

Supplementary Materials for “Statistical-Computational Phase Transitions in Planted Models: The High-Dimensional Setting”

Yudong Chen

Jiaming Xu

The University of California, Berkeley

University of Illinois at Urbana-Champaign

yudong.chen@eecs.berkeley.edu

jxu18@illinois.edu

Abstract

We provide the proofs for the theorems in the main paper.

1 Proofs for Planted Clustering

In this section, Theorems 1–6 refer to the theorems in the main paper. Equations are numbered continuously from the main paper. We let $n_1 := rK$ and $n_2 := n - rK$ be the numbers of non-isolated and isolated nodes, respectively.

1.1 Proof of Theorem 1

The proof relies on information theoretical arguments and the Fano’s inequality [4]. We use $D(\text{Ber}(p)\|\text{Ber}(q))$ to denote the KL divergence between two Bernoulli distributions with mean p and q . We first state an upper bound on $D(\text{Ber}(p)\|\text{Ber}(q))$, which is used later in the proof:

$$D(\text{Ber}(p)\|\text{Ber}(q)) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \stackrel{(a)}{\leq} p \frac{p-q}{q} + (1-p) \frac{q-p}{1-q} = \frac{(p-q)^2}{q(1-q)}, \quad (16)$$

where (a) follows from the inequality $\log x \leq x - 1, \forall x \geq 0$. Let $\mathbb{P}_{(Y^*, A)}$ be the joint distribution of Y^* and A when Y^* is sampled from \mathcal{Y} uniformly at random and A is generated according to the planted clustering model. Because the supremum is lower bounded by the average, we have

$$\inf_{\hat{Y}} \sup_{Y^* \in \mathcal{Y}} \mathbb{P}[\hat{Y} \neq Y^*] \geq \inf_{\hat{Y}} \mathbb{P}_{(Y^*, A)}[\hat{Y} \neq Y^*]. \quad (17)$$

Let $H(X)$ be the entropy of a random variable X and $I(X; Z)$ the mutual information between X and Z . By Fano’s inequality, we have for any \hat{Y} ,

$$\mathbb{P}_{(Y^*, A)}(\hat{Y} \neq Y^*) \geq 1 - \frac{I(Y^*; A) + 1}{\log |\mathcal{Y}|}. \quad (18)$$

Simple counting gives that $|\mathcal{Y}| = \binom{n}{n_1} \frac{n_1!}{r!(K!)^r}$. Note that $\binom{n}{n_1} \geq (\frac{n}{n_1})^{n_1}$ and $\sqrt{n}(\frac{n}{e})^n \leq n! \leq e\sqrt{n}(\frac{n}{e})^n$. It follows that

$$|\mathcal{Y}| \geq (n/n_1)^{n_1} \frac{\sqrt{n_1}(n_1/e)^{n_1}}{e\sqrt{r}(r/e)^r e^r K^{r/2}(K/e)^{n_1}} \geq \left(\frac{n}{K}\right)^{n_1} \frac{1}{e(r\sqrt{K})^r}.$$

This implies $\log |\mathcal{Y}| \geq \frac{1}{2}n_1 \log \frac{n}{K}$ under the assumption that $8 \leq K \leq n/2$ and $n \geq 32$. On the other hand, note that $H(A) \leq \binom{n}{2}H(A_{12})$ because the A_{ij} 's are identically distributed by symmetry. Furthermore, the A_{ij} 's are independent conditioned on Y^* , so $H(A|Y^*) = \binom{n}{2}H(A_{12}|Y_{12}^*)$. It follows that $I(Y^*; A) = H(A) - H(A|Y^*) \leq \binom{n}{2}I(Y_{12}^*; A_{12})$. We bound $I(Y_{12}^*; A_{12})$ below. Observe that

$$\mathbb{P}(Y_{12}^* = 1) = \frac{\binom{n-2}{K-2} \binom{n-K}{K} \cdots \binom{n-rK+K}{K} \frac{1}{(r-1)!}}{|\mathcal{Y}|} = \alpha := \frac{K}{n},$$

and thus $\mathbb{P}(A_{12} = 1) = \beta := \alpha p + (1 - \alpha)q$. It follows that

$$\begin{aligned} I(Y_{12}^*; A_{12}) &= \alpha D(\text{Ber}(p) \parallel \text{Ber}(\beta)) + (1 - \alpha)D(\text{Ber}(q) \parallel \text{Ber}(\beta)) \\ &\stackrel{(a)}{\leq} \alpha \frac{(p - \beta)^2}{\beta(1 - \beta)} + (1 - \alpha)q \frac{(q - \beta)^2}{\beta(1 - \beta)} \\ &= \frac{\alpha(1 - \alpha)(p - q)^2}{\beta(1 - \beta)} \stackrel{(b)}{\leq} \frac{\alpha(p - q)^2}{q(1 - q)}, \end{aligned}$$

where (a) follows from (16) and (b) follows because $\beta(1 - \beta) \geq \alpha p(1 - p) + (1 - \alpha)q(1 - q)$ due to the concavity of $x(1 - x)$. Hence we have $I(Y^*; A) \leq \frac{n_1(K-1)(p-q)^2}{2q(1-q)}$. Combining with (18), we obtain

$$\mathbb{P}_{(Y^*, A)}(Y \neq Y^*) \geq 1 - \frac{\frac{n_1(K-1)(p-q)^2}{q(1-q)} + 2}{n_1 \log \frac{n}{K}} \geq 3/4 - \frac{(K-1)(p-q)^2}{q(1-q) \log \frac{n}{K}}, \quad (19)$$

where the last inequality holds because $n_1 \log \frac{n}{K} \geq 8$ when $K \geq n/2$ and $n \geq 32$. The RHS above is at least $1/2$ when the first condition (1) in the theorem holds. Substituting into (17) proves sufficiency of (1).

We now turn to the third condition (3). Since $p > q$, we have $\beta \leq p$ by the definition of β . Moreover, $\beta \geq \alpha p$ and $1 - \beta = 1 - q - \alpha(p - q) \geq (1 - \alpha)(1 - q)$. It follows that

$$\begin{aligned} I(Y_{12}^*; A_{12}) &= \alpha p \log \frac{p}{\beta} + \alpha(1 - p) \log \frac{(1 - p)}{1 - \beta} + (1 - \alpha)D(\text{Ber}(q) \parallel \text{Ber}(\beta)) \\ &\leq \alpha p \log \frac{1}{\alpha} + (1 - \alpha) \frac{(q - \beta)^2}{\beta(1 - \beta)} = \alpha p \log \frac{1}{\alpha} + \frac{\alpha^2(1 - \alpha)(p - q)^2}{\beta(1 - \beta)} \\ &\leq \alpha p \log \frac{1}{\alpha} + \frac{\alpha(p - q)^2}{p(1 - q)} \leq \alpha p \log \frac{e}{\alpha}. \end{aligned}$$

where the first inequality follows from $p/\beta \leq 1/\alpha$, $1 - \beta \geq 1 - p$ and (16), the second inequality follows from $\beta(1 - \beta) \geq \alpha(1 - \alpha)p(1 - q)$, and the last inequality follows from $p - q \leq p(1 - q)$. By the definition of α , we have $\alpha \geq \frac{K(K-1)}{n(n-1)} \geq \frac{K}{en}$ when $K \geq 8$. It follows that $I(Y_{12}^*; A_{12}) \leq \alpha p \log \frac{e^2 n}{K}$, and thus $I(Y^*; A) \leq \frac{1}{2}n_1 K p \log \frac{e^2 n}{K}$. By equation (19), if $Kp \leq \frac{1}{16}$, i.e., the condition (3) in the theorem holds, then $\mathbb{P}(Y \neq Y^*) \geq 1/2$.

It remains to prove the sufficiency of the second condition (3). Let $\bar{M} = n - K$ and $\bar{\mathcal{Y}} = \{Y_0, Y_1, \dots, Y_{\bar{M}}\}$ be a subset of \mathcal{Y} with cardinality $\bar{M} + 1$, which is specified later. Let $\mathbb{P}_{(Y^*, A)}$ denote the joint distribution of Y^* and A when Y^* is sampled from $\bar{\mathcal{Y}}$ uniformly at random and then A is generated according to the planted clustering model. By Fano's inequality, we have

$$\inf_{\hat{Y}} \sup_{Y^* \in \bar{\mathcal{Y}}} \mathbb{P}[\hat{Y} \neq Y^*] \geq \inf_{\hat{Y}} \bar{\mathbb{P}}_{(Y^*, A)}[\hat{Y} \neq Y^*] \geq \inf_{\hat{Y}} \left\{ 1 - \frac{I(Y^*; A) + 1}{\log |\hat{\mathcal{Y}}|} \right\}. \quad (20)$$

We construct $\bar{\mathcal{Y}}$ as follows. Let Y_0 be the clustering matrix such that the clusters $\{C_l\}_{l=1}^r$ are given by $C_l = \{(l-1)K+1, \dots, lK\}$. Informally, each Y_i with $i \geq 1$ is obtained from Y_0 by swapping two nodes in two different clusters. More specifically, for each $i \in [\bar{M}]$, (i) if the $(K+i)$ -th node belongs to cluster C_l for some l , then Y_i has the first right cluster as $\{1, 2, \dots, K-1, K+i\}$ and the l -th right cluster as $D_l \setminus \{K+i\} \cup \{K\}$, and all the other clusters identical to Y_0 ; (ii) if the $(K+i)$ -th node is an isolated node in Y_0 , then Y_i has the first right cluster as $\{1, 2, \dots, K-1, K+i\}$ and node K as an isolated node, and all the other clusters identical to Y_0 .

Let \mathbb{P}_i be the distribution of the graph A conditioned on $Y^* = Y_i$. Note that each \mathbb{P}_i is a product of $\frac{1}{2}n(n-1)$ Bernoulli distributions. We have the following chain of inequalities:

$$\begin{aligned} I(Y^*; A) &\stackrel{(a)}{\leq} \frac{1}{(M+1)^2} \sum_{i,i'=0}^M D(\mathbb{P}_i \|\mathbb{P}_{i'}) \\ &\leq \max_{i,i'=0,\dots,M} D(\mathbb{P}_i \|\mathbb{P}_{i'}) \\ &\stackrel{(b)}{\leq} 3KD(\text{Ber}(p) \|\text{Ber}(q)) + 3KD(\text{Ber}(q) \|\text{Ber}(p)) \\ &\stackrel{(c)}{\leq} 3K(p-q)^2 \frac{1}{\min\{p(1-p), q(1-q)\}} \end{aligned}$$

where (a) follows from the convexity of KL divergence, (b) follows by our construction of $\{Y_i\}$, and (c) follows from (16). If (2) holds, then $I(Y; A) \leq \frac{1}{4} \log(n-K) = \frac{1}{4} \log |\bar{\mathcal{Y}}|$. Since $\log(n-K) \geq \log(n/2) \geq 4$ if $n \geq 128$, it follows from (20) that the minimax error probability is at least $1/2$.

1.2 Proof of Theorem 2

Let $\langle X, Y \rangle := \text{Tr}(X^\top Y)$ denote the inner product between two matrices. Assume that $p > q$ first. For any feasible solution $Y \in \mathcal{Y}$ of (4), we define $\Delta(Y) := \langle A, Y^* - Y \rangle$. To prove the theorem, it suffices to show that $\Delta(Y) > 0$ for all feasible Y with $Y \neq Y^*$. For simplicity, in this proof we use a different convention that $Y_{ii}^* = 0$ and $Y_{ii} = 0$ for all $i \in V$. Note that

$$\Delta(Y) = \langle \mathbb{E}[A], Y^* - Y \rangle + \langle A - \mathbb{E}[A], Y^* - Y \rangle, \quad (21)$$

where $\mathbb{E}[A] = q\mathbf{1}\mathbf{1}^\top + (p-q)Y^* - q\mathbf{I}$, $\mathbf{1}$ is the all one vector in \mathbb{R}^\times and \mathbf{I} is the $n \times n$ identity matrix. Let $d(Y) := \langle Y^*, Y^* - Y \rangle$; since $\sum_{i,j} Y_{ij} = \sum_{i,j} Y_{ij}^*$, we have

$$\langle \mathbb{E}[A], Y^* - Y \rangle = (p-q)d(Y). \quad (22)$$

On the other hand, observe that

$$\langle A - \mathbb{E}[A], Y^* - Y \rangle = 2 \underbrace{\sum_{(i<j): \substack{Y_{ij}^*=1 \\ Y_{ij}=0}} (A_{ij} - p)}_{T_1(Y)} - 2 \underbrace{\sum_{(i<j): \substack{Y_{ij}^*=0 \\ Y_{ij}=1}} (A_{ij} - q)}_{T_2(Y)}.$$

Here $T_1(Y)$ ($T_2(Y)$, resp.) is the sum of $\frac{1}{2}d(Y)$ i.i.d. centered Bernoulli random variables with parameter p (q , resp.). Let $\delta_1 = (p-q)/(2p)$ and $\delta_2 = (p-q)/(2q)$. By the Bernstein inequality, we have for each fix $Y \in \mathcal{Y}$,

$$\begin{aligned} \mathbb{P} \left\{ T_1(Y) \leq -\frac{\delta_1}{2} d(Y)p \right\} &\leq \exp \left(-\frac{\delta_1^2}{4(1-p) + 4\delta_1/3} d(Y)p \right) \stackrel{(a)}{\leq} \exp \left(-\frac{(p-q)^2}{20p(1-q)} d(Y) \right), \\ \mathbb{P} \left\{ T_2(Y) \geq \frac{\delta_2}{2} d(Y)q \right\} &\leq \exp \left(-\frac{\delta_2^2}{4(1-q) + 4\delta_2/3} d(Y)q \right) \stackrel{(b)}{\leq} \exp \left(-\frac{(p-q)^2}{20p(1-q)} d(Y) \right), \end{aligned}$$

where (a) and (b) hold because $p > q$ and $p - q \leq p(1 - q)$. It follows from the union bound that

$$\mathbb{P} \left\{ \frac{1}{2} \langle A - \mathbb{E}[A], Y^* - Y \rangle = T_1(Y) - T_2(Y) \leq -\frac{1}{2}(p - q)d(Y) \right\} \leq 2 \exp \left(-\frac{(p - q)^2}{20p(1 - q)} d(Y) \right).$$

This implies

$$\mathbb{P} \{ \Delta(Y) \leq 0 \} \leq 2 \exp \left(-\frac{(p - q)^2}{20p(1 - q)} d(Y) \right) \quad (23)$$

in view of (21) and (22). This bound holds for each fixed $Y \in \mathcal{Y}$.

We prove to bound the probability of the event $\{\exists Y \in \mathcal{Y} : Y \neq Y^*, \Delta(Y) \leq 0\}$. Note that $2(K - 1) \leq d(Y) \leq rK^2$ for any feasible $Y \neq Y^*$, where the lower bound is achieved by swapping a node in V_1 with a node in V_2 , and the upper bound follows from $\sum_{i,j} Y_{ij}^* \leq rK^2$. The key step is to upper-bound the cardinality of the set $\{Y \in \mathcal{Y} : d(Y) = t\}$ for each t . This is done in the following combinatorial lemma.

Lemma 1.1. *For each $t \in [2(K - 1), rK^2]$, we have*

$$|\{Y \in \mathcal{Y} : d(Y) = t\}| \leq \frac{25t^2}{K^2} n^{20t/K}. \quad (24)$$

We prove the lemma in Section 1.2.1. Combining the lemma with (23) and the union bound, we obtain

$$\begin{aligned} & \mathbb{P} \{ \exists Y \in \mathcal{Y} : Y \neq Y^*, \Delta(Y) \leq 0 \} \\ & \leq \sum_{t=2K-2}^{rK^2} \mathbb{P} \{ \exists Y \in \mathcal{Y} : d(Y) = t, \Delta(Y) \leq 0 \} \\ & \leq 2 \sum_{t=2K-2}^{rK^2} |\{ \exists Y \in \mathcal{Y} : d(Y) = t \}| \exp \left(-\frac{(p - q)^2 t}{20p(1 - q)} \right) \\ & \leq 2 \sum_{t=2K-2}^{rK^2} \frac{25t^2}{K^2} n^{20t/K} \exp \left(-\frac{(p - q)^2 t}{20p(1 - q)} \right) \\ & \stackrel{(a)}{\leq} 50 \sum_{t=2K-2}^{rK^2} n^2 n^{-5t/K} \\ & \leq 50rK^2 n^{-3} \leq 50n^{-1}, \end{aligned}$$

where (a) follows from the assumption that $(p - q)^2 K \geq C' p(1 - q) \log n$ for a large constant C' . This means Y^* is the unique optimal solution with high probability. A similar argument applies to the case with $q > p$. This completes the proof of the theorem.

1.2.1 Proof of Lemma 1.1

Recall that C_1^*, \dots, C_r^* are the true clusters associated with Y^* . Fix a $Y \in \mathcal{Y}$ with $d(Y) = t$. The cluster matrix Y defines a new ordered partition (C_1, \dots, C_{r+1}) of V according to the following procedure.

1. Let $C_{r+1} := \{i : Y_{ij} = 0, \forall j\}$.

2. The nodes in $V \setminus C_{r+1}$ can be further partitioned into r new clusters of size K , where nodes i and j are in the same cluster if and only if $Y_{ij} = 1$; we define an ordering C_1, \dots, C_r of these r new clusters as follows.

- (a) For each new cluster C , if there exists a $k \in [r]$ such that $|C \cap C_k^*| > K/2$, then we label this new cluster as C_k ; this label is unique because the cluster size is K .
- (b) The remaining clusters are labeled arbitrarily.

This new partition has the following properties:

- (A0) $(C_1, \dots, C_r, C_{r+1})$ is a partition of V , and $|C_k| = K$ for all $k \in [r]$.
- (A1) For every $k \in [r]$, either $|C_k \cap C_k^*| > K/2$, or $|C_{k'} \cap C_k^*| \leq K/2$ for all $k' \in [r]$;
- (A2) We have

$$\sum_{k=1}^r \left(|C_k^* \cap C_{r+1}|^2 - |C_k^* \cap C_{r+1}| + \sum_{\substack{k', k'' \in [r+1] \\ k' \neq k''}} |C_k^* \cap C_{k'}| |C_k^* \cap C_{k''}| \right) = t.$$

Here, Properties (A0) and (A1) are direct consequences of how we label the new clusters, and Property (A2) follows from the following equalities

$$\begin{aligned} t &= d(Y) = |\{(i, j) \in V \times V : Y_{ij}^* = 1, Y_{ij} = 0\}| \\ &= \sum_{k=1}^r |\{(i, j) : (i, j) \in C_k^* \times C_k^*, i \neq j, Y_{ij} = 0\}| \\ &= \sum_{k=1}^r |\{(i, j) : (i, j) \in C_k^* \times C_k^*, i \neq j, (i, j) \in C_{r+1} \times C_{r+1}\}| \\ &\quad + \sum_{k=1}^r \sum_{\substack{k', k'' \in [r+1] \\ k' \neq k''}} |\{(i, j) : (i, j) \in C_k^* \times C_k^*, (i, j) \in C_{k'} \times C_{k''}\}|. \end{aligned}$$

Since these properties are satisfied by any Y with $d(Y) = t$, we have

$$|\{Y \in \mathcal{Y} : d(Y) = t\}| \leq |\{(C_1, \dots, C_r, C_{r+1}) : \text{it has properties (A0)–(A2)}\}|. \quad (25)$$

To prove the lemma, it suffices to upper bound the right hand side of (25).

Fix an ordered partition $\mathcal{C} := (C_1, \dots, C_r, C_{r+1})$ with properties (A0)–(A2). We consider the first true cluster, C_1^* . Define $m_1 := \sum_{k' \in [r+1], k' \neq 1} |C_{k'} \cap C_1^*|$, which is the number of nodes in C_1^* that are misclassified by \mathcal{C} . We consider two cases.

- If $|C_1 \cap C_1^*| > K/10$, then

$$\sum_{\substack{k', k'' \in [r+1] \\ k' \neq k''}} |C_1^* \cap C_{k'}| |C_1^* \cap C_{k''}| \geq 2|C_1^* \cap C_1| \sum_{\substack{k'' \in [r+1] \\ k'' \neq 1}} |C_1^* \cap C_{k''}| > m_1 K/5.$$

- If $|C_1 \cap C_1^*| \leq K/10$, then by condition (A1) we must have $|C_{k'} \cap C_1^*| \leq K/2$ for all $1 \leq k' \leq r$

and $m_1 > 9K/10$. Hence,

$$\begin{aligned}
& \sum_{\substack{k', k'' \in [r+1] \\ k' \neq k''}} |C_1^* \cap C_{k'}| |C_1^* \cap C_{k''}| + |C_1^* \cap C_{r+1}|^2 - |C_1^* \cap C_{r+1}| \\
& \geq \sum_{\substack{k', k'' \in [r+1] \\ k' \neq k''}} \mathbb{I}_{\{k' \neq 1\}} \mathbb{I}_{\{k'' \neq 1\}} |C_{k'} \cap C_1^*| |C_{k''} \cap C_1^*| + |C_1^* \cap C_{r+1}|^2 - |C_1^* \cap C_{r+1}| \\
& = m_1^2 - \sum_{2 \leq k' \leq r} |C_{k'} \cap C_1^*|^2 - |C_1^* \cap C_{r+1}| \\
& \geq m_1^2 - \frac{1}{2} K m_1 \geq \frac{2}{5} m_1 K.
\end{aligned}$$

We conclude that we always have

$$\sum_{\substack{k', k'' \in [r+1] \\ k' \neq k''}} |C_1^* \cap C_{k'}| |C_1^* \cap C_{k''}| + |C_1^* \cap C_{r+1}|^2 - |C_1^* \cap C_{r+1}| \geq \frac{1}{5} m_1 K.$$

The above inequality holds if we replace C_1^* by C_k^* and m_1 by m_k (defined similarly) for each $k \in [r]$. It follows that

$$\sum_{k=1}^r \left(|C_k^* \cap C_{r+1}|^2 - |C_k^* \cap C_{r+1}| + \sum_{\substack{k', k'' \in [r+1] \\ k' \neq k''}} |C_k^* \cap C_{k'}| |C_k^* \cap C_{k''}| \right) \geq \frac{K}{5} \sum_{k=1}^r m_k.$$

Thus by Property (A2), we have $\sum_{k \in [r]} m_k \leq 5t/K$, i.e., the total number of misclassified nodes in V_1 is upper bounded by $5t/K$. This means that the total number of misclassified nodes in V_2 is also upper bounded by $5t/K$ because by our cluster size constraint, one misclassified node in V_2 must produce one misclassified node in V_1 . Therefore, the total number of misclassified nodes in V_1 can at most take $5t/K$ different values and the same is true for the total number of misclassified nodes in V_2 . For a fixed number of misclassified nodes in V_1 and V_2 , there are at most $n_1^{5t/K} n_2^{5t/K}$ different ways to choose these misclassified nodes. Each misclassified node in V_1 can be assigned to one of $r - 1$ different clusters or leave isolated, and each misclassified node in V_2 can be assigned to one of r different clusters. Hence, the right hand side of (25) is upper bounded by $\frac{25t^2}{K^2} n_1^{5t/K} n_2^{5t/K} r^{10t/K} \leq \frac{25t^2}{K^2} n^{20t/K}$, which proves the lemma.

1.3 Proof of Theorem 3

The proof uses several matrix norms. The spectral norm $\|X\|$ of a matrix X is the largest singular value of X . The nuclear norm $\|X\|_*$ is the sum of singular values. We also need the L_1 norm $\|X\|_1 = \sum_{i,j} |X_{ij}|$ and the L_∞ norm $\|X\|_\infty = \max_{i,j} |X_{ij}|$. Let $\langle X, Y \rangle = \text{Tr}(X^\top Y)$ denote the inner product between two matrices. For a vector x , $\|x\|_2$ is the usual Euclidean norm.

Define $\Delta(Y) = \langle Y^* - Y, A \rangle$. It suffices to show that $\Delta(Y) > 0$ for all feasible solution Y of the program (6)–(8) with $Y \neq Y^*$. Rewrite $\Delta(Y)$ as

$$\Delta(Y) = \langle \bar{A}, Y^* - Y \rangle + \langle A - \bar{A}, Y^* - Y \rangle = \langle \bar{A}, Y^* - Y \rangle + \lambda \langle W, Y^* - Y \rangle, \quad (26)$$

where $\bar{A} = q\mathbf{1}\mathbf{1}^\top + (p - q)Y^*$ and $W := (A - \bar{A})/\lambda$. Because $\sum_{i,j} Y_{ij} = rK^2 = \sum_{i,j} Y_{ij}^*$ and $Y_{ij} \in [0, 1]$, the first term in the RHS of (26) satisfies

$$\langle \bar{A}, Y^* - Y \rangle = (p - q)\langle Y^*, Y^* - Y \rangle = \frac{p - q}{2} \|Y^* - Y\|_1.$$

We now control the second term in (26). Note that $\text{Var}[A_{ij}] \leq p(1 - q)$. Define $\sigma^2 := \max\{p(1 - q), q(1 - p)\}$. Note that $\|\bar{A} - \mathbb{E}[A]\| \leq 1$. By assumption, $\sigma^2 \geq C' \log n / K$ for a constant C' . We need the following bound, which is proved in Section 1.3.1 to follow.

Lemma 1.2. *Under the notation above, if $\sigma^2 \geq C' \log n / K$ for a constant C' , then there exists a constant C such that with high probability $\|A - \mathbb{E}[A]\| \leq C \sqrt{\sigma^2 K \log n + q(1 - q)n}$.*

It follows that with high probability

$$\|A - \bar{A}\| \leq \|A - \mathbb{E}[A]\| + \|\bar{A} - \mathbb{E}[A]\| \leq C \sqrt{\sigma^2 K \log n + q(1 - q)n}$$

for a universal constant C . Define $\lambda := C \sqrt{\sigma K \log n + q(1 - q)n}$ and the normalized noise matrix W . Thus w.h.p. $\|W\| \leq 1$.

Let u_k be the normalized characteristic vector of cluster C_k^* , i.e., $u_k(i) = 1/\sqrt{K}$ if node i is in cluster C_k^* and $u_k(i) = 0$ otherwise. Let $U = [u_1, \dots, u_r]$. Then $Y^* = KU U^\top$ is the singular value decomposition of Y^* . Define the projections $\mathcal{P}_T(M) = UU^\top M + MU U^\top - UU^\top M U U^\top$ and $\mathcal{P}_{T^\perp}(M) = M - \mathcal{P}_T(M)$. Because $\|\mathcal{P}_{T^\perp}(W)\| \leq \|W\| \leq 1$ w.h.p., $UU^\top + \mathcal{P}_{T^\perp}(W)$ is a subgradient of $\|X\|_*$ at $X = Y^*$. Since $\|Y^*\|_* = n_1$, it follows that for any feasible Y ,

$$0 \geq \|Y\|_* - \|Y^*\|_* \geq \langle UU^\top + \mathcal{P}_{T^\perp}(W), Y - Y^* \rangle.$$

Substituting into (26), we obtain that for any feasible Y ,

$$\begin{aligned} \Delta(Y) &\geq \frac{p - q}{2} \|Y^* - Y\|_1 + \lambda \langle \mathcal{P}_T(W) - UU^\top, Y^* - Y \rangle \\ &\geq \left(\frac{p - q}{2} - \lambda \|UU^\top\|_\infty - \|\mathcal{P}_T(\lambda W)\|_\infty \right) \|Y^* - Y\|_1 \\ &= \left(\frac{p - q}{2} - \frac{\lambda}{K} - \|\mathcal{P}_T(\lambda W)\|_\infty \right) \|Y^* - Y\|_1, \end{aligned} \tag{27}$$

where the last equality holds because $\|UU^\top\|_\infty = 1/K$.

We proceed by bounding the term $\|\mathcal{P}_T(\lambda W)\|_\infty$ in (27). From the definition of \mathcal{P}_T , we have

$$\|\mathcal{P}_T(\lambda W)\|_\infty \leq \|UU^\top(\lambda W)\|_\infty + \|(\lambda W)UU^\top\|_\infty + \|UU^\top(\lambda W)UU^\top\|_\infty \leq 3\|UU^\top(\lambda W)\|_\infty.$$

Assume node i belongs to cluster k . Then

$$(UU^\top(\lambda W))_{ij} = (u_k u_k^\top(\lambda W))_{ij} = (1/K) \sum_{i' \in C_k^*} (\lambda W)_{i'j},$$

which is the average of K independent random variables. By Bernstein's inequality (Theorem A.1) we have with probability at least $1 - n^{-4}$,

$$\left| \sum_{i' \in C_k^*} (\lambda W)_{i'j} \right| \leq \sqrt{6\sigma^2 K \log n} + 2 \log n \leq C_1 \sigma \sqrt{K \log n},$$

for some constant C_1 , where the last inequality follows from the assumption (9). It follows from the union bound over all (i, j) that $\|\mathcal{P}_T(\lambda W)\|_\infty \leq 3\|UU^\top(\lambda W)\|_\infty \leq 3C_1 \sigma \sqrt{\log n / K}$ with probability at least $1 - n^{-2}$. Substituting back to (27), we conclude that with probability at least $1 - n^{-1}$,

$$\Delta(Y) \geq \left(\frac{p - q}{2} - \frac{\lambda}{K} - 3C_1 \sigma \sqrt{\log n / K} \right) \|Y^* - Y\|_1 > 0$$

for all feasible $Y \neq Y^*$, where the last inequality follows from the assumption (9).

1.3.1 Proof of Lemma 1.2

Let $R := \text{support}(Y^*)$ and $\mathcal{P}_R(\cdot) : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}$ be the operator which sets the entries outside of R to be zero. Let $B_1 = \mathcal{P}_R(A - \mathbb{E}[A])$ and $B_2 = A - \mathbb{E}[A] - B_1$. Then B_1 is a block-diagonal matrix with r blocks of size $K \times K$ and has entries with variance bounded by σ^2 . Applying the matrix Bernstein inequality [5], we get that with high probability, $\|B_1\| \leq c_1 \sqrt{\sigma^2 K \log n}$ for a constant c_1 . On the other hand, B_2 has entries with variance bounded by $\max\{q(1-q), c_2 \log n/n\}$ for a constant c_2 . By Lemma 1.3 below, we obtain that $\|B_2\| \leq c_3 \max\{\sqrt{q(1-q)n}, \sqrt{\log n}\}$ for a constant c_3 . It follows that

$$\begin{aligned} \|A - \mathbb{E}[A]\| &\leq \|B_1\| + \|B_2\| \leq c_1 \sqrt{\sigma^2 K \log n} + c_3 \max\{\sqrt{q(1-q)n}, \sqrt{\log n}\} \\ &\leq C \sqrt{\sigma^2 K \log n + q(1-q)n}, \end{aligned}$$

which completes the proof of the lemma.

Lemma 1.3. *Let M denote the $n \times n$ symmetric matrix such that M_{ij} ($1 \leq i < j \leq n$) are independent random variables with $\mathbb{P}(M_{ij} = 1 - p_{ij}) = p_{ij}$ and $\mathbb{P}(M_{ij} = -p_{ij}) = 1 - p_{ij}$, and $M_{ii} = 0$. Suppose that $\text{Var}[M_{ij}] \leq \sigma^2$ with $\sigma^2 \geq C' \log n/n$ for a constant C' , then with high probability $\|M\| \leq C\sigma\sqrt{n}$ for a constant C .*

Proof. If $\sigma^2 \geq \frac{\log^7 n}{n}$, then Theorem 8.4 in [1] implies that $\|M\| \leq 3\sigma\sqrt{n}$ w.h.p. If $C' \frac{\log n}{n} \leq \sigma^2 \leq \frac{\log^7 n}{n}$, then Lemma 2 in [2] implies that $\|M\| \leq C\sigma\sqrt{n}$ w.h.p. for some universal constant C . \square

1.4 Proof of Theorem 4

We first claim that $K(p-q) \leq c_2 \sqrt{Kp + qn}$ implies $K(p-q) \leq c_2 \sqrt{2qn}$ under the assumption that $K \leq n/2$ and $qn \geq c_1 \log n$. In fact, if $Kp \leq qn$, then the claim trivially holds. If $Kp > qn$, then $q < Kp/n \leq p/2$. It follows that

$$Kp/2 < K(p-q) \leq c_2 \sqrt{Kp + qn} \leq c_2 \sqrt{2Kp}.$$

Thus, $Kp < 8c_2^2$ which contradicts the assumption that $Kp > qn \geq c_1 \log n$. Therefore, $Kp > qn$ cannot hold. Hence, it suffices to show that if $K(p-q) \leq c_2 \sqrt{2qn}$, then Y^* is not an optimal solution of the convex program (6)–(8). We do this by showing that the optimality of Y^* implies $K(p-q) > c_2 \sqrt{2qn}$.

Let \mathbb{I} be the $n \times n$ all-one matrix. Let $\mathcal{R} := \text{support}(Y^*)$ and $\mathcal{A} := \text{support}(A)$. Recall that $Y^* = KUU^\top$ is the singular value decomposition of Y^* , and the orthogonal projection onto the space T is given by $\mathcal{P}_T(M) = UU^\top M + MUU^\top - UU^\top MUU^\top$.

Consider the Lagrangian

$$L(Y; \lambda, \mu, F, G) := -\langle A, Y \rangle + \lambda (\|Y\|_* - \|Y^*\|_*) + \eta (\langle \mathbb{I}, Y \rangle - rK^2) - \langle F, Y \rangle + \langle G, Y - \mathbb{I} \rangle,$$

where $\lambda \geq 0$, $\eta \in \mathbb{R}$, $F_{ij} \geq 0$ and $G_{ij} \leq 0$, $\forall i, j$ are the Lagrangian multipliers. Note that $Y = \frac{rK^2}{n^2} \mathbb{I}$ is strictly feasible so strong duality holds by Slater's Theorem. By standard convex analysis, if $Y = Y^*$ is an optimal solution, then there must exist some F, G and λ for which the following

KKT conditions hold:

$$\begin{aligned}
0 \in \frac{\partial L(Y; \lambda, \mu, F, G)}{\partial Y} \Big|_{Y=Y^*}, & \quad \left. \vphantom{\frac{\partial L(Y; \lambda, \mu, F, G)}{\partial Y}} \right\} \text{Stationary condition} \\
F_{ij} \geq 0, \forall (i, j), & \quad \left. \vphantom{F_{ij}} \right\} \text{Dual feasibility} \\
G_{ij} \geq 0, \forall (i, j), & \\
\lambda \geq 0, & \\
F_{ij} = 0, \forall (i, j) \in \mathcal{R}, & \quad \left. \vphantom{F_{ij}} \right\} \text{Complementary slackness} \\
G_{ij} = 0, \forall (i, j) \in \mathcal{R}^c. &
\end{aligned}$$

Recall that $M \in \mathbb{R}^{n \times n}$ is a subgradient of $\|X\|_*$ at $X = Y^*$ if and only if $\mathcal{P}_T(M) = UU^\top$ and $\|\mathcal{P}_{T^\perp}(M)\| \leq 1$. Let $H = F - G$; the KKT conditions imply that there exist some λ, η, W and H obeying

$$A - \lambda(UU^\top + W) - \eta\mathbb{I} + H = 0, \quad (28)$$

$$\lambda \geq 0, \quad (29)$$

$$P_T W = 0, \quad (30)$$

$$\|W\| \leq 1, \quad (31)$$

$$H_{ij} \leq 0, \forall (i, j) \in \mathcal{R}, \quad (32)$$

$$H_{ij} \geq 0, \forall (i, j) \in \mathcal{R}^c. \quad (33)$$

Now observe that $UU^\top WUU^\top = 0$ by (30). We left and right multiply (28) by UU^\top to obtain

$$\bar{A} - \lambda UU^\top - \eta\mathbb{I} + \bar{H} = 0,$$

where for any matrix $X \in \mathbb{R}^{n \times n}$, $\bar{X} := UU^\top XUU^\top$ is the matrix obtained by taking the average in each $K \times K$ block of X . Consider the last display equation on entries in \mathcal{R} and \mathcal{R}^c respectively. Applying the Bernstein inequality (Theorem A.1) for each entry of \bar{A}_{ij} , we get that with high probability,

$$p - \frac{\lambda}{K} - \eta + \bar{H}_{ij} \geq -\frac{c_3 \sqrt{p(1-p) \log n}}{K} - \frac{c_4 \log n}{2K^2} \stackrel{(a)}{\geq} -\frac{\epsilon_0}{8}, \quad \forall (i, j) \in \mathcal{R} \quad (34)$$

$$q - \eta + \bar{H}_{ij} \leq \frac{c_3 \sqrt{q(1-q) \log n}}{K} + \frac{c_4 \log n}{2K^2} \stackrel{(b)}{\leq} \frac{\epsilon_0}{8}, \quad \forall (i, j) \in \mathcal{R}^c \quad (35)$$

for some constants $c_3, c_4 > 0$, where (a) and (b) follow from the assumption $K \geq c_1 \sqrt{\log n}$ with c_1 sufficiently large. Using (32) and (33), we get

$$q - \frac{\epsilon_0}{8} \leq q - \frac{c_3 \sqrt{q(1-q) \log n}}{K} - \frac{c_4 \log n}{2K^2} \leq \eta \leq p + \frac{c_3 \sqrt{p(1-p) \log n}}{K} + \frac{c_4 \log n}{2K^2} - \frac{\lambda}{K} \leq p + \frac{\epsilon_0}{8} - \frac{\lambda}{K}. \quad (36)$$

It follows that

$$\begin{aligned}
\lambda &\leq K(p - q) + c_3(\sqrt{p(1-p) \log n} + \sqrt{q(1-q) \log n}) + \frac{c_4 \log n}{K} \\
&\leq 4 \max \left\{ K(p - q), c_3 \sqrt{p(1-p) \log n}, c_3 \sqrt{q(1-q) \log n}, \frac{c_4 \sqrt{\log n}}{c_1} \right\}. \quad (37)
\end{aligned}$$

On the other hand, (31), (30) and (28) imply

$$\begin{aligned}\lambda^2 &= \left\| \lambda(UU^\top + W) \right\|^2 \geq \frac{1}{n} \left\| \lambda(UU^\top + W) \right\|_F^2 \\ &= \frac{1}{n} \|A - \eta \mathbb{I} + H\|_F^2 \geq \frac{1}{n} \|A_{\mathcal{R}^c} - \eta \mathbb{I}_{\mathcal{R}^c} + H_{\mathcal{R}^c}\|_F^2 \geq \frac{1}{n} \sum_{(i,j) \in \mathcal{R}^c} (1 - \eta)^2 A_{ij},\end{aligned}$$

where $X_{\mathcal{R}^c}$ denotes that matrix obtained from X by setting the entries outside \mathcal{R}^c to zero. Using $\eta \leq 1 - \frac{7}{8}\epsilon_0$, which is a consequence of (36), (29) and the assumption $p \leq 1 - \epsilon_0$, we obtain

$$\lambda^2 \geq \frac{49}{64n} \epsilon_0^2 \sum_{(i,j) \in \mathcal{R}^c} A_{ij}, \quad (38)$$

Note that $\sum_{(i,j) \in \mathcal{R}^c} A_{ij}$ equals two times the sum of $\binom{n}{2} - r\binom{K}{2}$ i.i.d. Bernoulli random variables with parameter q . By the Chernoff bound of Binomial distributions and the assumption that $qn \geq c_1 \log n$, we know with high probability $\sum_{(i,j) \in \mathcal{R}^c} A_{ij} \geq Cqn^2$ for some constant C . It follows from (38) that $\lambda^2 \geq C'qn$ for some constant $C' > 0$. Combining with (37) and the assumption that $qn \geq c_1 \log n$, we conclude that $K(p - q) \geq C''\sqrt{qn}$ for some constant $C'' > 0$. This completes the proof of the theorem.

1.5 Proof of Theorem 5

The degree d_i of node i is distributed as $\text{Bin}(K - 1, p)$ plus an independent $\text{Bin}(n - K, q)$ if $i \in V_1$. Otherwise d_i is distributed as $\text{Bin}(n - 1, q)$ if $i \in V_2$. It follows that $\mathbb{E}[d_i] = (n - 1)q + (K - 1)(p - q)$ if $i \in V_1$ and $\mathbb{E}[d_i] = (n - 1)q$ if $i \in V_2$. If we define $\sigma^2 = Kp(1 - q) + nq(1 - q)$, then we further have $\text{Var}[d_i] \leq \sigma^2$. Set $t := (K - 1)|p - q|/2 \leq \sigma^2$; the Bernstein inequality (Theorem A.1) gives

$$\mathbb{P}\{|d_i - \mathbb{E}[d_i]| \geq t\} \leq 2 \exp\left(-\frac{t^2}{2\sigma^2 + 2t/3}\right) \leq 2 \exp\left(-\frac{(K - 1)^2(p - q)^2}{12\sigma^2}\right) \leq 2n^{-2},$$

where the last inequality follows from assumption (12). By the union bound, with probability at least $1 - 2n^{-1}$, $d_i > \frac{(p-q)K}{2} + qn$ for all nodes $i \in V_1$ and $d_i < \frac{(p-q)K}{2} + qn$ for all nodes $i \in V_2$. Therefore, all nodes in V_2 are correctly declared to be isolated with high probability.

The number of common neighbors S_{ij} between the nodes i and j is distributed as $\text{Bin}(K - 2, p^2)$ plus an independent $\text{Bin}(n - K, q^2)$ if i and j are in the same cluster, and it is distributed as $\text{Bin}(2(K - 1), pq)$ plus an independent $\text{Bin}(n - 2K, q^2)$ if i and j are in different clusters. Hence, $\mathbb{E}[S_{ij}]$ equals $(K - 2)p^2 + (n - K)q^2$ if i and j are in the same cluster and $2(K - 1)pq + (n - 2K)q^2$ otherwise. The difference in mean equals $K(p - q)^2 - 2p(p - q)$. Let $\sigma^2 = 2Kp^2(1 - q^2) + nq^2(1 - q^2)$. Then $\text{Var}[S_{ij}] \leq \sigma^2$. Set $t' := K(p - q)^2/3 \leq \sigma^2$. Assumption (13) implies that $t' > 2p(p - q)$. Applying the Bernstein inequality (Theorem A.1), we obtain

$$\mathbb{P}\{|S_{ij} - \mathbb{E}[S_{ij}]| \geq t'\} \leq 2 \exp\left(-\frac{t'^2}{2\sigma^2 + 2t'/3}\right) \leq 2 \exp\left(-\frac{K^2(p - q)^4}{27\sigma^2}\right) \leq 2n^{-3},$$

where the last inequality follows from the assumption (13). By union bound, with probability at least $1 - 2n^{-1}$, $S_{ij} > \frac{(p-q)^2K}{3} + 2Kpq + q^2n$ for all nodes i, j from the same cluster and $S_{ij} < \frac{(p-q)^2K}{3} + 2Kpq + q^2n$ for all nodes i, j from two different clusters. Therefore the simple algorithm returns the true clusters w.h.p.

1.6 Proof of Theorem 6

For simplicity we assume K and n_2 are even numbers. We partition V_1 into two equal-sized subsets V_{1+} and V_{1-} such that half of the nodes of each cluster are in V_{1+} . Similarly, V_2 is partitioned into two equal-sized subsets V_{2+} and V_{2-} . To prove the theorem, we need the following anti-concentration inequality.

Theorem 1.4 (Theorem 7.3.1 in [3]). *Let X_1, \dots, X_n be independent random variables such that $0 \leq X_i \leq 1$ for all i . Suppose $X = \sum_{i=1}^n X_i$ and $\sigma^2 := \sum_{i=1}^n \text{Var}[X_i] \geq 200$. Then for all $0 \leq t \leq \sigma^2/100$ and some universal constant $c > 0$, we have*

$$\mathbb{P}[X \geq \mathbb{E}[X] + t] \geq ce^{-t^2/(3\sigma^2)}.$$

Identifying isolated nodes For each node i in $V_{1+} \cup V_{2+}$, let d_{i+} be the number of its neighbors in $V_{1+} \cup V_{2+}$ and d_{i-} be the number of its neighbors in $V_{1-} \cup V_{2-}$, so $d_i = d_{i+} + d_{i-}$. We consider two cases.

Case 1: Suppose $(Kp + (n - K)q) \log n_1 \geq nq \log n_2$. In this case we have $(K - 1)^2(p - q)^2 \leq 2c_2(Kp + nq) \log n_1$ by (14). For each $i \in V_{1+}$, d_{i-} is distributed as $\text{Bin}(K/2, p)$ plus an independent $\text{Bin}((n - K)/2, q)$. Let $t := (K - 1)(p - q) + 2$, $\gamma_1 := \mathbb{E}[d_{i-}] - t = nq/2 + K(p - q)/2 - t$ and $\sigma_d^2 := \text{Var}[d_{i-}] = \frac{1}{2}Kp(1 - p) + \frac{1}{2}(n - K)q(1 - q)$. Since $K \leq n/2$, p, q are bounded away from 1 and $Kp + nq \geq Kp^2 + nq^2 \geq c_1 \log n$, we have $\sigma_d^2 \geq c' \log n \geq 200$. Combining with (14), we further have $\sigma_d^4 \geq c''K^2(p - q)^2/\log n_1 \cdot c' \log n \geq 100t^2$. We can thus apply Theorem 1.4 and get

$$\mathbb{P}[d_{i-} \leq \gamma_1] \geq c \exp\left(-\frac{t^2}{3\sigma_d^2}\right) = \exp\left(-\frac{((K - 1)(p - q) + 2)^2}{3(Kp(1 - p) + (n - K)q(1 - q))}\right) \geq cn_1^{-c'},$$

for some constant $c' > 0$ that can be made small by choosing c_2 above sufficiently small. Let $i^* := \arg \min_{i \in V_{1+}} d_{i-}$. Since the random variables $\{d_{i-} : i \in V_{1+}\}$ are mutually independent, we have

$$\mathbb{P}[d_{i^*-} \geq \gamma_1] = \prod_{i \in V_{1+}} \mathbb{P}[d_{i-} \geq \gamma_1] \leq (1 - cn_1^{-c'})^{n_1/2} \leq \exp\left(-cn_1^{1-c'}/2\right) \leq 1/4.$$

On the other hand, for each $i \in V_{1+}$, d_{i+} is distributed as $\text{Bin}(K/2 - 1, p)$ plus an independent $\text{Bin}((n - K)/2, q)$. Since the median of $\text{Bin}(n, p)$ is at most $np + 1$, we know that with probability at least $1/2$, $d_{i+} \leq \gamma_2 := nq/2 + K(p - q)/2 - p + 2$. Now observe that the two sets of random variables $\{d_{i+}, i \in V_{1+}\}$ and $\{d_{i-}, i \in V_{1+}\}$ are independent of each other, so d_{i+} is independent of i^* for each $i \in V_{1+}$. It follows that

$$\mathbb{P}[d_{i^*+} \leq \gamma_2] = \sum_{i \in V_{1+}} \mathbb{P}[d_{i+} \leq \gamma_2 | i^* = i] \mathbb{P}[i^* = i] = \sum_{i \in V_{1+}} \mathbb{P}[d_{i+} \leq \gamma_2] \mathbb{P}[i^* = i] \geq \frac{1}{2}.$$

Combining the last two display equations by the union bound, we obtain that with probability at least $1/4$,

$$d_{i^*} = d_{i^*-} + d_{i^*+} \leq \gamma_1 + \gamma_2 = (n - 1)q,$$

and thus node i^* will be incorrectly declared as an isolated node.

Case 2: Suppose $(Kp + nq) \log n_1 \leq nq \log n_2$. In this case we have $(K - 1)^2(p - q)^2 \leq 2c_2nq \log n_2$ by assumption. Define $i^* = \arg \max_{i \in V_{2+}} d_{i-}$. Using the same argument as in Case 1, we can show that with probability at least $1/4$, $d_{i^*} \geq (n - 1)q + (K - 1)(p - q)$ and thus node i^* will be incorrectly declared as a non-isolated node.

Recovering clusters For two nodes $i, j \in V_1$, let S_{ij+} be the number of their common neighbors in $V_{1+} \cup V_{2+}$ and S_{ij-} be the number of their common neighbors in $V_{1-} \cup V_{2-}$, so $S_{ij} = S_{ij+} + S_{ij-}$.

For each pair of nodes i, j in V_{1+} that are from the same cluster, S_{ij-} is distributed as $\text{Bin}(K/2, p^2)$ plus an independent $\text{Bin}((n-K)/2, q^2)$. Let $t' := K(p-q)^2 + 4$, $\gamma_3 := \mathbb{E}[S_{ij-}] - t' = nq^2/2 + K(p^2 - q^2)/2 - t'$, and $\sigma_S^2 := \text{Var}[S_{ij-}] = \frac{1}{2}Kp^2(1-p^2) + \frac{1}{2}(n-K)q^2(1-q^2)$. Since $K \leq n/2$, p, q are bounded away from 1 and $Kp^2 + nq^2 \geq c_1 \log n$, we have that $\sigma_S^2 \geq 200$ and $\sigma_S^2 \geq 100t'$. Theorem 1.4 implies that there exists a constant $c > 0$ such that

$$\mathbb{P}[S_{ij-} \leq \gamma_3] \geq c \exp\left(-\frac{t'^2}{3\sigma_S^2}\right) = c \exp\left(-\frac{(K(p-q)^2 + 4)^2}{3(Kp^2(1-p^2) + (n-K)q^2(1-q^2))}\right) \geq cn_1^{-c'},$$

where the constant $c' > 0$ can be made sufficiently small by choosing c_2 sufficiently small in the statement of the lemma. Without loss of generality, we may re-label the nodes such that $V_{1+} = \{1, 2, \dots, n_1/2\}$ and for each $k = 1, \dots, n_1/4$, the nodes $2k-1$ and $2k$ are in the same cluster. Note that the random variables $\{S_{(2k-1)2k-} : k = 1, 2, \dots, n_1/4\}$ are mutually independent. Let $i^* = -1 + 2 \arg \min_{k=1, 2, \dots, n_1/4} S_{(2k-1)2k-}$ and $j^* = i^* + 1$; it follows that

$$\mathbb{P}[S_{i^*j^*-} \geq \gamma_3] \leq (1 - cn_1^{-c'})^{n_1/4} \leq \exp(-cn_1^{1-c'}/4) \leq 1/4.$$

On the other hand, since S_{ij+} is distributed as $\text{Bin}(K/2 - 2, p^2)$ plus an independent $\text{Bin}((n-K)/2, q^2)$, we use the median argument to obtain that with probability at least $1/2$, $S_{ij+} \leq \gamma_4 := nq^2/2 + K(p^2 - q^2)/2 - 2p^2 + 2$. Because $\{S_{ij+}, i, j \in V_{1+}\}$ only depends on the edges between V_{1+} and $V_{1+} \cup V_{2+}$, and (i^*, j^*) only depends on the edges between V_{1+} and $V_{1-} \cup V_{2-}$, we know $\{S_{ij+}, i, j \in V_{1+}\}$ and (i^*, j^*) are independent of each other. It follows that $S_{i^*j^*+} \leq \gamma_4$ with probability at least $1/2$. It follows that with probability at least $1/4$,

$$S_{i^*j^*} = S_{i^*j^*-} + S_{i^*j^*+} \leq \gamma_3 + \gamma_4 = 2(K-1)pq + (n-2K)q^2$$

and thus the nodes i^*, j^* will be incorrectly assigned to two different clusters.

Appendices

A Standard Bernstein Inequality

Theorem A.1. *Let X_1, \dots, X_n be independent random variables such that $|X_i| \leq M$ almost surely. Let $\sigma^2 = \sum_{i=1}^n \text{Var}(X_i)$, then for any $t \geq 0$*

$$\mathbb{P}\left[\sum_{i=1}^n X_i \geq t\right] \leq \exp\left(\frac{-t^2}{2\sigma^2 + \frac{2}{3}Mt}\right).$$

A consequent of the above inequality is $\mathbb{P}\left[\sum_{i=1}^n X_i \geq \sqrt{2\sigma^2 u} + \frac{2Mu}{3}\right] \leq e^{-u}$ for any $u > 0$.

References

- [1] S. Chattergee. Matrix estimation by universal singular value thresholding. *arxiv:1212.1247*, 2012.
- [2] L. Massoulié and D.-C. Tomozei. Distributed user profiling via spectral methods. *arxiv:1109.3318*, 2011.
- [3] J. Matoušek and J. Vondrák. The probabilistic method, lecture notes. <http://kam.mff.cuni.cz/~matousek/prob-ln-2pp.ps.gz>, 2008.
- [4] K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4), 2011.
- [5] J. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.