011; Amini et al. 2013; Chen and Xu 2016).

Although most real world network datasets come with covariate information associated with nodes, existing approaches are primarily focused on using the network for inferring the hidden community memberships. Take for example the Mexican political elites network (described in detail in Section 4). This dataset comprises of 35 politicians (military or civilian) and their connections. The associated covariate for each politician is the year when one came into power. After the military coup in 1913, the political arena was dominated by the military. In 1946, the first civilian president since the coup was elected, signaling the shift of power from revolutionary armed forces to governmental financial elites. Hence those who came into power later are more likely to be civilians. Politicians who have similar number of connections to the military and civilian groups are hard to classify from the network alone. Here the temporal covariate is crucial in resolving which group they belong to. On the other hand, politicians who came into power around 1940s are ambiguous to classify using covariates, since both groups had equal presence in politics in the 1940s. Hence, the number of connections to the two groups in the network helps in classifying these nodes. Our method has higher accuracy in classifying these politicians than existing methods (Zhang, Levina, and Zhu 2016; Binkiewicz, Vogelstein, and Rohe 2017).

In statistics literature, there has been some interesting work on combining covariates and dense networks (average degree growing faster than logarithm of the number of nodes). In Binkiewicz, Vogelstein, and Rohe (2017), the authors present assortative covariate-assisted spectral clustering (ACASC) where one does spectral clustering on the gram matrix of the covariates plus the regularized graph Laplacian weighted by a tuning parameter. A joint criterion for community detection (JCDC) is proposed by Zhang, Levina, and Zhu (2016), which could be seen as a covariate reweighted Newman–Girvan modularity. This approach enables learning different influence on each covariate. In concurrent work, Weng and Feng (2016) provide a variational approach for community detection. Other notable works include Newman and Clauset (2016) and Zhang et al. (2017).

All of the above works are analyzed in the dense regime with strong separability conditions on the linkage probabilities. In

**CONTACT** Bowei Yan *yanbowei@gmail.com* Department of Statistics and Data Sciences GDC 7.504, 2317 Speedway D9800 Austin, Texas 78712-1823.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

contrast, we prove our result for sparse graphs where the average degree is constant and the covariates are finite dimensional sub-Gaussian mixtures with moderate separability conditions. In our setting, neither source can yield consistent clustering in the limit.

Leveraging information from multiple sources have been long studied in machine learning and data mining under the general envelop of multi-view clustering methods. Kumar, Rai, and Daume (2011) used a regularization framework so that the clustering adheres to the dissimilarity of clustering from each view. Liu et al. (2013) optimized the nonnegative matrix factorization loss function on each view, plus a regularization forcing the factors from each view to be close to each other. The only provable method was by Chaudhuri et al. (2009), where the authors obtain guarantees where the two views are mixtures of log-concave distributions. This algorithm does not apply to networks.

In this article, we propose a penalized optimization framework for community detection when node covariates are present. We take the sparse degree regime of SBMs, where one can only correctly cluster a nonvanishing fraction of nodes. Similarly, for covariates, we assume that the covariates are generated from a finite dimensional sub-Gaussian mixture with moderate separability conditions. We prove that our method leads to better bounds on clustering accuracy under weaker conditions on the separation between clusters from each source. As byproducts of our theoretical analysis, we obtain new asymptotic results for sparse networks under weak separability conditions and kernel clustering of finite dimensional mixture of sub-Gaussians. Using a variety of real world and simulated data examples, we show that our method often outperforms existing methods. We also illustrate in the simulations that, when the two sources only have partial and in some sense orthogonal information about the clusterings, combining them leads to better clustering than using the individual sources.

In Section 2, we introduce relevant notation and present our optimization framework. In Section 3, we present our main results, followed by experimental results on simulations and real world networks in Section 4. All proofs are deferred to the supplementary materials.

## 2. Problem Setup

In this section, we introduce our model and set up the convex relaxation framework. For clarity, we list population quantities in Table 1, random variables in Table 2 and other definitions and notations that will be used later in Table 3.

Assume $(C_1, \ldots, C_r)$ represent a $r$-partition for $n$ nodes $\{1, \ldots, n\}$. Let $m_i = |C_i|$ be the size of cluster $i$, and let $m_{\min}$ and $m_{\max}$ be the minimum and maximum cluster sizes, respectively. We use $\pi_i := m_i/n$, $\pi_{\min} = m_{\min}/n$, and $\alpha = m_{\max}/m_{\min}$. We denote by $A$ the $n \times n$ binary adjacency matrix and by $Y$ the $n \times d$ matrix of $d$ dimensional covariates. The generation of $A$ and $Y$ share the true and unknown membership matrix $Z = \{0, 1\}^{n\times r}$. We define the graph model as

$$\text{(Graph model)} \qquad P(A_{ij} = 1|Z) = Z_i^T B Z_j, \quad \text{for } i \neq j, \quad (1)$$

where $B$ is a $r \times r$ matrix of within and across cluster connection probabilities. Furthermore $A_{ii} = 0, \forall i \in [n]$. We consider the sparse regime where $n \max_{k\ell} B_{k\ell}$ is a constant and hence average expected degree is also a constant w.r.t $n$. Amini and Levina (2018) define two different classes of block models in terms of separability properties of $B$. We state this below.

*Definition 1.* A SBM is called *strongly assortative* if $\min_k B_{kk} > \max_{k\neq\ell} B_{k\ell}$, and is called *weakly assortative* if $\forall k \neq \ell$, $B_{kk} > B_{k\ell}$.

This distinction is important because the weakly assortative class of blockmodels is a superset of strongly assortative models,

**Table 1.** Population quantities used in the article.

| Notation | Mathematical definition | Explanation |
|---|---|---|
| $n$ | | Number of nodes |
| $d$ | | Dimensionality of covariates |
| $I_d$ | | Identity matrix of size $d \times d$ |
| $\text{diag}(v_1, \ldots, v_k) \in \mathbb{R}^{k\times k}$ | | Diagonal matrix with diagonal $(v_1, \ldots, v_k)$ |
| $r$ | $\Theta(1)$ | Number of clusters |
| $B \in [0, 1]^{r\times r}$ | $B_{ij} = \Theta(1/n), i, j \in [r]$ | Symmetric probability matrix in SBM |
| $Z \in \{0, 1\}^{n\times r}$ | | Latent class memberships |
| $m_i$ | $\sum_j Z(j, i)$ | Number of points in $i$th cluster |
| $\pi_i$ | $\frac{m_i}{n}$ | Proportion of points in $i$th cluster |
| $m_{\max}$ | $\max_k m_k, \Theta(n)$ | Largest cluster size |
| $m_{\min}$ | $\min_k m_k, \Theta(n)$ | Smallest cluster size |
| $\alpha$ | $m_{\max}/m_{\min}, \Theta(1)$ | Ratio between largest and smallest clusters |
| $C_k$ | $\{j : Z(j, k) = 1\}$ | Point set for $k$th cluster |
| $X_0 \in \mathbb{R}^{n\times n}$ | $Z\text{diag}(1/m_1, \ldots, 1/m_r)Z^T$ | Ground truth clustering matrix |
| $a_k$ | $nB_{kk}, \Theta(1)$ | Rescaled connection probability |
| $b_k$ | $n \max_{\ell\neq k} B_{k\ell}, \Theta(1)$ | Rescaled connection probability |
| $v_A \in \mathbb{R}_{\geq 0}$ | $\frac{2}{n-1} \sum_{i<j} Var(A_{ij}), \Theta(1)$ | Average variance of graph edges |
| $\mu_k, \Sigma_k$ | | Mean, covariance matrix for $Y_i$ if $i \in C_k$ |
| $\psi_k$ | | Sub-Gaussian parameter for $Y_i$ if $i \in C_k$ |
| $\psi_{\max}$ | $\max_{k\in[r]} \psi_k$ | Largest sub-Gaussian parameter across all clusters |
| $d_{k\ell}$ | $\|\mu_k - \mu_\ell\|$ | Distance between cluster centers for the covariates |
| $Q_K$ | Equation (7) | Reference matrix for the kernel |
| $\nu_k$ | Equation (6) | Separation in $Q_K$ |
| $\gamma$ | $\min_k(a_k - b_k + \lambda_0\nu_k), \Theta(1)$ | Separation of $ZBZ^T + \lambda_n K$. |
| $\eta$ | | Scale parameter in kernel function |

**Table 2.** Random variables used in the article.

| Notation | Mathematical definition | Explanation |
|---|---|---|
| $A \in \{0,1\}^{n\times n}$ | $A_{ij}\|i \in C_k, j \in C_\ell \sim \text{Ber}(B_{k\ell})$ | Adjacency matrix (symmetric) |
| $Y_i \in \mathbb{R}^d$ | | Covariate observation for $i$th point |
| $K \in [0,1]^{n\times n}$ | $K(i,j) = f(\|\|Y_i - Y_j\|\|_2^2)$ | Kernel matrix, symmetric and positive definite |

**Table 3.** Useful notations and definitions.

| Notation | Mathematical definition | Explanation |
|---|---|---|
| $\mathbf{1}_n$ | | All one vector of length $n$ |
| $E_n$ | $\mathbf{1}_n\mathbf{1}_n^T$ | All ones matrix of size $n \times n$ |
| $I_d$ | | Identity matrix of size $d \times d$ |
| $K_G$ | $\leq 1.783$ | Grothendieck's constant |
| $f(x) : \mathbb{R}_{\geq 0} \to [0,1]$ | $\exp(-\eta x)$ | Kernel function |
| $\mathcal{F}$ | $\{X \succeq 0,\ 0 \leq X \leq \frac{1}{m_{\min}},$ $X\mathbf{1}_n = \mathbf{1}_n,\ \text{trace}(X) = r\}$ | Feasible set of the SDP |
| $X_M$ | $\arg\max_X \langle M, X\rangle$ *s.t.* $X \in \mathcal{F}$ | Solution matrix of the SDP |
| $\theta_i(M)$ | | $i$th eigenvalue of $M$ |
| $\lambda_n, \lambda_0$ | $\lambda_n = \lambda_0/n, \lambda_0 = \Theta(1)$ | Tuning parameter between graph and covariates |

and most of the analysis are done in the stronger setting. To our knowledge, there has not been any work on weakly assortative blockmodels in the sparse setting. For the covariates, we define,

$$\text{(Covariate model)} \qquad Y_i = \sum_{a=1}^{r} Z_{ia}\mu_a + W_i, \qquad (2)$$

where $W_i$'s are mean zero $d$-dimensional sub-Gaussian vectors with covariance matrix $\Sigma_k$ and sub-Gaussian parameter $\psi_k$ (for $i \in C_k$). We use standard definitions of sub-Gaussian random variables and vectors as in Hsu, Kakade, and Zhang (2012) and Wainwright (2015), which are presented in details in Section A of the supplementary materials. Using properties of sub-Gaussian random variables, it can be shown that the operator norm of $\Sigma_k$ is no larger than $\psi_k$. We define the distance between clusters $C_k$ and $C_\ell$ as $d_{k\ell} = \|\mu_k - \mu_\ell\|$ and the separation as $d_{\min} = \min_{k\neq\ell} d_{k\ell}$.

### 2.1. Notation

We use $\|M\|_F$ and $\|M\|$ to denote the Frobenius and operator norms of a matrix $M \in \mathbb{R}^{n\times n}$, respectively. The $\ell_\infty$ norm is defined as: $\|M\|_\infty = \max_{i,j}|M_{ij}|$. For two matrices $M, Q \in \mathbb{C}^{m\times n}$, their inner product is $\langle M, Q\rangle = \text{trace}(M^T Q)$. From now on we use $I_n$ to denote the identity matrix of size $n$, $\mathbf{1}_n$ to represent the all one $n$-vector and $E_n, E_{n,k}$ to represent the all one matrix with size $n \times n$ and $n \times k$, respectively. We use standard order notations $O, o, \Omega, \omega$, etc. For example, $t(n) = \Theta(1/n)$ is to denote that $n \cdot t(n)$ is a constant w.r.t $n$. $\tilde{O}$ notation is used for implicit multiplicative terms logarithmic in $n$.

### 2.2. Optimization Framework

We now present our optimization framework. There are many available semidefinite programming (SDP) relaxations for clustering block models (Cai and Li 2015; Chen and Xu 2016; Amini and Levina 2018). The common element in all of these is maximizing the inner product between $A$ and $X$, for a positive semidefinite (p.s.d.) matrix $X$. Here $X$ is a stand-in for the clustering matrix $ZZ^T$. Unequal-sized clusters is usually tackled with an extra regularization term added to the objective function (see Cai and Li 2015; Hajek, Wu, and Xu 2016; Perry and Wein 2017, among others). While the above consistency results are for dense graphs, Guédon and Vershynin (2015) and Montanari and Sen (2016) show that in the sparse regime one can use SDP to obtain an error rate which is a constant w.r.t $n$ and depends on the gap between the within and across cluster probabilities.

SDPs are not limited to network clustering. Several convex relaxations for $k$-means type loss are proposed in the literature (see Peng and Wei 2007; Yan and Sarkar 2016; Mixon, Villar, and Ward 2017 for more references). In these settings one maximizes $\langle W, X\rangle$, for some p.s.d. matrix $X$, where $W$ is a matrix of similarities between pairwise data points. For classical $k$-means, $W_{ij}$ can be $Y_i^T Y_j$, whereas for $k$-means in the kernel space one uses a suitably defined kernel similarity function between the $i$th and $j$th data points (covariates). We analyze the widely used Gaussian kernel to allow for nonlinear boundaries between clusters. Let $K$ be the $n \times n$ kernel matrix whose $(i,j)$th entry is $K(i,j) = f(\|Y_i - Y_j\|_2^2)$, where $f(x) = \exp(-\eta x)$ for $x \geq 0$. This kernel function is upper bounded by 1 and is Lipschitz continuous w.r.t. the distance between two observations. Furthermore, in contrast to network based SDPs, the above uses $X$ as a stand-in for the normalized variant of the clustering matrix $ZZ^T$, that is, the desired solution is $(X_0)_{ij} = \frac{1(k=\ell)}{m_k}$, if $i \in C_k, j \in C_\ell$. It can be seen that $\|X_0\|_F^2 = r$.

In our optimization framework, we propose to add a $k$-means type regularization term to the network objective to maximize $\langle A + \lambda_n K, X\rangle$ *s.t.* $X \in \mathcal{F}$. This enforces that the estimated clusters are also consistent with the latent memberships in the covariate space. Here $\lambda_n$ is a tuning parameter (possibly depending on $n$) and the constraint set is $\mathcal{F} = \{X \succeq 0,\ 0 \leq X \leq \frac{1}{m_{\min}},\ X\mathbf{1}_n = \mathbf{1}_n,\ \text{trace}(X) = r\}$, which is similar to Peng and Wei (2007). The $m_{\min}$ in the constraint can be replaced by any lower bound on the smallest cluster size, and is mainly for convenience of the analysis. In the implementation, it suffices to enforce the elementwise positivity constraints, and other linear constraints. For ease of exposition, let

$$X_M = \arg\max_X \langle M, X\rangle \ \text{ s.t. } X \in \mathcal{F}. \qquad (3)$$

When $K(i,j) = Y_i^T Y_j$, then the nonconvex variant of the objective function naturally assumes a form similar to the work of ACASC (modulo normalization of $A$).

In the next section, we provide theoretical guarantees for the solution to the convex semidefinite relaxation problem in Equation (3).

## 3. Main Results

Typically in existing SDP literature for sparse networks or sub-Gaussian mixtures (Guédon and Vershynin 2015; Mixon, Villar, and Ward 2017), one obtains a relative error bound of the deviation of $X_M$ (the solution of the SDP) from the ideal clustering matrix $X_0$. This relative error typically has the form of a ratio of two quantities; the first measures deviation of the observed matrix $M$ from some suitably defined reference matrix, which leads to perfect clustering. The second quantity measures the separation between the different clusters. Similarly, our theoretical result shows that the relative error is a ratio of the combined deviation resulting from the network and the covariates, and a quantity, which is a nonlinear combination of the separations stemming from the two sources. We first present an informal version of the main result.

*Main theorem (Informal).* Let $X_{A+\lambda_n K}$ be the solution of (3). Let $s_G^k$ and $s_C^k$ be constants denoting the separations of the $k$th cluster from the other clusters defined in terms of the model parameters of the network and the covariates, respectively. If the tuning parameter $\lambda_n = \lambda_0/n$ for some constant $\lambda_0$, then for some constant $C$,

$$\|X_{A+\lambda_n K} - X_0\|_F^2 \le C \frac{c_G + \lambda_0 c_C}{\min_k \left(s_G^k + \lambda_0 s_C^k\right)},$$

where $c_G$ and $c_C$ represent the deviation of the graph adjacency matrix $A$ and the covariate kernel matrix $K$ from their corresponding reference matrices.

We will now make $c_G, c_C, s_G$, and $s_C$ concrete. In SBM, that is, when $M = A$, a natural choice of the reference matrix is $\mathbb{E}[A|Z]$ which is block-wise constant. In this case, the separation is given by $\min_k(B_{kk} - \max_\ell B_{k\ell})$, and leads to a result on weakly assortative sparse block models which we present in more details in Section 3.1. A similar route for the kernel matrix is difficult due to the pairwise dependencies of the entries of $K$. Hence, we introduce a novel choice of the reference matrix, which is not block-wise constant, but still has a small deviation from $K$. We show this in Section D of the supplementary materials.

To better understand the role of the separation parameter, we first present a key technical lemma bounding $\|X_M - X_0\|_F$. The main goal of this lemma is to establish an upper bound on the Frobenius norm difference between the ideal clustering matrix and the solution to the SDP with input matrix $M$.

*Lemma 1.* Let $X_M$ be defined by Equation (3) for some input matrix $M \in \mathbb{R}^{n\times n}$. Also let $Q$ be a reference matrix where $Q_{ij} = \beta_k^{(\text{in})}, \forall i,j \in C_k$, and $\beta_k^{(\text{out})} \ge Q_{ij} \ge 0, \forall i \in C_k, j \in C_\ell, k \ne \ell$. If $\min_k(\beta_k^{(\text{in})} - \beta_k^{(\text{out})}) > 0$, then

$$\|X_M - X_0\|_F^2 \le 2 \frac{\langle M - Q, X_M - X_0\rangle}{m_{\min} \min_k(\beta_k^{(\text{in})} - \beta_k^{(\text{out})})} \tag{4}$$

*Remark 1.* The above lemma formalizes the notion of the reference matrix $Q$ we have mentioned before. The deviation between $X_M$ and $X_0$ is small if $M - Q$ is small, and the separation between blocks in $Q$ is large. While the proof technique is inspired by Guédon and Vershynin (2015), the fact that we use different constraints and our reference matrix $Q$ does not have to be block-wise constant complicates the analysis. Moreover, our reference matrix can be weakly assortative instead of strongly assortative.

The results on networks, covariates and the combination of the two essentially reduce to identifying a good reference matrix ($Q$) for the input matrices $A$, $K$, and $A + \lambda K$, which

1. satisfies the properties of $Q$ in the above lemma;
2. has a large separation $\min_k(\beta_k^{(\text{in})} - \beta_k^{(\text{out})})$ increasing the denominator of Equation (4);
3. has a small deviation from $M$, thereby decreasing the numerator of Equation (4).

Now the main work is to choose the reference matrix $Q$ for $A + \lambda K$. As pointed out before, a common choice for reference matrix of $A$ is $\mathbb{E}[A|Z]$. For the covariates, define the set of "good" nodes as follows:

$$\mathcal{S} = \cup_{k=1}^r \mathcal{S}_k, \text{ where } \mathcal{S}_k = \{i \in C_k : \|Y_i - \mu_k\| \le \Delta_k\}. \tag{5}$$

The intuition is that the "good" nodes are easier to cluster, and the sub-matrix of $K$ induced by $\mathcal{S}$ resembles a reference matrix with a relatively large separation. $\Delta_k$ are selected such that the kernel matrix induced by the rows and columns in $\mathcal{S}$ is weakly assortative. Define

$$r_k := f(2\Delta_k), \quad s_k := \max_{\ell \ne k} f(d_{k\ell} - \Delta_k - \Delta_\ell),$$
$$\nu_k = r_k - s_k. \tag{6}$$

A simple use of triangle inequality gives $\min_{i,j\in\mathcal{S}_k} K_{ij} \ge r_k$ and $\max_{i\in\mathcal{S}_k, j\in\mathcal{S}_\ell, \ell\ne k} K_{ij} \le s_k$. When $\max\{3\Delta_k + \Delta_\ell, 3\Delta_\ell + \Delta_k\} \le d_{k\ell}$, the separation for cluster $k$ is $\nu_k > 0$. We define the reference matrix $Q_K$ as

$$(Q_K)_{ij} = \begin{cases} f(2\Delta_k), & \text{if } i,j \in C_k, \\ \min\{f(d_{k\ell} - \Delta_k - \Delta_\ell), K_{ij}\}, & \text{if } i \in C_k, j \in C_\ell, k \ne \ell. \end{cases} \tag{7}$$

The choice of $\Delta_k$ is crucial. A large $\Delta_k$ makes the size of nonseparable nodes $\mathcal{S}^c$ small, but drives down the separation $\nu_k$.

We are now ready to present our main result. As we will show in the proof, the new separation is $\gamma = \min_k \frac{(a_k - b_k) + \lambda_0 \nu_k}{n}$. Typically, in the general case with unequal sub-Gaussian parameters, one should benefit from using different $\Delta_k$'s for different clusters. For example for a cluster with a large $a_k - b_k$, we can afford to have a small $\nu_k$. To think in terms of $\Delta_k$, for this cluster

one can have a large $\Delta_k$, which will make $|\mathcal{S}_k|$ larger than before, but will not affect the separation $(a_k - b_k) + \lambda_0 \nu_k$ of cluster $k$ very detrimentally. We now present our first main theorem. All notations, $B, \psi_k, \pi_k, \pi_{\min}, d_{k\ell}, m_{\min}, g$, and $r$ are defined in Table 1; $f$ is defined in Table 3.

*Theorem 1.* Let $a_k = nB_{kk}, b_k = n\max_{\ell\neq k} B_{k\ell}$, $\nu_A := \frac{2}{n-1}\sum_{i<j}\text{var}(a_{ij}) \geq 9$. Take $\lambda_n = \lambda_0/n$, $m_k = n\pi_k$, $m_{\min} = n\pi_{\min}$, and $\pi_0 := \sum_k (m_k \exp(-\Delta_k^2/5\psi_k^2) + \sqrt{m_k \log m_k/2})/n$. Let $X_{A+\lambda_n K}$ be defined as in Equation (3). If $r, \pi_{\min} = \Theta(1)$ and $\min_k(a_k - b_k + \lambda_0\nu_k) > 0$, then, with probability tending to one,

$$\frac{\|X_{A+\lambda K} - X_0\|_F^2}{\|X_0\|_F^2} \leq 2K_G \frac{6\sqrt{\nu_A} + \lambda_0\left(2\pi_0 + \sum_k \pi_k^2(1 - f(2\Delta_k))\right)}{r\pi_{\min}^2 \min_k(a_k - b_k + \lambda_0\nu_k)},$$

where $\nu_k = f(2\Delta_k) - \max_{\ell\neq k} f(d_{k\ell} - \Delta_k - \Delta_\ell)$ (see Equation (6)) for some $\Delta_k, \Delta_\ell \geq 0$ and $\Delta_k + \Delta_\ell \leq d_{k\ell}$.

Here $K_G$ is the Grothendieck's constant. The best value of $K_G$ is still unknown, and the best known bound is $K_G \leq 1.783$ (Braverman et al. 2013). The proof can be found in Section E in the supplementary materials. Since we work in the sparse setting, we take $\lambda_n = \lambda_0/n$ for some constant $\lambda_0$. Hence, the right hand side of the bound of Theorem 1 is a constant, which decreases as the separation grows. Note that the assumption that $r$ and $\pi_{\min}$ are both $\Theta(1)$ is a common assumption in the sparse setting (Guédon and Vershynin 2015; Montanari and Sen 2016).

In general the upper bound depends on several parameters such as $\lambda_0$ and the scale parameter $\eta$ in the Gaussian kernel. We provide procedures for tuning $\lambda_0$ and $\eta$ in Section 4. The $\Delta_k$'s show up in the numerator as well as the denominator. Finding the optimal $\Delta_k$ is cumbersome in the general case with unequal $\psi_k$'s. In Section 3.2, we derive an upper bound by setting all $\Delta_k$'s to be equal for concreteness.

*Remark 2 (Connection to clustering accuracy).* Intuitively, a small Frobenius norm of $X - X_0$ for any matrix $X$ should result in good clustering if we conduct spectral clustering on $X$. Indeed, as shown in Section G of the supplementary materials, the average mis-clustering error rate per cluster can be upper bounded using $\|X - X_0\|_F$.

*Remark 3 (Practical implications).* All the theoretical results in this article are about the global optima of the convex optimization problem in Equation (3). There are many polynomial time commercial softwares to solve a SDP, for example, Mosek, SeDuMi, CVX, etc. Unfortunately even polynomial time complexity can be prohibitive for moderately large networks. This is why recently first-order methods like alternating direction method of multipliers (Amini and Levina 2018) have gained much popularity. However, first-order methods may not be convergent (Chen et al. 2016).

In this article, we use a widely used (Li, Qi, and Yu 2013; Rauhut and Stojanac 2015; Villar et al. 2016; Kushagra et al. 2017; Mixon, Villar, and Ward 2017) software SDPNAL+ Yang, Sun, and Toh (2015), which uses second-order information. The linear convergence properties of SDPNAL+ to the global optima are presented in Yang, Sun, and Toh (2015, Theorems 3.1 and 3.2). One can also use divide and conquer type approaches to compute SDP solutions using CVX, SeDuMi, or Mosek on smaller local subgraphs and merge them to obtain the global clustering as shown in Mukherjee, Sarkar, and Bickel (2017). A more detailed discussion on different algorithms for solving SDP's is deferred to Section 4.2.

Now we present two natural byproducts of our analysis, namely the result on graphs, that is, bounds on $\|X_0 - X_A\|_F$, and the result on covariate clustering, that is, bounds on $\|X_0 - X_K\|_F$.

### 3.1. Result on Sparse Graphs

While most community detection schemes give perfect clustering in the limit for dense networks (Amini et al. 2013; Cai and Li 2015; Chen and Xu 2016; Amini and Levina 2018; Yan, Sarkar, and Cheng 2018), in the sparse case no algorithm is consistent; however, semidefinite relaxations (among others) can achieve an error rate governed by the within and across cluster probabilities (Guédon and Vershynin 2015; Montanari and Sen 2016). Most analysis for the sparse network are done under strongly assortative settings.

*Proposition 1 (Analysis for graph).* Let $a_k, b_k$ defined as in Theorem 1 are positive constants and $\nu_A \geq 9$. Then with probability tending to 1, $\frac{\|X_A - X_0\|_F}{\|X_0\|_F} \leq \epsilon$, if $\min_k(a_k - b_k) \geq \frac{23\alpha^2 r\sqrt{\nu_A}}{\epsilon^2}$ where $\alpha := m_{\max}/m_{\min}$.

The proof of Proposition 1 is in Section C of the supplementary materials. Note that in the above result, $a_k - b_k$ is constant by definition, hence the error rate $\epsilon$ never goes to 0. In addition, both number of clusters $r$ and the ratio between largest and smallest cluster sizes $\alpha$ needs to be constant order w.r.t $n$ to guarantee the error rate does not increase when the size of the network grows.

*Remark 4.* In contrast to having $\min_k a_k - \max_k b_k$ (strong assortativity) in the denominator like Guédon and Vershynin (2015), we have $\min_k(a_k - b_k)$ (weak assortativity), which allows for a much broader parameter regime. It is important to note that the condition in Proposition 1 requires the graph to be fairly large. In fact, using a series of intricate and different techniques, Montanari and Sen (2016) show that the constant 23 can be improved to nearly match the information-theoretical lower bound in the two parameter setting (within block probability $a/n$ and across block probability $b/n$) and equal-sized clusters. In our general weakly assortative network and mixture of sub-Gaussian covariates setting, optimizing the constant would be much harder. We leave that for future work.

### 3.2. Result on Covariates

Analogous to the sparse graph setting, we present a result for covariates: while SDP with covariates is not consistent with finite signal-to-noise ratio, it achieves a small error rate if the cluster centers are far apart. Before delving into our analysis, we provide a brief overview of existing work.

For covariate clustering, it is common to make distributional assumptions, such as mixture models with well-separated centers. The most well-studied model is Gaussian mixture models, which can be inferred by expectation-maximization algorithm and its variants (Dasgupta and Schulman 2007). For EM algorithm there have been some local convergence results recently (Balakrishnan, Wainwright, and Yu 2017; Yan, Yin, and Sarkar 2017). The condition required for provable recovery on the separation is usually the minimum distance between clusters being greater than some multiple of the square root of dimension (or effective dimension).

Another popular technique is based on SDP relaxations. For example, Peng and Wei (2007) and Mixon, Villar, and Ward (2017) propose a SDP relaxation for $k$-means type clustering. To make the analysis concrete, we present Proposition 2 with $\Delta_k = \Delta$. The proof of Proposition 2 is deferred to Section D of supplementary materials.

*Proposition 2 (Analysis for covariates).* Let $K$ be the kernel matrix generated from kernel function $f$. Denote $v_k$ as in Equation (6). If $\frac{d_{\min}}{\psi_{\max}} > \max\left\{\sqrt{d}, \frac{180}{\sqrt{d}}\right\}$, then with properly chosen $\eta$, with probability at least $1 - \sum_k \frac{1}{m_k}$,

$$\frac{\|X_K - X_0\|_F^2}{\|X_0\|_F^2} \le C\alpha^2 d \frac{\psi_{\max}^2}{d_{\min}^2} \max\left\{\log\left(\frac{d_{\min}}{\psi_{\max}\sqrt{d}}\right), r\right\}.$$

*Remark 5 (Comparison with prior work).* In recent work, Mixon, Villar, and Ward (2017) show the effectiveness of SDP relaxation with $k$-means clustering for sub-Gaussian mixtures, provided the minimum distance between centers is greater than the standard deviation of the sub-Gaussian times the number of clusters $r$. We provide a dimensionality reduction scheme, which implies a weaker separation condition, in particular, $d_{\min} = \Omega(\sqrt{\min(r, d)})$. Our proof technique is new and involves carefully constructing a reference matrix for Lemma 1.

### 3.3. Analysis of Covariate Clustering When $d \gg r$

In high-dimensional statistical problems, the signal is often assumed to lie in a low-dimensional subspace or manifold. This is why much of Gaussian mixture modeling literature first computes some projection of the data onto a low-dimensional subspace (Vempala and Wang 2004). To reduce the dimensionality of the raw data, one could do a feature selection for the covariates (e.g., Jin, Ke, and Wang 2017; Verzelen and Arias-Castro 2017). In contrast, here we propose a much simpler dimensionality reduction step, which does not distort the pairwise distances between cluster means too much. The intuition is that, for clustering a sub-Gaussian mixture, if $d \gg r$, the effective dimensionality of the data is $r$ since the cluster means lie in an at most $r$-dimensional subspace.

Assume $\sum_k \pi_k \mu_k = 0$ for simplicity. We propose the following simple dimensionality reduction algorithm when $d \gg r$ in a spirit similar to Chaudhuri et al. (2009). We split up the sample into two random subsets $P_1$ and $P_2$ of sizes $n_1$ and $n - n_1$ and compute the top $r-1$ eigenvectors $U_{r-1}$ of the matrix $\hat{S} = \frac{\sum_{i\in P_1}(Y_i - \bar{Y})(Y_i - \bar{Y})^T}{n_1} \in \mathbb{R}^{d\times d}$, where $\bar{Y} = \frac{\sum_{i\in P_1} Y_i}{n_1}$. Now we project the covariates from subset $P_2$ onto this lower dimensional subspace as $Y_i' = U_{r-1}^T Y_i$ to get the low-dimensional projections. We take $n_1 = n/\log n$.

*Lemma 2.* Let $M := \sum_k \pi_k \mu_k \mu_k^T$. If $\sum_k \pi_k \mu_k = 0$, and the smallest eigenvalue of $M$ satisfies $\theta_{r-1}(M) \ge 5\psi_{\max}^2 + C\sqrt{\frac{d\log^2 n}{n}}$ for some constant $C$, the projected $Y_i'$ are also independent data points generated from a sub-Gaussian mixture in $r-1$ dimensions, with sub-Gaussian parameter upper bounded by $\psi_k$ if $i \in C_k$. Furthermore the minimum distance between the means in the $r-1$ dimensional space is at least $d_{\min}/2$ with probability at least $1 - \tilde{O}(r^2 n^{-d})$, where $d_{\min}$ is the separation in the original space.

The proof of this lemma is deferred to Section F of the supplementary materials. Typically $\theta_{r-1}(M)$ signifies the amount of signal. For example, for the simple case of mixture of two Gaussians with $\pi_1 = 1/2$, and $\mu_2 = -\mu_1$, $\theta_{r-1}(M) = \|\mu_1\|^2$, which is essentially $d_{\min}^2/4$. Hence, the condition on $\theta_{r-1}(M)$ essentially translates to a lower bound on the signal-to-noise ratio, that is, $d_{\min}^2 \ge 48\psi_{\max}^2 + C'\sqrt{\frac{d\log^2 n}{n}}$ for some constant $C'$. When $d > r$, if one applies Lemma 2 on the $r-1$ dimensional space, then as long as $d_{\min}^2 = \Omega(\psi_{\max}^2 r)$, the separation in the low-dimensional space also satisfies the separation condition in Proposition 2. Thus, the dimensionality reduction brings down the separation condition in Proposition 2 from $\Omega(\psi_{\max}\sqrt{d})$ to $\Omega(\psi_{\max}\sqrt{\min(r, d)})$.

The sample splitting is merely for theoretical convenience which ensures that the projection matrix and the projected data are independent, resulting in the fact that the final projection is also an independent sample from a sub-Gaussian mixture. To be concrete, the labels of $P_1$ do not matter asymptotically, since they incur a relative error in $\|X_0 - X_K\|_F/\|X_0\|_F$ less than $\sqrt{n^2/(m_{\min}^2 \log n)}/\sqrt{r} \le \sqrt{\alpha^2 r/\log n}$, where $\alpha$ and $r$ are both constants. In our setting, the relative error in Proposition 2 is a small but nonvanishing constant, and so this additional vanishing error term does not affect it. One can make $P_1$ much smaller, however, this will come at the expense of a slightly worse separation condition and tail probability in Lemma 2. However, this sample splitting step is not necessary in practice (Chaudhuri et al. 2009), and so we do not pursue this further.

*Remark 6.* We would like to point out that, in Binkiewicz, Vogelstein, and Rohe (2017), $d$ grows with $n$ to make the misclustering rate converge to zero. It is natural to wonder, why do dimensionality reduction? In other words, should large dimension somehow not help in the concentration of individual elements of the kernel matrix $K$? Indeed, the squared pairwise distance between two nodes, respectively, in communities $a$ and $b$, concentrates around $\|\mu_a - \mu_b\|^2 + 2d\sigma^2$ (for sub-Gaussian mixtures with equal covariance matrices $\sigma^2 I_d$) at rate $\sqrt{d}$. If $\|\mu_a - \mu_b\|$ grows faster than $\sqrt{d}$, indeed one should have consistent clustering, similar to settings in El Karoui (2010) and Yan and Sarkar (2016). However, in our case, we show that $\|\mu_a - \mu_b\|$ needs to be larger than just $\sqrt{\min(d, r)}$, which can be much smaller than $\sqrt{d}$ in high-dimensional settings. This can happen when the signal is embedded in some lower

dimensional manifold. In these cases, dimensionality reduction helps reduce noise, as we demonstrate with experiments in the supplementary materials (Section H).

We now present the tuning procedure, and experimental results.

## 4. Experiments

In this section, we present results on real and simulated data. The cluster labels in our method are obtained by spectral clustering of the solution matrix returned by the SDP. We will use SDP-comb, SDP-net, and SDP-cov to represent the labels estimated from $X_{A+\lambda_n K}$, $X_A$, and $X_K$, respectively. The latter two are used as references of graph-only and covariate-only clustering, respectively. Clustering performance is measured by normalized mutual information (NMI), which is defined as the mutual information of the two distributions divided by square root of the product of their entropies. We have also calculated classification accuracy which shows similar trends, so only NMIs are reported here. We compare the following methods: (1) covariate-assisted spectral clustering (ACASC) (Binkiewicz, Vogelstein, and Rohe 2017); (2) JCDC (Zhang, Levina, and Zhu 2016), (3) SDP-comb, (4) SDP-net, and (5) SDP-cov.

### 4.1. High-Dimensional Covariates

When the data are high-dimensional, we reduce the dimensionality of the high-dimensional covariates by projecting onto the top $r$ singular vectors of the sample covariance matrix of covariates $Y$ (as described in Section 3.3). One may wonder if it would help to do feature selection using contemporary methods (Jin, Ke, and Wang 2017). Note that feature selection will not work well if the means were to be rotated, whereas a PCA-based dimensionality reduction would. We provide a detailed discussion and more experiments comparing feature selection with our simple dimensionality reduction in Section H of the supplementary materials.

### 4.2. Implementation and Computational Cost

Semidefinite programs are used in a variety of practical applications ranging from inference in networks (Cai and Li 2015; Chen and Xu 2016; Amini and Levina 2018), control theory (Fares, Noll, and Apkarian 2002), and general clustering problems (Mixon, Villar, and Ward 2017). The widely used commercial softwares for solving them (CVX, SeDuMi, Mosek, etc.) typically use interior point methods (Vandenberghe and Boyd 1996). In a nutshell, a primal-dual interior point method iteratively uses Newton's method to solve for a sequence of points which converge to the optimal solution. For a standard SDP, these methods require roughly $O(mn^3 + m^2n^2 + m^3)$ operations (Monteiro and Zanjcomo 1999), where $m$ denotes the number of equality constraints and $n$ denotes the size of the problem. However, even polynomial time is not good enough for solving moderately large semidefinite programs. This is why designing large scale SDP solvers with both linear and nonlinear constraints has been an active area of research in the optimization community (Wen, Goldfarb, and Yin 2010; Monteiro, Ortiz, and Svaiter 2014; Zheng and Lafferty 2015).

In contrast to interior point methods, first-order methods are characterized by simple operations per iteration (Burer and Monteiro 2003; Renegar 2014): matrix-vector multiplications, vector dot products, and top eigenvalue-eigenvector pair computations, which are more scalable in practice. Despite being scalable, first-order methods are not necessarily convergent (Chen et al. 2016). To improve over existing first-order methods, recently Yang, Sun, and Toh (2015) introduced a second-order Newton-CG algorithm for solving conic programming coupled with a convergent 3-block alternating direction method of multipliers (Sun, Toh, and Yang 2014). The corresponding solver is SDPNAL+, which enables the authors to solve, for the first time, 95 difficult semidefinite relaxations of quadratic assignment problems to an accuracy of $10^{-6}$ efficiently, while existing first-order methods can only successfully solve about a third of these.

SDPNAL+ has been widely used by researchers for solving semidefinite programs (Li, Qi, and Yu 2013; Rauhut and Stojanac 2015; Villar et al. 2016; Kushagra et al. 2017; Mixon, Villar, and Ward 2017) and optimization toolboxes like Sum of Squares Optimization Toolbox for MATLAB (SOSTOOLS, Papachristodoulou et al. 2013). One of the problems tested by the authors is closely related to the relaxation we use (Equation (3)).

Solving the SDP for a network of 2000 nodes with average degree around 8, takes about 30 min on an Intel(R) Core(TM) i7-4770 CPU @ 3.40GHz processor. In comparison, ACASC takes less than 10 sec, and JCDC takes around 20 hr.

### 4.3. Choice of Tuning Parameters

In all experiments in Section 4, we set $m_{\min} = 1$. The impact of different choices of $m_{\min}$ is discussed in Section H.2 of the supplementary materials. We now present the tuning of the scale parameter in the kernel matrix $\eta$ and the tradeoff parameter between graph and covariates $\lambda_n$. In most of our experiments, the number of clusters is assumed known. In this section, we also provide a practical way to choose among candidates of $r$ when it is not given.

#### 4.3.1. Choice of $\eta$

We use the method proposed in (Shi, Belkin, and Yu 2009) to select the scale parameter $\eta$. The intuition is to keep enough (say 10%) of the data points in the "range" of the kernel for most (say 95%) data points. For each data point $Y_i$, we compute $q_i$ as the 10% quantile of $\|Y_i - Y_j\|, j \in [n]$, and the scale parameter as $\eta = \frac{1}{2w^2}$ and $w = \frac{95\%\text{ quantile of } q_i}{\sqrt{95\%\text{ quantile of } \chi^2_d}}$.

#### 4.3.2. Choice of $\lambda_n$

As $\lambda_n$ increases, the resulting $X_{A+\lambda_n K}$ clustering gradually changes from $X_A$ clustering to $X_K$ clustering. Our theoretical results show that, with the right $\lambda_n$, $X_{A+\lambda_n K}$ and $X_0$ should be close, and hence also have similar eigenvalues. Let $\theta_i(M)$ be the $i$th eigenvalue of matrix $M$. Define the eigen gap function for

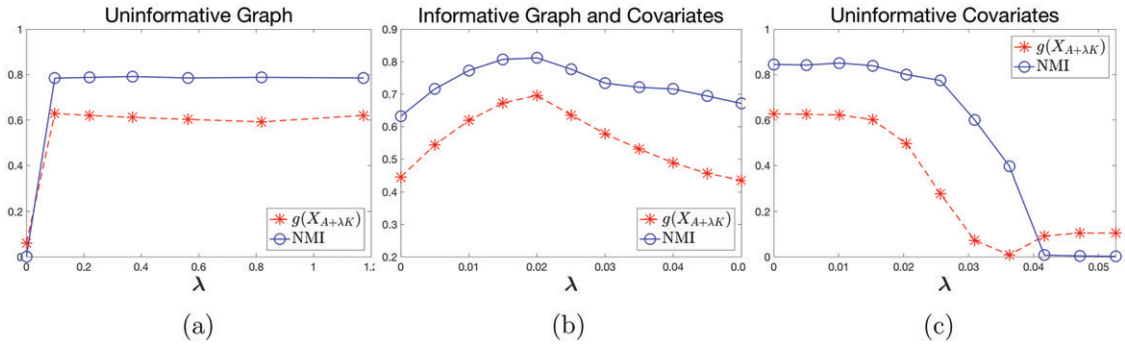



**Figure 1.** Tuning: (a) $B = 0.005E_3, n = 10^3, d = 6, d_{\min} = 15\sigma$; (b) $d = 6, d_{\min} = 1.3, \sigma = (1, 1, 5), B = 10^{-3} * (\text{diag}(4, 24, 24) + 4 * E_3)$; (c) $d = 6, d_{\min} = 0, B = 0.0144I_3 + 0.0016E_3$.

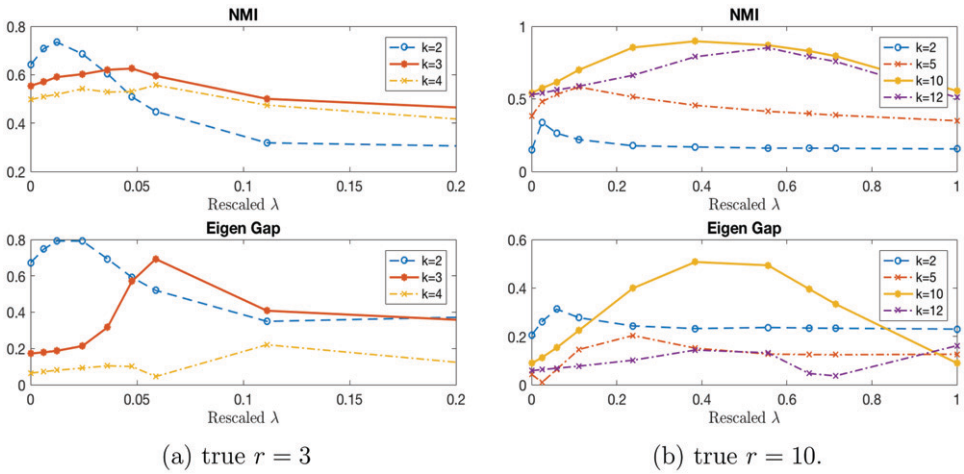



**Figure 2.** NMI and eigen gap $g_k(X_{A+\lambda_n K})$ (Equation (8)) for various $k$ against $\frac{\lambda_n}{1+\lambda_n}$.

clustering matrices

$$g_r(X) := (\theta_r(X) - \theta_{r+1}(X))/\theta_r(X). \tag{8}$$

Using Weyl's inequality and the fact that $\|X_{A+\lambda_n K} - X_0\|_{\text{op}} \le \|X_{A+\lambda_n K} - X_0\|_F$, we have: $\theta_r(X_0) - \|X_{A+\lambda_n K} - X_0\|_F \le \theta_r(X_{A+\lambda_n K}) \le \theta_r(X_0) + \|X_{A+\lambda_n K} - X_0\|_F$. Since $g_r(X_0) = 1$, we pick the $\lambda_n$ maximizing $g_r(X_{A+\lambda_n K})$. Figures 1(a)–(c), respectively, represent the situation where graph is uninformative (Erdős–Rényi), both are informative and covariates are uninformative. We plot $g_r(X_{A+\lambda_n K})$ and NMI of the clustering from $X_{A+\lambda_n K}$ with the true labels against $\lambda_n$. Figure 1 shows that $g_r(X_{A+\lambda_n K})$ and NMI of the predicted clustering have a similar trend, justifying the effectiveness of the tuning procedure.

### 4.3.3. Unknown Number of Clusters

It is hard to know the true number of clusters $r$ in many realistic settings. While only a few methods have been proposed for selecting $r$ under the sparse SBM (Le and Levina 2015), these are designed specifically for graph adjacency matrices and cannot be generalized to kernel similarity matrices resulting from the covariates. We observe that the eigen gap (Equation (8)) acts as an informative indicator for picking $r$. When $r$ is unknown, we run the SDP over a grid of $\{\lambda_n, k\}$, and choose the pair that maximizes the eigen gap $g_k(X_{A+\lambda_n K})$. As shown in Figure 2, we construct two settings to test the performance of using eigen gap to select $r$. In the first setting, the true model has 3 clusterings with proportion 3:4:5, the probability matrix is $B = 0.01 * \begin{bmatrix} 1.6 & 1.2 & 0.16 \\ 1.2 & 1.6 & 0.02 \\ 0.16 & 0.02 & 1.2 \end{bmatrix}$. The covariates are high-dimensional Gaussians centered at $\mu_1 = (0, 2, 0, \ldots, 0)$, $\mu_2 = (-1, -0.8, 0, \ldots, 0)$, $\mu_3 = (1, -0.8, 0, \ldots, 0)$ with covariance matrix $I_{100}$. We sample $n = 800$ data points, on which we run SDP with different choices of $\lambda_n$ and specified number of clusters $k$. For each pair of parameters, we compute the NMI and $g_k(X)$ and plot them on the upper and lower panels of Figure 2(a). On the $x$-axis we have a monotonically increasing function of $\lambda_n$ (which is $\frac{\lambda_n}{1+\lambda_n}$) for better illustration. As we can see, the eigen gap and NMI has a similar trend, hence picking the pair that optimizes the eigen gap $g_k(X_{A+\lambda_n K})$, will have a relatively high NMI as well. Note that, here the mis-specified $k = 2$ has a higher NMI than $k = 3$, which is the true value of $r$. This indicates that, even when the number of clusters is mis-specified, the SDP is still able to find structure that correlates with the underlying model. This phenomenon is also observed in several other works for dense graphs (Perry and Wein 2017; Yan, Sarkar, and Cheng 2018).

In the second scenario, we generate from a SBM with 10 equal-sized clusters, where $B = 0.046I_{10} + 0.004E_{10}$. The covariates are generated from a mixture of Gaussians with means $[3 * I_{10} \mid \mathbf{0}_{10,90}]$, where $[\cdot|\cdot]$ represents the horizontal stacking of matrices and $\mathbf{0}_{m,n}$ represents all 0 matrix of size $m \times n$. We conduct the same type of experiment as above and plot the NMI

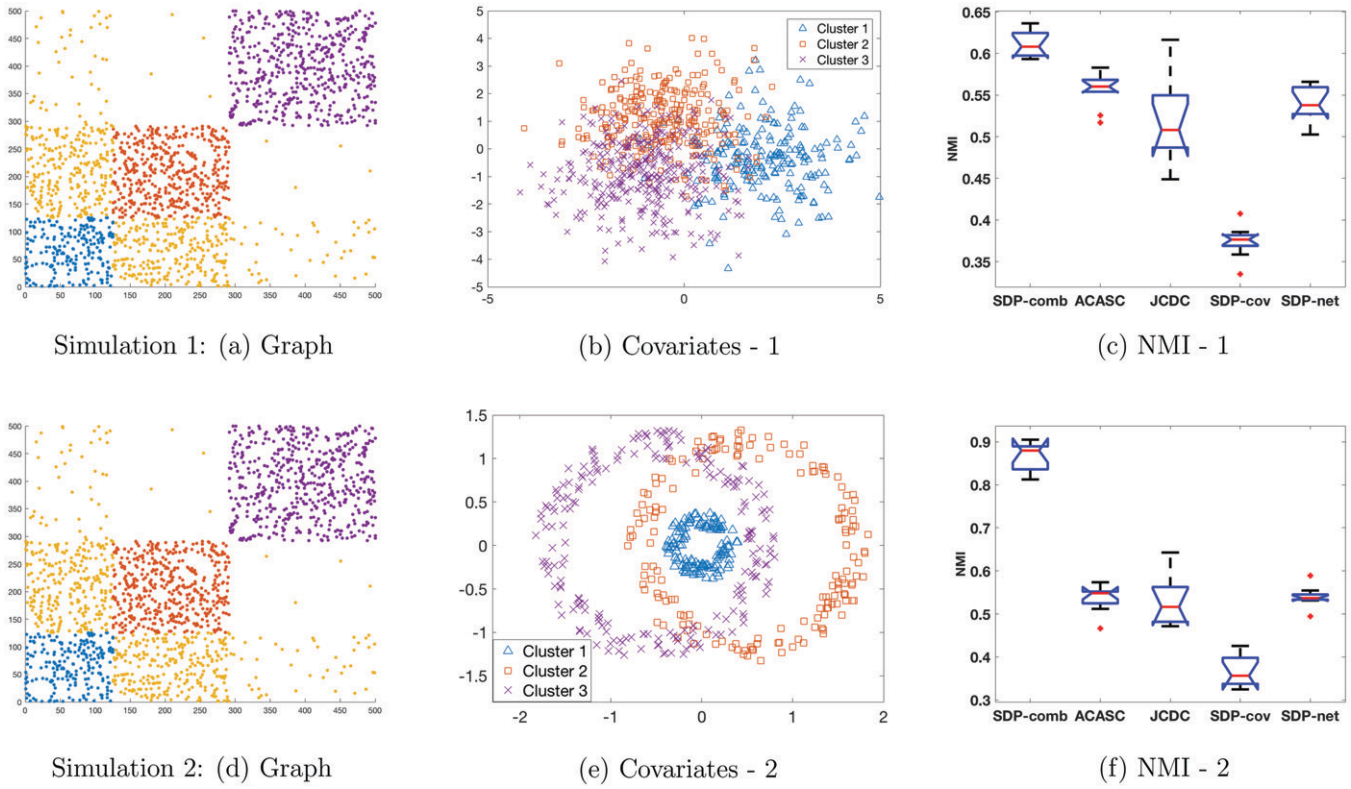



**Figure 3.** The first and second rows have results for isotropic Gaussian covariates and covariates lies on a nonlinear manifold, respectively. We plot the adjacency matrix $A$ in (a, d), where blue, red, and purple represent within cluster edges for 3 ground truth clusters, respectively, and yellow points represent inter-cluster edges. In (b, e) we plot covariates; different shapes and colors imply different clusters. (c, f) The boxplots for NMI.

and eigen gap. In this case, the eigen gap successfully recovered the true number of clusters.

### 4.4. Simulation Studies

In this section, we consider two simulation settings. In the first setting, we generate three clusters with sizes 3:4:5, with $n = 800$. The probability matrix is $B = 0.01 * \begin{bmatrix} 1.6 & 1.2 & 0.16 \\ 1.2 & 1.6 & 0.02 \\ 0.16 & 0.02 & 1.2 \end{bmatrix}$, and the covariates for each cluster are generated with 100 dimensional Gaussians, whose centers are only nonzero on the first two dimensions with $\mu_1 = (0, 2, 0, \ldots, 0)$, $\mu_2 = (-1, -0.8, 0, \ldots, 0)$, $\mu_3 = (1, -0.8, 0, \ldots, 0)$, and covariance $I_{100}$. This is the same setting as in the first simulation for unknown $r$. We first conduct dimensionality reduction by projecting onto the top $r$ singular vectors of the sample covariance matrix of $Y$; then we solve the SDP with the tuning procedure as in Section 4.3, and finally we get the cluster labels by spectral clustering on the solution clustering matrix. In this example, the network cannot separate out clusters one and two well, whereas the covariates can. On the other hand, clusters two and three are not well separated in the covariate space, while they are well-separated by the network parameters. The experiments are repeated on 10 independently generated samples and the boxplot for NMI is shown in Figure 3(c). SDP-comb performs better than ACASC and JCDC in these experiments, possibly because the latter are better suited for denser networks. The variance of JCDC is larger compared to the other methods, possibly because JCDC solves a nonconvex objective function via alternating minimization, and sometimes gets stuck at local optima.

In Figures 3(d)–(f), we examine covariates with nonlinear cluster boundaries. The graph used here is the same as above, and the covariates are two-dimensional, whose scatterplot is shown in Figure 3(e). Since the Gaussian kernel is more suitable for detecting nonlinear decision boundaries, SDP-comb performs better than ACASC, which uses the linear inner product kernel. In both simulations, SDP-comb outperforms others.

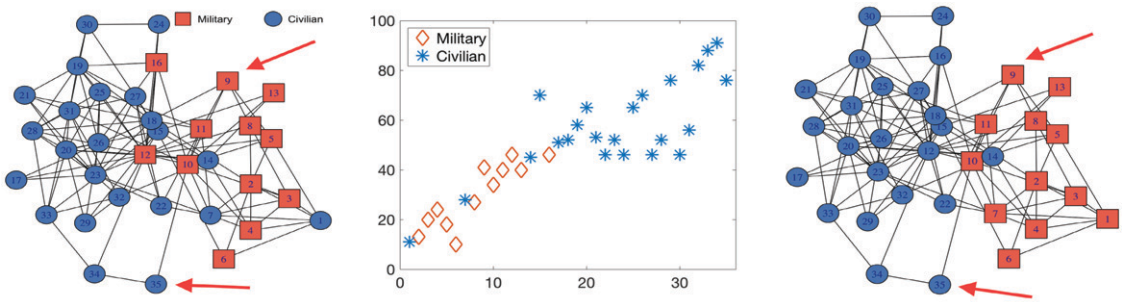
In these experiments, ACASC seems to perform similarly to SDP-net. In general, we noticed that when the network and covariates are "aligned," for example, both have separation for all clusters, ACASC performs better than when they have complementary information. For space constraints, we present more experiments in Section H of the supplementary materials.

### 4.5. Real World Networks

Now we present results on a real world social network and an ecological network. The performance of clustering is evaluated by NMI with the ground truth labels.

### 4.6. Mexican Political Elites

As discussed before, this network (Gil-Mendieta and Schmidt 1996) depicts the political, kinship, or business interactions between 35 Mexican presidents and close collaborators. The two ground truth clusters consist of politicians with military background and civilian background. The year in which a politician




(a) Ground truth (b) Node feature (c) Predicted by SDP-comb

**Figure 4.** Mexican political network.

**Table 4.** NMI with ground truth for various methods.

| Dataset | SDP-net | SDP-cov | SDP-comb | ACASC | JCDC |
|---|---|---|---|---|---|
| Mexican politicians | 0.37 | 0.43 | **0.46** | 0.37 | 0.25 |
| Weddell Sea | 0.36 | 0.22 | **0.51** | 0.32 | 0.42 |

The highest values in the corresponding row are highlighted.

first held a significant governmental position is used as a covariate. It gives a good indication of the labels as in Figure 4(b). This is because the military dominated the political arena after the revolution in the beginning of the twentieth century, and were succeeded by the civilians.

Table 4 shows that our method outperforms other covariate-assisted approaches in NMI. In Figures 4(a) and (c), node 35 has exactly one connection to each of the military and civilian groups, but seized power in the 90s, which strongly indicates a civilian background. Meanwhile, node 9 took power in 1940, a year when civilian and military had almost equal presence in politics, making it hard to detect node 9's political affiliation. Yet node 9 has more edges to the military group than the civilian group. By taking the graph structure into consideration, we can correctly assign the military label to it.

### 4.7. Weddell Sea Trophic Dataset

The next example we consider is an ecological network collected by Jacob et al. (2011), describing the marine ecosystem of Weddell Sea, a large bay off the coast of Antarctica. The dataset lists 489 marine species and their directed predator-prey interactions, as well as the average adult body mass for each of the species. We use a thresholded symmetrization of the directed graph as the adjacency matrix. Let $G$ be the directed graph, the $(i,j)$th entry of $GG^T$ captures the number of other species which $i$ and $j$ both feed on. We create binary matrices $A_\tau = 1(GG^T \geq \tau)$. Choosing different $\tau$'s between 1 to 10 gives similar clustering. We use $\tau = 5$.

All species are labeled into four categories based on their prey types. Autotrophs (e.g., plants) do not feed on anything. Herbivores feed on autotrophs. Carnivores feed on animals that are not autotrophs, and the remaining are omnivores, which feed on both autotrophs and other animals (herbivore, carnivore, or omnivores). Since body masses of species vary largely from nanograms to tons, we work with the normalized logarithm of mass following the convention in Newman and Clauset (2016). Figure 5(b) illustrates the log body mass for species. Without loss of generality, we order the nodes by their prey types.

In Figure 5(c), we plot $A_\tau$. Since the autotrophs do not feed on other species in this dataset, and since herbivores do not have too much overlap in the autotrophs they feed on, the upper left corner of the input network is extremely sparse. On the other side, the body sizes for autotrophs are much smaller than those of other prey types. Therefore, the kernel matrix clearly separates them out.

We see that SDP-net (Figure 5(e)) heavily misclusters the autotrophs since it only replies on the network. SDP-cov (Figure 5(f)) only takes the covariates into account and cannot distinguish herbivores from omnivores, since they possess similar body masses. However, SDP-comb (Figure 5(d)) achieves a significantly better NMI by combining both sources. Table 4 shows the NMI between predicted labels and the ground truth from SDP-comb, JCDC and ACASC. While JCDC and ACASC can only get as good as the best of graph or covariates, our method achieves a higher NMI.

## 5. Discussion

In this article, we propose a regularized convex optimization framework to infer community memberships jointly from sparse networks and covariates. Our theoretical bounds show that, the proposed method works under weak separability of clusters, which is a much broader parameter regime than those in most existing work. Our methodology leads to higher clustering accuracy especially when each source only reflects clustering structure on a subset of the nodes. We demonstrate the performance of our methodology on simulated and real networks, and show that it in general performs better than other state-of-the-art methods. While we limit ourselves to two sources for ease of exposition, our method can be easily generalized to multiple views or sources. Empirically, we demonstrate that our method works for covariates with nonlinear cluster boundaries, the theoretical analysis of which is part of future work.

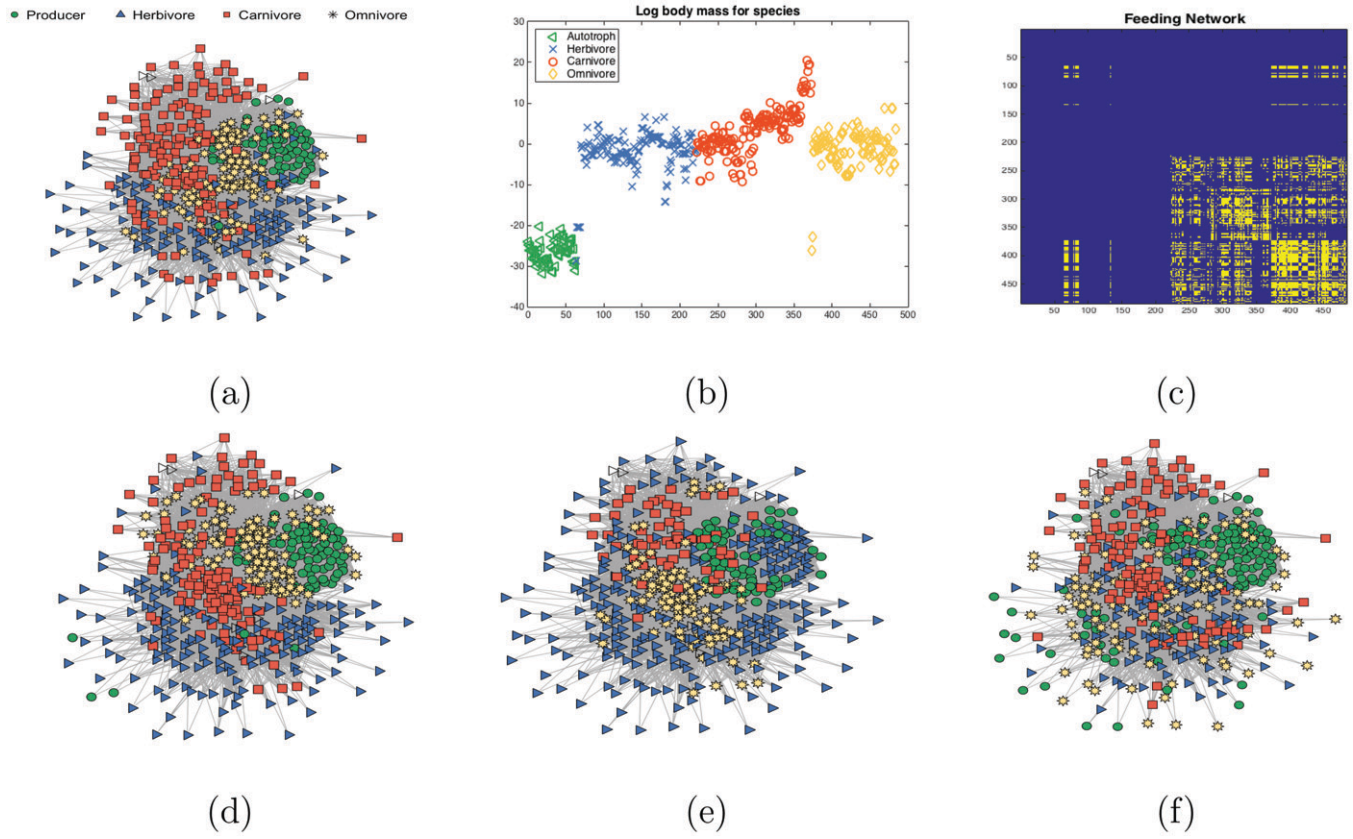



**Figure 5.** Weddell sea network: (a) true labels; (b) log body mass; (c) constructed adjacency matrix $A_\tau$; we show labels from (d) SDP-comb; (e) SDP-net; (f) SDP-cov.

## Supplementary Materials

**Title:** Additional proofs and experimental results (.pdf file).

## Acknowledgements

We thank Arash Amini and Yuan Zhang for generously sharing their code. We are grateful to Soumendu S. Mukherjee, Peter J. Bickel, David Choi, Harrison Zhou, and Qixing Huang for interesting discussions on our article.

## Funding

PS was partially supported by NSF grant DMS 1713082.

## References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), "Mixed Membership Stochastic Blockmodels," *Journal of Machine Learning Research*, 9, 1981–2014. [734]

Amini, A. A., Chen, A., Bickel, P. J., and Levina, E. (2013), "Pseudo-Likelihood Methods for Community Detection in Large Sparse Networks," *The Annals of Statistics*, 41, 2097–2122. [734,738]

Amini, A. A., and Levina, E. (2018), "On Semidefinite Relaxations for the Block Model," *The Annals of Statistics*, 46, 149–179. [735,736,738,740]

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017), "Statistical Guarantees for the EM Algorithm: From Population to Sample-Based Analysis," *The Annals of Statistics*, 45, 77–120. [739]

Binkiewicz, N., Vogelstein, J. T., and Rohe, K. (2017), "Covariate-Assisted Spectral Clustering," *Biometrika*, 104, 361–377. [734,739,740]

Braverman, M., Makarychev, K., Makarychev, Y., and Naor, A. (2013), "The Grothendieck Constant Is Strictly Smaller Than Krivine's Bound," in *Forum of Mathematics, Pi*, Cambridge University Press, 1, e4. doi: 10.1017/fmp.2013.4 [738]

Burer, S., and Monteiro, R. D. (2003), "A Nonlinear Programming Algorithm for Solving Semidefinite Programs via Low-Rank Factorization," *Mathematical Programming*, 95, 329–357. [740]

Cai, T. T., and Li, X. (2015), "Robust and Computationally Feasible Community Detection in the Presence of Arbitrary Outlier Nodes," *The Annals of Statistics*, 43, 1027–1059. [736,738,740]

Chaudhuri, K., Kakade, S. M., Livescu, K., and Sridharan, K. (2009), "Multi-View Clustering via Canonical Correlation Analysis," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, pp. 129–136. [735,739]

Chen, C., He, B., Ye, Y., and Yuan, X. (2016), "The Direct Extension of ADMM for Multi-Block Convex Minimization Problems Is Not Necessarily Convergent," *Mathematical Programming*, 155, 57–79. [738,740]

Chen, Y., and Xu, J. (2016), "Statistical-Computational Tradeoffs in Planted Problems and Submatrix Localization With a Growing Number of Clusters and Submatrices," *Journal of Machine Learning Research*, 17, 1–57. [734,736,738,740]

Dasgupta, S., and Schulman, L. (2007), "A Probabilistic Analysis of EM for Mixtures of Separated, Spherical Gaussians," *Journal of Machine Learning Research*, 8, 203–226. [739]

El Karoui, N. (2010), "On Information Plus Noise Kernel Random Matrices," *The Annals of Statistics*, 38, 3191–3216. [739]

Fares, B., Noll, D., and Apkarian, P. (2002), "Robust Control via Sequential Semidefinite Programming," *SIAM Journal on Control and Optimization*, 40, 1791–1820. [740]

Gil-Mendieta, J., and Schmidt, S. (1996), "The Political Network in Mexico," *Social Networks*, 18, 355–381. [734,742]

Guédon, O., and Vershynin, R. (2015), "Community Detection in Sparse Networks via Grothendieck's Inequality," *Probability Theory and Related Fields*, 165, 1025–1049. [736,737,738]

Hajek, B., Wu, Y., and Xu, J. (2016), "Achieving Exact Cluster Recovery Threshold via Semidefinite Programming," *IEEE Transactions on Information Theory*, 62, 2788–2797. [736]

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983), "Stochastic Blockmodels: First Steps," *Social Networks*, 5, 109–137. [734]

Hsu, D., Kakade, S. M., and Zhang, T. (2012), "A Tail Inequality for Quadratic Forms of Subgaussian Random Vectors," *Electronic Communications in Probability*, 17, 1–6. [736]

Jacob, U., Thierry, A., Brose, U., Arntz, W. E., Berg, S., Brey, T., Fetzer, I., Jonsson, T., Mintenbeck, K., Mollmann, C., and Petchey, O. L. (2011), "The Role of Body Size in Complex Food Webs: A Cold Case," *Advances in Ecological Research*, 45, 181–223. [734,743]

Jin, J., Ke, Z. T., and Wang, W. (2017), "Phase Transitions for High Dimensional Clustering and Related Problems," *The Annals of Statistics*, 45, 2151–2189. [739,740]

Kumar, A., Rai, P., and Daume, H. (2011), "Co-Regularized Multi-View Spectral Clustering," in *Advances in Neural Information Processing Systems* (Vol. 24), pp. 1413–1421. [735]

Kushagra, S., McNabb, N., Yu, Y., and Ben-David, S. (2017), "Provably Noise-Robust, Regularised *k*-Means Clustering," arXiv no. 1711.11247. [738,740]

Le, C. M., and Levina, E. (2015), "Estimating the Number of Communities in Networks by Spectral Methods," arXiv no. 1507.00827. [741]

Li, G., Qi, L., and Yu, G. (2013), "The *Z*-Eigenvalues of a Symmetric Tensor and Its Application to Spectral Hypergraph Theory," *Numerical Linear Algebra With Applications*, 20, 1001–1029. [738,740]

Liu, J., Wang, C., Gao, J., and Han, J. (2013), "Multi-View Clustering via Joint Nonnegative Matrix Factorization," in *Proceedings of the 2013 SIAM International Conference on Data Mining*, SIAM, pp. 252–260. [735]

Mixon, D. G., Villar, S., and Ward, R. (2017), "Clustering Subgaussian Mixtures by Semidefinite Programming," *Information and Inference*, 6, 389–415. [736,737,738,739,740]

Montanari, A., and Sen, S. (2016), "Semidefinite Programs on Sparse Random Graphs and Their Application to Community Detection," in *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, ACM, pp. 814–827. [736,738]

Monteiro, R. D. C., and Zanjcomo, P. (1999), "Implementation of Primal-Dual Methods for Semidefinite Programming Based on Monteiro and Tsuchiya Newton Directions and Their Variants," *Optimization Methods and Software*, 11, 91–140. [740]

Monteiro, R. D., Ortiz, C., and Svaiter, B. F. (2014), "A First-Order Block-Decomposition Method for Solving Two-Easy-Block Structured Semidefinite Programs," *Mathematical Programming Computation*, 6, 103–150. [740]

Mukherjee, S., Sarkar, P., and Bickel, P. (2017), "Two Provably Consistent Divide and Conquer Clustering Algorithms for Large Networks," arXiv no. 1708.05573. [738]

Newman, M. E., and Clauset, A. (2016), "Structure and Inference in Annotated Networks," *Nature Communications*, 7, 11863. [734,743]

Papachristodoulou, A., Anderson, J., Valmorbida, G., Prajna, S., Seiler, P., and Parrilo, P. A. (2013), "SOSTOOLS: Sum of Squares Optimization Toolbox for MATLAB," arXiv no. 1310.4716, available at *http://www.eng.ox.ac.uk/control/sostools*, *http://www.cds.caltech.edu/sostools*, and *http://www.mit.edu/~parrilo/sostools*. [740]

Peng, J., and Wei, Y. (2007), "Approximating *k*-Means-Type Clustering via Semidefinite Programming," *SIAM Journal on Optimization*, 18, 186–205. [736,739]

Perry, A., and Wein, A. S. (2017), "A Semidefinite Program for Unbalanced Multisection in the Stochastic Block Model," in *International Conference on Sampling Theory and Applications*, pp. 64–67. [736,741]

Rauhut, H., and Stojanac, Ž. (2015), "Tensor Theta Norms and Low Rank Recovery," arXiv no. 1505.05175. [738,740]

Renegar, J. (2014), "Efficient First-Order Methods for Linear Programming and Semidefinite Programming," arXiv no. 1409.5832. [740]

Rohe, K., Chatterjee, S., and Yu, B. (2011), "Spectral Clustering and the High-Dimensional Stochastic Blockmodel," *The Annals of Statistics*, 39, 1878–1915. [734]

Shi, T., Belkin, M., and Yu, B. (2009), "Data Spectroscopy: Eigenspaces of Convolution Operators and Clustering," *The Annals of Statistics*, 37, 3960–3984. [740]

Sun, D., Toh, K.-C., and Yang, L. (2014), "A Convergent Proximal Alternating Direction Method of Multipliers for Conic Programming With 4-Block Constraints," Technical Report. [740]

Vandenberghe, L., and Boyd, S. (1996), "Semidefinite Programming," *SIAM Review*, 38, 49–95. [740]

Vempala, S., and Wang, G. (2004), "A Spectral Algorithm for Learning Mixture Models," *Journal of Computer and System Sciences*, 68, 841–860. [739]

Verzelen, N., and Arias-Castro, E. (2017), "Detection and Feature Selection in Sparse Mixture Models," *The Annals of Statistics*, 45, 1920–1950. [739]

Villar, S., Bandeira, A. S., Blumberg, A. J., and Ward, R. (2016), "A Polynomial-Time Relaxation of the Gromov-Hausdorff Distance," arXiv no. 1610.05214. [738,740]

Wainwright, M. (2015), "Basic Tail and Concentration Bounds," available at *https://www.stat.berkeley.edu/.../Chap2_TailBounds_Jan22_2015.pdf*. [736]

Wen, Z., Goldfarb, D., and Yin, W. (2010), "Alternating Direction Augmented Lagrangian Methods for Semidefinite Programming," *Mathematical Programming Computation*, 2, 203–230. [740]

Weng, H., and Feng, Y. (2016), "Community Detection With Nodal Information," arXiv no. 1610.09735. [734]

Yan, B., and Sarkar, P. (2016), "On Robustness of Kernel Clustering," in *Advances in Neural Information Processing Systems*, pp. 3098–3106. [736,739]

Yan, B., Sarkar, P., and Cheng, X. (2018), "Provable Estimation of the Number of Blocks in Block Models," in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, PMLR (Vol. 84), pp. 1185–1194. [738,741]

Yan, B., Yin, M., and Sarkar, P. (2017), "Convergence of Gradient EM on Multi-Component Mixture of Gaussians," in *Advances in Neural Information Processing Systems*, pp. 6959–6969. [739]

Yang, L., Sun, D., and Toh, K.-C. (2015), "SDPNAL++: A Majorized Semismooth Newton-CG Augmented Lagrangian Method for Semidefinite Programming With Nonnegative Constraints," *Mathematical Programming Computation*, 7, 331–366. [738,740]

Zhang, Y., Chen, K., Sampson, A., Hwang, K., and Luna, B. (2017), "Node Features Adjusted Stochastic Block Model," Technical Report, Working Paper. [734]

Zhang, Y., Levina, E., and Zhu, J. (2016), "Community Detection in Networks With Node Features," *Electronic Journal of Statistics*, 10, 3153–3178. [734,740]

Zheng, Q., and Lafferty, J. (2015), "A Convergent Gradient Descent Algorithm for Rank Minimization and Semidefinite Programming From Random Linear Measurements," in *Advances in Neural Information Processing Systems*, pp. 109–117. [740]

# Supplementary Materials for

## Covariate Regularized Community Detection in Sparse Graphs

In this document we collect technical details and accompanying lemmas which are necessary for the main results in the paper *Covariate Regularized Community Detection in Sparse Graphs*. When we make references to equations or theorems etc. in the main document, we follow the numbering scheme of the main document, and the references do not have any alphabets in them.

In Sec. A we introduce the background for sub-gaussian random vectors. In Sec. B, we present the proof of Lemma 1. Proofs for graph only and covariates only SDP (Propositions 1 and 2) are shown in Sec. C and Sec. D respectively, and the proof of the main theorem is in Sec. E. Sec. F shows the results and proofs related to dimensionality reduction, and Sec. G establishes an upper bound of clustering error in terms of the Frobenius deviation of SDP solution to ground truth clustering matrix. Finally, we present additional experimental results in Sec. H.

# A Background on Sub-gaussian Random Vectors

In our analysis, we make use of some useful properties of sub-gaussian random vectors from (Hsu et al., 2012; Wainwright, 2015).

**Definition 1** (Sub-gaussian Random Variable)**.** *A random variable $X$ with mean $\mu$ is defined to be sub-gaussian if there exists a positive number $\psi$, such that the following holds:*

$$E[e^{\lambda(X-\mu)}] \leq e^{\psi^2\lambda^2/2} \qquad \text{For all } \lambda \in \mathbb{R}$$

*$\psi$ is also called the sub-gaussian parameter.*

This can be easily generalized to sub-gaussian random vectors (Hsu et al., 2012).

**Definition 2** (Sub-gaussian Random Vectors)**.** *A random vector $X \in \mathbb{R}^d$ with mean vector $\mu \in \mathbb{R}^d$ is defined to be sub-gaussian if there exists a positive number $\psi$, such that the following holds:*

$$E[e^{v^T(X-\mu)}] \leq e^{\psi^2\|v\|^2/2} \qquad \text{For all } v \in \mathbb{R}^d$$

*$\psi$ is also called the sub-gaussian parameter.*

# B Proof of Lemma 1

We start with the following lemma.

**Lemma B.1.** *For any $X$ that satisfies $X \succeq 0, X \geq 0, X\mathbf{1} = \mathbf{1}$, we have $\|X\|_F^2 \leq$ trace$(X)$.*

*Proof.* We first show that for all such $X$, the eigenvalues of $X$ are in $[0, 1]$. Let $v_i$ be the eigenvector of $X$ corresponding to the $i^{th}$ largest eigenvalue $\theta_i$. Since $X$ is positive semi-definite, $\theta_i \geq 0, \forall i$. Without loss of generality, let $i^* = \arg\max_i |v_1(i)|$, i.e. be the index of the entry with the largest absolute value of $v_1$. Since $Xv_1 = \theta_1 v_1$, and $\sum_j X_{ij} = 1, X_{ij} \geq 0$, we have:

$$|\theta_1 v_1(i^*)| = |\sum_j X_{i^*j} v_1(j)| \leq \sum_j X_{i^*j} |v_1(j)| \leq |v_1(i^*)|.$$

Therefore $|\theta_1| \leq 1$.

$$\|X\|_F^2 = \sum_i \theta_i^2 \leq \sum_i \theta_i = \text{trace}(X)$$

□

Now we are in position to prove Lemma 1.

*Proof of Lemma 1.* Note that both $X_0$ and $X_M$ are in the feasible set $\mathcal{F}$, by optimality, we have $\langle M, X_M \rangle \geq \langle M, X_0 \rangle$. We construct $Q$ as stated in the lemma to obtain: $\langle Q, X_M - X_0 \rangle$, $\langle M - Q, X_M - X_0 \rangle \geq \langle Q, X_0 - X_M \rangle$. Note that $Q$ is constant on diagonal blocks and upper bounded by $q_k$ on off-diagonal blocks, with respect to the clustering of nodes. Using the fact that $|C_k| = m_k$, we have:

$$\begin{aligned}
&\langle M, X_0 - X_M \rangle = \sum_k \sum_{i \in C_k} \left( \beta_k^{(in)} \sum_{j \in C_k} \left( \frac{1}{m_k} - (X_M)_{ij} \right) + \sum_{\ell \neq k} \sum_{j \in C_\ell} Q_{ij} (0 - (X_M)_{ij}) \right) \\
\geq &\sum_k \sum_{i \in C_k} \left( \beta_k^{(in)} \sum_{j \in C_k} \left( \frac{1}{m_k} - (X_M)_{ij} \right) - \beta_k^{(out)} \sum_{\ell \neq k} \sum_{j \in C_\ell} (X_M)_{ij} \right) \\
= &\sum_k \sum_{i \in C_k} \left( \beta_k^{(in)} \left( 1 - \sum_{j \in C_k} (X_M)_{ij} \right) - \beta_k^{(out)} \left( 1 - \sum_{j \in C_k} (X_M)_{ij} \right) \right) \\
= &\sum_k \sum_{i \in C_k} (\beta_k^{(in)} - \beta_k^{(out)}) \left( 1 - \sum_{j \in C_k} (X_M)_{ij} \right) \geq \min_k (\beta_k^{(in)} - \beta_k^{(out)}) \sum_k \sum_{i \in C_k} \left( 1 - \sum_{j \in C_k} (X_M)_{ij} \right)
\end{aligned}$$

The third line and last inequality uses the constraint that $\sum_j \hat{X}_{ij} = 1$, and $1 - \sum_{j\in C_k} \hat{X}_{ij} \geq 1 - \sum_j \hat{X}_{ij} = 0$. On the other hand,

$$\|X_M - X_0\|_F^2 = \|X_M\|_F^2 - \|X_0\|_F^2 + 2\langle X_0 - X_M, X_0\rangle$$

By Lemma B.1, and the fact that $\|X_0\|_F^2 = r$, we have $\|X_M\|_F^2 - \|X_0\|_F^2 \leq \text{trace}(X_M) - r = 0$. Since $\min_k(\beta_k^{(in)} - \beta_k^{(out)}) \geq 0$,

$$\begin{aligned}
&\|X_M - X_0\|_F^2 \leq 2\langle X_0 - X_M, X_0\rangle \\
=&2\sum_k \sum_{i\in C_k}\sum_{j\in C_k} \frac{1}{m_k}\left(\frac{1}{m_k} - (X_M)_{ij}\right) \\
=&2\sum_k \sum_{i\in C_k} \frac{1}{m_k}\left(1 - \sum_{j\in C_k}(X_M)_{ij}\right) \\
\leq&\frac{2}{m_{\min}}\sum_k \sum_{i\in C_k}\left(1 - \sum_{j\in C_k}(X_M)_{ij}\right) \\
\leq&\frac{2}{m_{\min}\min_k(\beta_k^{(in)} - \beta_k^{(out)})}\langle Q, X_0 - X_M\rangle \\
\leq&\frac{2}{m_{\min}\min_k(\beta_k^{(in)} - \beta_k^{(out)})}\langle M - Q, X_M - X_0\rangle
\end{aligned}$$

□

# C Proof of Proposition 1

We first introduce the following result on sparse graph with Grothendieck's inequality by Guédon and Vershynin (2015).

**Lemma C.2** (Guédon and Vershynin (2015)). *Let $\mathcal{M}_G^+ = \{X : X \succeq 0, diag(X) \preceq I_n\}$, $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ be a symmetric matrix whose diagonal entries equal 0, and entries above the diagonal are independent random variables satisfying $0 \leq a_{ij} \leq 1$. Let $P = \mathbb{E}[A|Z]$. Assume that $\bar{p} := \frac{2}{n(n-1)}\sum_{i<j} Var(a_{ij}) \geq \frac{9}{n}$. Define the $\ell_\infty \to \ell_1$ norm of a matrix $M$ as $\|M\|_{\ell_\infty\to\ell_1} = \max_{x,y\in\{\pm\}^n}\sum_{i,j} x_i y_j M_{ij}$. Then, with probability at least $1 - e^3 5^{-n}$, we have $\max_{X\in\mathcal{M}_G^+}|\langle A - P, X\rangle| \leq K_G\|A-P\|_{\ell_\infty\to\ell_1} \leq 3K_G\bar{p}^{1/2}n^{3/2}$, where $K_G$ is the Grothendieck's constant, and its best know upper bound is 1.783.*

*Proof of Proposition 1.* Notice that $A$ and $P := \mathbb{E}[A|Z]$ has zero diagonals. Therefore,

$$\begin{aligned}\langle P - Q, X_A - X_0\rangle &= \sum_k \sum_{i\in C_k} a_k/n \left(\frac{1}{m_k} - (X_A)_{ii}\right) \\ &\leq \sum_k p_k - p_{\min}\text{trace}(X_A) \leq r(p_{\max} - p_{\min})\end{aligned}, \tag{C.1}$$

where $p_{\max} = \max_k a_k/n$ and $p_{\min} = \min_k a_k/n$. Thus by Lemma 1 and Eq (C.1),

$$\|X_A - X_0\|_F^2 \leq \frac{2}{m_{\min}\min_k(a_k/n - b_k/n)}\left(\langle A - P, X_A - X_0\rangle + r(p_{\max} - p_{\min})\right)$$

In sparse regime, both $m_{\min}X_0$ and $m_{\min}X_A$ belong to the set $\mathcal{M}_G^+$. Let $v_A = n\bar{p} \geq 9$, applying Lemma C.2 we get with probability at least $1 - e^3 5^{-n}$,

$$\|X_A - X_0\|_F^2 \leq \frac{22\sqrt{n^2 v_A}}{m_{\min}^2 \min_k(a_k/n - b_k/n)} + \frac{2r(p_{\max} - p_{\min})}{m_{\min}\min_k(a_k/n - b_k/n)}$$

Substituting $p_k = a_k/n, q_k = b_k/n$, and using the fact that

$$\frac{2r(p_{\max} - p_{\min})}{m_{\min}\min_k(p_k - q_k)} = \frac{2rm_{\min}(p_{\max} - p_{\min})}{m_{\min}^2\min_k(p_k - q_k)} \leq \frac{2\max_k a_k}{m_{\min}^2\min_k(p_k - q_k)} = o(\sqrt{n^2 g}),$$

Recall that $\alpha := m_{\max}/m_{\min}$, we get with probability tending to 1,

$$\frac{\|\hat{X} - X_0\|_F^2}{\|X_0\|_F^2} \leq \frac{23n^2\sqrt{v_A}}{rm_{\min}^2\min_k(a_k - b_k)} \leq \frac{23\alpha^2 r\sqrt{v_A}}{\min_k(a_k - b_k)}.$$

□

# D Proof of Proposition 2

*Proof of Proposition 2.* Recall that by definition, for $i \in C_k$, $Y_i - \mu_k$ is sub-gaussian random vector with sub-gaussian parameter $\psi_k$. Using the following concentration inequality from Hsu et al. (2012) for sub-gaussian random vectors, we have:

$$\text{For } i \in C_k,\ P(\|Y_i - \mu_k\|_2^2 > \psi_k^2(d + 2\sqrt{td} + 2t)) \leq e^{-t}$$

We take $t = c_k^2 d$ for $c_k \geq 1$. Since $1 + 2c_k + 2c_k^2 \leq 5c_k^2$ for $c_k \geq 1$, we get $P(\|X - \mathbb{E}X\|^2 \leq 5c_k^2\psi_k^2 d) \geq 1 - \exp(-c_k^2 d)$. Let $\Delta_k = \sqrt{5}c_k\psi_k\sqrt{d}$, we divide the nodes into "good nodes"

(those close to their population mean) $\mathcal{S}_k$ and the rest as in Eq. (5), which we redefine for completeness:

$$\mathcal{S}_k = \{i \in C_k : \|Y_i - \mu_k\| \leq \Delta_k\}, \qquad \mathcal{S} = \cup_{k=1}^{r} \mathcal{S}_k. \tag{D.2}$$

Let $m_c^{(k)} = m_k - |\mathcal{S}_k|$. We want to bound $m_c^{(k)}$ with high probability. Note that $m_c^{(k)} = \sum_{i \in C_k} \mathbf{1}(\|Y_i - \mu_k\| \geq \Delta_k)$ is a sum of i.i.d random variables. Therefore, using the Hoeffding bound we have:

$$P\left(m_c^{(k)} - m_k P(i \notin \mathcal{S}_k) \geq m_k \delta\right) \leq \exp(-2m_k\delta^2)$$

Using $\delta = \sqrt{\log m_k / 2m_k}$, we have:

$$P\left(m_c^{(k)} - m_k P(i \notin \mathcal{S}_k) \geq \sqrt{m_k \log m_k / 2}\right) \leq \frac{1}{m_k}$$

Since $P(i \notin \mathcal{S}_k) \leq \exp(-c_k^2 d)$, we have:

$$P\left(m_c^{(k)} \geq m_k \exp(-c_k^2 d) + \sqrt{m_k \log m_k / 2})\right) \leq \frac{1}{m_k}$$

Finally, using union bound over all clusters we get:

$$P\left(m_c \geq \sum_k m_k e^{-c_k^2 d} + \sum_k \sqrt{m_k \log m_k / 2}\right) \leq \sum_k \frac{1}{m_k} \tag{D.3}$$

Now recall the reference matrix from Eq. (7). For completeness we redefine $Q_K$ below.

$$(Q_K)_{ij} = \begin{cases} f(2\Delta_k), & \text{if } i, j \in C_k \\ \min\{f(d_{k\ell} - \Delta_k - \Delta_\ell), K_{ij}\}, & \text{if } i \in C_k, j \in C_\ell, k \neq \ell \end{cases} \tag{D.4}$$

By Lemma 1, all diagonal blocks are blockwise constant and the off-diagonal blocks are upper bounded by $f(d_{k\ell} - \Delta_k - \Delta_\ell)$. Let $\nu_k = f(2\Delta_k) - \max_{\ell \neq k} f(d_{k\ell} - \Delta_k - \Delta_\ell)$, and $\gamma = \min_k \nu_k$. If $\nu_k \geq 0$, we have

$$\|X_K - X_0\|_F^2 \leq \frac{2}{m_{\min}\gamma} \langle K - Q_K, X_K - X_0 \rangle$$

Apply Grothendieck's inequality,

$$\|X_K - X_0\|_F^2 \leq \frac{2K_G}{m_{\min}^2 \gamma} \|K - Q_K\|_{\ell_\infty \to \ell_1} \tag{D.5}$$

Now it remains to bound the $\ell_\infty \to \ell_1$ norm of $K - Q_K$. Note that if $i \in S_k, j \in S_\ell, k \neq \ell$, then by a simple use of triangle inequality we have $K_{ij} \leq f(d_{k\ell} - \Delta_k - \Delta_\ell)$, so $K_{ij} = (Q_K)_{ij}$; and if $i, j \in S_k$, then $K_{ij} \geq f(2\Delta_k)$.

$$\begin{aligned}
\|K - Q_K\|_{\ell_\infty \to \ell_1} &= \max_{x,y \in \{\pm\}^n} \sum_{i,j} x_i y_j \left(K_{ij} - (Q_K)_{ij}\right) \\
&\leq \max_{x,y \in \{\pm\}^n} \sum_{i,j \in \mathcal{S}} x_i y_j \left(K_{ij} - (Q_K)_{ij}\right) + \max_{x,y \in \{\pm\}^n} \sum_{i \notin \mathcal{S} \cup j \notin \mathcal{S}} x_i y_j \left(K_{ij} - (Q_K)_{ij}\right) \\
&\overset{(i)}{\leq} \max_{x,y \in \{\pm\}^n} \sum_{i,j \in \mathcal{S}} x_i y_j \left(K_{ij} - (Q_K)_{ij}\right) + 2m_c n \\
&\overset{(ii)}{=} \max_{x,y \in \{\pm\}^n} \sum_{k} \sum_{i,j \in \mathcal{S}_k} x_i y_j \left(K_{ij} - f(2\Delta_k)\right) + 2m_c n \\
&\leq \sum_k m_k^2 (1 - f(2\Delta_k)) + 2m_c n
\end{aligned} \tag{D.6}$$

where $(i)$ is due to $|K_{ij} - (Q_K)_{ij}| \leq 1$, and $(ii)$ comes from the definition of $Q_K$. Now Eq. (D.5) follows as

$$\begin{aligned}
\|X_K - X_0\|_F^2 \leq & \frac{4K_G \left(\sum_k m_k^2 (1 - f(2\Delta_k)) + 2m_c n\right)}{m_{\min}^2 \gamma} \\
= & \frac{4K_G}{m_{\min}^2} \sum_k \left( m_k^2 \frac{1 - f(2\Delta_k)}{\gamma} + 2m_k n e^{-c_k^2 d}/\gamma \right) + \frac{\sqrt{2} K_G n}{m_{\min}^2 \gamma} \sum_k \sqrt{m_k \log m_k}
\end{aligned} \tag{D.7}$$

Recall that $f(x) = \exp(-\eta x^2)$, and $\gamma = \min_k \{f(2\Delta_k) - \max_{\ell \neq k} f(d_{k\ell} - \Delta_k - \Delta_\ell)\}$. For simplicity, we assume $c_k = c_0$. We take $c_0 = \sqrt{\log\left(\frac{d_{\min}^2}{\psi_{\max}^2 d}\right)/d}$ and the scale parameter $\eta = \frac{\phi}{20 c_0^2 \psi_{\max}^2 d}$, for some $\phi > 0$, which will be chosen later. Furthermore, we also define

$$\xi = \frac{d_{\min}}{2\sqrt{5} c_0 \psi_{\max} \sqrt{d}} - 1. \tag{D.8}$$

If $\xi > 1$, then $d_{\min} > 4\sqrt{5} c_0 \psi_{\max} \sqrt{d}$, and hence $\gamma > 0$. Also, since $\eta (d_{\min} - 2\sqrt{5} c_0 \psi_{\max} \sqrt{d})^2 = \phi \xi^2$, $\forall k, \ell \in [r]$, if $d_{\min} := \min_{k\ell} d_{k\ell} > 4\sqrt{5} c_0 \psi_{\max} \sqrt{d}$, then

$$\gamma \geq f(2\sqrt{5} c_0 \psi_{\max} \sqrt{d}) - f(d_{\min} - 2\sqrt{5} c_0 \psi_{\max} \sqrt{d}) = \exp(-\phi) - \exp(-\phi \xi^2).$$

and

$$1 - f(2\Delta_k) \leq 1 - f(2\sqrt{5}c_0\psi_{\max}\sqrt{d}) = 1 - \exp(\phi)$$

Recall $\alpha = \frac{m_{\max}}{m_{\min}}$,

$$\begin{aligned}
&\|X_K - X_0\|_F^2 \qquad\qquad (D.9)\\
\leq& 4K_G r\alpha^2 \cdot \frac{1 - f(2\sqrt{5}c_0\psi_{\max}\sqrt{d}) + 2r\exp(-c_0^2 d)}{\gamma} + \frac{2\sqrt{2}K_G m_{\max} r^2\sqrt{m_{\max}\log m_{\max}}}{\gamma m_{\min}^2}\\
\leq& \frac{4K_G r\alpha^2}{\gamma}\left(1 - \exp(-\phi) + \frac{2r\psi_{\max}^2 d}{d_{\min}^2} + r\sqrt{\log m_{\max}/2m_{\max}}\right)\\
\leq& 4K_G r\alpha^2\left(\underbrace{\frac{(1 - \exp(-\phi) + 2r\psi_{\max}^2 d/d_{\min}^2}{\exp(-\phi) - \exp(-\phi\xi^2)}}_{A} + \underbrace{\frac{r\sqrt{\log m_{\max}/2m_{\max}}}{\exp(-\phi) - \exp(-\phi\xi^2)}}_{B}\right) \qquad (D.10)
\end{aligned}$$

We will first bound part (A).

$$(A) = \frac{\exp(\phi) - 1 + \exp(\phi)\frac{2r\psi_{\max}^2 d}{d_{\min}^2}}{1 - \exp(\phi - \phi\xi^2)} \overset{(i)}{\leq} \frac{\phi + \frac{\phi^2}{2}\exp(\phi) + \exp(\phi)\frac{2r\psi_{\max}^2 d}{d_{\min}^2}}{1 - \exp(\phi - \phi\xi^2)} \qquad (D.11)$$

where $(i)$ uses the Mean value theorem: for $e^x - 1 \leq x + e^y x^2/2$ for $y \in [0, x]$. If $\frac{d_{\min}}{\psi_{\max}\sqrt{d}} > \max\left\{1, \frac{180}{d}\right\}$, using the fact that $\log x \leq \sqrt{x}$, we have:

$$\frac{d_{\min}^2}{\psi_{\max}^2 d} > \frac{180}{d^2}\frac{d_{\min}}{\psi_{\max}} > \frac{180}{d}\log\left(\frac{d_{\min}^2}{\psi_{\max}^2 d}\right) = 180c_0^2.$$

Using Eq. (D.8), we see that $\xi > \frac{\sqrt{180}}{2\sqrt{5}} - 1 = 2$, and hence $\gamma > 0$. Now we pick $\phi = \frac{\log\xi}{\xi^2}$.

Now we will use this to obtain a lower bound on $1 - \exp(\phi - \phi\xi^2)$. Since $\xi \geq 2$, we have $\xi^2/4 \geq 1$. Hence

$$\begin{aligned}
1 - \exp(\phi - \phi\xi^2) &\geq 1 - \exp(\phi\xi^2/4 - \phi\xi^2)\\
&= 1 - \exp(-\phi 3\xi^2/4) = 1 - \exp(-3\log\xi/4) = 1 - \xi^{-3/4}\\
&\geq 1 - 2^{-3/4} = .4
\end{aligned}$$

Using the fact that the function $\frac{\log x}{x^2}$ is monotonically decreasing when $x > 2$, we see that $\phi < \log 2/2^2$ and $\exp(\phi) \leq 1.2$. Furthermore,

$$\gamma \geq \exp(-\phi)(1 - \exp(\phi(1 - \xi^2))) \geq .3 \qquad (D.12)$$

Now Eq. (D.11) yields:

$$(A) \leq \frac{\phi + 1.2\left(\frac{\phi^2}{2} + \frac{2r\psi_{\max}^2 d}{d_{\min}^2}\right)}{.4} \leq \frac{c\log\xi}{\xi^2} + \frac{6r\psi_{\max}^2 d}{d_{\min}^2}$$
$$\overset{(ii)}{\leq} \frac{c'\log(\xi+1)}{(\xi+1)^2} + \frac{6r\psi_{\max}^2 d}{d_{\min}^2} \leq c''\frac{\psi_{\max}^2 d}{d_{\min}^2}\log\left(\frac{d_{\min}}{\psi_{\max}\sqrt{d}}\right) + \frac{6r\psi_{\max}^2 d}{d_{\min}^2},$$

for some constant $c$. To get $(ii)$, note that

$$\frac{\log\xi}{\xi^2} \leq \frac{\log(\xi+1)}{\xi^2} \leq \frac{2.25\log(\xi+1)}{(\xi+1)^2}, \forall \xi > 2$$

Finally, we bound (B) in Eq. (D.10) using Eq. (D.12).

$$(B) = \frac{r\sqrt{\log m_{\max}/2m_{\max}}}{\exp(-\phi) - \exp(-\phi\xi^2)} \leq c_1 r\sqrt{\frac{\log m_{\max}}{m_{\max}}}$$

for some constant $c_1 > 0$. Putting pieces together, we have

$$\frac{\|X_K - X_0\|_F^2}{\|X_0\|_F^2} \leq C\alpha^2 \max\left(\frac{\psi_{\max}^2 d}{d_{\min}^2}\max\left\{\log\left(\frac{d_{\min}}{\psi_{\max}\sqrt{d}}\right), r\right\}, r\sqrt{\frac{\log m_{\max}}{m_{\max}}}\right)$$

□

# E Analysis for $X_{A+\lambda_n K}$

*Proof of Theorem 1.* Let $K_I$ be defined as in Eq. (7). Let

$$\gamma = \min_k(a_k/n - b_k/n + \lambda_n(f(2\Delta_k) - \max_{\ell\neq k} f(d_{k\ell} - \Delta_k - \Delta_\ell))).$$

When $\gamma \geq 0$, Lemma 1 with $Q = ZBZ^T + \lambda_n K_I$, we have

$$\|X_{A+\lambda_n K} - X_0\|_F^2 \leq$$
$$\frac{2}{m_{\min}\gamma}\left(\langle A - P, X_{A+\lambda_n K} - X_0\rangle + r(\max_k a_k/n - \min_k a_k/n) + \lambda_n\langle K - K_I, X_{A+\lambda_n K} - X_0\rangle\right).$$

Now by Grothendieck's inequality on both $\langle A - P, X_{A+\lambda_n K} - X_0\rangle$ and $\langle K - K_I, X_{A+\lambda_n K} - X_0\rangle$, one gets,

$$\|X_{A+\lambda_n K} - X_0\|_F^2 \leq \frac{2K_G}{m_{\min}^2\gamma}\left(2\|A - P\|_{\ell_\infty\to\ell_1} + r(\max_k a_k/n - \min_k a_k/n) + 2\lambda_n\|K - K_I\|_{\ell_\infty\to\ell_1}\right)$$

By Lemma C.2 and Eq (D.6),

$$\|X_{A+\lambda_n K} - X_0\|_F^2 \leq \frac{4K_G}{m_{\min}^2 \gamma}\left(6\sqrt{n^3\bar{p}} + \lambda_n\left(2m_c n + \sum_k m_k^2(1 - f(2\Delta_k))\right)\right)$$

Using $\lambda_n = \lambda_0/n$, $m_k = n\pi_k$, $m_{\min} = n\pi_{\min}$, and $\pi_0 := \sum_k (m_k \exp(-\Delta_k^2/(5\psi_k^2)) + \sqrt{m_k \log m_k/2})/n$ in conjunction with Eq (D.3), we get with probability tending to 1,

$$\|X_{A+\lambda_n K} - X_0\|_F^2 \leq 4K_G \frac{6\sqrt{v_A} + \lambda_0\left(2\pi_0 + \sum_k \pi_k^2(1 - f(2\Delta_k))\right)}{\pi_{\min}^2 \min_k(a_k - b_k + \lambda_0 \nu_k)}$$

□

# F Analysis of Covariate Clustering when $d \gg r$

Before proving Lemma 2, we clearly state our assumptions and other useful lemmas.

**Assumption 1.** *We assume that $M$ is of rank $r-1$, i.e. the means are not collinear, or linearly dependent, other than the fact that they are centered.*

**Lemma F.3.** *Let $M = \sum_k \pi_k \mu_k \mu_k^T$ and $S$ be the covariance matrix of $n$ data points from a sub-gaussian mixture $\{\pi_k, \mu_k, \Sigma_k\}_{k=1}^r$, then $S = M + \sum_i \pi_i \Sigma_i$. Let $\hat{S}$ be the sample covariance matrix $\hat{S} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})^T}{n}$. We have $\|\hat{S} - S\| \leq C\sqrt{\frac{d \log n}{n}}$ for some constant $C$ with probability bigger than $1 - O(n^{-d})$.*

This is a direct consequence of Corollary 5.50 from Vershynin (2010). Note that while Vershynin (2010) use the Orlicz norm formulation of sub-gaussian random variables, we use the more classical moment generating function based formulation (Wainwright, 2015; Hsu et al., 2012). However, these definitions are equivalent in the sense that a sub-gaussian parameter under the classical Definition 1 is within a constant factor of the sub-gaussian norm used by (Vershynin, 2010) (see *e.g.* Lemma 5.5 of Vershynin (2010)). The main ingredient of the proof is provided below.

**Lemma F.4.** *Let $U_{r-1}$ be the top $r-1$ eigenvectors of $\hat{S}$ estimated using $P_1$, $\theta_{r-1}(M)$ be the $r-1^{th}$ eigenvalue of $M$, and $\psi_{\max} = \max\{\psi_1, \cdots, \psi_r\}$ be the largest sub-gaussian parameter of all mixture components. For any vector $v$ in the span of $\{\mu_i\}_{i=1}^r$, as long as $\theta_{r-1}(M) > 5\left(\psi_{\max}^2 + C\sqrt{\frac{d \log^2 n}{n}}\right)$ we have $\|U_{r-1}^T v\| \geq \|v\|/2$ with probability at least $1 - \tilde{O}(n^{-d})$.*

*Proof.* First note that $\theta_{r-1}(M) \geq 0$ since $M$ is positive semi-definite. For simplicity we will use $\theta$ to denote $\theta_{r-1}(M)$. Take $n_1 = \frac{n}{\log n}$ and $v$ to be a vector in the span of $\{\mu_i\}_{i=1}^r$. By definition, we have $\|Mv\| \geq \theta\|v\|$. Let $R = \hat{S} - S$. Denote $\bar{\Sigma} = \sum_i \pi_i \Sigma_i$, and as defined in the statement of Lemma F.3, $S = M + \bar{\Sigma}$. Thus, we have $\hat{S} = M + \bar{\Sigma} + R$.

Below we show that the operator norm of $\bar{\Sigma}$ is bounded by $\psi_{\max}^2$. Let $w$ be the unit principal eigenvector of $\Sigma_k$. Consider a sub-gaussian vector $y$ from cluster $k$. Since $y - \mu_k$ is sub-gaussian with parameter $\psi_k$, a straightforward application of Definition 2 shows that, $w^T(y - \mu_k)$ is also sub-gaussian with sub-gaussian parameter $\psi_k$.

$$\mathbb{E} \exp(w^T(y - \mu_k)) \leq \exp(\psi_k^2/2)$$

Furthermore, its variance is

$$\mathrm{Var}(w^T(y - \mu_k)) = w^T \Sigma_k w = \|\Sigma_k\|$$

By property of sub-gaussian distributions (Wainwright, 2015), the variance of a sub-gaussian random variable is less than or equal to the square of the sub-gaussian parameter, hence $\|\Sigma_k\| \leq \psi_k^2$. Finally, since the operator norm is convex, using Jensen's inequality, we have:

$$\|\bar{\Sigma}\| \leq \max_k \|\Sigma_k\| \leq \psi_{\max}^2 \tag{F.13}$$

Since $S$ is estimated from $P_1$ with $n_1$ points, applying Lemma F.3 with $n = n_1$ we get $\|R\| \leq \epsilon = C\sqrt{\frac{d \log n_1}{n_1}}$. By Weyl's inequality,

$$\|\hat{S}v\| = \|(M + R + \bar{\Sigma})v\| \geq (\theta - \psi_{\max}^2 - \epsilon)\|v\|. \tag{F.14}$$

Let $U_{r:d}$ be the eigenspace orthogonal to $U_{r-1}$.Assume the contradiction that $\|U_{r-1}^T v\| < \|v\|/2$. Then there has to be a unit $d$ dimensional vector $u \in \mathrm{span}(U_{r:d})$, such that $|u^T v| > \|v\|/2$. Let us write $u = c\frac{v}{\|v\|} + \sqrt{1 - c^2} v^\perp$, for $|c| > 1/2$ and some unit vector $v^\perp$ orthogonal to $v$. Using triangle inequality and Eq F.14, this yields:

$$\|\hat{S}u\| \geq \frac{\theta - \psi_{\max}^2 - \epsilon}{2} - \sqrt{1 - c^2}\|\hat{S}v^\perp\|.$$

Now we will provide an upper bound on $\|\hat{S}v^\perp\|$.

$$\|\hat{S}v^\perp\| \overset{(i)}{=} \|(M + \bar{\Sigma} + R)v^\perp\| \overset{(ii)}{\leq} \|\bar{\Sigma}\| + \epsilon \overset{(iii)}{\leq} (\psi_{\max}^2 + \epsilon).$$

In the above equation, $(i)$ is true since $\hat{S} = M + \bar{\Sigma} + R$; $(ii)$ is true since $v^\perp$ is orthogonal to the span of $M$, and $(iii)$ is true by Eq. (F.13). Hence

$$\|\hat{S}u\| \geq \frac{\theta - 3(\psi_{\max}^2 + \epsilon)}{2}. \tag{F.15}$$

On the other hand, we also have the following inequality.

$$\|\hat{S}u\| \overset{(i)}{\leq} |\theta_r(\hat{S})| \overset{(ii)}{\leq} \theta_r(S) + \epsilon \overset{(iii)}{\leq} \theta_r(M) + \|\bar{\Sigma}\| + \epsilon \overset{(iv)}{\leq} \psi_{\max}^2 + \epsilon.$$

(i) holds since $u \in \text{span}(U_{r:d})$, (ii) and (iii) holds by Weyl's inequality, and (iv) holds since $\theta_r(M) = 0$ and by Eq. (F.13). This contradicts with Eq. (F.15) since we assume $\theta > 5(\psi_{\max}^2 + \epsilon)$. The result is proven by contradiction. □

We are now ready to prove Lemma 2.

*Proof of Lemma 2.* Recall that $Y_i' = U_{r-1}^T Y_i$ where $U_{r-1}$ and $Y_i$ are from two different partitions and hence independent. Let $Z_i \in [r]$ denote that latent variable associated with $i$. Thus,

$$\mathbb{E}[Y_i' | Z_i = a, P_1] = U_{r-1}^T \mathbb{E}[Y_i | Z_i = a] = U_{r-1}^T \mu_a.$$

The means of the new mixture are $\mu_a' := U_{r-1}^T \mu_a$. The covariance matrix after the projection is $U_{r-1}^T \Sigma_k U_{r-1}$ for cluster $k$. For any $x \in \mathbb{R}^{r-1}$,

$$\mathbb{E}[\exp(x^T U_{r-1}^T Y_j)] \leq \exp\left(\frac{\psi_k^2 \|x^T U_{r-1}^T\|^2}{2}\right) = \exp\left(\frac{\psi_k^2 \|x\|^2}{2}\right)$$

Hence the sub-gaussian parameter for cluster $k$ after projection is no larger than $\psi_k$. Furthermore, using Lemma F.4 we have $\min_{k \neq \ell} \|\mu_k' - \mu_\ell'\| = \min_{k \neq \ell} \|U_{r-1}^T(\mu_k - \mu_\ell)\| \geq d_{\min}/2$. Since this requires an application of Lemma F.4 to each of the vectors $\mu_k - \mu_\ell$, $k, \ell \in [r]$, the success probability is at least $1 - \tilde{O}(r^2 n^{-d})$ by union bound. □

# G From $X$ to Cluster Labels

From some solution matrix $\hat{X}$, we can apply Spectral Clustering on it to get the cluster labels. To remind the reader, Spectral Clustering proceeds by computing top $r$ eigenvectors of $\hat{X}$ and then doing k-means clustering on these. Below we present a theorem that bounds the misclassification error by the Frobenius norm of matrix difference, given a $(1+\epsilon)$ approximate solution in the $k$-means step in spectral clustering. The proof uses Lemma 5.3 of Lei et al. (2015), which for completeness we include below.

**Lemma G.5** (Lemma 5.3 of Lei et al. (2015)). *Define $\mathbb{M}_{n,r} \in \{0,1\}^{n\times r}$ be the set of membership matrices, such that any element of it has only exactly one 1 on each row. Consider two matrices $U, \hat{U} \in \mathbb{R}^{n\times r}$ such that $U = \Theta^* M$ with $\Theta^* \in \mathbb{M}_{n,r}$, $M \in \mathbb{R}^{r\times r}$. Let $G_k = \{i : \Theta^*_{ik} = 1\}$, i.e. the nodes in the $k^{th}$ cluster induced by $\Theta^*$. Consider the k-means problem:*

$$\arg \min_{\Theta\in\mathbb{M}_{n,r}, M\in\mathbb{R}^{r\times r}} \|\hat{U} - \Theta M\|_F^2. \tag{G.16}$$

*Let $(\hat{\Theta}, \hat{M})$ be a $(1+\epsilon)$ approximate solution to Eq. (G.16) for $\epsilon > 0$:*

$$\|\hat{U} - \hat{\Theta}\hat{M}\|_F^2 \le (1+\epsilon) \min_{\Theta\in\mathbb{M}_{n,r}, M\in\mathbb{R}^{r\times r}} \|\hat{U} - \Theta M\|_F^2.$$

*Let $\bar{U} = \hat{\Theta}\hat{M}$. For any $\delta_k \le \min_{\ell\neq k} \|M_{\ell *} - M_{k*}\|$, define*

$$S_k = \left\{ i \in G_k : \|\bar{U}_{i*} - U_{i*}\| \ge \frac{\delta_k}{2} \right\}. \tag{G.17}$$

*Then*

$$\sum_{k=1}^{r} |S_k|\delta_k^2 \le 4(4+2\epsilon)\|U - \hat{U}\|_F^2.$$

*Moreover, if $(16+8\epsilon)\|U-\hat{U}\|_F^2 \le n_k\delta_k^2$ for all $k \in [r]$, then there exists a $r\times r$ permutation matrix $J$ such that $\hat{\Theta}_{G^*} = \Theta_{G^*}J$, where $G = \cup_{k=1}^r (G_k \setminus S_k)$ .*

The following lemma bounds the mis-clustering error by the Frobenius norm of $\hat{X} - X_0$.

**Lemma G.6.** *Consider the clustering label obtained from spectral clustering from some clustering matrix $\hat{X}$. If $64(2+\epsilon)\|\hat{X} - X_0\|_F^2 \le 1$, then there exists $S_k \subset G_k$, such that*

$$\frac{1}{r}\sum_{k=1}^{r} \frac{|S_k|}{m_k} \le 64(2+\epsilon)\frac{\|\hat{X} - X_0\|_F^2}{\|X_0\|_F^2},$$

*and all nodes in $\cup_{k=1}^r (C_k \setminus S_k)$ are correctly clustered.*

*Proof.* Let us first connect the quantities in our clustering problem to the general Lemma G.5. We will use the true clustering matrix $Z$ as $\Theta^*$, and thus, by definition $G_k$ becomes the true $k^{th}$ community $C_k$ (See Table 1). Let $V, \hat{V}$ be the eigenvectors of the ground truth clustering matrix $X_0$ and the solution of our SDP problem $X$. By Davis-Kahan theorem Yu et al. (2014), there exists a $r \times r$ rotation matrix $O$ such that,

$$\|\hat{V} - VO\|_F^2 \le \frac{8\|\hat{X} - X_0\|_F^2}{(\theta_r(X_0) - \theta_{r+1}(X_0))^2} = 8\|\hat{X} - X_0\|_F^2 \tag{G.18}$$

Given the structure of $X_0$, its top $r$ eigenvalues are all ones, and

$$V = Z\text{diag}((1/\sqrt{m_1}, \ldots, 1/\sqrt{m_r}))O',$$

where $O'$ is some $r \times r$ rotation matrix. We take $M = \text{diag}(1/\sqrt{m_1}, \ldots, 1/\sqrt{m_r})O'O$, $U = VO$, and $\hat{U} = \hat{V}$. Now $\|M_{\ell *} - M_{k*}\|^2 = \frac{1}{m_k} + \frac{1}{m_\ell}$. Taking $\delta_k = \sqrt{\frac{1}{m_k} + \frac{1}{m_{\max}}}$, and applying Lemma G.5, we have

$$\sum_{k=1}^{r} \frac{|S_k|}{m_k} \leq \sum_{k=1}^{r} |S_k| \left( \frac{1}{m_k} + \frac{1}{m_\ell} \right) = \sum_{k=1}^{r} |S_k| \delta_k^2 \leq 64(2+\epsilon)\|\hat{X} - X_0\|_F^2.$$

□

**Corollary G.1.** *Let* $\psi = 2K_G \frac{6\sqrt{v_A} + \lambda_0 \left(2\pi_0 + \sum_k \pi_k^2 (1 - f(2\Delta_k))\right)}{\pi_{\min}^2 \min_k (a_k - b_k + \lambda_0 \nu_k)}$. *Under the conditions of Theorem 1, if* $64(2+\epsilon)\psi \leq 1$, *denote* $S_k$ *be the set of mis-clustered node in cluster* $k$. *Then with probability going to 1,*

$$\sum_{k=1}^{r} \frac{|S_k|}{m_k} \leq 128(2+\epsilon) K_G \frac{6\sqrt{v_A} + \lambda_0 \left(2\pi_0 + \sum_k \pi_k^2 (1 - f(2\Delta_k))\right)}{\pi_{\min}^2 \min_k (a_k - b_k + \lambda_0 \nu_k)}.$$

*Proof.* It can be shown by combining Theorem 1 and Lemma G.6. □

# H Additional Experiments

In this section, we present some additional experiments that are left out of the main paper due to space limit.

## H.1 Additional Parameter Settings

In this section, we present simulation results where the information on graph and covariates are not complementary but each have separation for all clusters. Consider a SBM with $n = 800, r = 3$ and equal sized clusters. The graph is generated from SBM with $B = 1e-3 * (7.6 I_3 + 0.4 E_3)$. The covariates are generated from an isotropic Gaussian mixture, with each pair of centers at distance 2, and covariance matrix for each cluster is $I_d$ where $d = 100$. In this extremely sparse case, where average expected degree is around 2, JCDC does not work well, so we only report the result of SDP-comb, SDP-net, SDP-cov and ACASC in Figure A. From the figure we can see that ACASC returns comparable NMI as that of SDP-comb.

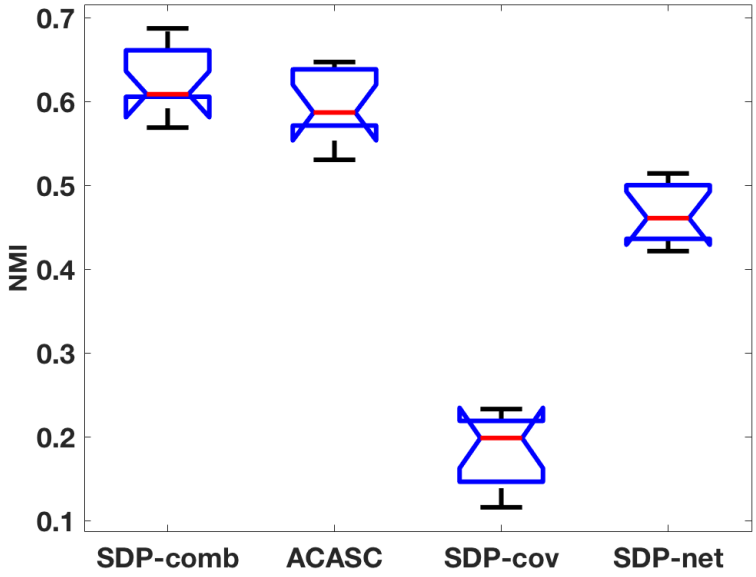



Figure A: NMI of various methods on aligned information from network and covariates.

## H.2 Choice of $m_{\min}$

For the experiment in Section H.1, we choose the true $m_{\min}$, which is $n/r$ in this example. If one specifies a $\frac{1}{m_{\min}}$ that is smaller than the true value, then the ground truth solution matrix will be excluded from the feasible set causing slower convergence and lower accuracy. On the other hand, if the specified value for $\frac{1}{m_{\min}}$ is larger than the true value, then we are searching over a larger set, and the SDP may return sub-optimal clustering result. To illustrate the effect of the choice of $m_{\min}$, we use the same parameter setting as in Section H.1, but with different values of specified $m_{\min}$, and plot the result in Figure B. From Figure B we see that if the element-wise upper bound is over-specified, then there exists an interval which allows similar clustering accuracy. However under-specifying the upper bound is more detrimental for the SDP.

Note that in the simulations and real data examples in Section 4 we set $m_{\min} = 1$. This choice works fine when the network is reasonably sparse, average degree 6 in a 800 node network. However, when the network is very sparse, for example in the setting for Section H.1, with average degree 2 for $n = 800$, the choice of $m_{\min}$ matters more. In principle, this can be tuned jointly with $\lambda$ as we tuned $r$ in Section 4.3.

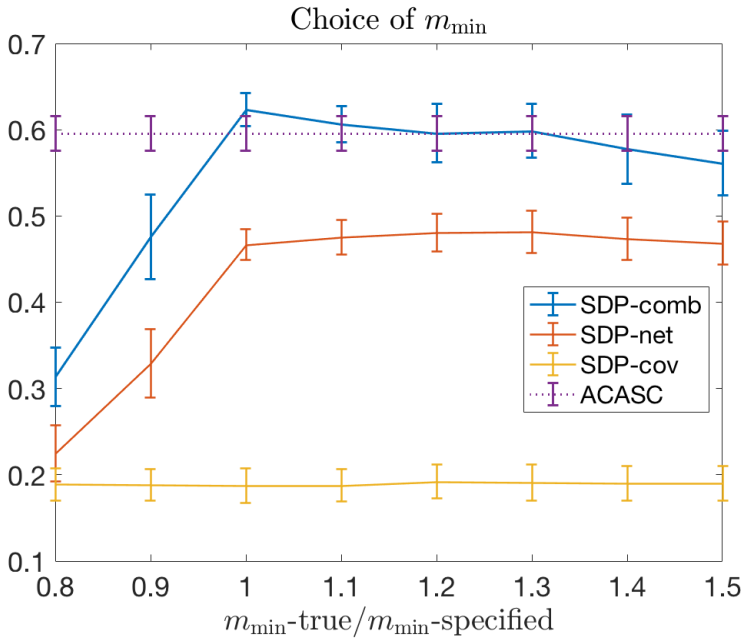



Figure B: Comparison of NMI on various choices of element-wise upper bound.

## H.3 Dimensionality reduction

In this section, we present experimental results for the dimensionality reduction algorithm proposed in Section 3.3 of the main paper. We construct a mixture of gaussians in a low dimensional subspace contaminated by high dimensional noise. The centers of the gaussians are $\mu_1 = (0, 3, 0 \cdots, 0)$, $\mu_2 = (-\sqrt{3}, 0, 0 \cdots, 0)$, $\mu_3 = (\sqrt{3}, 0, 0 \cdots, 0)$, and the covariance matrices are all equal to $I_d$. We compare the clustering results for increasing $d$ of the following methods:

1. (Unprocessed) The unprocessed features to construct the kernel matrix
2. (PCA-dim-reduce) Using dimensionality reduction as in Sec 3.3 by projecting onto the top $r$ singular vectors of the sample covariance matrix
3. (IF-PCA) Influential feature PCA proposed in Jin et al. (2016)
4. (Ground-truth) The ground truth covariates (first two dimensions).
5. (IF-PCA-on-rotated-data)A random rotation of the high dimensional data points, where the pairwise distances remain unchanged. Here the pairwise distances remain unchanged and hence Unprocessed, PCA-dim-reduce and Ground-truth perform identically as the original unrotated setting.

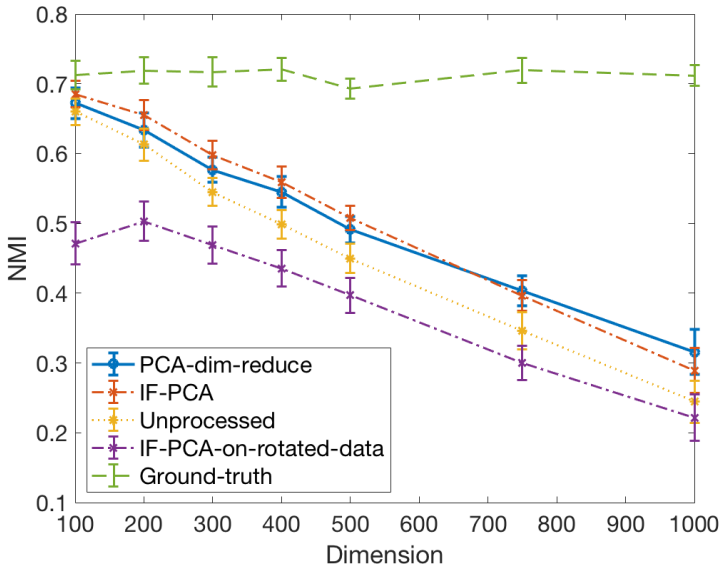



Figure C: NMI of clustering for dimensionality reduction and feature selection methods.

The average clustering NMI over 30 repetitions are reported in Figure C. The result shows that high dimensionality makes the clustering problem harder. Both dimensionality reduction and IF-PCA help improve the clustering result over the original high dimensional problem. Dimensionality reduction is robust to rotations, where IF-PCA on rotated space does not work well because the signal is no longer sparse in the rotated axes.

# References

Guédon, O. and Vershynin, R. (2015). Community detection in sparse networks via grothendieck's inequality. *Probability Theory and Related Fields*, pages 1–25.

Hsu, D., Kakade, S. M., and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab*, 17(52):1–6.

Jin, J., Wang, W., et al. (2016). Influential features pca for high dimensional clustering. *The Annals of Statistics*, 44(6):2323–2359.

Lei, J., Rinaldo, A., et al. (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Wainwright, M. (2015). Basic tail and concentration bounds. *URl: https://www. stat. berkeley. edu/.../Chap2_TailBounds_Jan22_2015. pdf (visited on 12/31/2017)*.

Yu, Y., Wang, T., and Samworth, R. J. (2014). A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323.