On the Fundamental Statistical Limit of Community Detection in Random Hypergraphs

Chung-Yi Lin, I (Eli) Chien, and I-Hsiang Wang
Graduate Institute of Communication Engineering and Department of Electrical Engineering,
National Taiwan University, Taipei, Taiwan
Email: {r05942127,b01901044,ihwang}@ntu.edu.tw

***Abstract*—TO BE CONSIDERED FOR THE 2017 IEEE JACK KEIL WOLF ISIT STUDENT PAPER AWARD.**

The problem of community detection in random hyper graphs is considered. We extend the Stochastic Block Model (SBM) from graphs to hypergraphs with d-uniform hyperedges, which we term "d-wise hyper stochastic block model" (d-hSBM), and consider a homogeneous and approximately equal-sized K community case. For $d = 3$, we fully characterize the exponentially decaying rate of the minimax risk in recovering the underlying communities, where the loss function is the mis-match ratio between the true community assignment and the recovered one. It turns out that the rate function is a weighted combination of several divergence terms, each of which is the Rényi divergence of order $\frac{1}{2}$ between two Bernoulli distributions. The Bernoulli distributions involved in the characterization of the rate function are those governing the random instantiation of hyperedges in d-hSBM. The lower bound is set by finding a smaller parameter space where we can analyze the risk, while the upper bound is achieved with the Maximum Likelihood estimator. The technical contribution is to show that upper bound has the same decaying rate as the lower bound, which involves careful bounding of the various probabilities of errors. Finally, we relate the minimax risk to the recovery criterion under the Bayesian framework and derive a threshold condition for exact recovery.

I. INTRODUCTION

The problem of community detection has received great attention recently across many applications, including social science, biology, and computer science. A standard way to investigate this problem is to formulate it in a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where each node $i \in \mathcal{V} = [n] \triangleq \{1, \ldots, n\}$ is assigned a community (label) $\sigma(i) \in [K] \triangleq \{1, \ldots, K\}$. The *community assignment* $\sigma : [n] \to [K]$ is hidden while the graph $\mathcal{G}$ is revealed. Each edge in the graph models *pairwise* interaction between the two nodes. The goal of the community detection problem is to determine σ from $\mathcal{G}$, by leveraging the fact that the nodes with the same community are more likely to have interaction and hence more likely to be connected by an edge.

In a statistical setting, a canonical framework for studying the community detection problem is the *stochastic block model* (SBM) [1] (also known as planted partition model [2]). SBM is a random graph model generating randomly connected edges from a set of labeled nodes. The presence of the $\binom{n}{2}$ edges is governed by $\binom{n}{2}$ Bernoulli random variables, parameter of each of them depends on the community assignments of the two nodes in the corresponding edge. Under the SBM framework, the community detection problem is to reconstruct the hidden labels of all nodes given a single realization of the random graph. The fundamental statistical limits of community detection in SBM have been characterized recently. One line of work takes a Bayesian perspective, where the unknown labeling σ of nodes in $\mathcal{V}$ is assumed to be distributed according to certain prior, and one of the most common assumption is i.i.d. over nodes. Along this line, the fundamental limit for exact recovery is completed characterized [3] in the full generality, while partial recovery remains open. See the survey [4] for more details and references therein. A second line of work takes a minimax perspective, and the goal is to characterize the minimax risk in community detection. In [5], a tight asymptotic characterization of the minimax risk for community detection in SBM is found.

In this work we explore the fundamental statistical limit of leveraging *beyond-pairwise* interactions to recover hidden community structures. In many applications, interaction among a group with more than two entities happens frequently, and it is natural to model such interaction by a *hyperedge* in a hypergraph. It can be expected that the recoverability is enhanced, and we would like to quantify the performance gain in the statistical setting. In particular, we consider a random hypergraph model called *d-wise hyper stochastic block model* (d-hSBM), which can be thought of as a natural extension of SBM to the hypergraph setting. In a d-hSBM, the presence of an order-d hyperedge $\mathsf{e} \subseteq [n]$ (where $|\mathsf{e}| = d$) is governed by a Bernoulli random variable with parameter θ_{e}, and the presence of different hyperedges are mutually independent.

Our main contribution in this conference paper is the characterization of the asymptotic minimax risk for community detection in 3-hSBM. To the best of our knowledge, this is the first piece of work that studies community detection in random hypergraphs. We focus on a homogeneous setting where θ_{e} only depends on the number of communities in e. It turns out that the minimax risk decays to 0 exponentially fast as $n \to \infty$, similar to that in SBM [5], while the exponent is roughly n-times of that in SBM (order-wise). This makes sense because the total number of hyperedges is also roughly n-times of SBM ($\binom{n}{3}$ versus $\binom{n}{2}$). Consequently, to guarantee strong consistency (exact recovery), in 3-hSBM the connection probability θ should at least $\Theta(\frac{\log n}{n^2})$ instead of $\Theta(\frac{\log n}{n})$ as in SBM [3]. For the general d-hSBM, we expect the boost in the exponent will be n^{d-2} times over SBM. Furthermore, let p, q, r denote the connecting probability θ_{e} when the number of communities in the hyperedge e is $1, 2, 3$ respectively. It

turns out that the exponent (normalized by n^2) is a weighted combination of I_{pq} and I_{qr}, where I_{xy} denote the Rényi divergence of order $\frac{1}{2}$ between $\mathrm{Ber}(x)$ and $\mathrm{Ber}(y)$. Hence, the exponent does not depend on the divergence between $\mathrm{Ber}(p)$ and $\mathrm{Ber}(r)$ explicitly. The procedures and techniques in our proof are mainly inspired by [5]. However, in the derivation of some technical lemmas, we will use slightly different methods to make the argument more concise.

II. Problem Formulation

Notations: Let $|S|$ denote the cardinality of the set S, $[N] \triangleq \{1, 2, \ldots, N\}$ for $N \in \mathbb{N}$, and $\bar{t} \triangleq 1 - t$. $\mathcal{S}_n$ is the symmetric group of degree n, which contains all the permutations from $[n]$ to $[n]$. $d_H(\mathbf{x}, \mathbf{y})$ is the Hamming distance between two vectors $\mathbf{x}$ and $\mathbf{y}$. We say that two functions $f(n)$ and $g(n)$ are asymtotically equal, denoted as $f \asymp g$ (as $n \to \infty$), if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$. $\mathcal{R}.\mathcal{V}.$ is to mean "Random Variable" for short.

A. Hypergraph Model

The 3-hSBM is formulated as follows. Consider a 3-uniform hypergraph with n nodes, each belonging to one and only one of the K ($K \geq 2$) communities. Denote σ as the community assignment, and $\sigma(i)$ is the community assignment of the i-th node. We observe the connectivity of the network in an order-3 hyperedge form, which means that the underlying adjacency relation is an $n \times n \times n$ random tensor $A_{i,j,k}$ determined by a set of i.i.d. Bernoulli random variables with success probability $\{\theta_{i,j,k}\}$. We denote $\mathbf{p} = (p, q, r) \in (0, 1)^3$ for the probability parameter. Let $n_k = |\{i \mid \sigma(i) = k\}|$ be the size of the k-th community for $k \in [K]$. The parameter space we consider is a homogeneous and approximately equal-sized case where each $n_k \approx \lfloor \frac{n}{K} \rfloor \triangleq n'$. Formally,

$$\Theta^0(n, K, \mathbf{p}, \eta) \triangleq \Big\{(\sigma, \{\theta_{i,j,k}\}) \mid \sigma : [n] \to [K], \\ n_k \in [(1-\eta)n', (1+\eta)n'] \ \forall k \in [K]\Big\}$$

η is a parameter that determines how much n_k could vary. We assume that $\eta \geq \frac{1}{n'}$ since it is more interesting when the sizes of each community are not exactly equal. As for the success probability $\{\theta_{i,j,k}\}$, we assume that $\theta_{i,j,k} \neq 0$ if and only if $|\{i, j, k\}| = 3$ (no self-loop). Also, $\theta_{i,j,k} = \theta_{\sigma(i),\sigma(j),\sigma(k)}\ \forall \sigma \in \mathcal{S}_3$ (symmetry). Finally, the block struture and homogeneity are characterized by

$$\theta_{i,j,k} = \begin{cases} p, \text{ if } |\{\sigma(i), \sigma(j), \sigma(k)\}| = 1 \text{ (relation } \textit{all-same}) \\ q, \text{ if } |\{\sigma(i), \sigma(j), \sigma(k)\}| = 2 \text{ (relation } \textit{two-one}) \\ r, \text{ if } |\{\sigma(i), \sigma(j), \sigma(k)\}| = 3 \text{ (relation } \textit{all-diff}) \end{cases}$$

Interchangeably, we would write $(i, j, k) \overset{\sigma}{\sim} \alpha$, $(i, j, k) \overset{\sigma}{\sim} \beta$, and $(i, j, k) \overset{\sigma}{\sim} \gamma$ to indicate the *all-same*, *two-one* and *all-diff* relation within nodes i, j, k under the assignment σ.

B. Performance Measure

We use the mis-match ratio to measure the performance of community detection. To tackle the issue of label permutaion, we introduce the concept of equivalence class. For any $\sigma \in \Theta^0$ and $\delta \in \mathcal{S}_K$, let $\delta \circ \sigma$ be a new assignment with $(\delta \circ \sigma)(i) = \delta(\sigma(i))\ \forall i \in [n]$. The equivalence class of σ is defined as $\Gamma(\sigma) = \{\sigma' \mid \exists\, \delta \in \mathcal{S}_K \text{ s.t. } \sigma' = \delta \circ \sigma\}$. Then, for any $\sigma_1, \sigma_2 \in \Theta^0$, we can define the distance between σ_1 and σ_2 as $d(\sigma_1, \sigma_2) \triangleq \inf_{\sigma_1' \in \Gamma(\sigma_1), \sigma_2' \in \Gamma(\sigma_2)} d_H(\sigma_1', \sigma_2')$, the distance between the corresponding equivalence class $\Gamma(\sigma_1)$ and $\Gamma(\sigma_2)$. Note that $d(\sigma_1, \sigma_2)$ is also equivalent to $\inf_{\sigma_2' \in \Gamma(\sigma_2)} d_H(\sigma_1, \sigma_2')$, which is the minimum distance achievable over all relabeling of σ_2 only.

The mis-match ratio is the loss function, counting the proportion of misclassified nodes and minimized over all possible label permutations, defined as $\ell(\sigma, \hat{\sigma}) \triangleq \frac{d(\sigma, \hat{\sigma})}{n}$. Also, we use $\mathrm{R}_\sigma(\hat{\sigma}) \triangleq \mathbb{E}_\sigma \ell(\sigma, \hat{\sigma})$ to denote the corresponding risk function. Therefore, the minimax risk we try to characterize for the parameter space $\Theta^0(n, K, \mathbf{p}, \eta)$ is denoted as

$$\mathrm{R}^* \triangleq \inf_{\hat{\sigma}} \sup_{\sigma \in \Theta^0(n, K, \mathbf{p}, \eta)} \mathrm{R}_\sigma(\hat{\sigma})$$

Remark. Since we want to study the asymtotic behavior of the minimax risk, we view $\mathrm{R}^* = \mathrm{R}^*(n)$ as a function of n (i.e. the number of nodes). In addition, we also couple the parameter K, $\mathbf{p}$ and η with n. The relationship between K and n is stated as the sufficient condition in our main theorem, while the interplay between η and n is that $\eta = \eta_n \downarrow 0$ as $n \to \infty$. $\mathbf{p} = (p, q, r)$ can be a fixed constant or it can scale with n. It it noteworthy that $r < q < p$ is not required in our work though a practical assumption as it may be.

III. Main Results

A. Contributions

The minimax rate for the parameter space $\Theta^0(n, K, \mathbf{p}, \eta)$ under the 3-hSBM is as follows.

Main Theorem: *If*

$$\frac{\frac{(n')^2}{2} I_{pq} + (K-2)(n')^2 I_{qr}}{\log K} \to \infty \tag{1}$$

then

$$\log \mathrm{R}^*(n) \asymp -\Big(\frac{(n')^2}{2} I_{pq} + (K-2)(n')^2 I_{qr}\Big) \tag{2}$$

where the key quantity $I_{xy} \triangleq -2 \log\big(\sqrt{xy} + \sqrt{1-x}\sqrt{1-y}\big)$ is the Rényi divergence of order $\frac{1}{2}$ between two Bernoulli distributions $\mathrm{Ber}(x)$ and $\mathrm{Ber}(y)$.

The lower bound of the minimax risk is depicted as follows.

Theorem 3.1: *If* $\frac{(n')^2}{2} I_{pq} + (K-2)(n')^2 I_{qr} \to \infty$, *then*

$$\mathrm{R}^*(n) \geq \exp\Big(-(1 + o(1))\Big(\frac{(n')^2}{2} I_{pq} + (K-2)(n')^2 I_{qr}\Big)\Big)$$

Proof: The main idea is to construct a smaller parameter space where we can analyze the risk. We pass the global Bayesian risk to a local one and trasform the local Bayesian risk into the risk function of a hypothesis testing problem. With some known results from large deviation, we could get the desired bound. Detailed proofs are given in section IV. ■

Let $L(\sigma; A)$ denote the log-likelihood of an observation A given that the underlying assignment is σ. For upper bound, it can be shown that the *Maximum Likelihood estimator*

$$\hat{\sigma}_{\mathrm{MLE}} = \arg\max_{\sigma} L(\sigma; A)$$

is a rate-optimal procedure.

Theorem 3.2: *If* $\frac{\frac{(n')^2}{2} I_{pq} + (K-2)(n')^2 I_{qr}}{\log K} \to \infty$, *then*

$$\mathrm{R}^*(n) \leq \exp\Big(-(1-o(1))\big(\frac{(n')^2}{2} I_{pq} + (K-2)(n')^2 I_{qr}\big)\Big)$$

Proof: The key step is to get a tight bound on the probability of error event for a fixed candidate assignment σ with $d(\sigma, \sigma_0) = m$, i.e., having m different assignments compared to the truth assignment σ_0. First, we bound $\mathbb{P}\{L(\sigma; A) \geq L(\sigma_0; A)\}$ using the Chernoff Bound. Then, we demonstrate that the number of disagreements are at least in the order of $\frac{(n')^2}{2} I_{pq} + (K-2)(n')^2 I_{qr}$. In the final part, we apply the Union Bound and show that $\mathrm{R}_{\sigma_0}(\hat{\sigma}_{\mathrm{MLE}})$ would be dominated by the R.H.S. of (2) no matter what relative order between $-\big(\frac{(n')^2}{2} I_{pq} + (K-2)(n')^2 I_{qr}\big)$ and $\log n$. Rigorous arguments are provided in section V. ∎

B. Improvements with Hyperedge Observation

For ordinary SBM, the main result of [5] under the homogeneous and approximately equal-sized parameter space $\Theta^0(n, K, p, q)$ is as follows.

$$\text{If } \frac{n' I_{pq}}{\log K} \to \infty, \text{ then } \log \mathrm{R}^*(n) \asymp -n' I_{pq}$$

Compared to (2), we can see that the minimax rate will decay n-times faster than in [5]. An explanation is that, in pairwise interaction we could only make use of $\binom{n}{2}$ random edges, while we have $\binom{n}{3}$ random hyperedges for 3-hSBM. This means that under the same connection probability, fewer nodes are needed to guarantee the same risk in 3-hSBM. Equivalently, given the same number of nodes, using hypergraph allows us to use a connecting probability $\mathbf{p}$ that is n-times smaller while still achieving the same risk.

C. Implications to Exact Recovery

Here, we want to explore the relationship between our minimax result and those under the Bayesian criterion. We will use the definitions in [4]. Besides, we denote π as the prior distribution for σ and $\mathrm{R}_\pi(\hat{\sigma}) \triangleq \mathbb{E}_{\sigma\sim\pi} \mathrm{R}_\sigma(\hat{\sigma})$ as the corresponding Bayesian risk.

First, we look at an interesting observation. In our main theorem, the condition (1) essentially requires that the exponent of the minimax risk R^* should dominate $\log K$. In light of (2), this further leads to $\mathrm{R}^* < \frac{1}{K}$ asymtotically. As noted in [4], if we assume a uniform prior over the appearance of each community (i.e. $\sigma(i) \sim \mathrm{Unif}([K])\ \forall i \in [n]$), then the mis-match ratio is only of interest when it is less than $\frac{1}{K}$. Intuitively, condition (1) restricts us to a more meaningful "better-than-random-guess" Maximum Likelihood estimator in the achievability part of the proof.

Next, we show that how the phase transition phenomenon can be derived from (2). Exact recovery requires

$$\mathbb{E}_{\sigma\sim\pi}\mathbb{P}_\sigma\left\{\sup_{\hat{\sigma}'\in\Gamma(\hat{\sigma})} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\sigma(i) = \hat{\sigma}'(i)\} = 1\right\} = 1 - o(1)$$

For a homogeneous parameter space (*symmetric* SBM in [4]) with K communities embedded, [6] shows that the threshold condition using the Chernoff-Hellinger divergence reduces to

$$(\sqrt{a} - \sqrt{b})^2 > K \tag{3}$$

when a, b, K are constants not scaling with n. In [5], the authors relate the minimax risk R^* in SBM to (3) by rewriting R^* in the form of a threshold variable $\rho \triangleq \liminf_{n\to\infty} \frac{-\log \mathrm{R}^*}{\log n}$. If $\rho > 1$, then there exists a small constant $c > 0$ such that $\frac{-\log \mathrm{R}^*}{\log n} > 1 + c$. We can immediately obtain an upper bound for the error probability.

$$\mathbb{E}_{\sigma\sim\pi}\mathbb{P}_\sigma\left\{\inf_{\hat{\sigma}'\in\Gamma(\hat{\sigma})} \bigcup_{i=1}^{n} \{\sigma(i) \neq \hat{\sigma}(i)\}\right\} \leq n\mathrm{R}_\pi(\hat{\sigma}) \leq n\mathrm{R}^* < n^{-c}$$

When $(p, q) = \frac{\log n}{n}(a, b)$ for connectivity concern, the Rényi divergence I_{pq} is equivalent to $(1+o(1))(\sqrt{a}-\sqrt{b})^2 \frac{\log n}{n}$. Thus, we can see that the condition $\rho > 1$ is identical to (3).

Similarly, we can also identify (3) with $\rho > 1$. Since the connecting probability required (for the same risk) in 3-hSBM is now n-times smaller, we could allow the probability parameter $\mathbf{p}$ to be in the $\Theta(\frac{\log n}{n^2})$ regime. Specifically, let $\mathbf{p} = (p, q, r) = \frac{\log n}{n^2}(a, b, c)$. Based on the above observation and note that $I_{xy} = (1 + o(1))(\sqrt{x} - \sqrt{y})^2 \frac{\log n}{n^2}$ for $x, y = o(\frac{\log n}{n^2})$, we have the following corollary.

Corollary 3.1: *Exact recovery is solvable if*

$$\frac{1}{2K}(\sqrt{a} - \sqrt{b})^2 + \left(1 - \frac{2}{K}\right)(\sqrt{b} - \sqrt{c})^2 > K$$

when a, b, c, K are constants not scaling with n.

Remark. The thershold condition derived from the minimax result provide only a *sufficient* condition for exact recovery in 3-hSBM. However, for SBM, the solvability of exact recovery is fully characterized in [6] by the condition $I > 1$ (or (3) for symmetric cases). We expect that $\rho > 1$ would also be necessary for exact recovery in 3-hSBM.

IV. Minimax Lower Bound

Consider a least favorable sub-parameter space of Θ^0:

$$\Theta_1^L(n, K, \mathbf{p}, \eta) = \Big\{(\sigma, \{\theta_{i,j,k}\}) \in \Theta^0(n, K, \mathbf{p}, \eta) \mid \forall k \in [K] \\ n_k \in \{n' - 1, n', n' + 1\}, n_{\sigma(1)} = n' + 1\Big\}$$

That is, each community only takes on one of the three possible sizes $n' - 1$, n' or $n' + 1$. We further require that the size of the community where the first node lies in is n'. Compared with Θ^0, Θ_1^L is small and specific enough for us to do the lower bound analysis. On the other hand, it is large and general enough to match the desired minimax risk order.

To start with, we introduce the notation of *local loss* which focuses on only one node. Let $S_\sigma(\hat{\sigma}) = \{\sigma' \in \Gamma(\hat{\sigma}) \mid d_H(\sigma,\sigma') = d(\sigma,\hat{\sigma})\}$ be the set of all permutations in the equivalence class of $\hat{\sigma}$ that achieve the mimimum distance. The local loss function for each $i \in [n]$ is defined as the proportion of false labeling of node i in $S_\sigma(\hat{\sigma})$: $\ell(\sigma(i),\hat{\sigma}(i)) \triangleq \sum_{\sigma' \in S_\sigma(\hat{\sigma})} \frac{\mathbb{1}\{\sigma(i)\neq\hat{\sigma}(i)\}}{|S_\sigma(\hat{\sigma})|}$. It turns out that it is rather easy to study the local loss. Since Θ_1^L is closed under permutation, we can apply the *global-to-local* lemma in [5].

Lemma 4.1 (*Lemma 2.1* in [5])**:** *Let Θ be any homogeneous parameter space that is closed under permutation. Let* Unif *be the uniform prior over all the elements in Θ. Define the global Bayesian risk as $\mathrm{R}_{\sigma\sim\mathrm{Unif}}(\hat{\sigma}) = \frac{1}{|\Theta|}\sum_{\sigma\in\Theta}\mathbb{E}_\sigma\ell(\sigma,\hat{\sigma})$ and the local Bayesian risk $\mathrm{R}_{\sigma\sim\mathrm{Unif}}(\hat{\sigma}(1)) = \frac{1}{|\Theta|}\sum_{\sigma\in\Theta}\mathbb{E}_\sigma\ell(\sigma(1),\hat{\sigma}(1))$ for the first node. Then*

$$\inf_{\hat{\sigma}} \mathrm{R}_{\sigma\sim\mathrm{Unif}}(\hat{\sigma}) = \inf_{\hat{\sigma}} \mathrm{R}_{\sigma\sim\mathrm{Unif}}(\hat{\sigma}(1))$$

Second, we relate the local Bayesian risk to a hypothesis testing problem. The two competing assignments disagreed on only one node directly translate into the hypothesis candidates in the testing problem.

Lemma 4.2:

$$\mathrm{R}_{\sigma\sim\mathrm{Unif}}(\hat{\sigma}(1)) \geq \mathbb{P}\left\{\sum_{i=1}^{n_{pq}} C_{pq}(X_i - Y_i) + \sum_{i=1}^{n_{qr}} C_{qr}(Z_i - W_i) \geq 0\right\}$$

where $n_{pq} \triangleq \frac{1}{2}n'(n'-1), n_{qr} \triangleq n'(n-2n'-1)$, $f(t) \triangleq t/\bar{t}$ for $C_{xy} \triangleq \log\frac{f(x)}{f(y)}$, and $X_i \overset{i.i.d.}{\sim} \mathrm{Ber}(q), Y_i \overset{i.i.d.}{\sim} \mathrm{Ber}(p), Z_i \overset{i.i.d.}{\sim} \mathrm{Ber}(r), W_i \overset{i.i.d.}{\sim} \mathrm{Ber}(q)$ are all mutually indepedent $\mathcal{R}.\mathcal{V}.s.$

With the aid of the *Rozovsky* lower bound [7], we are able to prove the following lemma.

Lemma 4.3: *If $n_{pq}I_{pq} + n_{qr}I_{qr} \to \infty$, then*

$$\mathbb{P}\left\{\sum_{i=1}^{n_{pq}} C_{pq}(X_i - Y_i) + \sum_{i=1}^{n_{qr}} C_{qr}(Z_i - W_i) \geq 0\right\} \geq \exp\big(-(1+o(1))(n_{pq}I_{pq} + n_{qr}I_{qr})\big) \quad (4)$$

Lemma 4.2 and 4.3 are proved in Appendix A in [8].

Proof of Theorem 3.1: Finally, since the Bayesian risk always lower bounds the minimax risk, we have

$$\begin{aligned}
\mathrm{R}^*(n) &= \inf_{\hat{\sigma}} \sup_{\sigma\in\Theta^0} \mathbb{E}_\sigma\ell(\sigma,\hat{\sigma}) \geq \inf_{\hat{\sigma}} \sup_{\sigma\in\Theta_1^L} \mathbb{E}_\sigma\ell(\sigma,\hat{\sigma}) \\
&\geq \inf_{\hat{\sigma}} \mathrm{R}_{\sigma\sim\mathrm{Unif}}(\hat{\sigma}) = \inf_{\hat{\sigma}} \mathrm{R}_{\sigma\sim\mathrm{Unif}}(\hat{\sigma}(1)) \\
&\geq \mathbb{P}\left\{\sum_{i=1}^{n_{pq}} C_{pq}(X_i - Y_i) + \sum_{i=1}^{n_{qr}} C_{qr}(Z_i - W_i) \geq 0\right\} \\
&\geq \exp\big(-(1+o(1))(n_{pq}I_{pq} + n_{qr}I_{qr})\big) \\
&\geq \exp\Big(-(1+o(1))\big(\frac{(n')^2}{2}I_{pq} + (K-2)(n')^2 I_{qr}\big)\Big)
\end{aligned}$$

■

V. Minimax Upper Bound

In this section, we use σ_0 to denote the true assignment while σ is any other competitor in ML estimation. Toward our goal, we will find a tight bound on the following probability

$$P_m \triangleq \mathbb{P}\left\{\exists\,\sigma\in\Theta^0 \mid d(\sigma,\sigma_0) = m,\, L(\sigma;A) \geq L(\sigma_0;A)\right\} \quad (5)$$

for $m \in [n]$. First of all, we derive the expression for the log-likelihood function. To simplify, denote $S_\alpha \triangleq \{(i,j,k) \mid i<j<k,\, (i,j,k) \overset{\sigma}{\sim} \alpha\}$. S_β and S_γ are defined in the same way. With some manipulations, we have

Claim 5.1:

$$\begin{aligned}
L(\sigma;A) = &\left(C_{pq}\sum_{(i,j,k)\in S_\alpha} A_{i,j,k} - \lambda_{pq}\sum_{(i,j,k)\in S_\alpha} 1\right) \\
&- \left(C_{qr}\sum_{(i,j,k)\in S_\gamma} A_{i,j,k} - \lambda_{qr}\sum_{(i,j,k)\in S_\gamma} 1\right) + f(\sigma^c)
\end{aligned}$$

where $f(\sigma^c)$ are those terms that are invariant to the choice of the community assignment σ and $\lambda_{xy} \triangleq \log(\bar{x}/\bar{y})$.

Similarly, we can obtain the expression $L(\sigma_0;A)$. Then,

Claim 5.2: *Event $\{L(\sigma;A) \geq L(\sigma_0;A)\}$ is equivalent to*

$$\begin{aligned}
&C_{pq}\Big(\sum_{i=1}^{n(\alpha,\beta_0)} V_i - \sum_{i=1}^{n(\beta,\alpha_0)} U_i\Big) \\
&+C_{qr}\Big(\sum_{i=1}^{n(\beta,\gamma_0)} W_i - \sum_{i=1}^{n(\beta,\beta_0)} V_i\Big) + C_{pr}\Big(\sum_{i=1}^{n(\alpha,\gamma_0)} W_i - \sum_{i=1}^{n(\gamma,\alpha_0)} U_i\Big) \\
&\geq C_{pq}\lambda_{pq}\big(n(\alpha,\beta_0) - n(\beta,\alpha_0)\big) + C_{qr}\lambda_{qr}\big(n(\beta,\gamma_0) - n(\gamma,\beta_0)\big) \\
&\qquad + C_{pr}\lambda_{pr}\big(n(\alpha,\gamma_0) - n(\gamma,\alpha_0)\big)
\end{aligned}$$

where $S_{\alpha,\beta_0} \triangleq \{(i,j,k) \mid (i,j,k) \overset{\sigma}{\sim} \alpha, (i,j,k) \overset{\sigma_0}{\sim} \beta\}$ for $n(\alpha,\beta_0) \triangleq |S_{\alpha,\beta_0}|$ (other related notations are similarly defined) and $U_i \overset{i.i.d.}{\sim} \mathrm{Ber}(p), V_i \overset{i.i.d.}{\sim} \mathrm{Ber}(q), W_i \overset{i.i.d.}{\sim} \mathrm{Ber}(r)$ are all mutually independent $\mathcal{R}.\mathcal{V}.s.$

The following diagram may help clarify our definitions:

	all-same	*two-one*	*all-diff*
σ	α	β	γ
σ_0	α_0	β_0	γ_0

Note that $n(\alpha,\beta_0)$ and $n(\beta,\alpha_0)$ have the same struture and order by symmetry. We use $n(\alpha,\beta) \triangleq \frac{1}{2}\big(n(\alpha,\beta_0)+n(\beta,\alpha_0)\big)$ to denote their average. $n(\beta,\gamma)$, $n(\alpha,\gamma)$ are similarly defined.

The next step is to use Chernoff Bound to put a limit on the probability of error event $\mathsf{P_e}(\sigma) \triangleq \mathbb{P}\{L(\sigma;A) \geq L(\sigma_0;A)\}$.

Lemma 5.1:

$$\mathsf{P_e}(\sigma) \leq \exp\Big(-\big(n(\alpha,\beta)I_{pq} + n(\beta,\gamma)I_{qr}\big)\Big)$$

Third, we apply some knowledge from convex analysis (see, for example, [9]) to obtain the combinatorial bounds on $n(\alpha,\beta)$ and $n(\beta,\gamma)$.

Lemma 5.2: *Let σ be an assignment with $d(\sigma,\sigma_0)=m$ where $m\in[n]$. Then there exists a constant c independent of n such that $\eta'=c\eta$ and the following holds $\forall K\geq 2$: for $m\leq \frac{n'}{2}$,*

$$n(\alpha,\beta)\geq(1-\eta')\Big(1-\frac{\frac{m}{1-\eta'}}{n'}\Big)\Big(1-\frac{\frac{1}{1-\eta'}}{n'}\Big)\frac{(n')^2}{2}m$$
$$n(\beta,\gamma)\geq(1-\eta')\Big(1-\frac{\frac{m}{1-\eta'}}{n'}\Big)^2(K-2)(n')^2m$$

and for $m>\frac{n'}{2}$,

$$n(\alpha,\beta)\geq\frac{(1-\eta')}{(K')^2}\Big(1-\frac{\frac{(K')^2}{1-\eta'}}{n'}\Big)\frac{(n')^2}{2}m$$
$$n(\beta,\gamma)\geq\frac{7(1-\eta')}{81}(K-2)(n')^2m$$

where $K'=\max\{3,K\}$ in case of $K=2$.

Lemma 5.2, together with Lemma 5.1, immediately imply an upper bound on $\mathsf{P_e}(\sigma)$ for each given σ.

Lemma 5.3: *Let $\sigma\in\Theta^0$ be an arbitrary assignment with $d(\sigma,\sigma_0)=m$. We have the following.*

$$\mathbb{P}\big\{L(\sigma;A)\geq L(\sigma_0;A)\big\}$$
$$\leq\begin{cases}\exp\Big(-(1-\eta')\big(1-\frac{\frac{m}{1-\eta'}}{n'}\big)^2 \\ \big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)m\Big),\ \text{if } m\leq\frac{n'}{2}\\ \exp\Big(-(1-\eta')\big(\frac{1}{(K')^2}\big(1-\frac{\frac{(K')^2}{1-\eta'}}{n'}\big)\frac{(n')^2}{2}I_{pq}\\ \quad+\frac{7}{81}(K-2)(n')^2I_{qr}\big)m\Big),\ \text{if } m>\frac{n'}{2}\end{cases}$$

Then, we apply Union Bound on P_m defined in (5). We use the cardinality of $\{\Gamma(\sigma)\}$ (the equivalence class of σ) rather than the quite larger counterpart $\{\sigma\in\Theta^0\mid d(\sigma,\sigma_0)=m\}$.

Lemma 5.4 (*Proposition 5.2* in [5])**:** *The cardinality of equivalent class that has distance m from σ_0 is upper bounded by*

$$\big|\{\Gamma\mid\exists\sigma\in\Gamma\ s.t.\ d(\sigma,\sigma_0)=m\}\big|\leq\min\Big\{\big(\frac{enK}{m}\big)^m,K^n\Big\}$$

In light of Lemma 5.4, we further have $P_m\leq\big|\{\Gamma\mid\exists\sigma\in\Gamma\text{ s.t. }d(\sigma,\sigma_0)=m\}\big|\cdot\mathbb{P}\{L(\sigma;A)\geq L(\sigma_0;A)\}$. Let $\eta'=\eta'_n\downarrow 0$ in Lemma 5.2 as $n\to\infty$ and recall the requirement (1) in our main theorem. We can make η' decay slow enough and dominate $\log K$ and other lower(constant) order terms by $\eta'\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)$. Considering the order between $\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}$ and $\log n$, we are going to find different m_0s and m's so that

Lemma 5.5: *For the following three different cases:*

1) $\liminf\limits_{n\to\infty}\frac{\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}}{\log n}>1$
2) $\limsup\limits_{n\to\infty}\frac{\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}}{\log n}<1$
3) $\limsup\limits_{n\to\infty}\frac{\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}}{\log n}=1+o(1)$

We have $\frac{m_0}{n}\leq\exp\big(-(1-o(1))(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr})\big)\triangleq R$. Besides, $\frac{m}{n}P_m\leq Q_{m1}\ \forall m\in(m_0,m']$ where $Q_{m_0}=o(R)$ and $\frac{m}{n}P_m\leq Q_{m2}\ \forall m\in(m',n]$ where $Q_{m'}=o(R)$ for some fast decaying geometric series $\{Q_{m1}\}$ and $\{Q_{m2}\}$.

Proof of Theorem 3.1: Combining, we have $\forall\sigma_0\in\Theta^0$,

$$\begin{aligned}\mathrm{R}_{\sigma_0}(\hat\sigma_{\mathrm{MLE}})=\mathbb{E}_{\sigma_0}\ell(\sigma_0,\hat\sigma_{\mathrm{MLE}})&\leq\sum_{m=1}^n\frac{m}{n}P_m\\&=\sum_{m=1}^{m_0}\frac{m}{n}P_m+\sum_{m>m_0}^{m'}\frac{m}{n}P_m+\sum_{m>m'}^{n}\frac{m}{n}P_m\\&\leq\frac{m_0}{n}+c_1Q_{m_0}+c_2Q_{m'}=o(R)\end{aligned}$$

Essentially, we can conclude that $\mathrm{R}^*(n)=o(R)$. ∎

Proofs of Claim 5.1, 5.2, Lemma 5.1, 5.2 and 5.5 are rather involved. We defer the details to Appendix B in [8].

VI. Extensions

There are still a lot of interesting issues awaiting to be explored. We list a few possible directions for future work.

- *Extension to general parameter space:* In [5], the authors consider a general parameter space Θ beyond the homogeneous and approximately equal-sized special case. We hope to finish this generalization in 3-`hSBM` soonly after.
- *Efficient algorithm:* After characterizing the statistical limit of community detection in 3-`hSBM`, we are now constructing an efficient algorithm that achieves it. Inspired by [10], we develop a 2-step algorithm which seems to achieve the statistical limit found in this paper.
- *Extension to higher order hypergraph:* `hSBM` of higher order would definitely be an interest of research for years to come. We predict that the main difficulty for higher order extension would still lie in the combinatorial lower bound (Lemma 5.2). Moreover, we could even consider `hSBM` with *mixing* order of hyperedges. The mixture model could possibly provides more insight to the community detection problem in `hSBM`.

References

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

[2] A. Condon and R. M. Karp, "Algorithms for graph partitioning on the planted partition model," *Random Structures and Algorithms*, vol. 18, no. 2, pp. 116–140, March 2001.

[3] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 471–487, 2016.

[4] E. Abbe, "Community detection and the stochastic block model: recent developments," 2016.

[5] A. Y. Zhang and H. H. Zhou, "Minimax rates of community detection in stochastic block models," *Ann. Statist.*, vol. 44, no. 5, pp. 2252–2280, 10 2016.
http://dx.doi.org/10.1214/15-AOS1428

[6] E. Abbe and C. Sandon, "Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery," *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pp. 670–688, Oct 2015.

[7] L. Rozovsky, "A lower bound of large-deviation probabilities for the sample mean under the cramér condition," *Journal of Mathematical Sciences*, vol. 118, no. 6, pp. 5624–5634, 2003.

[8] C.-Y. Lin, I. E. Chien, and I.-H. Wang, "On the fundamental statistical limit of community detection in random hypergraphs," 2017.
http://homepage.ntu.edu.tw/%7Eihwang/Research/Eprints/isit17cd.pdf

[9] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1997.

[10] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou, "Achieving optimal misclassification proportion in stochastic block model," *arXiv preprint arXiv:1505.03772*, 2015.

APPENDIX A
PROOF OF AUXILIARY RESULTS IN LOWER BOUND

A. Proof of Lemma 4.2

First recall that

$$\mathrm{R}_{\sigma\sim\mathrm{Unif}}(\hat{\sigma}(1)) = \frac{1}{|\Theta_1^L|}\sum_{\sigma\in\Theta_1^L}\mathbb{E}_\sigma \ell(\sigma(1),\hat{\sigma}(1))$$

In order to connect $\mathrm{R}_{\sigma\sim\mathrm{Unif}}(\hat{\sigma}(1))$ with the risk function of a hypothesis testing problem, we shall find an equivalent form of $\mathbb{E}_\sigma\ell(\sigma(1),\hat{\sigma}(1))$. The idea is to find another assignment σ' such that $d(\sigma,\sigma') = d_H(\sigma(1),\sigma'(1)) = \mathbb{1}\{\sigma(1)\neq\sigma'(1)\} = 1$. σ' is the most indistinguishable opponent against σ in the sense that their assignments differ by only one node. Specifically, for each $\sigma_0\in\Theta_1^L$, we construct a new assignment $\sigma[\sigma_0]$ based on σ_0:

$$\sigma[\sigma_0](1) = \operatorname*{arg\,min}_{2\le i\le n}\left\{n_{\sigma_0(i)} = n'\right\}$$

and $\sigma[\sigma_0](i) = \sigma_0(i)$ for $2\le i\le n$. Note that $\{i \mid n_{\sigma_0(i)} = n'\}\neq\varnothing\,\forall\sigma_0\in\Theta_1^L$ and $\sigma[\sigma_0]\in\Theta_1^L$. In addition, for any $\sigma_1,\,\sigma_2\in\Theta_1^L$, we can see that $\sigma_1\neq\sigma_2$ if and only if $\sigma[\sigma_1]\neq\sigma[\sigma_2]$. Therefore, $\{\sigma_0\mid\sigma_0\in\Theta_1^L\} = \{\sigma_0\mid\sigma_0\in\Theta_1^L\}$ and thus

$$\begin{aligned}\mathrm{R}_{\sigma\sim\mathrm{Unif}}(\hat{\sigma}(1)) &= \frac{1}{|\Theta_1^L|}\sum_{\sigma_0\in\Theta_1^L}\mathbb{E}_{\sigma_0}\ell(\sigma_0(1),\hat{\sigma}(1))\\ &= \frac{1}{|\Theta_1^L|}\sum_{\sigma_0\in\Theta_1^L}\frac{1}{2}\big(\mathbb{E}_{\sigma_0}\ell(\sigma_0(1),\hat{\sigma}(1)) + \mathbb{E}_{\sigma[\sigma_0]}\ell(\sigma[\sigma_0](1),\hat{\sigma}(1))\big)\end{aligned}$$

In the testing problem, we can use the optimal Bayes risk as a lower bound. Let $\hat{\sigma}_{\mathrm{Bayes}}$ be an assignment that achieves the minimum Bayes risk $\inf_{\hat{\sigma}}\frac{1}{2}\big(\mathbb{E}_{\sigma_0}\ell(\sigma_0(1),\hat{\sigma}(1)) + \mathbb{E}_{\sigma[\sigma_0]}\ell(\sigma[\sigma_0](1),\hat{\sigma}(1))\big)$. Notice that $\hat{\sigma}_{\mathrm{Bayes}}(1)$ is a Bayes estimator concerning the 0-1 loss, indicating that $\hat{\sigma}_{\mathrm{Bayes}}(1)$ must to be the mode of the posterior distribution. Roughly speaking, the team who has a larger value of sum of the supporting $\mathcal{R}.\mathcal{V}$.s wins the test. Denote $J_0 = \big\{u\in[n]/\{1\}\mid\sigma_0(u) = \sigma_0(1)\big\}$, $J_1 = \big\{u\in[n]\mid\sigma_0(u) = \sigma[\sigma_0](1)\big\}$ and $J_2 = \big\{u\in[n]\mid u\notin J_0\cup J_1\big\}$. If the underlying assignment is σ_0, the set of supporting random variables is divided into

$$\begin{aligned}A_{1,j,k}\sim\mathrm{Ber}(p) &: S_1 = \{(j,k)\mid j<k;\,j,k\in J_0\}\\ A_{1,j,k}\sim\mathrm{Ber}(q) &: S_2 = \{(j,k)\mid j\in J_0,k\in J_1\},\ S_3 = \{(j,k)\mid j\in J_0,k\in J_2\},\\ &\quad\ S_4 = \{(j,k)\mid j<k;\,j,k\in J_1\},\ S_5 = \{(j,k)\mid j<k;\,j,k\in J_2,\sigma(j) = \sigma(k)\}\\ A_{1,j,k}\sim\mathrm{Ber}(r) &: S_6 = \{(j,k)\mid j\in J_1,k\in J_2\},\ S_7 = \{(j,k)\mid j<k;\,j,k\in J_2,\sigma(j)\neq\sigma(k)\}\end{aligned}$$

Otherwise, if $\sigma = \sigma[\sigma_0]$, the situation becomes

$$\begin{aligned}A_{1,j,k}\sim\mathrm{Ber}(p) &: \{(j,k)\mid j<k;\,j,k\in J_1\} = S_4\\ A_{1,j,k}\sim\mathrm{Ber}(q) &: \{(j,k)\mid j\in J_1,k\in J_0\} = S_2',\ \{(j,k)\mid j\in J_1,k\in J_2\} = S_6,\\ &\quad\ \{(j,k)\mid j<k;\,j,k\in J_0\} = S_4,\ \{(j,k)\mid j<k;\,j,k\in J_2,\sigma(j) = \sigma(k)\} = S_5\\ A_{1,j,k}\sim\mathrm{Ber}(r) &: \{(j,k)\mid j\in J_0,k\in J_2\} = S_6,\ \{(j,k)\mid j<k;\,j,k\in J_2,\sigma(j)\neq\sigma(k)\} = S_7\end{aligned}$$

First of all, observe that $|S_1| = |S_4| = \frac{1}{2}n'(n'-1) = n_{pq}$ and $|S_3| = |S_6| = n'(n-2n'-1) = n_{qr}$. Moreover, $S_2 = S_2'$ because of the symmetry of $A_{i,j,k}$. Third, S_5 and S_7 are independent of the hypothesis testing (denoted as σ^c).

As a result, the posterior probability is

$$\mathbb{P}\{A\mid\sigma_0\} = \prod_{(j,k)\in S_1}p^{A_{1,j,k}}\bar{p}^{\bar{A}_{1,j,k}}\prod_{(j,k)\in S_3}q^{A_{1,j,k}}\bar{q}^{\bar{A}_{1,j,k}}\prod_{(j,k)\in S_4}q^{A_{1,j,k}}\bar{q}^{\bar{A}_{1,j,k}}\prod_{(j,k)\in S_6}r^{A_{1,j,k}}\bar{r}^{\bar{A}_{1,j,k}}\cdot f(\sigma^c)$$

and the log-likelihood function becomes

$$\begin{aligned}L(\sigma_0;A) = \log\mathbb{P}\{A\mid\sigma_0\} &= \sum_{(j,k)\in S_1}\left(A_{1,j,k}\log\frac{p}{\bar{p}} + \log\bar{p}\right) + \sum_{(j,k)\in S_3}\left(A_{1,j,k}\log\frac{q}{\bar{q}} + \log\bar{q}\right)\\ &\quad + \sum_{(j,k)\in S_4}\left(A_{1,j,k}\log\frac{q}{\bar{q}} + \log\bar{q}\right) + \sum_{(j,k)\in S_6}\left(A_{1,j,k}\log\frac{r}{\bar{r}} + \log\bar{r}\right) + \log f(\sigma^c)\end{aligned}$$

Similarly, by interchanging the role of J_0 and J_1, we can obtain $L(\sigma[\sigma_0]; A)$. Hence,

$$\begin{aligned}\mathbb{E}_{\sigma_0}\ell(\sigma_0(1), \hat{\sigma}_{\text{Bayes}}(1)) &= \mathbb{P}_{\sigma_0}\{L(\sigma[\sigma_0]; A) \geq L(\sigma_0; A)\} \\ &= \mathbb{P}\left\{\sum_{i=1}^{n_{pq}} C_{pq}(X_i - Y_i) + \sum_{i=1}^{n_{qr}} C_{qr}(Z_i - W_i) \geq 0\right\}\end{aligned} \tag{6}$$

after some manipulations. By symmetry, the situation is exactly the same for $\mathbb{E}_{\sigma[\sigma_0]}\ell(\sigma[\sigma_0](1), \hat{\sigma}_{\text{Bayes}}(1))$. Finally, since (6) holds for all $\sigma_0 \in \Theta_1^L$ and $\inf(\cdot)$ is a concave function, we have

$$\begin{aligned}\mathrm{R}_{\sigma\sim\text{Unif}}(\hat{\sigma}(1)) &\geq \inf_{\hat{\sigma}} \mathrm{R}_{\sigma\sim\text{Unif}}(\hat{\sigma}(1)) \\ &= \inf_{\hat{\sigma}} \frac{1}{|\Theta_1^L|} \sum_{\sigma_0\in\Theta_1^L} \frac{1}{2}\big(\mathbb{E}_{\sigma_0}\ell(\sigma_0(1), \hat{\sigma}(1)) + \mathbb{E}_{\sigma[\sigma_0]}\ell(\sigma[\sigma_0](1), \hat{\sigma}(1))\big) \\ &\geq \frac{1}{|\Theta_1^L|} \sum_{\sigma_0\in\Theta_1^L} \inf_{\hat{\sigma}} \frac{1}{2}\big(\mathbb{E}_{\sigma_0}\ell(\sigma_0(1), \hat{\sigma}(1)) + \mathbb{E}_{\sigma[\sigma_0]}\ell(\sigma[\sigma_0](1), \hat{\sigma}(1))\big) \\ &= \frac{1}{|\Theta_1^L|} \sum_{\sigma_0\in\Theta_1^L} \mathbb{P}\left\{\sum_{i=1}^{n_{pq}} C_{pq}(X_i - Y_i) + \sum_{i=1}^{n_{qr}} C_{qr}(Z_i - W_i) \geq 0\right\} \\ &= \mathbb{P}\left\{\sum_{i=1}^{n_{pq}} C_{pq}(X_i - Y_i) + \sum_{i=1}^{n_{qr}} C_{qr}(Z_i - W_i) \geq 0\right\}\end{aligned}$$

B. Proof of Lemma 4.3

We can break the L.H.S of (4) dirctly into

$$\mathbb{P}\left\{\sum_{i=1}^{n_{pq}} C_{pq}(X_i - Y_i) + \sum_{i=1}^{n_{qr}} C_{qr}(Z_i - W_i) \geq 0\right\} \geq \mathbb{P}\left\{\sum_{i=1}^{n_{pq}} C_{pq}(X_i - Y_i) \geq 0\right\} \cdot \mathbb{P}\left\{\sum_{i=1}^{n_{qr}} C_{qr}(Z_i - W_i) \geq 0\right\}$$

Though naïve, we could still arrive at the same order as the minimax rate. By symmetry, it is enough to focus on

$$\mathbb{P}\left\{\sum_{i=1}^{n_{pq}} C_{pq}(X_i - Y_i) \geq 0\right\}$$

in (4). Here, we utilize a result from large deviation.

Conseder i.i.d. $\mathcal{R}.\mathcal{V}.$s $\{X_i\}_{i=1}^n$ where each $X_i \sim X$. We assume X is nondegenerate and that

$$\mathbb{E}X^2 e^{\lambda X} < \infty \tag{7}$$

for some $\lambda > 0$. The former condition ensures, for $0 < u \leq \lambda$, the existence of the functions $m(u) \triangleq \big(\log L_X(u)\big)'$, $\sigma^2(u) \triangleq m'(u)$ and $Q(u) \triangleq um(u) - \log L_X(u)$ where $L_X(u) \triangleq \mathbb{E}e^{uX}$ is the *Moment Generating Function* (MGF) of the random variable X. Recall some known results:

$$\lim_{u\downarrow 0} m(u) = m(0) = \mathbb{E}X < \infty$$

and

$$\sup_{0<u\leq\lambda} (ux - \log L_X(u)) = Q(u^*) \tag{8}$$

for $m(0) < x \leq m(\lambda)$, where u^* is the unique solution of the equation

$$m(u) = x \tag{9}$$

Note that it is the sup-achieving condition in (8). The main theorem goes as follows.

Theorem 1.1 (*Theorem 1* in [7])**:** *$\forall x$ such that $m(0) < x \leq m(\lambda)$ and $\forall n \geq 1$, the relation*

$$e^{-nQ(u^*)} \geq \mathbb{P}\left\{\sum_{i=1}^{n} X_i \geq nx\right\} \geq e^{-nQ(u^*) - c\left(1+\sqrt{nQ(u^*)}\right)}$$

holds, where the constant c does not depend on x and n.

The first inequality is essentially the Chernoff Bound, while here we use the second one, i.e., the lower bound result.

First, we identify that $X = C_{pq}(X_i - Y_i)$ and $n = n_{pq}$ for our problem. Besides, since $X < \infty$, we can take λ large enough so that (7) holds. The MGF now becomes

$$L_X(u) = \mathbb{E}e^{uX} = \mathbb{E}\big[e^{uC_{pq}X_i}\big] \cdot \mathbb{E}\big[e^{-uC_{pq}Y_i}\big]$$

Also, since $m(0) = \mathbb{E}X < 0$, we make a trick here to take $x = 0$. The corresponding optimalilty condition (9) becomes

$$\begin{aligned} m(u) = x = 0 &\Leftrightarrow \frac{L'_X(u)}{L_X(u)} = 0 \\ &\Leftrightarrow L'_X(u) = 0 \end{aligned}$$

It can be shown that $u^* = \frac{1}{2}$ and the supremum achieved is

$$\begin{aligned} Q(u^*) &= \sup_{0<u\leq\lambda} (ux - \log L_X(u)) \\ &= -\log L_X(u^*) \\ &= I_{pq} \end{aligned}$$

Combining the expressions for C_{pq} and C_{qr}, we can conclude that

$$\begin{aligned} \mathbb{P}\left\{\sum_{i=1}^{n_{pq}} C_{pq}(X_i - Y_i) + \sum_{i=1}^{n_{qr}} C_{qr}(Z_i - W_i) \geq 0\right\} &\geq \mathbb{P}\left\{\sum_{i=1}^{n_{pq}} C_{pq}(X_i - Y_i) \geq 0\right\} \cdot \mathbb{P}\left\{\sum_{i=1}^{n_{qr}} C_{qr}(Z_i - W_i) \geq 0\right\} \\ &\geq e^{-n_{pq}I_{pq} - c_{pq}\left(1+\sqrt{n_{pq}I_{pq}}\right)} \cdot e^{-n_{qr}I_{qr} - c_{qr}\left(1+\sqrt{n_{qr}I_{qr}}\right)} \\ &= exp\left(-\Big(n_{pq}I_{pq} + n_{qr}I_{qr} + c_{pq}\big(1 + \sqrt{n_{pq}I_{pq}}\big) + c_{qr}\big(1 + \sqrt{n_{qr}I_{qr}}\big)\Big)\right) \\ &\geq exp\left(-\Big(n_{pq}I_{pq} + n_{qr}I_{qr} + c\big(2 + \sqrt{2(n_{pq}I_{pq} + n_{qr}I_{qr})}\big)\Big)\right) \end{aligned}$$

where $c = \max\{c_{pq}, c_{qr}\}$ is independent of n'. Finally, since we assume that $n_{pq}I_{pq} + n_{qr}I_{qr} \to \infty$, the second term with the constant c in the above equation would be dominated by the first term. We have the desired asymtotic result in consequence.

APPENDIX B
PROOF OF AUXILIARY RESULTS IN UPPER BOUND

A. *Proof of Claim 5.1*

$$\begin{aligned} L(\sigma; A) &= \log p \sum_{S_\alpha} A_{i,j,k} + \log \bar{p} \sum_{S_\alpha} \bar{A}_{i,j,k} + \log q \sum_{S_\beta} A_{i,j,k} + \log \bar{q} \sum_{S_\beta} \bar{A}_{i,j,k} + \log r \sum_{S_\gamma} A_{i,j,k} + \log \bar{r} \sum_{S_\gamma} \bar{A}_{i,j,k} \\ &= \log \frac{p}{\bar{p}} \sum_{S_\alpha} A_{i,j,k} + \log \bar{p} \sum_{S_\alpha} 1 + \log \frac{q}{\bar{q}} \sum_{S_\beta} A_{i,j,k} + \log \bar{q} \sum_{S_\beta} + \log \frac{r}{\bar{r}} \sum_{S_\gamma} A_{i,j,k} + \log \bar{r} \sum_{S_\gamma} 1 \end{aligned}$$

Write $\sum_{S_\beta} A_{i,j,k} = \sum A_{i,j,k} - \sum_{S_\alpha} A_{i,j,k} - \sum_{S_\gamma} A_{i,j,k}$. After some manipulations, we can arrive at the expression

$$L(\sigma; A) = \left(\log \frac{f(p)}{f(q)} \sum_{S_\alpha} A_{i,j,k} - \log \frac{\bar{q}}{\bar{p}} \sum_{S_\alpha} 1\right) - \left(\log \frac{f(q)}{f(r)} \sum_{S_\gamma} A_{i,j,k} - \log \frac{\bar{r}}{\bar{q}} \sum_{S_\gamma} 1\right) + \Big(\log f(q) \sum A_{i,j,k} + \log \bar{q} \sum 1\Big)$$

where the last term is invariant to the choice of σ or σ_0.

B. *Proof of Claim 5.2*

Recall that the *error-event* equation we try to simplify is

$$L(\sigma; A) \geq L(\sigma_0; A) \tag{10}$$

When rearranging terms in (10), we would encouter terms like $\sum_{S_\alpha} - \sum_{S_{\alpha_0}}$. Hence, the computaion of the difference of the cardinality between S_α and S_{α_0} is required. It is not hard to see that $|S_\alpha| - |S_{\alpha_0}| = n(\alpha, \beta_0) + n(\alpha, \gamma_0) - n(\beta, \alpha_0) - n(\gamma, \alpha_0)$.

These are the numbers of remaining $\mathcal{R}.\mathcal{V}$.s after cancellation in (10). Likewise, we can do the same calculations between S_γ and S_{γ_0}. Details out, we have

$$\begin{aligned}
&\Rightarrow C_{pq}\left(\sum_{i=1}^{n(\alpha,\beta_0)} V_i + \sum_{i=1}^{n(\alpha,\gamma_0)} W_i - \sum_{i=1}^{n(\beta,\alpha_0)+n(\gamma,\alpha_0)} U_i\right) - C_{qr}\left(\sum_{i=1}^{n(\gamma,\alpha_0)} U_i + \sum_{i=1}^{n(\gamma,\beta_0)} V_i - \sum_{i=1}^{n(\alpha,\gamma_0)+n(\beta,\gamma_0)} W_i\right) \\
&\quad \geq C_{pq}\lambda_{pq}\Big(n(\alpha,\beta_0)+n(\alpha,\gamma_0)-n(\beta,\alpha_0)-n(\gamma,\alpha_0)\Big) + C_{qr}\lambda_{qr}\Big(n(\gamma,\alpha_0)+n(\gamma,\beta_0)-n(\alpha,\gamma_0)-n(\beta,\gamma_0)\Big) \\
&\Rightarrow C_{pq}\left(\sum_{i=1}^{n(\alpha,\beta_0)} V_i - \sum_{i=1}^{n(\beta,\alpha_0)} U_i\right) + C_{qr}\left(\sum_{i=1}^{n(\beta,\gamma_0)} W_i - \sum_{i=1}^{n(\beta,\beta_0)} V_i\right) + (C_{pq}+C_{qr})\left(\sum_{i=1}^{n(\alpha,\gamma_0)} W_i - \sum_{i=1}^{n(\gamma,\alpha_0)} U_i\right) \\
&\quad \geq C_{pq}\lambda_{pq}\Big(n(\alpha,\beta_0)-n(\beta,\alpha_0)\Big) + C_{qr}\lambda_{qr}\Big(n(\beta,\gamma_0)-n(\gamma,\beta_0)\Big) + (C_{pq}\lambda_{pq}+C_{qr}\lambda_{qr})\Big(n(\alpha,\gamma_0)-n(\gamma,\alpha_0)\Big)
\end{aligned}$$

The result follows because $C_{pq}+C_{qr}=C_{pr}$ and $C_{pq}\lambda_{pq}+C_{qr}\lambda_{qr}=C_{pr}\lambda_{pr}$.

C. Proof of Lemma 5.1

Recall the *Chernoff Bound* states that

$$\mathbb{P}\{X \geq x\} \leq \min_{u>0} \frac{L_X(u)}{e^{ux}} = e^{-u^* x} L_X(u^*)$$

where $L_X(u)$ is again the MGF of the $\mathcal{R}.\mathcal{V}.$ X, and $u^* = \frac{1}{2}$ for our problem.

It turns out we could separate out the expression to three terms with respect to C_{pq}, C_{qr} and C_{qr}, respectively, due to the fact that addition in the exponent translates into multiplication in the exponential. Consequently, we can focus on the first term

$$E_{pq} : C_{pq}\left(\sum_{i=1}^{n(\alpha,\beta_0)} V_i + \sum_{i=1}^{n(\beta,\alpha_0)} U_i\right) \geq C_{pq}\lambda_{pq}\big(n(\alpha,\beta_0)-n(\beta,\alpha_0)\big)$$

and the result follows by symmetry.

$$\begin{aligned}
\mathbb{P}\{E_{pq}\} &\leq \Big(\mathbb{E}[e^{u^* C_{pq} V}]\Big)^{n(\alpha,\beta_0)} \Big(\mathbb{E}[e^{-u^* C_{pq} U}]\Big)^{n(\beta,\alpha_0)} \cdot e^{-u^* C_{pq}\lambda_{pq}\left(n(\alpha,\beta_0)-n(\beta,\alpha_0)\right)} \\
&= \Big(\mathbb{E}[e^{u^* C_{pq} V}]\mathbb{E}[e^{-u^* C_{pq} U}]\Big)^{\frac{1}{2}\left(n(\alpha,\beta_0)+n(\beta,\alpha_0)\right)} \cdot \left(\frac{\mathbb{E}[e^{u^* C_{pq} V}]^{\frac{1}{2}}}{\mathbb{E}[e^{-u^* C_{pq} U}]^{\frac{1}{2}}} e^{-u^* C_{pq}\lambda_{pq}}\right)^{n(\alpha,\beta_0)-n(\beta,\alpha_0)} \\
&= \big(e^{-I_{pq}}\big)^{\frac{1}{2}(n(\alpha,\beta_0)+n(\beta,\alpha_0))} \cdot 1^{n(\alpha,\beta_0)-n(\beta,\alpha_0)} \\
&= exp\Big(-\frac{1}{2}\big(n(\alpha,\beta_0)+n(\beta,\alpha_0)\big) I_{pq}\Big)
\end{aligned}$$

Note that we could drop out the term $n(\alpha,\gamma)I_{pr}$ in the following derivations.

D. Proof of Lemma 5.2

To derive the order of $n(\alpha,\beta)$ and $n(\beta,\gamma)$, it is equivalent to focus on the order of $n(\alpha,\beta_0)$ and $n(\beta,\gamma_0)$ by symmetry. Let $n'_k = |\{i \mid \sigma(i)=k\}|$ denote the population in the k-th community under a candidate assignment σ, and n^0_k is the corresponding quantity for the true assignment σ_0. From the difinition of Θ^0, we have $(1-\eta)n' \leq n'_k,\ n^0_k \leq (1+\eta)n'$. In addition, to study the mis-classification in the k-th community, we use the notations as follows.

$$\begin{aligned}
m_k &= \big|\{i \mid \sigma(i)=k,\ \sigma_0(i)\neq k\}\big|, \ \sum_k m_k = m \\
m_{k,k'} &= \big|\{i \mid \sigma(i)=k,\ \sigma_0(i)=k'\}\big|, \ \sum_{k'\neq k} m_{k,k'} = m_k \\
m_k + m_{k,k} &= n'_k
\end{aligned}$$

An important observation is that $\forall k' \neq k$, the following holds.

$$m_{k,k'} \leq \frac{2}{3}(1+\eta)n'$$

Otherwise, if $m_{k,k'} > \frac{2}{3}(1+\eta)n'$, then $m_{k',k'} \leq n^0_k - m_{k,k'} < \frac{1}{3}(1+\eta)n'$. Exchanging the label of k and k', we could recover at least $m_{k,k'} - (n'_k - m_{k,k'}) - m_{k',k'} > 0$ node. This contradicts our assumption that $d(\sigma,\sigma_0)=m$ since we can achieve a smaller distance after a simple permutation of labels.

Here, we mainly utilize a few techniques from convex analysis to help us find combinatorial lower bounds:

- *Maximum Principle:* A convex function attains its maximum at the extreme points of the underlying convex set.
- For a convex optimization problem where we try to minimize a convex function, if
  - the objective function is separable and isopotric in all the coordinates
  - every inequality constraint is affine, separable and isopotric in all the coordinates
  - the uniform point (with all its coordinates being equal) lies inside the feasible region

  then the uniform point satisfies the corresponding KKT conditions, i.e., it is a primal optimal point.

For either $n(\alpha, \beta_0)$ or $n(\beta, \gamma_0)$, we will partition out into two cases: $m \leq \frac{n'}{2}$ and $m > \frac{n'}{2}$.

1) Case $m \leq \frac{n'}{2}$:

- $n(\alpha, \beta_0)$

  Define $n_k(\alpha, \beta_0) \triangleq |\{(i,j,k) \in S_{\alpha,\beta_0} \mid \sigma(i) = k\}|$. Obviously, $n(\alpha, \beta_0) = \sum_{k=1}^{K} n_k(\alpha, \beta_0)$. We have

$$\begin{aligned} n_k(\alpha, \beta_0) &\geq |\{(i,j) \mid \sigma(i) = k, \sigma_0(i) = k;\, \sigma(j) = k, \sigma_0(j) = k\}| \cdot |\{i \mid \sigma(i) = k, \sigma_0(i) \neq k\}| \\ &+ |\{i \mid \sigma(i) = k, \sigma_0(i) = k\}| \cdot |\{(i,j) \mid \sigma(i) = k, \sigma_0(i) \neq k;\, \sigma(j) = k, \sigma_0(j) \neq k\}| \\ &= \binom{n'_k - m_k}{2} m_k + (n'_k - m_k)\binom{m_k}{2} \end{aligned}$$

  Thus,

$$\begin{aligned} n(\alpha, \beta_0) &\geq \sum_k \binom{n'_k - m_k}{2} m_k + (n'_k - m_k)\binom{m_k}{2} \\ &= \sum_k \frac{(n'_k - m_k)^2 m_k - (n'_k - m_k) m_k + (n'_k - m_k) m_k^2 - (n'_k - m_k) m_k}{2} \\ &= \sum_k \frac{(n'_k)^2 m_k - (n'_k) m_k^2 - 2(n'_k - m_k) m_k}{2} \\ &= \frac{\sum_k \left[(n'_k)^2 - 2n'_k\right] m_k - (n'_k - 2)\sum_k m_k^2}{2} \\ &\geq \frac{\left[(n'_k)^2 - 2n'_k\right] m - (n'_k - 2) m^2}{2} \\ &= \frac{(n'_k - 2)(n'_k - m) m}{2} \\ &\geq \Big(1 - \frac{2}{n'_k}\Big)\Big(1 - \frac{m}{n'_k}\Big)\frac{(n'_k)^2}{2} m \\ &\geq (1 - 2\eta)\Big(1 - \frac{\frac{m}{1-\eta}}{n'}\Big)\Big(1 - \frac{\frac{2}{1-\eta}}{n'}\Big)\frac{(n')^2}{2} m \\ &\geq (1 - 2\eta)\Big(1 - \frac{\frac{m}{1-2\eta}}{n'}\Big)\Big(1 - \frac{\frac{1}{1-2\eta}}{n'}\Big)\frac{(n')^2}{2} m \\ &= (1 - \eta')\Big(1 - \frac{\frac{m}{1-\eta'}}{n'}\Big)\Big(1 - \frac{\frac{1}{1-\eta'}}{n'}\Big)\frac{(n')^2}{2} m \end{aligned}$$

- $n(\beta, \gamma_0)$

  Similarly, let $n_l(\beta, \gamma_0) \triangleq |\{(i,j,k) \in S_{\beta,\gamma_0} \mid \sigma(\mathsf{Mode}\{\sigma(i), \sigma(j), \sigma(k)\}) = l\}|$ where $\mathsf{Mode}\{\{s_i\}_{i=1}^n\} = s_m$ with $m = \arg\max_i |\{j \mid s_j = s_i\}|$ is the mode (majority) within a finite set $\{s_i\}$. We have

$$\begin{aligned} n_k(\beta, \gamma_0) &\geq |\{i \mid \sigma(i) = k, \sigma_0(i) = k\}| \cdot |\{i \mid \sigma(i) = k, \sigma_0(i) = k' \text{ for some } k' \neq k\}| \cdot |\{i \mid \sigma(i) \neq k, \sigma_0(i) \notin \{k, k'\}\}| \\ &= m_{k,k} \sum_{k' \neq k} m_{k,k'}(n - n'_k - n_k^0 - n_{k'}^0 + m_{k,k'} + m_{k,k}) \\ &\geq (n'_k - m_k) m_k \big(n - n_k^0 - n_{k'}^0 - m_k\big) \\ &\geq (n'_k - m_k) m_k \big(K(1-\eta)n' - 2(1+\eta)n' - m_k\big) \\ &= (n'_k - m_k) m_k \Big((K-2)\big(1 - \frac{(K+2)\eta}{K-2}\big)n' - m_k\Big) \\ &\geq (n'_k - m_k) m_k \big((K-2)(1-5\eta)n' - m_k\big) \end{aligned}$$

Thus,

$$\begin{aligned}
n(\beta,\gamma_0) &\geq \sum_k (n'_k - m_k)m_k\big((K-2)(1-5\eta)n' - m_k\big) \\
&= (K-2)(1-5\eta)n'\sum_k (n'_k - m_k)m_k - \sum_k (n'_k - m_k)m_k^2 \\
&\geq (K-2)(1-5\eta)n'(n'_k m - m^2) - (n'_k - m)m^2 \\
&\geq \big[(K-2)(1-5\eta)n' - m\big]\big[(1-\eta)n' - m\big]m \\
&\geq (1-\eta)(1-5\eta)\big(1 - \frac{m}{(K-2)(1-5\eta)n'}\big)\big(\frac{m}{(1-\eta)n'}\big)(K-2)(n')^2 m \\
&\geq (1-6\eta)\Big(1 - \frac{\frac{m}{1-5\eta}}{n'}\Big)\Big(1 - \frac{\frac{m}{1-\eta}}{n'}\Big)(K-2)(n')^2 m \\
&= (1-\eta')\Big(1 - \frac{\frac{m}{1-\eta'}}{n'}\Big)^2 (K-2)(n')^2 m
\end{aligned}$$

2) Case $m > \frac{n'}{2}$:

- $n(\alpha,\beta_0)$

$$\begin{aligned}
n_k(\alpha,\beta_0) &= \sum_{k_1\neq k_2}\binom{m_{k,k_1}}{2} m_{k,k_2} \\
&= \sum_{k_1}\sum_{k_2\neq k_1}\frac{m_{k,k_1}^2 - m_{k,k_1}}{2} m_{k,k_2} \\
&= \sum_{k_1}\frac{m_{k,k_1}^2 - m_{k,k_1}}{2}(n'_k - m_{k,k_1}) \\
&= \frac{1}{2}\big((m_{k,k}^2 - m_{k,k})(n'_k - m_{k,k})\big) + \frac{1}{2}\sum_{k_1\neq k}\big((m_{k,k_1}^2 - m_{k,k_1})(n'_k - m_{k,k_1})\big) \\
&= \frac{1}{2}((n'_k - m_k)^2 - (n'_k - m_k))m_k + \frac{1}{2}\Big((n'_k+1)\sum_{k_1\neq k} m_{k,k_1}^2 - n'_k m_k - \sum_{k_1\neq k} m_{k,k_1}^3\Big)
\end{aligned}$$

Hence,

$$\frac{n_k(\alpha,\beta_0)}{m_k} = \frac{1}{2}(n'_k - m_k)(n'_k - m_k - 1) + \frac{1}{2}\Big(\frac{\sum_{k'\neq k}(n'_k + 1 - m_{k,k'})m_{k,k'}^2}{m_k} - n'_k\Big)$$

If $m_k \leq \frac{2(1+\eta)}{3}n'$, then

$$\begin{aligned}
\frac{n_k(\alpha,\beta_0)}{m_k} &\geq \frac{1}{2}(n'_k - m_k)(n'_k - m_k - 1) \\
&\geq \frac{1}{2}\big((1-\eta)n' - \frac{2(1+\eta)}{3}n'\big)\big((1-\eta)n' - \frac{2(1+\eta)}{3}n' - 1\big) \\
&= \frac{1}{2}\frac{1-5\eta}{3}n'\big(\frac{1-5\eta}{3}n' - 1\big) \\
&\geq \frac{1-10\eta}{9}\Big(1 - \frac{\frac{3}{1-5\eta}}{n'}\Big)\frac{(n')^2}{2} \\
&\geq \frac{1-\eta'}{9}\Big(1 - \frac{\frac{9}{1-\eta'}}{n'}\Big)\frac{(n')^2}{2}
\end{aligned}$$

For $\frac{2(1+\eta)}{3}n' < m_k \leq n'_k$, we first first look at the case $K \geq 3$.

$$\begin{aligned}
\frac{n_k(\alpha,\beta_0)}{m_k} &\geq \frac{1}{2}\Big(\frac{\sum_{k'\neq k}(n'_k + 1 - m_{k,k'})m_{k,k'}^2}{m_k} - n'_k\Big) \\
&\geq \frac{1}{2}\Big(\frac{\sum_{k'\neq k}(n'_k + 1 - \frac{m_k}{K-1})\big(\frac{m_k}{K-1}\big)^2}{m_k} - n'_k\Big) \\
&\geq \frac{1}{2}\Big(\big(n'_k + 1 - \frac{m_k}{K-1}\big)\frac{m_k}{K-1} - n'_k\Big)
\end{aligned}$$

$$
\begin{aligned}
&\geq \frac{1}{2}\Big(((1-\eta)n' + 1 - \frac{(1+\eta)n'}{K-1})\frac{(1+\eta)n'}{K-1} - (1+\eta)n'\Big) \\
&= \frac{1}{2}\frac{(1+\eta)n'}{K-1}\Big((1-\eta)n' + 1 - \frac{(1+\eta)n'}{K-1} - (K-1)\Big) \\
&\geq \frac{1}{2}\frac{(1+\eta)n'}{K-1}\Big(\frac{(K-2)-(K+2)\eta}{K-1}n' - (K-2)\Big) \\
&\geq \frac{(1+\eta)}{K-1}\frac{1-(K+2)\eta}{K-1}\Big(1 - \frac{\frac{(K-1)(K-2)}{1-(K+2)\eta}}{n'}\Big)\frac{(n')^2}{2} \\
&\geq \frac{1-(2K+3)\eta}{(K-1)^2}\Big(1 - \frac{\frac{(K-1)(K-2)}{1-(K+2)\eta}}{n'}\Big)\frac{(n')^2}{2} \\
&\geq \frac{(1-\eta')}{K^2}\Big(1 - \frac{\frac{K^2}{1-\eta'}}{n'}\Big)\frac{(n')^2}{2}
\end{aligned}
$$

In the last inequality, we can always make η' decay fast enough so that $(2K+3)\eta'$ still goes to 0 even though K grwos without bound. If $K=2$, then m_k cannot be greater than $\frac{1}{2}n'_k$. Otherwise, we can just exchange the (only two) labels and get a fewer number of mis-classifications. As a result, m_k always falls in the first regime $m_k \leq \frac{2(1+\eta)}{3}n'$ in two-community case and the lower bound obtained corresponds to the above expression with $K=3$. Writing $K' = \max\{3, K\}$, we have

$$
\begin{aligned}
\frac{n_k(\alpha,\beta_0)}{m_k} &\geq \frac{(1-\eta')}{(K')^2}\Big(1 - \frac{\frac{(K')^2}{1-\eta'}}{n'}\Big)\frac{(n')^2}{2} \\
\Rightarrow n_k(\alpha,\beta_0) &\geq \frac{(1-\eta')}{(K')^2}\Big(1 - \frac{\frac{(K')^2}{1-\eta'}}{n'}\Big)\frac{(n')^2}{2}m_k \\
\Rightarrow n(\alpha,\beta_0) = \sum_k n_k(\alpha,\beta_0) &\geq \frac{(1-\eta')}{(K')^2}\Big(1 - \frac{\frac{(K')^2}{1-\eta'}}{n'}\Big)\frac{(n')^2}{2}\sum_k m_k = \frac{(1-\eta')}{(K')^2}\Big(1 - \frac{\frac{(K')^2}{1-\eta'}}{n'}\Big)\frac{(n')^2}{2}m
\end{aligned}
$$

- $n(\beta,\gamma_0)$
  Similarly, write

$$
\begin{aligned}
n_k(\beta,\gamma_0) &= \sum_{l\neq k}\frac{1}{2!}\sum_{k_1\neq k_2\neq k_3} m_{k,k_1}m_{k,k_2}m_{l,k_3} \\
&= \sum_{l\neq k}\frac{1}{2}\sum_{k_3} m_{l,k_3}\sum_{k_1\neq k_2\neq k_3} m_{k,k_1}m_{k,k_2} \\
&= \sum_{l\neq k}\frac{1}{2}\sum_{k_3} m_{l,k_3}\big((n'_k - m_{k,k_3})^2 - \sum_{k'\neq k_3}(m_{k,k'})^2\big) \\
&= \sum_{l\neq k}\sum_{k_3} m_{l,k_3}\Big(\frac{1}{2}\big((n'_k)^2 - \sum_{k'}(m_{k,k'})^2\big) - \big(n'_k m_{k,k_3} - (m_{k,k_3})^2\big)\Big) \\
&= (1) - (2) \\
(1) &\Rightarrow \sum_{l\neq k} n'_l\frac{1}{2}\Big((n'_k)^2 - \sum_{k'}(m_{k,k'})^2\Big) \\
&= \frac{1}{2}(n - n'_k)\Big((n'_k)^2 - (n'_k - m_k)^2 - \sum_{k'\neq k}(m_{k,k'})^2\Big) \\
&\geq \big((K-1) - (K+1)\eta\big)n'\Big(n'_k m_k - \frac{1}{2}\big(m_k^2 + \sum_{k'\neq k}(m_{k,k'})^2\big)\Big) \\
(2) &\Rightarrow \sum_{k_3}(n^0_{k_3} - m_{k,k_3})m_{k,k_3}(n'_k - m_{k,k_3}) \\
&= (n^0_k - m_{k,k})m_{k,k}(n'_k - m_{k,k}) + \sum_{k_3\neq k}(n^0_{k_3} - m_{k,k_3})m_{k,k_3}(n'_k - m_{k,k_3}) \\
&\leq (n'_k - m_k)m_k(m_k + 2\eta n') + \sum_{k'\neq k}(n'_k - m_{k,k'})m_{k,k'}(n'_k - m_{k,k'} + 2\eta n')
\end{aligned}
$$

If $m_k \le \frac{2(1+\eta)}{3}n'$, we separate out the case $K=3$. Note that although we assume $K \ge 2$, if there are only two communities to dichotomize, then $n(\beta,\gamma) \equiv 0$ and is meaningless. For $K=3$,

$$\begin{aligned}
\frac{n_k(\beta,\gamma_0)}{m_k} &\ge (2-4\eta)n'\Big(n'_k - \frac{1}{2}\big(m_k + \frac{\frac{m_k^2}{2}}{m_k}\big)\Big) - (n'_k - m_k)(m_k + 2\eta n') - \frac{(n'_k - \frac{m_k}{2})m_k(n'_k - \frac{m_k}{2} + 2\eta n')}{m_k} \\
&\ge (2-4\eta)n'\Big((1-\eta)n' - \frac{3}{4}\frac{2(1+\eta)}{3}n'\Big) - \Big((1+\eta)n' - \frac{2(1+\eta)}{3}n'\Big)\Big(\frac{2(1+\eta)}{3}n' + 2\eta n'\Big) \\
&\qquad - \Big((1+\eta)n' - \frac{1}{2}\frac{2(1+\eta)}{3}n'\Big)\Big((1+\eta)n' - \frac{1}{2}\frac{2(1+\eta)}{3}n' + 2\eta n'\Big) \\
&= (2-4\eta)n'\frac{1-3\eta}{2}n' - \frac{1+\eta}{3}\frac{2+8\eta}{3} - \frac{2(1+\eta)}{3}\frac{2+8\eta}{3} \\
&\ge \frac{1-\eta'}{3}(K-2)(n')^2
\end{aligned}$$

$\forall K \ge 4$, notice that the following expression increases in K, which means that the minimum is attained when $K=4$.

$$\begin{aligned}
\frac{n_k(\beta,\gamma_0)}{m_k} &\ge ((K-1) - (K+1)\eta')n'\Big(n'_k - \frac{1}{2}\big(m_k + \frac{m_k^2}{m_k}\big)\Big) - (n'_k - m_k)(m_k + 2\eta' n') \\
&\qquad - \frac{(n'_k - \frac{m_k}{K-1})m_k(n'_k - \frac{m_k}{K-1} + 2\eta' n')}{m_k} \\
&\ge (3-5\eta')n'\Big((1-\eta')n' - \frac{2(1+\eta')}{3}n'\Big) - \Big((1+\eta')n' - \frac{2(1+\eta')}{3}n'\Big)\Big(\frac{2(1+\eta')}{3}n' + 2\eta' n'\Big) \\
&\qquad - \Big((1+\eta')n' - \frac{1}{3}\frac{2(1+\eta')}{3}n'\Big)\Big((1+\eta')n' - \frac{1}{3}\frac{2(1+\eta')}{3}n' + 2\eta' n'\Big) \\
&= (3-5\eta')n'\frac{1-5\eta'}{3}n' - \frac{1+\eta'}{3}\frac{2+8\eta'}{3} - \frac{7(1+\eta')}{9}\frac{7+25\eta'}{9} \\
&\ge \frac{7(1-\eta)}{81}(K-2)(n')^2
\end{aligned}$$

On the other side, if $\frac{2(1+\eta)}{3}n' < m_k \le n'_k$, then

$$\begin{aligned}
\frac{n_k(\beta,\gamma_0)}{m_k} &\ge ((K-1) - (K+1)\eta)n' \cdot \Big(n'_k - \frac{1}{2}\big(m_k + \frac{\big(\frac{2(1+\eta)}{3}n'\big)^2 + \big(m_k - \frac{2(1+\eta)}{3}n'\big)^2}{m_k}\big)\Big) - (n'_k - m_k)(m_k + 2\eta n') \\
&\qquad - \frac{(n'_k - \frac{m_k}{K-1})m_k(n'_k - \frac{m_k}{K-1} + 2\eta n')}{m_k} \\
&\ge (2-4\eta)n'\Big((1-\eta)n' - \frac{1}{2}\big((1+\eta)n' + \big(\frac{2}{3}\big)^2(1+\eta)n' + \big(\frac{1}{3}\big)^2(1+\eta)n'\big)\Big) \\
&\qquad - \big((1+\eta)n' - \frac{1}{2}(1+\eta)n'\big)\big((1+\eta)n' - \frac{1}{2}(1+\eta)n' + 2\eta n'\big) \\
&= (2-4\eta)n'\frac{2-16\eta}{9}n' - \frac{1+\eta}{2}\frac{1+5\eta}{2} \\
&\ge \frac{7(1-\eta')}{36}(K-2)(n')^2
\end{aligned}$$

We can see that $\frac{7(1-\eta')}{81}(K-2)(n')^2$ is the overall minimum. The result follows by multiplying back m_k and summing over all $k \in [K]$.

E. Proof of Lemma 5.5

We focus on the case with $K \to \infty$. For finite value K, the proof is almost identical but with different values of m_0 and m' for each of the three scenarios. From Lemma 5.3, we can see that the coefficient of I_{qr} is always greater than that of I_{pq} (order-wise). Therefore, we could focus on the term $(n')^2 I_{qr}$ when analyzing the error probability.

1) Case $\liminf_{n\to\infty} \frac{\frac{(n')^2}{2}I_{pq} + (K-2)(n')^2 I_{qr}}{\log n} > 1$*:* Then there exists small $\epsilon > 0$ such that

$$\frac{(1-2\eta')\big(1 - \frac{\frac{\epsilon}{3}}{1-\eta'}\big)^2\big(\frac{(n')^2}{2}I_{pq} + (K-2)(n')^2 I_{qr}\big)}{\log n} > 1 + \frac{\epsilon}{3}$$

Take $m_0 = 1$, we have

$$\frac{m_0}{n} = \frac{1}{n} \le exp\Big(-(1-2\eta')\big(1 - \frac{\frac{\epsilon}{3}}{1-\eta'}\big)^2\big(\frac{(n')^2}{2}I_{pq} + (K-2)(n')^2 I_{qr}\big)\Big) \triangleq R$$

Take $m' = \frac{\epsilon}{3}n'$. Then, for $m \in (m_0, m']$,

$$\begin{aligned}
\frac{m}{n}P_m &\le \frac{m}{n}\binom{n}{m}K^m \cdot exp\Big(-(1-\eta')\big(1-\frac{\frac{m}{1-\eta'}}{n'}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)m\Big) \\
&\le \left(\frac{e(n-1)K}{m_0-1}\right)^{m-1} K \cdot exp\Big(-(1-\eta')\big(1-\frac{\frac{m'}{1-\eta'}}{n'}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)m\Big) \\
&\le exp\Big(-(1-2\eta')\big(1-\frac{\frac{\epsilon}{3}}{1-\eta'}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\Big) \\
&\qquad \cdot\left(\frac{enK}{2-1}exp\Big(-(1-\eta')\big(1-\frac{\frac{\epsilon}{3}}{1-\eta'}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\Big)\right)^{m-1} \\
&\le R\left(nexp\Big(-(1-2\eta')\big(1-\frac{\frac{\epsilon}{3}}{1-\eta'}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\Big)\right)^{m-1} \\
&\le R\left(exp\big(-\frac{\epsilon}{3}\log n\big)\right)^{m-1} \\
&= Rn^{-\epsilon\frac{(m-1)}{3}}
\end{aligned}$$

For $m \in (m', n]$,

$$\begin{aligned}
\frac{m}{n}P_m &\le \left(\frac{enK}{m}\cdot exp\Big(-(1-\eta')\big(\frac{1}{(K')^2}\big(1-\frac{\frac{(K')^2}{1-\eta'}}{n'}\big)\frac{(n')^2}{2}I_{pq}+\frac{7}{81}(K-2)(n')^2I_{qr}\big)\Big)\right)^m \\
&\le \left(\frac{enK}{m'}\cdot exp\Big(-(1-\eta')\frac{7}{81}(K-2)(n')^2I_{qr}\Big)\right)^m \\
&= \left(\frac{3eK^2}{\epsilon}\right)^{18} exp\Big(-(1-\eta')\big(\frac{42}{81}(K-2)(n')^2I_{qr}+\frac{84}{81}(K-2)(n')^2I_{qr}\big)\Big) \\
&\qquad\qquad \cdot\left(\frac{3eK^2}{\epsilon}exp\Big(-(1-\eta')\frac{7}{81}(K-2)(n')^2I_{qr}\Big)\right)^{m-18} \\
&\le R\left(exp\Big(-(1-2\eta')\frac{7}{81}(K-2)(n')^2I_{qr}\Big)\right)^{m-18} \\
&\le Rn^{-\epsilon\frac{7(m-12)}{81}}
\end{aligned}$$

2) Case $\limsup_{n\to\infty}\frac{\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}}{\log n} < 1$*:* Then there exists small $\epsilon > 0$ such that

$$\frac{(1-2\eta)\big(1-\frac{K^{-\epsilon}}{1-\eta}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)}{\log n} < 1-\epsilon$$

Take $m_0 = n\cdot exp\Big(-(1-K^{-\epsilon/2})(1-2\eta)\big(1-\frac{K^{-\epsilon}}{1-\eta}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\Big)$, which satisfies $m_0 \ge (nK)^{\epsilon/2}$ and $m_0 = o(\frac{n}{K^2}) = o(\frac{n'}{K})$. Then

$$\frac{m_0}{n} = exp\Big(-(1-K^{-\epsilon/2})(1-2\eta)\big(1-\frac{K^{-\epsilon}}{1-\eta}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\Big) \triangleq R$$

Take $m' = K^{-\epsilon}n'$. Then, for $m \in (m_0, m']$,

$$\begin{aligned}
\frac{m}{n}P_m &\le \left(\frac{enK}{m_0}\cdot exp\Big(-(1-\eta)\big(1-\frac{\frac{m'}{1-\eta}}{n'}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\Big)\right)^m \\
&\le \left(\frac{enK}{m_0}exp\Big(-(1-\eta)\big(1-\frac{K^{-\epsilon}}{1-\eta}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\Big)\right)^m \\
&\le \Bigg(exp\Big(-(1-2\eta)\big(1-\frac{K^{-\epsilon}}{1-\eta}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big) \\
&\quad +(1-K^{-\epsilon/2})(1-2\eta)\big(1-\frac{K^{-\epsilon}}{1-\eta}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\Big)\Bigg)^m
\end{aligned}$$

$$
\begin{aligned}
&\le exp\Big(-\frac{m_0}{K^{\epsilon/2}}(1-2\eta)\big(1-\frac{K^{-\epsilon}}{1-\eta}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\Big)\\
&\le exp\Big(-n^{\epsilon/2}(1-2\eta)\big(1-\frac{K^{-\epsilon}}{1-\eta}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\Big)\\
&= o(R)
\end{aligned}
$$

For $m \in (m', n]$,

$$
\begin{aligned}
\frac{m}{n}P_m &\le \left(\frac{enK}{m'}exp\Big(-(1-\eta)\big(\frac{1}{(K')^2}\big(1-\frac{\frac{(K')^2}{1-\eta}}{n'}\big)\frac{(n')^2}{2}I_{pq}+\frac{7}{81}(K-2)(n')^2I_{qr}\big)\Big)\right)^m\\
&= \left(eK^2K^{\epsilon}exp\Big(-(1-\eta)\frac{7}{81}(K-2)(n')^2I_{qr}\Big)\right)^m\\
&\le \left(exp\Big(-(1-2\eta)\frac{7}{81}(K-2)(n')^2I_{qr}\Big)\right)^m\\
&\le exp\Big(-\frac{6}{81}(K-2)(n')^2I_{qr}m\Big)\\
&= exp\Big(-\frac{4}{81}\cdot\frac{3}{2}(K-2)(n')^2I_{qr}m\Big)\\
&= o(R)
\end{aligned}
$$

3) Case $\limsup_{n\to\infty}\frac{\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}}{\log n}=1+o(1)$*:* Then there exists a small positive sequence $\omega=\omega_n\downarrow 0$ with $\omega\ge\frac{1}{\sqrt{\log n}}$ such that

$$
\left|\frac{(1-2\eta)\big(1-\frac{\omega^2}{1-\eta}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)}{\log n}-1\right|\ll\omega
$$

Take $m_0=n\cdot exp\Big(-(1-\omega)(1-2\eta)\big(1-\frac{\omega^2}{1-\eta}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\Big)$, which satisfies $m_0\ge n^{\omega/2}$ and $m_0=o(\frac{n'}{\log n})$. Then

$$
\frac{m_0}{n}=exp\Big(-(1-\omega)(1-2\eta)\big(1-\frac{\omega^2}{1-\eta}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\Big)\triangleq R
$$

Take $m'=\omega^2n'$. Then, for $m\in(m_0,m']$,

$$
\begin{aligned}
\frac{m}{n}P_m &\le \left(\frac{enK}{m_0}\cdot exp\Big(-(1-\eta)\big(1-\frac{\frac{m'}{1-\eta}}{n'}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\Big)\right)^m\\
&\le \left(\frac{enK}{m_0}exp\Big(-(1-\eta)\big(1-\frac{\omega^2}{1-\eta}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\Big)\right)^m\\
&\le \Big(exp\Big(-(1-2\eta)\big(1-\frac{\omega^2}{1-\eta}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\\
&\qquad +(1-\omega)(1-2\eta)\big(1-\frac{\omega^2}{1-\eta}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\Big)\Big)^m\\
&= exp\Big(-m_0\omega(1-2\eta)\big(1-\frac{\omega^2}{1-\eta}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\Big)\\
&\le exp\Big(-\frac{n^{\omega/2}}{\sqrt{\log n}}(1-2\eta)\big(1-\frac{\omega^2}{1-\eta}\big)^2\big(\frac{(n')^2}{2}I_{pq}+(K-2)(n')^2I_{qr}\big)\Big)\\
&= o(R)
\end{aligned}
$$

For $m\in(m',n]$,

$$
\begin{aligned}
\frac{m}{n}P_m &\le \left(\frac{enK}{m'}exp\Big(-(1-\eta)\big(\frac{1}{(K')^2}\big(1-\frac{\frac{(K')^2}{1-\eta}}{n'}\big)\frac{(n')^2}{2}I_{pq}+\frac{7}{81}(K-2)(n')^2I_{qr}\big)\Big)\right)^m\\
&= \left(eK^2\log n\cdot exp\Big(-(1-\eta)\frac{7}{81}(K-2)(n')^2I_{qr}\Big)\right)^m
\end{aligned}
$$

$$
\begin{aligned}
&\leq \left(exp\Big(-(1-2\eta)\frac{7}{81}(K-2)(n')^2 I_{qr} \Big) \right)^m \\
&\leq exp\Big(-\frac{6}{81}(K-2)(n')^2 I_{qr} m \Big) \\
&= exp\Big(-\frac{4}{81} \cdot \frac{3}{2}(K-2)(n')^2 I_{qr} m \Big) \\
&= o(R)
\end{aligned}
$$