*Submitted to the Annals of Statistics*
arXiv: arXiv:0000.0000

# RATE-OPTIMAL GRAPHON ESTIMATION

By Chao Gao, Yu Lu and Harrison H. Zhou

*Yale University*

Network analysis is becoming one of the most active research areas in statistics. Significant advances have been made recently on developing theories, methodologies and algorithms for analyzing networks. However, there has been little fundamental study on optimal estimation. In this paper, we establish optimal rate of convergence for graphon estimation. For the stochastic block model with $k$ clusters, we show that the optimal rate under the mean squared error is $n^{-1}\log k + k^2/n^2$. The minimax upper bound improves the existing results in literature through a technique of solving a quadratic equation. When $k \leq \sqrt{n\log n}$, as the number of the cluster $k$ grows, the minimax rate grows slowly with only a logarithmic order $n^{-1}\log k$. A key step to establish the lower bound is to construct a novel subset of the parameter space and then apply Fano's lemma, from which we see a clear distinction of the nonparametric graphon estimation problem from classical nonparametric regression, due to the lack of identifiability of the order of nodes in exchangeable random graph models. As an immediate application, we consider nonparametric graphon estimation in a Hölder class with smoothness $\alpha$. When the smoothness $\alpha \geq 1$, the optimal rate of convergence is $n^{-1}\log n$, independent of $\alpha$, while for $\alpha \in (0,1)$, the rate is $n^{-\frac{2\alpha}{\alpha+1}}$, which is, to our surprise, identical to the classical nonparametric rate.

**1. Introduction.** Network analysis [20] has gained cosiderable research interests in both theories [7] and applications [51, 19]. A lot of recent work has been focusing on studying networks from a nonparametric perspective [7], following the deep advancement in exchangeable arrays [3, 30, 32, 14]. In this paper, we study the fundamental limits in estimating the underlying generating mechanism of network models, called graphon. Though various algorithms have been proposed and analyzed [10, 45, 52, 2, 9], it is not clear whether the convergence rates obtained in these works can be improved, and not clear what the differences and connections are between nonparametric graphon estimation and classical nonparametric regression. The results obtained in this paper provide answers to those questions. We found many existing results in literature are not sharp. Nonparametric graphon estima-

*MSC 2010 subject classifications:* Primary 60G05
*Keywords and phrases:* network, graphon, stochastic block model, nonparametric regression, minimax rate

tion can be seen as nonparametric regression without knowing design. When the smoothness of the graphon is small, the minimax rate of graphon estimation is identical to that of nonparametric regression. This is surprising, since graphon estimation seems to be a more difficult problem, for which the design is not observed. When the smoothness is high, we show that the minimax rate does not depend on the smoothness anymore, which provides a clear distinction between nonparametric graphon estimation and nonparametric regression.

We consider an undirected graph of $n$ nodes. The connectivity can be encoded by an adjacency matrix $\{A_{ij}\}$ taking values in $\{0,1\}^{n\times n}$. The value of $A_{ij}$ stands for the presence or the absence of an edge between the $i$-th and the $j$-th nodes. The model in this paper is $A_{ij} = A_{ji} \sim \text{Bernoulli}(\theta_{ij})$ for $1 \leq j < i \leq n$, where

$$\theta_{ij} = f(\xi_i, \xi_j), \quad i \neq j \in [n]. \tag{1.1}$$

The sequence $\{\xi_i\}$ are random variables sampled from a distribution $\mathbb{P}_\xi$ supported on $[0,1]^n$. A common choice for the probability $\mathbb{P}_\xi$ is i.i.d. uniform distribution on $[0,1]$. In this paper, we allow $\mathbb{P}_\xi$ to be any distribution, so that the model (1.1) is studied to its full generality. Given $\{\xi_i\}$, we assume $\{A_{ij}\}$ are independent for $1 \leq j < i \leq n$, and adopt the convention that $A_{ii} = 0$ for each $i \in [n]$. The nonparametric model (1.1) is inspired by the advancement of graph limit theory [37, 14, 36]. The function $f(x,y)$, which is assumed to be symmetric, is called graphon. This concept plays a significant role in network analysis. Since graphon is an object independent of the network size $n$, it gives a natural criterion to compare networks of different sizes. Moreover, model based prediction and testing can be done through graphon [35]. Besides nonparametric models, various parametric models have been proposed on the matrix $\{\theta_{ij}\}$ to capture different aspects of the network [28, 29, 44, 43, 24, 27, 1, 33].

The model (1.1) has a close relation to the classical nonparametric regression problem. We may view the setting (1.1) as modeling the mean of $A_{ij}$ by a regression function $f(\xi_i, \xi_j)$ with design $\{(\xi_i, \xi_j)\}$. In a regression problem, the design points $\{(\xi_i, \xi_j)\}$ are observed, and the function $f$ is estimated from the pair $\{(\xi_i, \xi_j), A_{ij}\}$. In contrast, in the graphon estimation setting, $\{(\xi_i, \xi_j)\}$ are latent random variables, and $f$ can only be estimated from the response $\{A_{ij}\}$. This causes an identifiability problem, because without observing the design, there is no way to associate the value of $f(x,y)$ with $(x,y)$. In this paper, we consider the following loss function

$$\frac{1}{n^2} \sum_{i,j\in[n]} (\hat{\theta}_{ij} - \theta_{ij})^2$$

to overcome the identifiability issue. This is identical to the loss function widely used in the classical nonparametric regression problem with the form

$$\frac{1}{n^2}\sum_{i,j\in[n]}\left(\hat{f}(\xi_i,\xi_j)-f(\xi_i,\xi_j)\right)^2.$$

Even without observing the design $\{(\xi_i,\xi_j)\}$, it is still possible to estimate the matrix $\{\theta_{ij}\}$ by exploiting its underlying structure modeled by (1.1).

We first consider $\{\theta_{ij}\}$ of a block structure. This stochastic block model, proposed by [29], is serving as a standard data generating process in network community detection problem [7, 47, 4, 31, 34, 8]. We denote the parameter space for $\{\theta_{ij}\}$ by $\Theta_k$, where $k$ is the number of clusters in the stochastic block model. In total, there are an order of $k^2$ number of blocks in $\{\theta_{ij}\}$. The value of $\theta_{ij}$ only depends on the clusters that the $i$-th and the $j$-th nodes belong to. The exact definition of $\Theta_k$ is given in Section 2.2. For this setting, the minimax rate for estimating the matrix $\{\theta_{ij}\}$ is as follows.

THEOREM 1.1. *Under the stochastic block model, we have*

$$\inf_{\hat{\theta}}\sup_{\theta\in\Theta_k}\mathbb{E}\left\{\frac{1}{n^2}\sum_{i,j\in[n]}(\hat{\theta}_{ij}-\theta_{ij})^2\right\}\asymp\frac{k^2}{n^2}+\frac{\log k}{n},$$

*for any* $1\leq k\leq n$.

The convergence rate has two terms. The first term $k^2/n^2$ is due to the fact that we need to estimate an order of $k^2$ number of unknown parameters with an order of $n^2$ number of observations. The second term $n^{-1}\log k$, which we coin as the clustering rate, is the error induced by the lack of identifiability of the order of nodes in exchangeable random graph models. Namely, it is resulted from the unknown clustering structure of the $n$ nodes. This term grows logarithmically as the number of clusters $k$ increases, which is different from what is obtained in literature [10] based on lower rank matrix estimation.

We also study the minimax rate of estimating $\{\theta_{ij}\}$ modeled by the relation (1.1) with $f$ belonging to a Hölder class $\mathcal{F}_\alpha(M)$ with smoothness $\alpha$. The class $\mathcal{F}_\alpha(M)$ is rigorously defined in Section 2.3. The result is stated in the following theorem.

THEOREM 1.2. *Consider the Hölder class $\mathcal{F}_\alpha(M)$, defined in Section 2.3. We have*

$$\inf_{\hat{\theta}}\sup_{f\in\mathcal{F}_\alpha(M)}\sup_{\xi\sim\mathbb{P}_\xi}\mathbb{E}\left\{\frac{1}{n^2}\sum_{i,j\in[n]}(\hat{\theta}_{ij}-\theta_{ij})^2\right\}\asymp\begin{cases}n^{-\frac{2\alpha}{\alpha+1}}, & 0<\alpha<1,\\ \frac{\log n}{n}, & \alpha\geq 1,\end{cases}$$

*where the expectation is jointly over* $\{A_{ij}\}$ *and* $\{\xi_i\}$.

The approximation of piecewise block function to an $\alpha$-smooth graphon $f$ yields an additional error at the order of $k^{-2\alpha}$ (see Lemma 2.1). In view of the minimax rate in Theorem 1.1, picking the best $k$ to trade off the sum of the three terms $k^{-2\alpha}$, $k^2/n^2$, and $n^{-1}\log k$ gives the minimax rate in Theorem 1.2.

The minimax rate reveals a new phenomenon in nonparametric estimation. When the smoothness parameter $\alpha$ is smaller than 1, the optimal rate of convergence is the typical nonparametric rate. Note that the typical nonparametric rate is $N^{-\frac{2\alpha}{2\alpha+d}}$ [49], where $N$ is the number of observations and $d$ is the function dimension. Here, we are in a two-dimensional setting with number of observations $N \asymp n^2$ and dimension $d = 2$. Then the corresponding rate is $N^{-\frac{2\alpha}{2\alpha+d}} \asymp n^{-\frac{2\alpha}{\alpha+1}}$. Surprisingly, in Theorem 1.2 for the regime $\alpha \in (0,1)$, we get the exact same nonparametric minimax rate, though we are not given the knowledge of the design $\{(\xi_i, \xi_j)\}$. The cost of not observing the design is reflected in the case with $\alpha \geq 1$. In this regime, the smoothness of the function does not help improve the rate anymore. The minimax rate is dominated by $n^{-1}\log n$, which is essentially contributed by the logarithmic cardinality of the set of all possible assignments of $n$ nodes to $k$ clusters. A distinguished feature of Theorem 1.2 to note is that we do not impose any assumption on the distribution $\mathbb{P}_\xi$.

To prove Theorem 1.1 and Theorem 1.2, we develop a novel lower bound argument (see Section 3.3 and Section 4.2), which allows us to correctly obtain the packing number of all possible assignments. The packing number characterizes the difficulty brought by the ignorance of the design $\{(\xi_i, \xi_j)\}$ in the graphon model or the ignorance of clustering structure in the stochastic block model. Such argument may be of independent interest, and we expect its future applications in deriving minimax rates of other network estimation problems.

Our work on optimal graphon estimation is closely connected to a growing literature on nonparametric network analysis. For estimating the matrix $\{\theta_{ij}\}$ of stochastic block model, [10] viewed $\{\theta_{ij}\}$ as a rank-$k$ matrix and applied singular value thresholding on the adjacency matrix. The convergence rate obtained is $\sqrt{k/n}$, which is not optimal compared with the rate $n^{-1}\log k + k^2/n^2$ in Theorem 1.1. For nonparametric graphon estimation, [52] considered estimating $f$ in a Hölder class with smoothness $\alpha$ and obtained the rate $\sqrt{n^{-\alpha/2}\log n}$ under a closely related loss function. The work by [9] obtained the rate $n^{-1}\log n$ for estimating a Lipschitz $f$, but they imposed strong assumptions on $f$. Namely, they assumed

$L_2|x-y| \leq |g(x)-g(y)| \leq L_1|x-y|$ for some constants $L_1, L_2$, with $g(x) = \int_0^1 f(x,y)dy$. Note that this condition excludes the stochastic block model, for which $g(x)-g(y) = 0$ when different $x$ and $y$ are in the same cluster. Local asymptotic normality for stochastic block model was established in [6]. A method of moment via tensor decomposition was proposed by [5].

**Organization.** The paper is organized as follows. In Section 2, we state the main results of the paper, including both upper and lower bounds for stochastic block model and nonparametric graphon estimation. Section 3 is a discussion section, where we discuss possible generalization of the model, relation to nonparametric regression without knowing design and lower bound techniques used in network analysis. The main body of the technical proofs are presented in Section 4, and the remaining proofs are stated in the supplementary material [15].

**Notation.** For any positive integer $d$, we use $[d]$ to denote the set $\{1, 2, ..., d\}$. For any $a, b \in \mathbb{R}$, let $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. The floor function $\lfloor a \rfloor$ is the largest integer no greater than $a$, and the ceiling function $\lceil a \rceil$ is the smallest integer no less than $a$. For any two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n \asymp b_n$ means there exists a constant $C > 0$ independent of $n$, such that $C^{-1}b_n \leq a_n \leq Cb_n$ for all $n$. For any $\{a_{ij}\}, \{b_{ij}\} \in \mathbb{R}^{n\times n}$, we denote the $\ell_2$ norm by $||a|| = \sqrt{\sum_{i,j\in[n]} a_{ij}^2}$ and the inner product by $\langle a, b\rangle = \sum_{i,j\in[n]} a_{ij}b_{ij}$. Given any set $S$, $|S|$ denotes its cardinality, and $\mathbb{I}\{x \in S\}$ stands for the indicator function which takes value 1 when $x \in S$ and takes value 0 when $x \notin S$. For a metric space $(T, \rho)$, the covering number $\mathcal{N}(\epsilon, T, \rho)$ is the smallest number of balls with radius $\epsilon$ and centers in $T$ to cover $T$, and the packing number $\mathcal{M}(\epsilon, T, \rho)$ is the largest number of points in $T$ that are at least $\epsilon$ away from each other. The symbols $\mathbb{P}$ and $\mathbb{E}$ stand for generic probability and expectation, whenever the distribution is clear from the context.

## 2. Main Results.

In this section, we present the main results of the paper. We first introduce the estimation procedure in Section 2.1. The minimax rates of stochastic block and nonparametric graphon estimation are stated in Section 2.2 and Section 2.3, respectively.

### 2.1. *Methodology.*

We are going to propose an estimator for both stochastic block model and nonparametric graphon estimation under Hölder smoothness. To introduce the estimator, let us define the set $\mathcal{Z}_{n,k} = \{z : [n] \to [k]\}$

to be the collection of all possible mappings from $[n]$ to $[k]$ with some integers $n$ and $k$. Given a $z \in \mathcal{Z}_{n,k}$, the sets $\{z^{-1}(a) : a \in [k]\}$ form a partition of $[n]$, in the sense that $\cup_{a\in[k]} z^{-1}(a) = [n]$ and $z^{-1}(a) \cap z^{-1}(b) = \varnothing$ for any $a \neq b \in [k]$. In other words, $z$ defines a clustering structure on the $n$ nodes. It is easy to see that the cardinality of $\mathcal{Z}_{n,k}$ is $k^n$. Given a matrix $\{\eta_{ij}\} \in \mathbb{R}^{n\times n}$, and a partition function $z \in \mathcal{Z}_{n,k}$, we use the following notation to denote the block average on the set $z^{-1}(a) \times z^{-1}(b)$. That is,

$$\bar{\eta}_{ab}(z) = \frac{1}{|z^{-1}(a)||z^{-1}(b)|} \sum_{i\in z^{-1}(a)} \sum_{j\in z^{-1}(b)} \eta_{ij}, \quad \text{for } a \neq b \in [k], \tag{2.1}$$

and when $|z^{-1}(a)| > 1$,

$$\bar{\eta}_{aa}(z) = \frac{1}{|z^{-1}(a)|(|z^{-1}(a)| - 1)} \sum_{i\neq j\in z^{-1}(a)} \eta_{ij}, \quad \text{for } a \in [k]. \tag{2.2}$$

For any $Q = \{Q_{ab}\} \in \mathbb{R}^{k\times k}$ and $z \in \mathcal{Z}_{n,k}$, define the objective function

$$L(Q, z) = \sum_{a,b\in[k]} \sum_{\substack{(i,j)\in z^{-1}(a)\times z^{-1}(b) \\ i\neq j}} (A_{ij} - Q_{ab})^2.$$

For any optimizer of the objective function,

$$(\hat{Q}, \hat{z}) \in \underset{Q\in\mathbb{R}^{k\times k}, z\in\mathcal{Z}_{n,k}}{\operatorname{argmin}} L(Q, z), \tag{2.3}$$

the estimator of $\theta_{ij}$ is defined as

$$\hat{\theta}_{ij} = \hat{Q}_{\hat{z}(i)\hat{z}(j)}, \quad i > j, \tag{2.4}$$

and $\hat{\theta}_{ij} = \hat{\theta}_{ji}$ for $i < j$. Set the diagonal element by $\hat{\theta}_{ii} = 0$. The procedure (2.4) can be understood as first clustering the data by an estimated $\hat{z}$ and then estimating the model parameters via block averages. By the least squares formulation, it is easy to observe the following property.

PROPOSITION 2.1. *For any minimizer $(\hat{Q}, \hat{z})$, the entries of $\hat{Q}$ has representation*

$$\hat{Q}_{ab} = \bar{A}_{ab}(\hat{z}), \tag{2.5}$$

*for all $a, b \in [k]$.*

The representation of the solution (2.5) shows that the estimator (2.4) is essentially doing a histogram approximation after finding the optimal cluster assignment $\hat{z} \in \mathcal{Z}_{n,k}$ according to the least squares criterion (2.3). In the classical nonparametric regression problem, it is known that a simple histogram estimator cannot achieve optimal convergence rate for $\alpha > 1$ [49]. However, we are going to show that this simple histogram estimator achieves optimal rates of convergence under both stochastic block model and nonparametric graphon estimation settings.

Similar estimators using the Bernoulli likelihood function have been proposed and analyzed in the literature [7, 57, 52, 45]. Instead of using the likelihood function of Bernoulli distribution, the least squares estimator (2.3) can be viewed as maximizing Gaussian likelihood. This allows us to obtain optimal convergence rates with cleaner analysis.

2.2. *Stochastic Block Model.* In the stochastic block model setting, each node $i \in [n]$ is associated with a label $a \in [k]$, indicating its cluster. The edge $A_{ij}$ is a Bernoulli random variable with mean $\theta_{ij}$. The value of $\theta_{ij}$ only depends on the clusters of the $i$-th and the $j$-th nodes. We assume $\{\theta_{ij}\}$ is from the following parameter space,

$$\begin{aligned}\Theta_k &= \Big\{ \{\theta_{ij}\} \in [0,1]^{n\times n} : \theta_{ii} = 0,\ \theta_{ij} = Q_{ab} = Q_{ba} \\ &\qquad \text{for } (i,j) \in z^{-1}(a) \times z^{-1}(b) \text{ for some } Q_{ab} \in [0,1] \text{ and } z \in \mathcal{Z}_{n,k} \Big\}.\end{aligned}$$

Namely, the partition function $z$ assigns cluster to each node, and the value of $Q_{ab}$ measures the intensity of link between the $a$-th and the $b$-th clusters. The least squares estimator (2.3) attains the following convergence rate for estimating $\{\theta_{ij}\}$.

THEOREM 2.1. *For any constant $C' > 0$, there is a constant $C > 0$ only depending on $C'$, such that*

$$\frac{1}{n^2} \sum_{i,j\in[n]} (\hat{\theta}_{ij} - \theta_{ij})^2 \leq C \left( \frac{k^2}{n^2} + \frac{\log k}{n} \right),$$

*with probability at least $1 - \exp(-C' n \log k)$, uniformly over $\theta \in \Theta_k$. Furthermore, we have*

$$\sup_{\theta\in\Theta_k} \mathbb{E} \left\{ \frac{1}{n^2} \sum_{i,j\in[n]} (\hat{\theta}_{ij} - \theta_{ij})^2 \right\} \leq C_1 \left( \frac{k^2}{n^2} + \frac{\log k}{n} \right),$$

*for all $k \in [n]$ with some universal constant $C_1 > 0$.*

Theorem 2.1 characterizes different convergence rates for $k$ in different regimes. Suppose $k \asymp n^\delta$ for some $\delta \in [0,1]$. Then the convergence rate in Theorem 2.1 is

$$\frac{k^2}{n^2} + \frac{\log k}{n} \asymp \begin{cases} n^{-2} & k = 1, \\ n^{-1} & \delta = 0, k \geq 2, \\ n^{-1}\log n & \delta \in (0, 1/2], \\ n^{-2(1-\delta)} & \delta \in (1/2, 1]. \end{cases} \tag{2.6}$$

The result completely characterizes the convergence rates for stochastic block model with any possible number of clusters $k$. Depending on whether $k$ is small, moderate, or large, the convergence rates behave differently.

The convergence rate, in terms of $k$, has two parts. The first part $k^2/n^2$ is called the nonparametric rate. It is determined by the number of parameters and the number of observations of the model. For the stochastic block model with $k$ clusters, the number of parameters is $k(k+1)/2 \asymp k^2$ and the number of observations is $n(n+1)/2 \asymp n^2$. The second part $n^{-1}\log k$ is called the clustering rate. Its presence is due to the unknown labels of the $n$ nodes. Our result shows the clustering rate is logarithmically depending on the number of clusters $k$. From (2.6), we observe that when $k$ is small, the clustering rate dominates. When $k$ is large, the nonparametric rate dominates.

To show that the rate in Theorem 2.1 cannot be improved, we obtain the following minimax lower bound.

THEOREM 2.2. *There exists a universal constant $C > 0$, such that*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta_k} \mathbb{P}\left\{ \frac{1}{n^2} \sum_{i,j \in [n]} (\hat{\theta}_{ij} - \theta_{ij})^2 \geq C\left(\frac{k^2}{n^2} + \frac{\log k}{n}\right) \right\} \geq 0.8,$$

*and*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta_k} \mathbb{E}\left\{ \frac{1}{n^2} \sum_{i,j \in [n]} (\hat{\theta}_{ij} - \theta_{ij})^2 \right\} \geq C\left(\frac{k^2}{n^2} + \frac{\log k}{n}\right),$$

*for any $k \in [n]$.*

The upper bound of Theorem 2.1 and the lower bound of Theorem 2.2 immediately imply the minimax rate in Theorem 1.1.

2.3. *Nonparametric Graphon Estimation.* Let us proceed to nonparametric graphon estimation. For any $i \neq j$, $A_{ij}$ is sampled from the following process,

$$(\xi_1, ..., \xi_n) \sim \mathbb{P}_\xi, \quad A_{ij}|(\xi_i, \xi_j) \sim \text{Bernoulli}(\theta_{ij}), \quad \text{where } \theta_{ij} = f(\xi_i, \xi_j).$$

For $i \in [n]$, $A_{ii} = \theta_{ii} = 0$. Conditioning on $(\xi_1, ..., \xi_n)$, $A_{ij}$ is independent across $i, j \in [n]$. To completely specify the model, we need to define the function class of $f$ on $[0,1]^2$. Since $f$ is symmetric, we only need to specify its value on $\mathcal{D} = \{(x,y) \in [0,1]^2 : x \geq y\}$. Define the derivative operator by

$$\nabla_{jk} f(x,y) = \frac{\partial^{j+k}}{(\partial x)^j (\partial y)^k} f(x,y),$$

and we adopt the convention $\nabla_{00} f(x,y) = f(x,y)$. The Hölder norm is defined as

$$||f||_{\mathcal{H}_\alpha} = \max_{j+k \leq \lfloor \alpha \rfloor} \sup_{x,y \in \mathcal{D}} |\nabla_{jk} f(x,y)| + \max_{j+k = \lfloor \alpha \rfloor} \sup_{(x,y) \neq (x',y') \in \mathcal{D}} \frac{|\nabla_{jk} f(x,y) - \nabla_{jk} f(x',y')|}{(|x-x'| + |y-y'|)^{\alpha - \lfloor \alpha \rfloor}}.$$

The Hölder class is defined by

$$\mathcal{H}_\alpha(M) = \left\{ |f||_{\mathcal{H}_\alpha} \leq M : f(x,y) = f(y,x) \text{ for } x \geq y \right\},$$

where $\alpha > 0$ is the smoothness parameter and $M > 0$ is the size of the class, which is assumed to be a constant. When $\alpha \in (0,1]$, a function $f \in \mathcal{H}_\alpha(M)$ satisfies the Lipschitz condition

$$|f(x,y) - f(x',y')| \leq M(|x-x'| + |y-y'|)^\alpha, \tag{2.7}$$

for any $(x,y), (x',y') \in \mathcal{D}$. In the network model, the graphon $f$ is assumed to live in the following class,

$$\mathcal{F}_\alpha(M) = \left\{ 0 \leq f \leq 1 : f \in \mathcal{H}_\alpha(M) \right\}.$$

We have mentioned that the convergence rate of graphon estimation is essentially due to the stochastic block model approximation of $f$ in a Hölder class. This intuition is established by the following lemma, whose proof is given in the supplementary material [15].

LEMMA 2.1. *There exists $z^* \in \mathcal{Z}_{n,k}$, sastisfying,*

$$\frac{1}{n^2} \sum_{a,b \in [k]} \sum_{\{i \neq j : z^*(i) = a, z^*(j) = b\}} \left( \theta_{ij} - \bar{\theta}_{ab}(z^*) \right)^2 \leq CM^2 \left( \frac{1}{k^2} \right)^{\alpha \wedge 1},$$

*for some universal constant $C > 0$.*

The graph limit theory [37] suggests $\mathbb{P}_\xi$ to be an i.i.d. uniform distribution on the interval $[0,1]$. For the estimating procedure (2.3) to work, we allow $\mathbb{P}_\xi$ to be any distribution. The upper bound is attained over all distributions $\mathbb{P}_\xi$ uniformly. Combining Lemma 2.1 and Theorem 2.1 in an appropriate manner, we obtain the convergence rate for graphon estimation by the least squares estimator (2.3).

THEOREM 2.3. *Choose $k = \lceil n^{\frac{1}{\alpha\wedge 1+1}}\rceil$. Then for any $C' > 0$, there exists a constant $C > 0$ only depending on $C'$ and $M$, such that*

$$\frac{1}{n^2}\sum_{i,j\in[n]}(\hat{\theta}_{ij}-\theta_{ij})^2 \le C\left(n^{-\frac{2\alpha}{\alpha+1}}+\frac{\log n}{n}\right),$$

*with probability at least $1-\exp(-C'n)$, uniformly over $f\in\mathcal{F}_\alpha(M)$ and $\mathbb{P}_\xi$. Furthermore,*

$$\sup_{f\in\mathcal{F}_\alpha(M)}\sup_{\mathbb{P}_\xi}\mathbb{E}\left\{\frac{1}{n^2}\sum_{i,j\in[n]}(\hat{\theta}_{ij}-\theta_{ij})^2\right\} \le C_1\left(n^{-\frac{2\alpha}{\alpha+1}}+\frac{\log n}{n}\right),$$

*for some other constant $C_1 > 0$ only depending on $M$. Both the probability and the expectation are jointly over $\{A_{ij}\}$ and $\{\xi_i\}$.*

Similar to Theorem 2.1, the convergence rate of Theorem 2.3 has two parts. The nonparametric rate $n^{-\frac{2\alpha}{\alpha+1}}$, and the clustering rate $n^{-1}\log n$. Note that the clustering rates in both theorems are identical because $n^{-1}\log n \asymp n^{-1}\log k$ under the choice $k=\lceil n^{\frac{1}{\alpha\wedge 1+1}}\rceil$. An interesting phenomenon to note is that the smoothness index $\alpha$ only plays a role in the regime $\alpha\in(0,1)$. The convergence rate is always dominated by $n^{-1}\log n$ when $\alpha\ge 1$.

In order to show the rate of Theorem 2.3 is optimal, we need a lower bound over the class $\mathcal{F}_\alpha(M)$ and over all $\mathbb{P}_\xi$. To be specific, we need to show

$$\inf_{\hat{\theta}}\sup_{f\in\mathcal{F}_\alpha(M)}\sup_{\mathbb{P}_\xi}\mathbb{E}\left\{\frac{1}{n^2}\sum_{i,j\in[n]}(\hat{\theta}_{ij}-\theta_{ij})^2\right\} \ge C\left(n^{-\frac{2\alpha}{\alpha+1}}+\frac{\log n}{n}\right), \tag{2.8}$$

for some constant $C>0$. In fact, the lower bound we obtained is stronger than (2.8) in the sense that it holds for a subset of the space of probabilities on $\{\xi_i\}$. The subset $\mathcal{P}$ requires the sampling points $\{\xi_i\}$ to well cover the interval $[0,1]$ for $\{f(\xi_i,\xi_j)\}_{i,j\in[n]}$ to be good representatives of the whole function $f$. For each $a\in[k]$, define the interval

$$U_a = \left[\frac{a-1}{k},\frac{a}{k}\right). \tag{2.9}$$

We define the distribution class by

$$\mathcal{P} = \left\{ \mathbb{P}_\xi : \mathbb{P}_\xi \left( \frac{\lambda_1 n}{k} \leq \sum_{i=1}^n \mathbb{I}\{\xi_i \in U_a\} \leq \frac{\lambda_2 n}{k}, \text{ for any } a \in [k] \right) > 1 - \exp(-n^\delta) \right\},$$

for some positive constants $\lambda_1, \lambda_2$ and some arbitrary small constant $\delta \in (0,1)$. Namely, for each interval $U_a$, it contains roughly $n/k$ observations. By applying standard concentration inequality, it can be shown that the i.i.d. uniform distribution on $\{\xi_i\}$ belongs to the class $\mathcal{P}$.

THEOREM 2.4. *There exists a constant $C > 0$ only depending on $M, \alpha$, such that*

$$\inf_{\hat{\theta}} \sup_{f \in \mathcal{F}_\alpha(M)} \sup_{\mathbb{P}_\xi \in \mathcal{P}} \mathbb{P} \left\{ \frac{1}{n^2} \sum_{i,j \in [n]} (\hat{\theta}_{ij} - \theta_{ij})^2 \geq C \left( n^{-\frac{2\alpha}{\alpha+1}} + \frac{\log n}{n} \right) \right\} \geq 0.8,$$

*and*

$$\inf_{\hat{\theta}} \sup_{f \in \mathcal{F}_\alpha(M)} \sup_{\mathbb{P}_\xi \in \mathcal{P}} \mathbb{E} \left\{ \frac{1}{n^2} \sum_{i,j \in [n]} (\hat{\theta}_{ij} - \theta_{ij})^2 \right\} \geq C \left( n^{-\frac{2\alpha}{\alpha+1}} + \frac{\log n}{n} \right),$$

*where the probability and expectation are jointly over $\{A_{ij}\}$ and $\{\xi_i\}$.*

The proof of Theorem 2.4 is given in the supplementary material [15]. The minimax rate in Theorem 1.2 is an immediate consequence of Theorem 2.3 and Theorem 2.4.

## 3. Discussion.

3.1. *More General Models.* The results in this paper assume symmetry on the graphon $f$ and the matrix $\{\theta_{ij}\}$. Such assumption is naturally made in the context of network analysis. However, these results also hold under more general models. We may consider a slightly more general version of (1.1) as

$$\theta_{ij} = f(\xi_i, \eta_j), \quad 1 \leq i, j \leq n,$$

with $\{\xi_i\}$ and $\{\eta_j\}$ sampled from $\mathbb{P}_\xi$ and $\mathbb{P}_\eta$ respectively, and the function $f$ is not necessarily symmetric. To be specific, let us redefine the Hölder norm $||\cdot||_{\mathcal{H}_\alpha}$ by replacing $\mathcal{D}$ with $[0,1]^2$ in its original definition in Section 2.3. Then, we consider the function class

$$\mathcal{F}'_\alpha(M) = \{0 \leq f \leq 1 : ||f||_{\mathcal{H}_\alpha} \leq M\}.$$

The minimax rate for this class is stated in the following theorem without proof.

THEOREM 3.1. *Consider the function class $\mathcal{F}'_\alpha(M)$ with $\alpha > 0$ and $M > 0$. We have*

$$\inf_{\hat{\theta}} \sup_{f\in\mathcal{F}'_\alpha(M)} \sup_{\substack{\xi\sim\mathbb{P}_\xi \\ \eta\sim\mathbb{P}_\eta}} \mathbb{E}\left\{\frac{1}{n^2}\sum_{i,j\in[n]}(\hat{\theta}_{ij}-\theta_{ij})^2\right\} \asymp \begin{cases} n^{-\frac{2\alpha}{\alpha+1}}, & 0<\alpha<1, \\ \frac{\log n}{n}, & \alpha\geq 1, \end{cases}$$

*where the expectation is jointly over $\{A_{ij}\}$, $\{\xi_i\}$ and $\{\eta_j\}$.*

Similarly, we may generalize the stochastic block model by the parameter space

$$\Theta_{kl}^{asym} = \Big\{\{\theta_{ij}\}\in[0,1]^{n\times m}: \ \theta_{ij}=Q_{ab} \text{ for } (i,j)\in z_1^{-1}(a)\times z_2^{-1}(b)$$
$$\text{with some } Q_{ab}\in[0,1],\ z_1\in\mathcal{Z}_{n,k} \text{ and } z_2\in\mathcal{Z}_{m,l}\Big\}.$$

Such model naturally arises in the contexts of biclustering [25, 41, 11, 39] and matrix organization [18, 13, 17], where symmetry of the model is not assumed. Under such extension, we can show that a similar minimax rate as in Theorem 1.1 as follows.

THEOREM 3.2. *Consider the parameter space $\Theta_{kl}^{asym}$ and assume $\log k \asymp \log l$. We have*

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta_{kl}^{asym}} \mathbb{E}\left\{\frac{1}{nm}\sum_{\substack{i\in[n]\\ j\in[m]}}(\hat{\theta}_{ij}-\theta_{ij})^2\right\} \asymp \frac{kl}{nm}+\frac{\log k}{m}+\frac{\log l}{n},$$

*for any $1\leq k\leq n$ and $1\leq l\leq m$.*

The lower bounds of Theorem 3.1 and Theorem 3.2 are directly implied by viewing the symmetric parameter spaces as subsets of the asymmetric ones. For the upper bound, we propose a modification of the least squares estimator in Section 2.1. Consider the criterion function

$$L^{asym}(Q,z_1,z_2)=\sum_{(a,b)\in[k]\times[l]}\ \sum_{(i,j)\in z_1^{-1}(a)\times z_2^{-1}(b)}(A_{ij}-Q_{ab})^2.$$

For any $(\hat{Q},\hat{z}_1,\hat{z}_2)\in\operatorname{argmin}_{Q\in\mathbb{R}^{k\times l},z_1\in\mathcal{Z}_{n,k},z_2\in\mathcal{Z}_{m,l}} L(Q,z_1,z_2)$, define the estimator of $\theta_{ij}$ by

$$\hat{\theta}_{ij}=\hat{Q}_{\hat{z}_1(i)\hat{z}_2(j)}, \quad \text{for all } (i,j)\in[n]\times[m].$$

Using the same proofs of Theorem 2.1 and Theorem 2.3, we can obtain the upper bounds.

3.2. *Nonparametric Regression without Knowing Design.* The graphon estimation problem is closely related to the classical nonparametric regression problem. This section explores their connections and differences to bring better understandings of both problems. Namely, we study the problem of nonparametric regression without observing the design. First, let us consider the one-dimensional regression problem

$$y_i = f(\xi_i) + z_i, \quad i \in [n],$$

where $\{\xi_i\}$ are sampled from some $\mathbb{P}_\xi$, and $z_i$ are i.i.d. $N(0,1)$ variables. A nonparametric function estimator $\hat{f}$ estimates the function $f$ from the pairs $\{(\xi_i, y_i)\}$. For Hölder class with smoothness $\alpha$, the minimax rate under the loss $\frac{1}{n}\sum_{i\in[n]}\left(\hat{f}(\xi_i) - f(\xi_i)\right)^2$ is at the order of $n^{-\frac{2\alpha}{2\alpha+1}}$ [49]. However, when the design $\{\xi_i\}$ is not observed, the minimax rate is at a constant order. To see this fact, let us consider a closely related problem

$$y_i = \theta_i + z_i, \quad i \in [n],$$

where we assume $\theta \in \Theta_2$. The parameter space $\Theta_2$ is defined as a subset of $[0,1]^n$ with $\{\theta_i\}$ that can only take two possible values $q_1$ and $q_2$. It can be viewed as a one-dimensional version of stochastic block model. We can show that

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta_2} \mathbb{E}\left\{\frac{1}{n}\sum_{i\in[n]}(\hat{\theta}_i - \theta_i)^2\right\} \asymp 1.$$

The upper bound is achieved by letting $\hat{\theta}_i = y_i$ for each $i \in [n]$. To see the lower bound, we may fix $q_1 = 1/4$ and $q_2 = 1/2$. Then the problem is reduced to $n$ independent two-point testing problems between $N(1/4, 1)$ and $N(1/2, 1)$ for each $i \in [n]$. It is easy to see that each testing problem contributes to an error at the order of a constant, which gives the lower bound of a constant order. This leads to a constant lower bound for the original regression problem by using the embedding technique in the proof of Theorem 2.4, which shows that $\Theta_2$ is a smaller space than a Hölder class on a subset of $[n]$. Thus, 1 is also a lower bound for the regression problem without knowing design.

In contrast to the one-dimensional problem, we can show that a two-dimensional nonparametric regression without knowing design is more informative. Consider

$$y_{ij} = f(\xi_i, \xi_j) + z_{ij}, \quad i, j \in [n],$$

where $\{\xi_i\}$ are sampled from some $\mathbb{P}_\xi$, and $z_{ij}$ are i.i.d. $N(0,1)$ variables. Let us consider the Hölder class $\mathcal{H}'_\alpha(M) = \{f : ||f||_{\mathcal{H}_\alpha} \leq M\}$ with Hölder norm $||\cdot||_{\mathcal{H}_\alpha}$ defined in Section 3.1. When the design $\{\xi_i\}$ is known, the minimax rate under the loss $\frac{1}{n^2}\sum_{i,j\in[n]}\left(\hat{f}(\xi_i,\xi_j) - f(\xi_i,\xi_j)\right)^2$ is at the order of $n^{-\frac{2\alpha}{\alpha+1}}$. When the design is unknown, the minimax rate is stated in the following theorem.

THEOREM 3.3. *Consider the Hölder class $\mathcal{H}'_\alpha(M)$ for $\alpha > 0$ and $M > 0$. We have*

$$\inf_{\hat{f}} \sup_{f\in\mathcal{H}'_\alpha(M)} \sup_{\mathbb{P}_\xi} \mathbb{E}\left\{\frac{1}{n^2}\sum_{i,j\in[n]}\left(\hat{f}(\xi_i,\xi_j) - f(\xi_i,\xi_j)\right)^2\right\} \asymp \begin{cases} n^{-\frac{2\alpha}{\alpha+1}}, & 0<\alpha<1, \\ \frac{\log n}{n}, & \alpha \geq 1, \end{cases}$$

*where the expectation is jointly over $\{A_{ij}\}$ and $\{\xi_i\}$.*

The minimax rate is identical to that of Theorem 1.2, which demonstrates the close relation between nonparametric graphon estimation and nonparametric regression without knowing design. The proof of this result is similar to the proofs of Theorem 2.3 and Theorem 2.4, and is omitted in the paper. One simply needs to replace the Bernoulli analysis by the corresponding Gaussian analysis in the proof. Compared with the rate for one-dimensional regression without knowing design, the two-dimensional minimax rate is more interesting. It shows that the ignorance of design only matters when $\alpha \geq 1$. For $\alpha \in (0,1)$, the rate is exactly the same as the case when the design is known.

The main reason for the difference between the one-dimensional and the two-dimensional problems is that the form of $\{(\xi_i,\xi_j)\}$ implicitly imposes more structure. To illustrate this point, let us consider the following two-dimensional problem

$$y_{ij} = f(\xi_{ij}) + z_{ij}, \quad i,j \in [n],$$

where $\xi_{ij} \in [0,1]^2$ and $\{\xi_{ij}\}$ are sampled from some distribution. It is easy to see that this is equivalent to the one-dimensional problem with $n^2$ observations and the minimax rate is at the order of a constant. The form $\{(\xi_i,\xi_j)\}$ implies that the lack of identifiability caused by the ignorance of design is only resulted from row permutation and column permutation, and thus it is more informative than the design $\{\xi_{ij}\}$.

3.3. *Lower Bound for Finite $k$.* A key contribution of the paper lies in the proof of Theorem 2.2, where we establish the lower bound $k^2/n^2 + n^{-1}\log k$ (especially the $n^{-1}\log k$ part) via a novel construction. To better understand the main idea behind the construction, we present the analysis for a finite $k$ in this section. When $2 \leq k \leq O(1)$, the minimax rate becomes $n^{-1}$. To prove this lower bound, it is sufficient to consider the parameter space $\Theta_k$ with $k = 2$. Let us define

$$Q = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} + \frac{c}{\sqrt{n}} \\ \frac{1}{2} + \frac{c}{\sqrt{n}} & \frac{1}{2} \end{bmatrix},$$

for some $c > 0$ to be determined later. Define the subspace

$$T = \left\{\{\theta_{ij}\} \in [0,1]^{n\times n} : \theta_{ij} = Q_{z(i)z(j)} \text{ for some } z \in \mathcal{Z}_{n,2}\right\}.$$

It is easy to see that $T \subset \Theta_2$. With a fixed $Q$, the set $T$ has a one-to-one correspondence with $\mathcal{Z}_{n,2}$. Let us define the collection of subsets $\mathcal{S} = \{S : S \subset [n]\}$. For any $z \in \mathcal{Z}_{n,2}$, it induces a partition $\{z^{-1}(1), z^{-1}(2)\}$ on the set $[n]$. This corresponds to $\{S, S^c\}$ for some $S \in \mathcal{S}$. With this observation, we may rewrite $T$ as

$$\begin{aligned} T \;=\; & \Bigg\{\{\theta_{ij}\} \in [0,1]^{n\times n} : \theta_{ij} = \frac{1}{2} \text{ for } (i,j) \in (S\times S)\cup(S^c\times S^c), \\ & \theta_{ij} = \frac{1}{2} + \frac{c}{\sqrt{n}} \text{ for } (i,j) \in (S\times S^c)\cup(S^c\times S), \text{ with some } S \in \mathcal{S}\Bigg\}. \end{aligned}$$

The subspace $T$ characterizes the difficulty of the problem due to the ignorance of the clustering structure $\{S, S^c\}$ of the $n$ nodes. Such difficulty is central in the estimation problem of network analysis. We are going to use Fano's lemma (Proposition 4.1) to lower bound the risk. Then, it is sufficient to upper bound the KL diameter $\sup_{\theta,\theta'\in T} D(\mathbb{P}_\theta||\mathbb{P}_{\theta'})$ and lower bound the packing number $\mathcal{M}(\epsilon, T, \rho)$ for some appropriate $\epsilon$ and the metric $\rho(\theta,\theta') = n^{-1}||\theta-\theta'||$. Using Proposition 4.2, we have

$$\sup_{\theta,\theta'\in T} D(\mathbb{P}_\theta||\mathbb{P}_{\theta'}) \leq \sup_{\theta,\theta'\in T} 8||\theta-\theta'||^2 \leq 8c^2 n.$$

To obtain a lower bound for $\mathcal{M}(\epsilon, T, \rho)$, note that for $\theta, \theta' \in T$ associated with $S, S' \in \mathcal{S}$, we have

$$n^2\rho^2(\theta,\theta') = \frac{2c^2}{n}|S\Delta S'|\left(n - |S\Delta S'|\right),$$

where $A\Delta B$ is the symmetric difference defined as $(A \cap B^c) \cup (A^c \cap B)$. By viewing $|S\Delta S'|$ as the Hamming distance of the corresponding indicator functions of the sets, we can use the Varshamov-Gilbert bound (Lemma 4.5) to pick $S_1, ..., S_N \subset \mathcal{S}$ satisfying

$$\frac{1}{4}n \leq |S_i \Delta S_j| \leq \frac{3}{4}n, \quad \text{for } i \neq j \in [N],$$

with $N \geq \exp(c_1 n)$, for some $c_1 > 0$. Hence, we have

$$\mathcal{M}(\epsilon, T, \rho) \geq N \geq \exp(c_1 n), \quad \text{with } \epsilon^2 = \frac{c^2}{8n}.$$

Applying (4.9) of Proposition 4.1, we have

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta_2} \mathbb{P}\left\{ \frac{1}{n^2} \sum_{i,j \in [n]} (\hat{\theta}_{ij} - \theta_{ij})^2 \geq \frac{c^2}{32n} \right\} \geq 1 - \frac{8c^2 n + \log 2}{c_1 n} \geq 0.8,$$

where the last inequality holds by choosing a sufficiently small $c$. Note that the above derivation ignores the fact that $\theta_{ii} = 0$ for $i \in [n]$ for the sake of clear presentation. The argument can be easily made rigorous with slight modification. Thus, we prove the lower bound for a finite $k$. For $k$ growing with $n$, a more delicate construction is stated in Section 4.2.

3.4. *Application to Link Prediction.* An important application of Theorem 2.1 and Theorem 2.3 is link prediction. The link prediction or the network completion problem [21, 38, 56] has practical significances. Instead of observing the whole adjacency matrix, we observe $\{A_{ij} : (i,j) \in \Omega\}$ for some $\Omega \subset [n] \times [n]$. The goal is to infer the unobserved edges. One example is the biological network. Scientific study showed that only 80% of the molecular interactions in cells of Yeast are known [54]. Accurate prediction of those unseen interactions can greatly reduce the costs of biological experiments. To tackle the problem of link prediction, we consider a modification of the constrained least square program, which is defined as

$$\min \ ||\theta||^2 - \frac{2n^2}{|\Omega|} \sum_{(i,j) \in \Omega} A_{ij}\theta_{ij}, \quad \text{s.t. } \theta \in \Theta_k. \tag{3.1}$$

The estimator $\hat{\theta}$ obtained from solving (3.1) takes advantage of the underlying block structure of the network, and is an extension to (2.3). The number $\hat{\theta}_{ij}$ can be interpreted as how likely there is an edge between $i$ and $j$. To analyze the theoretical performance of (3.1), let us assume the set $\Omega$ is obtained by uniformly sampling with replacement from all edges. In other words, $\Omega$ may contain some repeated elements.

THEOREM 3.4. *Assume $|\Omega|/n^2 \geq c$ for a constant $c \in (0,1]$. For any constant $C' > 0$, there exists some constant $C > 0$ only depending on $C'$ and $c$ such that*

$$\frac{1}{n^2}\sum_{i,j\in[n]}(\hat{\theta}_{ij}-\theta_{ij})^2 \leq C\left(\frac{k^2}{n^2}+\frac{\log k}{n}\right),$$

*with probability at least $1-\exp(-C'n\log k)$ uniformly over $\theta \in \Theta_k$ for all $k \in [n]$.*

The result of Theorem 3.4 assumes $|\Omega|/n^2 \geq c$. For example, when $|\Omega|/n^2 = 1/2$, we only observe at most half of the edges. Theorem 3.4 gives rate-optimal link prediction of the rest of the edges. In contrast, the low-rank matrix completion approach, though extensively studied and applied in literature, only gives a rate $k/n$, which is inferior to that of Theorem 3.4.

In the case where the assumption of stochastic block model is not natural [48], we may consider a more general class of networks generated by a smooth graphon. This is also a useful assumption to do link prediction. Using the same estimator (3.1) with $k = \lceil n^{\frac{1}{\alpha\wedge 1+1}} \rceil$, we can obtain the error

$$\frac{1}{n^2}\sum_{i,j\in[n]}(\hat{\theta}_{ij}-\theta_{ij})^2 \leq C\left(n^{-\frac{2\alpha}{2\alpha+1}}+\frac{\log n}{n}\right),$$

with probability at least $1-\exp(-C'n)$ uniformly over $f \in \mathcal{F}_\alpha(M)$ and $\mathbb{P}_\xi$, which extends Theorem 2.3. The proof of Theorem 3.4 is nearly identical to that of Theorem 2.1 and is omitted in the paper.

3.5. *Minimax Rate for Operator Norm.* The minimax rates in the paper are all studied under the $\ell_2$ norm, which is the Frobenius norm for a matrix. It is also interesting to investigate the minimax rate under the matrix operator norm. Recall that for a matrix $U$, its operator norm $\|U\|_{\text{op}}$ is the largest singular value.

THEOREM 3.5. *For the stochastic block model $\Theta_k$ with $k \geq 2$, we have*

$$\inf_{\hat{\theta}}\sup_{\theta\in\Theta_k}\mathbb{E}\|\hat{\theta}-\theta\|_{\text{op}}^2 \asymp n.$$

Interestingly, the result of Theorem 3.5 does not depend on $k$ as long as $k \geq 2$. The optimal estimator is the adjacency matrix itself $\hat{\theta} = A$, whose bound under the operator norm can be derived from standard random matrix theory [50]. The lower bound is directly implied from Theorem 2.2 by the following argument,

$$\begin{aligned}\inf_{\hat\theta}\sup_{\theta\in\Theta_k}\mathbb{E}\|\hat\theta-\theta\|_{\mathrm{op}}^2 &\gtrsim \inf_{\hat\theta}\sup_{\theta\in\Theta_2}\mathbb{E}\|\hat\theta-\theta\|_{\mathrm{op}}^2\\ &\gtrsim \inf_{\hat\theta\in\Theta_2}\sup_{\theta\in\Theta_2}\mathbb{E}\|\hat\theta-\theta\|_{\mathrm{op}}^2 \gtrsim \inf_{\hat\theta}\sup_{\theta\in\Theta_2}\mathbb{E}||\hat\theta-\theta||^2. \end{aligned}\tag{3.2}$$

The first inequality is because $\Theta_2$ is a smaller model than $\Theta_k$ for $k \geq 2$. The second inequality is because of the fact that we can always project the estimator into the parameter space without compromising the convergence rate. Then, for $\hat\theta, \theta \in \Theta_2$, $\hat\theta - \theta$ is a matrix with rank at most 4, and we have the inequality $||\hat\theta-\theta||^2 \leq 4\|\hat\theta-\theta\|_{\mathrm{op}}^2$, which gives the last inequality. Finally, $\inf_{\hat\theta}\sup_{\theta\in\Theta_2}\mathbb{E}||\hat\theta-\theta||^2 \gtrsim n$ by Theorem 2.2 implies the desired conclusion.

Theorem 3.5 suggests that estimating $\theta$ under the operator norm is not a very interesting problem, because the estimator does not need to take advantage of the structure of the space $\Theta_k$. Due to recent advances in community detection, a more suitable parameter space for the problem is $\Theta(\beta)\cap\Theta_k$, where

$$\Theta(\beta) = \left\{\theta = \theta^T = \{\theta_{ij}\} \in [0,1]^{n\times n} : \theta_{ii} = 0, \max_{ij}\theta_{ij} \leq \beta\right\}.$$

The parameter $\beta$ is understood to be the sparsity of the network because a smaller $\beta$ leads to less edges of the graph.

THEOREM 3.6. *For $n^{-1} \leq \beta \leq 1$ and $k \geq 2$, we have*

$$\inf_{\hat\theta}\sup_{\theta\in\Theta(\beta)\cap\Theta_k}\mathbb{E}\|\hat\theta-\theta\|_{\mathrm{op}}^2 \asymp \inf_{\hat\theta}\sup_{\theta\in\Theta(\beta)}\mathbb{E}\|\hat\theta-\theta\|_{\mathrm{op}}^2 \asymp \beta n.$$

The lower bound of Theorem 3.6 can be obtained in a similar way by combining the argument in (3.2) and a modified version of Theorem 2.2 (see the supplementary material [15]). When $\beta \geq n^{-1}\log n$, the upper bound is still achieved by the adjacency matrix, as is proved in Theorem 5.2 of [34]. For $n^{-1} \leq \beta < n^{-1}\log n$, one needs to replace the rows and columns that have high degrees by zeros in $A$, and the upper bound is achieved by this trimmed adjacency matrix. This is recently established in [12].

3.6. *Relation to Community Detection.* Community detection is another important problem in network analysis. The parameter estimation result established in this paper has some consequences in community detection, especially for the results under the operator norm in Theorem 3.5 and Theorem

3.6. Recent works in community detection [34, 12] show that the bound for $\|\hat{\theta}-\theta\|_{\mathrm{op}}^2$ can be used to derive the misclassification error of spectral clustering algorithm applied on the matrix $\hat{\theta}$. Recall that the spectral clustering algorithm applies $k$-means to the leading singular vectors of the matrix $\hat{\theta}$. Theorem 3.5 justifies the use of adjacency matrix as $\hat{\theta}$ in spectral clustering because of its minimax optimality under the operator norm. Moreover, when the network is in a sparse regime with $n^{-1} \leq \beta < n^{-1}\log n$, [12] suggests to use the trimmed adjacency matrix as $\hat{\theta}$ for spectral clustering. According to Theorem 3.6, the trimmed adjacency matrix is an optimal estimator of $\theta$ under the operator norm.

On the other hand, the connection between the minimax rates under the $\ell_2$ norm and community detection is not that close. We illustrate this point by the case when $k=2$. Let us consider $\theta \in \Theta_2$, then $\theta_{ij} = Q_{z(i)z(j)}$ for some $2\times 2$ symmetric matrix $Q$ and $z$ is the label function. Suppose the within community connection probability is greater than the between community connection probability by a margin of $s$. Namely, assume $Q_{11}\wedge Q_{22} - Q_{12} \geq s > 0$. Then, for the estimator $\hat{\theta}_{ij} = \hat{Q}_{\hat{z}(i)\hat{z}(j)}$ with error $\frac{1}{n^2}\sum_{i,j\in[n]}(\hat{\theta}_{ij}-\theta_{ij})^2 \leq \epsilon^2$, the number of mis-clustered nodes under $\hat{z}$ is roughly bounded by $O\left((n\epsilon/s)^2\right)$. This is because when two nodes that have the same labels under $z$ are clustered into different communities or when two nodes belong to different communities are clustered into the same one, an estimation error of $O(s^2)$ must occur. Conversely, bounds on community detection can lead to an improved bound for parameter estimation. Specifically, when $\left(\sqrt{Q_{11}\wedge Q_{22}}-\sqrt{Q_{12}}\right)^2 > 2n^{-1}\log n$ and $|z^{-1}(1)| = |z^{-1}(2)| = n/2$, [42, 23] show that there exists a strongly consistent estimator of $z$ in the sense that the misclassification error is 0 with high probability. In this case, the estimation error of $\theta$ under the loss $\frac{1}{n^2}\sum_{i,j\in[n]}(\hat{\theta}_{ij}-\theta_{ij})^2$ can be improved to $n^{-2}$ from $n^{-1}$.

Generally, parameter estimation and community detection are different problems of network analysis. When $\{Q_{ab}\}_{a,b\in[k]}$ all take the same value, it is impossible to do community detection, but parameter estimation would be easy. Thus, good parameter estimation result does not necessarily imply consistent community detection. General minimax rates of the community detection problem are recently established in [55, 16].

## 4. Proofs.

We present the proofs of the main results in this section. The upper bounds Theorem 2.1 and Theorem 2.3 are proved in Section 4.1. The lower bound Theorem 2.2 is proved in Section 4.2.

4.1. *Proofs of Theorem 2.1 and Theorem 2.3.* This section is devoted to proving the upper bounds. We first prove Theorem 2.1 and then prove Theorem 2.3.

Let us first give an outline of the proof of Theorem 2.1. In the definition of the class $\Theta_k$, we denote the true value on each block by $\{Q^*_{ab}\} \in [0,1]^{k\times k}$ and the oracle assignment by $z^* \in \mathcal{Z}_{n,k}$ such that $\theta_{ij} = Q^*_{z^*(i)z^*(j)}$ for any $i \neq j$. To facilitate the proof, we introduce the following notation. For the estimated $\hat{z}$, define $\{\tilde{Q}_{ab}\} \in [0,1]^{k\times k}$ by $\tilde{Q}_{ab} = \bar{\theta}_{ab}(\hat{z})$, and also define $\tilde{\theta}_{ij} = \tilde{Q}_{\hat{z}(i)\hat{z}(j)}$ for any $i \neq j$. The diagonal elements $\{\tilde{\theta}_{ii}\}$ are defined as zero for all $i \in [n]$. By the definition of the estimator (2.3), we have

$$L(\hat{Q}, \hat{z}) \leq L(Q^*, z^*),$$

which can be rewritten as

$$||\hat{\theta} - A||^2 \leq ||\theta - A||^2. \tag{4.1}$$

The left side of (4.1) can be decomposed as

$$||\hat{\theta} - \theta||^2 + 2\left\langle \hat{\theta} - \theta, \theta - A \right\rangle + ||\theta - A||^2. \tag{4.2}$$

Combining (4.1) and (4.2), we have

$$||\hat{\theta} - \theta||^2 \leq 2\left\langle \hat{\theta} - \theta, A - \theta \right\rangle. \tag{4.3}$$

The right side of (4.3) can be bounded as

$$\begin{aligned}
\left\langle \hat{\theta} - \theta, A - \theta \right\rangle &= \left\langle \hat{\theta} - \tilde{\theta}, A - \theta \right\rangle + \left\langle \tilde{\theta} - \theta, A - \theta \right\rangle \\
&\leq ||\hat{\theta} - \tilde{\theta}|| \left| \left\langle \frac{\hat{\theta} - \tilde{\theta}}{||\hat{\theta} - \tilde{\theta}||}, A - \theta \right\rangle \right| \qquad (4.4) \\
&\quad + \left( ||\tilde{\theta} - \hat{\theta}|| + ||\hat{\theta} - \theta|| \right) \left| \left\langle \frac{\tilde{\theta} - \theta}{||\tilde{\theta} - \theta||}, A - \theta \right\rangle \right|. \qquad (4.5)
\end{aligned}$$

Using Lemma 4.1-4.3, the following three terms

$$||\hat{\theta} - \tilde{\theta}||, \quad \left| \left\langle \frac{\hat{\theta} - \tilde{\theta}}{||\hat{\theta} - \tilde{\theta}||}, A - \theta \right\rangle \right|, \quad \left| \left\langle \frac{\tilde{\theta} - \theta}{||\tilde{\theta} - \theta||}, A - \theta \right\rangle \right| \tag{4.6}$$

can all be bounded by $C\sqrt{k^2 + n\log k}$ with probability at least $1 - 3\exp(-C'n\log k)$. Combining these bounds with (4.4), (4.5) and (4.3), we get

$$||\hat{\theta} - \theta||^2 \leq C_1\left(k^2 + k\log n\right),$$

with probability at least $1 - 3\exp(-C'n \log k)$. This gives the conclusion of Theorem 2.1. The details of the proof is stated in the later part of the section. To prove Theorem 2.3, we use Lemma 2.1 to approximate the nonparametric graphon by the stochastic block model. With similar arguments above, we get

$$||\hat{\theta} - \theta||^2 \leq C_2 \left(k^2 + k \log n + n^2 k^{-2(\alpha\wedge 1)}\right),$$

with high probability. Choosing the best $k$ gives the conclusion of Theorem 2.3.

Before stating the complete proofs, let us first present the following lemmas, which bound the three terms in (4.6), respectively. The proofs of the lemmas will be given in the supplementary material [15].

LEMMA 4.1. *For any constant $C' > 0$, there exists a constant $C > 0$ only depending on $C'$, such that*

$$||\hat{\theta} - \tilde{\theta}|| \leq C\sqrt{k^2 + n \log k},$$

*with probability at least $1 - \exp(-C'n \log k)$.*

LEMMA 4.2. *For any constant $C' > 0$, there exists a constant $C > 0$ only depending on $C'$, such that*

$$\left|\left\langle \frac{\tilde{\theta} - \theta}{||\tilde{\theta} - \theta||}, A - \theta \right\rangle\right| \leq C\sqrt{n \log k},$$

*with probability at least $1 - \exp(-C'n \log k)$.*

LEMMA 4.3. *For any constant $C' > 0$, there exists a constant $C > 0$ only depending on $C'$, such that*

$$\left|\left\langle \frac{\hat{\theta} - \tilde{\theta}}{||\hat{\theta} - \tilde{\theta}||}, A - \theta \right\rangle\right| \leq C\sqrt{k^2 + n \log k},$$

*with probability at least $1 - \exp(-C'n \log k)$.*

PROOF OF THEOREM 2.1. Combining the bounds for (4.6) with (4.4), (4.5) and (4.3), we have

$$||\hat{\theta} - \theta||^2 \leq 2C||\hat{\theta} - \theta||\sqrt{k^2 + n \log k} + 4C^2\left(k^2 + n \log k\right),$$

with probability at least $1 - 3\exp(-C'n \log k)$. Solving the above equation, we get

$$||\hat{\theta} - \theta||^2 \leq C_1\left(k^2 + n \log k\right),$$

with probability at least $1-3\exp(-C'n\log k)$. This proves the high probability bound. To get the bound in expectation, we use the following inequality

$$\begin{aligned}
& \mathbb{E}n^{-2}||\hat{\theta}-\theta||^2 \\
\leq\ & \mathbb{E}\left(n^{-2}||\hat{\theta}-\theta||^2\mathbb{I}\{n^{-2}||\hat{\theta}-\theta||^2\leq\epsilon^2\}\right)+\mathbb{E}\left(n^{-2}||\hat{\theta}-\theta||^2\mathbb{I}\{n^{-2}||\hat{\theta}-\theta||^2>\epsilon^2\}\right) \\
\leq\ & \epsilon^2+\mathbb{P}\left(n^{-2}||\hat{\theta}-\theta||^2>\epsilon^2\right)\leq\epsilon^2+3\exp(-C'n\log k),
\end{aligned}$$

where $\epsilon^2=C_1\left(\frac{k^2}{n^2}+\frac{\log k}{n}\right)$. Since $\epsilon^2$ is the dominating term, the proof is complete. □

To prove Theorem 2.3, we need to redefine $z^*$ and $Q^*$. We choose $z^*$ to be the one used in Lemma 2.1, which implies a good approximation of $\{\theta_{ij}\}$ by the stochastic block model. With this $z^*$, define $Q^*$ by letting $Q^*_{ab}=\bar{\theta}_{ab}(z^*)$ for any $a,b\in[k]$. Finally, we define $\theta^*_{ij}=Q^*_{z^*(i)z^*(j)}$ for all $i\neq j$. The diagonal elements $\theta^*_{ii}$ are set as zero for all $i\in[n]$. Note that for the stochastic block model, we have $\theta=\theta^*$. The proof of Theorem 2.3 requires another lemma.

LEMMA 4.4. *For any constant $C'>0$, there exists a constant $C>0$ only depending on $C'$, such that*

$$\left|\left\langle\frac{\tilde{\theta}-\theta^*}{||\tilde{\theta}-\theta^*||},A-\theta\right\rangle\right|\leq C\sqrt{n\log k},$$

*with probability at least $1-\exp(-C'n\log k)$.*

The proof of Lemma 4.4 is identical to the proof of Lemma 4.2, and will be omitted in the paper.

PROOF OF THEOREM 2.3. Using the similar argument as outlined in the beginning of this section, we get

$$||\hat{\theta}-\theta^*||^2\leq 2\left\langle\hat{\theta}-\theta^*,A-\theta^*\right\rangle,$$

whose right side can be bounded as

$$\begin{aligned}
& \left\langle\hat{\theta}-\theta^*,A-\theta^*\right\rangle \\
=\ & \left\langle\hat{\theta}-\tilde{\theta},A-\theta\right\rangle+\left\langle\tilde{\theta}-\theta^*,A-\theta\right\rangle+\left\langle\hat{\theta}-\theta^*,\theta-\theta^*\right\rangle \\
\leq\ & ||\hat{\theta}-\tilde{\theta}||\left|\left\langle\frac{\hat{\theta}-\tilde{\theta}}{||\hat{\theta}-\tilde{\theta}||},A-\theta\right\rangle\right|+\left(||\tilde{\theta}-\hat{\theta}||+||\hat{\theta}-\theta^*||\right)\left|\left\langle\frac{\tilde{\theta}-\theta^*}{||\tilde{\theta}-\theta^*||},A-\theta\right\rangle\right| \\
& +||\hat{\theta}-\theta^*||||\theta-\theta^*||
\end{aligned}$$

To better organize what we have obtained, let us introduce the notation

$$L = ||\hat{\theta} - \theta^*||, \quad R = ||\tilde{\theta} - \hat{\theta}||, \quad B = ||\theta - \theta^*||,$$

$$E = \left| \left\langle \frac{\hat{\theta} - \tilde{\theta}}{||\hat{\theta} - \tilde{\theta}||}, A - \theta \right\rangle \right|, \quad F = \left| \left\langle \frac{\tilde{\theta} - \theta^*}{||\tilde{\theta} - \theta^*||}, A - \theta \right\rangle \right|.$$

Then, by the derived inequalities, we have

$$L^2 \leq 2RE + 2(L+R)F + 2LB.$$

It can be rearranged as

$$L^2 \leq 2(F+B)L + 2(E+F)R.$$

By solving this quadratic inequality of $L$, we can get

$$L^2 \leq \max\{16(F+B)^2, 4R(E+F)\}. \tag{4.7}$$

By Lemma 2.1, Lemma 4.1, Lemma 4.3 and Lemma 4.4, for any constant $C' > 0$, there exist constants $C$ only depending on $C', M$, such that

$$B^2 \leq Cn^2 \left(\frac{1}{k^2}\right)^{\alpha \wedge 1}, \quad F^2 \leq Cn \log k,$$

$$R^2 \leq C(k^2 + n \log k), \quad E^2 \leq C(k^2 + n \log k),$$

with probability at least $1 - \exp(-C'n)$. By (4.7), we have

$$L^2 \leq C_1 \left( n^2 \left(\frac{1}{k^2}\right)^{\alpha \wedge 1} + k^2 + n \log k \right) \tag{4.8}$$

with probability at least $1 - \exp(-C'n)$ for some constant $C_1$. Hence, there is some constant $C_2$ such that

$$\begin{aligned} \frac{1}{n^2} \sum_{ij} (\hat{\theta}_{ij} - \theta_{ij})^2 &\leq \frac{2}{n^2} \left( L^2 + B^2 \right) \\ &\leq C_2 \left( \left(\frac{1}{k^2}\right)^{\alpha \wedge 1} + \frac{k^2}{n^2} + \frac{\log k}{n} \right), \end{aligned}$$

with probability at least $1 - \exp(-C'n)$. When $\alpha \geq 1$, we choose $k = \lceil \sqrt{n} \rceil$, and the bound is $C_3 n^{-1} \log n$ for some constant $C_3$ only depending on $C'$ and $M$. When $\alpha < 1$, we choose $k = \lceil n^{\frac{1}{\alpha+1}} \rceil$. Then the bound is $C_4 n^{-\frac{2\alpha}{\alpha+1}}$ for some constant $C_4$ only depending on $C'$ and $M$. This completes the proof. $\square$

4.2. *Proof of Theorem 2.2.* This section is devoted to proving the lower bounds. For any probability measures $\mathbb{P}, \mathbb{Q}$, define the Kullback-Leibler divergence by $D(\mathbb{P}||\mathbb{Q}) = \int \left(\log \frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P}$. The chi-squared divergence is defined by $\chi^2(\mathbb{P}||\mathbb{Q}) = \int \left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P} - 1$. To prove minimax lower bounds, we need the following proposition.

PROPOSITION 4.1. *Let $(\Theta, \rho)$ be a metric space and $\{\mathbb{P}_\theta : \theta \in \Theta\}$ be a collection of probability measures. For any totally bounded $T \subset \Theta$, define the Kullback-Leibler diameter and the chi-squared diameter of $T$ by*

$$d_{\mathrm{KL}}(T) = \sup_{\theta,\theta' \in T} D(\mathbb{P}_\theta||\mathbb{P}_{\theta'}), \quad d_{\chi^2}(T) = \sup_{\theta,\theta' \in T} \chi^2(\mathbb{P}_\theta||\mathbb{P}_{\theta'}).$$

*Then*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_\theta \left\{ \rho^2\left(\hat{\theta}(X), \theta\right) \geq \frac{\epsilon^2}{4} \right\} \geq 1 - \frac{d_{\mathrm{KL}}(T) + \log 2}{\log \mathcal{M}(\epsilon, T, \rho)}, \tag{4.9}$$

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_\theta \left\{ \rho^2\left(\hat{\theta}(X), \theta\right) \geq \frac{\epsilon^2}{4} \right\} \geq 1 - \frac{1}{\mathcal{M}(\epsilon, T, \rho)} - \sqrt{\frac{d_{\chi^2}(T)}{\mathcal{M}(\epsilon, T, \rho)}}, \tag{4.10}$$

*for any $\epsilon > 0$.*

The inequality (4.9) is the classical Fano's inequality. The version we present here is by [53]. The inequality (4.10) is a generalization of the classical Fano's inequality by using chi-squared divergence instead of KL divergence. It is due to [22]. We use it here as an alternative of Assouad's lemma to get the corresponding in-probability lower bound. In this section, the parameter is a matrix $\{\theta_{ij}\} \in [0,1]^{n\times n}$. The metric we consider is

$$\rho^2(\theta, \theta') = \frac{1}{n^2} \sum_{ij} (\theta_{ij} - \theta'_{ij})^2.$$

Let us give bounds for KL divergence and chi-squared divergence under random graph model. Let $P_{\theta_{ij}}$ denote the probability of Bernoulli$(\theta_{ij})$. Given $\theta = \{\theta_{ij}\} \in [0,1]^{n\times n}$, the probability $\mathbb{P}_\theta$ stands for the product measure $\otimes_{i,j\in[n]} P_{\theta_{ij}}$ throughout this section.

PROPOSITION 4.2. *For any $\theta, \theta' \in [1/2, 3/4]^{n\times n}$, we have*

$$D(\mathbb{P}_\theta||\mathbb{P}_{\theta'}) \leq 8 \sum_{ij} (\theta_{ij} - \theta'_{ij})^2, \quad \chi^2(\mathbb{P}_\theta||\mathbb{P}'_\theta) \leq \exp\left(8 \sum_{ij} (\theta_{ij} - \theta'_{ij})^2\right). \tag{4.11}$$

The proposition will be proved in the supplementary material [15]. We also need the following Varshamov-Gilbert bound. The version we present here is due to [40, Lemma 4.7].

LEMMA 4.5. *There exists a subset* $\{\omega_1, ..., \omega_N\} \subset \{0,1\}^d$ *such that*

$$\rho_H(\omega_i, \omega_j) \triangleq ||\omega_i - \omega_j||^2 \geq \frac{d}{4}, \quad \text{for any } i \neq j \in [N], \tag{4.12}$$

*for some* $N \geq \exp(d/8)$.

PROOF OF THEOREM 2.2. By the definition of the parameter space $\Theta_k$, we rewrite the minimax rate as

$$\begin{aligned} & \inf_{\hat{\theta}} \sup_{\theta \in \Theta_k} \mathbb{P}\left\{ \frac{1}{n^2} \sum_{ij} (\hat{\theta}_{ij} - \theta_{ij})^2 \geq \epsilon^2 \right\} \\ = \; & \inf_{\hat{\theta}} \sup_{Q = Q^T \in [0,1]^{k \times k}} \sup_{z \in \mathcal{Z}_{n,k}} \mathbb{P}\left\{ \frac{1}{n^2} \sum_{i \neq j} (\hat{\theta}_{ij} - Q_{z(i)z(j)})^2 \geq \epsilon^2 \right\}. \end{aligned}$$

If we fix a $z \in \mathcal{Z}_{n,k}$, it will be direct to derive the lower bound $k^2/n^2$ for estimating $Q$. On the other hand, if we fix $Q$ and let $z$ vary, it will become a new type of convergence rate due to the unknown label and we name it as the clustering rate, which is at the order of $n^{-1} \log k$. In the following arguments, we will prove the two different rates separately and then combine them together to get the desired in-probability lower bound.

Without loss of generality, we consider the case where both $n/k$ and $k/2$ are integers. If they are not, let $k' = 2\lfloor k/2 \rfloor$ and $n' = \lfloor n/k' \rfloor k'$. By restricting the unknown parameters to the smaller class $Q' = (Q')^T \in [0,1]^{k' \times k'}$ and $z' \in \mathcal{Z}_{n',k'}$, the following lower bound argument works for this smaller class. Then it also provides a lower bound for the original larger class.

**Nonparametric Rate.** First we fix a $z \in \mathcal{Z}_{n,k}$. For each $a \in [k]$, we define $z^{-1}(a) = \{(a-1)n/k + 1, ..., an/k\}$. Let $\Omega = \{0,1\}^d$ be the set of all binary sequences of length $d = k(k-1)/2$. For any $\omega = \{\omega_{ab}\}_{1 \leq b < a \leq k} \in \Omega$, define a $k \times k$ matrix $Q^\omega = (Q^\omega_{ab})_{k \times k}$ by

$$Q^\omega_{ab} = Q^\omega_{ba} = \frac{1}{2} + \frac{c_1 k}{n} \omega_{ab}, \quad \text{for } a > b \in [k] \quad \text{and} \quad Q^\omega_{aa} = \frac{1}{2}, \quad \text{for } a \in [k], \tag{4.13}$$

where $c_1$ is a constant that we are going to specify later. Define $\theta^\omega = (\theta^\omega_{ij})_{n \times n}$ with $\theta^\omega_{ij} = Q^\omega_{z(i)z(j)}$ for $i \neq j$ and $\theta^\omega_{ii} = 0$. The subspace we consider is

$T_1 = \{\theta^\omega : \omega \in \Omega\} \subset \Theta_k$. To apply (4.10), we need to upper bound $\sup_{\theta,\theta' \in T_1} \chi^2(\mathbb{P}_\theta || \mathbb{P}_{\theta'})$ and lower bound $\mathcal{M}(\epsilon, T_1, \rho)$. For any $\theta^\omega, \theta^{\omega'} \in T_1$, from (4.11) and (4.13), we get

$$\begin{aligned} \chi^2(\mathbb{P}_{\theta^\omega} || \mathbb{P}_{\theta^{\omega'}}) &= \exp\left(8 \sum_{i,j\in[n]} (\theta^\omega_{ij} - \theta^{\omega'}_{ij})^2\right) \\ &\leq \exp\left(\frac{8n^2}{k^2} \sum_{a,b\in[k]} (Q^\omega_{ab} - Q^{\omega'}_{ab})^2\right) \leq \exp(8c_1^2 k^2), \end{aligned} \tag{4.14}$$

where we choose sufficiently small $c_1$ so that $\theta^\omega_{ij}, \theta^{\omega'}_{ij} \in [1/2, 3/4]$ is satisfied. To lower bound the packing number, we reduce the metric $\rho(\theta^\omega, \theta^{\omega'})$ to $\rho_H(\omega, \omega')$ defined in (4.12). In view of (4.13), we get

$$\rho^2(\theta^\omega, \theta^{\omega'}) \geq \frac{1}{k^2} \sum_{1\leq b<a\leq k} (Q^\omega_{ab} - Q^{\omega'}_{ab})^2 = \frac{c_1^2}{n^2} \rho_H(\omega, \omega'). \tag{4.15}$$

By Lemma 4.5, we can find a subset $S \subset \Omega$ that satisfies the following properties: (a) $|S| \geq \exp(d/8)$ and (b) $\rho_H(\omega, \omega') \geq d/4$ for any $\omega, \omega' \in S$. From (4.15), we have

$$\mathcal{M}(\epsilon, T_1, \rho) \geq |S| \geq \exp(d/8) = \exp(k(k-1)/16),$$

with $\epsilon^2 = \frac{c_1 k(k-1)}{8n^2}$. By choosing sufficiently small $c_1$, together with (4.14), we get

$$\inf_{\hat\theta} \sup_{\theta\in T_1} \mathbb{P}\left\{\frac{1}{n^2} \sum_{ij} (\hat\theta_{ij} - \theta_{ij})^2 \geq \frac{C_1 k^2}{n^2}\right\} \geq 0.9, \tag{4.16}$$

by (4.10) for sufficiently large $k$ with some constant $C_1 > 0$. When $k$ is not sufficiently large, i.e. $k \leq O(1)$, then it is easy to see that $n^{-2}$ is always the correct order of lower bound. Since $n^{-2} \asymp k^2/n^2$ when $k \leq O(1)$, $k^2/n^2$ is also a valid lower bound for small $k$.

**Clustering Rate.** We are going to fix a $Q$ that has the following form

$$Q = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}, \tag{4.17}$$

where $B$ is a $(k/2) \times (k/2)$ matrix. By Lemma 4.5, when $k$ is sufficiently large, we can find $\{\omega_1, ..., \omega_{k/2}\} \subset \{0,1\}^{k/2}$ such that $\rho_H(\omega_a, \omega_b) \geq k/8$ for all $a \neq b \in [k/2]$. Fixing such $\{\omega_1, ..., \omega_{k/2}\}$, define $B = (B_1, B_2, ..., B_{k/2})$ by letting $B_a = \frac{1}{2} + \sqrt{\frac{c_2 \log k}{n}}\omega_a$ for $a \in [k/2]$. With such construction, it is easy to see that for any $a \neq b \in [k/2]$,

$$||B_a - B_b||^2 \geq \frac{c_2 k \log k}{8n}. \tag{4.18}$$

Define a subset of $\mathcal{Z}_{n,k}$ by

$$\begin{aligned} \mathcal{Z} \quad = \quad & \Big\{ z \in \mathcal{Z}_{n,k} : |z^{-1}(a)| = \frac{n}{k} \text{ for } a \in [k], \\ & z^{-1}(a) = \left\{ \frac{(a-1)n}{k} + 1, ..., \frac{an}{k} \right\} \text{ for } a \in [k/2] \Big\}. \end{aligned}$$

For each $z \in \mathcal{Z}$, define $\theta^z$ by $\theta^z_{ij} = Q_{z(i)z(j)}$ for $i \neq j$ and $\theta^z_{ii} = 0$. The subspace we consider is $T_2 = \{\theta^z : z \in \mathcal{Z}\} \subset \Theta_{n,k}$. To apply (4.9), we need to upper bound $\sup_{\theta,\theta \in T_2} D(\mathbb{P}_\theta || \mathbb{P}_{\theta'})$ and lower bound $\log \mathcal{M}(\epsilon, T_2, \rho)$. By (4.11), for any $\theta, \theta' \in T_2$,

$$D(\mathbb{P}_\theta || \mathbb{P}_{\theta'}) \leq 8 \sum_{ij} (\theta_{ij} - \theta'_{ij})^2 \leq 8n^2 c_2 \frac{\log k}{n} = 8c_2 n \log k. \tag{4.19}$$

Now we are going to give a lower bound of the packing number $\log \mathcal{M}(\epsilon, T_2, \rho)$ with $\epsilon^2 = (c_2 \log k)/(48n)$ for the $c_2$ in (4.18). Due to the construction of $B$, there is a one-to-one correspondence between $T_2$ and $\mathcal{Z}$. Thus, $\log \mathcal{M}(\epsilon, T_2, \rho) = \log \mathcal{M}(\epsilon, \mathcal{Z}, \rho_1)$ for some metric $\rho_1$ on $\mathcal{Z}$ defined by $\rho_1(z, w) = \rho(\theta^z, \theta^w)$. Given any $z \in \mathcal{Z}$, define its $\epsilon$-neighborhood by $B(z, \epsilon) = \{w \in \mathcal{Z} : \rho_1(z, w) \leq \epsilon\}$. Let $S$ be the packing set in $\mathcal{Z}$ with cardinality $\mathcal{M}(\epsilon, \mathcal{Z}, \rho_1)$. We claim that $S$ is also the covering set of $\mathcal{Z}$ with radius $\epsilon$, because otherwise there is some point in $\mathcal{Z}$ which is at least $\epsilon$ away from every point in $S$, contradicting the definition of $\mathcal{M}(\epsilon, \mathcal{Z}, \rho_1)$. This implies the fact $\cup_{z \in S} B(z, \epsilon) = \mathcal{Z}$, which leads to

$$|\mathcal{Z}| \leq \sum_{z \in S} |B(z, \epsilon)| \leq |S| \max_{z \in S} |B(z, \epsilon)|.$$

Thus, we have

$$\mathcal{M}(\epsilon, \mathcal{Z}, \rho_1) = |S| \geq \frac{|\mathcal{Z}|}{\max_{z \in S} |B(z, \epsilon)|}. \tag{4.20}$$

Let us upper bound $\max_{z \in S} |B(z, \epsilon)|$ first. For any $z, w \in \mathcal{Z}$, by the construction of $\mathcal{Z}$, $z(i) = w(i)$ when $i \in [n/2]$ and $|z^{-1}(a)| = n/k$ for each

$a \in [k]$. Hence,

$$\begin{aligned}
\rho_1^2(z,w) &\geq \frac{1}{n^2}\sum_{1\leq i\leq n/2<j\leq n}(Q_{z(i)z(j)}-Q_{w(i)w(j)})^2 \\
&= \frac{1}{n^2}\sum_{n/2<j\leq n}\sum_{1\leq a\leq k/2}\sum_{i\in z^{-1}(a)}(Q_{az(j)}-Q_{aw(j)})^2 \\
&= \frac{1}{n^2}\sum_{n/2<j\leq n}\frac{n}{k}||B_{z(j)}-B_{w(j)}||^2 \\
&\geq \frac{c_2\log k}{8n^2}|\{j: w(j)\neq z(j)\}|,
\end{aligned}$$

where the last inequality is due to (4.18). Then for any $w \in B(z,\epsilon)$, $|\{j : w(j) \neq z(j)\}| \leq n/6$ under the choice $\epsilon^2 = (c_2 \log k)/(48n)$. This implies

$$|B(z,\epsilon)| \leq \binom{n}{n/6}k^{n/6} \leq (6e)^{n/6}k^{n/6} \leq \exp\left(\frac{1}{4}n\log k\right).$$

Now we lower bound $|\mathcal{Z}|$. Note that by Stirling's formula,

$$|\mathcal{Z}| = \frac{(n/2)!}{[(n/k)!]^{k/2}} = \exp\left(\frac{1}{2}n\log k + o(n\log k)\right) \geq \exp\left(\frac{1}{3}n\log k\right).$$

By (4.20), we get $\log\mathcal{M}(\epsilon,T,\rho) = \log\mathcal{M}(\epsilon,\mathcal{Z},\rho_1) \geq (1/12)n\log k$. Together with (4.19) and using (4.9), we have

$$\inf_{\hat{\theta}}\sup_{\theta\in T_2}\mathbb{P}\left\{\frac{1}{n^2}\sum_{ij}(\hat{\theta}_{ij}-\theta_{ij})^2 \geq \frac{C_2\log k}{n}\right\} \geq 0.9, \tag{4.21}$$

with some constant $C_2 > 0$ for sufficiently small $c_2$ and sufficiently large $k$. When $k$ is not sufficiently large but $2 \leq k \leq O(1)$, the argument in Section 3.3 gives the desired lower bound at the order of $n^{-1} \asymp n^{-1}\log k$. When $k = 1$, $n^{-1}\log k = 0$ is still a valid lower bound.

**Combining the Bounds.** Finally, let us combine (4.16) and (4.21) to get the desired in-probability lower bound in Theorem 2.2 with $C = \big(C_1 \wedge C_2\big)/2$.

For any $\theta \in \Theta_k$, by union bound, we have

$$
\begin{aligned}
& \mathbb{P}\left\{\frac{1}{n^2}\sum_{ij}(\hat{\theta}_{ij}-\theta_{ij})^2 \geq C\left(\frac{k^2}{n^2}+\frac{\log k}{n}\right)\right\} \\
\geq\ & 1-\mathbb{P}\left\{\frac{1}{n^2}\sum_{ij}(\hat{\theta}_{ij}-\theta_{ij})^2 \leq \frac{C_1k^2}{n^2}\right\}-\mathbb{P}\left\{\frac{1}{n^2}\sum_{ij}(\hat{\theta}_{ij}-\theta_{ij})^2 \leq \frac{C_2\log k}{n}\right\} \\
=\ & \mathbb{P}\left\{\frac{1}{n^2}\sum_{ij}(\hat{\theta}_{ij}-\theta_{ij})^2 \geq \frac{C_1k^2}{n^2}\right\}+\mathbb{P}\left\{\frac{1}{n^2}\sum_{ij}(\hat{\theta}_{ij}-\theta_{ij})^2 \geq \frac{C_2\log k}{n}\right\}-1.
\end{aligned}
$$

Taking sup on both sides, and using the fact $\sup_{z,Q}\Big(f(z)+g(Q)\Big)=\sup_z f(z)+\sup_Q g(Q)$, we have

$$
\begin{aligned}
& \sup_{\theta\in\Theta_k}\mathbb{P}\left\{\frac{1}{n^2}\sum_{ij}(\hat{\theta}_{ij}-\theta_{ij})^2 \geq C\left(\frac{k^2}{n^2}+\frac{\log k}{n}\right)\right\} \\
\geq\ & \sup_{\theta\in T_1}\mathbb{P}\left\{\frac{1}{n^2}\sum_{ij}(\hat{\theta}_{ij}-\theta_{ij})^2 \geq \frac{C_1k^2}{n^2}\right\}+\sup_{\theta\in T_2}\mathbb{P}\left\{\frac{1}{n^2}\sum_{ij}(\hat{\theta}_{ij}-\theta_{ij})^2 \geq \frac{C_2\log k}{n}\right\}-1,
\end{aligned}
$$

for any estimator $\hat{\theta}$. Plugging the lower bounds (4.16) and (4.21), we obtain the desired result. A Markov's inequality argument leads to the lower bound in expectation. □

**Acknowledgements.** We want to thank Zongming Ma for helpful discussion on the relation between graphon estimation and link prediction, and to thank the Associate Editor and the Referee for their constructive comments and suggestions that lead to the improvement of the paper.

## SUPPLEMENTARY MATERIAL

**Supplement A: Supplement to "Rate-Optimal Graphon Estimation"**
(url to be specified). In the supplement, we prove Theorem 2.4, Lemma 2.1, Lemma 4.1, Lemma 4.2, Lemma 4.3, Proposition 4.2 and Theorem 3.6.

### References.

[1] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. In *Advances in Neural Information Processing Systems*, pages 33–40, 2009.

[2] Edoardo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.
[3] David J Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.
[4] Arash A Amini, Aiyou Chen, Peter J Bickel, and Elizaveta Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013.
[5] Anima Anandkumar, Rong Ge, Daniel Hsu, and Sham M Kakade. A tensor spectral approach to learning mixed membership community models. *arXiv preprint arXiv:1302.2684*, 2013.
[6] Peter Bickel, David Choi, Xiangyu Chang, and Hai Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41(4):1922–1943, 2013.
[7] Peter J Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.
[8] Tony Cai and Xiaodong Li. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *arXiv preprint arXiv:1404.6000*, 2014.
[9] Stanley H Chan and Edoardo M Airoldi. A consistent histogram estimator for exchangeable graph models. *arXiv preprint arXiv:1402.1888*, 2014.
[10] Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *arXiv preprint arXiv:1212.1247*, 2012.
[11] Yizong Cheng and George M Church. Biclustering of expression data. In *Ismb*, volume 8, pages 93–103, 2000.
[12] Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in the sparse graphs: A spectral algorithm with optimal rate of recovery. *arXiv preprint arXiv:1501.05021*, 2015.
[13] Ronald R Coifman and Matan Gavish. Harmonic analysis of digital data bases. In *Wavelets and Multiscale Analysis*, pages 161–197. Springer, 2011.
[14] Persi Diaconis and Svante Janson. Graph limits and exchangeable random graphs. *arXiv preprint arXiv:0712.2749*, 2007.
[15] Chao Gao, Yu Lu, and Harrison H. Zhou. Supplement to "rate-optimal graphon estimation". 2015.
[16] Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Achieving optimal misclassification proportion in stochastic block model. *arXiv preprint arXiv:1505.03772*, 2015.
[17] Matan Gavish and Ronald R Coifman. Sampling, denoising and compression of matrices by coherent matrix organization. *Applied and Computational Harmonic Analysis*, 33(3):354–369, 2012.
[18] Matan Gavish, Boaz Nadler, and Ronald R Coifman. Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 367–374, 2010.
[19] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.
[20] Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2010.

[21] Roger Guimerà and Marta Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078, 2009.

[22] Aditya Guntuboyina. Lower bounds for the minimax risk using-divergences, and applications. *IEEE Transactions on Information Theory*, 57(4):2386–2399, 2011.

[23] Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming. *arXiv preprint arXiv:1412.6156*, 2014.

[24] Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

[25] John A Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972.

[26] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[27] Peter Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, pages 657–664, 2008.

[28] Paul W Holland and Samuel Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76 (373):33–50, 1981.

[29] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[30] Douglas N Hoover. Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*, 2, 1979.

[31] Antony Joseph and Bin Yu. Impact of regularization on spectral clustering. *arXiv preprint arXiv:1312.1733*, 2013.

[32] Olav Kallenberg. On the representation theorem for exchangeable arrays. *Journal of Multivariate Analysis*, 30(1):137–154, 1989.

[33] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

[34] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2014.

[35] James Lloyd, Peter Orbanz, Zoubin Ghahramani, and Daniel Roy. Random function priors for exchangeable arrays with applications to graphs and relational data. 2013.

[36] László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

[37] László Lovász and Balázs Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957, 2006.

[38] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.

[39] Sara C Madeira and Arlindo L Oliveira. Biclustering algorithms for biological data analysis: a survey. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 1(1):24–45, 2004.

[40] Pascal Massart. *Concentration inequalities and model selection*, volume 1896. Springer, 2007.

[41] Boris Mirkin. *Mathematical classification and clustering: From how to what and why.* Springer, 1998.

[42] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*, 2014.

[43] Mark EJ Newman and Elizabeth A Leicht. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences*, 104(23):9564–9569,

2007.

[44] Krzysztof Nowicki and Tom A B Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

[45] Sofia C Olhede and Patrick J Wolfe. Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences*, 111(41): 14722–14727, 2014.

[46] David Pollard. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–86. JSTOR, 1990.

[47] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

[48] Purnamrita Sarkar, Deepayan Chakrabarti, and Michael Jordan. Nonparametric link prediction in dynamic networks. *arXiv preprint arXiv:1206.6394*, 2012.

[49] Alexandre B Tsybakov. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

[50] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[51] Stanley Wasserman. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[52] Patrick J Wolfe and Sofia C Olhede. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.

[53] Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

[54] Haiyuan Yu, Pascal Braun, Muhammed A Yıldırım, Irma Lemmens, Kavitha Venkatesan, Julie Sahalie, Tomoko Hirozane-Kishikawa, Fana Gebreab, Na Li, Nicolas Simonis, et al. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, 2008.

[55] Anderson Y Zhang and Harrison H Zhou. Minimax rates of community detection in stochastic block model. http://www.stat.yale.edu/~yz482/community_detection_minimax.pdf, 2015.

[56] Xue Zhang, Xiaojie Wang, Chengli Zhao, Dongyun Yi, and Zheng Xie. Degree-corrected stochastic block models and reliability in networks. *Physica A: Statistical Mechanics and its Applications*, 393:553–559, 2014.

[57] Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, 2012.

SUPPLEMENT TO "RATE-OPTIMAL GRAPHON ESTIMATION"

BY Chao Gao, Yu Lu and Harrison H. Zhou

Yale University

# APPENDIX A: ADDITIONAL PROOFS

In this supplement, we prove Theorem 2.4, Lemma 2.1, Lemma 4.1, Lemma 4.2, Lemma 4.3, Proposition 4.2 and Theorem 3.6.

**A.1. Proof of Theorem 2.4.** We use the same idea as in the proof of Theorem 2.2. First we are going to fix a $\mathbb{P}_\xi \in \mathcal{P}$ and construct a subset of $\mathcal{F}_\alpha(M)$ to get the nonparametric rate $n^{-2\alpha/(\alpha+1)}$. Then we fix a $f \in \mathcal{F}_\alpha(M)$ and let $\mathbb{P}_\xi$ vary to get the clustering rate $n^{-1}\log n$. Since our target is the sum of the two rates, it is sufficient to prove the nonparametric rate lower bound for $\alpha \in (0,1)$ and prove the clustering rate lower bound for $\alpha \geq 1$.

**Nonparametric Rate.** We assume $\alpha \in (0,1)$ in this part. Consider the the fixed design $(\xi_1, ..., \xi_n) = (1/n, ..., n/n)$. This can be viewed as a degenerated distribution belonging to the set $\mathcal{P}$. Then it is sufficient to lower bound

$$\inf_{\hat{\theta}} \sup_{f\in\mathcal{F}_\alpha(M)} \mathbb{P}\left\{\frac{1}{n^2}\sum_{i,j\in[n]}\Big(\hat{\theta}_{ij} - f\big(i/n, j/n\big)\Big)^2 \geq \frac{\epsilon^2}{4}\right\}, \tag{A.1}$$

for $\epsilon^2 = cn^{-\frac{2\alpha}{\alpha+1}}$ with some $c > 0$ to be determined later. This can be viewed as a classical nonparametric regression problem, but with Bernoulli observations. We are going to apply (4.9) in Proposition 4.1. Our lower bound argument essentially follows the construction in [49, Sec 2.6.1]. To facilitate the presentation, we introduce the following function

$$K(x,y) = (1-2|x|)(1-2|y|)\mathbb{I}\left\{|x| \leq 1/2, |y| \leq 1/2\right\}.$$

Let us take $k = \lceil c_1 n^{\frac{1}{\alpha+1}} \rceil$ for some constant $c_1 > 0$ to be determined later. For any $a, b \in [k]$, define the function

$$\phi_{ab}(x,y) = Lk^{-\alpha}K\left(kx - a + \frac{1}{2}, ky - b + \frac{1}{2}\right). \tag{A.2}$$

By such construction, we have

PROPOSITION A.1. *Assume $\alpha \in (0,1)$. For some $L > 0$ depending on $\alpha, M$, the function (A.2) satisfies*

1. $\phi_{ab}(x,y) \in \mathcal{H}_\alpha(M)$.
2. $\sum_{i,j\in[n]} \phi_{ab}^2(i/n, j/n) \geq L^2 n^2 k^{-2\alpha-2}/9$.

The proposition will be proved in Appendix A.2. Let $\Omega = \{0,1\}^d$ be the set of all binary sequences of length $d = k(k+1)/2$. For any $\omega = \{\omega_{ab}\}_{1\leq b\leq a\leq k} \in \Omega$, we define the function $f^\omega$ by

$$f^\omega(x,y) = f^\omega(y,x) = \frac{1}{2} + \sum_{1\leq b\leq a\leq k} \omega_{ab}\phi_{ab}(x,y), \quad \text{for} \quad x \geq y. \tag{A.3}$$

The subspace we consider is $\mathcal{F}' = \{f^\omega : \omega \in \Omega\}$. Since $K$ is bounded and the collection $\{\phi_{ab}\}$ have disjoint supports and belong to $\mathcal{H}_\alpha(M)$, we have $\mathcal{F}' \subset \mathcal{F}_\alpha(M)$ when $M > 1$. For the case $M \leq 1$, we may replace the $1/2$ in (A.3) by some sufficiently small number so that $\mathcal{F}' \subset \mathcal{F}_\alpha(M)$ is still true. We choose to use $1/2$ so that the following analysis can be presented in a cleaner way. To apply (4.9), we first upper bound $\sup_{f,f'} D(\mathbb{P}_f||\mathbb{P}_{f'})$. For any $f \in \mathcal{F}'$, denote $f_{ij} = f(i/n, j/n)$ and by our construction $1/4 \leq f_{ij} \leq 3/4$ for sufficiently small $L$. Then from (4.11) in Proposition 4.2, for any $f, f' \in \mathcal{F}'$ we have

$$D(\mathbb{P}_f||\mathbb{P}_{f'}) \leq 8 \sum_{i,j\in[n]} (f_{ij} - f'_{ij})^2 \leq 8n^2L^2k^{-2\alpha} \leq 8L^2c_1^{-2}n^{\frac{2}{\alpha+1}}. \tag{A.4}$$

Next, we lower bound the packing number of $\mathcal{F}'$. For any $f^\omega, f^{\omega'} \in \mathcal{F}'$, we have

$$\begin{aligned} \rho^2(f^\omega, f^{\omega'}) &\geq \frac{1}{n^2} \sum_{i,j\in[n]} \sum_{1\leq b\leq a\leq k} (\omega_{ab} - \omega'_{ab})^2 \phi_{ab}^2(i/n, j/n) \\ &= \frac{1}{n^2} \sum_{1\leq b\leq a\leq k} (\omega_{ab} - \omega'_{ab})^2 \left( \sum_{i,j\in[n]} \phi_{ab}^2(i/n, j/n) \right) \\ &\geq \frac{1}{9} L^2 k^{-2\alpha-2} \rho_H(\omega, \omega'), \end{aligned}$$

where we have used Proposition A.1 in the last inequality above, and the distance $\rho_H$ is defined in (4.12). By Lemma 4.5, we may choose a subset $S \subset \Omega$ such that $|S| \geq \exp(k^2/16)$ and $\rho_H(\omega, \omega') \geq k^2/8$ for any $\omega \neq \omega' \in S$. Then if we set $\epsilon^2 = cn^{-\frac{2\alpha}{\alpha+1}}$ for some sufficiently small $c$, we have $\log \mathcal{M}(\epsilon, \mathcal{F}', \rho) \geq k^2/16$. By (4.9), we get

$$\inf_{\hat{\theta}} \sup_{f\in\mathcal{F}'} \mathbb{P}\left\{ \frac{1}{n^2} \sum_{i,j\in[n]} (\hat{\theta}_{ij} - f(i/n, j/n))^2 \geq \frac{\epsilon^2}{4} \right\} \geq 0.9, \tag{A.5}$$

by choosing sufficient large $c_1$.

**Clustering Rate.** We assume $\alpha \geq 1$ in this part. We are going to reduce the problem to the clustering rate of stochastic block model. First we are going to construct an $f \in \mathcal{F}_\alpha(M)$ that mimics $Q$ in the stochastic block model. For some $\delta \in (0,1)$ to be specified later, define $k = 2\lfloor n^\delta/2 \rfloor$. To construct a function $f \in \mathcal{H}_\alpha(M)$, we need the following smooth function $K(x)$ that is infinitely differentiable,

$$K(x) = C_K \exp\left(-\frac{1}{1-64x^2}\right)\mathbb{I}\left\{|x| < \frac{1}{8}\right\},$$

where $C_K > 0$ is a constant such that $\int K(x)dx = 1$. The function $K$ is a positive symmetric mollifier, based on which we define the following function

$$\psi(x) = \int_{-3/8}^{3/8} K(x-y)dy.$$

The function $\psi(x)$ is called a smooth cutoff function. It can be viewed as the convolution of $K(x)$ and $\mathbb{I}\{|x| \leq 3/8\}$. The support of $\psi(x)$ is $[-1/2, 1/2]$. Since $K(x)$ is supported on $[-1/8, 1/8]$ and the value of its integral is 1, $\psi(x)$ is 1 on the interval $[-1/4, 1/4]$. Moreover, the smoothness property of $K(x)$ is inherited by $\psi(x)$. Recall the $k \times k$ matrix $Q$ defined in (4.17), and define

$$f(x,y) = \sum_{a,b \in [k]} \left(Q_{ab} - \frac{1}{2}\right)\psi\left(kx - a + \frac{1}{2}\right)\psi\left(ky - b + \frac{1}{2}\right) + \frac{1}{2}.$$

It is easy to verify that $f \in \mathcal{F}_\alpha(M)$ as long as we choose sufficiently small $\delta$ depending on $\alpha \geq 1$ and $M > 1$. The case $M \geq 1$ requires some modification on the definition of $Q_{ab}$, and is omitted in the paper. The definition of $f$ implies that for any $a, b \in [k]$,

$$f(x,y) \equiv Q_{ab}, \qquad \text{when } (x,y) \in \left[\frac{a-3/4}{k}, \frac{a-1/4}{k}\right] \times \left[\frac{b-3/4}{k}, \frac{b-1/4}{k}\right].$$

Therefore, in a sub-domain, $f$ is a piecewise constant function. To be specific, define

$$I = \left(\bigcup_{a=1}^{k}\left[\frac{n(a-3/4)}{k}, \frac{n(a-1/4)}{k}\right]\right)\bigcap [n].$$

The values of $f(i/n, j/n)$ on $(i,j) \in I \times I$ form a stochastic block model. Let $\Pi_n$ be the set of all permutations on $[n]$. Define a subset by $\Pi'_n = \{\sigma \in$

$\Pi_n : \sigma(i) = i$ for $i \in [n] \backslash I\}$. In other words, any $\sigma \in \Pi'_n$ can be viewed as a permutation on $I$. Note that for any permutation $\sigma \in \Pi'_n$, the degenerated distribution

$$\mathbb{P}_\sigma\left((\xi_1, ..., \xi_n) = \big(\sigma(1)/n, ..., \sigma(n)/n\big)\right) = 1$$

belongs to the set $\mathcal{P}$. Then the minimax risk has lower bound

$$\begin{aligned}
&\inf_{\hat\theta} \sup_{f \in \mathcal{F}_\alpha(M)} \sup_{\mathbb{P}_\xi \in \mathcal{P}} \mathbb{P}\left\{\frac{1}{n^2}\sum_{i,j\in[n]}(\hat\theta_{ij} - \theta_{ij})^2 \geq \epsilon^2\right\} \\
\geq\ & \inf_{\hat\theta} \max_{\sigma \in \Pi'_n} \mathbb{P}\left\{\frac{1}{n^2}\sum_{i,j\in[n]}\Big(\hat\theta_{ij} - f\big(\sigma(i)/n, \sigma(j)/n\big)\Big)^2 \geq \epsilon^2\right\} \\
\geq\ & \inf_{\hat\theta} \max_{\sigma \in \Pi'_n} \mathbb{P}\left\{\frac{1}{n^2}\sum_{i,j\in I}\Big(\hat\theta_{ij} - f\big(\sigma(i)/n, \sigma(j)/n\big)\Big)^2 \geq \epsilon^2\right\}. \qquad \text{(A.6)}
\end{aligned}$$

The form (A.6) is a case of stochastic block model with fixed $Q$ and varying $z$. To see this, for any $\sigma \in \Pi'_n$, let us define $z : I \to [k]$ satisfying $z(i) = \lceil n^{-1}k\sigma(i)\rceil$ for each $i \in I$. Then we collect all such $z$ to form the set $\mathcal{Z}_{I,k}$. For any $i, j \in I$, as long as $(i, j) \in z^{-1}(a) \times z^{-1}(b)$, we have $\theta_{ij} = f\big(\sigma(i)/n, \sigma(j)/n\big) = Q_{ab}$. Using the same argument in the proof of Theorem 2.2, we can get the same result of (4.21),

$$\inf_{\hat\theta} \max_{\sigma \in \Pi'_n} \mathbb{P}\left\{\frac{1}{n^2}\sum_{i,j\in I}\Big(\hat\theta_{ij} - f\big(\sigma(i)/n, \sigma(j)/n\big)\Big)^2 \geq \epsilon^2\right\} \geq 0.9, \qquad \text{(A.7)}$$

with $\epsilon_2^2 = \frac{c_2 \log k}{n} \geq \frac{c_3 \log n}{n}$, for some $c_2, c_3 > 0$. Finally, applying the same argument in the proof of Theorem 2.2 to combine (A.5) and (A.7), the proof is complete.

### A.2. Proofs of Some Auxiliary Results.

Proof of Lemma 2.1. Define $z^* : [n] \to [k]$ by

$$(z^*)^{-1}(a) = \{i \in [n] : \xi_i \in U_a\},$$

for $U_a$ defined in (2.9). We use the notation $n_a^* = |(z^*)^{-1}(a)|$ for each $a \in [k]$ and $Z_{ab}^* = \{(u, v) : z^*(u) = a, z^*(v) = b\}$ for $a, b \in [k]$. By such construction of $z^*$, for $i,j$ such that $\xi_i \in U_a, \xi_j \in U_b$ with $a \neq b$, we have

$$\begin{aligned}
& |f(\xi_i,\xi_j) - \bar{\theta}_{ab}(z^*)| \\
= & \left| f(\xi_i,\xi_j) - \frac{1}{n_a^* n_b^*} \sum_{(u,v)\in Z_{ab}^*} f(\xi_u,\xi_v) \right| \\
\leq & \frac{1}{n_a^* n_b^*} \sum_{(u,v)\in Z_{ab}^*} |f(\xi_i,\xi_j) - f(\xi_u,\xi_v)| \\
\leq & \frac{1}{n_a^* n_b^*} \sum_{(u,v)\in Z_{ab}^*} M\big(|\xi_i - \xi_u| + |\xi_j - \xi_u|\big)^{\alpha\wedge 1} \\
\leq & C_1 M k^{-(\alpha\wedge 1)}.
\end{aligned}$$

The second inequality above is because of the Lipschitz condition (2.7) for $\alpha \in (0,1]$. When $\alpha > 1$, any function $f \in \mathcal{H}_\alpha(M)$ satisfies (2.7) for $\alpha = 1$. Similar results also hold for the case $a = b$. Summing over $a,b \in [k]$, the proof is complete. □

PROOF OF LEMMA 4.1. By the definitions of $\hat{\theta}_{ij}$ and $\tilde{\theta}_{ij}$, we have

$$\hat{\theta}_{ij} - \tilde{\theta}_{ij} = \hat{Q}_{\hat{z}(i)\hat{z}(j)} - \tilde{Q}_{\hat{z}(i)\hat{z}(j)} = \bar{A}_{ab}(\hat{z}) - \bar{\theta}_{ab}(\hat{z})$$

for any $(i,j) \in \hat{z}^{-1}(a) \times \hat{z}^{-1}(b)$ and $i \neq j$. We also have $\hat{\theta}_{ii} - \tilde{\theta}_{ii} = 0$ for any $i \in [n]$. Then

$$\begin{aligned}
\sum_{ij} (\hat{\theta}_{ij} - \tilde{\theta}_{ij})^2 &\leq \sum_{a,b\in[k]} \left|\hat{z}^{-1}(a)\right| \left|\hat{z}^{-1}(b)\right| \Big(\bar{A}_{ab}(\hat{z}) - \bar{\theta}_{ab}(\hat{z})\Big)^2 \\
&\leq \max_{z\in\mathcal{Z}_{n,k}} \sum_{a,b\in[k]} \left|z^{-1}(a)\right| \left|z^{-1}(b)\right| \Big(\bar{A}_{ab}(z) - \bar{\theta}_{ab}(z)\Big)^2. \qquad \text{(A.8)}
\end{aligned}$$

For any $a,b \in [k]$ and $z \in \mathcal{Z}_{n,k}$, define $n_a = \left|z^{-1}(a)\right|$ and $V_{ab}(z) = n_a n_b \Big(\bar{A}_{ab}(z) - \bar{\theta}_{ab}(z)\Big)^2$. Then, (A.8) is bounded by

$$\max_{z\in\mathcal{Z}_{n,k}} \sum_{a,b\in[k]} \mathbb{E}V_{ab}(z) + \max_{z\in\mathcal{Z}_{n,k}} \sum_{a,b\in[k]} \Big(V_{ab}(z) - \mathbb{E}V_{ab}(z)\Big). \qquad \text{(A.9)}$$

We are going to bound the two terms separately. For the first term, when

$a \neq b$, we have

$$\begin{aligned}\mathbb{E}V_{ab}(z) &= n_a n_b \mathbb{E}\left(\frac{1}{n_a n_b}\sum_{i\in z^{-1}(a),j\in z^{-1}(b)}(A_{ij}-\theta_{ij})\right)^2 \\ &= \frac{1}{n_a n_b}\sum_{i\in z^{-1}(a),j\in z^{-1}(b)}\mathrm{Var}(A_{ij}) \leq 1,\end{aligned}$$

where we have used the fact that $\mathbb{E}A_{ij} = \theta_{ij}$ and $\mathrm{Var}(A_{ij}) = \theta_{ij}(1-\theta_{ij}) \leq 1$. Similar conclusions can be made for diagonal $V_{aa}(z)$ by using the definition (2.2). Summing over $a, b \in [k]$, we get

$$\max_{z\in\mathcal{Z}_{n,k}} \sum_{a,b\in[k]} \mathbb{E}V_{ab}(z) \leq C_1 k^2, \tag{A.10}$$

for some universal constant $C_1 > 0$. By Hoeffding inequality [26] and $0 \leq A_{ij} \leq 1$, for any $t > 0$ we have

$$\mathbb{P}\big(V_{ab}(z) > t\big) = \mathbb{P}\left(\left|\frac{1}{n_a n_b}\sum_{i\in z^{-1}(a),j\in z^{-1}(b)}(A_{ij}-\theta_{ij})\right| > \sqrt{\frac{t}{n_a n_b}}\right) \leq 2\exp(-2t).$$

Thus, $V_{ab}(z)$ $(a \neq b)$ is a sub-exponential random variable with constant sub-exponential parameter. Again, similar conclusions can be obtained for diagonal $V_{aa}(z)$. By Bernstein's inequality for sub-exponential variables [50, Prop 5.16], we have

$$\mathbb{P}\left(\sum_{a,b\in[k]}\big(V_{ab}(z) - \mathbb{E}V_{ab}(z)\big) > t\right) \leq \exp\left(-C_2 \min\left\{\frac{t^2}{k^2}, t\right\}\right),$$

for some universal constant $C_2 > 0$. Applying union bound and using the fact that $\log|\mathcal{Z}_{n,k}| \leq n\log k$, we have

$$\mathbb{P}\left(\max_{z\in\mathcal{Z}_{n,k}}\sum_{a,b\in[k]}\big(V_{ab}(z) - \mathbb{E}V_{ab}(z)\big) > t\right) \leq \exp\left(-C_2 \min\left\{\frac{t^2}{k^2}, t\right\} + n\log k\right).$$

Thus, for any $C_3 > 0$, there exists $C_4 > 0$ only depending on $C_2$ and $C_3$, such that

$$\max_{z\in\mathcal{Z}_{n,k}}\sum_{a,b\in[k]}\left(V_{ab}(z) - \mathbb{E}V_{ab}(z)\right) \leq C_3\left(n\log k + \sqrt{nk^2\log k}\right) \tag{A.11}$$

with probability at least $1 - \exp(-C_4 n \log k)$. Plugging the bounds (A.10) and (A.11) into (A.9), we obtain

$$\begin{aligned}\sum_{ij}(\hat{\theta}_{ij} - \tilde{\theta}_{ij})^2 &\leq (C_3 + C_1)\left(k^2 + n\log k + \sqrt{nk^2 \log k}\right) \\ &\leq 2(C_3 + C_1)\left(k^2 + n \log k\right)\end{aligned}$$

with probability at least $1 - \exp(-C_4 n \log k)$. The proof is complete. □

PROOF OF LEMMA 4.2. Note that

$$\tilde{\theta}_{ij} - \theta_{ij} = \sum_{a,b\in[k]} \bar{\theta}_{ab}(\hat{z})\mathbb{I}\{(i,j) \in \hat{z}^{-1}(a) \times \hat{z}^{-1}(b)\} - \theta_{ij}$$

is a function of the partition $\hat{z}$, then we have

$$\left|\sum_{ij} \frac{\tilde{\theta}_{ij} - \theta_{ij}}{\sqrt{\sum_{ij}(\tilde{\theta}_{ij} - \theta_{ij})^2}}(A_{ij} - \theta_{ij})\right| \leq \max_{z\in\mathcal{Z}_{n,k}} \left|\sum_{ij} \gamma_{ij}(z)(A_{ij} - \theta_{ij})\right|,$$

where

$$\gamma_{ij}(z) \propto \sum_{a,b\in[k]} \bar{\theta}_{ab}(z)\mathbb{I}\{(i,j) \in z^{-1}(a) \times z^{-1}(b)\} - \theta_{ij}$$

satisfies $\sum_{ij} \gamma_{ij}(z)^2 = 1$. By Hoeffding's inequality [50, Prop 5.10] and union bound, we have

$$\begin{aligned}&\mathbb{P}\left(\max_{z\in\mathcal{Z}_{n,k}} \left|\sum_{ij} \gamma_{ij}(z)(A_{ij} - \theta_{ij})\right| > t\right) \\ \leq\ & \sum_{z\in\mathcal{Z}_{n,k}} \mathbb{P}\left(\left|\sum_{ij} \gamma_{ij}(z)(A_{ij} - \theta_{ij})\right| > t\right) \\ \leq\ & |\mathcal{Z}_{n,k}| \exp(-C_1 t^2) \\ \leq\ & \exp(-C_1 t^2 + n \log k),\end{aligned}$$

for some universal constant $C_1 > 0$. Choosing $t \propto \sqrt{n \log k}$, the proof is complete. □

To prove Lemma 4.3, we need the following auxiliary result, whose proof will be given after the proof of Lemma 4.3.

LEMMA A.1. *Let $\mathcal{B} \subset \left\{a \in \mathbb{R}^{n\times n} : \sum_{ij} a_{ij}^2 \leq 1\right\}$. Assume for any $a, b \in \mathcal{B}$,*

$$\frac{a-b}{||a-b||} \in \mathcal{B}. \tag{A.12}$$

*Then, we have*

$$\mathbb{P}\left(\sup_{a\in\mathcal{B}}\left|\sum_{ij} a_{ij}(A_{ij}-\theta_{ij})\right| > t\right) \leq \mathcal{N}\Big(1/2, \mathcal{B}, ||\cdot||\Big)\exp(-Ct^2),$$

*for some universal constant $C > 0$.*

PROOF OF LEMMA 4.3. For each $z \in \mathcal{Z}_{n,k}$, define the set $\mathcal{B}_z$ by

$$\mathcal{B}_z = \left\{\{c_{ij}\} : c_{ij} = Q_{ab} \text{ if } (i,j) \in z^{-1}(a)\times z^{-1}(b) \text{ for some } Q_{ab}, \text{ and } \sum_{ij} c_{ij}^2 \leq 1\right\}.$$

In other words, $\mathcal{B}_z$ collects those piecewise constant matrices determined by $z$. Thus, we have the bound

$$\left|\sum_{ij}\frac{\tilde{\theta}_{ij}-\hat{\theta}_{ij}}{\sqrt{\sum_{ij}(\tilde{\theta}_{ij}-\hat{\theta}_{ij})^2}}(A_{ij}-\theta_{ij})\right| \leq \max_{z\in\mathcal{Z}_{n,k}}\sup_{c\in\mathcal{B}_z}\left|\sum_{ij} c_{ij}(A_{ij}-\theta_{ij})\right|.$$

Note that for each $z \in \mathcal{Z}_{n,k}$, $\mathcal{B}_z$ satisfies the condition (A.12). Thus, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{z\in\mathcal{Z}_{n,k}}\sup_{c\in\mathcal{B}_z}\left|\sum_{ij} c_{ij}(A_{ij}-\theta_{ij})\right| > t\right)\\
\leq &\sum_{z\in\mathcal{Z}_{n,k}}\mathbb{P}\left(\sup_{c\in\mathcal{B}_z}\left|\sum_{ij} c_{ij}(A_{ij}-\theta_{ij})\right| > t\right)\\
\leq &\sum_{z\in\mathcal{Z}_{n,k}}\mathcal{N}\Big(1/2, \mathcal{B}_z, ||\cdot||\Big)\exp(-C_1t^2),
\end{aligned}$$

for some universal $C_1 > 0$, where the last inequality is due to Lemma A.1. Since $\mathcal{B}_z$ has a degree of freedom $k^2$, we have $\mathcal{N}\Big(1/2, \mathcal{B}_z, ||\cdot||\Big) \leq \exp(C_2k^2)$ for all $z \in \mathcal{Z}_{n,k}$, which is a direct consequence of covering number in $\mathbb{R}^{k^2}$ [46, Lemma 4.1]. Finally, by $|\mathcal{Z}_{n,k}| \leq \exp(n\log k)$, we have

$$\mathbb{P}\left(\max_{z\in\mathcal{Z}_{n,k}}\sup_{c\in\mathcal{B}_z}\sum_{i\neq j} c_{ij}(A_{ij}-\theta_{ij}) > t\right) \leq \exp(-C_1t^2 + C_2k^2 + n\log k).$$

Choosing $t^2 \propto k^2 + n \log k$, the proof is complete. □

PROOF OF LEMMA A.1. Let $\mathcal{B}'$ be a $1/2$-net of $\mathcal{B}$ such that $|\mathcal{B}'| \leq \mathcal{N}\Big(1/2, \mathcal{B}, ||\cdot||\Big)$ and for any $a \in \mathcal{B}$, there is $b \in \mathcal{B}'$ satisfying

$$||a - b|| \leq 1/2. \tag{A.13}$$

Thus,

$$\begin{aligned} |\langle a, A - \theta \rangle| &\leq |\langle a - b, A - \theta \rangle| + |\langle b, A - \theta \rangle| \\ &\leq ||a - b|| \left| \left\langle \frac{a - b}{||a - b||}, A - \theta \right\rangle \right| + |\langle b, A - \theta \rangle| \\ &\leq \frac{1}{2} \sup_{a \in \mathcal{B}} |\langle a, A - \theta \rangle| + |\langle b, A - \theta \rangle|, \end{aligned} \tag{A.14}$$

where the inequality (A.14) is due to (A.13) and the assumption (A.12). Taking sup and max on both sides, we have

$$\sup_{a \in \mathcal{B}} \left| \sum_{ij} a_{ij}(A_{ij} - \theta_{ij}) \right| \leq 2 \max_{b \in \mathcal{B}'} \left| \sum_{ij} b_{ij}(A_{ij} - \theta_{ij}) \right|.$$

Using Hoeffding's inequality [50, Prop 5.16] and union bound, the proof is complete. □

PROOF OF PROPOSITION 4.2. By definition,

$$\begin{aligned} D(\mathbb{P}_\theta || P_{\theta'}) &= \sum_{ij} \left( \theta_{ij} \log \frac{\theta_{ij}}{\theta'_{ij}} + (1 - \theta_{ij}) \log \frac{1 - \theta_{ij}}{1 - \theta'_{ij}} \right) \\ &\leq \sum_{ij} \frac{(\theta_{ij} - \theta'_{ij})^2}{\theta'_{ij}(1 - \theta'_{ij})} \\ &\leq 8 \sum_{ij} (\theta_{ij} - \theta'_{ij})^2, \end{aligned}$$

where we using the inequality $\log x \leq x - 1$ for $x > 0$ for the first inequality and the fact that $1/4 \leq \theta'_{ij} \leq 3/4$ for the second inequality. We then bound the chi-squared divergence in the same way,

$$\begin{aligned}\chi^2(\mathbb{P}_\theta||P_{\theta'}) &= \prod_{ij}\left(1+\frac{(\theta_{ij}-\theta'_{ij})^2}{\theta'_{ij}(1-\theta'_{ij})}\right)-1\\ &\leq \exp\left(\sum_{ij}\log\left(1+8(\theta_{ij}-\theta'_{ij})^2\right)\right)-1\\ &\leq \exp\left(8\sum_{ij}(\theta_{ij}-\theta'_{ij})^2\right).\end{aligned}$$

The proof is complete. □

Proof of Proposition A.1. By the definition of $K(x,y)$, we have

$$\begin{aligned}|K(x_1,y_1)-K(x_2,y_2)| &\leq 2\,||x_1|-|x_2||\,|1-2|y_1||+2\,||y_1|-|y_2||\,|1-2|x_2||\\ &\leq 2\left(|x_1-x_2|+|y_1-y_2|\right).\end{aligned}$$

For any $(x_1,y_1),(x_2,y_2)$ in the support of $\phi_{ab}$, we have

$$\begin{aligned}|\phi_{ab}(x_1,y_1)-\phi_{ab}(x_2,y_2)| &\leq 2Lk^{1-\alpha}\left(|x_1-x_2|+|y_1-y_2|\right)\\ &\leq 2L\left(|x_1-x_2|+|y_1-y_2|\right)^\alpha,\end{aligned}$$

where we have used $\alpha\in(0,1)$ and $|x_1-x_2|+|y_1-y_2|\leq k^{-1}$ in the last inequality. This means $\phi_{ab}\in\mathcal{H}_\alpha(M)$ for some sufficiently small $L$. This proves the first claim. For the second one, note that,

$$\sum_{i,j\in[n]}\phi_{ab}^2(i/n,j/n)=L^2k^{-2\alpha}\sum_{i,j\in[n]}K^2\left(\frac{ki}{n}-a+\frac{1}{2},\frac{kj}{n}-b+\frac{1}{2}\right),$$

where

$$\sum_{i,j\in[n]}K^2\left(\frac{ki}{n}-a+\frac{1}{2},\frac{kj}{n}-b+\frac{1}{2}\right)=\left(\sum_{\frac{n(a-1)}{k}<i\leq\frac{na}{k}}\left(1-2\left|\frac{ki}{n}-a+\frac{1}{2}\right|\right)^2\right)^2,$$

and

$$\begin{aligned}&\sum_{\frac{n(a-1)}{k}<i\leq\frac{na}{k}}\left(1-2\left|\frac{ki}{n}-a+\frac{1}{2}\right|\right)^2=\sum_{\frac{n(a-1)}{k}<i\leq\frac{na}{k}}\left(1-\frac{2k}{n}\left|i-\frac{(a-1/2)n}{k}\right|\right)^2\\ &= 4\sum_{0\leq t\leq\frac{n}{2k}}\left(1-\frac{2k}{n}t\right)^2\geq 2\int_0^{n/2k}\left(1-\frac{2k}{n}t\right)^2dt=\frac{n}{3k}.\end{aligned}$$

The proof is complete. □

**A.3. Proof of Theorem 3.6.** The upper bound is given in [34, 12]. It is sufficient to consider the lower bound. Without loss of generality, we may assume $\beta < 1/2$. The case $\beta \geq 1/2$ can be treated as in the proof of Theorem 2.2. Consider the set

$$T = \Bigg\{ \{\theta_{ij}\} \in [0,1]^{n\times n} : \theta_{ij} = \beta \text{ for } (i,j) \in (S\times S) \cup (S^c \times S^c), \\ \theta_{ij} = \beta - \frac{c\sqrt{\beta}}{\sqrt{n}} \text{ for } (i,j) \in (S\times S^c) \cup (S^c \times S), \text{ with some } S \in \mathcal{S} \Bigg\},$$

where $\mathcal{S} = \{S : S \subset [n]\}$ and $c \in (0, 1/2)$ is some constant to be determined later. Since $\beta \geq n^{-1}$, we must have $T \subset \Theta(\beta)$ and $T$ is our least favorable subset consisting of matrices with rank at most 2. The argument (3.2) implies that

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta(\beta)} \mathbb{E}\|\hat{\theta}-\theta\|_{\text{op}}^2 \geq \inf_{\hat{\theta}} \sup_{\theta\in\Theta(\beta)\cap\Theta_k} \mathbb{E}\|\hat{\theta}-\theta\|_{\text{op}}^2 \gtrsim \inf_{\hat{\theta}} \sup_{\theta\in T} \mathbb{E}||\hat{\theta}-\theta||^2.$$

Hence, it is sufficient to lower bound $\inf_{\hat{\theta}} \sup_{\theta\in T} \mathbb{E}||\hat{\theta}-\theta||^2$, which will be established similarly according to the argument in Section 3.3. Specifically, using the argument in the proof of Proposition 4.2, we have

$$\sup_{\theta,\theta'\in T} D(\mathbb{P}_\theta||\mathbb{P}_{\theta'}) \leq \sup_{\theta,\theta'\in T} \sum_{ij} \frac{(\theta_{ij}-\theta'_{ij})^2}{\theta'_{ij}(1-\theta'_{ij})} \leq 4c^2 n.$$

Moreover, for the distance $\rho(\theta,\theta') = n^{-1}||\theta-\theta'||$ and any $\theta,\theta' \in T$ associated with $S, S' \in \mathcal{S}$, we have

$$n^2\rho^2(\theta,\theta') = \frac{2c^2\beta}{n}|S\Delta S'|(n-|S\Delta S'|).$$

According to the argument in Section 3.3, this implies $\mathcal{M}(\epsilon, T, \rho) \geq N \geq \exp(c_1 n)$ for some $c_1 > 0$ with $\epsilon^2 = \frac{c^2\beta}{8n}$. Finally, applying (4.9) of Proposition 4.1, we have the desired result by letting $c$ be sufficiently small.

Department of Statistics
Yale University
New Haven, CT 06511
E-mail: chao.gao@yale.edu
E-mail: yu.lu@yale.edu
E-mail: huibin.zhou@yale.edu
URL: http://www.stat.yale.edu/~hz68/