The Annals of Statistics
2017, Vol. 45, No. 1, 316–354
DOI: 10.1214/16-AOS1454

ORACLE INEQUALITIES FOR NETWORK MODELS AND SPARSE GRAPHON ESTIMATION

BY OLGA KLOPP[*,1], ALEXANDRE B. TSYBAKOV[†,2] AND NICOLAS VERZELEN[‡,3]

Université Paris Ouest, ENSAE, UMR CNRS 9194[†] and INRA, UMR 729, MISTEA[‡]*

Inhomogeneous random graph models encompass many network models such as stochastic block models and latent position models. We consider the problem of statistical estimation of the matrix of connection probabilities based on the observations of the adjacency matrix of the network. Taking the stochastic block model as an approximation, we construct estimators of network connection probabilities—the ordinary block constant least squares estimator, and its restricted version. We show that they satisfy oracle inequalities with respect to the block constant oracle. As a consequence, we derive optimal rates of estimation of the probability matrix. Our results cover the important setting of sparse networks. Another consequence consists in establishing upper bounds on the minimax risks for graphon estimation in the L_2 norm when the probability matrix is sampled according to a graphon model. These bounds include an additional term accounting for the "agnostic" error induced by the variability of the latent unobserved variables of the graphon model. In this setting, the optimal rates are influenced not only by the bias and variance components as in usual nonparametric problems but also include the third component, which is the agnostic error. The results shed light on the differences between estimation under the empirical loss (the probability matrix estimation) and under the integrated loss (the graphon estimation).

1. Introduction. Consider a network defined as an undirected graph with n nodes. Assume that we observe the values $\mathbf{A}_{ij} \in \{0, 1\}$ where $\mathbf{A}_{ij} = 1$ is interpreted as the fact that the nodes i and j are connected and $\mathbf{A}_{ij} = 0$ otherwise. We set $\mathbf{A}_{ii} = 0$ for all $1 \le i \le n$ and we assume that $\mathbf{A}_{ij}$ is a Bernoulli random variable with parameter $(\boldsymbol{\Theta}_0)_{ij} = \mathbb{P}(\mathbf{A}_{ij} = 1)$ for $1 \le j < i \le n$. The random variables $\mathbf{A}_{ij}$, $1 \le j < i \le n$, are assumed independent. We denote by $\mathbf{A}$ the adjacency matrix, that is, the $n \times n$ symmetric matrix with entries $\mathbf{A}_{ij}$ for $1 \le j < i \le n$ and zero di-

Received July 2015; revised February 2016.

[1]Supported by the ANR under the grant Labex MME-DIIA (ANR11-LBX-0023-01).

[2]Supported by GENES, the ANR under the grants IPANEMA (ANR-13-BSH1-0004-02) and Labex ECODEC (ANR-11-LABEX-0047), and the "Chaire Economie et Gestion des Nouvelles Données", under the auspices of Institut Louis Bachelier, Havas-Media and Paris-Dauphine.

[3]Supported by the ANR under the grant Calibration (ANR-2011-BS01-010-01).

MSC2010 subject classifications. Primary 62G05; secondary 60C05.

Key words and phrases. Inhomogeneous random graph, networks, oracle inequality, sparse graphon, sparsity, stochastic block model.

agonal entries. Similarly, we denote by $\mathbf{\Theta}_0$ the $n \times n$ symmetric matrix with entries $(\mathbf{\Theta}_0)_{ij}$ for $1 \le j < i \le n$ and zero diagonal entries. This is a matrix of probabilities associated to the graph; the nodes i and j are connected with probability $(\mathbf{\Theta}_0)_{ij}$. The model with such observations $\mathbf{A}' = (\mathbf{A}_{ij},\, 1 \le j < i \le n)$ is a special case of inhomogeneous random graph model that we will call for definiteness the *network sequence model*[1] to emphasize a parallel with the Gaussian sequence model.

1.1. *Graphons and sparse graphon models.* Networks arise in many areas such as information technology, social life and genetics. These real-life networks are in permanent movement and often their size is growing. Therefore, it is natural to look for a well-defined "limiting object" independent of the network size n and such that a stochastic network can be viewed as a partial observation of this limiting object. Such objects, called the *graphons*, play a central role in the recent theory of graph limits introduced by Lovász and Szegedy [12]. For a detailed description of this theory, we refer to the monograph by Lovász [11]. Graphons are symmetric measurable functions $W : [0, 1]^2 \to [0, 1]$. In the sequel, the space of graphons is denoted by $\mathcal{W}$. The main message is that every graph limit can be represented by a graphon. This beautiful theory of graph limits was developed for *dense graphs*, that is, graphs with the number of edges comparable to the square of the number of vertices.

Graphons give a natural way of generating random graphs [9, 11]. The probability that two distinct nodes i and j are connected in the graphon model is the random variable

$$(\mathbf{\Theta}_0)_{ij} = W_0(\xi_i, \xi_j), \tag{1}$$

where $\xi_1, \dots, \xi_n$ are unobserved (latent) i.i.d. random variables uniformly distributed on $[0, 1]$. As above, the diagonal entries of $\mathbf{\Theta}_0$ are zero. Conditionally on $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$, the observations $\mathbf{A}_{ij}$ for $1 \le j < i \le n$ are assumed to be independent Bernoulli random variables with success probabilities $(\mathbf{\Theta}_0)_{ij}$. For any positive integer n, a graphon function W_0 defines a probability distribution on graphs of size n. Note that this model is different from the network sequence model since the observations $\mathbf{A}_{ij}$ are no longer independent. If W_0 is a step-function with k steps, we obtain an analog of the stochastic block model with k groups. More generally, many exchangeable distributions on networks [9], including random dot product graphs [14] and some geometric random graphs [13] can be expressed as graphons.

Given an observed adjacency matrix $\mathbf{A}'$ sampled according to the model (1), the graphon function W_0 is not identifiable. The topology of a network is invariant with respect to any change of labelling of its nodes. Consequently, for any given function $W_0(\cdot, \cdot)$ and a measure-preserving bijection $\tau : [0, 1] \to [0, 1]$ (with respect to

[1] In some recent papers, it is also called the inhomogeneous Erdős–Rényi model, which is somewhat ambiguous since the words "Erdős–Rényi model" designate a homogeneous graph.

the Lebesgue measure), the functions $W_0(x, y)$ and $W_0^\tau(x, y) := W_0(\tau(x), \tau(y))$ define the same probability distribution on random graphs. The equivalence classes of graphons defining the same probability distribution are characterized in a slightly more involved way based on the following result.

PROPOSITION 1.1 ([11], Section 10). *Two graphons U and W in $\mathcal{W}$ are called weakly isomorphic if there exist measure preserving maps ϕ, ψ: $[0, 1] \to [0, 1]$ such that $U^\phi = W^\psi$ almost everywhere. Two graphons U and W define the same probability distribution if and only if they are weakly isomorphic.*

This proposition motivates considering the equivalence classes of graphons that are weakly isomorphic. The corresponding quotient space is denoted by $\widetilde{\mathcal{W}}$.

It is easy to see that the expected number of edges in model (1) is a constant times the squared number of vertices, which corresponds to the dense case. Networks observed in the real life are often *sparse* in the sense that the number of edges is of the order $o(n^2)$ as the number n of vertices tends to infinity. The situation in the sparse case is more complex than in the dense case. Even the almost dense case with $n^{2-o(1)}$ edges is rather different from the dense case. The extremely sparse case corresponds to the *bounded degree* graph where all degrees are smaller than a fixed positive integer. It is an opposite extreme to the dense graph. Networks that occur in the applications are usually between these two extremes described by dense and bounded degree graphs. They often correspond to inhomogeneous networks with density of edges tending to 0 but with the maximum degree tending to infinity as n grows.

For a given $\rho_n > 0$, one can modify the definition (1) to get a random graph model with $O(\rho_n n^2)$ edges. It is usually assumed that $\rho_n \to 0$ as $n \to \infty$. The adjacency matrix $\mathbf{A}'$ is sampled according to graphon $W_0 \in \mathcal{W}$ with scaling parameter ρ_n if for all $j < i$,

$$(\mathbf{\Theta}_0)_{ij} = \rho_n W_0(\xi_i, \xi_j), \tag{2}$$

where $\rho_n > 0$ is the scale parameter that can be interpreted as the expected proportion of nonzero edges. Alternatively, model (2) can be considered as a graphon model (1) that has been sparsified in the sense that its edges have been independently removed with probability $1 - \rho_n$ and kept with probability ρ_n. This sparse graphon model was considered in [2, 3, 16, 17].

1.2. *Our results.* This paper has two contributions. First, we study optimal rates of estimation of the probability matrix $\mathbf{\Theta}_0$ under the Frobenius norm from an observation $\mathbf{A}' = (\mathbf{A}_{ij}, 1 \le j < i \le n)$ in the network sequence model. We estimate $\mathbf{\Theta}_0$ by a block-constant matrix and we focus on deriving oracle inequalities with optimal rates (with possibly nonpolynomial complexity of estimation methods). Note that estimating $\mathbf{\Theta}_0$ by a $k \times k$ block constant matrix is equivalent to fitting a stochastic block model with k classes. Estimation of $\mathbf{\Theta}_0$ has already been

considered by [5, 6, 17] but convergence rates obtained there are far from being optimal. More recently, Gao, Lu and Zhou [10] have established the minimax estimation rates for $\mathbf{\Theta}_0$ on classes of block constant matrices and on the smooth graphon classes. Their analysis is restricted to the dense case (1) corresponding to $\rho_n = 1$ when dealing with model (2). In this paper, we explore the general setting of model (2). In particular, our aim is to understand the behavior of least squares estimators when the probabilities $(\mathbf{\Theta}_0)_{ij}$ can be arbitrarily small. This will be done via developing oracle inequalities with respect to the block constant oracle. Two estimators will be considered—the ordinary block constant least squares estimator, and its restricted version where the estimator is chosen in the ℓ_∞ cube of a given radius. As corollaries, we provide an extension for the *sparse* graphon model of some minimax results in [10] and we quantify the impact of the scaling parameter ρ_n on the optimal rates of convergence.

Second, we consider estimation of the graphon function W_0 based on observation $\mathbf{A}'$. In view of Proposition 1.1, graphons are not identifiable and can only be estimated up to weak isomorphisms. Hence, we study estimation of W_0 in the quotient space $\widetilde{\mathcal{W}}$ of graphons. In order to contrast this problem with the estimation of $\mathbf{\Theta}_0$, one can invoke an analogy with the random design nonparametric regression. Suppose that we observe (y_i, ξ_i), $i = 1, \dots, n$, that are independently sampled according to the model $y = f(\xi) + \varepsilon$ where f is an unknown regression function, ε is a zero mean random variable and ξ is distributed with some density h on $[0, 1]$. Given a sample of (y_i, ξ_i), estimation of f with respect to the *empirical* loss is equivalent to estimation of the vector $(f(\xi_1), \dots, f(\xi_n))$ in, for instance, the Euclidean norm. On the other hand, estimation under the integrated loss consists in constructing an estimator $\hat{f}$ such that the integral $\int (\hat{f}(t) - f(t))^2 h(t)\, dt$ is small. Following this analogy, estimation of $\mathbf{\Theta}_0$ corresponds to an empirical loss problem whereas the graphon estimation corresponds to an integrated loss problem. However, as opposed to nonparametric regression, in the graphon models (1) and (2) the design $\xi_1, \dots, \xi_n$ is not observed, which makes it quite challenging to derive the convergence rates in these settings.

In Section 3, we obtain L_2 norm nonasymptotic estimation rates for graphon estimation on classes of step functions (analogs of stochastic block models) and on classes of smooth graphons in model (2). This result improves upon previously known bounds by Wolfe and Olhede [16]. For classes of step function graphons, we also provide a matching minimax lower bound allowing one to characterize the regimes such that graphon optimal estimation rates are slower than probability matrix estimation rates. In a work parallel to ours, Borgs, Chayes and Smith [4] have analyzed the rates of convergence for estimation of step function graphons under the privacy model. Apart from the issues of privacy that we do not consider here, their results in our setting provide a weaker version of the upper bound in Corollary 3.3, with a suboptimal rate, which is the square root of the rate of Corollary 3.3 in the moderately sparse zone. We also mention the paper by Choi [7] devoted to the convergence of empirical risk functions associated to the graphon model.

1.3. *Notation.* We provide a brief summary of the notation used throughout this paper:

- For a matrix $\mathbf{B}$, we denote by $\mathbf{B}_{ij}$ [or by $\mathbf{B}_{i,j}$, or by $(\mathbf{B})_{ij}$] its (i, j)th entry.
- For an integer m, set $[m] = \{1, \dots, m\}$.
- Let n, k and n_0 be integers such that $2 \le k \le n$, $n_0 \le n$. We denote by $\mathcal{Z}_{n,k,n_0}$ the set of all mappings z from $[n]$ to $[k]$ such that $\min_{a=1,\dots,k} |z^{-1}(a)| \ge n_0$ (the minimal size of each "community" is n_0). For brevity, we set $\mathcal{Z}_{n,k} = \mathcal{Z}_{n,k,1}$.
- We denote by $\mathbb{R}^{k\times k}_{\mathrm{sym}}$ the class of all symmetric $k \times k$ matrices with real-valued entries.
- The inner product between matrices $\mathbf{D}, \mathbf{B} \in \mathbb{R}^{n\times n}$ will be denoted by $\langle \mathbf{D}, \mathbf{B}\rangle = \sum_{i,j} \mathbf{D}_{ij}\mathbf{B}_{ij}$.
- Denote by $\|B\|_F$ and by $\|B\|_\infty$ the Frobenius norm and the entry-wise supremum norm of matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$, respectively.
- We denote by $\lfloor x \rfloor$ the maximal integer less than $x \ge 0$ and by $\lceil x \rceil$ the smallest integer greater than or equal to x.
- We denote by $\mathbb{E}$ the expectation with respect to the distribution of $\mathbf{A}$ if we consider the network sequence model and the expectation with respect to the joint distribution of $(\boldsymbol{\xi}, \mathbf{A})$ if we consider the graphon model.
- We denote by C positive constants that can vary from line to line. These are absolute constants unless otherwise mentioned.
- We denote by λ the Lebesgue measure on the interval $[0, 1]$.

2. Probability matrix estimation. In this section, we deal with the network sequence model and we obtain optimal rates of estimation in the Frobenius norm of the probability matrix $\boldsymbol{\Theta}_0$. Fixing some integer $k > 0$, we estimate the matrix $\boldsymbol{\Theta}_0$ by a block constant matrix with $k \times k$ blocks.

2.1. *Least squares estimator.* First, we study the least squares estimator $\widehat{\boldsymbol{\Theta}}$ of $\boldsymbol{\Theta}_0$ in the collection of all $k \times k$ block constant matrices with block size larger than some given integer n_0. Recall that we denote by $\mathcal{Z}_{n,k,n_0}$ the set of all possible mappings z from $[n]$ to $[k]$ such that $\min_{a=1,\dots,k} |z^{-1}(a)| \ge n_0$. For any $z \in \mathcal{Z}_{n,k,n_0}$, $\mathbf{Q} \in \mathbb{R}^{k\times k}_{\mathrm{sym}}$, we define the residual sum of squares

$$L(\mathbf{Q}, z) = \sum_{(a,b)\in[k]\times[k]} \sum_{(i,j)\in z^{-1}(a)\times z^{-1}(b),\, j<i} (\mathbf{A}_{ij} - \mathbf{Q}_{ab})^2$$

and consider the least squares estimator of $(\mathbf{Q}, z)$:

$$(\widehat{\mathbf{Q}}, \hat{z}) \in \arg\min_{\mathbf{Q}\in\mathbb{R}^{k\times k}_{\mathrm{sym}}, z\in\mathcal{Z}_{n,k,n_0}} L(\mathbf{Q}, z).$$

The block constant least squares estimator of $(\boldsymbol{\Theta}_0)_{ij}$ is defined as $\widehat{\boldsymbol{\Theta}}_{ij} = \widehat{\mathbf{Q}}_{\hat{z}(i)\hat{z}(j)}$ for all $i > j$. Note that $\widehat{\boldsymbol{\Theta}}_{ij} \in [0, 1]$. Finally, we denote by $\widehat{\boldsymbol{\Theta}}$ the symmetric matrix with entries $\widehat{\boldsymbol{\Theta}}_{ij}$ for all $i > j$ and with zero diagonal entries. According to the

stochastic block models terminology, $\widehat{\mathbf{Q}}$ stands for the estimated connection probabilities whereas $\hat{z}$ is the estimated partition of the of nodes into k communities. The only difference between this estimator and the one considered in [10] is the restriction $|z^{-1}(a)| \geq n_0$. This prevents the partition z from being too unbalanced. A common choice of n_0 is of the order n/k, which is referred to as balanced partition. Taking larger n_0 when we have additional information allows us to obtain simpler estimators since we reduce the number of configurations.

2.2. *Restricted least squares estimator.* Given $r \in (0, 1]$, consider the least squares estimator restricted to the l_∞ ball of radius r,

$$(\widehat{\mathbf{Q}}_r, \hat{z}_r) \in \arg \min_{\mathbf{Q} \in \mathbb{R}^{k\times k}_{\mathrm{sym}} : \|\mathbf{Q}\|_\infty \leq r, z \in \mathcal{Z}_{n,k}} L(\mathbf{Q}, z).$$

The estimator $\widehat{\mathbf{\Theta}}^r$ of matrix $\mathbf{\Theta}_0$ is defined as the symmetric matrix with entries $\widehat{\mathbf{\Theta}}^r_{ij} = (\widehat{\mathbf{Q}}_r)_{\hat{z}_r(i)\hat{z}_r(j)}$ for all $i > j$, and with zero diagonal entries. Note that here we consider any partitions, including really unbalanced ones ($n_0 = 1$).

2.3. *Oracle inequalities.* Let $\mathbf{\Theta}_{*,n_0}$ be the best Frobenius norm approximation of $\mathbf{\Theta}_0$ in the collection of matrices

$$\mathcal{T}_{n_0}[k] = \big\{\mathbf{\Theta} : \exists z \in \mathcal{Z}_{n,k,n_0}, \mathbf{Q} \in \mathbb{R}^{k\times k}_{\mathrm{sym}} \text{ such that } \mathbf{\Theta}_{ij} = \mathbf{Q}_{z(i)z(j)}, i \neq j, \text{ and } \mathbf{\Theta}_{ii} = 0 \ \forall i\big\}.$$

In particular, $\mathcal{T}_1[k]$ is the set of all probability matrices corresponding to k-class stochastic block models without group size restriction. For brevity, we write $\mathcal{T}_1[k] = \mathcal{T}[k]$, and $\mathbf{\Theta}_{*,1} = \mathbf{\Theta}_*$.

PROPOSITION 2.1. *Consider the network sequence model. There exist positive absolute constants C_1, C_2, C_3 such that for $n_0 \geq 2$,*

$$\mathbb{E}\bigg[\frac{1}{n^2}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F^2\bigg] \leq \frac{C_1}{n^2}\|\mathbf{\Theta}_0 - \mathbf{\Theta}_{*,n_0}\|_F^2 + C_2\|\mathbf{\Theta}_0\|_\infty\bigg(\frac{\log(k)}{n} + \frac{k^2}{n^2}\bigg) + \frac{C_3\log(n/n_0)}{n_0}\bigg(\frac{\log(k)}{n} + \frac{k^2}{n^2}\bigg). \tag{3}$$

To discuss some consequences of this oracle inequality, we consider the case of balanced partitions.

COROLLARY 2.2. *Consider the network sequence model. Let $n_0 \geq Cn/k$ for some $C > 0$ (balanced partition), $n_0 \geq 2$, and $\|\mathbf{\Theta}_0\|_\infty \geq \frac{k\log(k)}{n}$. Then, there exist positive absolute constants C_1, C_2, such that*

$$\mathbb{E}\bigg[\frac{1}{n^2}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F^2\bigg] \leq \frac{C_1}{n^2}\|\mathbf{\Theta}_0 - \mathbf{\Theta}_{*,n_0}\|_F^2 + C_2\|\mathbf{\Theta}_0\|_\infty\bigg(\frac{\log(k)}{n} + \frac{k^2}{n^2}\bigg). \tag{4}$$

In particular, if $\mathbf{\Theta}_0 \in \mathcal{T}_{n_0}[k]$ (i.e., when we have a k-class stochastic block model with balanced communities), the rate of convergence is of the order $\|\mathbf{\Theta}_0\|_\infty(\frac{\log(k)}{n} + \frac{k^2}{n^2})$. In [10], the rate $(\frac{\log(k)}{n} + \frac{k^2}{n^2})$ is shown to be minimax over all stochastic block models with k blocks, with no restriction on $\|\mathbf{\Theta}_0\|_\infty$ except the obvious one, $\|\mathbf{\Theta}_0\|_\infty \le 1$. We prove in the next subsection that the rate $\|\mathbf{\Theta}_0\|_\infty(\frac{\log(k)}{n} + \frac{k^2}{n^2})$ obtained in Corollary 2.2 exhibits the optimal dependency on $\|\mathbf{\Theta}_0\|_\infty$. Before doing this, we provide an oracle inequality for the restricted least squares estimator $\widehat{\mathbf{\Theta}}^r$.

PROPOSITION 2.3. *Consider the network sequence model. There exist positive absolute constants C_1 and C_2 such that the following holds. If $\|\mathbf{\Theta}_0\|_\infty \le r$, then*

$$\mathbb{E}\left[\frac{1}{n^2}\|\widehat{\mathbf{\Theta}}^r - \mathbf{\Theta}_0\|_F^2\right] \le \frac{C_1}{n^2}\|\mathbf{\Theta}_0 - \mathbf{\Theta}_*\|_F^2 + C_2 r\left(\frac{\log(k)}{n} + \frac{k^2}{n^2}\right). \tag{5}$$

As opposed to Corollary 2.2, the risk bound (5) is applicable for unbalanced partitions and for arbitrarily small values of $\|\mathbf{\Theta}_0\|_\infty$. However, this restricted least squares estimator requires the knowledge of an upper bound of $\|\mathbf{\Theta}_0\|_\infty$. Whereas the mean value of $(\mathbf{\Theta}_0)_{ij}$ is easily inferred from the data, the maximal value $\|\mathbf{\Theta}_0\|_\infty$ is difficult to estimate. If the matrix $\mathbf{\Theta}_0$ satisfies the sparse graphon model (2), one can set $r = r_n = u_n\overline{A}$ where $\overline{A} = 2\sum_{j<i}\mathbf{A}_{i,j}/(n(n-1))$ is the edge density of the graph and u_n is any sequence that tends to infinity (e.g., $u_n = \log\log n$). For n large enough, $\overline{A}/\rho_n$ is close to $\int_{[0,1]^2} W_0(x,y)\,dx\,dy$ in the sparse graphon model (2) with probability close to one and, therefore, r_n is greater than $\|\mathbf{\Theta}_0\|_\infty$. The price to pay for this data-driven choice of r_n is that the rate in the risk bound (5) is multiplied by u_n.

2.4. *Stochastic block models.* Given an integer k and any $\rho_n \in (0,1]$, consider the set of all probability matrices corresponding to k-class stochastic block model with connection probability uniformly smaller than ρ_n:

$$\mathcal{T}[k,\rho_n] = \{\mathbf{\Theta}_0 \in \mathcal{T}[k] : \|\mathbf{\Theta}_0\|_\infty \le \rho_n\}. \tag{6}$$

Gao, Lu and Zhou [10] have proved that the minimax estimation rate over $\mathcal{T}[k,1]$ is of order the $\frac{k^2}{n^2} + \frac{\log(k)}{n}$. The following proposition extends their lower bound to arbitrarily small $\rho_n > 0$.

PROPOSITION 2.4. *Consider the network sequence model. For all $k \le n$ and all $0 < \rho_n \le 1$,*

$$\inf_{\widehat{\mathbf{T}}}\sup_{\mathbf{\Theta}_0\in\mathcal{T}[k,\rho_n]}\mathbb{E}_{\mathbf{\Theta}_0}\left[\frac{1}{n^2}\|\widehat{\mathbf{T}} - \mathbf{\Theta}_0\|_F^2\right] \ge C\min\left(\rho_n\left(\frac{\log(k)}{n} + \frac{k^2}{n^2}\right), \rho_n^2\right), \tag{7}$$

where $\mathbb{E}_{\mathbf{\Theta}_0}$ denotes the expectation with respect to the distribution of $\mathbf{A}$ when the underlying probability matrix is $\mathbf{\Theta}_0$ and $\inf_{\widehat{\mathbf{T}}}$ is the infimum over all estimators.

If $\rho_n \geq \frac{\log(k)}{n} + \frac{k^2}{n^2}$, the minimax rate of estimation is of the order $\rho_n(\frac{\log(k)}{n} + \frac{k^2}{n^2})$. This rate is achieved by the restricted least squares estimator with $r \asymp \rho_n$ and by the least squares estimator if the partition is balanced and $\rho_n \geq k\log(k)/n$. For really sparse graphs (ρ_n smaller than $\frac{\log(k)}{n} + \frac{k^2}{n^2}$), the estimation problem becomes rather trivial since both the null estimator $\widehat{\mathbf{T}} = 0$ and the constant least squares estimator $\widehat{\mathbf{\Theta}}$ with all entries $\widehat{\mathbf{\Theta}}_{ij} = \overline{A}$ achieve the optimal rate ρ_n^2.

2.5. *Smooth graphons.* We now use Propositions 2.1 and 2.3 to obtain rates of convergence for the probability matrix estimation when $\mathbf{\Theta}_0$ satisfies the sparse graphon model (2) and the graphon $W_0 \in \mathcal{W}$ is smooth. For any $\alpha > 0$, $L > 0$, define the class of α-Hölder continuous functions $\Sigma(\alpha, L)$ as the set of all functions $W : [0,1]^2 \to [0,1]$ such that

$$\left|W(x', y') - \mathcal{P}_{\lfloor\alpha\rfloor}((x, y), (x' - x, y' - y))\right| \leq L\left(|x' - x|^{\alpha - \lfloor\alpha\rfloor} + |y' - y|^{\alpha - \lfloor\alpha\rfloor}\right)$$

for all $(x', y'), (x, y) \in [0,1]^2$, where $\lfloor\alpha\rfloor$ is the maximal integer less than α, and the function $(x', y') \mapsto \mathcal{P}_{\lfloor\alpha\rfloor}((x, y), (x' - x, y' - y))$ is the Taylor polynomial of degree $\lfloor\alpha\rfloor$ at point (x, y). It follows from the standard embedding theorems that, for any $W \in \Sigma(\alpha, L)$ and any $(x', y'), (x, y) \in [0,1]^2$,

$$\left|W(x', y') - W(x, y)\right| \leq M\left(|x' - x|^{\alpha\wedge 1} + |y' - y|^{\alpha\wedge 1}\right), \tag{8}$$

where M is a constant depending only on α and L. In the following, we will use only this last property of $W \in \Sigma(\alpha, L)$.

The following two propositions give bounds on the bias terms $\|\mathbf{\Theta}_0 - \mathbf{\Theta}_{*,n_0}\|_F^2$ and $\|\mathbf{\Theta}_0 - \mathbf{\Theta}_*\|_F^2$.

PROPOSITION 2.5. *Consider the graphon model* (2) *with* $W_0 \in \Sigma(\alpha, L)$ *where* $\alpha, L > 0$. *Let* $n_0 \geq 2$ *and* $k = \lfloor n/n_0 \rfloor$. *Then*

$$\mathbb{E}\left(\frac{1}{n^2}\|\mathbf{\Theta}_0 - \mathbf{\Theta}_{*,n_0}\|_F^2\right) \leq CM^2\rho_n^2\left(\frac{1}{k^2}\right)^{\alpha\wedge 1}. \tag{9}$$

We will also need the following proposition proved in [10], Lemma 2.1.

PROPOSITION 2.6. *Consider the graphon model* (2) *with* $W_0 \in \Sigma(\alpha, L)$ *where* $\alpha, L > 0$. *Then, almost surely,*

$$\frac{1}{n^2}\|\mathbf{\Theta}_0 - \mathbf{\Theta}_*\|_F^2 \leq CM^2\rho_n^2\left(\frac{1}{k^2}\right)^{\alpha\wedge 1}. \tag{10}$$

COROLLARY 2.7. *Consider the graphon model* (2) *with* $W_0 \in \Sigma(\alpha, L)$ *where* $\alpha, L > 0$ *and* $0 < \rho_n \leq 1$. *Set* $k = \lceil(\rho_n^{1/2}n)^{1/(1+\alpha\wedge 1)}\rceil$.

(i) *Assume that* $\rho_n \geq n^{-2(\alpha\wedge 1)/(1+2(\alpha\wedge 1))}(\log n)^{2(1+\alpha\wedge 1)}$ *and there exists* $n_0 \geq 2$ *such that* $k = \lfloor n/n_0 \rfloor$. *Then there exists a positive constant* C *depending only on* L *and* α *such that the least squares estimator* $\widehat{\boldsymbol{\Theta}}$ *constructed with this choice of* n_0 *satisfies*

$$\mathbb{E}\left[\frac{1}{n^2}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F^2\right] \leq C\left\{\rho_n^{(2+\alpha\wedge 1)/(1+\alpha\wedge 1)} n^{-2(\alpha\wedge 1)/(1+\alpha\wedge 1)} + \frac{\rho_n \log n}{n}\right\}. \tag{11}$$

(ii) *Assume that* $r \geq \rho_n \geq Cn^{-2}$. *Then there exists a positive constant* C *depending only on* L *and* α *such that the restricted least squares estimator* $\widehat{\boldsymbol{\Theta}}^r$ *satisfies*

$$\mathbb{E}\left[\frac{1}{n^2}\|\widehat{\boldsymbol{\Theta}}^r - \boldsymbol{\Theta}_0\|_F^2\right] \leq C\left\{r^{(2+\alpha\wedge 1)/(1+\alpha\wedge 1)} n^{-2(\alpha\wedge 1)/(1+\alpha\wedge 1)} + \frac{r \log n}{n}\right\}. \tag{12}$$

Corollary 2.7 extends the results obtained in Gao, Lu and Zhou [10] to arbitrary $\rho_n \in (0, 1]$. To simplify the discussion assume that $\alpha \leq 1$. There are two ingredients in the rates of convergence in Corollary 2.7, the nonparametric rate $\rho_n^{(2+\alpha)/(1+\alpha)} n^{-2\alpha/(1+\alpha)}$ and the clustering rate $\frac{\rho_n \log n}{n}$. The smoothness index α has an impact on the rate only if $\alpha \in (0, 1)$ and only if the network is not too sparse, that is, if $\rho_n \geq Cn^{\alpha-1}(\log n)^{1+\alpha}$.

In [10], Gao, Lu and Zhou prove a lower bound showing that the rate $n^{-2(\alpha\wedge 1)/(1+\alpha\wedge 1)} + \frac{\log(n)}{n}$ is optimal if $\rho_n = 1$. Following the same lines as in Proposition 2.4, one can readily extend their result to prove that the rate in (11) is minimax optimal for $\rho_n \geq n^{-2+\varepsilon}$ with an arbitrarily small $\varepsilon > 0$.

3. Graphon estimation problem. In this section, our purpose is to estimate the graphon function $W_0(\cdot, \cdot)$ in the sparse graphon model (2).

3.1. *From probability matrix estimation to graphon estimation.* We start by associating a graphon to any $n \times n$ probability matrix $\boldsymbol{\Theta}$. This provides a way of deriving an estimator of $f_0(\cdot, \cdot) = \rho_n W_0(\cdot, \cdot)$ from an estimator of $\boldsymbol{\Theta}_0$.

Given a $n \times n$ matrix $\boldsymbol{\Theta}$ with entries in $[0, 1]$, define the *empirical graphon* $\widetilde{f}_{\boldsymbol{\Theta}}$ as the following piecewise constant function:

$$\widetilde{f}_{\boldsymbol{\Theta}}(x, y) = \boldsymbol{\Theta}_{\lceil nx \rceil, \lceil ny \rceil} \tag{13}$$

for all x and y in $(0, 1]$. The empirical graphons associated to the least squares estimator and to the restricted least squares estimator with threshold r will be denoted by $\widehat{f}$ and $\widehat{f}_r$, respectively:

$$\widehat{f} = \widetilde{f}_{\widehat{\boldsymbol{\Theta}}}, \qquad \widehat{f}_r = \widetilde{f}_{\widehat{\boldsymbol{\Theta}}^r}.$$

They will be used as graphon estimators. For any estimator $\check{f}$ of $f_0 = \rho_n W_0$, define the squared error

$$\delta^2(\check{f}, f_0) := \inf_{\tau \in \mathcal{M}} \int\int_{(0,1)^2} \left| f_0\big(\tau(x), \tau(y)\big) - \check{f}(x, y)\right|^2 dx\, dy, \tag{14}$$

where $\mathcal{M}$ is the set of all measure-preserving bijections $\tau : [0,1] \to [0,1]$. It has been proved in [11], Chapters 8 and 13, that $\delta(\cdot,\cdot)$ defines a metric on the quotient space $\widetilde{\mathcal{W}}$ of graphons.

The following lemma is a simple consequence of the triangle inequality.

LEMMA 3.1. *Consider the graphon model* (2). *For any* $W_0 \in \mathcal{W}$, $\rho_n > 0$, *and any estimator* $\widehat{\mathbf{T}}$ *of* $\boldsymbol{\Theta}_0$ *such that* $\widehat{\mathbf{T}}$ *is an* $n \times n$ *matrix with entries in* $[0,1]$, *we have*

$$\mathbb{E}\big[\delta^2(\widetilde{f}_{\widehat{\mathbf{T}}}, f_0)\big] \le 2\mathbb{E}\bigg[\frac{1}{n^2}\|\widehat{\mathbf{T}} - \boldsymbol{\Theta}_0\|_F^2\bigg] + 2\mathbb{E}\big[\delta^2(\widetilde{f}_{\boldsymbol{\Theta}_0}, f_0)\big]. \tag{15}$$

The bound on the integrated risk in (15) is a sum of two terms. The first term containing $\|\widehat{\mathbf{T}} - \boldsymbol{\Theta}_0\|_F^2$ has been considered in Section 2 for $\widehat{\mathbf{T}} = \widehat{\boldsymbol{\Theta}}$ and $\widehat{\mathbf{T}} = \widehat{\boldsymbol{\Theta}}^r$. It is the *estimation error* term. The second term containing $\delta^2(\widetilde{f}_{\boldsymbol{\Theta}_0}, f_0)$ measures the distance between the true graphon f_0 and its discretized version sampled at the unobserved random design points $\xi_1, \dots, \xi_n$. We call it the *agnostic error*. The behavior of $\delta^2(\widetilde{f}_{\boldsymbol{\Theta}_0}, f_0)$ depends on the topology of the considered graphons as shown below for two examples, namely, the step function graphons and the smooth graphons.

3.2. *Step function graphons*. Define $\mathcal{W}[k]$ the collection of k-step graphons, that is, the subset of graphons $W \in \mathcal{W}$ such that for some $\mathbf{Q} \in \mathbb{R}^{k\times k}_{\mathrm{sym}}$ and some $\phi : [0,1] \to [k]$,

$$W(x,y) = \mathbf{Q}_{\phi(x),\phi(y)} \qquad \text{for all } x, y \in [0,1]. \tag{16}$$

A step function $W \in \mathcal{W}[k]$ is called balanced if $\lambda(\phi^{-1}(1)) = \cdots = \lambda(\phi^{-1}(k)) = 1/k$ where λ is the Lebesgue measure on $[0,1]$. The agnostic error associated to step function graphons is evaluated as follows.

PROPOSITION 3.2. *Consider the graphon model* (2). *For all integers* $k \le n$, $W_0 \in \mathcal{W}[k]$ *and* $\rho_n > 0$ *we have*

$$\mathbb{E}\big[\delta^2(\widetilde{f}_{\boldsymbol{\Theta}_0}, f_0)\big] \le C\rho_n^2\sqrt{\frac{k}{n}}.$$

Combining this result with Lemma 3.1 and Propositions 2.1 and 2.3, we obtain the following risk bounds for the least squares and restricted least squares graphon estimators.

COROLLARY 3.3. *Consider the graphon model* (2) *with* $W_0 \in \mathcal{W}[k]$. *There exist absolute constants* C_1 *and* C_2 *such that the following holds.*

(i) *If* $\rho_n \geq \frac{k\log(k)}{n}$, $n_0 = \lfloor n/2k \rfloor$, *and the step function* W_0 *is balanced, then the least squares graphon estimator* $\widehat{f}$ *constructed with this choice of* n_0 *satisfies*

$$\mathbb{E}\big[\delta^2(\widehat{f}, f_0)\big] \leq C_1\bigg[\rho_n\bigg(\frac{k^2}{n^2} + \frac{\log(k)}{n}\bigg) + \rho_n^2\sqrt{\frac{k}{n}}\bigg].$$

(ii) *If* $\rho_n \leq r$, *the restricted least squares graphon estimator* $\widehat{f}_r$ *satisfies*

$$\mathbb{E}\big[\delta^2(\widehat{f}_r, f_0)\big] \leq C_2\bigg[r\bigg(\frac{k^2}{n^2} + \frac{\log(k)}{n}\bigg) + \rho_n^2\sqrt{\frac{k}{n}}\bigg]. \tag{17}$$

As an immediate consequence of this corollary, we get the following upper bound on the minimax risk:

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}[k]} \mathbb{E}_{W_0}\big[\delta^2(\widehat{f}, f_0)\big] \leq C_3\bigg\{\bigg[\rho_n\bigg(\frac{k^2}{n^2} + \frac{\log(k)}{n}\bigg) + \rho_n^2\sqrt{\frac{k}{n}}\bigg] \wedge \rho_n^2\bigg\}, \tag{18}$$

where C_3 is an absolute constant. Here, $\mathbb{E}_{W_0}$ denotes the expectation with respect to the distribution of observations $\mathbf{A}' = (\mathbf{A}_{ij}, 1 \leq j < i \leq n)$ when the underlying sparse graphon is $\rho_n W_0$ and $\inf_{\widehat{f}}$ is the infimum over all estimators. The bound (18) follows from (17) with $r = \rho_n$ and the fact that the risk of the null estimator $\widetilde{f} \equiv 0$ is smaller than ρ_n^2.

The next proposition shows that the upper bound (18) is optimal in a minimax sense (up to a logarithmic factor in k in one of the regimes).

PROPOSITION 3.4. *There exists a universal constant* $C > 0$ *such that for any sequence* $\rho_n > 0$ *and any positive integer* $2 \leq k \leq n$,

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}[k]} \mathbb{E}_{W_0}\big[\delta^2(\widehat{f}, f_0)\big] \geq C\bigg\{\bigg[\rho_n\bigg(\frac{k^2}{n^2} + \frac{1}{n}\bigg) + \rho_n^2\sqrt{\frac{k-1}{n}}\bigg] \wedge \rho_n^2\bigg\}, \tag{19}$$

where $\mathbb{E}_{W_0}$ *denotes the expectation with respect to the distribution of observations* $\mathbf{A}' = (\mathbf{A}_{ij}, 1 \leq j < i \leq n)$ *when the underlying sparse graphon is* $\rho_n W_0$ *and* $\inf_{\widehat{f}}$ *is the infimum over all estimators.*

The proof is given in Section 4.9. Showing the rate $\rho_n^2\sqrt{(k-1)/n}$ relies on two arguments. First, we construct a family of graphons that are well separated with respect to the $\delta(\cdot,\cdot)$ distance. The difficulty comes from the fact that this is not a classical L_2 distance but a minimum of the L_2 distance over all measure-preserving transformations. Second, we show that the behavior of the Kullback–Leibler divergences between the graphons in this family is driven by the randomness of the latent variables $\xi_1, \ldots, \xi_n$ while the nature of the connection probabilities can be neglected. This argument is similar to the "data processing" inequality.

Proposition 3.4 does not cover the case $k = 1$ (Erdős–Rényi models). In fact, in this case the constant graphon estimator $\widehat{f} = \widetilde{f_{\bar{A}}}$ achieves the optimal rate, which

is equal to ρ_n/n^2. The suboptimality of the bound of Proposition 3.2 for $k=1$ is due to the fact that the diagonal blocks of $\widehat{f}$ and $\widehat{f}^r$ are all equal to zero. By carefully defining the diagonal blocks of these estimators, we could have achieved the optimal bound $\rho_n^2\sqrt{(k-1)/n}$ but this would need a modification of the notation and of the proofs.

The bounds (18) and (19) imply that there are three regimes depending on the sparsity parameter ρ_n:

(i) *Weakly sparse graphs*: $\rho_n \geq \frac{\log(k)}{\sqrt{kn}} \vee (\frac{k}{n})^{3/2}$. The minimax risk is of the order $\rho_n^2\sqrt{k/n}$, and thus it is driven by the agnostic error arising from the lack of knowledge of the design.

(ii) *Moderately sparse graphs*: $\frac{\log(k)}{n} \vee (\frac{k}{n})^2 \leq \rho_n \leq \frac{\log(k)}{\sqrt{kn}} \vee (\frac{k}{n})^{3/2}$. The risk bound (18) is driven by the probability matrix estimation error. The upper bound (18) is of the order $\rho_n(\frac{k^2}{n^2} + \frac{\log(k)}{n})$, which is the optimal rate of probability matrix estimation, cf. Proposition 2.4. Due to (19), it is optimal up to $\log(k)$ factor with respect to the $\delta(\cdot,\cdot)$ distance.

(iii) *Highly sparse graphs*: $\rho_n \leq \frac{\log(k)}{n} \vee (\frac{k}{n})^2$. The minimax risk is of the order ρ_n^2, and it is attained by the null estimator.

In a work parallel to ours, Borgs, Chayes and Smith [4] provide an upper bound for the risk of step function graphon estimators in the context of privacy. If the partitions are balanced, Borgs, Chayes and Smith [4] obtain the bound on the agnostic error as in Proposition 3.2. When there is no privacy issues, comparing the upper bound of [4] with that of Corollary 3.3, we see that it has a suboptimal rate, which is the square root of the rate of Corollary 3.3 in the moderately sparse zone. Note also that the setting in [4] is restricted to balanced partitions while we consider more general partitions.

3.3. *Smooth graphons*. We now derive bounds on the mean squared error of smooth graphon estimation. The analysis will be based on the results of Section 2 and on the following bound for the agnostic error associated to smooth graphons.

PROPOSITION 3.5. *Consider the graphon model* (2) *with* $W_0 \in \Sigma(\alpha, L)$ *where* $\alpha, L > 0$ *and* $\rho_n > 0$. *Then*

$$\mathbb{E}\big[\delta^2(\widetilde{f}_{\boldsymbol{\Theta}_0}, f_0)\big] \leq C\frac{\rho_n^2}{n^{\alpha\wedge 1}}, \tag{20}$$

where the constant C *depends only on* L *and* α.

Combining Proposition 3.5 with Lemma 3.1 and with Propositions 2.1 and 2.3, we obtain the following risk bounds for the least squares and restricted least squares graphon estimators.

COROLLARY 3.6. *Consider the graphon model (2) with $W_0 \in \Sigma(\alpha, L)$ where $\alpha, L > 0$ and $0 < \rho_n \le 1$. Fix $k = \lceil (\rho_n^{1/2} n)^{1/(1+\alpha\wedge 1)} \rceil$.*

(i) *Assume that $\rho_n \ge n^{-2(\alpha\wedge 1)/(1+2(\alpha\wedge 1))} (\log n)^{2(1+\alpha\wedge 1)}$, and there exists $n_0 \ge 2$ such that $k = \lfloor n/n_0 \rfloor$. Then there exists a positive constant C_1 depending only on L and α such that the least squares graphon estimator $\widehat{f}$ constructed with this choice of n_0 satisfies*

$$\mathbb{E}\big[\delta^2(\widehat{f}, f_0)\big] \le C_1 \left\{ \rho_n^{(2+\alpha\wedge 1)/(1+\alpha\wedge 1)} n^{-2(\alpha\wedge 1)/(1+\alpha\wedge 1)} + \frac{\rho_n \log n}{n} + \frac{\rho_n^2}{n^{\alpha\wedge 1}} \right\}. \tag{21}$$

(ii) *Assume that $r \ge \rho_n \ge C n^{-2}$. Then there exists a positive constant C_2 depending only on L and α such that the restricted least squares graphon estimator $\widehat{f}_r$ satisfies*

$$\mathbb{E}\big[\delta^2(\widehat{f}_r, f_0)\big] \le C_2 \left\{ r^{(2+\alpha\wedge 1)/(1+\alpha\wedge 1)} n^{-2(\alpha\wedge 1)/(1+\alpha\wedge 1)} + \frac{r \log n}{n} + \frac{\rho_n^2}{n^{\alpha\wedge 1}} \right\}. \tag{22}$$

For the purpose of the discussion, assume that $r \asymp \rho_n$. If $\rho_n \le n^{\alpha-1} \log(n)$, the rate of convergence is of the order $\rho_n \log(n)/n$, the same as that of the probability matrix estimation risk $\mathbb{E}[\|\widehat{\mathbf{\Theta}}^r - \mathbf{\Theta}_0\|_F^2/n^2]$, cf. Corollary 2.7. Observe that the above condition is always satisfied when $\alpha \ge 1$. If $\rho_n \ge n^{\alpha-1} \log(n)$, the rate of convergence in (22) for $\alpha \le 1$ is of the order ρ_n^2/n^α due to the agnostic error. This is slower than the optimal nonparametric rate $\rho_n^{(2+\alpha)/(1+\alpha)} n^{-2\alpha/(1+\alpha)}$ for probability matrix estimation. We conjecture that this loss is unavoidable when considering graphon estimation with the $\delta^2(\cdot,\cdot)$ error measure. We also note that the rates in (22) are faster than those obtained by Wolfe and Olhede [16] for the maximum likelihood estimator. In some cases, the improvement in the rate is up to $n^{-\alpha/2}$.

4. Proofs. In this section, we will repeatedly use Bernstein's inequality that we state here for reader's convenience: Let $X_1, \dots, X_N$ be independent zero-mean random variables. Suppose that $|X_i| \le M$ almost surely, for all i. Then, for any $t > 0$,

$$\mathbb{P}\left\{ \sum_{i=1}^N X_i \ge \sqrt{2t \sum_{i=1}^N \mathbb{E}(X_i^2)} + \frac{2M}{3} t \right\} \le e^{-t}. \tag{23}$$

In the proofs of the upper bounds, to shorten the notation, we assume that instead of the $n(n-1)/2$ observations $\mathbf{A}'$ we have the symmetrized observation matrix $\mathbf{A}$, and thus

$$L(\mathbf{Q}, z) = \frac{1}{2} \sum_{(a,b)\in[k]\times[k]} \sum_{(i,j)\in z^{-1}(a)\times z^{-1}(b), j\ne i} (\mathbf{A}_{ij} - \mathbf{Q}_{ab})^2.$$

This does not change the estimators nor the results. The changes are only in the values of constants that do not explicitly appear in the risk bounds.

4.1. *Proof of Proposition 2.1.* Since $\widehat{\mathbf{\Theta}}$ is a least squares estimator, we have

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F^2 \le \|\mathbf{\Theta}_0 - \mathbf{\Theta}_*\|_F^2 + 2\langle \widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0, \mathbf{E}\rangle + 2\langle \mathbf{\Theta}_0 - \mathbf{\Theta}_*, \mathbf{E}\rangle, \tag{24}$$

where $\mathbf{E} = \mathbf{A} - \mathbf{\Theta}_0$ is the noise matrix. The last summand on the right-hand side of (24) has mean zero. So, to prove the proposition it suffices to bound the expectation of $\langle \widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0, \mathbf{E}\rangle$. For any $z \in \mathcal{Z}_{n,k,n_0}$, denote by $\widetilde{\mathbf{\Theta}}_z$ the best Frobenius norm approximation of $\mathbf{\Theta}_0$ in the collection of matrices

$$\mathcal{T}_z := \big\{\mathbf{\Theta} : \exists \mathbf{Q} \in \mathbb{R}^{k\times k}_{\mathrm{sym}} \text{ such that } \mathbf{\Theta}_{ij} = \mathbf{Q}_{z(i)z(j)},\, i \ne j, \text{ and } \mathbf{\Theta}_{ii} = 0\ \forall i\big\}.$$

We have the following decomposition:

$$\langle \widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0, \mathbf{E}\rangle = \langle \widetilde{\mathbf{\Theta}}_{\hat z} - \mathbf{\Theta}_0, \mathbf{E}\rangle + \langle \widehat{\mathbf{\Theta}} - \widetilde{\mathbf{\Theta}}_{\hat z}, \mathbf{E}\rangle = (I) + (II).$$

In this decomposition, (I) is the error due to mis-clustering and (II) is the error due to the Bernoulli noise. We bound each of these errors separately.

Control of (I). We apply Bernstein's inequality (23) together with the union bound over all $z \in \mathcal{Z}_{n,k,n_0}$ and we use that $\langle \widetilde{\mathbf{\Theta}}_z - \mathbf{\Theta}_0, \mathbf{E}\rangle = 2\sum_{1\le j<i\le n}(\widetilde{\mathbf{\Theta}}_z - \mathbf{\Theta}_0)_{ij}\mathbf{E}_{ij}$. Since the variance of $\mathbf{E}_{ij}$ satisfies $\mathrm{Var}(\mathbf{E}_{ij}) \le \|\mathbf{\Theta}_0\|_\infty$ while $\mathbf{E}_{ij} \in [-1,1]$, and the cardinality of $\mathcal{Z}_{n,k,n_0}$ satisfies $|\mathcal{Z}_{n,k,n_0}| \le k^n$, we obtain

$$\begin{aligned}\mathbb{P}\Big[\langle \widetilde{\mathbf{\Theta}}_{\hat z} - \mathbf{\Theta}_0, \mathbf{E}\rangle &\ge 2\|\widetilde{\mathbf{\Theta}}_{\hat z} - \mathbf{\Theta}_0\|_F \sqrt{\|\mathbf{\Theta}_0\|_\infty\big(n\log(k) + t\big)} \\ &\quad + \frac{4}{3}\|\widetilde{\mathbf{\Theta}}_{\hat z} - \mathbf{\Theta}_0\|_\infty\big(n\log(k) + t\big)\Big] \\ &\le e^{-t},\end{aligned}$$

for all $t > 0$. Since the entries of $\widetilde{\mathbf{\Theta}}_{\hat z}$ are equal to averaged entries of $\mathbf{\Theta}_0$ over blocks, we obtain that $\|\widetilde{\mathbf{\Theta}}_{\hat z} - \mathbf{\Theta}_0\|_\infty \le \|\mathbf{\Theta}_0\|_\infty$. Using this observation, decoupling the term $2\|\widetilde{\mathbf{\Theta}}_{\hat z} - \mathbf{\Theta}_0\|_F\sqrt{\|\mathbf{\Theta}_0\|_\infty(n\log(k)+t)}$ via the elementary inequality $2uv \le u^2 + v^2$ and then integrating with respect to t we obtain

$$\mathbb{E}\Big[\langle \widetilde{\mathbf{\Theta}}_{\hat z} - \mathbf{\Theta}_0, \mathbf{E}\rangle - \frac{1}{8}\|\widetilde{\mathbf{\Theta}}_{\hat z} - \mathbf{\Theta}_0\|_F^2\Big] \le C\|\mathbf{\Theta}_0\|_\infty n\log(k). \tag{25}$$

Control of (II). The control of the error due to the Bernoulli noise is more involved. We first consider the intersection of $\mathcal{T}_z$ with the unit ball in the Frobenius norm and construct a $1/4$-net on this set. Then, using the union bound and Bernstein's inequality, we can write a convenient bound on $\langle \mathbf{\Theta}, \mathbf{E}\rangle$ for any $\mathbf{\Theta}$ from this net. Finally we control (II) using a bound on $\langle \mathbf{\Theta}, \mathbf{E}\rangle$ on the net and a bound on the supremum norm $\|\widetilde{\mathbf{A}}_{\hat z} - \widetilde{\mathbf{\Theta}}_{\hat z}\|_\infty$. We control the supremum norm $\|\widetilde{\mathbf{A}}_{\hat z} - \widetilde{\mathbf{\Theta}}_{\hat z}\|_\infty$ using the definition of $\hat z$ and Bernstein's inequality.

For any $z \in \mathcal{Z}_{n,k,n_0}$, define the set $\mathcal{T}_{z,1} = \{\mathbf{\Theta} \in \mathcal{T}_z : \|\mathbf{\Theta}\|_F \le 1\}$ and denote by $\widetilde{\mathbf{A}}_z$ the best Frobenius norm approximation of $\mathbf{A}$ in $\mathcal{T}_z$. Then $\widetilde{\mathbf{E}}_z = \widetilde{\mathbf{A}}_z - \widetilde{\mathbf{\Theta}}_z$ is the projection of $\mathbf{E}$ onto $\mathcal{T}_z$. Notice that $\frac{\widetilde{\mathbf{A}}_z - \widetilde{\mathbf{\Theta}}_z}{\|\widetilde{\mathbf{A}}_z - \widetilde{\mathbf{\Theta}}_z\|_F} = \frac{\widetilde{\mathbf{E}}_z}{\|\widetilde{\mathbf{E}}_z\|_F}$ maximizes $\langle \mathbf{\Theta}', \mathbf{E}\rangle$ over all

$\mathbf{\Theta}' \in \mathcal{T}_{z,1}$. Denote by $\mathcal{C}_z$ the minimal 1/4-net on $\mathcal{T}_{z,1}$ in the Frobenius norm. To each $\mathbf{V} \in \mathcal{C}_z$, we associate

$$\widetilde{\mathbf{V}} \in \arg \min_{\mathbf{\Theta}' \in \mathcal{T}_{z,1} \cap \mathcal{B}_{\|\cdot\|_F}(\mathbf{V}, 1/4)} \|\mathbf{\Theta}'\|_\infty,$$

which is a matrix minimizing the entry-wise supremum norm over the Frobenius ball $\mathcal{B}_{\|\cdot\|_F}(\mathbf{V}, 1/4)$ of radius 1/4 centered at $\mathbf{V}$. Finally, define $\widetilde{\mathcal{C}}_z := \{\widetilde{\mathbf{V}} : \mathbf{V} \in \mathcal{C}_z\}$.

A standard bound for covering numbers implies that $\log|\widetilde{\mathcal{C}}_z| \le Ck^2$, where $|S|$ denotes the cardinality of set S. By Bernstein's inequality combined with the union bound, we find that with probability greater than $1 - e^{-t}$,

$$\langle \mathbf{\Theta}, \mathbf{E} \rangle \le 2\sqrt{\|\mathbf{\Theta}_0\|_\infty \big(n \log(k) + k^2 + t\big)} + \frac{4}{3}\|\mathbf{\Theta}\|_\infty \big(n \log(k) + k^2 + t\big)$$

simultaneously for all $\mathbf{\Theta} \in \widetilde{\mathcal{C}}_z$ with any $z \in \mathcal{Z}_{n,k,n_0}$. Here, we have used that $\|\mathbf{\Theta}\|_F \le 1$ for all $\mathbf{\Theta} \in \widetilde{\mathcal{C}}_z$.

Assume w.l.o.g. that $\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}} \ne 0$. By the definition of $\widetilde{\mathcal{C}}_z$, there exists $\mathbf{\Theta} \in \widetilde{\mathcal{C}}_{\hat{z}}$, such that

$$\left\| \mathbf{\Theta} - \frac{\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}}{\|\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}\|_F} \right\|_F \le \frac{1}{2} \quad \text{and} \quad \|\mathbf{\Theta}\|_\infty \le \frac{\|\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}\|_\infty}{\|\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}\|_F}.$$

Note that for this $\mathbf{\Theta}$, the matrix $2(\mathbf{\Theta} - \frac{\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}}{\|\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}\|_F})$ belongs to $\mathcal{T}_{z,1}$. Thus,

$$\left\langle \frac{\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}}{\|\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}\|_F}, \mathbf{E} \right\rangle \le \langle \mathbf{\Theta}, \mathbf{E} \rangle + \frac{1}{2} \max_{\mathbf{\Theta}' \in \mathcal{T}_{\hat{z},1}} \langle \mathbf{\Theta}', \mathbf{E} \rangle = \langle \mathbf{\Theta}, \mathbf{E} \rangle + \frac{1}{2} \left\langle \frac{\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}}{\|\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}\|_F}, \mathbf{E} \right\rangle$$

since $\frac{\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}}{\|\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}\|_F}$ maximizes $\langle \mathbf{\Theta}', \mathbf{E} \rangle$ over all $\mathbf{\Theta}' \in \mathcal{T}_{\hat{z},1}$. Using the last two displays, we obtain that

$$\begin{aligned} \langle \widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}, \mathbf{E} \rangle &\le 4\|\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}\|_F \sqrt{\|\mathbf{\Theta}_0\|_\infty \big(n \log(k) + k^2 + t\big)} \\ &\quad + \frac{8}{3}\|\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}\|_\infty \big(n \log(k) + k^2 + t\big) \end{aligned} \tag{26}$$

with probability greater than $1 - e^{-t}$.

Next, we are looking for a bound on $\|\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}\|_\infty$. Since $\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}} = \widetilde{\mathbf{E}}_{\hat{z}}$ we have

$$[\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}]_{ij} = \frac{\sum_{l' \ne l : \hat{z}(l') = \hat{z}(i), \hat{z}(l) = \hat{z}(j)} \mathbf{E}_{l'l}}{|(l', l) : l' \ne l, \hat{z}(l') = \hat{z}(i), \hat{z}(l) = \hat{z}(j)|}$$

for all $i \ne j$. Consequently, we have $\|\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\mathbf{\Theta}}_{\hat{z}}\|_\infty \le \sup_{m=n_0,\dots,n} \sup_{s=n_0,\dots,n} X_{ms}$ where

$$X_{ms} := \sup_{\mathcal{V}_1 : |\mathcal{V}_1| = m} \sup_{\mathcal{V}_2 : |\mathcal{V}_2| = s} \frac{|\sum_{(i,j) \in \mathcal{V}_1 \times \mathcal{V}_2 : i \ne j} \mathbf{E}_{ij}|}{ms - |\mathcal{V}_1 \cap \mathcal{V}_2|}.$$

Since $n_0 \geq 2$, we have $ms - |\mathcal{V}_1 \cap \mathcal{V}_2| \geq ms - m \wedge s \geq ms/2$ for all $m, s \geq n_0$. Furthermore, $|\{\mathcal{V}_1 : |\mathcal{V}_1| = m\}| \leq \binom{n}{m} \leq (en/m)^m$. Therefore, Bernstein's inequality combined with the union bound over $\mathcal{V}_1, \mathcal{V}_2$ leads to

$$\begin{aligned}\mathbb{P}\Bigg[X_{ms} \leq C\Bigg(\sqrt{\|\boldsymbol{\Theta}_0\|_\infty \frac{m\log(en/m) + s\log(en/s) + t}{ms}} \\ + \frac{m\log(en/m) + s\log(en/s) + t}{ms}\Bigg)\Bigg] \\ \geq 1 - 2e^{-t}\end{aligned}$$

for any $t > 0$. From a union bound over all integers $m, s \in [n_0, n]$, we conclude that

$$\mathbb{P}\Bigg[\|\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\boldsymbol{\Theta}}_{\hat{z}}\|_\infty \leq C\Bigg(\|\boldsymbol{\Theta}_0\|_\infty + \frac{\log(n/n_0) + t}{n_0}\Bigg)\Bigg] \geq 1 - 2e^{-t}. \tag{27}$$

Decoupling (26) by use of the inequality $2uv \leq u^2 + v^2$ and combining the result with (27), we obtain

$$\begin{aligned}\langle \widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\boldsymbol{\Theta}}_{\hat{z}}, \mathbf{E}\rangle \leq \frac{\|\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\boldsymbol{\Theta}}_{\hat{z}}\|_F^2}{16} + C\Bigg(\|\boldsymbol{\Theta}_0\|_\infty\big(n\log(k) + k^2 + t\big) \\ + \frac{\log(n/n_0)(n\log(k) + k^2) + t(n\log(k) + k^2 + t)}{n_0}\Bigg)\end{aligned}$$

with probability greater than $1 - 3e^{-t}$. Integrating with respect to t leads to

$$\begin{aligned}\mathbb{E}\Bigg[\langle \widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\boldsymbol{\Theta}}_{\hat{z}}, \mathbf{E}\rangle - \frac{\|\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\boldsymbol{\Theta}}_{\hat{z}}\|_F^2}{16}\Bigg] \\ \leq C\Bigg(\|\boldsymbol{\Theta}_0\|_\infty\big(n\log(k) + k^2\big) + \frac{\log(n/n_0)(n\log(k) + k^2)}{n_0}\Bigg).\end{aligned} \tag{28}$$

Now, note that $\widehat{\boldsymbol{\Theta}} = \widetilde{\mathbf{A}}_{\hat{z}}$, and hence $\|\widetilde{\boldsymbol{\Theta}}_{\hat{z}} - \boldsymbol{\Theta}_0\|_F \leq \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F$ by definition of $\widetilde{\boldsymbol{\Theta}}_{\hat{z}}$. Thus, $\|\widetilde{\mathbf{A}}_{\hat{z}} - \widetilde{\boldsymbol{\Theta}}_{\hat{z}}\|_F = \|\widehat{\boldsymbol{\Theta}} - \widetilde{\boldsymbol{\Theta}}_{\hat{z}}\|_F \leq 2\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F$. These remarks, together with (25) and (28), imply

$$\begin{aligned}\mathbb{E}\big[\langle \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0, \mathbf{E}\rangle\big] \leq \frac{3}{8}\mathbb{E}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F^2 + C\Bigg(\|\boldsymbol{\Theta}_0\|_\infty\big(n\log(k) + k^2\big) \\ + \frac{\log(n/n_0)(n\log(k) + k^2)}{n_0}\Bigg).\end{aligned}$$

The result of the proposition now follows from the last inequality and (24).

4.2. *Proof of Proposition 2.3.* We first follow the lines of the proof of Proposition 2.1. As there, we have

$$\|\widehat{\mathbf{\Theta}}^r - \mathbf{\Theta}_0\|_F^2 \le \|\mathbf{\Theta}_0 - \mathbf{\Theta}_*\|_F^2 + 2\langle \widetilde{\mathbf{\Theta}}_{\hat{z}_r} - \mathbf{\Theta}_0, \mathbf{E}\rangle + 2\langle \widehat{\mathbf{\Theta}}^r - \widetilde{\mathbf{\Theta}}_{\hat{z}_r}, \mathbf{E}\rangle + 2\langle \mathbf{\Theta}_0 - \mathbf{\Theta}_*, \mathbf{E}\rangle.$$

Here, $\mathbb{E}\langle \mathbf{\Theta}_0 - \mathbf{\Theta}_*, \mathbf{E}\rangle = 0$, and analogously to (25),

$$\mathbb{E}\Big[\langle \widetilde{\mathbf{\Theta}}_{\hat{z}_r} - \mathbf{\Theta}_0, \mathbf{E}\rangle - \frac{1}{8}\|\widetilde{\mathbf{\Theta}}_{\hat{z}_r} - \mathbf{\Theta}_0\|_F^2\Big] \le C\|\mathbf{\Theta}_0\|_\infty n\log(k). \tag{29}$$

It remains to obtain a bound on the expectation of $\langle \widehat{\mathbf{\Theta}}^r - \widetilde{\mathbf{\Theta}}_{\hat{z}_r}, \mathbf{E}\rangle$, which corresponds to the term (II) in the proof of Proposition 2.1 (the error due to the Bernoulli noise). To do this, we consider the subset $\mathcal{A}$ of matrices in $\mathcal{T}_{\hat{z}_r}$ such that their supremum norm is bounded by $2r$ and Frobenius norm is bounded by $\|\widehat{\mathbf{\Theta}}^r - \widetilde{\mathbf{\Theta}}_{\hat{z}_r}\|_F$. As $\widehat{\mathbf{\Theta}}^r - \widetilde{\mathbf{\Theta}}_{\hat{z}_r}$ belongs to this set it is enough to obtain a bound on the supremum of $\langle \mathbf{\Theta}, \mathbf{E}\rangle$ for $\mathbf{\Theta} \in \mathcal{A}$. To control this supremum, we construct a finite subset $\mathcal{C}_z^*$ that approximates well the set $\mathcal{A}$ both in the Frobenius norm and in the supremum norm.

Consider the set

$$\mathcal{A} = \big\{\mathbf{\Theta} \in \mathcal{T}_{\hat{z}_r} : \|\mathbf{\Theta}\|_\infty \le 2r, \|\mathbf{\Theta}\|_F \le \|\widehat{\mathbf{\Theta}}^r - \widetilde{\mathbf{\Theta}}_{\hat{z}_r}\|_F\big\}.$$

Then

$$\langle \widehat{\mathbf{\Theta}}^r - \widetilde{\mathbf{\Theta}}_{\hat{z}_r}, \mathbf{E}\rangle \le \max_{\mathbf{\Theta}\in\mathcal{A}} \langle \mathbf{\Theta}, \mathbf{E}\rangle := \langle \hat{\mathbf{T}}, \mathbf{E}\rangle_F, \tag{30}$$

where $\hat{\mathbf{T}}$ is a matrix in $\mathcal{A}$ that achieves the maximum. If $\|\hat{\mathbf{T}}\|_F < 2r$, we have a trivial bound $\langle \hat{\mathbf{T}}, \mathbf{E}\rangle_F \le 2rn$ since all components of $\mathbf{E}$ belong to $[-1, 1]$. Thus, it suffices to consider the case $\|\hat{\mathbf{T}}\|_F \ge 2r$. In order to bound $\langle \hat{\mathbf{T}}, \mathbf{E}\rangle_F$ in this case, we construct a finite subset of $\mathcal{T}_{\hat{z}_r}$ that approximates well $\hat{\mathbf{T}}$ both in the Frobenius norm and in the supremum norm.

For each $z \in \mathcal{Z}_{n,k}$, let $\mathcal{C}_z$ be a minimal $1/4$-net of $\mathcal{T}_{z,1}$ in the Frobenius norm. Set $\varepsilon_0 = r$ and $\varepsilon_q = 2^q\varepsilon_0$ for any integer $q = 1, \dots, q_{\max}$, where $q_{\max}$ is the smallest integer such that $2rn \le \varepsilon_{q_{\max}}$. Clearly, $q_{\max} \le C\log(n)$. For any $\mathbf{V} \in \mathcal{C}_z$, any $q = 0, \dots, q_{\max}$, and any matrix $\mathbf{U} \in \{-1, 0, 1\}^{k\times k}$, define a matrix $\mathbf{V}^{q,\mathbf{U},z} \in \mathbb{R}^{n\times n}$ with elements $\mathbf{V}_{ij}^{q,\mathbf{U},z}$ such that $\mathbf{V}_{ii}^{q,\mathbf{U},z} = 0$ for all $i \in [n]$, and for all $i \ne j$,

$$\mathbf{V}_{ij}^{q,\mathbf{U},z} = \operatorname{sign}(\mathbf{V}_{ij})\big(|\varepsilon_q\mathbf{V}_{ij}| \wedge (2r)\big)\big(1 - |\mathbf{U}_{z(i)z(j)}|\big) + r\mathbf{U}_{z(i)z(j)}. \tag{31}$$

Finally, denote by $\mathcal{C}_z^*$ the set of all such matrices:

$$\mathcal{C}_z^* := \big\{\mathbf{V}^{q,\mathbf{U},z} : \mathbf{V} \in \mathcal{C}_z, q = 0, \dots, q_{\max}, \mathbf{U} \in \{-1, 0, 1\}^{k\times k}\big\}.$$

For any $z \in \mathcal{Z}_{n,k}$ we have $\log|\mathcal{C}_z^*| \le C(k^2 + \log\log(n))$ while $\log|\mathcal{Z}_{n,k}| \le n\log(k)$. Also, the variance of $\mathbf{E}_{ij}$ satisfies $\operatorname{Var}(\mathbf{E}_{ij}) \le \|\mathbf{\Theta}_0\|_\infty \le r$, and $\mathbf{E}_{ij} \in [-1, 1]$. Thus,

from Bernstein's inequality combined with the union bound we obtain that, with probability greater than $1 - e^{-t}$,

$$\langle \mathbf{V}, \mathbf{E} \rangle \le 2\|\mathbf{V}\|_F \sqrt{r\big(k^2 + n\log(k) + t\big)} + \frac{4}{3} r\big(k^2 + n\log(k) + t\big)$$

simultaneously for all matrices $\mathbf{V} \in \mathcal{C}_z^*$ and all $z \in \mathcal{Z}_{n,k}$. Here, we have used that $\|\mathbf{V}\|_\infty \le 2r$ for all $\mathbf{V} \in \mathcal{C}_z^*$, cf. (31). It follows that, with probability greater than $1 - e^{-t}$,

$$\langle \mathbf{V}, \mathbf{E} \rangle - \frac{1}{50}\|\mathbf{V}\|_F^2 \le Cr\big(k^2 + n\log(k) + t\big) \tag{32}$$

simultaneously for all matrices $\mathbf{V} \in \mathcal{C}_z^*$ and all $z \in \mathcal{Z}_{n,k}$.

We now use the following lemma proved in Section 4.10.

LEMMA 4.1. *If $\|\hat{\mathbf{T}}\|_F \ge 2r$ there exists a matrix $\hat{\mathbf{V}}$ in $\mathcal{C}_{\hat{z}_r}^*$, such that*:

- $\|\hat{\mathbf{T}} - \hat{\mathbf{V}}\|_F \le \|\hat{\mathbf{T}}\|_F/4$,
- $\|\hat{\mathbf{T}} - \hat{\mathbf{V}}\|_\infty \le r$.

Thus, on the event $\|\hat{\mathbf{T}}\|_F \ge r$, as a consequence of Lemma 4.1 we obtain that $2(\hat{\mathbf{T}} - \hat{\mathbf{V}}) \in \mathcal{A}$. This and the definition of $\hat{\mathbf{T}}$ imply $\langle \hat{\mathbf{T}} - \hat{\mathbf{V}}, \mathbf{E} \rangle \le \langle \hat{\mathbf{T}}, \mathbf{E} \rangle_F/2$, so that $\langle \hat{\mathbf{T}}, \mathbf{E} \rangle_F \le 2\langle \hat{\mathbf{V}}, \mathbf{E} \rangle$, and thus

$$\langle \widehat{\mathbf{\Theta}}^r - \widetilde{\mathbf{\Theta}}_{\hat{z}_r}, \mathbf{E} \rangle/2 \le \langle \hat{\mathbf{V}}, \mathbf{E} \rangle.$$

Furthermore, by Lemma 4.1

$$\|\hat{\mathbf{V}}\|_F \le 5\|\hat{\mathbf{T}}\|_F/4 \le 5\big\|\widehat{\mathbf{\Theta}}^r - \widetilde{\mathbf{\Theta}}_{\hat{z}_r}\big\|/4.$$

These remarks (recall that they hold on the event $\|\hat{\mathbf{T}}\|_F \ge r$), and (32) yield

$$\mathbb{P}\bigg[\bigg(\langle \widehat{\mathbf{\Theta}}^r - \widetilde{\mathbf{\Theta}}_{\hat{z}_r}, \mathbf{E} \rangle - \frac{\|\widehat{\mathbf{\Theta}}^r - \widetilde{\mathbf{\Theta}}_{\hat{z}_r}\|_F^2}{16}\bigg) \mathbb{1}_{\{\|\hat{\mathbf{T}}\|_F \ge r\}} \ge Cr\big(k^2 + n\log(k) + t\big)\bigg] \le e^{-t}$$

for all $t > 0$. Integrating with respect to t and using that $\langle \widehat{\mathbf{\Theta}}^r - \widetilde{\mathbf{\Theta}}_{\hat{z}_r}, \mathbf{E} \rangle \le rn$ for $\|\hat{\mathbf{T}}\|_F < r$, we obtain

$$\mathbb{E}_{\mathbf{\Theta}_0}\bigg[\langle \widehat{\mathbf{\Theta}}^r - \widetilde{\mathbf{\Theta}}_{\hat{z}_r}, \mathbf{E} \rangle - \frac{\|\widehat{\mathbf{\Theta}}^r - \widetilde{\mathbf{\Theta}}_{\hat{z}_r}\|_F^2}{16}\bigg] \le Cr\big(n\log(k) + k^2\big).$$

In view of this inequality and (29), the proof is now completed by the same argument as in Proposition 2.1.

4.3. *Proof of Proposition 2.4.* The proof follows the lines of Theorem 2.2 in [10]. So, for brevity, we only outline the differences from that proof. The remaining details can be found in [10]. First, note that to prove the theorem it suffices to obtain separately the lower bounds of the order $\rho_n(k/n)^2 \wedge \rho_n^2$ and of the order $\rho_n \log(k)/n \wedge \rho_n^2$. Next, note that the Kullback–Leibler divergence $\mathcal{K}(\rho_n p, \rho_n q)$ between two Bernoulli distributions with parameters $\rho_n p$ and $\rho_n q$ such that $1/4 < p, q < 3/4$ satisfies

$$\begin{aligned}\mathcal{K}(\rho_n p, \rho_n q) &= \rho_n q \log\left(\frac{q}{p}\right) + (1-\rho_n q)\log\left(\frac{1-\rho_n q}{1-\rho_n p}\right)\\ &\le \rho_n \frac{q(q-p)}{p} + \rho_n \frac{1-\rho_n q}{1-\rho_n p}(p-q)\\ &= \rho_n \frac{(q-p)^2}{p(1-\rho_n p)} \le \frac{16}{3}\rho_n (q-p)^2.\end{aligned} \tag{33}$$

The main difference from the proof in [10] is that now the matrices $\mathbf{\Theta}_0$ defining the probability distributions in the Fano lemma depend on ρ_n. Namely, to prove the $\rho_n(k/n)^2 \wedge \rho_n^2$ bound we consider matrices of connection probabilities with elements $\frac{\rho_n}{2} + c_1\sqrt{\rho_n}(\frac{k}{n} \wedge \sqrt{\rho_n})\omega_{ab}$ with suitably chosen $\omega_{ab} \in \{0, 1\}$ and $c_1 > 0$ small enough (for $\rho_n = 1$, this coincides with the definition of elements $\mathbf{Q}^{\omega}_{ab}$ in [10]). Then the squared Frobenius distance between matrices $\mathbf{\Theta}_0$ in the corresponding set is of the order $n^2(\rho_n(k/n)^2 \wedge \rho_n^2)$, which leads to the desired rate, whereas in view of (33) the Kullback–Leibler divergences between the probability measures in this set are bounded by $Cn^2\rho_n(\frac{k^2}{n^2\rho_n} \wedge 1) \le Ck^2$. Thus, the lower bound of the order $\rho_n(k/n)^2 \wedge \rho_n^2$ follows.

The argument used to prove the lower bound of the order $\rho_n \log(k)/n \wedge \rho_n^2$ is quite analogous. To this end, we modify the corresponding construction of [10] only in that we take the connection probabilities of the form $\frac{\rho_n}{2} + c_2\sqrt{\rho_n}(\sqrt{\frac{\log(k)}{n}} \wedge \sqrt{\rho_n})\omega_a$ with suitably chosen $\omega_a \in \{0, 1\}$ and $c_2 > 0$ small enough (for $\rho_n = 1$ this coincides with the definition of probabilities B_a in the proof of [10]).

4.4. *Proof of Proposition 2.5.* We prove that there exists a random matrix $\overline{\mathbf{\Theta}}$ measurable with respect to $\xi_1, \dots, \xi_n$ and with values in $\mathcal{T}_{n_0}[k]$ satisfying

$$\mathbb{E}\left(\frac{1}{n^2}\|\mathbf{\Theta}_0 - \overline{\mathbf{\Theta}}\|_F^2\right) \le CM^2\rho_n^2\left(\frac{1}{k^2}\right)^{\alpha\wedge 1}. \tag{34}$$

Obviously, this implies (9). To obtain such $\overline{\mathbf{\Theta}}$, we construct a balanced partition z^* where the first $k-1$ classes contain exactly n_0 elements and the last class contains $n - n_0(k-1)$ elements. Then we construct $\overline{\mathbf{\Theta}}$ using block averages on the blocks given by z^*.

Let $n = n_0 k + r$ where r is a remainder term between 0 and $n_0 - 1$. Define $z^* : [n] \to [k]$ by

$$(z^*)^{-1}(a) = \{i \in [n] : \xi_i = \xi_{(j)} \text{ for some } j \in [(a-1)n_0 + 1, an_0]\}$$

for each $a \in \{1, \dots, k-1\}$ and

$$(z^*)^{-1}(k) = \{i \in [n] : \xi_i = \xi_{(j)} \text{ for some } j \in [(k-1)n_0 + 1, n]\},$$

where $\xi_{(j)}$ denotes the jth order statistic. Note that with this partition the first $k-1$ classes contain n_0 elements and the last class contains $n_0 + r$ elements. We define

$$n^*_{ab} = \begin{cases} n_0^2, & \text{if } a \neq b, a \neq k, b \neq k, \\ (n_0 + r)n_0, & \text{if } a = k \text{ or } b = k \text{ and } a \neq b, \\ (n_0 - 1)n_0, & \text{if } a = b \text{ and } a \neq k, \\ (n_0 + r)(n_0 + r - 1), & \text{if } a = b = k. \end{cases}$$

Using the partition z^*, we define the block average

$$\mathbf{Q}^*_{ab} = \frac{1}{n^*_{ab}} \sum_{i \in (z^*)^{-1}(a), j \in (z^*)^{-1}(b), i \neq j} \rho_n W(\xi_i, \xi_j).$$

Finally, the approximation $\overline{\mathbf{\Theta}}$ of $\mathbf{\Theta}_0$ is defined as a symmetric matrix with entries $\overline{\mathbf{\Theta}}_{ij} = \mathbf{Q}^*_{z^*(i)z^*(j)}$ for all $i \neq j$ and $\overline{\mathbf{\Theta}}_{ii} = 0$ for all i. We have

$$\begin{aligned} \mathbb{E}\Big(\frac{1}{n^2}\|\mathbf{\Theta}_0 - \overline{\mathbf{\Theta}}\|_F^2\Big) &= \frac{1}{n^2} \sum_{a \in [k], b \in [k]} \mathbb{E} \sum_{i \in (z^*)^{-1}(a), j \in (z^*)^{-1}(b), i \neq j} (\mathbf{\Theta}_{ij} - \mathbf{Q}^*_{ab})^2 \\ &= \frac{1}{n^2} \sum_{a \in [k], b \in [k]} \mathbb{E} \sum_{i \in (z^*)^{-1}(a), j \in (z^*)^{-1}(b), i \neq j} \Bigg(\rho_n W(\xi_i, \xi_j) \\ &\quad - \frac{\sum_{u \in (z^*)^{-1}(a), v \in (z^*)^{-1}(b), u \neq v} \rho_n W(\xi_u, \xi_v)}{n^*_{ab}}\Bigg)^2. \end{aligned}$$

Define $J_a = [(a-1)n_0 + 1, an_0]$ if $a < k$ and $J_k = [(k-1)n_0 + 1, n]$. By definition of z^*, we have

$$\begin{aligned} \mathbb{E}\Big(\frac{1}{n^2}\|\mathbf{\Theta}_0 - \overline{\mathbf{\Theta}}\|_F^2\Big) &= \frac{\rho_n^2}{n^2} \sum_{a \in [k], b \in [k]} \mathbb{E} \sum_{i \in J_a, j \in J_b, i \neq j} \Bigg(W(\xi_{(i)}, \xi_{(j)}) \\ &\quad - \frac{\sum_{u \in J_a, v \in J_b, u \neq v} W(\xi_{(u)}, \xi_{(v)})}{n^*_{ab}}\Bigg)^2 \\ &\leq \frac{\rho_n^2}{n^2} \sum_{a \in [k], b \in [k]} \sum_{i \in J_a, j \in J_b, i \neq j} \Bigg(\frac{1}{n^*_{ab}} \sum_{u \in J_a, v \in J_b, u \neq v} \mathbb{E}\big(W(\xi_{(i)}, \xi_{(j)}) \\ &\quad - W(\xi_{(u)}, \xi_{(v)})\big)^2\Bigg). \end{aligned} \tag{35}$$

Using (8) and Jensen's inequality, we obtain

$$\mathbb{E}\big(W(\xi_{(i)},\xi_{(j)})-W(\xi_{(u)},\xi_{(v)})\big)^2 \le M^2\mathbb{E}\big[\big(|\xi_{(i)}-\xi_{(u)}|^{\alpha'}+|\xi_{(j)}-\xi_{(v)}|^{\alpha'}\big)^2\big]$$
$$\le 2M^2\big(\big(\mathbb{E}|\xi_{(i)}-\xi_{(u)}|^2\big)^{\alpha'}+\big(\mathbb{E}|\xi_{(j)}-\xi_{(v)}|^2\big)^{\alpha'}\big),$$

where we set for brevity $\alpha' = \alpha \wedge 1$. Note that by definition of z^* we have $|i-u| < 2n_0$ and $|j-v| < 2n_0$. Therefore, application of Lemma 4.10 (see Section 4.10) leads to

$$\mathbb{E}\big(W(\xi_{(i)},\xi_{(j)})-W(\xi_{(u)},\xi_{(v)})\big)^2 \le C(n_0/n)^{2\alpha'} \le C(1/k)^{2\alpha'},$$

where we have used that $k = \lfloor n/n_0 \rfloor$. Plugging this bound into (35) proves the proposition.

4.5. *Proof of Proposition 3.2.* For any $W_0 \in \mathcal{W}[k]$, we first construct an ordered graphon W' isomorphic to W_0 and we set $f' = \rho_n W'$. Then we construct an ordered empirical graphon $\widehat{f}'$ isomorphic to $\widetilde{f}_{\mathbf{\Theta}_0}$ and we estimate the $\delta(\cdot,\cdot)$-distance between these two ordered versions f' and $\widehat{f}'$.

Consider the matrix $\mathbf{\Theta}_0'$ with entries $(\mathbf{\Theta}_0')_{ij} = \rho_n W(\xi_i,\xi_j)$ for all i, j. As opposed to $\mathbf{\Theta}_0$, the diagonal entries of $\mathbf{\Theta}_0'$ are not constrained to be null. By the triangle inequality, we get

$$\mathbb{E}\big[\delta^2(\widetilde{f}_{\mathbf{\Theta}_0}, f_0)\big] \le 2\mathbb{E}\big[\delta^2(\widetilde{f}_{\mathbf{\Theta}_0}, \widetilde{f}_{\mathbf{\Theta}_0'})\big] + 2\mathbb{E}\big[\delta^2(\widetilde{f}_{\mathbf{\Theta}_0'}, f_0)\big]. \tag{36}$$

Since the entries of $\mathbf{\Theta}_0$ coincide with those of $\mathbf{\Theta}_0'$ outside the diagonal, the difference $\widetilde{f}_{\mathbf{\Theta}_0} - \widetilde{f}_{\mathbf{\Theta}_0'}$ is null outside of a set of measure $1/n$. Also, the entries of $\mathbf{\Theta}_0'$ are smaller than ρ_n. It follows that $\mathbb{E}[\delta^2(\widetilde{f}_{\mathbf{\Theta}_0}, \widetilde{f}_{\mathbf{\Theta}_0'})] \le \rho_n^2/n$. Hence, it suffices to prove that

$$\mathbb{E}\big[\delta^2(\widetilde{f}_{\mathbf{\Theta}_0'}, f_0)\big] \le C\rho_n^2\sqrt{k/n}.$$

We prove this inequality by induction on k. The result is trivial for $k = 1$ as $\delta^2(\widehat{f}_{\mathbf{\Theta}_0'}, f_0) = 0$. Fix some $k > 1$ and assume that the result is valid for $\mathcal{W}[k-1]$. Consider any $W_0 \in \mathcal{W}[k]$ and let $\mathbf{Q} \in \mathbb{R}^{k\times k}_{\mathrm{sym}}$ and $\phi : [0,1] \to [k]$ be associated to W_0 as in definition (16). We assume w.l.o.g. that all the rows of $\mathbf{Q}$ are distinct and that $\lambda_a := \lambda(\phi^{-1}(a))$ is positive for all $a \in [k]$, since otherwise W_0 belongs to $\mathcal{W}[k-1]$. For any $b \in [k]$, define the cumulative distribution function

$$F_\phi(b) = \sum_{a=1}^{b} \lambda_a$$

and set $F_\phi(0) = 0$. For any $(a,b) \in [k]\times[k]$ define $\Pi_{ab}(\phi) = [F_\phi(a-1), F_\phi(a)) \times [F_\phi(b-1), F_\phi(b))$ where $\mathbb{1}_A(\cdot)$ denotes the indicator function of set A. Finally, we consider the ordered graphon

$$W'(x,y) = \sum_{a=1}^{k}\sum_{b=1}^{k} \mathbf{Q}_{ab}\mathbb{1}_{\Pi_{ab}(\phi)}(x,y).$$

Obviously, $f' = \rho_n W'$ is weakly isomorphic to $f_0 = \rho_n W_0$. Let

$$\widehat{\lambda}_a = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\xi_i \in \phi^{-1}(a)\}}$$

be the (unobserved) empirical frequency of group a. Here, $\xi_1, \dots, \xi_n$ are the i.i.d. uniformly distributed random variables in the graphon model (2).

Note that the relations $\sum_{a=1}^{k} \lambda_a = \sum_{a=1}^{k} \widehat{\lambda}_a = 1$ imply

$$\sum_{a:\lambda_a > \widehat{\lambda}_a} (\lambda_a - \widehat{\lambda}_a) = \sum_{a:\widehat{\lambda}_a > \lambda_a} (\widehat{\lambda}_a - \lambda_a). \tag{37}$$

Consider a function $\psi : [0, 1] \to [k]$ such that:

(i) $\psi(x) = a$ for all $a \in [k]$ and $x \in [F_\phi(a-1), F_\phi(a-1) + \widehat{\lambda}_a \wedge \lambda_a)$,
(ii) $\lambda(\psi^{-1}(a)) = \widehat{\lambda}_a$ for all $a \in [k]$.

Such a function ψ exists. Indeed, for each a such that $\lambda_a > \widehat{\lambda}_a$, conditions (i) and (ii) are trivially satisfied if we take $\psi^{-1}(a) = [F_\phi(a-1), F_\phi(a-1) + \widehat{\lambda}_a)$, and there is an interval of Lebesgue measure $\lambda_a - \widehat{\lambda}_a$ left nonassigned. Summing over all such a, we see that there is a union of intervals with Lebesgue measure $m_+ := \sum_{a:\lambda_a > \widehat{\lambda}_a} (\lambda_a - \widehat{\lambda}_a)$ left nonassigned. On the other hand, for a such that $\lambda_a < \widehat{\lambda}_a$, we must have $\psi(x) = a$ for $x \in [F_\phi(a-1), F_\phi(a-1) + \lambda_a)$ to satisfy (i), while to meet condition (ii) we need additionally to assign $\psi(x) = a$ for x on a set of Lebesgue measure $\widehat{\lambda}_a - \lambda_a$. Summing over all such a, we need additionally to find a set of Lebesgue measure $m_- := \sum_{a:\widehat{\lambda}_a > \lambda_a} (\lambda_a - \widehat{\lambda}_a)$ to make such assignments. But this set is readily available as a union of nonassigned intervals for all a such that $\lambda_a > \widehat{\lambda}_a$ since $m_+ = m_-$ by virtue of (37).

Finally, define the graphon $\widehat{f}'(x, y) = \mathbf{Q}_{\psi(x),\psi(y)}$. Notice that in view of (ii) $\widehat{f}'$ is weakly isomorphic to the empirical graphon $\widetilde{f}_{\mathbf{\Theta}_0}$. Since $\delta(\cdot, \cdot)$ is a metric on the quotient space $\widetilde{\mathcal{W}}$,

$$\delta^2(\widetilde{f}_{\Theta_0}, f_0) = \delta^2(\widehat{f}', f') \le \int_{[0,1]^2} |\widehat{f}'(x, y) - f'(x, y)|^2 \, dx \, dy$$

$$\le \rho_n^2 \int_{[0,1]^2} \mathbb{1}_{\{f'(x,y) \neq \widehat{f}'(x,y)\}} \, dx \, dy.$$

The two functions $f_0(x, y)$ and $\widehat{f}(x, y)$ are equal except possibly the case when either x or y belongs to one of the intervals $[F_\phi(a-1) + \widehat{\lambda}_a \wedge \lambda_a, F_\phi(a-1) + \lambda_a)$ for $a \in [k]$. Hence, the Lebesgue measure of the set $\{(x, y) : f'(x, y) \neq \widehat{f}'(x, y)\}$ is not greater than $2m_+ = m_+ + m_- = \sum_{a=1}^{k} |\lambda_a - \widehat{\lambda}_a|$. Thus,

$$\delta^2(\widehat{f}_{\mathbf{\Theta}'_0}, f_0) \le \rho_n^2 \sum_{a=1}^{k} |\lambda_a - \widehat{\lambda}_a|.$$

Since $\xi_1, \dots, \xi_n$ are i.i.d. uniformly distributed random variables, $n\widehat{\lambda}_a$ has a binomial distribution with parameters (n, λ_a). By the Cauchy–Schwarz inequality, we

get $\mathbb{E}[|\lambda_a - \widehat{\lambda}_a|] \le \sqrt{\frac{\lambda_a(1-\lambda_a)}{n}}$. Applying again the Cauchy–Schwarz inequality, we conclude that

$$\mathbb{E}\big[\delta^2(\widehat{f}_{\mathbf{\Theta}_0'}, f_0)\big] \le \frac{\rho_n^2}{\sqrt{n}} \sum_{a=1}^{k} \sqrt{\lambda_a} \le \rho_n^2 \sqrt{\frac{k}{n}}.$$

4.6. *Proof of Proposition* 3.5. Arguing as in the proof of Proposition 3.2, we have

$$\mathbb{E}\big[\delta^2(\widetilde{f}_{\mathbf{\Theta}_0}, f_0)\big] \le \frac{2\rho_n^2}{n} + 2\mathbb{E}\big[\delta^2(\widetilde{f}_{\mathbf{\Theta}_0'}, f_0)\big],$$

where we recall that $\mathbf{\Theta}_0'$ is defined by $(\mathbf{\Theta}_0')_{ij} = \rho_n W(\xi_i, \xi_j)$ for all i, j. Hence, it suffices to prove that

$$\mathbb{E}\big[\delta^2(\widetilde{f}_{\mathbf{\Theta}_0'}, f_0)\big] \le C \frac{\rho_n^2}{n^{\alpha \wedge 1}}.$$

We have

$$\delta^2(\widetilde{f}_{\mathbf{\Theta}_0'}, f_0) = \inf_{\tau \in \mathcal{M}} \sum_{i,j=1}^{n} \int_{(j-1)/n}^{j/n} \int_{(i-1)/n}^{i/n} \big| f_0\big(\tau(x), \tau(y)\big) - \big(\mathbf{\Theta}_0'\big)_{ij} \big|^2 \, dx \, dy.$$

The infimum over all measure-preserving bijections is smaller than the minimum over the subclass of measure-preserving bijections τ satisfying the following property:

$$\int_{(j-1)/n}^{j/n} \int_{(i-1)/n}^{i/n} W\big(\tau(x), \tau(y)\big) \, dx \, dy = \int_{(\sigma(j)-1)/n}^{\sigma(j)/n} \int_{(\sigma(i)-1)/n}^{\sigma(i)/n} W(x, y) \, dx \, dy$$

for some permutation $\sigma = (\sigma(1), \dots, \sigma(n))$ of $\{1, \dots, n\}$. Such τ correspond to permutations of intervals $[(i-1)/n, i/n]$ in accordance with σ. For $(x, y) \in [(\sigma(i)-1)/n, \sigma(i)/n] \times [(\sigma(j)-1)/n, \sigma(j)/n]$ we use the bound

$$\begin{aligned} \big|\rho_n W_0(x, y) - \rho_n W_0(\xi_i, \xi_j)\big| \le{}& \left|\rho_n W_0(x, y) - \rho_n W_0\left(\frac{\sigma(i)}{n+1}, \frac{\sigma(j)}{n+1}\right)\right| \\ &+ \left|\rho_n W_0\left(\frac{\sigma(i)}{n+1}, \frac{\sigma(j)}{n+1}\right) - \rho_n W_0(\xi_{(\sigma(i))}, \xi_{(\sigma(j))})\right| \\ &+ \big|\rho_n W_0(\xi_{(\sigma(i))}, \xi_{(\sigma(j))}) - \rho_n W_0(\xi_i, \xi_j)\big|, \end{aligned} \tag{38}$$

where $\xi_{(m)}$ denotes the mth largest element of the set $\{\xi_1, \dots, \xi_n\}$. We choose a random permutation $\sigma = (\sigma(1), \dots, \sigma(n))$ such that $\xi_{\sigma^{-1}(1)} \le \xi_{\sigma^{-1}(2)} \le \cdots \le \xi_{\sigma^{-1}(n)}$. With this choice of σ, we have that $(\xi_{(\sigma(i))}, \xi_{(\sigma(j))}) = (\xi_i, \xi_j)$ and $|\rho_n W_0(\xi_{(\sigma(i))}, \xi_{(\sigma(j))}) - \rho_n W_0(\xi_i, \xi_j)| = 0$ almost surely.

For the first summand in (38), as $W_0(\cdot, \cdot)$ satisfies (8) and

$$\left(\frac{\sigma(i)}{n+1}, \frac{\sigma(j)}{n+1}\right) \in \big[(\sigma(i)-1)/n, \sigma(i)/n\big] \times \big[(\sigma(j)-1)/n, \sigma(j)/n\big]$$

we get

$$\left|W_0(x,y) - W_0\Big(\frac{\sigma(i)}{n+1}, \frac{\sigma(j)}{n+1}\Big)\right| \le 2Ln^{-\alpha'}, \tag{39}$$

where we set for brevity $\alpha' = \alpha \wedge 1$. To evaluate the contribution of the second summand on the right-hand side of (38), we use that

$$\begin{aligned}&\left|W_0\Big(\frac{\sigma(i)}{n+1}, \frac{\sigma(j)}{n+1}\Big) - W_0(\xi_{(\sigma(i))}, \xi_{(\sigma(j))})\right| \\ &\qquad \le M\Big(\Big|\frac{\sigma(i)}{n+1} - \xi_{(\sigma(i))}\Big|^{\alpha'} + \Big|\frac{\sigma(j)}{n+1} - \xi_{(\sigma(j))}\Big|^{\alpha'}\Big).\end{aligned} \tag{40}$$

Squaring, integrating and taking expectation we obtain

$$\begin{aligned}&\mathbb{E}\Big[\sum_{i,j}\int_{(\sigma(j)-1)/n}^{\sigma(j)/n}\int_{(\sigma(i)-1)/n}^{\sigma(i)/n}\Big|\frac{\sigma(i)}{n+1} - \xi_{(\sigma(i))}\Big|^{2\alpha'} dx\,dy\Big] \\ &\quad = \frac{1}{n}\mathbb{E}\Big[\sum_{m=1}^{n}\Big|\frac{m}{n+1} - \xi_{(m)}\Big|^{2\alpha'}\Big] \\ &\quad \le \max_{m=1,\dots,n}\mathbb{E}\Big[\Big|\frac{m}{n+1} - \xi_{(m)}\Big|^{2\alpha'}\Big] \le \max_{m=1,\dots,n}\big(\mathrm{Var}(\xi_{(m)})\big)^{\alpha'} \le Cn^{-\alpha'},\end{aligned} \tag{41}$$

where we have used the relations $\mathbb{E}(\xi_{(m)}) = \frac{m}{n+1}$, $\mathrm{Var}(\xi_{(m)}) \le C/n$, and Jensen's inequality. The contribution corresponding to the second summand on the right-hand side of (40) is evaluated analogously. Combining (39)–(41) with (38), we get

$$\mathbb{E}[\delta^2(\widetilde{f}_{\boldsymbol{\Theta}_0'}, f_0)] \le C\rho_n^2 n^{-\alpha'}.$$

4.7. *Proof of Corollary 2.7.* To prove the first part of the corollary, notice that its assumptions imply that the partition is balanced and $\rho_n \ge k\log(k)/n$. Thus, similarly to Corollary 2.2, we obtain from Proposition 2.1 that

$$\mathbb{E}\Big[\frac{1}{n^2}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F^2\Big] \le \frac{C}{n^2}\|\boldsymbol{\Theta}_0 - \boldsymbol{\Theta}_{*,n_0}\|_F^2 + C\rho_n\Big(\frac{\log(k)}{n} + \frac{k^2}{n^2}\Big).$$

Using Proposition 2.5 to bound the expectation of the first summand on the right-hand side, we get

$$\mathbb{E}\Big[\frac{1}{n^2}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F^2\Big] \le C\Big\{\frac{\rho_n^2}{k^{2(\alpha\wedge 1)}} + \rho_n\Big(\frac{\log(k)}{n} + \frac{k^2}{n^2}\Big)\Big\}. \tag{42}$$

Now (11) follows from (42) by taking $k = \lceil(\rho_n n^2)^{1/(2(1+\alpha\wedge 1))}\rceil$. Bound (12) for the restricted least squares estimator follows from Propositions 2.3 and 2.6.

4.8. *Proof of Corollary* 3.3. To prove part (i) of the corollary, we control the size of each block of $\mathbf{\Theta}_0$. For any a in $[k]$, the number N_a of nodes belonging to block a is a binomial random variable with parameters $(n, 1/k)$ since the graphon W is balanced. By Bernstein's inequality,

$$\mathbb{P}\left[N_a - \frac{n}{k} \le -t\right] \le \exp\left[-\frac{t^2/2}{n/k + t/3}\right].$$

Taking $t = n/(2k)$ in the above inequality, we obtain

$$\mathbb{P}\left[N_a \le \frac{n}{2k}\right] \le \exp[-Cn/k].$$

This inequality and the union bound over all $a \in [k]$ imply that the size of all blocks of $\mathbf{\Theta}_0$ is greater than n_0 with probability greater than $1 - k\exp(-Cn/k)$. Together with Propositions 2.1 and 3.2, this yields

$$\mathbb{E}\big[\delta^2(\widehat{f}, f_0)\big] \le C'\left(\rho_n \frac{k^2}{n^2} + \rho_n \frac{\log(k)}{n} + \rho_n^2\sqrt{\frac{k}{n}}\right) + k\exp(-Cn/k),$$

where the last summand is negligible. The second part of the corollary is a straightforward consequence of Propositions 2.1 and 2.3.

4.9. *Proof of Proposition* 3.4. It is enough to prove separately the following three minimax lower bounds:

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}[k]} \mathbb{E}_{W_0}\big[\delta^2(\widehat{f}, \rho_n W_0)\big] \ge C\rho_n^2\sqrt{\frac{k-1}{n}}, \tag{43}$$

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}[k]} \mathbb{E}_{W_0}\big[\delta^2(\widehat{f}, \rho_n W_0)\big] \ge C\left(\rho_n \frac{k^2}{n^2} \wedge \rho_n^2\right), \tag{44}$$

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}[2]} \mathbb{E}_{W_0}\big[\delta^2(\widehat{f}, \rho_n W_0)\big] \ge C\left(\frac{\rho_n}{n} \wedge \rho_n^2\right). \tag{45}$$

4.9.1. *Proof of* (43). Without loss of generality, it suffices to prove (43) for $k = 2$ and for all $k = 16\bar{k}$ with integer $\bar{k}$ large enough. Indeed, for any $k > 2$, there exists $k' \le k$ such that $k - k' \le 15$ and k' is either a multiple of 16 or equal to 2, so that

$$\begin{aligned}\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}[k]} \mathbb{E}_{W_0}\big[\delta^2(\widehat{f}; f_0)\big] &\ge \inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}[k']} \mathbb{E}_{W_0}\big[\delta^2(\widehat{f}; f_0)\big] \\ &\ge C\rho_n^2\sqrt{\frac{k'-1}{n}} \ge \frac{C}{4}\rho_n^2\sqrt{\frac{k-1}{n}}.\end{aligned}$$

We first consider k which is a multiple of 16. The case $k = 2$ is sketched afterward.

The proof follows the general scheme of reduction to testing of finite number of hypotheses (cf., e.g., [15]). The main difficulty is to obtain the necessary lower bound for the distance $\delta(\cdot,\cdot)$ between the graphons generating the hypotheses since this distance is defined as a minimum over all measure-preserving bijections. We start by constructing a matrix $\mathbf{B}$ from which the step-function graphons will be derived.

LEMMA 4.2. *Fix $\eta_0 = 1/16$ and assume that k is a multiple of 16 and is greater than some constant. There exists a $k \times k$ symmetric $\{-1, 1\}$ matrix $\mathbf{B}$ satisfying the following two properties*:

- *For any $(a, b) \in [k]$ with $a \neq b$, the inner product between the two columns $\langle \mathbf{B}_{a,\cdot}, \mathbf{B}_{b,\cdot} \rangle$ satisfies*

$$\left|\langle \mathbf{B}_{a,\cdot}, \mathbf{B}_{b,\cdot} \rangle\right| \leq k/4. \tag{46}$$

- *For any two subsets X and Y of $[k]$ satisfying $|X| = |Y| = \eta_0 k$ and $X \cap Y = \varnothing$ and any labelings $\pi_1 : [\eta_0 k] \to X$ and $\pi_2 : [\eta_0 k] \to Y$, we have*

$$\sum_{a,b=1}^{\eta_0 k} [\mathbf{B}_{\pi_1(a),\pi_1(b)} - \mathbf{B}_{\pi_2(a),\pi_2(b)}]^2 \geq \eta_0^2 k^2/2. \tag{47}$$

Note that any Hadamard matrix satisfies condition (46) since its columns are orthogonal. Unfortunately, the second condition (47) seems difficult to check for such matrices. This is why we adopt a probabilistic approach in the proof of this lemma showing that, with positive probability, a symmetric matrix with independent Rademacher entries satisfies the above conditions.

The graphon "hypotheses" that we consider in this proof are generated by the matrix of connection probabilities $\mathbf{Q} := (\mathbf{J} + \mathbf{B})/2$ where $\mathbf{B}$ is a matrix from Lemma 4.2 and $\mathbf{J}$ is a $k \times k$ matrix with all entries equal to 1.

Fix some $\varepsilon < 1/(4k)$. Denote by $\mathcal{C}_0$ the collection of vectors $u \in \{-\varepsilon, \varepsilon\}^k$ satisfying $\sum_{a=1}^k u_a = 0$. For any $u \in \mathcal{C}_0$, define the cumulative distribution F_u on $\{0, \ldots, k\}$ by the relations $F_u(0) = 0$ and $F_u(a) = a/k + \sum_{b=1}^a u_b$ for $a \in [k]$. Then, set $\Pi_{ab}(u) = [F_u(a-1), F_u(a)) \times [F_u(b-1), F_u(b))$ and define the graphon $W_u \in \mathcal{W}[k]$ by

$$W_u(x, y) = \sum_{(a,b)\in[k]\times[k]} \mathbf{Q}_{ab} \mathbb{1}_{\Pi_{ab}(u)}(x, y).$$

The graphon W_u is slightly unbalanced as the weight of each class is either $1/k - \varepsilon$ or $1/k + \varepsilon$.

Let $\mathbb{P}_{W_u}$ denote the distribution of observations $\mathbf{A}' := (\mathbf{A}_{ij}, 1 \leq j < i \leq n)$ sampled according to the sparse graphon model (2) with $W_0 = W_u$. Since the matrix $\mathbf{Q}$ is fixed the difficulty in distinguishing between the distributions $\mathbb{P}_{W_u}$ and $\mathbb{P}_{W_v}$ for $u \neq v$ comes from the randomness of the design points $\xi_1, \ldots, \xi_n$ in the graphon

model (2) rather than from the randomness of the realization of $\mathbf{A}'$ conditionally on $\xi_1, \dots, \xi_n$. The following lemma gives a bound on the Kullback–Leibler divergences $\mathcal{K}(\mathbb{P}_{W_u}, \mathbb{P}_{W_v})$ between $\mathbb{P}_{W_u}$ and $\mathbb{P}_{W_v}$.

LEMMA 4.3. *For all $u, v \in \mathcal{C}_0$, we have*

$$\mathcal{K}(\mathbb{P}_{W_u}, \mathbb{P}_{W_v}) \le 16nk^2\varepsilon^2/3.$$

Next, we need the following combinatorial result in the spirit of the Varshamov–Gilbert lemma ([15], Chapter 2). It shows that there exists a large subset of $\mathcal{C}_0$ composed of vectors u that are well separated in the Hamming distance. We state this result in terms of the sets $\mathcal{A}_u := \{a \in [k] : u_a = \varepsilon\}$ where $u \in \mathcal{C}_0$. Notice that, by definition of $\mathcal{C}_0$, we have $|\mathcal{A}_u| = k/2$ for all $u \in \mathcal{C}_0$.

LEMMA 4.4. *There exists a subset $\mathcal{C}$ of $\mathcal{C}_0$ such that $\log|\mathcal{C}| \ge k/16$ and*

$$|\mathcal{A}_u \Delta \mathcal{A}_v| > k/4 \tag{48}$$

for any $u \ne v \in \mathcal{C}$.

Lemmas 4.4 and 4.2 are used to obtain the following lower bound on the distance $\delta(W_u, W_v)$ between two distinct graphons in $\mathcal{C}$. This lemma is the main ingredient of the proof.

LEMMA 4.5. *For all $u, v \in \mathcal{C}$ such that $u \ne v$, the graphons W_u and W_v are well separated in the $\delta(\cdot,\cdot)$ distance*:

$$\delta^2(W_u, W_v) \ge \eta_0^2 k\varepsilon/2,$$

so that

$$\delta^2(\rho_n W_u, \rho_n W_v) \ge \rho_n^2\eta_0^2 k\varepsilon/2 \qquad \forall u, v \in \mathcal{C} : u \ne v. \tag{49}$$

Now, choose ε such that $\varepsilon^2 = \frac{3}{(16)^3 nk}$. Then it follows from Lemmas 4.3 and 4.4 that

$$\mathcal{K}(\mathbb{P}_{W_u}, \mathbb{P}_{W_v}) \le \frac{1}{16}\log|\mathcal{C}| \qquad \forall u, v \in \mathcal{C} : u \ne v. \tag{50}$$

In view of Theorem 2.7 in [15], inequalities (49) and (50) imply that

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}[k]} \mathbb{E}_{W_0}\big[\delta^2(\widehat{f}, \rho_n W_0)\big] \ge C\rho_n^2\sqrt{\frac{k}{n}},$$

where $C > 0$ is an absolute constant. This completes the proof for the case when k is a large enough multiple of 16.

Let now $k = 2$. Then we reduce the lower bound to the problem of testing two hypotheses. We consider the matrix $\mathbf{B} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$ and the two corresponding

graphons W_{u_1} and W_{u_2} with $u_1 = (\varepsilon, -\varepsilon)$ and $u_2 = (-\varepsilon, \varepsilon)$. Adapting the argument of Lemma 4.5, one can prove that $\delta^2(\rho_n W_{u_1}, \rho_n W_{u_2}) \geq \rho_n^2 \varepsilon$. Moreover, exactly as in Lemma 4.3, the Kullback–Leibler divergence between $\mathbb{P}_{W_{u_1}}$ and $\mathbb{P}_{W_{u_2}}$ is bounded by $Cn\varepsilon^2$. Taking ε of the order $n^{-1/2}$ and using Theorem 2.2 from [15], we conclude that

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}[2]} \mathbb{E}_{W_0}\big[\delta^2(\widehat{f}, \rho_n W_0)\big] \geq C\rho_n^2 \sqrt{\frac{1}{n}},$$

where $C > 0$ is an absolute constant.

PROOF OF LEMMA 4.2. Let $\mathbf{B}$ be a $k \times k$ symmetric random matrix such that its elements $\mathbf{B}_{a,b}$, $a \leq b \in [k]$, are independent Rademacher random variables. It suffices to prove that $\mathbf{B}$ satisfies properties (46) and (47) with positive probability. Fix $a \neq b$. Then $\langle \mathbf{B}_{a,\cdot}, \mathbf{B}_{b,\cdot} \rangle$ is distributed as a sum of k independent Rademacher variables. By Hoeffindg's inequality,

$$\mathbb{P}\big[\big|\langle \mathbf{B}_{a,\cdot}, \mathbf{B}_{b,\cdot} \rangle\big| \geq k/4\big] \leq 2\exp[-k/32].$$

By the union bound, property (46) is satisfied for all $a \neq b$ with probability greater than $1 - 2k^2 \exp[-k/32]$. For k larger than some absolute constant, this probability is greater than $3/4$.

Fix any two subsets X and Y of $[k]$ such that $|X| = |Y| = \eta_0 k$ and $X \cap Y = \varnothing$. Let π_1 and π_2 be any two labelings of X and Y. Then define $T_{\pi_1,\pi_2} := \sum_{a,b=1}^{\eta_0 k} [\mathbf{B}_{\pi_1(a),\pi_1(b)} - \mathbf{B}_{\pi_2(a),\pi_2(b)}]^2$. By symmetry of $\mathbf{B}$, $T_{\pi_1,\pi_2}/8$ is greater than $\sum_{a<b} [\mathbf{B}_{\pi_1(a),\pi_1(b)} - \mathbf{B}_{\pi_2(a),\pi_2(b)}]^2/4$ (we have put aside the diagonal terms) where all the summands $[\mathbf{B}_{\pi_1(a),\pi_1(b)} - \mathbf{B}_{\pi_2(a),\pi_2(b)}]^2/4$ are independent Bernoulli random variables with parameter $1/2$ since $\pi_1([\eta_0 k]) \cap \pi_2([\eta_0 k]) = \varnothing$. Thus, $T_{\pi_1,\pi_2}/8$ is stochastically greater than a binomial random variable with parameters $\eta_0 k(\eta_0 k - 1)/2$ and $1/2$. Applying again Hoeffding's inequality, we find

$$\mathbb{P}\left[T_{\pi_1,\pi_2}/8 \leq \frac{\eta_0^2 k^2}{8} - \frac{\eta_0 k}{4}\right] \leq \exp\left[-\frac{\eta_0^2 k^2}{32}\right].$$

For k large enough, we have $\eta_0 k/4 \leq \eta_0^2 k^2/16$ so that $T_{\pi_1,\pi_2} \geq \eta_0^2 k^2/2$ with probability greater than $1 - \exp(-\eta_0^2 k^2/32)$. There are less than $k^{2\eta_0 k}$ such maps (π_1, π_2) so that property (47) is satisfied with probability greater than $1 - \exp(2\eta_0 k \log(k) - \eta_0^2 k^2/32)$. Again, this probability is greater than $3/4$ for k large enough. Applying once again the union bound, we find that properties (46) and (47) are satisfied with probability greater than $1/2$. □

PROOF OF LEMMA 4.4. Let $\mathcal{C}$ be a maximal subset $\mathcal{C}_0$ of points u such that the corresponding $\mathcal{A}_u$ are $k/4$-separated with respect to the measure of symmetric difference distance. By maximality of $\mathcal{C}$, the union of all balls in the Hamming

distance centered at $u \in \mathcal{C}$ with radii $k/4$ covers $\mathcal{C}_0$. Denoting by $\mathcal{B}[u, k/4]$ such a ball, we obtain by a volumetric argument

$$|\mathcal{B}[u, k/4]|/|\mathcal{C}_0| \geq |\mathcal{C}|^{-1}. \tag{51}$$

If we endow $\mathcal{C}_0$ with the uniform probability, then $|\mathcal{B}[u, k/4]||\mathcal{C}_0|^{-1}$ is the probability to draw a point from $\mathcal{B}[u, k/4]$. Note that a point v is in $\mathcal{B}[u, k/4]$ if $|A_v \setminus A_u| \leq k/8$. One can construct the set A_v doing $k/2$ sampling without replacement from the set $[k]$. We call an $a \in [k]$ a success if $a \in A_u$. There are $k/2$ of them. Let S_k denotes the number of success, then S_k follows an hypergeometric distribution with parameters $(k, k/2, k/2)$ and we have

$$\frac{|\mathcal{B}[u, k/4]|}{|\mathcal{C}_0|} = \mathbb{P}[S_k \geq 3k/8].$$

Now we only have to bound the deviations of S_k. It follows from [1], page 173, that S_k has the same distribution as the random variable $\mathbb{E}[\eta|\mathcal{B}]$ where η is a binomial random variable with parameters $(k/2, 1/2)$ and $\mathcal{B}$ is some suitable σ-algebra. Thus, by a convexity argument, we obtain that, for any $\lambda > 0$,

$$\mathbb{E}e^{\lambda(S_k - k/4)} \leq \mathbb{E}e^{\lambda(\eta - k/4)} \leq e^{\lambda^2 k/16},$$

where the last bound is due to Hoeffding's inequality. Applying the exponential Markov inequality and choosing $\lambda = 1$, we obtain

$$P\left[S_k - \frac{k}{4} \geq k/8\right] \leq \exp[-k/16]$$

and, in view of (51), we conclude that $|\mathcal{C}| \geq \exp[k/16]$. $\square$

PROOF OF LEMMA 4.5. Let u and v be two different vectors in $\mathcal{C}$ and let τ be a measure preserving bijection $[0, 1] \mapsto [0, 1]$. We aim to prove that, for any τ,

$$\int |W_u(x, y) - W_v(\tau(x), \tau(y))|^2 \, dx \, dy \geq \eta_0^2 k\varepsilon/2, \tag{52}$$

where we recall that $\eta_0 = 1/16$.

If x and x' correspond to two different classes of W_u, that is, $x \in [F_u(a-1), F_u(a))$ and $x' \in [F_u(b-1), F_u(b))$ for some $a \neq b$, then the inner product between $W_u(x, \cdot)$ and $W_u(x', \cdot)$ satisfies

$$\begin{aligned} \left| \int (W_u(x, y) - 1/2)(W_u(x', y) - 1/2) \, dy \right| &= \left| \frac{1}{4} \sum_{c=1}^{k} \left(\frac{1}{k} + u_c \right) \mathbf{B}_{ac} \mathbf{B}_{bc} \right| \\ &\leq \frac{1}{4k} \langle \mathbf{B}_{a,\cdot}, \mathbf{B}_{b,\cdot} \rangle + \frac{1}{4} k\varepsilon \\ &\leq 1/8, \end{aligned} \tag{53}$$

since we assume that $4k\varepsilon \leq 1$.

For any $a, b \in [k]$, define ω_{ab} the Lebesgue measure of $[F_u(a-1), F_u(a)] \cap \tau([F_v(b-1), F_v(b)])$. Since τ is measure preserving, $\sum_b \omega_{ab} = 1/k + u_a$ and $\sum_a \omega_{ab} = 1/k + v_b$. For any a and $b \in [k]$ define $h_{u,a}(y) = W_u(F_u(a-1), y) - 1/2$ and $k_{v,b}(y) = W_v(F_v(b-1), \tau(y)) - 1/2$. Then we have

$$\int \left|W_u(x,y) - W_v\big(\tau(x), \tau(y)\big)\right|^2 dx\, dy = \sum_{a=1}^{k}\sum_{b=1}^{k} \omega_{a,b} \int \left|h_{u,a}(y) - k_{v,b}(y)\right|^2 dy.$$

Let $\langle \cdot, \cdot \rangle_2$ and $\| \cdot \|_2$ denote the standard inner product and the Euclidean norm in $L_2([0,1])$. By definition of W_u we have that $|k_{v,a}(y)| = 1/2$ for all $y \in [0,1]$ and any $a \in [k]$ which implies $\|k_{v,a}\|_2 = 1/2$. Now for $b_1 \neq b_2$, $\|k_{v,b_1} - k_{v,b_2}\|_2^2 \geq 1/2 - 1/4 = 1/8$ by (53). By the triangle inequality, this implies that

$$\|h_{u,a} - k_{v,b_1}\|_2^2 + \|h_{u,a} - k_{v,b_2}\|_2^2 \geq \frac{\|k_{v,b_1} - k_{v,b_2}\|_2^2}{2} \geq 1/16.$$

As a consequence, for any $a \in [k]$ there exists at most one $b \in [k]$ such that $\|h_{u,a} - k_{v,b}\|^2 < 1/32$. If such an index b exists, it is denoted by $\pi(a)$. Exactly the same argument shows that for any $b \in [k]$ there exists at most one $a \in [k]$ such that $\|h_{u,a} - k_{v,b}\|^2 < 1/32$ which implies that there exists no $a \neq a'$ such that $\pi(a) = \pi(a')$. Thus, it is possible to extend π to a permutation of $[k]$. We get

$$\begin{aligned}\int \left|W_u(x,y) - W_v\big(\tau(x), \tau(y)\big)\right|^2 dx\, dy &\geq \frac{1}{32}\sum_{a=1}^{k}\sum_{b \neq \pi(a)} \omega_{a,b} \\ &= \frac{1}{32}\sum_{a=1}^{k}(1/k + u_a - \omega_{a,\pi(a)}).\end{aligned}$$

If the sum $\sum_{a=1}^{k} 1/k + u_a - \omega_{a,\pi(a)}$ is greater than $k\varepsilon/16$, then (52) is satisfied. Thus, we can assume in the sequel that $\sum_{a=1}^{k} 1/k + u_a - \omega_{a,\pi(a)} \leq k\varepsilon/16$. Using that $\omega_{a,b} \leq (1/k + u_a) \wedge (1/k + v_b)$ and the cardinality of the collection $\{a \in [k] : u_a > 0\}$ is $k/2$ we deduce that the collection $\{a \in [k] : u_a > 0, v_{\pi(a)} > 0 \text{ and } \omega_{a,\pi(a)} \geq 1/k\}$ has cardinality greater than $7k/16$. Since for $u \neq v \in \mathcal{C}$, $|\mathcal{A}_u \cap \mathcal{A}_v| \leq 3k/8$ (Lemma 4.4), there exist subsets $A \subset \mathcal{A}_u$ and $B \subset \mathcal{A}_v$ of cardinality $\eta_0 k$ (recall that $\eta_0 = 1/16$) such that $\pi(A) = B$, $A \cap B = \varnothing$, and $\omega_{a,\pi(a)} \geq 1/k$ for all $a \in A$. Hence,

$$\begin{aligned}&\int \left|W_u(x,y) - W_v\big(\tau(x), \tau(y)\big)\right|^2 dx\, dy \\ &\quad\geq \sum_{a_1 \in A}\sum_{a_2 \in A} \int_{[F_u(a_1-1), F_u(a_1)) \times [F_u(a_2-1), F_u(a_2))} \left|W_u(x,y)\right. \\ &\quad\quad \left. - W_v(\tau(x), \tau(y)\right|^2 dx\, dy\end{aligned}$$

$$\geq \sum_{a_1\in A}\sum_{a_2\in A} \omega_{a_1,\pi(a_1)}\omega_{a_2,\pi(a_2)}[\mathbf{Q}_{a_1,a_2}-\mathbf{Q}_{\pi(a_1),\pi(a_2)}]^2$$

$$\geq \frac{1}{4k^2}\sum_{a_1\in A}\sum_{a_2\in A}[\mathbf{B}_{a_1,a_2}-\mathbf{B}_{\pi(a_1),\pi(a_2)}]^2,$$

where the last inequality follows from the facts that $\mathbf{Q}=(\mathbf{J}+\mathbf{B})/2$ and $\omega_{a,\pi(a)}\geq 1/k$. Finally, we apply the property (47) of $\mathbf{B}$ to conclude that

$$\int |W_u(x,y)-W_v(\tau(x),\tau(y))|^2\,dx\,dy \geq \eta_0^2/8 \geq \eta_0^2 k\varepsilon/2.$$

This proves (52), and thus the lemma. □

PROOF OF LEMMA 4.3. For $u\in\mathcal{C}_0$, let $\zeta(u)=(\zeta_1(u),\dots,\zeta_n(u))$ be the vector of n i.i.d. random variables with the discrete distribution on $[k]$ defined by $\mathbb{P}[\zeta_1(u)=a]=1/k+u_a$ for any $a\in[k]$. Let $\mathbf{\Theta}_0$ be the $n\times n$ symmetric matrix with elements $(\mathbf{\Theta}_0)_{ii}=0$ and $(\mathbf{\Theta}_0)_{ij}=\rho_n\mathbf{Q}_{\zeta_i(u),\zeta_j(u)}$ for $i\neq j$. Assume that, conditionally on $\zeta(u)$, the adjacency matrix $\mathbf{A}$ is sampled according to the network sequence model with such probability matrix $\mathbf{\Theta}_0$. Notice that in this case the observations $\mathbf{A}'=(\mathbf{A}_{ij},1\leq j<i\leq n)$ have the probability distribution $\mathbb{P}_{W_u}$. Using this remark and introducing the probabilities $\alpha_{\mathbf{a}}(u)=\mathbb{P}[\zeta(u)=\mathbf{a}]$ and $p_{A\mathbf{a}}=\mathbb{P}[\mathbf{A}'=A|\zeta(u)=\mathbf{a}]$ for $\mathbf{a}\in[k]^n$, we can write the Kullback–Leibler divergence between $\mathbb{P}_{W_u}$ and $\mathbb{P}_{W_v}$ in the form

$$\mathcal{K}(\mathbb{P}_{W_u},\mathbb{P}_{W_v})=\sum_A\sum_{\mathbf{a}} p_{A\mathbf{a}}\alpha_{\mathbf{a}}(u)\log\biggl(\frac{\sum_{\mathbf{a}}p_{A\mathbf{a}}\alpha_{\mathbf{a}}(u)}{\sum_{\mathbf{a}}p_{A\mathbf{a}}\alpha_{\mathbf{a}}(v)}\biggr),$$

where the sums in $\mathbf{a}$ are over $[k]^n$ and the sum in A is over all triangular upper halves of matrices in $\{0,1\}^{n\times n}$. Since the function $(x,y)\mapsto x\log(x/y)$ is convex, we can apply Jensen's inequality to get

(54)
$$\mathcal{K}(\mathbb{P}_{W_u},\mathbb{P}_{W_v})\leq\sum_{\mathbf{a}}\alpha_{\mathbf{a}}(u)\log\biggl(\frac{\alpha_{\mathbf{a}}(u)}{\alpha_{\mathbf{a}}(v)}\biggr)=n\sum_{a\in[k]}(1/k+u_a)\log\biggl(\frac{1/k+u_a}{1/k+v_a}\biggr),$$

where the last equality follows from the fact that $\alpha_{\mathbf{a}}(u)$ are n-product probabilities. Since the Kullback–Leibler divergence does not exceed the chi-square divergence, we obtain

$$\sum_{a\in[k]}(1/k+u_a)\log\biggl(\frac{1/k+u_a}{1/k+v_a}\biggr)\leq\sum_{a\in[k]}\frac{(u_a-v_a)^2}{1/k+v_a}\leq 16k^2\varepsilon^2/3,$$

where last inequality uses that $|v_a|\leq\varepsilon\leq 1/(4k)$, and $|u_a-v_a|\leq 2\varepsilon$. Combining this with (54) proves the lemma. □

4.9.2. *Proof of* (44). As in the proof of Proposition 2.4, we use here Fano's method. The main difference is that the graphon separation in $\delta(\cdot,\cdot)$ distance is more difficult to handle than the matrix separation in the Frobenius distance.

First, note that it is sufficient to prove (44) for $k > k_0$ where k_0 is any fixed integer. Indeed, if $k \le k_0$, the lower bound (44) immediately follows from (45).

Let $\mathcal{C}_0$ denote the set of all symmetric $k \times k$ matrices with entries in $\{-1, 1\}$. The graphon hypotheses that we consider in this proof are generated by matrices of intra-class connection probabilities of the form $\mathbf{Q_B} := (\mathbf{J} + \varepsilon\mathbf{B})/2$ where $\mathbf{B} \in \mathcal{C}_0$, $\varepsilon \in (0, 1/2)$, and $\mathbf{J}$ is a $k \times k$ matrix with all entries equal to 1. Given a matrix $\mathbf{Q_B}$, define the graphon $W_{\mathbf{B}} \in \mathcal{W}[k]$ by the formula

$$W_{\mathbf{B}}(x, y) = \sum_{(a,b)\in[k]\times[k]} (\mathbf{Q_B})_{ab} \mathbf{1}_{[(a-1)/k,a/k)\times[(b-1)/k,b/k)}(x, y).$$

As in the previous proof, we use the following combinatorial result in the spirit of the Varshamov–Gilbert lemma [15], Chapter 2, that grants the existence of a large subset $\mathcal{C}$ of $\mathcal{C}_0$ such that the matrices $\mathbf{B} \in \mathcal{C}$ are well separated in some sense. Given any two permutations π and π' of $[k]$ and any matrix $\mathbf{B}$, we denote by $\mathbf{B}^{\pi,\pi'}$ a matrix with entries $\mathbf{B}^{\pi,\pi'}_{ab} = \mathbf{B}_{\pi(a)\pi'(b)}$.

LEMMA 4.6. *Let k_0 be an integer large enough. For any $k > k_0$, there exists a subset $\mathcal{C}$ of $\mathcal{C}_0$ satisfying* $\log|\mathcal{C}| \ge k^2/32$ *and such that*

$$\|\mathbf{B}_1 - \mathbf{B}_2^{\pi,\pi'}\|_F^2 \ge \frac{k^2}{2} \tag{55}$$

for all permutations π, π' and all $\mathbf{B}_1 \ne \mathbf{B}_2 \in \mathcal{C}$.

We assume in the rest of this proof that k is greater than k_0. As noticed above, it is enough to prove (44) only in this case. We choose a maximal subset $\mathcal{C}$ satisfying the properties stated in Lemma 4.6. The next lemma shows that the separation between matrices $\mathbf{B}$ in $\mathcal{C}$ translates into separation between the corresponding graphons $W_{\mathbf{B}}$.

LEMMA 4.7. *Let $\mathcal{C}$ be a maximal set satisfying the properties stated in Lemma 4.6. For any two distinct $\mathbf{B}_1$ and $\mathbf{B}_2$ in $\mathcal{C}$, we have*

$$\delta^2(\rho_n W_{\mathbf{B}_1}, \rho_n W_{\mathbf{B}_2}) \ge \rho_n^2\varepsilon^2/8. \tag{56}$$

Finally, in order to apply Fano's method, we need to have an upper bound on the Kullback–Leibler divergence between the distributions $\mathbb{P}_{W_{\mathbf{B}}}$ for $\mathbf{B} \in \mathcal{C}$. It is given in the next lemma.

LEMMA 4.8. *For any $\mathbf{B}_1$ and $\mathbf{B}_2$ in $\mathcal{C}$, we have*

$$\mathcal{K}(\mathbb{P}_{W_{\mathbf{B}_1}}, \mathbb{P}_{W_{\mathbf{B}_2}}) \le 3n^2\rho_n\varepsilon^2.$$

Choosing now $\varepsilon = (\frac{k}{n\sqrt{\rho_n}} \wedge 1)/32$ we deduce from Lemmas 4.8 and 4.6 that

$$\mathcal{K}(\mathbb{P}_{W_{\mathbf{B}_1}}, \mathbb{P}_{W_{\mathbf{B}_2}}) \le \frac{1}{10}\log|\mathcal{C}| \qquad \forall \mathbf{B}_1, \mathbf{B}_2 \in \mathcal{C} : \mathbf{B}_1 \ne \mathbf{B}_2. \tag{57}$$

In view of Theorem 2.7 in [15], inequalities (56) and (57) imply that

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}[k]} \mathbb{E}_{W_0}\big[\delta^2(\widehat{f}, \rho_n W_0)\big] \ge C\bigg(\rho_n \frac{k^2}{n^2} \wedge \rho_n^2\bigg),$$

where $C > 0$ is an absolute constant. This completes the proof of (44).

PROOF OF LEMMA 4.6. Define the pseudo-distance d_Π on $\mathcal{C}_0$ by $d_\Pi(\mathbf{B}_1, \mathbf{B}_2) := \min_{\pi,\pi'} \|\mathbf{B}_1 - \mathbf{B}_2^{\pi,\pi'}\|_F$, where the minimum is taken over all permutations of $[k]$. Let $\mathcal{C}$ be a maximal subset $\mathcal{C}_0$ of matrices $\mathbf{B}$ that are $\sqrt{k(k+1)/2}$-separated with respect to d_Π. By maximality of $\mathcal{C}$, the union of all balls in d_Π distance centered at $\mathbf{B} \in \mathcal{C}$ with radii $\sqrt{k(k+1)/2}$ covers $\mathcal{C}_0$. Denoting by $\mathcal{B}[\mathbf{B}, \sqrt{k(k+1)/2}, d_\Pi]$ such a ball, we obtain by a volumetric argument

$$\big|\mathcal{B}\big[\mathbf{B}, \sqrt{k(k+1)/2}, d_\Pi\big]\big|/|\mathcal{C}_0| \ge |\mathcal{C}|^{-1}.$$

Since $d_\Pi(\mathbf{B}_1, \mathbf{B}_2)$ is the infimum over all $k!^2$ permutations of the distance $\|\mathbf{B}_1 - \mathbf{B}_2^{\pi,\pi'}\|_F$, we have

$$\big|\mathcal{B}\big[\mathbf{B}, \sqrt{k(k+1)/2}, d_\Pi\big]\big| \le k!^2 \big|\mathcal{B}\big[\mathbf{B}, \sqrt{k(k+1)/2}, \|\cdot\|_F\big]\big|,$$

where $\mathcal{B}[\mathbf{B}, \sqrt{k(k+1)/2}, \|\cdot\|_F]$ is the ball centered at $\mathbf{B}$ with respect to the Frobenius distance. Given $\mathbf{B}_1$ and $\mathbf{B}_2$ in $\mathcal{C}_0$, $\|\mathbf{B}_1 - \mathbf{B}_2\|_F^2/4$ is equal to the Hamming distance between $\mathbf{B}_1$ and $\mathbf{B}_2$. Consider the ball $\mathcal{B}[\mathbf{B}, k(k+1)/8, d_H]$ centered at $\mathbf{B}$ with respect to the Hamming distance. As matrices in $\mathcal{C}_0$ are symmetric, $\mathcal{C}_0$ is in bijection with $\{0,1\}^{k(k+1)/2}$. Using Hamming bound and Varshamov–Gilbert lemma [15], Chapter 2, we obtain

$$2^{-k(k+1)/2}\big|\mathcal{B}\big[\mathbf{B}, k(k+1)/8, d_H\big]\big| \le |C'|^{-1} \le e^{-k(k+1)/16},$$

where C' is a maximal subset of matrices $\mathbf{B}$ that are $\frac{k(k+1)}{4}$-separated with respect to the Hamming distance. We then conclude that

$$|\mathcal{C}| \ge \frac{|\mathcal{C}_0|}{k!^2 |\mathcal{B}[\mathbf{B}, \sqrt{k(k+1)/2}, \|\cdot\|_F]|} \ge \exp\big[k\big\{(k+1)/16 - 2\log(k)\big\}\big],$$

which is larger $\exp(k^2/32)$ for k large enough. □

PROOF OF LEMMA 4.7. Let $\mathbf{B}_1$ and $\mathbf{B}_2$ be two distinct matrices in $\mathcal{C}$ and let τ be a measure preserving bijection $[0,1] \mapsto [0,1]$. Our aim is to prove that, for any such τ,

$$\int \big|W_{\mathbf{B}_1}(x,y) - W_{\mathbf{B}_2}\big(\tau(x), \tau(y)\big)\big|^2 \, dx\, dy \ge \varepsilon^2/8. \tag{58}$$

For any $a, b \in [k]$, let ω_{ab} denote the Lebesgue measure of $[(a-1)/k, a/k] \cap \tau([(b-1)/k, b/k])$. Since τ is measure preserving, $\sum_b \omega_{ab} = 1/k$ and $\sum_a \omega_{ab} = 1/k$. Hence, the $k \times k$ matrix $k\omega$, where $\omega = (\omega_{ab})_{a,b\in[k]}$, is doubly stochastic. For any permutation π of $[k]$, denote $H(\pi)$ the corresponding permutation matrix. By the Birkhoff–von Neumann theorem [8], $k\omega$ is a convex combination of permutation matrices, that is, there exist positive numbers γ_π such that $\sum_\pi \gamma_\pi = 1/k$, and $\omega = \sum_\pi \gamma_\pi H(\pi)$ where the summation runs over all permutations. Using these remarks, we obtain

$$\begin{aligned}
&\int \bigl|W_{\mathbf{B}_1}(x,y) - W_{\mathbf{B}_2}\bigl(\tau(x),\tau(y)\bigr)\bigr|^2\,dx\,dy \\
&\quad= \frac{\varepsilon^2}{4} \sum_{a_1,a_2,b_1,b_2\in[k]} \omega_{a_1a_2}\omega_{b_1b_2}\bigl[(\mathbf{B}_1)_{a_1b_1} - (\mathbf{B}_2)_{a_2b_2}\bigr]^2 \\
&\quad= \frac{\varepsilon^2}{4} \sum_{\pi_1,\pi_2} \sum_{a_1,a_2,b_1,b_2\in[k]} \gamma_{\pi_1}\gamma_{\pi_2} H(\pi_1)_{a_1a_2} H(\pi_2)_{b_1b_2}\bigl[(\mathbf{B}_1)_{a_1b_1} - (\mathbf{B}_2)_{a_2b_2}\bigr]^2 \\
&\qquad\text{(by the definition of permutation matrices)} \\
&\quad= \frac{\varepsilon^2}{4} \sum_{\pi_1,\pi_2} \gamma_{\pi_1}\gamma_{\pi_2} \sum_{a,b}\bigl[(\mathbf{B}_1)_{ab} - (\mathbf{B}_2)_{\pi_1(a)\pi_2(b)}\bigr]^2 \\
&\quad= \frac{\varepsilon^2}{4} \sum_{\pi_1,\pi_2} \gamma_{\pi_1}\gamma_{\pi_2} \bigl\|\mathbf{B}_1 - \mathbf{B}_2^{\pi_1,\pi_2}\bigr\|_F^2 \\
&\quad\ge \sum_{\pi_1,\pi_2} \gamma_{\pi_1}\gamma_{\pi_2} \frac{\varepsilon^2 k^2}{8} = \frac{\varepsilon^2}{8},
\end{aligned}$$

where we have used Lemma 4.6 and the property $\sum_\pi \gamma_\pi = 1/k$. □

PROOF OF LEMMA 4.8. The proof is quite similar to that of Lemma 4.3. Fix two matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ in $\mathcal{C}$. Let $\zeta = (\zeta_1, \dots, \zeta_n)$ be the vector of n i.i.d. random variables with uniform distribution on $[k]$. Let $\mathbf{\Theta}_1$ be the $n \times n$ symmetric matrix with elements $(\mathbf{\Theta}_1)_{ii} = 0$ and $(\mathbf{\Theta}_1)_{ij} = \rho_n[1 + (\mathbf{B}_1)_{\zeta_i,\zeta_j}]/2$ for $i \ne j$. Assume that, conditionally on $\zeta(u)$, the adjacency matrix $\mathbf{A}$ is sampled according to the network sequence model with such probability matrix $\mathbf{\Theta}_1$. Notice that in this case the observations $\mathbf{A}' = (\mathbf{A}_{ij}, 1 \le j < i \le n)$ have the probability distribution $\mathbb{P}_{W_{\mathbf{B}_1}}$. Using this remark, we introduce the probabilities $\alpha_{\mathbf{a}} = \mathbb{P}[\zeta = \mathbf{a}]$ and $p^{(1)}_{A\mathbf{a}} = \mathbb{P}[\mathbf{A}' = A|\zeta = \mathbf{a}]$ for $\mathbf{a} \in [k]^n$. Next, we introduce the analogous probabilities $p^{(2)}_{A\mathbf{a}}$ for a matrix $\mathbf{\Theta}_2$ depending on $\mathbf{B}_2$ in the same way as $\mathbf{\Theta}_1$ depends on $\mathbf{B}_1$. The Kullback–Leibler divergence between $\mathbb{P}_{W_{\mathbf{B}_1}}$ and $\mathbb{P}_{W_{\mathbf{B}_2}}$ has the form

$$\mathcal{K}(\mathbb{P}_{W_{\mathbf{B}_1}}, \mathbb{P}_{W_{\mathbf{B}_2}}) = \sum_A \sum_{\mathbf{a}} \alpha_{\mathbf{a}} p^{(1)}_{A\mathbf{a}} \log\biggl(\frac{\sum_{\mathbf{a}} \alpha_{\mathbf{a}} p^{(1)}_{A\mathbf{a}}}{\sum_{\mathbf{a}} \alpha_{\mathbf{a}} p^{(2)}_{A\mathbf{a}}}\biggr),$$

where the sums in $\mathbf{a}$ are over $[k]^n$ and the sum in A is over all triangular upper halves of matrices in $\{0,1\}^{n\times n}$. Since the function $(x,y)\mapsto x\log(x/y)$ is convex, we can apply Jensen's inequality to get

$$\mathcal{K}(\mathbb{P}_{W_{\mathbf{B}_1}},\mathbb{P}_{W_{\mathbf{B}_2}})\le\sum_{\mathbf{a}}\alpha_{\mathbf{a}}\sum_{A}p^{(1)}_{A\mathbf{a}}\log\Big(\frac{p^{(1)}_{A\mathbf{a}}}{p^{(2)}_{A\mathbf{a}}}\Big).$$

Here, the sum in A for fixed $\mathbf{a}$ is the Kullback–Leibler divergence between two $n(n-1)/2$-products of Bernoulli measures, each of which has success probability either $\rho_n(1+\varepsilon)/2$ or $\rho_n(1-\varepsilon)/2$. Thus, for $p=(1+\varepsilon)/2$ and $q=(1-\varepsilon)/2$ we have

$$\mathcal{K}(\mathbb{P}_{W_{\mathbf{B}_1}},\mathbb{P}_{W_{\mathbf{B}_2}})\le\frac{n(n-1)}{2}\rho_n\kappa(p,q),\tag{59}$$

where $\kappa(p,q)$ is the Kullback–Leibler divergence between the Bernoulli measures with success probabilities p and q. Since the Kullback–Leibler divergence does not exceed the chi-square divergence, we obtain $\kappa(p,q)\le(p-q)^2(p^{-1}+q^{-1})=4\varepsilon^2/(1-\varepsilon^2)\le 16\varepsilon^2/3$ for any $\varepsilon<1/2$. The lemma now follows by substitution of this bound on $\kappa(p,q)$ into (59). □

4.9.3. *Proof of* (45). We use here a reduction to the problem of testing two simple hypotheses by Le Cam's method. Fix some $0<\varepsilon\le 1/4$. Let W_1 be the constant graphon with $W_1(x,y)\equiv 1/2$, and let $W_2\in\mathcal{W}[2]$ be the 2-step graphon with $W_2(x,y)=1/2+\varepsilon$ if $x,y\in[0,1/2)^2\cup[1/2,1]^2$ and $W_2(x,y)=1/2-\varepsilon$ elsewhere. Obviously, we have

$$\delta^2(\rho_nW_1,\rho_nW_2)=\rho_n^2\varepsilon^2.\tag{60}$$

We have

$$\begin{aligned}
&\inf_{\widehat{f}}\max_{W_0\in\{W_1,W_2\}}\mathbb{E}_{W_0}\big[\delta^2(\widehat{f},\rho_nW_0)\big]\\
&\quad\ge\frac{1}{2}\int\big(\delta^2(\widehat{f},\rho_nW_1)+\delta^2(\widehat{f},\rho_nW_2)\big)\min(d\mathbb{P}_{W_1},d\mathbb{P}_{W_2})\\
&\quad\ge\frac{\delta^2(\rho_nW_1,\rho_nW_2)}{4}\int\min(d\mathbb{P}_{W_1},d\mathbb{P}_{W_2})\\
&\quad\ge\frac{\rho_n^2\varepsilon^2}{8}\exp\big(-\chi^2(\mathbb{P}_{W_2},\mathbb{P}_{W_1})\big),
\end{aligned}\tag{61}$$

where $\chi^2(\mathbb{P}_{W_2},\mathbb{P}_{W_1})$ is the chi-square divergence between $\mathbb{P}_{W_2}$ and $\mathbb{P}_{W_1}$. In the last inequality, we have used (2.24) and (2.26) from [15], and (60). Finally, the following lemma allows us to conclude the proof by setting $\varepsilon=\sqrt{\frac{c_0}{n\rho_n}}\wedge\frac{1}{4}$.

LEMMA 4.9. *There exists an absolute constant $c_0>0$ such that $\chi^2(\mathbb{P}_{W_2},\mathbb{P}_{W_1})\le 1/4$ if ε satisfies $n\rho_n\varepsilon^2\le c_0$.*

PROOF. Let $L(\mathbf{A}')$ be likelihood ratio of $\mathbb{P}_{W_2}$ with respect to $\mathbb{P}_{W_1}$. Since $\chi^2(\mathbb{P}_{W_2}, \mathbb{P}_{W_1}) = \mathbb{E}_{W_1}[L(\mathbf{A}')^2] - 1$ it remains to prove that $\mathbb{E}_{W_1}[L(\mathbf{A}')^2] \le 5/4$.

For the sake of brevity, we write in what follows $\mathbb{E}[\cdot] = \mathbb{E}_{W_1}[\cdot]$. We also set $p_0 := \rho_n/2$, $p_1 := \rho_n(1/2 + \varepsilon)$ and $p_2 := \rho_n(1/2 - \varepsilon)$.

As the graphon $\rho_n W_2$ is a 2-step function, we may assume that the components of $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$ are i.i.d. uniformly distributed on $\{0, 1\}$. Given $\boldsymbol{\xi}$, define the collection $S := \{\{a, b\} : \xi_a = \xi_b\}$ of subsets of indices with identical position. For $\{i, j\}$ in S (resp., S^c), $\mathbf{A}_{i,j}$ follows a Bernoulli distribution with parameter p_1 (resp., p_2). Denote μ the distribution of S and $n^{(2)} = n(n-1)/2$ the number of subsets of size 2 of $[n]$. Then the likelihood has the form

$$
\begin{aligned}
L(\mathbf{A}') &= \int L_S(\mathbf{A}')\, d\mu(S), \\
L_S(\mathbf{A}') &:= \left(\frac{1-p_1}{1-p_0}\right)^{|S|} \left(\frac{1-p_2}{1-p_0}\right)^{n^{(2)}-|S|} \prod_{\{a,b\}\in S} \left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^{\mathbf{A}_{ab}} \\
&\quad \times \prod_{\{a,b\}\in S^c} \left(\frac{p_2(1-p_0)}{p_0(1-p_2)}\right)^{\mathbf{A}_{ab}}.
\end{aligned}
$$

By Fubini's theorem, we may write

$$
\mathbb{E}\big[L^2(\mathbf{A}')\big] = \int \mathbb{E}\big[L_{S_1}(\mathbf{A}')L_{S_2}(\mathbf{A}')\big]\, d\mu(S_1)\, d\mu(S_2),
$$

where

$$
\begin{aligned}
&\mathbb{E}\big[L_{S_1}(\mathbf{A}')L_{S_2}(\mathbf{A}')\big] \\
&\quad = \left(\frac{1-p_1}{1-p_0}\right)^{|S_1|+|S_2|} \left(\frac{1-p_2}{1-p_0}\right)^{2n^{(2)}-|S_1|-|S_2|} \mathbb{E}\Bigg[\prod_{\{a,b\}\in S_1\cap S_2} \left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^{2\mathbf{A}_{ab}} \\
&\qquad \times \prod_{\{a,b\}\in S_1^c\cap S_2^c} \left(\frac{p_2(1-p_0)}{p_0(1-p_2)}\right)^{2\mathbf{A}_{ab}} \prod_{\{a,b\}\in S_1\Delta S_2} \left(\frac{p_1p_2(1-p_0)^2}{p_0^2(1-p_1)(1-p_2)}\right)^{\mathbf{A}_{ab}}\Bigg] \\
&\quad = \left[1 + \frac{(p_1-p_0)^2}{p_0(1-p_0)}\right]^{|S_1\cap S_2|} \left[1 + \frac{(p_2-p_0)^2}{p_0(1-p_0)}\right]^{|S_1^c\cap S_2^c|} \\
&\qquad \times \left[1 + \frac{p_1p_2 + p_0^2 - p_1p_0 - p_2p_0}{p_0(1-p_0)}\right]^{|S_1\Delta S_2|}.
\end{aligned}
$$

Using the definition of p_0, p_1 and p_2, we find

$$
\begin{aligned}
&\mathbb{E}\big[L_{S_1}(\mathbf{A}')L_{S_2}(\mathbf{A}')\big] \\
&\quad = \left[1 + \frac{\rho_n^2\varepsilon^2}{p_0(1-p_0)}\right]^{|S_1\cap S_2|+|S_1^c\cap S_2^c|} \left[1 - \frac{\rho_n^2\varepsilon^2}{p_0(1-p_0)}\right]^{|S_1\Delta S_2|}
\end{aligned}
$$

$$\leq \left[1 + \frac{\rho_n^2 \varepsilon^2}{p_0(1-p_0)}\right]^{|S_1 \cap S_2| + |S_1^c \cap S_2^c| - |S_1 \Delta S_2|}$$
$$\leq \exp\bigl[(2|S_1 \cap S_2| + 2|S_1^c \cap S_2^c| - n^{(2)}) 4\rho_n \varepsilon^2\bigr].$$

Thus, to bound the second moment of $L(\mathbf{A}')$, it suffices to control an exponential moment of $T := |S_1 \cap S_2| + |S_1^c \cap S_2^c|$ where S_1 and S_2 are independent and distributed according to μ. To handle this quantity, we denote by $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ and by $\boldsymbol{\xi}' = (\xi_1', \ldots, \xi_n')$ the positions for the first and second sample corresponding to S_1 and S_2, respectively. Next, for any $i, j \in \{0,1\}^2$, define

$$N_{ij} := \bigl|\{a : \xi_a = i \text{ and } \xi_a' = j\}\bigr|.$$

Then we have $2|S_1 \cap S_2| + n = N_{00}^2 + N_{01}^2 + N_{10}^2 + N_{11}^2$ and $2|S_1^c \cap S_2^c| = 2N_{00}N_{11} + 2N_{01}N_{10}$ so that $2T + n = (N_{00} + N_{11})^2 + (N_{01} + N_{10})^2$. Define the random variable $Z := N_{00} + N_{11} - n/2$. It has a centered binomial distribution with parameters n and $1/2$. We have

$$2T - n^{(2)} = (n/2 + Z)^2 + (n/2 - Z)^2 - n - n^{(2)} = 2Z^2 - n/2.$$

Plugging this identity into the expression for $\mathbb{E}[L^2(\mathbf{A}')]$, we conclude that

$$\mathbb{E}\bigl[L^2(\mathbf{A}')\bigr] \leq \mathbf{E}\bigl[\exp\bigl(8\rho_n \varepsilon^2 Z^2\bigr)\bigr],$$

where $\mathbf{E}[\cdot]$ stands for the expectation with respect to the distribution of Z. By Hoeffding's inequality, Z has a sub-Gaussian distribution with sub-Gaussian norm smaller than $n^{1/2}$. Consequently, $\mathbb{E}[L^2(\mathbf{A}')] \leq 5/4$ as soon as $8\rho_n n\varepsilon^2$ is smaller than some numerical constant. □

4.10. *Technical lemmas.*

LEMMA 4.10. *Let $X_1, \ldots, X_n$ be i.i.d. uniformly distributed variables on $[0,1]$ and $X_{(i)}$ is the ith element of the ordered sample $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. Then, for any $n_0 \leq n$, and $0 \leq s < n_0$,*

$$\mathbb{E}(X_{(i)} - X_{(i+s)})^2 = \frac{s(s+1)}{(n+1)(n+2)} \leq \left(\frac{n_0}{n}\right)^2.$$

PROOF. Note that $X_{(i)} - X_{(i+s)} \sim \mathrm{Beta}(s, n-s+1)$. For $Y \sim \mathrm{Beta}(\beta, \gamma)$ we have

$$\mathbb{E}\bigl(Y^2\bigr) = \frac{\beta(\beta+1)}{(\beta+\gamma+1)(\beta+\gamma)}$$

which implies the lemma. □

PROOF OF LEMMA 4.1. Since $\|\hat{\mathbf{T}}\|_\infty \leq 2r$, the definition of $q_{\max}$ and the fact that $\|\hat{\mathbf{T}}\|_F \geq r$ imply $\varepsilon_0 \leq \|\hat{\mathbf{T}}\|_F \leq \varepsilon_{q_{\max}}$. Denote by $\hat{q}$ the integer such that

$2^{-1}\varepsilon_{\hat{q}} \le \|\hat{\mathbf{T}}\|_F < \varepsilon_{\hat{q}}$. Let $\hat{\mathbf{V}}_0$ be a matrix minimizing $\|\hat{\mathbf{T}}/\|\hat{\mathbf{T}}\|_F - \mathbf{V}\|_F$ over all $\mathbf{V}$ in $\mathcal{C}_{\hat{z}_r}$. Then take $\hat{\mathbf{V}} = \hat{\mathbf{V}}_0^{\hat{q},\hat{\mathbf{U}},\hat{z}_r}$ where $\hat{\mathbf{U}}$ is a minimizer of $\|\hat{\mathbf{T}} - \hat{\mathbf{V}}_0^{\hat{q},\mathbf{U},\hat{z}_r}\|_F$ over all $\mathbf{U} \in \{-1,0,1\}^{k\times k}$. Notice that $\hat{\mathbf{U}}$ is also a minimizer of $\|\hat{\mathbf{T}} - \hat{\mathbf{V}}_0^{\hat{q},\mathbf{U},\hat{z}_r}\|_\infty$ over all $\mathbf{U} \in \{-1,0,1\}^{k\times k}$ since both $\hat{\mathbf{T}}$ and $\hat{\mathbf{V}}_0^{\hat{q},\mathbf{U},\hat{z}_r}$ are block-constant matrices with the same block structure determined by $\hat{z}_r$.

Denoting by $\mathbf{0}$ the zero $k \times k$ matrix, we have

$$\begin{aligned}
\|\hat{\mathbf{T}} - \hat{\mathbf{V}}\|_F &\le \big\|\hat{\mathbf{T}} - \hat{\mathbf{V}}_0^{\hat{q},\mathbf{0},\hat{z}_r}\big\|_F \qquad (\text{since } \hat{U} \text{ is a minimizer of } \big\|\hat{\mathbf{T}} - \hat{\mathbf{V}}_0^{\hat{q},\mathbf{U},\hat{z}_r}\big\|_F)\\
&= \|\hat{\mathbf{T}} - \varepsilon_{\hat{q}}\hat{\mathbf{V}}_0\|_F \qquad (\text{since } \|\hat{\mathbf{T}}\|_\infty \le 2r)\\
&= \big\|\hat{\mathbf{T}} - \varepsilon_{\hat{q}}\hat{\mathbf{T}}/\|\hat{\mathbf{T}}\|_F\big\|_F + \varepsilon_{\hat{q}}\big\|\hat{\mathbf{T}}/\|\hat{\mathbf{T}}\|_F - \hat{\mathbf{V}}_0\big\|_F\\
&\le \big(\|\hat{\mathbf{T}}\|_F - \varepsilon_{\hat{q}}\big) + \varepsilon_{\hat{q}}\big\|\hat{\mathbf{T}}/\|\hat{\mathbf{T}}\|_F - \hat{\mathbf{V}}_0\big\|_F\\
&\le \big(\|\hat{\mathbf{T}}\|_F - \varepsilon_{\hat{q}}\big) + \frac{\varepsilon_{\hat{q}}}{4} \qquad (\text{since } \mathcal{C}_{\hat{z}_r} \text{ is a } 1/4\text{-net})\\
&\le \|\hat{\mathbf{T}}\|_F/4.
\end{aligned}$$

Next, since $\hat{\mathbf{V}} = \hat{\mathbf{V}}_0^{\hat{q},\hat{\mathbf{U}},\hat{z}_r}$ minimizes $\|\hat{\mathbf{T}} - \hat{\mathbf{V}}_0^{\hat{q},\mathbf{U},\hat{z}_r}\|_\infty$ over $\mathbf{U}$ we have

$$\|\hat{\mathbf{T}} - \hat{\mathbf{V}}\|_\infty \le \big\|\hat{\mathbf{T}} - \hat{\mathbf{V}}_0^{\hat{q},\mathbf{U}^*,\hat{z}_r}\big\|_\infty,$$

where $\mathbf{U}^*$ is the matrix with elements defined by the relation $\mathbf{U}^*_{ab} = \operatorname{sign}(\hat{\mathbf{T}}_{ij})$ if $i \in \hat{z}_r^{-1}(a)$, $j \in \hat{z}_r^{-1}(b)$ for $(a,b) \in [k] \times [k]$. Thus, all the entries of $\hat{\mathbf{V}}_0^{\hat{q},\mathbf{U}^*,\hat{z}_r}$ in each block are either equal to r or to $-r$ depending on whether the value of $\hat{\mathbf{T}}$ on this block is positive or negative, respectively. Since $\|\hat{\mathbf{T}}\|_\infty \le 2r$, we obtain that $\|\hat{\mathbf{T}} - \hat{\mathbf{V}}_0^{\hat{q},\mathbf{U}^*,\hat{z}_r}\|_\infty \le r$. □

REFERENCES

[1] ALDOUS, D. J. (1985). Exchangeability and related topics. In *École D'été de Probabilités de Saint-Flour, XIII—1983*. *Lecture Notes in Math*. **1117** 1–198. Springer, Berlin. MR0883646

[2] BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

[3] BICKEL, P. J., CHEN, A. and LEVINA, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist*. **39** 2280–2301. MR2906868

[4] BORGS, C., CHAYES, J. and SMITH, A. (2015). Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems* 1369–1377. Nips Foundation. Available at http://books.nips.cc.

[5] CHAN, S. H. and AIROLDI, E. M. (2014). A consistent histogram estimator for exchangeable graph models. In *Proceedings of the* 31*st International Conference on Machine Learning* 208–216. Omnipress, Madison, WI.

[6] CHATTERJEE, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist*. **43** 177–214. MR3285604

[7] CHOI, D. (2015). Co-clustering of nonsmooth graphons. Preprint. Available at arXiv:1507.06352.

[8] CHVÁTAL, V. (1983). *Linear Programming*. W. H. Freeman and Company, New York. MR0717219

[9] DIACONIS, P. and JANSON, S. (2008). Graph limits and exchangeable random graphs. *Rend. Mat. Appl.* (7) **28** 33–61. MR2463439

[10] GAO, C., LU, Y. and ZHOU, H. H. (2014). Rate-optimal graphon estimation. Preprint. Available at arXiv:1410.5837.

[11] LOVÁSZ, L. (2012). *Large Networks and Graph Limits*. *American Mathematical Society Colloquium Publications* **60**. Amer. Math. Soc., Providence, RI. MR3012035

[12] LOVASZ, L. and SZEGEDY, B. (2004). Limits of dense graph sequences. ArXiv Mathematics E-prints.

[13] PENROSE, M. (2003). *Random Geometric Graphs*. *Oxford Studies in Probability* **5**. Oxford Univ. Press, Oxford. MR1986198

[14] TANG, M., SUSSMAN, D. L. and PRIEBE, C. E. (2013). Universally consistent vertex classification for latent positions graphs. *Ann. Statist.* **41** 1406–1430. MR3113816

[15] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. Revised and extended from the 2004 French original, translated by Vladimir Zaiats. MR2724359

[16] WOLFE, P. J. and OLHEDE, S. C. (2013). Nonparametric graphon estimation. Preprint. Available at arXiv:1309.5936.

[17] XU, J., MASSOULIÉ, L. and LELARGE, M. (2014). Edge label inference in generalized stochastic block models: From spectral theory to impossibility results. Preprint. Available at arXiv:1406.6897.

O. KLOPP
UNIVERSITÉ PARIS OUEST
200 AVENUE DE LA RÉPUBLIQUE
92001 NANTERRE
FRANCE

A. B. TSYBAKOV
ENSAE, UMR CNRS 9194
3, AV. PIERRE LAROUSSE
92240 MALAKOFF
FRANCE

N. VERZELEN
INRA, UMR 729, MISTEA
2 PLACE VIALA, BAT. 29
34060 MONTPELLIER CEDEX 1
FRANCE
E-MAIL: nicolas.verzelen@supagro.inra.fr