CREST
CENTER FOR RESEARCH
IN ECONOMICS AND STATISTICS

Série des Documents de Travail

n° 2017-42

Optimal graphon estimation in cut distance

O. KLOPP[1]
N. VERZELEN[2]

[1] ESSEC Business School ; CREST. E-mail : kloppolga@math.cnrs.fr
[2] INRA, Montpellier. E-mail : nicolas.verzelen@inra.fr

Optimal graphon estimation in cut distance

Olga Klopp, ESSEC Business School and CREST

and

Nicolas Verzelen, INRA, Montpellier

December 8, 2017

Abstract

We consider the twin problems of estimating the connection probability matrix of an inhomogeneous random graph and the graphon function of graphon random graph. We establish the minimax estimation rates with respect to the cut metric for classes of block constant matrices and classes of step function graphons. Surprisingly, our results imply that, from the minimax point of view, the raw data, that is the adjacency matrix of the observed graph, is already optimal and more involved procedures cannot improve the convergence rates for this metric.

1 Introduction

Networks is an old problem considered (almost independently) in many different fields: social sciences, computer sciences, statistical physics, biology. Recently there was an increasing interest for model-based statistical analysis of networks. It is customary to modelize networks by inhomogeneous random graphs which are a natural generalization of the classical Erdös-Rényi model $\mathcal{G}(n,p)$ [3]. In inhomogeneous random graphs model, the single parameter p is replaced by a symmetric $n \times n$ matrix of connection probabilities $\mathbf{\Theta}_0 = (\mathbf{\Theta}_{ij})$ with $0 \leq \mathbf{\Theta}_{ij} \leq 1$. Here vertices i and j are connected by an edge with probability $\mathbf{\Theta}_{ij}$ and these events are independent for all pairs (i,j) with $i < j$. We assume that the diagonal entries of $\mathbf{\Theta}_0$ are zero.

A benchmark example is the stochastic block model [13]. In the stochastic block model a random graph on the set of vertices $\{1,\dots,n\}$ is defined as follows: the set of vertices is partitioned into k disjoint communities $\{C_1,\dots,C_k\}$ and a symmetric matrix of inter-communities probabilities connections $Q = (Q_{ij}) \in \mathbb{R}^{k\times k}_{\text{sym}}$ is given. Then, for each pair of distinct vertices $i \in C_a$ and $j \in C_b$, an edge is drawn independently with probability Q_{ab}. The idea is that the probability of an edge between two nodes depends only on blocks to which they belong. The stochastic block model is a useful statistical model that allows to generate networks with community structure.

Usually in applications the number of vertices of the network is growing and its number of blocks can be unbounded. In such situations, a non-parametric model where the number of parameters need not be fixed (or finite), called the graphon model, is more suitable. Graphons are symmetric measurable functions $W : [0,1]^2 \to [0,1]$. In the sequel, the space of graphons is denoted by $\mathcal{W}^+$. Given a graphon $W_0 \in \mathcal{W}^+$, we generate a graph on n vertices in the following way:

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

where $1 \geq \rho_n > 0$ is the scale parameter that can be interpreted as the expected proportion of non-zero edges and $\xi_1, \ldots, \xi_n$ are unobserved (latent) i.i.d. random variables uniformly distributed on $[0, 1]$. As above, the diagonal entries of $\mathbf{\Theta}_0$ are zero. In the case when W_0 is a step-function with k steps, we obtain an analog of the stochastic block model with k groups. The case of dense graph correspond to ρ_n that does not depend on the number of vertices and in the case of sparse graphs we have $\rho_n \to 0$ when $n \to \infty$. This model was recently studied by a number of authors, see, e.g., [1, 2, 9, 10, 22].

In the present paper we consider the problem of estimating the matrix of connection probabilities $\mathbf{\Theta}_0$ and the graphon function $f_0 = \rho_n W_0$ from a single observation of a graph. Suppose that we observe the adjacency matrix $\boldsymbol{A} = (\boldsymbol{A}_{ij})$ of an inhomogeneous random graph with unknown expectation $\mathbf{\Theta}_0 = \mathbb{E}(\boldsymbol{A})$. Here $\boldsymbol{A}_{ij} \in \{0, 1\}$ where $\boldsymbol{A}_{ij} = 1$ is interpreted as the fact that the nodes i and j are connected and $\boldsymbol{A}_{ij} = 0$ otherwise. We set $\boldsymbol{A}_{ii} = 0$ for all $1 \leq i \leq n$. From an observation $\boldsymbol{A}' = (\boldsymbol{A}_{ij},\, 1 \leq j < i \leq n)$ we would like to estimate the matrix of connection probabilities $\mathbf{\Theta}_0$ or the graphon function W_0 when the graph is generated according to the graphon model (1).

The problem of graphon estimation is more challenging than the problem of probability matrix estimation, in particular, because of the identifiability issue: multiple graphons can lead to the same distribution on the space of graphs of size n. The reason for it is the invariance of the topology of a network with respect to any change of labelling of its nodes. Specifically, two graphons U and W define the same probability distribution if and only if they are weakly isomorphic [16]. 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 $U(\phi(x), \phi(y)) = W(\psi(x), \psi(y))$ almost everywhere. This motivates considering the equivalence classes of graphons that are weakly isomorphic. The corresponding quotient space is denoted by $\widetilde{\mathcal{W}}^+$.

The problem of the estimation of $\mathbf{\Theta}_0$ was previously considered in a number of papers. For instance, [8, 9, 21] obtain sub-optimal convergence rates for this problem. Gao et al. [12] 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 with constant $\|\mathbf{\Theta}_0\|_\infty$. More recently, Klopp et al [15] extended their results to a more challenging sparse case when $\|\mathbf{\Theta}_0\|_\infty$ depends on n and goes to zero when $n \to \infty$. The problem of graphon estimation for classes of smooth graphons was considered by Wolfe and Olhede [20] and by Klopp et al [15] where faster rates of convergence are obtained. For classes of step function graphons, Borgs et al [7] obtain suboptimal rate of convergence and Klopp et al [15] obtain minimax optimal rates of convergency for this problem.

All these papers consider the risk measured in Frobenius norm for probability matrix estimation and in l_2-norm for graphon estimation problem. A natural question here is, are there some other norms that are more suitable for this problem? A better candidate here could be the cut distance which plays a central role in the random graph theory.

1.1 Cut metric

One of the fundamental questions in graph theory is the following one: what does it mean for two large graphs to be similar or close? There are different ways of defining the distance of two graphs. For example, the edit distance is defined as normalized Hamming distance of the edge sets. One of the troubles with this notion of distance is that it does not reflect well structural similarities between two graphs. For instance, the edit distance between two independent graphs drawn from the Erdös-Rényi model $\mathcal{G}(n, p)$ with $p = 1/2$ is close to $1/2$ with high probability.

Another notion of distance between two graphs is the sampling distance. Two graphs are close in the sampling distance if random sampling of "small" induced subgraphs does not distinguish them faithfully. This notion of distance is much more suitable than the edit distance but still does not reflect directly structural similarity of graphs. For more details see [16].

Another notion of distance, called *cut distance*, turns out to be more suitable for comparing two graphs. It has multiple advantages: it can be defined for two graphs with different number of nodes and reflects structural similarities between two graphs. In particular, the cut distance of two random graphs with the same edge density will be small. It also happens that the cut distance is equivalent to the sampling distance in a topological sense. This is one of the main results of [16].

The cut norm is also a cornerstone in recent limit graphs theory introduced by Lovász and Szegedy [17] and further developed by, e.g., [5, 6]. One can consider a sequence of graphs of growing size and ask whenever this sequence converge to any meaningful limit. The basic fact is that every graph limit can be represented by a graphon and it turns out that the cut metric is also a key tool for studying the convergence of graph sequences. More generally, several equivalent metrics arise naturally in the study of the limiting behaviour of graph sequences. One possible way of studying a large graph $\mathcal{G}$ is by using homomorphisms. For example, we can count the number of copies of various "small" graphs in $\mathcal{G}$. Probing a large graph with a small graph leads to a notion of convergence of a sequence of graphs [17] and it is equivalent to the notion of convergence in the cut metric.

We start by defining the cut norm of a matrix $\boldsymbol{A} = (a_{ij}) \in \mathbb{R}^{n\times n}$ introduced by Frieze and Kannan [11]:

$$\|\boldsymbol{A}\|_\square = \frac{1}{n^2} \max_{S,T\subset[n]} \Big| \sum_{i\in S, j\in T} a_{ij} \Big|.$$

Note that in this definition one can take $S = T$, $S \cap T = \emptyset$ or $T = \bar{S}$ (see [14]). It is also easy to see that

$$\|\boldsymbol{A}\|_\square \le \frac{1}{n^2}\|\boldsymbol{A}\|_1 \le \frac{1}{n}\|\boldsymbol{A}\|_2$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ are the usual l_1 and l_2-norms of a matrix. For any two graphs $\mathcal{G}$ and $\mathcal{G}'$ the norm of the difference of their adjacency matrices $\boldsymbol{A}_\mathcal{G}$ and $\boldsymbol{A}_{\mathcal{G}'}$ defines a distance between two graphs. For instance $\|\boldsymbol{A}_\mathcal{G} - \boldsymbol{A}_{\mathcal{G}'}\|_1$ is the edit distance and $d_\square(\mathcal{G},\mathcal{G}') = \|\boldsymbol{A}_\mathcal{G} - \boldsymbol{A}_{\mathcal{G}'}\|_\square$ is the cut distance between $\mathcal{G}$ and $\mathcal{G}'$ as labeled graphs. For two unlabeled graphs with the same number of nodes we define their distance by

$$\delta_\square(\mathcal{G},\mathcal{G}') = \min_{\hat{\mathcal{G}},\hat{\mathcal{G}}'} d_\square(\hat{\mathcal{G}},\hat{\mathcal{G}}')$$

where $\hat{\mathcal{G}}$ and $\hat{\mathcal{G}}'$ range over all labelings of $\mathcal{G}$ and $\mathcal{G}'$.

By analogy with the matrix cut norm we can define the cut norm of a graphon $W \in \mathcal{W}$ this definition writes

$$\|W\|_\square = \sup_{S,T\subset[0,1]} \left| \int_{S\times T} W(x,y)\mathrm{d}x\mathrm{d}y \right| \tag{2}$$

Olga: on n'a pas definit $\mathcal{W}$

where the supremum is taken over all measurable subsets S and T. As in the case of matrices one can take $S = T$, $S \cap T = \emptyset$ or $T = [0,1] \setminus S$ (see [14]) and

$$\|W\|_\square \le \|W\|_1 \le \|W\|_2 \le \|W\|_\infty \le 1$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ denote l_1 and l_2-norms of a graphon. In the opposite direction, we have trivially $\|W\|_2 \le \sqrt{\|W\|_1}$. Hence, l_1 and l_2-norms define the same topology on the space of graphons $\mathcal{W}^+$. The cut norm is continuous in this topology and there are easy examples showing that convergence in cut norm does not imply convergence in l_1 or l_2-norms.

$\mathcal{W}$ noyau a valeur dans $[-1,1]$. $\mathcal{W}^+$ noyaux a valeur dans $[0,1]$

Olga: $\mathcal{W}^+$ is also symmetric now...

The space of kernels $\mathcal{W}^+$ corresponds to the labeled graphs. In order to define a distance on the quotient space $\widetilde{\mathcal{W}^+}$ we take the infimum over all measure preserving bijections:

$$\delta_\square(W_1, W_2) = \inf_{\tau\in\mathcal{M}} \|W_1 - W_2^\tau\|_\square \tag{3}$$

where $\mathcal{M}$ is the set of all measure-preserving bijections $\tau : [0,1] \to [0,1]$ and $W^\tau(x,y) = W(\tau(x), \tau(y))$. This construction can be applied to any other norm N that is invariant under measure preserving maps:

$$\delta_N(W_1, W_2) = \inf_{\tau\in\mathcal{M}} \|W_1 - W_2^\tau\|_N. \tag{4}$$

Beyond the cut norm, interesting cases are l_2- and l_1-norms. The corresponding unlabeled distances are δ_2 and δ_1.

The cut-norm is closely related to several other norms that are often used. In particular, it is equivalent to the operator $l_\infty \to L_1$ norm:

$$\|W\|_{\infty\to 1} = \sup_{\|f\|_\infty, \|g\|_\infty \le 1} \left| \int_{S\times T} W(x,y) f(x) g(y) \mathrm{d}x \mathrm{d}y \right| \tag{5}$$

where the supremum is taken over all (real-valued) functions f and g with values in $[-1,1]$. We have the following simple relation (see [14]):

$$\|W\|_\square \le \|W\|_{\infty\to 1} \le 4\|W\|_\square. \tag{6}$$

1.2 Step graphons

We consider the class of step graphons. Let $\mathcal{W}^+[k]$ be the collection of $k-$step graphons, that is the subset of graphons $W \in \mathcal{W}$ such that for some $\boldsymbol{Q} \in \mathbb{R}^{k\times k}_{\mathrm{sym}}$ and some $\phi : [0,1] \to [k]$, $\mathcal{W}^+[k]$ est a valeur dans $\mathbb{R}$ ou $[0,1]$?

$$W(x,y) = \boldsymbol{Q}_{\phi(x),\phi(y)} \quad \text{for all } x, y \in [0,1]\ . \tag{7}$$

1.3 Related literature and our contribution

Olga: peut-etre les gens ont envie de lire ca avant la partie un peu technique avec les defs pour la cut norm ??

The main focus of this paper is to establish the minimax optimal rates of convergence for the problem of estimating the matrix of connection probabilities $\boldsymbol{\Theta}_0$ and the graphon function $f_0 = \rho_n W_0$ for the risk measured in the *cut metric*. The minimax rate of convergence characterizes the fundamental limitation of the estimation accuracy. It also captures the interdependence between the different parameters in the model. In the present paper we prove minimax optimal rates of convergence for classes of block constant matrices (Section 2) and for classes of step function graphons (Section 3). Surprisingly, our results imply that from the minimax point of view there is

no need to look for more involved estimators then the adjacency matrix and associated empirical graphon.

Previously upper bounds for the risk measured in the cut distance for graphon estimation problem was obtained in [4]. In [4] authors deal with a more general case of unbounded graphons and consider three algorithms for producing a block model approximation of the unknown graphon function. In particular, they show an oracle inequality for the risk of the least cut norm estimator measured in the cut distance. For block constant graphons with block size larger then κn where $\kappa \in \left[\frac{\log n}{n}, 1\right]$ this oracle inequality imply the following upper bound on the risk:

$$\sqrt{\rho_n/n} + \rho_n \sqrt{\log n/\kappa n}.$$

In the case of balanced partition, that is $k \sim \kappa^{-1}$, this upper bound coincide with the minimax optimal rate of convergence obtained in Section 3 up to a logarithmic factor and it is suboptimal for unbalanced partitions and weakly sparse graphs. Note that unlike [4], we consider bounded block constant graphons with no restriction on the block size.

An important part of our analysis is to establish lower bounds for the minimax estimation in the cut distance of step function graphons which requires the development of new technical tools. One of the main results of the paper is the Proposition 3 which is of the independent interest. This result considerable improves the Second Sampling Lemma for Graphons for relevant in applications class of step function graphons. In particular, let $W \in \mathcal{W}^+$ be a graphon. We denote by $\mathbb{G}(n, W)$ random sampled graph on n vertices which satisfies the graphon model (1). For $\rho_n = 1$, the Second Sampling Lemma for Graphons states:

Lemma 1 (Lemma 10.16, [16]). *Let $n \geq 1$, and let $W \in \mathcal{W}^+$ be a graphon. Then with probability at least $1 - \exp\{-n/(2\log n)\}$,*

$$\delta_\square\left(\mathbb{G}(n, W), W\right) \leq \frac{22}{\sqrt{\log(n)}}.$$

We prove that, when W is a k-step graphon and $k \leq n$, this can be improved to the following upper bound:

$$\mathbb{E}\left[\delta_\square\left(\mathbb{G}(n, W), W\right)\right] \leq C\sqrt{\frac{k}{n\log(k)}}$$

where C is a numerical constant (independent of n and k).

The paper is organized as follows. In Subsection 1.1 we recall basic definitions and facts related to the cut metric. Subsection 1.4 contains notation used in the paper. The problem of estimating the matrix of connection probabilities is considered in Section 2. We study the problem of graphon function estimation in Section 3. Appendix contains all the proofs where in Appendix A we recall some basic facts and results that are often used in the proofs. In particular in Subsection A.1 we introduce general spaces of kernels.

1.4 Notation

We provide a brief summary of the notation used throughout this paper.

- For a matrix $\boldsymbol{B}$, $\boldsymbol{B}_{ij}$ (or $\boldsymbol{B}_{i,j}$, or $(\boldsymbol{B})_{ij}$) is its (i,j)th entry. We denote by $\boldsymbol{B}_{i,\cdot}$ and by $\boldsymbol{B}_{\cdot,j}$ its ith row and jth column respectively.
- For an integer m, set $[m] = \{1, \ldots, m\}$.

- We denote by $\mathbb{R}^{k\times k}_{\text{sym}}$ the class of all symmetric $k\times k$ matrices with real-valued entries.
- $\langle \boldsymbol{D}, \boldsymbol{B}\rangle = \sum_{i,j} \boldsymbol{D}_{ij}\boldsymbol{B}_{ij}$ is the inner product between matrices $\boldsymbol{D}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$.
- Denote by $\|B\|_2$ and by $\|B\|_1$ the Frobenius norm and the l_1-norm of matrix $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ respectively. Let $\|B\|_\infty$ be the entry-wise supremum norm of matrix $\boldsymbol{B} \in \mathbb{R}^{n\times n}$.
- $\|W\|_{\infty\to 1}$ is $l_\infty \to L_1$ norm of the operator T_W associated to W and it is defined by (5).
- We denote by $\lfloor x \rfloor$ the maximal integer less than $x \geq 0$ and by $\lceil x \rceil$ the smallest integer greater than or equal to x.
- $\mathbb{1}_A(\cdot)$ denotes the indicator function of a set A.
- We denote by $\mathbb{E}$ the expectation with respect to the distribution of $\boldsymbol{A}$ if we consider the network sequence model and the expectation with respect to the joint distribution of $(\boldsymbol{\xi}, \boldsymbol{A})$ 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 $[0, 1]$.
- $\mathcal{W}^+$ is the space of graphons and $\widetilde{\mathcal{W}^+}$ is the quotient space of graphons.
- $\delta_\square(.,.)$ cut distance in the graphons spaces defined by (3).
- $\delta_1(.,.)$ and $\delta_2(.,.)$ defined by(4) are respectively l_1 and l_2 distances on the quotient space of graphons $\widetilde{\mathcal{W}^+}$.
- $\widetilde{f}_{\boldsymbol{\Theta}}$ is the empirical graphon associated to $\boldsymbol{\Theta}$ and it is defined by (9).
- For two positive functions f and g, we write $f \asymp g$ when there exist two positive numerical constants C and C' such $Cg \leq f \leq C'g$.

2 Probability matrix estimation

The following proposition, proved in section B, bounds the cut distance between $\boldsymbol{\Theta}_0$ and the sampled adjacency matrix $\boldsymbol{A}$. The proof is based on Bernstein inequality. Similar results already appear in the literature (see, e.g., [16, Lemma 10.11] in the case of dense graphs).

Proposition 1. *For any probability matrix $\boldsymbol{\Theta}_0$ with $\|\boldsymbol{\Theta}_0\|_\infty \geq 1/n$,*

$$\mathbb{E}\,\|\boldsymbol{A} - \boldsymbol{\Theta}_0\|_\square \leq 12\sqrt{\frac{\|\boldsymbol{\Theta}_0\|_\infty}{n}}.$$

This implies that the adjacency matrix $\boldsymbol{A}$ is $\sqrt{\|\boldsymbol{\Theta}_0\|_\infty/n}$-close in the cut-distance to the probability matrix $\boldsymbol{\Theta}_0$. This bound is valid for all matrices $\boldsymbol{\Theta}_0$. It turns out that no estimator can perform much better than $\boldsymbol{A}$, even on some simple classes of parameters $\boldsymbol{\Theta}_0$.

Let n, k be integers such that $2 \leq k \leq n$ and $\mathcal{T}[k]$ be the set of all probability matrices corresponding to k-class stochastic block models:

$$\mathcal{T}[k] = \{\boldsymbol{\Theta}_0 : \ \exists\, z \in \mathcal{Z}_{n,k},\ \boldsymbol{Q} \in \mathbb{R}^{k\times k}_{\text{sym}} \text{ such that } \boldsymbol{\Theta}_{ij} = \boldsymbol{Q}_{z(i)z(j)},\ i \neq j, \ \text{ and } \boldsymbol{\Theta}_{ii} = 0\ \forall i\}$$

where we denote by $\mathcal{Z}_{n,k}$ the set of all mappings z from $[n]$ to $[k]$. For any $\rho_n \in (0,1]$, consider the set of all probability matrices corresponding to k-class stochastic block models with connection probability uniformly bounded by ρ_n:

$$\mathcal{T}[k,\rho_n] = \{\boldsymbol{\Theta}_0 \in \mathcal{T}[k] : \ \|\boldsymbol{\Theta}_0\|_\infty \leq \rho_n\}\,.$$

In other words $\mathcal{T}[k,\rho_n]$ is made of matrices that, up to a permutation of its rows and its columns, is block constants with k blocks. The following Proposition, proved in section C, gives a lower bound on the minimax risk over the class $\mathcal{T}[2,\rho_n]$ of block-constant matrices with only two blocks:

Proposition 2. *The minimax risk measured in cut norm satisfies*

$$\inf_{\widehat{\boldsymbol{\Theta}}} \sup_{\boldsymbol{\Theta}_0 \in \mathcal{T}[2,\rho_n]} \mathbb{E}_{\boldsymbol{\Theta}_0}\left[\left\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\right\|_\square\right] \geq C \min\left(\sqrt{\frac{\rho_n}{n}}, \rho_n\right)$$

where $\mathbb{E}_{\boldsymbol{\Theta}_0}$ denotes the expectation with respect to the distribution of $\boldsymbol{A}$ when the underlying probability matrix is $\boldsymbol{\Theta}_0$.

Comparing Proposition 2 with Proposition 1 we observe that, the raw data $\boldsymbol{A}$ is minimax optimal for the class $\mathcal{T}[2,\rho_n]$ for all $\rho_n \geq 1/n$. As a consequence, there is no need to look for a more involved estimator. Since for $\rho_n \leq 1/n$ the constant estimator $\widehat{\boldsymbol{\Theta}} = 0$ satisfies $\mathbb{E}_{\boldsymbol{\Theta}_0}[\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_\square] \leq \rho_n$ and using that the collections $\mathcal{T}[k,\rho_n]$ are nested, the two previous propositions imply that the optimal estimation rate for risk measured in cut norm for stochastic block models with $k \geq 2$ is given by

$$\inf_{\widehat{\boldsymbol{\Theta}}} \sup_{\boldsymbol{\Theta}_0 \in \mathcal{T}[k,\rho_n]} \mathbb{E}_{\boldsymbol{\Theta}_0}\left[\left\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\right\|_\square\right] \asymp \min\left(\sqrt{\frac{\rho_n}{n}}, \rho_n\right).$$

Until now, we left aside the specific case of constant matrices $\mathcal{T}[1,\rho_n]$ which correspond to Erdös-Renyi random graphs. It turns out that the situation is quite different for this simple class. For a constant matrix $\boldsymbol{\Theta}_0$, estimating $\boldsymbol{\Theta}_0$ given $\boldsymbol{A}$ amounts to infer the parameter p of a Bernoulli distribution given a sample of size $n(n-1)/2$. From this analogy, we consider the constant matrix $\overline{\boldsymbol{A}} \equiv \sum_{i,j} \boldsymbol{A}_{ij}/(n(n-1))$. Then, it is straightforward to prove that

$$\mathbb{E}_{\boldsymbol{\Theta}_0}\left[\|\boldsymbol{\Theta}_0 - \overline{\boldsymbol{A}}\|_\square\right] \leq \sqrt{\frac{2\rho_n}{n(n-1)}}\,,$$

which is $\sqrt{n}$-faster than what is achieved by the adjacency matrix $\boldsymbol{A}$. Continuing the analogy with the problem of Bernoulli parameter estimation, one may easily get the following minimax lower bound:

$$\inf_{\widehat{\boldsymbol{\Theta}}} \sup_{\boldsymbol{\Theta}_0 \in \mathcal{T}[1,\rho_n]} \mathbb{E}_{\boldsymbol{\Theta}_0}\left[\left\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\right\|_\square\right] \geq C \min\left(\frac{\sqrt{\rho_n}}{n}, \rho_n\right)$$

which assesses that the $\sqrt{\rho_n}/n$-rate achieved by $\overline{\boldsymbol{A}}$ is optimal.

2.1 Comparison with Frobenius norm estimation

Qu'est-ce qu'on veut dmontrer dans cette section

Peut-etre pourrait-on affiner un peu les resultats qui suivent prendre en compte le cas $k \geq \sqrt{n}$..... Y reflechir... Regarder les papiers de Borgs et Chayes (et les commenter)

This is in sharp contrast with what has been established for Frobenius estimation rate in [15]

$$\inf_{\widehat{\mathbf{\Theta}}} \sup_{\mathbf{\Theta}_0 \in \mathcal{T}[k,\rho_n]} \mathbb{E}_{\mathbf{\Theta}_0} \left[\frac{1}{n} \left\| \widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0 \right\|_2 \right] \asymp \min \left(\sqrt{\frac{\rho_n \log(k)}{n}} + \frac{\sqrt{\rho_n} k}{n}, \rho_n \right) . \tag{8}$$

As expected, the cut norm convergence rate is faster than the frobenius norm estimation rate. When the number of blocks remains small ($k \leq \sqrt{n \log(n)}$), the gain is $\log(k)$ factor, whereas for larger k when k is large $k \geq \sqrt{n \log(n)}$ the gain is $\sqrt{n}/k$. Most importantly, the optimal Frobenius convergence rate (8) is only known to be achieved by involved estimators based on the least squares and restricted least squares (see [15]) criteria.

Using $\| \cdot \|_\square \leq \frac{1}{n} \| \cdot \|_2$ we see that the least squares estimator and the restricted least square estimator considered in [15] (both are minimax optimal in the Frobenius norm) are also minimax optimal (up to a logarithmic factor $\log(k)$) in the cut-norm when the number of blocks is small compared to the number of nodes $k \leq \sqrt{n}$:

$$\mathbb{E} \| \widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0 \|_\square \leq \frac{1}{n} \mathbb{E} \| \widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0 \|_2 \leq C \sqrt{\frac{\| \mathbf{\Theta}_0 \|_\infty \log(k)}{n}}$$

where we use $\widehat{\mathbf{\Theta}}$ for the least squares or for the restricted least square estimator.

3 Graphon estimation problem

In this section we estimate the graphon function $W_0(\cdot, \cdot)$ in the sparse graphon model (1). Following [15], we start by associating a graphon to any $n \times n$ probability matrix $\mathbf{\Theta}_0$. Then, we can estimate graphon function $f_0(\cdot, \cdot) = \rho_n W_0(\cdot, \cdot)$ using empirical graphon associated to an estimator of $\mathbf{\Theta}_0$. Given a $n \times n$ matrix $\mathbf{\Theta}$ with entries in $[0, 1]$, we define the *empirical graphon* $\widetilde{f}_{\mathbf{\Theta}}$ as the following piecewise constant function:

$$\widetilde{f}_{\mathbf{\Theta}}(x, y) = \mathbf{\Theta}_{\lceil nx \rceil, \lceil ny \rceil} \tag{9}$$

for all x and y in $(0, 1]$.

For any estimator $\widehat{T}$ of $\mathbf{\Theta}_0$ and any norm N that is invariant under measure preserving maps the triangle inequality implies

$$\mathbb{E} \left[\delta_N(\widetilde{f}_{\widehat{\boldsymbol{T}}}, f_0) \right] \leq \mathbb{E} \left[\| \widehat{\boldsymbol{T}} - \mathbf{\Theta}_0 \|_N \right] + \mathbb{E} \left[\delta_N \left(\widetilde{f}_{\mathbf{\Theta}_0}, f_0 \right) \right] . \tag{10}$$

So, we have two parts in the bound on the risk in (10). The first term is the *estimation error* term $\| \widehat{\boldsymbol{T}} - \mathbf{\Theta}_0 \|_N$. It has been considered in Section 2 for $\widehat{\boldsymbol{T}} = A$ and the cut norm. The second term $\delta_N(\widetilde{f}_{\mathbf{\Theta}_0}, f_0)$ is the *agnostic error*. It measures the δ_N-distance between the true graphon f_0 and its discretized version sampled at the unobserved random design points $\xi_1, \ldots, \xi_n$. The behavior of $\delta_N(\widetilde{f}_{\mathbf{\Theta}_0}, f_0)$ depends on the topology of the considered class of graphons.

The following proposition, proved in Section D, gives the upper bound on the agnostic error, measured in $\delta_\square$-distance, associated to step function graphons:

Proposition 3 (Agnostic error measured in cut distance)**.** *Consider the graphon model* (1)*. For all integers $k \geq 2$, all positive integers n, all $W_0 \in \mathcal{W}^+[k]$ and $\rho_n > 0$, we have*

$$\mathbb{E} \left[\delta_\square(\widetilde{f}_{\mathbf{\Theta}_0}, f_0) \right] \leq C \rho_n \begin{cases} \sqrt{\dfrac{k}{n \log(k)}} & \text{if } k \leq n, \\ \sqrt{\dfrac{1}{\log(n)}} & \text{if } k > n. \end{cases}$$

Note that the case $k > n$ is a consequence of Lemma 1 [16], so that we effectively only have to consider the case $k \leq n$. The proof combines two ideas. First, we build W and $\widehat{W}$ as the representatives of W_0 and $\widetilde{f}_{\boldsymbol{\Theta}_0}$ in the quotient space $\widetilde{\mathcal{W}}^+$ such that W and $\widehat{W}$ match everywhere except on a set of Lebesgue measure of order at most $\sqrt{k/n}$. This allows us to control the risk to the rate $\sqrt{k/n}$. In order to recover the correct logarithmic factor $\sqrt{\log(k)}$, we rely on weak Szemeredy Lemma. Here, the key idea is to build a cut-norm approximation of a distorted transformation of W where the weights of the group have been modified to take into account the geometry of the sampling error.

As an immediate consequence of (10), Propositions 1 and 3, we get the following upper bound on the risk of the empirical graphon $\widetilde{f}_{\boldsymbol{A}}$. For any $k \geq 2$, it holds that

$$\sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0}\left[\delta_\square\left(\widetilde{f}_{\boldsymbol{A}}, f_0\right)\right] \leq C\left[\rho_n\left(\sqrt{\frac{k}{n\log(k)}} \wedge \frac{1}{\sqrt{\log(n)}}\right) + \sqrt{\frac{\rho_n}{n}}\right], \tag{11}$$

where C is an absolute constant. Here, $\mathbb{E}_{W_0}$ denotes the expectation with respect to the distribution of observations $\boldsymbol{A} = (\boldsymbol{A}_{ij}, 1 \leq j < i \leq n)$ when the underlying sparse graphon is $f_0 = \rho_n W_0$. The following Proposition gives matching lower bound for $2 \leq k \leq n$:

Proposition 4. *There exists a universal constant $C > 0$ such that for any sequence $\rho_n > 0$ and any positive integer $2 \leq k \leq n$,*

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0}\left[\delta_\square\left(\widehat{f}, f_0\right)\right] \geq C\left\{\left[\rho_n\sqrt{\frac{k}{n\log(k)}} + \sqrt{\frac{\rho_n}{n}}\right] \wedge \rho_n\right\}, \tag{12}$$

where $\mathbb{E}_{W_0}$ denotes the expectation with respect to the distribution of $\boldsymbol{A}$ when the underlying sparse graphon is $f_0 = \rho_n W_0$ and $\inf_{\widehat{f}}$ is the infimum over all estimators.

Since the collections $\mathcal{W}^+[k]$ are nested, it follows that for all $k \geq n$, one has

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0}\left[\delta_\square\left(\widehat{f}, f_0\right)\right] \geq C\left\{\left[\rho_n\sqrt{\frac{1}{\log(n)}} + \sqrt{\frac{\rho_n}{n}}\right] \wedge \rho_n\right\}.$$

In view of (11) and (12), we observe that, as long as, $\rho_n \geq 1/n$, the empirical graphon $\widetilde{f}_{\boldsymbol{A}}$ is minimax optimal over all classes $\mathcal{W}^+[k]$, $k \geq 2$. For sparser graphs ($\rho_n \leq 1/n$), the trivial estimator $\widehat{f} \equiv 0$ achieves the optimal rate ρ_n.

Note that there are two distinct regimes in the minimax convergence rate. When $\rho_n \geq \log(k)/k$ (weakly sparse graphs or large number of groups), the agnostic error dominates and the minimax risk is of order $\rho_n\sqrt{k/(n\log(k))}$. For moderately sparse graphs or equivalently a small number of steps ($n^{-1} \leq \rho_n \leq \log(k)/k$), the error arising from the probability matrix $\boldsymbol{\Theta}_0$ estimation dominates and the minimax risk is of order $\sqrt{\rho_n/n}$.

ajouter un resultat d'estimation avec du biais?: graphon W_0 proche de $\mathcal{W}[k]$ mais pas dans $\mathcal{W}[k]$...

As in the previous section, we left aside the specific case of constant graphons $\mathcal{W}^+[1]$. Note that for a graphon $W_0 \in \mathcal{W}^+[1]$ the agnostic error is always zero and the loss comes from probability matrix estimation problem. Following the arguments of the previous section, we derive that the graphon $\widetilde{f}_{\boldsymbol{A}}$ converges to $\rho_n W_0$ at the rate $\sqrt{\rho_n}/n$ which is optimal as soon as $\rho_n \geq 1/n^2$.

3.1 Comparison with l_1 and l_2 distance

Est-ce qu'on garde toute la discussion qui suit. A relire et a modifier eventuellement Corresponding results in δ_2-distance were obtained in [15], see Proposition 3.2. Actually the proof of Proposition 3.2 in [15] can be easily modified to get also an upper bound on the agnostic error measured in l_1-distance:

Proposition 5 (Agnostic error measured in l_1-distance)**.** *Consider the garphon model. For all integer $k \leq n$, $W_0 \in \mathcal{W}^+[k]$ and $\rho_n > 0$, we have*

$$\mathbb{E}\left[\delta_1(\widetilde{f}_{\boldsymbol{\Theta}_0}, f_0)\right] \leq C\rho_n\sqrt{\frac{k}{n}}.$$

The proof of Proposition 5 follows the lines of the proof of Proposition 3.2 in [15] with δ^2 replaced by δ_1.

The following proposition proved in in Section F gives a lower bound on the minimax rate of convergence measured in δ_1-distance:

Proposition 6. *There exists a universal constants $C_1, C_2 > 0$ such that for any sequence $\rho_n > 0$ and any positive integer $2 \leq k \leq n$,*

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0}\left[\delta_1\left(\widehat{f}, f\right)\right] \geq C\left\{\left[\rho_n\sqrt{\frac{k}{n}} + \sqrt{\rho_n}\frac{k}{n} + \sqrt{\frac{\rho_n}{n}}\right] \wedge \rho_n\right\} \tag{13}$$

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

Using $\frac{1}{n^2}\|A\|_1 \leq \frac{1}{n}\|A\|_2$ and the upper bound on the estimation error of restricted least squares estimator $\widehat{\boldsymbol{\Theta}}^r$ introduced in [15] we get that

$$\mathbb{E}\left[\frac{1}{n^2}\|\widehat{\boldsymbol{\Theta}}^r - \boldsymbol{\Theta}_0\|_1\right] \leq \mathbb{E}\left[\frac{1}{n}\|\widehat{\boldsymbol{\Theta}}^r - \boldsymbol{\Theta}_0\|_2^2\right] \leq C\sqrt{\|\boldsymbol{\Theta}_0\|_\infty}\left(\sqrt{\frac{\log(k)}{n}} + \frac{k}{n}\right). \tag{14}$$

Then, Proposition 3, (10) and (14) imply the following upper bound on the minimax risk measured in δ_1-distance:

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0}\left[\delta_1\left(\widehat{f}, f\right)\right] \leq \left\{\left[\rho_n\sqrt{\frac{k}{n}} + \sqrt{\rho_n}\frac{k}{n} + \sqrt{\frac{\rho_n \log(k)}{n}}\right] \wedge \rho_n\right\}. \tag{15}$$

Thus, Proposition 6 implies that this upper bound is minimax optimal (up to a logarithmic factor in k in one of the regimes). We have three zones in (15). The first one corresponds to the case of weakly sparse graphs with $\rho_n \geq \left(\frac{1}{k} \vee \frac{k}{n}\right)$. In this case, the agnostic error dominates and the optimal bound is $\rho_n\sqrt{\frac{k}{n}}$. The second case corresponds to the case of moderately sparse graphs with $\frac{1}{n} \vee \left(\frac{k}{n}\right)^2 \leq \rho_n \leq \left(\frac{1}{k} \vee \frac{k}{n}\right)$. Here the probability matrix estimation error dominates and the bound is $\sqrt{\frac{\rho_n}{n}} + \sqrt{\rho_n}\frac{k}{n}$ which is optimal up to a $\log(k)$ factor. In the case of highly sparse graphs with $\rho_n \leq \frac{1}{n} \vee \left(\frac{k}{n}\right)^2$ the optimal bound is ρ_n wich corresponds to the risk of the null estimator $\tilde{f} \equiv 0$.

Comparing (15) with the minimax optimal rate of convergence for risk measured in δ_2-distance obtained in [15] (up to a logarithmic factor):

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0} \left[\delta_2 \left(\widehat{f}, f \right) \right] \asymp \min \left(\frac{\sqrt{\rho_n} k}{n} + \sqrt{\frac{\rho_n}{n}} + \rho_n \left(\frac{k}{n} \right)^{1/4}, \rho_n \right) \tag{16}$$

we see that for weakly sparse graphs with $\rho_n \geq \frac{1}{\sqrt{kn}} \vee \left(\frac{k}{n}\right)^{3/2}$ the rate of convergence in δ_1-distance is faster. For graphs with $\rho_n \leq \frac{1}{\sqrt{kn}} \vee \left(\frac{k}{n}\right)^{3/2}$ we get the same rate of convergence in δ_1 and δ_2-distances.

Comparing the minimax optimal rate of convergence in δ_1-distance (15) and the minimax optimal rate of convergence in cut distance (12) we get the same rate of convergence (up to a $\log(k)$) factor except for the case of large $k \gtrsim \sqrt{n}$ and $\rho_n \leq k/n$ where the gain in the cut distance is $k/\sqrt{n}$ factor. Now we compare the minimax optimal rate of convergence in δ_2-distance (16) and the minimax optimal rate of convergence in cut distance (12). First consider the case of small number of blocks $k \leq \sqrt{n}$: for weakly sparse graphs with $\rho_n \geq 1/\sqrt{nk}$ the rate of convergence in $\delta_\square$-distance is faster with the gain of $(k/n)^{1/4}$ factor; for $\rho_n \leq 1/\sqrt{nk}$ we get the same rates of convergence in δ_2- and $\delta_\square$-distances. For $k \geq \sqrt{n}$ the rate of convergence in $\delta_\square$-distance is faster.

A Proof methods

In this section for readers convenience we summarize some basic facts and fundamental results that we use in the proofs.

A.1 Non-symmetric kernels

At some point, we will need to work with non-symmetric kernels and with kernel defined on general measurable subsets of $\mathbb{R}$. In this section we define the corresponding spaces. Let $\mathcal{X}$ and $\mathcal{Y}$ denote two bounded measurable subsets of $\mathbb{R}$. Then, $\mathcal{W}_{\mathcal{X},\mathcal{Y}}$ refers to the collection of bounded measurable functions $W : \mathcal{X} \times \mathcal{Y} \to [-1, 1]$. We will denote by $\mathcal{W}^+_{\mathcal{X},\mathcal{Y}}$ the collection of bounded measurable and non-negative functions $W : \mathcal{X} \times \mathcal{Y} \to [0, 1]$. Let $\mathcal{W}_{\mathcal{X},\mathcal{Y}}[k]$ be the collection of $k-$step graphons, that is the subset of kernels $W \in \mathcal{W}_{\mathcal{X},\mathcal{Y}}$ such that for some $\boldsymbol{Q} \in \mathbb{R}^{k\times k}$ and some $\phi_1 : \mathcal{X} \to [k]$, $\phi_2 : \mathcal{X} \to [k]$,

$$W(x, y) = \boldsymbol{Q}_{\phi_1(x),\phi_2(y)} \quad \text{for all } (x, y) \in \mathcal{X} \times \mathcal{Y} \, . \tag{17}$$

The cut norm also readily extends to kernel $W \in \mathcal{W}_{\mathcal{X},\mathcal{Y}}$ in the following way:

$$\|W\|_\square := \sup_{X\subset\mathcal{X},\ Y\subset\mathcal{Y}} \left| \int_{X\times Y} W(x, y)dxdy \right| \tag{18}$$

where the supremum is taken over all measurable subsets X and Y.

A.2 Concentration inequalities

In the proofs we repeatedly use Bernstein's inequality. We state it here for the readers' convenience. Let $X_1, \ldots, X_N$ be independent zero-mean random variables. Suppose that $|X_i| \leq M$ almost surely, for all i. Then, for any $t > 0$,

$$\mathbb{P}\left\{ \sum_{i=1}^N X_i \geq t \right\} \leq \exp\Big[- \frac{t^2}{2\sum_i \mathbb{E}[X_i^2] + 2Mt/3} \Big] \, . \tag{19}$$

We shall also rely on the bounded difference inequality (also called Mc Diarmid inequality).

Lemma 2 (Bounded difference inequality)**.** *Let $X_1, \ldots, X_n$ denote n independent real random variables. Assume that $g : \mathbb{R}^n \to \mathbb{R}$ is a measurable function satisfying, for some positive constants $(c_i)_{1 \leq i \leq n}$, the bounded difference condition*

$$|g(x_1, \ldots, x_i, \ldots, x_n) - g(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i \ ,$$

for all $x = (x_1, \ldots, x_i, \ldots, x_n) \in \mathbb{R}^n$, $x' = (x_1, \ldots, x_i', \ldots, x_n) \in \mathbb{R}^n$ and all $i \in [n]$. Then, the random variable $Z = g(X_1, \ldots, X_n)$ satisfies

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + t] \leq \exp\left[-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right] \ ,$$

for all $t > 0$.

A.3 Fano's lemma

In the sequel, $\mathcal{KL}(.,.)$ denotes the Kullback-Leibler divergence between two distributions. In this manuscript, the proofs of the minimax lower bounds proofs all rely on Fano's method. The following version of Fano's lemma is borrowed from [19].

Lemma 3. *[19, Theorem 2.7] Consider a parametric model $\mathbb{P}_\theta$, with $\theta \in \Theta$ and a metric $d(.,.)$ on Θ. Assume that Θ contains elements $\theta_1, \ldots, \theta_M$, $M \geq 3$, such that for all $j, k \in [M]$ with $j \neq k$*

(i) $d(\theta_j, \theta_k) \geq s > 0$,

(ii) $\mathcal{KL}(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_k}) \leq \log(M)/32$.

Then, we have

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[d(\widehat{\theta}, \theta)\right] \geq Cs \ ,$$

where the constant $C > 0$ is numeric.

A.4 Khintchine's inequality

Next, we state a particular case of Khintchine's inequality that turns out to be useful for bounding the cut norm of step kernels in terms of their l_1 norm.

Lemma 4. *[18] Let $\epsilon_1, \ldots, \epsilon_p$ be i.i.d. Rademacher random variables and let $x_1, \ldots, x_p$ be some real numbers. Then,*

$$\mathbb{E}\left[\left|\sum_{i=1}^p \epsilon_i x_i\right|\right] \geq \frac{1}{\sqrt{2}} \left[\sum_{i=1}^p x_i^2\right]^{1/2} \ . \tag{20}$$

We use this result to prove the following lower bound on the cut norm of step kernels:

Lemma 5. *Let $U : \mathcal{X} \times \mathcal{Y} \mapsto [-1, 1]$ denote a measurable $q_1 \times q_2$-step function. Then,*

$$\|U\|_\square \geq \frac{1}{4\sqrt{2q_2}} \|U\|_1 \ . \tag{21}$$

Proof of Lemma 5. There exists partitions of $\mathcal{X} = \mathcal{X}_1 \cup \dots \mathcal{X}_{q_1}$ and $\mathcal{Y} = \mathcal{Y}_1 \cup \dots \mathcal{Y}_{q_2}$ such that, for any fixed $y \in \mathcal{Y}$, $U(x,y)$ is constant over $\mathcal{X}_i$ for all $i \in [q_1]$ and, for any fixed $y \in \mathcal{X}$, $U(x,y)$ is constant over $\mathcal{Y}_i$ for all $i \in [q_2]$. For any $a \in [q_1]$ (resp. $b \in [q_2]$), denote x_a (resp. y_b) any element of $\mathcal{X}_a$ (resp. $\mathcal{Y}_b$). By definition of $\|U\|_\square$,

$$\begin{aligned}
\|U\|_\square &= \sup_{S\subset\mathcal{X},T\subset\mathcal{Y}} \Big| \int_{S,T} U(x,y)dxdy \Big| \\
&= \sup_{S\subset\mathcal{X},T\subset\mathcal{Y}} \sum_{a=1}^{q_1}\sum_{b=1}^{q_2} \Big| \lambda(S\cap\mathcal{X}_a)\lambda(T\cap\mathcal{Y}_b) U(x_a,y_b) \Big| \\
&= \sup_{\epsilon\in[0,1]^{q_1}} \sup_{\epsilon'\in[0,1]^{q_2}} \Big| \sum_{a=1}^{q_1}\sum_{b=1}^{q_2} \epsilon_a \lambda(\mathcal{X}_a)\epsilon'_b\lambda(\mathcal{Y}_b) U(x_a,y_b) \Big| \,,
\end{aligned}$$

where we used in the last line that the value of the sum only depends on S and T through the quantities $\lambda(S\cap\mathcal{X}_a)$ and $\lambda(T\cap\mathcal{Y}_b)$. Since the maximum of a linear function on a convex set is achieved at an extremal point, it follows that

$$\begin{aligned}
\|U\|_\square &= \sup_{\epsilon\in\{0,1\}^{q_1},\ \epsilon'\in\{0,1\}^{q_2}} \Big| \sum_{a=1}^{q_1}\sum_{b=1}^{q_2} \epsilon_a \lambda(\mathcal{X}_a)\epsilon'_b\lambda(\mathcal{Y}_b) U(x_a,y_b) \Big| \\
&\geq \frac{1}{4} \sup_{\epsilon\in\{-1,1\}^{q_1},\epsilon'\in\{-1,1\}^{q_2}} \Big| \sum_{a\in[q_1],b\in[q_2]} \epsilon_a\epsilon'_b\lambda(\mathcal{X}_a)\lambda(\mathcal{Y}_b) U[x_a,y_b] \Big| \\
&\geq \frac{1}{4} \sup_{\epsilon'\in\{-1,1\}^{q_2}} \sum_{a\in[q_1]} \lambda(\mathcal{X}_a) \left| \sum_{b\in[q_2]} \epsilon'_b\lambda(\mathcal{Y}_b) U[x_a,y_b] \right|
\end{aligned}$$

where we use (6) and take $\epsilon_a = \operatorname{sign}\sum_{b\in[q_2]} \epsilon'_b\lambda(\mathcal{Y}_b)U[x_a,y_b]$. Let $v = (v_1,\dots,v_{q_2})$ denote iid Rademacher random variables and let $\mathbb{E}_v[.]$ denotes the expectation with respect to v. Now, Khintchine's inequality (20) and Cauchy-Schwarz inequality imply

$$\begin{aligned}
\sup_{\epsilon'\in\{-1,1\}^{q_2}} \sum_{a\in[q_1]} \lambda(\mathcal{X}_a) \Big| \sum_{b\in[q_2]} \epsilon_b\lambda(\mathcal{Y}_b)U[x_a,y_b] \Big| &\geq \mathbb{E}_v \Big[\sum_{a\in[q_1]} \lambda(\mathcal{X}_a) \Big| \sum_{b\in[q_2]} v_b\lambda(\mathcal{Y}_b)U[x_a,y_b] \Big| \Big] \\
&\geq \frac{1}{\sqrt{2}} \sum_{a\in[q_1]} \lambda(\mathcal{X}_a) \left(\sum_{b\in[q_2]} \lambda^2(\mathcal{Y}_b)U^2[x_a,y_b] \right)^{1/2} \\
&\geq \frac{1}{\sqrt{2q_1}} \sum_{a\in[q_1]}\sum_{b\in[q_2]} \lambda(\mathcal{X}_a)\lambda(\mathcal{Y}_b) \left| U[x_a,y_b] \right| \\
&= \frac{1}{\sqrt{2q_2}} \|U\|_1 \,.
\end{aligned}$$

□

B Proof of Proposition 1

Since the diagonals of $\boldsymbol{A}$ and $\boldsymbol{\Theta}$ are both zero, it suffices to control the supremum over disjoints subsets S and T (see, e.g., [5])

$$\|\boldsymbol{A} - \boldsymbol{\Theta}_0\|_\square \leq \frac{4}{n^2} \max_{S\cap T=\emptyset} \Big| \sum_{i\in S, j\in T} (\boldsymbol{A}_{ij} - \boldsymbol{\Theta}_{ij}) \Big| \,.$$

Let S and T be any two disjoint subsets of $[n]$. Using Bernstein's inequality (19) we have that

$$\mathbb{P}\left\{\left|\sum_{i\in S,j\in T} \boldsymbol{A}_{ij} - \boldsymbol{\Theta}_{ij}\right| \geq 3\sqrt{\|\boldsymbol{\Theta}_0\|_\infty} n^{3/2}\right\} \leq 2\exp\left(\frac{-\frac{9}{2}\|\boldsymbol{\Theta}_0\|_\infty n^3}{\|\boldsymbol{\Theta}_0\|_\infty |S||T| + \sqrt{\frac{\|\boldsymbol{\Theta}_0\|_\infty}{n}} n^2}\right)$$
$$\leq 2\exp\left(-\frac{9}{4}n\right)$$

where we use $\|\boldsymbol{\Theta}_0\|_\infty \geq 1/n$. Now, using that the number of disjoint pairs (S,T) is 3^n and the union bound, we get that the probability that $|\sum_{i\in S,j\in T} \boldsymbol{A}_{ij} - \boldsymbol{\Theta}_{ij}|$ exceeds $3\sqrt{\|\boldsymbol{\Theta}_0\|_\infty} n^{3/2}$ for some (S,T) is bounded by $2\exp(-n)$. Hence, we have

$$\|A - \boldsymbol{\Theta}_0\|_\square \leq 4 \sup_{S\cap T=\emptyset} \frac{1}{n^2}\left|\sum_{(i,j)\in S\times T} \boldsymbol{A}_{ij} - \boldsymbol{\Theta}_{ij}\right| \leq 12\sqrt{\frac{\|\boldsymbol{\Theta}_0\|_\infty}{n}}$$

with probability $1 - 2e^{-n}$. Now bounding the distance by 1 in the exceptional case we get the statement of Proposition 1.

C Proof of Proposition 2

This proof is based on Fano's method. To apply Fano's Lemma (Lemma 3), it is enough to check that there exists a finite subset Ω of $\mathcal{T}[2,\rho_n]$ such that for any two distinct $\boldsymbol{\Theta}, \boldsymbol{\Theta}'$ in Ω we have

(a) $\|\boldsymbol{\Theta} - \boldsymbol{\Theta}'\|_\square \geq C\sqrt{\rho_n}\left(\frac{1}{\sqrt{n}} \wedge \sqrt{\rho_n}\right)$ and

(b) $\mathcal{KL}(\mathbb{P}_{\boldsymbol{\Theta}}, \mathbb{P}_{\boldsymbol{\Theta}'}) \leq \log(|\Omega|)/32$

for some constants $C > 0$. To prove it, we fix some $\rho_n/4 > \epsilon > 0$. For any $u \in \{-1,1\}^n$, define $\boldsymbol{\Theta}_u$ by $(\boldsymbol{\Theta}_u)_{i,j} = \rho_n/2 + u(i)u(j)\epsilon$ where $u = (u(1),\ldots,u(n))$. Obviously, we have

$$\{\boldsymbol{\Theta}_u \ : \ u \in \{-1,1\}^n\} \subset \mathcal{T}[2,\rho_n].$$

Denote $V_u := \{i \in [n] \ : \ u(i) = 1\}$ and $\bar{V}_u$ its complementary. Then, if we take $S := V_u \setminus V_v$ and $T := V_v \cap V_u$, we obtain

$$\Big|\sum_{i\in S,j\in T} (\boldsymbol{\Theta}_u - \boldsymbol{\Theta}_v)_{ij}\Big| = 2\epsilon |V_u \setminus V_v||V_v \cap V_u\ |.$$

By symmetry, we derive that

$$n^2\|\boldsymbol{\Theta}_u - \boldsymbol{\Theta}_v\|_\square \geq 2\epsilon \max\{|V_u \setminus V_v|, |V_v \setminus V_u|\} \max\{|\bar{V}_u \cap \bar{V}_v|, |V_v \cap V_u|\}$$
$$\geq \frac{\epsilon}{2}|V_u \triangle V_v|(n - |V_u \triangle V_v|)\ ,$$

where $A\triangle B$ is the symmetric difference of A and B. We can use Varshamov-Gilbert bound (see, e.g., [19, Lemma 2.9]) to pick $u_1,\ldots,u_N$ satisfying

$$\frac{n}{4} \leq |V_{u_i} \triangle V_{u_j}| \leq \frac{3n}{4} \quad \text{for} \quad i \neq j \in [N]$$

with $N \geq \exp(c_1 n)$ for some $c_1 > 0$. Let $\Omega = \{\mathbf{\Theta}_{u_i} \ : \ i = 1, \ldots, N\}$, hence we have $\log|\Omega| \geq c_1 n$ and

$$\|\mathbf{\Theta}_{u_i} - \mathbf{\Theta}_{u_j}\|_\square \geq \epsilon/14$$

which proves (a) when one takes ϵ as defined in (22) below.

To prove (b) we use the definition of Kullback-Leibler divergence $\mathcal{KL}(\mathbb{P}_{\mathbf{\Theta}_u}, \mathbb{P}_{\mathbf{\Theta}_v})$ and $\log x \leq x-1$ for $x > 0$ to get

$$\begin{aligned}\mathcal{KL}(\mathbb{P}_{\mathbf{\Theta}_u}, \mathbb{P}_{\mathbf{\Theta}_v}) &= \sum_{ij} (\mathbf{\Theta}_u)_{i,j} \log\left(\frac{(\mathbf{\Theta}_u)_{i,j}}{(\mathbf{\Theta}_v)_{i,j}}\right) + (1 - (\mathbf{\Theta}_u)_{i,j}) \log\left(\frac{(1-\mathbf{\Theta}_u)_{i,j}}{1-(\mathbf{\Theta}_v)_{i,j}}\right) \\ &\leq \sum_{ij} \frac{((\mathbf{\Theta}_u)_{i,j} - (\mathbf{\Theta}_v)_{i,j})^2}{(\mathbf{\Theta}_v)_{i,j}\,(1-(\mathbf{\Theta}_v)_{i,j})}.\end{aligned}$$

Now, $(\mathbf{\Theta}_v)_{i,j} \geq \rho_n/4$ and $\rho_n \leq 1$ imply

$$\mathcal{KL}(\mathbb{P}_{\mathbf{\Theta}_{u_i}}, \mathbb{P}_{\mathbf{\Theta}_{u_j}}) \leq \frac{16}{3\rho_n} \sum_{ij} ((\mathbf{\Theta}_u)_{i,j} - (\mathbf{\Theta}_v)_{i,j})^2 \leq \frac{16 n^2 \epsilon^2}{3\rho_n}.$$

For a choice

$$\epsilon = c_2 \sqrt{\rho_n} \left(\frac{1}{\sqrt{n}} \wedge \sqrt{\rho_n}\right) \tag{22}$$

with a suitable constant $c_2 > 0$, we have that

$$\mathcal{KL}(\mathbb{P}_{\mathbf{\Theta}_{u_i}}, \mathbb{P}_{\mathbf{\Theta}_{u_j}}) \leq \log|\Omega|/32$$

which proves (b).

D Proof of Proposition 3

Note that both $f_0 = \rho_n W_0$ and $\tilde{f}_{\mathbf{\Theta}_0}$ are proportional to ρ_n, so without loss of generality we can assume that $\rho_n = 1$. For $k \geq n/2$, the result is a straightforward consequence of the second Sampling Lemma for Graphons of [16] stated in Lemma 1. Given any graphon $W_0 \in \mathcal{W}^+[k]$, one can always divide some of the steps into smaller in such a way that W_0 is a $2k$-step graphon whose weights are all less or equal to $1/k$. Thus, we only need to prove the results for all graphons $W_0 \in \mathcal{W}^+[k]$ with $32 \leq k \leq n$ and such that its weights are all smaller or equal to $2/k$.

Let $\mathbf{\Theta}'_0$ be the matrix with entries $(\mathbf{\Theta}'_0)_{ij} = W(\xi_i, \xi_j)$ 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 have

$$\mathbb{E}\left[\delta_\square\left(\widetilde{f}_{\mathbf{\Theta}_0}, W_0\right)\right] \leq \mathbb{E}\left[\delta_\square\left(\widetilde{f}_{\mathbf{\Theta}_0}, \widetilde{f}_{\mathbf{\Theta}'_0}\right)\right] + \mathbb{E}\left[\delta_\square\left(\widetilde{f}_{\mathbf{\Theta}'_0}, W_0\right)\right]. \tag{23}$$

As 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 $1/n$. Since $\|W_0\|_\infty \leq 1$, $\mathbb{E}[\delta_\square(\widetilde{f}_{\mathbf{\Theta}_0}, \widetilde{f}_{\mathbf{\Theta}'_0})] \leq 1/n$. Thus, we only need to prove that

$$\mathbb{E}[\delta_\square(\widetilde{f}_{\mathbf{\Theta}'_0}, W_0)] \leq C\sqrt{\frac{k}{n\log(k)}}\,. \tag{24}$$

We first need to build two suitable representations of W_0 and $\widetilde{f}_{\mathbf{\Theta}'_0}$ in the quotient space $\widetilde{\mathcal{W}}^+$.

Step 1: Construction of a suitable representation W of W_0 in $\widetilde{\mathcal{W}}^+$. In the sequel, we denote $q_1 := \lfloor \sqrt{k} \rfloor$. Here, we want to choose W in such a way that a distortion of W is well approximated in cut norm by a q_1-step kernel. We use the following lemma which is based on a variation of Szemeredy lemma. Let $\boldsymbol{Q}_0 \in \mathbb{R}^{k\times k}_{\mathrm{sym}}$ and $\phi_0 : [0,1] \to [k]$ be associated to W_0 as in definition (7).

Lemma 6. *There exist a permutation π of $[k]$ and a partition $\mathcal{P} = (P_1, \dots, P_{q_1})$ of $[k]$ made of successive intervals such that the following holds. Let $\mathbf{Q}$ be the matrix obtained from $\mathbf{Q}_0$ by jointly applying the permutation π to its rows and its columns. Denote by $\phi = \pi \circ \phi_0$, and for $a = 1, \dots, k$, $\lambda_a := \lambda(\phi^{-1}(a))$. There are two matrices $\mathbf{Q}^{(ap)}$ and $\mathbf{Q}^{(ap,+)} \in [0,1]^{k\times k}$ that are q_1-block-constant according to the partition $\mathcal{P}$ and that satisfy*

$$\sup_{\epsilon\in\{0,1\}^k,\ \epsilon'\in\{0,1\}^k} \Big| \sum_{a,b=1}^{k} \epsilon_a\epsilon'_b\lambda_b\sqrt{\lambda_a}\big(\boldsymbol{Q}_{ab} - \boldsymbol{Q}^{(ap)}_{ab}\big)\Big| \ \le\ C\sqrt{\frac{k}{\log(k)}}\ , \tag{25}$$

$$\sup_{\epsilon\in\{0,1\}^k,\ \epsilon'\in\{0,1\}^k} \Big| \sum_{a,b=1}^{k} \epsilon_a\epsilon'_b\sqrt{\lambda_b}\sqrt{\lambda_a}\big(\boldsymbol{Q}_{ab} - \boldsymbol{Q}^{(ap,+)}_{ab}\big)\Big| \ \le\ C\frac{k}{\sqrt{\log(k)}}\ . \tag{26}$$

Invoking Lemma 6, we consider the graphons

$$W(x,y) := \boldsymbol{Q}_{\phi(x)\phi(y)}\ , \quad W_1(x,y) := \boldsymbol{Q}^{(ap)}_{\phi(x)\phi(y)}, \quad W_1^+(x,y) := \boldsymbol{Q}^{(ap,+)}_{\phi(x)\phi(y)}\ . \tag{27}$$

Obviously, W is weakly isomorphic to W_0.

Step 2: Construction of a suitable representation $\widehat{W}$ of $\widetilde{f}_{\boldsymbol{\Theta}'_0}$ in the quotient space $\widetilde{\mathcal{W}}^+$. Recall that $\xi_1, \dots, \xi_n$ are the i.i.d. uniformly distributed random variables in the graphon model (1) and that ϕ is defined in the previous step. For $a = 1, \dots, k$, 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 the group a corresponding to a finer partition of $[0,1]$ given by ϕ. For $a = 1, \dots, q_1$, let

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

be the (unobserved) empirical frequency of the group a corresponding to a coarser partition P of $[0,1]$ given by $\mathcal{P}\circ\phi$.

The relations $\sum_{a=1}^k \lambda_a = \sum_{a=1}^k \hat{\lambda}_a = 1$ imply

$$\sum_{a:\lambda_a>\widehat{\lambda}_a} (\lambda_a - \widehat{\lambda}_a) = \sum_{a:\widehat{\lambda}_a>\lambda_a} (\hat{\lambda}_a - \lambda_a) \quad \text{and} \quad \sum_{a:\omega_a>\widehat{\omega}_a} (\omega_a - \widehat{\omega}_a) = \sum_{a:\widehat{\omega}_a>\omega_a} (\hat{\omega}_a - \omega_a). \tag{28}$$

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

(i) For all $a \in [k]$, $\lambda(\{x,\ \psi(x) = \phi(x) = a\}) = \widehat{\lambda}_a \wedge \lambda_a$,

(ii) for all $a \in [q_1]$, $\lambda\Big[\{x\ , \psi(x) \in P_a \text{ and } \phi(x) \in P_a\}\Big] = \omega_a \wedge \widehat{\omega}_a$,

(iii) for all $a \in [k]$, $\lambda(\psi^{-1}(a)) = \widehat{\lambda}_a$.

Such a function ψ exists. First we construct ψ to satisfy (i) and (iii). For each a such that $\lambda_a > \widehat{\lambda}_a$, conditions (i) and (iii) are trivially satisfied if we take $\psi^{-1}(a)$ to be any subset of $\phi^{-1}(a)$ of Lebesgue measure $\widehat{\lambda}(a)$ and there is a subset of $\phi^{-1}(a)$ of Lebesgue measures $\lambda_a - \widehat{\lambda}_a$ left non-assigned. Summing over all such a, we see that there is a union of subsets with Lebesgue measure $m_+ := \sum_{a:\lambda_a > \widehat{\lambda}_a} (\lambda_a - \widehat{\lambda}_a)$ left non-assigned. On the other hand, for a such that $\lambda_a < \widehat{\lambda}_a$, we must have $\psi(x) = a$ for $x \in \phi^{-1}(a)$ to satisfy (i), while to meet condition (iii) we need additionally to assign $\psi(x) = a$ for x on a set of Lebesgue measure $\hat{\lambda}_a - \lambda_a$. 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 the union of non-assigned intervals for all a such that $\lambda_a > \widehat{\lambda}_a$ since $m_+ = m_-$ by virtue of (28). To ensure that condition (ii) is satisfied, we assign as a priority $\psi(x)$ to values belonging to the same partition element as $\phi(x)$. Again, (28) ensures that this is possible.

Finally, define the graphons $\widehat{W}(x,y) = \boldsymbol{Q}_{\psi(x),\psi(y)}$, $\widehat{W}_1(x,y) = \boldsymbol{Q}^{(ap)}_{\psi(x),\psi(y)}$, and $\widehat{W}_1^+(x,y) = \boldsymbol{Q}^{(ap,+)}_{\psi(x),\psi(y)}$ where $\boldsymbol{Q}$, $\boldsymbol{Q}^{(ap)}$, and $\boldsymbol{Q}^{(ap,+)}$ are as in (27). Notice that in view of (iii) $\widehat{W}$ is weakly isomorphic to the empirical graphon $\widetilde{f}_{\boldsymbol{\Theta}'_0}$. Let $\mathcal{R} = \{x\ , \phi(x) \neq \psi(x)\}$. Since W and $\widehat{W}$ match on $\mathcal{R}^c \times \mathcal{R}^c$, the purpose of (i) is to minimize the Lebesgue measure of the support of $W - \widehat{W}$. With properties (i) and (iii) alone, it would be possible to prove that $\mathbb{E}[\|W - \widehat{W}\|_\square \leq C\sqrt{k/n}$ as the Lebesgue measure of its support is at most of order $\sqrt{k/n}$. We will improve this rate by a logarithmic term as (ii) will enforce that the cut norm of $W - \widehat{W}$ is much smaller than its Lebesgue measure.

Step 3: Control of the cut norm. Since $\delta_\square(\cdot,\cdot)$ is a metric on the quotient space $\widetilde{\mathcal{W}}^+$,

$$\delta_\square(W_0, \widetilde{f}_{\boldsymbol{\Theta}'_0}) \ \leq \ \left\|W - \widehat{W}\right\|_\square = \sup_{S,T} \left| \int_{S\times T} (W(x,y) - \widehat{W}(x,y)) dxdy \right| .$$

By definition of ψ, the two functions $W(x,y)$ and $\widehat{W}(x,y)$ are equal except possibly when either x or y belongs to $\mathcal{R}$. As a consequence of triangular inequality and of the symmetry of $W - \widehat{W}$, we get

$$\begin{aligned} \left\|W - \widehat{W}\right\|_\square \ &\leq \ 2 \sup_{S\subset\mathcal{R}, T\subset\mathcal{R}^c} \left| \int_{S\times T} \left(W(x,y) - \widehat{W}(x,y)\right) dxdy \right| \\ &\quad + \sup_{S,T\subset\mathcal{R}} \left| \int_{S\times T} \left(W(x,y) - \widehat{W}(x,y)\right) dxdy \right| \\ &= 2 \left\| (W - \widehat{W})\Big|_{\mathcal{R}\times\mathcal{R}^c} \right\|_\square + \left\| (W - \widehat{W})\Big|_{\mathcal{R}\times\mathcal{R}} \right\|_\square . \end{aligned} \tag{29}$$

First, we focus on $\mathbb{E}[\|(W - \widehat{W})|_{\mathcal{R}\times\mathcal{R}^c}\|_\square]$, the second term being handled similarly at the end of the proof. For a and b in $[k]$, we write $a \sim_P b$ (resp. $a \nsim_P b$) when a and b belongs (resp. do not belong) to the same element of the partition P. Define

$$\mathcal{R}_2 := \{x,\ \psi(x) \nsim_P \phi(x)\} .$$

Obviously, we have $\mathcal{R}_2 \subset \mathcal{R}$. Property (ii) of ψ, implies that $\lambda(\mathcal{R}_2) = \sum_{a=1}^{q_1} (\omega_a - \widehat{\omega}_a)_+$. We shall rely on the decomposition $W = W_1 + (W - W_1)$ and $\widehat{W} = \widehat{W}_1 + (\widehat{W} - \widehat{W}_1)$. For any $x \in \mathcal{R} \setminus \mathcal{R}_2$, we have by definition (27) of W_1 that $(W_1 - \widehat{W}_1)(x,y) = 0$. Together with the triangular inequality,

this yields

$$\left\|(W-\widehat{W})|_{\mathcal{R}\times\mathcal{R}^c}\right\|_\square \le \left\|(W_1-\widehat{W}_1)\Big|_{\mathcal{R}_2\times\mathcal{R}^c}\right\|_\square + \left\|(W-W_1)\Big|_{\mathcal{R}\times\mathcal{R}^c}\right\|_\square + \left\|(\widehat{W}-\widehat{W}_1)\Big|_{\mathcal{R}\times\mathcal{R}^c}\right\|_\square . \quad (30)$$

To control the first expression in the rhs, we simply bound the cut norm of the difference by its l_1 norm

$$\left\|(W_1-\widehat{W}_1)\Big|_{\mathcal{R}_2\times\mathcal{R}^c}\right\|_\square \le \left\|(W_1-\widehat{W}_1)\Big|_{\mathcal{R}_2\times\mathcal{R}^c}\right\|_1 \le \lambda(\mathcal{R}_2)\left\|W_1-\widehat{W}_1\right\|_\infty \le \lambda(\mathcal{R}_2)\ ,$$

since W_1 and $\widehat{W}_1$ take values in $[0,1]$. Then, relying on the fact that $n\widehat{\omega}_a$ is distributed as a Binomial random variable with parameters (n,ω_a) and on Cauchy-Schwarz inequality, we get $\mathbb{E}\,|\omega_a-\widehat{\omega}_a| \le \sqrt{\frac{\omega_a(1-\omega_a)}{n}}$ and

$$\begin{aligned}\mathbb{E}\left[\left\|(W_1-\widehat{W}_1)\Big|_{\mathcal{R}_2\times\mathcal{R}^c}\right\|_\square\right] &\le \mathbb{E}\left[\sum_{a=1}^{q_1}|\omega_a-\widehat{\omega}_a|\right] \\ &\le \sum_{a=1}^{q_1}\sqrt{\frac{\omega_a(1-\omega_a)}{n}} \le \sqrt{\frac{q_1}{n}} \le \frac{k^{1/4}}{\sqrt{n}}, \end{aligned} \quad (31)$$

where we used again Cauchy-Schwarz in the last line. Let us turn to the second and third expressions in (30). To this end, we introduce a new kernel function U. For $a=1,\dots,k$, define $\widehat{\lambda}^\delta_a = |\lambda_a-\widehat{\lambda}_a|$ and the functions $F_{\widehat{\lambda}^\delta}\ :\ [k]\to\left[0,\sum_a|\lambda_a-\widehat{\lambda}_a|\right]$ and $F_\phi\ :\ [k]\mapsto[0,1]$ by

$$\begin{aligned} F_\phi(b) &= \sum_{a=1}^{b}\lambda_a \quad \text{and set} \quad F_\phi(0)=0 \\ F_{\widehat{\lambda}^\delta}(b) &= \sum_{a=1}^{b}\widehat{\lambda}^\delta_a \quad \text{and set} \quad F_{\widehat{\lambda}^\delta}(0)=0\ . \end{aligned} \quad (32)$$

For any $a,b\in[k]$, set $\widehat{\Pi}_{a,b}=[F_{\widehat{\lambda}^\delta}(a-1),F_{\widehat{\lambda}^\delta}(a))\times[F_\phi(b-1),F_\phi(b))$ and let U be a $k\times k$ step kernel on $[0,\sum_a|\widehat{\lambda}_a-\lambda_a|]\times[0,1]$ defined by

$$U(x,y) := \sum_{a,b=1}^{k}\left[\mathbf{Q}_{ab}-\mathbf{Q}^{(ap)}_{ab}\right]\mathbb{1}_{\widehat{\Pi}_{a,b}}(x,y).$$

By definition of $\mathcal{R}$ and of the function ψ, we have for any $a\in[k]$, $\lambda(\phi^{-1}(a))\cap\mathcal{R})=(\lambda_a-\widehat{\lambda})_+$ and $\lambda(\psi^{-1}(a))\cap\mathcal{R}^c)=\lambda_a\wedge\widehat{\lambda}$. As a consequence, the restriction of $(W-W_1)$ to $\mathcal{R}\times\mathcal{R}^c$ is, up to a measure preserving bijection of its rows and of its columns, equal to the restriction of U to the set $(\cup_{a:\ \lambda_a>\widehat{\lambda}_a}[F_{\widehat{\lambda}^\delta}(a-1),F_{\widehat{\lambda}^\delta}(a)))\times(\cup_a[F_\phi(a-1),F_\phi(a-1)+\widehat{\lambda}_a\wedge\lambda_a)$. This entails that

$$\left\|(W-W_1)\Big|_{\mathcal{R}\times\mathcal{R}^c}\right\|_\square \le \|U\|_\square\,. \quad (33)$$

On the other hand, for any $(x,y)\in\mathcal{R}\times\mathcal{R}^c$,

$$(\widehat{W}-\widehat{W}_1)(x,y)=\mathbf{Q}_{\psi(x)\psi(y)}-\mathbf{Q}^{(ap)}_{\psi(x)\psi(y)}=\mathbf{Q}_{\psi(x)\phi(y)}-\mathbf{Q}^{(ap)}_{\psi(x)\phi(y)}\ ,$$

by definition of $\mathcal{R}$. In view of the definition of ψ, for any $a \in [k]$ we have $\lambda(\phi^{-1}(a)) \cap \mathcal{R}) = (\widehat{\lambda} - \lambda_a)_+$. As a consequence, the restriction of $(\widehat{W} - \widehat{W}_1)$ to $\mathcal{R} \times \mathcal{R}^c$ is, up to a measure preserving bijection of its rows and of its columns, equal to the restriction of U to the set $(\cup_{a:\ \lambda_a < \widehat{\lambda}_a} [F_{\widehat{\lambda}^\delta}(a-1), F_{\widehat{\lambda}^\delta}(a))) \times (\cup_a [F_\phi(a-1), F_\phi(a-1) + \widehat{\lambda}_a \wedge \lambda_a)$. This implies that $\|(\widehat{W} - \widehat{W}_1)|_{\mathcal{R}\times\mathcal{R}^c}\|_\square \le \|U\|_\square$. Thus, we only have to control $\mathbb{E}[\|U\|_\square]$.

Step 4: Control of $\mathbb{E}[\|U\|_\square]$. Define the sets $\mathcal{B}_1 := \prod_{a=1}^k [0, |\widehat{\lambda}_a - \lambda_a|]$ and $\mathcal{B}_2 := \prod_{a=1}^k [0, |\lambda_a|]$. Then, the cut norm of U writes as

$$\begin{aligned} \|U\|_\square &\le \sup_{\gamma \in \mathcal{B}_1, \gamma' \in \mathcal{B}_2} \left| \sum_{a,b=1}^k \gamma_a \gamma'_b \left(\mathbf{Q}_{ab} - \mathbf{Q}^{(ap)}_{ab} \right) \right| \\ &\le \sup_{S,T \in [k]} \left| \sum_{a \in S, b \in T} \lambda_b |\widehat{\lambda}_a - \lambda_a| \left(\mathbf{Q}_{ab} - \mathbf{Q}^{(ap)}_{ab} \right) \right| , \end{aligned} \tag{34}$$

since the supremum of a linear function on a convex set is achieved at an extremal point. The random variable $|\widehat{\lambda}_a - \lambda_a|$ is in expectation of the order $\sqrt{\lambda_a/n}$. If we could replace each $|\widehat{\lambda}_a - \lambda_a|$ by $\sqrt{\lambda_a/n}$ in (34), then thanks to (25), we could prove that $\|U\|_\square$ is (up to a multiplicative constant) less than $\sqrt{k/(n\log(k))}$. Unfortunately, if we directly applied Bernstein inequality or the bounded difference inequality to simultaneously control $|\widehat{\lambda}_a - \lambda_a|$ over all $a \in [k]$ or to simultaneously control $\sum_{a\in S, b\in T} \lambda_b |\widehat{\lambda}_a - \lambda_a| (\mathbf{Q}_{ab} - \mathbf{Q}^{(ap)}_{ab})$ over all $S, T \subset [k]$, we would lose at least a logarithmic factor.

To bypass this issue, we adapt Lemma 10.9 of [16], which is a key point in the proof of sampling Lemma for graphons (Lemma 10.5 in [16]). Given a bounded non-symmetric kernel $W \in \mathcal{W}_{\mathcal{X},\mathcal{Y}}$, let us define the following one-side version of the cut norm:

$$\|W\|^+_\square = \sup_{X \subset \mathcal{X},\ Y \subset \mathcal{Y}} \int_{X\times Y} W(x,y)dxdy ,$$

where we take the supremum without any absolute value. As a consequence, the cut norm $\|W\|_\square$ is the maximum $\|W\|^+_\square$ and $\|-W\|^+_\square$.

Lemma 7. *Let $W \in \mathcal{W}_{[0,u],[0,v]}[k]$ and let $\boldsymbol{Q} \in \mathbb{R}^{k\times k}$, $\phi_1 : [0,u] \to [k]$ and $\phi_2 : [0,v] \to [k]$ be associated to W as in (17). For $a = 1, \dots, k$, define $\alpha_a := \lambda(\phi_1^{-1}(\{a\}))$ and $\beta_a := \lambda(\phi_2^{-1}(\{a\}))$. Given any subset $R \subset [k]$, let*

$$R^{l,W} := \left\{ b, \sum_{a\in R} \alpha_a \boldsymbol{Q}_{ab} > 0 \right\} , \quad R^{r,W} := \left\{ a, \sum_{b\in R} \beta_b \boldsymbol{Q}_{ab} > 0 \right\} . \tag{35}$$

Finally, we define for any $S, T \subset [k]$, $W[S,T] := \sum_{a\in S, b\in T} \alpha_a \beta_b \boldsymbol{Q}_{ab}$. Then, for any integer q with $1 \le q \le k$, we have

$$\|W\|^+_\square \le \max_{R_i \subset [k], |R_i| \le q} W\left[R_2^{r,W}, R_1^{l,W}\right] + \frac{u\sqrt{k\sum_{a=1}^k \beta_a^2} + v\sqrt{k\sum_{a=1}^k \alpha_a^2}}{\sqrt{q}} . \tag{36}$$

Note that in contrast to Equation (34) where one considers a supremum of 2^{2k} sums, only k^{2q} terms are involved in (36) up to the price of an additive term of order $q^{-1/2}$. The difficulty is that we will apply this lemma to U for which these k^{2q} will turn out to be random.

In the sequel, we fix $q = \lfloor \sqrt{k} \rfloor$ and apply Lemma 7 to U. Then, we can take $u = v = 1$. Since $\sum_{a=1}^k \lambda_a = 1$ and since we assumed at the beginning of the proof that the weigts λ_a are

all smaller than $2/k$, it follows that $(k\sum_{a=1}^k \lambda_a^2)^{1/2} \leq \sqrt{2}$. Let M and N denote the random variables $M := \sum_{a=1}^k |\widehat{\lambda}_a - \lambda_a|$ and $N := \left(\sum_{a=1}^k k|\widehat{\lambda}_a - \lambda_a|^2\right)^{1/2}$. Both M and N are functions of the independent random variables $(\xi_1, \dots, \xi_n)$. Besides, if we change the values of one of these ξ_i' the value of M changes by at most $2/n$ and the value of N changes by at most $\sqrt{2k}/n$. As a consequence, we may apply the bounded difference inequality (Lemma (2)) to these two random variables. Then, with probability larger than $1 - 2\exp(-\sqrt{k}/\log(k))$, one has

$$\sum_{a=1}^k \left|\widehat{\lambda}_a - \lambda_a\right| \leq \mathbb{E}\left[\sum_{a=1}^k |\widehat{\lambda}_a - \lambda_a|\right] + \sqrt{\frac{2k^{1/2}}{n\log(k)}} \leq C\sqrt{\frac{k}{n}}\,, \tag{37}$$

$$\left(k\sum_{a=1}^k |\widehat{\lambda}_a - \lambda_a|^2\right)^{1/2} \leq \mathbb{E}\left[\left(k\sum_{a=1}^k |\widehat{\lambda}_a - \lambda_a|^2\right)^{1/2}\right] + \sqrt{\frac{2k^{3/2}}{n\log(k)}} \leq Ck^{1/4}\sqrt{\frac{k}{n\log(k)}}\,. \tag{38}$$

In (37)-(38) we bound the expectation using that, 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) and the Cauchy-Schwarz inequality:

$$\mathbb{E}\left[\sum_{a=1}^k |\widehat{\lambda}_a - \lambda_a|\right] \leq \sum_{a=1}^k \sqrt{\frac{\lambda_a(1-\lambda_a)}{n}} \leq \sqrt{\frac{k}{n}} \quad \text{and}$$

$$\mathbb{E}\left[\left(k\sum_{a=1}^k |\widehat{\lambda}_a - \lambda_a|^2\right)^{1/2}\right] \leq \sqrt{k\sum_{a=1}^k \frac{\lambda_a(1-\lambda_a)}{n}} \leq \sqrt{\frac{k}{n}}.$$

Bound (38) and $(k\sum_{a=1}^k \lambda_a^2)^{1/2} \leq \sqrt{2}$, implies that for U, with probability larger than $1 - 2\exp(-\sqrt{k}/\log(k))$,

$$\frac{\sqrt{k\sum_{a=1}^k \beta_a^2} + \sqrt{k\sum_{a=1}^k \alpha_a^2}}{k^{1/4}} \leq C\sqrt{\frac{k}{n\log(k)}}. \tag{39}$$

Fix any two subsets $R_1, R_2 \subset [k]$ of size less or equal to q. In view of (36), one needs to control the following random variable

$$Z_{R_1,R_2} := U\left[R_2^{r,U}, R_1^{l,U}\right] = \sum_{a\in R_2^{r,U}} \left|\widehat{\lambda}_a - \lambda_a\right| \sum_{b\in R_1^{l,U}} \lambda_b(\boldsymbol{Q}_{ab} - \boldsymbol{Q}_{ab}^{ad})\,. \tag{40}$$

It is done in the following Lemma:

Lemma 8. *Let R_1, R_2 be two subsets of $[k]$ of size less or equal to q and Z_{R_1,R_2} given by (40). Then, we have that with probability larger than $1 - 3\exp(-\sqrt{k}/\log(k))$,*

$$\max_{R_1,R_2\,:\,|R_1|\leq q|R_2|\leq q} Z_{R_1,R_2} \leq C\sqrt{\frac{k}{n\log(k)}}.$$

Now, it follows from Lemma 7 together with (39) and Lemma 8 that, with probability larger than $1 - 5\exp(-\sqrt{k}/\log(k))$,

$$\|U\|_\square^+ \leq C\sqrt{\frac{k}{n\log(k)}}.$$

Controlling analogously $\| - U\|_{\square}^{+}$, we conclude that there exists an event $\mathcal{A}$ of probability larger than $1 - 10\exp(-\sqrt{k}/\log(k))$ such that, on $\mathcal{A}$,

$$\|U\|_{\square} \leq C\sqrt{\frac{k}{n\log(k)}} \ .$$

To finish the control of $\mathbb{E}[\|U\|_{\square}]$, we use the rough bound $\|U\|_{\square} \leq \|U\|_1 \leq \sum_{a=1}^{k} |\widehat{\lambda}_a - \lambda_a|$ on the complementary event $\bar{\mathcal{A}}$.

$$\begin{aligned}
\mathbb{E}[\|U\|_{\square}] &\leq E\left[\|U\|_{\square}\mathbb{1}_{\mathcal{A}}\right] + \mathbb{E}\left[\|U\|_{\square}\mathbb{1}_{\bar{\mathcal{A}}}\right] \\
&\leq C\sqrt{\frac{k}{n\log(k)}} + \sqrt{\mathbb{P}\left(\bar{\mathcal{A}}\right)}\left[\mathbb{E}\left(\sum_{a=1}^{k}|\widehat{\lambda}_a - \lambda_a|\right)^2\right]^{1/2} \\
&\leq C\sqrt{\frac{k}{n\log(k)}} + C'e^{-\sqrt{k}/(2\log(k))}\sqrt{\frac{k}{n}} \leq C''\sqrt{\frac{k}{n\log(k)}} \qquad (41)
\end{aligned}$$

where we use (37). Now, using the decomposition (30), (31) and (33), we can conclude that

$$\mathbb{E}\left[\left\|(W - \widehat{W})|_{\mathcal{R}\times\mathcal{R}^c}\right\|_{\square}\right] \leq C\sqrt{\frac{k}{n\log(k)}}.$$

The following lemma gives a corresponding bound on the second term $\left\|(W - \widehat{W})\right|_{\mathcal{R}\times\mathcal{R}}\|_{\square}$ in (29). The proof is somewhat analogous to that of the control of $\left\|(W - \widehat{W})|_{\mathcal{R}\times\mathcal{R}^c}\right\|_{\square}$ and is differed to the end of the section.

Lemma 9. *We have*

$$\mathbb{E}\left[\left\|(W - \widehat{W})|_{\mathcal{R}\times\mathcal{R}}\right\|_{\square}\right] \leq C\sqrt{\frac{k}{n\log(k)}} \ .$$

In view of (29), we have proved Proposition 3. □

Proof of Lemma 6. For $a \in [k]$, we denote $(\lambda_0)_a = \lambda(\phi_0^{-1}(a))$ and $u_a = \frac{\sqrt{(\lambda_0)_a}}{\sum_b \sqrt{(\lambda_0)_b}}$. For any $b \in [k]$, define the cumulative distribution functions $F_0(b) = \sum_{a=1}^{b}(\lambda_0)_a$ and $F_1(b) = \sum_{a=1}^{b} u_a$. For $a, b \in [k]$, let $(\Pi_d)_{ab} = [F_0(a-1), F_0(a)) \times [F_1(b-1), F_1(b))$ and $(\Pi_d^+)_{ab} = [F_1(a-1), F_1(a)) \times [F_1(b-1), F_1(b))$. Finally we consider the (non nessarily symmetric) kernels W_d and W_d^+ defined by

$$W_d(x,y) = \sum_{a=1}^{k}\sum_{b=1}^{k}(\boldsymbol{Q}_0)_{ab}\mathbb{1}_{(\Pi_d)_{ab}}(x,y) \ , \quad W_d^+(x,y) = \sum_{a=1}^{k}\sum_{b=1}^{k}(\boldsymbol{Q}_0)_{ab}\mathbb{1}_{(\Pi_d^+)_{ab}}(x,y) \ .$$

In comparison to W_0, the length of the steps in W_d and W_d^+ has been modified.

Lemma 10. *Let $W \in \mathcal{W}_{[0,1],[0,1]}$ be a k-step kernel defined by*

$$W(x,y) = \sum_{a=1}^{k}\sum_{b=1}^{k}\boldsymbol{Q}_{ab}\mathbb{1}_{S_a\times T_b}(x,y)$$

where $\boldsymbol{Q} \in [0,1]^{k\times k}$ and $(S_1, \ldots, S_k)$ and $(T_1, \ldots, T_k)$ are two partitions of $[0,1]$ into a finite number of measurable sets. For any integer $q_0 \geq 2$, there exist a q_0-step kernel $W^{(ap)} \in \mathcal{W}_{[0,1],[0,1]}^{+}$ satisfying

(i) For any $(a,b) \in [k]$, $W^{(ap)}$ is constant on $S_a \times T_b$.

(ii) $\left\|W - W^{(ap)}\right\|_\square \leq \frac{C}{\sqrt{\log(q_0)}}$.

The second property (ii) is just the consequence of the weak Regularity Lemma for kernels [11] (see also Corollary 9.13 in [16]). The first property, (i), follows from the explicit construction of the approximate kernel by Kannan and Frieze (see the proof of Lemma 9.10 in [16]). For the sake of completeness, we give the details in the end of this section.

Fix $q_0 = \lfloor k^{1/4} \rfloor$. Note that $q_0 \geq 2$ since we assume that $k \geq 16$. We denote by $W_d^{(ap)}$ and $W_d^{(ap,+)}$ the q_0-step kernels given by Lemma 10 to respectively approximate W_d and $W_d^{(+)}$. In virtue of Property (i), there exist two matrices $\boldsymbol{Q}_0^{(ap)}$ and $\boldsymbol{Q}_0^{(ap,+)}$ in $[0,1]^{k\times k}$ such that

$$W_d^{(ap)}(x,y) = \sum_{a=1}^k \sum_{b=1}^k (\boldsymbol{Q}_0^{(ap)})_{ab} \mathbb{1}_{(\Pi_d)_{ab}}(x,y) \text{ and } W_d^{(ap,+)}(x,y) = \sum_{a=1}^k \sum_{b=1}^k (\boldsymbol{Q}_0^{(ap,+)})_{ab} \mathbb{1}_{(\Pi_d^+)_{ab}}(x,y) \ .$$

There exist two partitions $\mathcal{P}_d$ and $\mathcal{P}_d^+$ of $[k]$ such that $\boldsymbol{Q}_0^{(ap)}$ is block constant according to $\mathcal{P}_d$ and $\boldsymbol{Q}_0^{(ap,+)}$ is block constant according to $\mathcal{P}_d^+$. Let $\mathcal{P}^*$ be the coarsest partition that refines both $\mathcal{P}$ and $\mathcal{P}_d^+$. As a consequence, $\mathcal{P}^*$ is made of less than $q_0^2 \leq q_1$ subsets. By possibly refining $\mathcal{P}^*$, we may assume without loss of generality that $\mathcal{P}^* = (P_1^*, \dots, P_{q_1}^*)$ is made of exactly q_1 elements. Let π be a permutation of $[k]$ transforming $\mathcal{P}^*$ in a partition $\mathcal{P} = (P_1, \dots, P_{q_1})$ with $P_a = \{\pi(b), b \in P_a^*\}$ made of consecutive intervals. Denoting $\boldsymbol{\Pi}$ the corresponding permutation matrix, we finally take

$$\boldsymbol{Q} = \boldsymbol{\Pi}^T \mathbf{Q}_0 \boldsymbol{\Pi}, \quad \boldsymbol{Q}^{(ap)} = \boldsymbol{\Pi}^T \boldsymbol{Q}_0^{(ap)} \boldsymbol{\Pi}, \quad \text{and} \quad \boldsymbol{Q}^{(ap,+)} = \boldsymbol{\Pi}^T \boldsymbol{Q}_0^{(ap,+)} \boldsymbol{\Pi} \ .$$

Now we are ready to prove (25) and (26). Recall that we denote $\phi = \pi \circ \phi_0$ and $\lambda_a := \lambda(\phi^{-1}(a))$ for $a \in [k]$. Define the sets $\mathcal{B}_1 := \prod_{a=1}^k [0, u_{\pi(a)}]$ and $\mathcal{B}_2 := \prod_{a=1}^k [0, \lambda_a]$. Since $W_d - W_d^{(ap)}$ is a k-step function, its cut norm writes as

$$\begin{aligned} \|W_d - W_d^{(ap)}\|_\square &= \sup_{\gamma \in \mathcal{B}_1 \gamma \in \mathcal{B}_2} \left| \sum_{a,b} \gamma_a \gamma_b \left(\boldsymbol{Q}_{ab} - \boldsymbol{Q}_{ab}^{(ap)} \right) \right| && (42) \\ &= \sup_{\epsilon \in \{0,1\}^k,\ \epsilon' \in \{0,1\}^k} \left| \sum_{a,b} \epsilon_a \epsilon'_b \lambda_b u_{\pi(a)} \left(\boldsymbol{Q}_{ab} - \boldsymbol{Q}_{ab}^{(ap)} \right) \right| \leq \frac{C}{\sqrt{\log(q_0)}} && (43) \end{aligned}$$

since the supremum is achieved at an extremal point of the convex and in the last inequality we use property (ii) of Lemma 10. Now (42) and the definition of $u_{\pi(a)}$ imply

$$\sup_{\epsilon \in \{0,1\}^k,\ \epsilon' \in \{0,1\}^k} \left| \sum_{a,b} \epsilon_a \epsilon'_b \lambda_b \sqrt{\lambda_a} \left(\boldsymbol{Q}_{ab} - \boldsymbol{Q}_{ab}^{(ap)} \right) \right| \leq C \frac{\sum_{b\in[k]} \sqrt{\lambda_b}}{\sqrt{\log(q_0)}} \leq C' \sqrt{\frac{k}{\log(k)}} \ ,$$

by Cauchy-Schwarz inequality. We have proved (25). The second inequality (26) is derived similarly. □

Proof of Lemma 10. We adapt the proof of the weak Regularity Lemma for symmetric kernels [16, Lemma 9.9] to nonsymmetric ones. We use the following extension of Lemma 9.11(a) in [16].

Lemma 11. *For every $W \in \mathcal{W}_{[0,1],[0,1]}[k]$ such that*

$$W(x,y) = \sum_{a=1}^{k}\sum_{b=1}^{k} \boldsymbol{Q}_{ab}\mathbb{1}_{S_a\times T_b}(x,y)$$

where $\boldsymbol{Q} \in \mathbb{R}^{k\times k}$ and $\mathcal{P} = \{(S_1,\dots,S_k),(T_1,\dots,T_k)\}$ are two partitions of $[0,1]$ into a finite number of measurable sets, there are two sets $\mathcal{A},\mathcal{B} \subset [k]$ and a real number $0 \le a \le \max_{a,b}|\boldsymbol{Q}_{ab}|$ such that, for $S' = \cup_{a\in\mathcal{A}} S_a$ and $T' = \cup_{b\in\mathcal{B}} T_b$,

$$\|W - a\mathbb{1}_{S'\times T'}\|_2^2 \le \|W\|_2^2 - \|W\|_\square^2 \ .$$

Now we apply Lemma 11 repeatedly, to get pairs of sets S_i', T_i' and real numbers a_i such that for any positive integer j, $W_j = W - \sum_{i=1}^{j} a_i\mathbb{1}_{S_i'\times T_i'}$ we have

$$\|W_j\|_2^2 \le \|W\|_2^2 - \sum_{i=1}^{j-1}\|W_i\|_\square^2.$$

Fix some integer $k_0 > 0$. Since the right-hand side of the above equation remains nonnegative, there exists $0 \le i < k_0$ with $\|W_i\|_\square^2 \le 1/k_0$. Now putting $a_l = 0$ for $l > i$ we get that for any $W \in \mathcal{W}_{[0,1],[0,1]}[k]$ and any $k_0 \ge 1$ there are k_0 pairs of subsets $S_i', T_i' \subset [0,1]$ and k_0 real numbers a_i such that

$$\left\| W - \sum_{i=1}^{k_0} a_i\mathbb{1}_{S_i'\times T_i'} \right\|_\square < \frac{1}{\sqrt{k_0}}. \tag{44}$$

Note that the approximation $W^{ap} = \sum_{i=1}^{k_0} a_i\mathbb{1}_{S_i'\times T_i'}$ is a step function with at most 2^{k_0} steps and $a_i \ge O$, for all i. On the other hand, by construction we have that for any $(a,b) \in [k]$, $W^{(ap)}$ is constant on all sets of the form $S_a \times T_b$. We conclude by taking $k_0 = \lfloor \log(q_0)/\log(2) \rfloor$.

□

Proof of Lemma 11. This lemma is proved in [16, Lemma 9.11] for symmetric kernels. For readers convenience we get the details here. Let W be a k-step kernel and let $(S_1,\dots,S_k),(T_1,\dots,T_k)$ be two measurable partitions of $[0,1]$ such that W is constant on each set $S_i \times T_j$. Relying on a convexity argument as in the proof of Lemma 6, the cut norm is achieved for measurable sets S and T that are unions of S_i and T_j respectively, that is

$$\|W\|_\square = \left| \int_{S\times T} W(x,y)dxdy \right| \ ,$$

where $S = \cup_{a\in\mathcal{A}} S_a$ and $T = \cup_{b\in\mathcal{B}} T_b$ with $\mathcal{A}$, $\mathcal{B} \subset [k]$. Let $\mathbf{a} = \frac{1}{\lambda(S)\lambda(T)}\|W\|_\square$. Then, we have

$$\|W - \mathbf{a}\mathbb{1}_{S\times T}\|_2^2 = \|W\|_2^2 - \frac{1}{\lambda(S)\lambda(T)}\|W\|_\square^2 \le \|W\|_2^2 - \|W\|_\square^2$$

which completes the proof. □

Proof of Lemma 7. This proof follows closely that of Lemma 10.9 in [16]. It is easy to see that

$$\|W\|_\square^+ = \max_{S,T\subset[k]} W[S,T]$$

so we only need to bound these expressions. Let Q and Q' be independent uniformly chosen q-subset of $[k]$ and let $\mathbb{E}_Q$ (resp. $\mathbb{E}_{Q'}$) denote the expectation with respect to Q(resp. Q'). We shall prove that, for any $S, T \subset [k]$

$$W[S,T] \leq \mathbb{E}_Q\left[W[(Q\cap T)^{r,W},T]\right] + \frac{u\sqrt{k\sum_{a=1}^k \beta_a^2}}{\sqrt{q}} \ . \tag{45}$$

By symmetry, this will imply

$$W[S,T] \leq \mathbb{E}_{Q'}\left[W[S,(Q'\cap S)^{l,W}]\right] + \frac{v\sqrt{k\sum_{a=1}^k \alpha_a^2}}{\sqrt{q}} \ ,$$

so that gathering both inequalities yields to

$$W[S,T] \leq \mathbb{E}_{Q,Q'}\left[W[(Q\cap T)^{r,W},(Q'\cap (Q\cap T)^{r,W})^{l,W}]\right] + \frac{u\sqrt{k\sum_{a=1}^k \beta_a^2} + v\sqrt{k\sum_{a=1}^k \alpha_a^2}}{\sqrt{q}} \ .$$

Since the above expectation is less or equal to $\sup_{R_i,\ |R_i|\leq q} W\left[R_2^{r,W}, R_1^{l,W}\right]$, this will conclude the proof. Thus, we only have to show (45). Note that using $W[S,T] \leq W[T^{r,W},T]$ it suffices to prove that

$$\mathbb{E}_Q\left[W[T^{r,W}\setminus (Q\cap T)^{r,W},T]\right] - \mathbb{E}_Q\left[W[(Q\cap T)^{r,W}\setminus T^{r,W},T]\right] \leq \frac{u\sqrt{k\sum_{a=1}^k \beta_a^2}}{\sqrt{q}} \ . \tag{46}$$

Let us denote Z the above difference of expectations. For any $a\in[k]$, write $B_a = \sum_{b\in T}\beta_b \boldsymbol{Q}_{ab}$ and $A_a = \sum_{b\in T\cap Q}\beta_b \boldsymbol{Q}_{ab}$. By the definition (35), we have that B_a is non-negative for $a\in T^{r,W}$ and $B_a \leq 0$ if $a\notin T^{r,W}$. In the same way, $A_a > 0$ for $a\in (Q\cap T)^{r,W}$ and $A_a \leq 0$ for $a\notin (Q\cap T)^{r,W}$. Denoting $\mathbb{P}_Q$ the probability with respect to Q, we obtain

$$\begin{aligned} Z &= \mathbb{E}_Q\left(\sum_{a\in T^{r,W}} \mathbb{1}_{\{a\notin (Q\cap T)^{r,W}\}}\alpha_a B_a + \sum_{a\notin T^{r,W}} \mathbb{1}_{\{a\in (Q\cap T)^{r,W}\}}\alpha_a |B_a|\right) \\ &= \sum_{a\in T^{r,W}} \mathbb{P}_Q[A_a\leq 0]\alpha_a B_a + \sum_{a\notin T^{r,W}} \mathbb{P}_Q[A_a>0]\alpha_a|B_a|. \end{aligned} \tag{47}$$

Now, using $\mathbb{E}_Q[A_a] = qB_a/k$, it follows from the Chebyshev inequality that, for $a\in T^{r,W}$, we have $\mathbb{P}_Q[A_a<0] \leq \mathrm{Var}_Q[A_a]/\,\mathbb{E}_Q^2[A_a]$. Since a probability is smaller or equal to one, it follows that $\mathbb{P}_Q[A_a<0] \leq \sqrt{\mathrm{Var}_Q[A_a]}/|\,\mathbb{E}_Q[A_a]|$. Similarly, for $a\notin T^{r,W}$ we also have that $\mathbb{P}_Q[A_a>0] \leq \sqrt{\mathrm{Var}_Q[A_a]}/|\,\mathbb{E}_Q[A_a]|$. Coming back to Z, this yields

$$Z \leq \sum_{a\in[k]} \alpha_a|B_a|\frac{\mathrm{Var}_Q^{1/2}[A_a]}{|E_Q[A_a]|} = \frac{k}{q}\sum_{a\in[k]}\alpha_a\,\mathrm{Var}_Q^{1/2}[A_a] \leq \frac{ku}{q}\max_{a\in[k]}\mathrm{Var}_Q^{1/2}[A_a] \ .$$

Working out the variance, we get $\mathrm{Var}_Q[A_a] \leq \frac{q}{k}\sum_{b\in T}\beta_b^2\boldsymbol{Q}_{ab}^2 \leq q(\sum_{b\in[k]}\beta_b^2)/k$, which concludes the proof. □

Proof of Lemma 8. Note that in (40), the definition of Z_{R_1,R_2}, the set $R_2^{r,U}$ is deterministic whereas the set $R_1^{l,U}$ only depends on $(\widehat{\lambda}_a)_{a\in R_1}$. We can upper bound Z_{R_1,R_2} in the following way:

$$Z_{R_1,R_2} \leq \sum_{a\in R_2^{r,U}\setminus R_1} \left|\widehat{\lambda}_a - \lambda_a\right| \sum_{b\in R_1^{l,U}} \lambda_b \left(\boldsymbol{Q}_{ab} - \boldsymbol{Q}^{ad}_{ab}\right) + \sum_{a\in R_1} \left|\widehat{\lambda}_a - \lambda_a\right| \tag{48}$$

where we use $\left|\sum_b \lambda_b(\boldsymbol{Q}_{ab} - \boldsymbol{Q}^{ad}_{ab})\right| \leq 1$. We set

$$T_{R_1,R_2} = \sum_{a\in R_2^{r,U}\setminus R_1} \left|\widehat{\lambda}_a - \lambda_a\right| \sum_{b\in R_1^{l,U}} \lambda_b \left(\boldsymbol{Q}_{ab} - \boldsymbol{Q}^{ad}_{ab}\right).$$

Conditionally to $(\widehat{\lambda}_a)_{a\in R_1}$, T_{R_1,R_2} is distributed as a function of $n - n\sum_{a\in R_1}\widehat{\lambda}_a$ i.i.d. random variables ξ_i' such that $\mathbb{P}[\xi' = a] = \lambda_a/(1-\sum_{a\in R_1}\lambda_a)$ for any $a \in [k]\setminus R_1$. Besides, if we change the values of one of these ξ_i' the value of this expression changes by at most $2/n$. It then follows from the bounded difference inequality (Lemma (2)) that, for any $t > 0$.

$$\mathbb{P}\left\{T_{R_1,R_2} \geq \mathbb{E}\left[T_{R_1,R_2}|(\widehat{\lambda}_a)_{a\in R_1}\right] + \sqrt{\frac{2t}{n}}\Big|(\widehat{\lambda}_a)_{a\in R_1}\right\} \geq 1 - e^{-t}. \tag{49}$$

Let us bound this conditional expectation:

$$\begin{aligned}\mathbb{E}\left[T_{R_1,R_2}|(\widehat{\lambda}_a)_{a\in R_1}\right] &= \sum_{a\in R_2^{r,U}\setminus R_1} \mathbb{E}\left[\left|\widehat{\lambda}_a - \lambda_a\right| |(\widehat{\lambda}_c)_{c\in R_1}\right] \sum_{b\in R_1^{l,U}} \lambda_b \left(\boldsymbol{Q}_{ab} - \boldsymbol{Q}^{ad}_{ab}\right) \\ &\leq \sup_{S\subset[k]\setminus R_1, T\subset[k]} \sum_{a\in S, b\in T} \mathbb{E}\left[\left|\widehat{\lambda}_a - \lambda_a\right| |(\widehat{\lambda}_c)_{c\in R_1}\right] \lambda_b(\boldsymbol{Q}_{ab} - \boldsymbol{Q}^{ad}_{ab}) \,. \end{aligned} \tag{50}$$

Now, using Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[\left|\widehat{\lambda}_a - \lambda_a\right| |(\widehat{\lambda}_c)_{c\in R_1}\right] \leq \sqrt{\frac{\lambda_a}{n\left(1-\sum_{b\in R_1}\lambda_b\right)}} \leq \sqrt{\frac{\lambda_a}{n\left(1-2q/k\right)}} \leq 2\sqrt{\frac{\lambda_a}{n}}\,,$$

where we used that $\lambda_b \leq 2/k$, $|R_1| \leq q \leq k^{1/2}$ and $k \geq 8$. The supremum in (50) is achieved for subsets (S^*, T^*) such that for all $a \in S^*$, $\sum_{b\in T^*}\lambda_b(\boldsymbol{Q}_{ab} - \boldsymbol{Q}^{ad}_{ab})$ is non-negative (otherwise this contradicts the optimality of S^*, T^*). As a consequence, we can plug the upper bounds on $\mathbb{E}\left[\left|\widehat{\lambda}_a - \lambda_a\right| (\widehat{\lambda}_c)_{c\in R_1}\right]$ into (50):

$$\mathbb{E}\left[T_{R_1,R_2}|(\widehat{\lambda}_a)_{a\in R_1}\right] \leq \frac{2}{\sqrt{n}} \sup_{S\subset[k]\setminus R_1, T\subset[k]} \sum_{a\in S, b\in T} \sqrt{\lambda_a}\lambda_b(\boldsymbol{Q}_{ab} - \boldsymbol{Q}^{ad}_{ab}) \leq C\sqrt{\frac{k}{n\log(k)}}\,,$$

where we used the property (25) of $\boldsymbol{Q}^{ad}$. Coming back to (49) and integrating the deviation inequality with respect to $(\widehat{\lambda}_a)_{a\in R_1}$, we conclude that, for any $t > 0$

$$\mathbb{P}\left[T_{R_1,R_2} \geq C\sqrt{\frac{k}{n\log(k)}} + \sqrt{\frac{2t}{n}}\right] \geq 1 - e^{-t}.$$

Fixing $t = 2\log(k)q + \sqrt{k}/\log(k)$ and taking an union bound over all possible R_1, R_2, we derive that

$$\max_{R_1,R_2\,,|R_1|\leq q|R_2|\leq q} Z_{R_1,R_2} \leq C\sqrt{\frac{k}{n\log(k)}} + q\max_{a=1,\dots k}\left|\widehat{\lambda}_a - \lambda_a\right| \tag{51}$$

on an event of probability higher than $1-\exp(-\sqrt{k}/\log(k))$.

Next we bound $\max_{a=1,\dots k}|\widehat{\lambda}_a-\lambda_a|$. Recall that $n\widehat{\lambda}_a$ has a binomial distribution with parameters $(n,\,\lambda_a)$ and $\lambda_a\le 2/k$. For any $a\in[k]$, applying Bernstein inequality to $|\widehat{\lambda}_a-\lambda_a|$ we get

$$\mathbb{P}\left\{n|\widehat{\lambda}_a-\lambda_a|\ge t\right\}\le 2\exp\left(-\frac{t^2}{4n/k+2t/3}\right).$$

Taking $t=\sqrt{n/\log(k)}$ and applying the union bound, we derive that with probability larger than $1-2\exp(-\sqrt{k}/\log(k))$

$$\sqrt{k}\max_{a=1,\dots k}|\widehat{\lambda}_a-\lambda_a|\le C\sqrt{k/(n\log(k))}. \tag{52}$$

Bound (51) together with (52) imply the statement of Lemma 8. □

Proof of Lemma 9. As the control of $(W-\widehat{W})|_{\mathcal{R}\times\mathcal{R}}$ is quite similar to that of $(W-\widehat{W})|_{\mathcal{R}\times\mathcal{R}^c}$, we only sketch the main steps. Relying on the graphon W_1^+ (defined in (27)), we have the following decomposition:

$$\|(W-\widehat{W})|_{\mathcal{R}\times\mathcal{R}}\|_\square\le\|(W_1^+-\widehat{W}_1^+)\big|_{\mathcal{R}\times\mathcal{R}}\|_\square+\|(W-W_1^+)\big|_{\mathcal{R}\times\mathcal{R}}\|_\square+\|(\widehat{W}-\widehat{W}_1^+)\big|_{\mathcal{R}\times\mathcal{R}}\|_\square. \tag{53}$$

Since $(W_1^+-\widehat{W}_1^+)(x,y)$ is zero except if $x\in\mathcal{R}_2$ or $y\in\mathcal{R}_2$, we bound the first expression by its l_1 norm as for $W_1-\widehat{W}_1$:

$$\mathbb{E}\left[\|(W_1^+-\widehat{W}_1^+)\big|_{\mathcal{R}\times\mathcal{R}}\|_\square\right]\le 2\,\mathbb{E}[\lambda(\mathcal{R}_2)]\le 2\frac{k^{1/4}}{\sqrt{n}}. \tag{54}$$

The two last expressions in (53) are bounded by the cut norm of a kernel V defined as follows. For any $a,b\in[k]$, define $\widetilde{\Pi}_{a,b}=[F_{\widehat{\lambda}^\delta}(a-1),F_{\widehat{\lambda}^\delta}(a))\times[F_{\widehat{\lambda}^\delta}(b-1),F_{\widehat{\lambda}^\delta}(b))$ where $F_{\widehat{\lambda}^\delta}(.)$ has been defined in (32). Let V be the $k\times k$ step kernel on $\left[0,\sum_a|\widehat{\lambda}_a-\lambda_a|\right]^2$ given by

$$V(x,y):=\sum_{a,b=1}^{k}\left[\mathbf{Q}_{ab}-\mathbf{Q}_{ab}^{(ap,+)}\right]\mathbb{1}_{\widetilde{\Pi}_{a,b}}(x,y).$$

Now, as for the restrictions of $W-W_1$ and $\widehat{W}-\widehat{W}_1$ to $\mathcal{R}\times\mathcal{R}^c$, we have

$$\left\|(W-W_1^+)\big|_{\mathcal{R}\times\mathcal{R}}\right\|_\square\bigvee\left\|(\widehat{W}-\widehat{W}_1^+)\big|_{\mathcal{R}\times\mathcal{R}}\right\|_\square\le\|V\|_\square\,. \tag{55}$$

Thus, it boils down to controlling $\mathbb{E}\left[\|V\|_\square\right]$. Since V is a k-step kernel, its cut norm writes as

$$\|V\|_\square=\sup_{S,T\subset[k]}\Big|\sum_{a\in S,\ b\in T}|\widehat{\lambda}_a-\lambda_a||\widehat{\lambda}_a-\lambda_a|\big(\mathbf{Q}_{ab}-\mathbf{Q}_{ab}^{(ap,+)}\big)\Big|.$$

As for the kernel U in the main proof, we rely on the Lemma 7. The random variables $\sum_a|\widehat{\lambda}_a-\lambda_a|$ and $(\sum_a|\widehat{\lambda}_a-\lambda_a|^2)^{1/2}$ are controlled as in (37) and (38).

Fix any two subsets $R_1,R_2\subset[k]$ of size less or equal to q and define

$$Z_{R_1,R_2}:=V[R_2^{r,V},R_1^{l,V}]=\sum_{a\in R_2^{r,V}}\sum_{b\in R_1^{l,V}}|\widehat{\lambda}_a-\lambda_a||\widehat{\lambda}_b-\lambda_b|(\boldsymbol{Q}_{ab}-\boldsymbol{Q}_{ab}^{ad,+})\,.$$

The set $R_1^{l,V}$ only depends on $(\widehat{\lambda}_a)_{a\in R_1}$ and $R_2^{r,V}$ only depends on $(\widehat{\lambda}_a)_{a\in R_2}$. We have

$$Z_{R_1,R_2} \leq \sum_{a\in R_2^{r,V}\setminus(R_1\cup R_2)} \sum_{b\in R_1^{l,V}\setminus(R_1\cup R_2)} |\widehat{\lambda}_a-\lambda_a||\widehat{\lambda}_b-\lambda_b|(\boldsymbol{Q}_{ab}-\boldsymbol{Q}_{ab}^{ad,+}) + 4\sum_{a\in R_1\cup R_2}|\widehat{\lambda}_a-\lambda_a|\ ,$$

since $\sum_{a\in[k]}|\widehat{\lambda}_a-\lambda_a|\leq 2$. We set

$$T_{R_1,R_2} = \sum_{a\in R_2^{r,V}\setminus(R_1\cup R_2)} \sum_{b\in R_1^{l,V}\setminus(R_1\cup R_2)} |\widehat{\lambda}_a-\lambda_a||\widehat{\lambda}_b-\lambda_b|(\boldsymbol{Q}_{ab}-\boldsymbol{Q}_{ab}^{ad,+}).$$

Write $R := R_1\cup R_2$ and $\widehat{\lambda}_{\{R\}} := (\widehat{\lambda}_a)_{a\in R}$. Conditionally to $\widehat{\lambda}_{\{R\}}$, T_{R_1,R_2} is a function of $n - n\sum_{a\in R}\widehat{\lambda}_a$ independent random variables. Besides, if we change the values of one of these independent random variables the value of T_{R_1,R_2} changes by at most $4/n$. Hence, the bounded difference inequality enforces that, for any $t>0$,

$$\mathbb{P}\left[T_{R_1,R_2}\geq \mathbb{E}[T_{R_1,R_2}|\widehat{\lambda}_{\{R\}}] + 8\sqrt{\frac{2t}{n}}\Big|\widehat{\lambda}_{\{R\}}\right]\geq 1-e^{-t}. \tag{56}$$

The conditional expectation is upper bounded by

$$\mathbb{E}[T_{R_1,R_2}|\widehat{\lambda}_{\{R\}}] \leq \sup_{S\subset[k]\setminus R,\, T\subset[k]\setminus R}\sum_{a\in S,b\in T}\mathbb{E}\left[|\widehat{\lambda}_a-\lambda_a||\widehat{\lambda}_b-\lambda_b|\big|\widehat{\lambda}_{\{R\}}\right](\boldsymbol{Q}_{ab}-\boldsymbol{Q}_{ab}^{ad,+})\ . \tag{57}$$

Here, unfortunately, we cannot directly replace $\mathbb{E}\left[|\widehat{\lambda}_a-\lambda_a||\widehat{\lambda}_b-\lambda_b|\big|\widehat{\lambda}_{\{R\}}\right]$ by an upper bound of it because this expression does not factorize. We shall prove that $\mathbb{E}\left[|\widehat{\lambda}_a-\lambda_a||\widehat{\lambda}_b-\lambda_b|\big|\widehat{\lambda}_{\{R\}}\right]$ is, up to a small loss, close to a product of expectations.

Write $N := n - n\sum_{c\in R}\widehat{\lambda}_c$, $\lambda_R := \sum_{c\in R}\lambda_c$ and $\widehat{\lambda}_R = \sum_{c\in R}\widehat{\lambda}_c$. Note that $n\widehat{\lambda}_R$ has a binomial distribution with parameters $(n,\,\lambda_R)$. Applying Bernstein inequality to $|\widehat{\lambda}_R-\lambda_R|$ we get

$$\mathbb{P}\left\{n|\widehat{\lambda}_R-\lambda_R|\geq t\right\}\leq 2\exp\left(-\frac{t^2}{4n/\sqrt{k}+2t/3}\right). \tag{58}$$

Let $\mathcal{R} = \{|\widehat{\lambda}_R-\lambda_R|\leq \frac{1}{\sqrt{n\log(k)}}\}$. Taking $t=\sqrt{n/\log(k)}$ in (58) we have that

$$\mathbb{P}(\mathcal{R})\geq 1-2e^{-\sqrt{k}/\log(k)}.$$

In what follows we assume that the event $\mathcal{R}$ is true. Take any two distinct elements a and b of $[k]\setminus R$. We shall prove that the conditional expectations $\mathbb{E}[|\widehat{\lambda}_a-\lambda_a|\widehat{\lambda}_b-\lambda_b|\widehat{\lambda}_{\{R\}}]$ are close to the products $\mathbb{E}[|\widehat{\lambda}_a-\lambda_a|\widehat{\lambda}_{\{R\}}]\,\mathbb{E}[|\widehat{\lambda}_b-\lambda_b|\widehat{\lambda}_{\{R\}}]$. It is easy to see that conditionally on $(\widehat{\lambda}_{\{R\}},\widehat{\lambda}_a)$, $n\widehat{\lambda}_b$ follows the Binomial distribution with parameters $((N-n\widehat{\lambda}_a),\lambda_b/(1-\lambda_R-\lambda_a))$. On the other hand, conditionally on $\widehat{\lambda}_{\{R\}}$, $n\widehat{\lambda}_b$ follows the Binomial distribution with parameters $(N,\lambda_b/(1-\lambda_R))$. Let $z_1,z_2,\ldots,$ be a sequence a independent Bernoulli random variables with parameters $\lambda_b/(1-\lambda_a-\lambda_R)$, $w_1,w_2\ldots,$ be an independent sequence of Bernoulli random variables with parameters $(1-\lambda_a-\lambda_R)/(1-\lambda_R)$ and $v_1,v_2,\ldots,$ be an independent sequence of Bernoulli random variables with parameters $\lambda_b/(1-\lambda_R)$. We define the following random variables

$$X := \sum_{i=1}^{N-n\widehat{\lambda}_a} z_i\ , \qquad Y := \sum_{i=1}^{N-n\widehat{\lambda}_a} z_iw_i + \sum_{i=1}^{n\widehat{\lambda}_a} v_i$$

where we use $\lambda_c \le 2/k$ and $|R| \le 2\sqrt{k}$. It is easy to see that X follows the Binomial distribution with parameters $(N - n\widehat{\lambda}_a)$ and $\lambda_b/(1-\lambda_R-\lambda_a)$ and Y follows the Binomial distribution with parameters N and $\lambda_b/(1-\lambda_R)$. Hence, we have that

$$\left|\mathbb{E}\left[|\widehat{\lambda}_a-\lambda_a||\widehat{\lambda}_b-\lambda_b|\middle|\widehat{\lambda}_{\{R\}}\right]-\mathbb{E}\left[|\widehat{\lambda}_a-\lambda_a|\widehat{\lambda}_{\{R\}}\right]\mathbb{E}[|\widehat{\lambda}_b-\lambda_b|\middle|\widehat{\lambda}_{\{R\}}]\right| \le \frac{1}{n}\mathbb{E}\left[|X-Y||\widehat{\lambda}_a-\lambda_a|\middle|\widehat{\lambda}_{\{R\}}\right]. \quad (59)$$

Relying our coupling between X and Y, we obtain

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left[|X-Y|\middle|\widehat{\lambda}_{\{R\}},\widehat{\lambda}_a\right] &\le \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{N-n\widehat{\lambda}_a} z_i(1-\omega_i)\middle|\widehat{\lambda}_{\{R\}},\widehat{\lambda}_a\right]+\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n\widehat{\lambda}_a} v_i\middle|\widehat{\lambda}_{\{R\}},\widehat{\lambda}_a\right]\\ &= \frac{N-n\widehat{\lambda}_a}{n}\frac{\lambda_b\lambda_a}{(1-\lambda_R)(1-\lambda_a-\lambda_R)}+\widehat{\lambda}_a\frac{\lambda_b}{1-\lambda_R}\\ &\le \frac{\lambda_b\lambda_a}{(1-\lambda_R)(1-\lambda_a-\lambda_R)}+\frac{\widehat{\lambda}_a\lambda_b}{1-\lambda_R}. \end{aligned} \quad (60)$$

On the other hand, conditionally on $\widehat{\lambda}_{\{R\}}$, $n\widehat{\lambda}_a$ follows the Binomial distribution with parameters $(N, \lambda_a/(1-\lambda_R))$ so that Cauchy-Schwarz inequality implies

$$\begin{aligned}\mathbb{E}\left[\left|\widehat{\lambda}_a-\lambda_a\right||\widehat{\lambda}_{\{R\}}\right] &= \mathbb{E}\left[\left|\widehat{\lambda}_a-\frac{N\lambda_a}{(1-\lambda_R)n}+\frac{N\lambda_a}{(1-\lambda_R)n}-\lambda_a\right||\widehat{\lambda}_{\{R\}}\right]\\ &\le \mathbb{E}\left[\left|\widehat{\lambda}_a-\frac{N\lambda_a}{(1-\lambda_R)n}\right||\widehat{\lambda}_{\{R\}}\right]+\left|\frac{N\lambda_a}{(1-\lambda_R)n}-\lambda_a\right|\\ &\le C\sqrt{\lambda_a/n}+\lambda_a|\lambda_R-\widehat{\lambda}_R|\\ &\le C\sqrt{\lambda_a/n}+\frac{4}{\sqrt{kn\log(k)}} \end{aligned} \quad (61)$$

where we use that $\lambda_a \le 2/k$ and the definition of the event $\mathcal{R}$. Similarly we compute

$$\mathbb{E}\left[\widehat{\lambda}_a|\widehat{\lambda}_a-\lambda_a|\middle|\widehat{\lambda}_{\{R\}}\right] \le C\left(\frac{1}{kn}+\frac{1}{k\sqrt{kn}}\right) \quad (62)$$

Plugging (60 – 62) into (59) we get

$$\begin{aligned}&\left|\mathbb{E}\left[|\widehat{\lambda}_a-\lambda_a||\widehat{\lambda}_b-\lambda_b|\middle|\widehat{\lambda}_{\{R\}}\right]-\mathbb{E}\left[|\widehat{\lambda}_a-\lambda_a|\widehat{\lambda}_{\{R\}}\right]\mathbb{E}[|\widehat{\lambda}_b-\lambda_b|\middle|\widehat{\lambda}_{\{R\}}]\right|\\ &\le \mathbb{E}\left[\frac{\lambda_b\lambda_a}{(1-\lambda_R)(1-\lambda_a-\lambda_R)}|\widehat{\lambda}_a-\lambda_a|\middle|\widehat{\lambda}_{\{R\}}\right]+\mathbb{E}\left[\widehat{\lambda}_a\frac{\lambda_b}{1-\lambda_R}|\widehat{\lambda}_a-\lambda_a|\middle|\widehat{\lambda}_{\{R\}}\right]\\ &\le C\left[\frac{1}{k^{5/2}n^{1/2}}+\frac{1}{nk^2}\right], \end{aligned}$$

where we use $\lambda_b, \lambda_a \le 2/k$. For $a = b$, (61) implies that the above difference is of order $(kn)^{-1}$. Going back to (57), we obtain that

$$\mathbb{E}\left[T_{R_1,R_2}|\widehat{\lambda}_{\{R\}}\right] \le \sup_{S\subset[k]\setminus R,\, T\subset[k]\setminus R}\sum_{a\in S, b\in T}\mathbb{E}\left[|\widehat{\lambda}_a-\lambda_a|\middle|\widehat{\lambda}_{\{R\}}\right]\mathbb{E}\left[|\widehat{\lambda}_b-\lambda_b|\middle|\widehat{\lambda}_{\{R\}}\right](\boldsymbol{Q}_{ab}-\boldsymbol{Q}^{ad,+}_{ab})+\frac{C}{\sqrt{n}}.$$

Take S^* and T^* being two sets maximizing the above expression. Then, for all $a \in S^*$ we have that $\sum_{b\in T^*}\mathbb{E}\left[|\widehat{\lambda}_b-\lambda_b|\middle|\widehat{\lambda}_R\right](\boldsymbol{Q}_{ab}-\boldsymbol{Q}^{ad,+}_{ab})$ is non-negative. As a consequence, using (61), we have that

$$\mathbb{E}\left[T_{R_1,R_2}|\widehat{\lambda}_{\{R\}}\right] \le C\sup_{S\subset[k]\setminus R,\, T\subset[k]\setminus R}\sum_{a\in S, b\in T}\left(\sqrt{\frac{\lambda_a}{n}}+\frac{4}{\sqrt{kn\log(k)}}\right)\mathbb{E}\left[|\widehat{\lambda}_b-\lambda_b|\middle|\widehat{\lambda}_{\{R\}}\right](\boldsymbol{Q}_{ab}-\boldsymbol{Q}^{ad,+}_{ab})+\frac{C'}{\sqrt{n}},$$

as soon as the event $\mathcal{R}$ holds. The same reasoning and $|\boldsymbol{Q}_{ab} - \boldsymbol{Q}^{ad,+}_{ab}| \leq 2$ leads to

$$\begin{aligned}\mathbb{E}\left[T_{R_1,R_2}|\widehat{\lambda}_{\{R\}}\right] &\leq C \sup_{S\subset[k]\setminus R, T\subset[k]\setminus R} \sum_{a\in S, b\in T} \frac{\sqrt{\lambda_b\lambda_a}}{n}\left(\boldsymbol{Q}_{ab} - \boldsymbol{Q}^{ad,+}_{ab}\right) + C''\left(\frac{k}{n\sqrt{\log(k)}} + \frac{1}{\sqrt{n}}\right) \\ &\leq C\sqrt{\frac{k}{n\log(k)}}\,,\end{aligned}$$

as soon as the event $\mathcal{R}$ holds. Going back to (56) and integrating the deviation inequality with respect to $\widehat{\lambda}_{\{R\}}$, we conclude that

$$\mathbb{P}\left[T_{R_1,R_2} \geq C\sqrt{\frac{k}{n\,\log(k)}} + 8\sqrt{\frac{2t}{n}}\right] \geq 1 - e^{-t} - \mathbb{P}[\overline{\mathcal{R}}] \geq 1 - e^{-t} - 2e^{-\sqrt{k}/\log(k)}$$

where we use $\mathbb{P}(\mathcal{R}) \geq 1 - 2e^{-\sqrt{k}/\log(k)}$. From this point the proof is identical to that of the main proof: we fix $t = 2\log(k)q + \sqrt{k}/\log(k)$ and take an union bound over all possible R_1 and R_2 to derive that

$$\max_{R_1,R_2:|R_1|\leq q,|R_2|\leq q} Z_{R_1,R_2} \leq C\sqrt{\frac{k}{n\,\log(k)}} + 4q \max_{a=1,\dots,k} |\widehat{\lambda}_a - \lambda_a|$$

on an event of probability higher than $1 - 3\exp(-\sqrt{k}/\log(k))$. Then, as in the main proof, Lemma 7 together with (39) and (52) enforce that $\|V\|^+_\square \leq C\sqrt{K/(n\log(k))}$ with probability larger than $1 - 7\exp(-\sqrt{k}/\log(k))$. By symmetry, we can find an event $\mathcal{A}$ of probability larger than $1 - 14\exp(-\sqrt{k}/\log(k))$ such that, on $\mathcal{A}$,

$$\|V\|_\square \leq C\sqrt{\frac{k}{n\,\log(k)}}\,.$$

In order to control $\mathbb{E}[\|V\|_\square]$, on the complementary event $\bar{\mathcal{A}}$ we use the rough bound

$$\|V\|_\square \leq \|V\|_1 \leq \sum_{a,b=1}^k |\widehat{\lambda}_a - \lambda_a||\widehat{\lambda}_b - \lambda_a| \leq 2\sum_{a=1}^k |\widehat{\lambda}_a - \lambda_a|$$

which implies

$$\begin{aligned}\mathbb{E}[\|V\|_\square] &\leq E[\|V\|_\square \mathbb{1}_{\mathcal{A}}] + \mathbb{E}[\|V\|_\square \mathbb{1}_{\bar{\mathcal{A}}}] \\ &\leq C\sqrt{\frac{k}{n\,\log(k)}} + 2\,\mathbb{P}^{1/2}\left[\bar{\mathcal{A}}\right]\left[\mathbb{E}\left(\sum_{a=1}^k |\widehat{\lambda}_a - \lambda_a|\right)^2\right]^{1/2} \\ &\leq C\sqrt{\frac{k}{n\,\log(k)}} + C'e^{-\sqrt{k}/(2\log(k))}\frac{\sqrt{k}}{\sqrt{n}} \leq C''\sqrt{\frac{k}{n\,\log(k)}}\end{aligned}$$

where we use (37). Together with the decomposition (53), (54) and (55), we conclude that

$$\mathbb{E}\left[\left\|(W - \widehat{W})|_{\mathcal{R}\times\mathcal{R}}\right\|_\square\right] \leq C\sqrt{\frac{k}{n\log(k)}}\,.$$

□

E Proof of Proposition 4

It is enough to prove separately the following two minimax lower bounds:

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0}[\delta_\square(\widehat{f}, \rho_n W_0)] \geq C\rho_n \sqrt{\frac{k}{n\log(k)}}\,, \tag{63}$$

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}[2]} \mathbb{E}_{W_0}[\delta_\square(\widehat{f}, \rho_n W_0)] \geq C\left(\sqrt{\frac{\rho_n}{n}} \wedge \rho_n\right). \tag{64}$$

The proof of (64) is identical to the proof of (45) in [15] so we just sketch the main idea. Fix some $0 < \epsilon \leq 1/4$. We consider W_1 to be the constant graphon with $W_1(x,y) \equiv 1/2$, and $W_2 \in \mathcal{W}[2]$ to be the 2-step graphon with $W_2(x,y) = 1/2 + \epsilon$ if $x, y \in [0,1/2)^2 \cup [1/2,1]^2$ and $W_2(x,y) = 1/2 - \epsilon$ elsewhere. Obviously, we have $\delta_\square[\rho_n W_1, \rho_n W_2] = \rho_n \epsilon$. Then, standard testing arguments [19] ensure that the minimax risk $\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}[2]} \mathbb{E}_{W_0}[\delta_\square(\widehat{f}, \rho_n W_0)]$ is at least of the order $\rho_n \epsilon$ when ϵ is chosen small enough so that the χ^2-distance $\chi^2(\mathbb{P}_{W_2}, \mathbb{P}_{W_1})$ is smaller than $1/4$. According to Lemma 4.9 in [15], this is the case when ϵ is small in front of $(\rho_n n)^{-1/2}$ which proves (64).

Henceforth, we only focus on (63). We first consider the case of k multiple of 32 and such that $k \geq C_0$ and $k \leq C_1 n$ for some Causufficiently large numerical constants C_0 and C_1. As the collections $\mathcal{W}^+[k]$ are nested this will imply (63) for all $k \in [32C_0, n]$. Afterwards, it will suffice to show (63) for $k = 2$ to prove the proposition. So, we assume that k is a multiple of 32, k is large enough and that k is small in front of n. Define $k_1 := k/2$, $M_k := \lceil 128\log(k) \rceil$, $\eta_0 := 1/16$ and $\eta_1 := 7/8$.

As for Proposition 2, we will rely on Fano's method (Lemma 3). Hence, we shall build a collection (W_u) of graphons that are well-spaced in cut distance and such that the Kullback-Leibler divergence between the associated distribution $\mathbb{P}_{W_u}$ remains small enough. All the graphons considered in this collection will be based on a $k_1 \times M_k$ matrix $\boldsymbol{B}$ such that (i) the rows of B are almost orthogonal and (ii) such that the l_1 distance between permutation and convex combinations of the columns of $\boldsymbol{B}$ are bounded from below.

Lemma 12. *For k large enough, there exists a matrix $\boldsymbol{B} \in \{-1,1\}^{k_1 \times M_k}$ satisfying the following two properties:*

(i) For any $(a,b) \in [k_1]$ with $a \neq b$, the inner product of two columns $\langle \boldsymbol{B}_{a,\cdot}, \boldsymbol{B}_{b,\cdot} \rangle$ satisfies

$$|\langle \boldsymbol{B}_{a,\cdot}, \boldsymbol{B}_{b,\cdot} \rangle| \leq M_k/4. \tag{65}$$

(ii) For any two subsets X and Y of $[k_1]$ satisfying $|X| = |Y| = \eta_0 k_1$ and $X \cap Y = \emptyset$, any labelings $\pi_1 : [\eta_0 k_1] \to X$ and $\pi_2 : [\eta_0 k_1] \to Y$, any subset Z of $[M_k]$ of size larger than $\eta_1 M_k$ and any $Z \times M_k$ stochastic matrix ω, we have

$$\sum_{a=1}^{\eta_0 k_1} \sum_{b \in Z} \Big| \boldsymbol{B}_{\pi_1(a),b} - \sum_{c \in [M_k]} \omega_{b,c} \boldsymbol{B}_{\pi_2(a),c} \Big| \geq C M_k k_1, \tag{66}$$

for some universal constant $C > 0$.

Taking $\boldsymbol{B}$ as in Lemma 12, we define the connection probability matrix $\boldsymbol{Q} := (\boldsymbol{J} + \boldsymbol{B})/2$ where $\boldsymbol{J}$ is the $k_1 \times M_k$ matrix with all entries equal to 1.

Fix some $\epsilon < 1/(8k_1)$ and denote by $\mathcal{C}_0$ the collection of vectors $u \in \{-\epsilon, \epsilon\}^{k_1}$ satisfying $\sum_{a=1}^{k_1} u_a = 0$. For any $u \in \mathcal{C}_0$, define the cumulative distribution F_u on $\{0, \ldots, k_1\}$ by the relations

$F_u(0) = 0$ and $F_u(a) = a/(2k_1) + \sum_{b=1}^{a} u_b$ for $a \in [k_1]$ and the cumulative distribution G on $\{0, \dots, M_k\}$ by $G(0) = 1/2$ and $G(b) = 1/2 + b/(2M_k)$. Note that F_u take values in $[0, 1/2]$ and G takes values in $[1/2, 1]$. Then, set $\Pi_{ab}(u) = [F_u(a-1), F_u(a)) \times [G(b-1), G(b))$ and define the graphon $W_u \in \mathcal{W}[k_1 + M_k]$ by

$$W_u(x,y) = \begin{cases} \sum_{(a,b)\in[k_1]\times[M_k]} \boldsymbol{Q}_{ab} \mathbb{1}_{\Pi_{ab}(u)}(x,y) & \text{if } x \in [0,1/2] \text{ and } y \in (1/2,1] \\ W_u(y,x) & \text{if } x \in (1/2,1] \text{ and } y \in [0,1/2] \\ 0 & \text{else .} \end{cases}$$

Note that W_u is a fairly unbalanced $(k_1 + M_k)$-step graphon: M_k of its steps have a large weight of order $1/\log(k)$. Besides, the k_1 smaller steps are slightly unbalanced as the weight of each class is either $1/k - \epsilon$ or $1/k + \epsilon$. The purpose of these M_k big steps is to make the cut distances between W_u and W_v the largest possible. see fig.

Next, we shall consider a subcollection $\mathcal{C}$ of $\mathcal{C}_0$ such that the graphons W_u with $u \in \mathcal{C}$ are well spaced. The following combinatorial result is in the spirit of the Varshamov-Gilbert lemma [19, Lemma 2.9]. It is borrowed from [15] (Lemma 4.4). For $u \in \mathcal{C}_0$, let $\mathcal{A}_u := \{a \in [k_1] : \ u_a = \epsilon\}$. Notice that, by definition of $\mathcal{C}_0$, we have $|\mathcal{A}_u| = k_1/2$ for all $u \in \mathcal{C}_0$.

Lemma 13. *There exists a subset $\mathcal{C}$ of $\mathcal{C}_0$ such that $\log|\mathcal{C}| \geq k_1/16$ and*

$$|\mathcal{A}_u \Delta \mathcal{A}_v| > k_1/4 \ . \tag{67}$$

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

Lemmas 12 and 13 are used to obtain the following lower bound on the distance $\delta_\square(W_u, W_v)$ between two distinct graphons with u and v in $\mathcal{C}$. This lemma is the main ingredient of the proof.

Lemma 14. *There exists two positive universal constants C_1 and C_2 such that if $k\epsilon \leq C_2$ then, for any $(u, v) \in \mathcal{C}$ with $u \neq v$, we have*

$$\delta_\square\big(W_u, W_v\big) \geq C_1 \frac{k\epsilon}{\sqrt{M_k}} \tag{68}$$

which implies

$$\delta_\square(\rho_n W_u, \rho_n W_v) \geq C_1 \rho_n \frac{k\epsilon}{\sqrt{M_k}}. \tag{69}$$

Note that for any u and v in $\mathcal{C}$ it is possible to build a measure-preserving transformation τ such that $W_u - W_v^\tau$ is null expect on a measurable set of Lebesgue measure of order $k\epsilon$ (see the proof of Proposition 3 in Section D for such construction). Hence, the l_1 norm of $W_u - W_v^\tau$ is of order $k\epsilon$. Lemma 14 states, that by taking the infimum over all τ and by considering the weaker norm $\|.\|_\square$, one still has a lower bound of the same order. The $M_k^{-1/2}$ factor arises as a consequence of Lemma 5. See the proof for more details.

To apply Fano's method, we need to upper bound the Kullback-Leible divergence between the distribution corresponding to any two graphon W_u and W_v with u and v in $\mathcal{C}$. Let $\mathbb{P}_{W_u}$ denote the distribution of $\boldsymbol{A}$ sampled according to the sparse graphon model (1) with $W_0 = W_u$. Since the matrix $\boldsymbol{Q}$ 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 $\xi_1, \dots, \xi_n$ in the graphon model (1) rather than from the randomness of the realization of the adjacency matrix $\boldsymbol{A}$ conditionally on $\xi_1, \dots, \xi_n$. The following lemma gives an upper bound on the Kullback-Leibler divergences $\mathcal{KL}(\mathbb{P}_{W_u}, \mathbb{P}_{W_v})$:

Lemma 15. *For all $u, v \in \mathcal{C}_0$ we have*

$$\mathcal{KL}(\mathbb{P}_{W_u}, \mathbb{P}_{W_v}) \leq 32nk_1^2\epsilon^2/3.$$

Now, choose ϵ such that $\epsilon^2 = \frac{3}{(16)^3 nk_1}$. When k is small in front of n, this choice of ϵ satisfies the conditions of Lemma 14. Then it follows from Lemmas 13 and 15 that

$$\mathcal{KL}(\mathbb{P}_{W_u}, \mathbb{P}_{W_v}) \leq \frac{1}{16}\log|\mathcal{C}|, \quad \forall\, u, v \in \mathcal{C} : u \neq v. \tag{70}$$

In view Fano's Lemma (Lemma 3), inequalities (69) and (70) imply that

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}^+[k]} \mathbb{E}_{W_0}[\delta_\square(\widehat{f}, \rho_n W_0)] \geq C\rho_n \sqrt{\frac{k}{n\log(k)}}$$

where $C > 0$ is an absolute constant. This completes the proof for k large enough.

Now we turn to the case $k = 2$. We reduce the lower bound to the problem of testing two hypotheses. Consider the matrix $\boldsymbol{B} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$. Given $u \in \{-\epsilon, +\epsilon\}$ define $F_u(0) = 0$, $F_u(1) = 1/2 + u$ and $F_u(2) = 1$. Then, we set $\Pi_{ab}(u) = [F_u(a-1), F_u(a)) \times [F_u(b-1), F_u(b))$ for any $a, b \in \{1, 2\}$ and define graphons

$$W_u(x, y) := \sum_{a,b=1}^{2} \frac{(\boldsymbol{B}_{ab} + 1)}{2} \mathbb{1}_{\Pi_{ab}(u)}(x, y).$$

For any measure preserving bijection τ, $(W_\epsilon - W_{-\epsilon}^\tau)$ is a four-step graphon. Thanks to Lemma 5, we deduce that $\delta_\square(W_\epsilon, W_{-\epsilon}) \geq C\delta_1(W_\epsilon, W_{-\epsilon})$. Then, it is not hard to see that $\delta_1(W_\epsilon, W_{-\epsilon}) \geq C'\epsilon$ so that $\delta_\square(\rho_n W_\epsilon, \rho_n W_{-\epsilon}) \geq C'\rho_n\epsilon$. Moreover, exactly as in Lemma 15, the Kullback-Leibler divergence between $\mathbb{P}_{W_\epsilon}$ and $\mathbb{P}_{W_{-\epsilon}}$ is bounded by $Cn\epsilon^2$. Taking ϵ of the order $n^{-1/2}$, this divergence is small. It is therefore impossible to reliably distinguish $\mathbb{P}_{W_\epsilon}$ from $\mathbb{P}_{W_{-\epsilon}}$ and the estimation error is at least of order $\rho_n\epsilon$. More formally, we use Theorem 2.2 from [19] to conclude that

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

where $C > 0$ is an absolute constant.

Proof of Lemma 12. Let $\boldsymbol{B}$ be a $k_1 \times M_k$ random matrix whose entries are independent Rademacher variables. We shall prove that, with positive probability, $\boldsymbol{B}$ satisfies both (65) and (66).

Fix $a \neq b$. Then, $\langle \boldsymbol{B}_{a,\cdot}, \boldsymbol{B}_{b,\cdot} \rangle$ is distributed as a sum of k_1 independent Rademacher variables. Using Hoeffding's inequality, we have that

$$\mathbb{P}\left[|\langle \boldsymbol{B}_{a,\cdot}, \boldsymbol{B}_{b,\cdot} \rangle| \geq M_k/4\right] \leq 2\exp[-M_k/32].$$

By the union bound, property (65) is satisfied for all $a \neq b$ with probability greater than $1 - k_1^2\exp[-M_k/32]$. Since $M_k \geq 128\log(k)$, for k greater than some absolute constant, this probability is greater than $3/4$.

Turning to (66), we first fix X, Y, Z, π_1, π_2, and ω. Let

$$T_{X,Y,Z,\pi_1,\pi_2,\omega} := \sum_{a=1}^{\eta_0 k_1} \sum_{b \in Z} \Big| \boldsymbol{B}_{\pi_1(a),b} - \sum_{c \in [M_k]} \omega_{b,c} \boldsymbol{B}_{\pi_2(a),c} \Big|.$$

We have that, conditionally on $(\boldsymbol{B}_{b,c})_{b\in Y,c\in[M_k]}$, $T_{X,Y,Z,\pi_1,\pi_2,\omega}$ stochastically dominates a binomial distribution with parameters $(\eta_0 k_1)\times|Z|$ and $1/2$. Then, Hoeffding's inequality yields

$$\mathbb{P}\left\{T_{X,Y,Z,\pi_1,\pi_2,\omega}\leq \eta_0 k_1|Z|/4\right\}\leq 2\exp(-\eta_0\eta_1 k_1 M_k/8).$$

Given any integer $Z\in[\eta_1 M_k, M_k]$, define Ω_Z the collection of $Z\times[M_k]$ stochastic matrices taking values in the discrete set $\{0,1/(8M_k),2/(8M_k),\dots,1\}$. Since $X,Y\subset[k_1]$ and $Z\subset M_k$, it is easy to see that the cardinality of the set of all possible tuples $(X,Y,Z,\pi_1,\pi_2,\omega)$ with $\omega\in\Omega_Z$ is bounded by

$$2^{2k_1+M_k}\left((\eta_0 k_1)!\right)^2(8M_k+1)^{M_k^2}.$$

Now, taking the union bound, we derive that, simultaneously for all such parameters,

$$T_{X,Y,Z,\pi_1,\pi_2,\omega}>\eta_0 k_1|Z|/4$$

with probability greater than $1-2^{2k_1+M_k+1}(\eta_0 k_1)!^2(8M_k+1)^{M_k^2}\exp[-\eta_0\eta_1 k_1 M_k/8]$. Using Stirling's approximation and $\eta_1 M_k\geq 64\log(k)$ we get that this probability is larger than $3/4$ for k large enough.

Finally, let us consider a general case, when matrix ω does not necessarily belong to Ω_Z. Observe that in this case, there exists a matrix $\omega'\in\Omega_Z$ such that $\max_{b\in Z}\sum_{c\in[M_k]}|\omega_{b,c}-\omega'_{b,c}|\leq 1/8$. This implies that

$$T_{X,Y,Z,\pi_1,\pi_2,\omega}\geq T_{X,Y,Z,\pi_1,\pi_2,\omega'}-\frac{|Y||Z|}{8}\geq\eta_0\eta_1 k_1 M_k/8\ .$$

We have proved that (66) holds with probability larger than $3/4$. As a consequence, $\boldsymbol{B}$ satisfies both (65) and (66) with probability larger than $1/2$. □

Proof of Lemma 14. We fix u and v, two different vectors in $\mathcal{C}$, and fix τ, a measure-preserving bijection on $[0,1]\to[0,1]$. We shall prove that for $k\epsilon$ small enough

$$\|W_u-W_v^\tau\|_\square\geq C\frac{k\epsilon}{\sqrt{M_k}}\ . \tag{71}$$

Since $\delta_\square\big(W_u,W_v\big)=\inf_\tau\|W_u(.,.)-W_v(\tau.,\tau.)\|_\square$ both (68) and (69) straightforwardly follow from (71). We denote

$$\begin{aligned}&\mathcal{B}_{11}:=\tau^{-1}\left([0,1/2]\right)\cap[0,1/2),\quad \mathcal{B}_{12}:=\tau^{-1}\left([0,1/2]\right)\cap(1/2,1],\\ &\mathcal{B}_{21}:=\tau^{-1}\left((1/2,1]\right)\cap[0,1/2],\quad \mathcal{B}_{22}:=\tau^{-1}\left((1/2,1]\right)\cap(1/2,1].\end{aligned} \tag{72}$$

Since τ is measure-preserving, we have

$$\lambda(\mathcal{B}_{11})=\lambda(\mathcal{B}_{22})=1/2-\lambda(\mathcal{B}_{12})=1/2-\lambda(\mathcal{B}_{21}). \tag{73}$$

Now, we consider three cases (i) $\lambda(\mathcal{B}_{12})\leq k_1\epsilon/64$, (ii) $k_1\epsilon/64<\lambda(\mathcal{B}_{12})\leq 1/2-k_1\epsilon/64$ and (iii) $\lambda(\mathcal{B}_{12})>1/2-k_1\epsilon/64$. In the Case (i) we shall focus on the restriction of W_u and W_v^τ on $\mathcal{B}_{11}\times\mathcal{B}_{22}$ so that these restrictions are $k_1\times M_k$-step functions. In the Case (ii), we focus on restrictions to $\mathcal{B}_{21}\times\mathcal{B}_{22}$, so that W_v^τ is constant on this restriction. In the pathological case (iii), we introduce a subset such that the restriction of W_u is a $M_k\times k_1$-step function and the restriction of W_v^τ is a $k_1\times M_k$-step function.

Case (i). We focus our attention on coordinates (x, y) in $\mathcal{B}_{11} \times \mathcal{B}_{22}$. Recall that the cumulative distribution function G is defined by $G(0) = 1/2$ and $G(b) = 1/2 + b/(2M_k)$ for $b \in [M_k]$. For any $(r, s) \in [M_k]^2$, define

$$\omega_{r,s} := \lambda\big\{[G(r-1), G(r)) \cap \tau^{-1}\left([G(s-1), G(s))\right)\big\}.$$

By definition of $\omega_{r,s}$, for any $r \in [M_k]$, we have

$$\omega_{r\bullet} := \sum_{s\in[M_k]} \omega_{r,s} \le 1/(2M_k) \quad \text{and} \quad \sum_{r,s} \omega_{r,s} = \lambda(\mathcal{B}_{22}).$$

Let $\mathcal{R}$ denote the sets of $r \in [M_k]$ such that $[G(r-1), G(r))$ has a large intersection with $\tau^{-1}([1/2, 0]$:

$$\mathcal{R} := \{r \in [M_k] \text{ s.t. } \omega_{r\bullet} \ge 3/(7M_k)\} \quad \text{and} \quad \mathcal{Y} := \cup_{r\in\mathcal{R}}[G(r-1), G(r)) \bigcap \mathcal{B}_{22}. \tag{74}$$

Denote $\bar{\mathcal{R}}$ the complementary set of $\mathcal{R}$. We have that $\lambda(\mathcal{B}_{22}) = 1/2 - \lambda(\mathcal{B}_{12}) \ge 1/2 - k_1\epsilon/64 \ge \frac{27}{56}$ for $k_1\epsilon$ small enough. Hence, it follows that

$$\frac{27}{56} \le \sum_{r,s} \omega_{r,s} = \sum_{r\in[M_k]} \omega_{r\bullet} = \sum_{r\in\mathcal{R}} \omega_{r\bullet} + \sum_{r\in\bar{\mathcal{R}}} \omega_{r\bullet} \tag{75}$$

which implies that $|\mathcal{R}| \ge 3M_k/4$ and $\lambda(\mathcal{Y}) = \sum_{r\in\mathcal{R}} \omega_{r\bullet} \ge 9/28$.

Now, denoting $\mathcal{X} := \mathcal{B}_{11}$, we define a new kernel $\overline{W}_v^\tau : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ by

$$\begin{aligned}
\overline{W}_v^\tau(x, y) &:= \sum_{r\in\mathcal{R}} \mathbb{1}_{\{y\in[G(r-1),G(r))\}} \frac{1}{\lambda\big\{[G(r-1), G(r)) \cap \mathcal{Y}\big\}} \int_{[G(r-1),G(r))\cap\mathcal{Y}} W_v(\tau(x), \tau(z))dz \\
&= \sum_{a=1}^{k_1} \sum_{r\in\mathcal{R}} \mathbb{1}_{\{y\in[G(r-1),G(r))\}} \mathbb{1}_{\{\tau(x)\in[F_v(a-1),F_v(a))\}} \sum_{s\in[M_k]} \frac{\omega_{r,s}}{\omega_{r\bullet}} \frac{(1 + \mathbf{B}_{as})}{2}.
\end{aligned} \tag{76}$$

We can view $\overline{W}_v^\tau$ as a smoothed version of the restriction of W_v^τ to $\mathcal{X} \times \mathcal{Y}$. The marginal functions $\overline{W}_v^\tau(x, .)$ are step functions with at most $|\mathcal{R}| \le M_k$ steps of the form $[G(r-1), G(r)) \cap \mathcal{B}_{22}$. Moreover, on each interval $[G(r-1), G(r)) \cap \mathcal{B}_{22}$, $\overline{W}_v^\tau(x, y)$ is equal to the mean of $W_v^\tau(x, z)$ for z ranging on this set. Equipped with this notation, we can control the cut distance between W_u and W_v^τ in terms of the l_1 distance between the restriction of W_u to $\mathcal{X} \times \mathcal{Y}$ and $\overline{W}_v^\tau$. For ease of notation, we still write W_u for for the restriction of W_u to $\mathcal{X} \times \mathcal{Y}$ when there is no ambiguity.

The following lemma provides a lower bound of the cut norm $\|W_u - W_v^\tau\|_\square$ in terms of the l_1 norm of $\|W_u - \overline{W}_v^\tau\|_1$.

Lemma 16. *For any u, v in $\mathcal{C}$ and any measure-preserving transformation τ, we have*

$$\|W_u - W_v^\tau\|_\square \ge \frac{1}{4\sqrt{2M_k}} \|W_u - \overline{W}_v^\tau\|_1 \ , \tag{77}$$

where $\overline{W}_v^\tau$ is defined in (76).

In view of Lemma 16 it is enough to control the l_1 norm $\|W_u - \overline{W}_v^\tau\|_1$. We can do it in a similar way as it is done in the proof of Lemma 4.5 in [15]. For $a \ne b$ and any $x \in [F_u(a-1), F_u(a)) \cap \mathcal{X}$

and $x' \in [F_u(b-1), F_u(b)) \cap \mathcal{X}$, the inner product between $W_u(x,.)$ and $W_v(x',.)$ satisfies

$$\begin{aligned}
&\left| \int_{\mathcal{Y}} (W_u(x,y) - 1/2)(W_v(x',y) - 1/2) dy \right| \\
&\leq \left| \int_{[1/2,1]} (W_u(x,y) - 1/2)(W_v(x',y) - 1/2) dy \right| + \frac{1}{4} \lambda\big\{ [1/2,1] \setminus \mathcal{Y} \big\} \\
&\leq \frac{1}{8M_k} \left| \sum_{c=1}^{M_k} \boldsymbol{B}_{ac} \boldsymbol{B}_{bc} \right| + \frac{5}{112} \leq \frac{1}{32} + \frac{5}{112}
\end{aligned} \tag{78}$$

where we used (65) in the last line. For any $a, b \in [k_1]$, let ψ_{ab} denote the Lebesgue measure of the set

$$[F_u(a-1), F_u(a)) \cap \tau^{-1}([F_v(b-1), F_v(b))) \cap \mathcal{X}.$$

Since τ is measure preserving, it follows that $\sum_b \psi_{ab} \leq 1/(2k_1) + u_a$ and $\sum_a \psi_{ab} \leq 1/(2k_1) + v_b$. For any $y \in \mathcal{Y}$, we set

$$h_{u,a}(y) := W_u(F_u(a-1), y) - 1/2 \quad \text{and} \quad k_{v,b}(y) := \overline{W}_v^{\tau}(\tau^{-1}(F_v(b-1)), y) - 1/2 \ .$$

Equipped with this notation, we have

$$\int_{\mathcal{X}\times\mathcal{Y}} |W_u(x,y) - \overline{W}_v^{\tau}(x,y)| dx dy = \sum_{a=1}^{k_1} \sum_{b=1}^{k_1} \psi_{a,b} \int_{\mathcal{Y}} |h_{u,a}(y) - k_{v,b}(y)| dy.$$

Now take any $a_1 \neq a_2$. By (78), $|h_{u,a}(y)| = 1/2$ and using the triangle inequality, we derive that

$$\begin{aligned}
\|h_{u,a_1} - k_{v,b}\|_1 + \|h_{u,a_2} - k_{v,b}\|_1 &\geq \|h_{u,a_1} - h_{u,a_2}\|_1 \\
&\geq \|h_{u,a_1} - h_{u,a_2}\|_2^2 \\
&\geq 2\left[\frac{1}{4}\lambda(\mathcal{Y}) - \frac{1}{32} - \frac{5}{112} \right] \geq \frac{1}{112} \ ,
\end{aligned}$$

where we used $\lambda(\mathcal{Y}) \geq 9/28$ in the last line. As a consequence, for any $b \in [k_1]$ there exists at most one $a \in [k_1]$ such that $\|h_{u,a} - k_{v,b}\|_1 < 1/224$. If such index a exists, it is denoted by $\pi(b)$. Then, it is possible to extend π as a function from $[k_1]$ to $[k_1]$. Since $\sum_{a,b} \psi_{a,b} = \lambda(\mathcal{X})$, we get

$$\begin{aligned}
\|W_u - \overline{W}_v^{\tau}\|_1 &\geq \frac{1}{224} \sum_{b=1}^{k_1} \sum_{a \neq \pi(b)} \psi_{a,b} = \frac{1}{224} \sum_{b=1}^{k_1} \left[\big(1/(2k_1) + v_b - \psi_{\pi(b),b}\big) - \left(\frac{1}{2} - \lambda[\mathcal{X}] \right) \right] \\
&= \frac{1}{224} \sum_{b=1}^{k_1} \left[\big(1/(2k_1) + v_b - \psi_{\pi(b),b}\big) - \lambda[\mathcal{B}_{1,2}] \right] \\
&\geq \frac{1}{224} \sum_{b=1}^{k_1} \left[\big(1/(2k_1) + v_b - \psi_{\pi(b),b}\big) - k_1\epsilon/64 \right] \ ,
\end{aligned}$$

since $\lambda[\mathcal{B}_{1,2}] \leq k_1\epsilon/64$. If the sum $\sum_{b=1}^{k_1} 1/(2k_1) + v_b - \psi_{\pi(b),b}$ is greater than $k_1\epsilon/32$, then (71) is satisfied. Thus, we can assume in the sequel that $\sum_{b=1}^{k_1} 1/(2k_1) + v_b - \psi_{\pi(b),b} \leq k_1\epsilon/32$.

Using that $\psi_{a,b} \leq (1/(2k_1) + u_a) \wedge (1/(2k_1) + v_b)$ and that the cardinality of the collection $\{b \in [k_1] : v_b > 0\}$ is $k_1/2$ we deduce that the collection $\{b \in [k_1] : v_b > 0,\ u_{\pi(b)} > 0$ and $\psi_{\pi(b),b} \geq 1/(2k_1)\}$ has cardinality greater than $7k_1/16$. Now, Lemma 13 implies that $|\mathcal{A}_u \cap \mathcal{A}_v| \leq 3k_1/8$

for $u \neq v \in \mathcal{C}$. Then, there exist subsets $A \subset \mathcal{A}_u$ and $B \subset \mathcal{A}_v$ of cardinality $\eta_0 k_1$ (recall that $\eta_0 = 1/16$) such that $\pi(B) = A$, $A \cap B = \emptyset$, and $\psi_{\pi(b),b} \geq 1/(2k_1)$ for all $b \in B$. The condition $\psi_{\pi(b),b} \geq 1/(2k_1)$ implies that π is injective on B. Hence,

$$\begin{aligned}
\|W_u - \overline{W}_v^\tau\|_1 &\geq \sum_{b\in B} \psi_{\pi(b),b} \int_{\mathcal{Y}} \left|h_{u,\pi(b)}(y) - k_{v,b}(y)\right| dy \\
&\geq \frac{C}{k_1 M_k} \sum_{b\in B} \sum_{c\in\mathcal{R}} \left| \boldsymbol{Q}_{\pi(b),c} - \frac{\sum_{d\in[M_k]} \omega_{b,d} \boldsymbol{Q}_{b,d}}{\omega_{b,\bullet}} \right| \\
&= \frac{C'}{k_1 M_k} \sum_{b\in B} \sum_{c\in\mathcal{R}} \left| \boldsymbol{B}_{\pi(b),c} - \frac{\sum_{d\in[M_k]} \omega_{c,d} \boldsymbol{B}_{b,d}}{\omega_{c,\bullet}} \right| ,
\end{aligned}$$

where the second inequality follows from $\psi_{\pi(b),b} \geq 1/(2k_1)$ and the fact that $h_{u,\pi(b)}$ and $k_{v,b}$ are step functions with steps larger than $3/(7M_k)$ (see (74), the definition of $\mathcal{R}$ and $\mathcal{Y}$). Finally, we apply the property (66) of $\boldsymbol{B}$ to conclude that

$$\int |W_u(x,y) - \overline{W}_v^\tau(x,y)| dxdy \geq C \geq C' k_1 \epsilon,$$

which, together with Lemma 16, proves (71).

Case (ii). Now we assume that $k_1\epsilon/64 < \lambda(\mathcal{B}_{12}) < 1/2 - k_1\epsilon/64$. Take $\mathcal{X} = \mathcal{B}_{21}$ and $\mathcal{Y} = \mathcal{B}_{22}$. We have that, on $\mathcal{X}\times\mathcal{Y}$, W_v^τ is constant and equals $1/2$. Denote U the restriction of $W_u - 1/2$ to $\mathcal{X}\times\mathcal{Y}$. Then, it follows that $\|W_u - W_v^\tau\|_\square \geq \|U\|_\square$. The kernel U is at most $k_1 \times M_k$ step function. By Lemma 5, we obtain

$$\|U\|_\square \geq \frac{1}{4\sqrt{2M_k}} \|U\|_1 = \frac{1}{8\sqrt{2M_k}} \lambda(\mathcal{X})\lambda(\mathcal{Y}) = \frac{1}{8\sqrt{2M_k}} \lambda(\mathcal{X}) \left(\frac{1}{2} - \lambda(\mathcal{X})\right) ,$$

where the last equality follows from (73). Using $\lambda(\mathcal{X}) = \lambda(\mathcal{B}_{12})$ and $x(1/2-x) \geq 1/4 \min(x, (1/2-x))$ we obtain (71).

Case (iii). Now we assume that $\lambda(\mathcal{B}_{12}) \geq 1/2 - k\epsilon/64$ and take $\mathcal{X} = \mathcal{B}_{21}$ and $\mathcal{Y} = \mathcal{B}_{12}$ so that $\lambda(\mathcal{X}) = \lambda(\mathcal{B}_{12}) \geq 1/2 - k_1\epsilon/64$. Define the smoothed kernel $\overline{W}_v^\tau : \mathcal{X}\times\mathcal{Y} \to [0,1]$ by

$$\overline{W}_v^\tau(x,y) := \sum_{a=1}^{M_r} \mathbb{1}_{\{y\in[G(a-1),G(a))\}} \frac{1}{\lambda\{[G(a-1),G(a))\cap\mathcal{Y}\}} \int_{[G(a-1),G(a))\cap\mathcal{Y}} W_v(\tau(x),\tau(z)) dz .$$

As a consequence, $\overline{W}_v^\tau$ is $M_k \times M_k$ block-constant on subsets of the form $\left(\tau^{-1}[G(a-1),G(a))\cap\mathcal{X}\right) \times \left([G(b-1),G(b))\cap\mathcal{Y}\right)$. Arguing as in the proof of Lemma 16, we derive that

$$\|W_u - W_v^\tau\|_\square \geq \frac{1}{4\sqrt{2M_k}} \|W_u - \overline{W}_v^\tau\|_1 . \tag{79}$$

For any a such that $[F_u(a-1), F_u(a)) \cap \mathcal{X} \neq \emptyset$ define the function $h_{u,a}$ on $\mathcal{Y}$ by $h_{u,a}(y) := W_u(F_u(a-1), y) - 1/2$. Arguing as in Case (i), we observe that $\|h_{u,a_1} - h_{u,a_2}\|_1 \geq 1/112$ for any $a_1 \neq a_2$. We have that the kernel $\overline{W}_v^\tau$ is a $M_k \times M_k$ step function. Hence, there exists a partition $(\mathcal{X}_b)_{b=1,\dots,M_k}$ of $\mathcal{X}$ and M_k functions $k_b(y)$ such that $\left(\overline{W}_v^\tau - 1/2\right)(x,y) = \sum_{b=1}^{M_k} \mathbb{1}_{x\in\mathcal{X}_b} k_b(y)$. Then, the triangular inequality ensures that, for any $a_1 \neq a_2$ and any $b \in [M_k]$, we have $\|h_{u,a_1} - k_b\|_1 +$

$\|h_{u,a_1} - k_b\|_1 \geq \|h_{u,a_1} - h_{u,a_2}\|_1 \geq 1/112$. As a consequence, for any $b \in [M_k]$ there exists at most one a, which we will denote by $\pi(b)$, such that $\|h_{u,\pi(b)} - k_b\|_1 \leq 1/224$.

$$\begin{aligned}
\left\|W_u - \overline{W}_v^\tau\right\|_1 &= \sum_{b=1}^{M_k}\sum_{a=1}^{k_1} \lambda(\mathcal{X}_b \cap \big[F_u(a-1), F_u(a)) \cap \mathcal{X}\big) \|h_{u,a} - k_b\|_1 \\
&\geq \frac{1}{224}\sum_{b=1}^{M_k} \lambda\Big[\mathcal{X}_b \setminus \big[F_u(\pi(b)-1), F_u(\pi(b))) \cap \mathcal{X}\big]\Big] \\
&\geq \frac{1}{224}\left[\lambda(\mathcal{X}) - \sum_{b=1}^{M_k} \frac{1}{2k_1} + u_{\pi(b)}\right] \\
&\geq \frac{1}{224}\left[\lambda(\mathcal{X}) - \frac{M_k}{2k_1} - M_k\epsilon\right] \geq C',
\end{aligned}$$

where we used $\lambda(\mathcal{X}) \geq 1/4$, $M_k/k \leq 1/8$, and that $M_k\epsilon \leq k\epsilon$ is small enough. Together with (79), we obtain the desired result (71). □

Proof of Lemma 16. We first prove that $\|W_u - \overline{W}_v^\tau\|_\square \leq \|W_u - W_v^\tau\|_\square$. Fix any measurable subset $S \subset \mathcal{X}$. Since functions $\left[W_u - \overline{W}_v^\tau\right](x,\cdot)$ are constant on each set $[G(r-1), G(r)) \cap \mathcal{Y}$, the supremum $\sup_{T\subset\mathcal{Y}} \left|\int_{S\times T} W_u(x,y) - \overline{W}_v^\tau(x,y)dxdy\right|$ is achieved by a subset T which is an union of some of $[G(r-1), G(r)) \cap \mathcal{Y}$, that is $T = \cup_{r\in\mathcal{R}'\subset\mathcal{R}}[G(r-1), G(r)) \cap \mathcal{Y}$. For such T, the definition (76) of $\overline{W}_v^\tau$ implies $\int_{S\times T} \overline{W}_v^\tau(x,y)dxdy = \int_{S\times T} W_v^\tau(x,y)dxdy$ so that

$$\sup_{T\subset\mathcal{Y}} \left|\int_{S\times T} W_u(x,y) - \overline{W}_v^\tau(x,y)dxdy\right| \leq \sup_{T\subset\mathcal{Y}} \left|\int_{S\times T} W_u(x,y) - W_v^\tau(x,y)dxdy\right| .$$

Taking the supremum over all S leads to $\|W_u - \overline{W}_v^\tau\|_\square \leq \|W_u - W_v^\tau\|_\square$. By definition of W_u and $\overline{W}_v^\tau$ we have that U is a $k_1^2 \times M_k$ step function. Then, Lemma 5 allows us to conclude

$$\|W_u - W_v^\tau\|_\square \geq \frac{1}{4\sqrt{2M_k}} \|W_u - \overline{W}_v^\tau\|_1 .$$

□

Proof of Lemma 15. The proof Lemma 15 follows the lines of the proof of of Lemma 4.3 in [15] and we give it here for completeness. For $u \in \mathcal{C}_0$, let $\zeta(u) = (\zeta_1(u), \ldots, \zeta_n(u))$ be the vector of n i.i.d. random variables with the discrete distribution on $[k_1 + M_k]$ defined by

$$\mathbb{P}[\zeta_1(u) = a] = \begin{cases} 1/(2k_1) + u_a & \text{if} \quad a \in [k_1] \\ 1/(2M_k) & \text{if} \quad k_1 + 1 \leq a \leq M_k + k_1 \end{cases} \tag{80}$$

Let $\boldsymbol{\Theta}_0$ be the $n \times n$ symmetric matrix with elements $(\boldsymbol{\Theta}_0)_{ii} = 0$ and $(\boldsymbol{\Theta}_0)_{ij} = \rho_n \boldsymbol{Q}_{\zeta_i(u),\zeta_j(u)}$ for $i \neq j$. Assume that, conditionally on $\zeta(u)$, the adjacency matrix $\boldsymbol{A}$ is sampled according to the network sequence model with such probability matrix $\boldsymbol{\Theta}_0$. Notice that in this case the observations $\boldsymbol{A}' = (\boldsymbol{A}_{ij}, 1 \leq j < i \leq n)$ have the probability distribution $\mathbb{P}_{W_u}$. Using this remark and introducing the probabilities $\alpha_{\boldsymbol{a}}(u) = \mathbb{P}[\zeta(u) = \boldsymbol{a}]$ and $p_{A\boldsymbol{a}} = \mathbb{P}[\boldsymbol{A}' = A|\zeta(u) = \boldsymbol{a}]$ for $\boldsymbol{a} \in [k_1 + M_k]^n$, we can write the Kullback-Leibler divergence between $\mathbb{P}_{W_u}$ and $\mathbb{P}_{W_v}$ in the form

$$\mathcal{KL}(\mathbb{P}_{W_u}, \mathbb{P}_{W_v}) = \sum_A \sum_{\boldsymbol{a}} p_{A\boldsymbol{a}}\alpha_{\boldsymbol{a}}(u) \log\left(\frac{\sum_{\boldsymbol{a}} p_{A\boldsymbol{a}}\alpha_{\boldsymbol{a}}(u)}{\sum_{\boldsymbol{a}} p_{A\boldsymbol{a}}\alpha_{\boldsymbol{a}}(v)}\right)$$

where the sums in $\boldsymbol{a}$ are over $[k_1 + M_k]^n$ 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

$$\mathcal{KL}(\mathbb{P}_{W_u}, \mathbb{P}_{W_v}) \leq \sum_{\boldsymbol{a}} \alpha_{\boldsymbol{a}}(u) \log\left(\frac{\alpha_{\boldsymbol{a}}(u)}{\alpha_{\boldsymbol{a}}(v)}\right) = n \sum_{a\in[k_1+M_k]} \mathbb{P}[\zeta_1(u)=a] \log\left(\frac{\mathbb{P}[\zeta_1(u)=a]}{\mathbb{P}[\zeta_1(v)=a]}\right)$$

where the last equality follows from the fact that $\alpha_{\boldsymbol{a}}(u)$ are n-product probabilities. Using (80) we get

$$\mathcal{KL}(\mathbb{P}_{W_u}, \mathbb{P}_{W_v}) \leq n \sum_{a\in[k_1]} (1/(2k_1)+u_a) \log\left(\frac{1/(2k_1)+u_a}{1/(2k_1)+v_a}\right) , \tag{81}$$

which equals $n/2$ times the Kullback-Leibler divergence between two discrete distribution. Since the Kullback-Leibler divergence is less than the chi-square divergence we obtain

$$\sum_{a\in[k_1]} (1/k_1+2u_a) \log\left(\frac{1/k_1+2u_a}{1/k_1+2v_a}\right) \leq \sum_{a\in[k_1]} \frac{4(u_a-v_a)^2}{1/k_1+2v_a} \leq 64k^2\epsilon^2/3,$$

where last inequality uses that $|v_a| \leq \epsilon \leq 1/(8k_1)$, and $|u_a - v_a| \leq 2\epsilon$. Combining this with (81) proves the lemma.

□

F Proof of Proposition 6

It is enough to prove separately the following three minimax lower bounds.

$$\inf_{\widehat{f}} \sup_{W_0\in\mathcal{W}^+[k]} \mathbb{E}_{W_0}[\delta_1(\widehat{f}, \rho_n W_0)] \geq C\rho_n\sqrt{\frac{k-1}{n}} , \tag{82}$$

$$\inf_{\widehat{f}} \sup_{W_0\in\mathcal{W}^+[k]} \mathbb{E}_{W_0}[\delta_1(\widehat{f}, \rho_n W_0)] \geq C\left(\sqrt{\rho_n}\frac{k}{n} \wedge \rho_n\right), \tag{83}$$

$$\inf_{\widehat{f}} \sup_{W_0\in\mathcal{W}[2]} \mathbb{E}_{W_0}[\delta_1(\widehat{f}, \rho_n W_0)] \geq C\left(\sqrt{\frac{\rho_n}{n}} \wedge \rho_n\right). \tag{84}$$

The proof of (82) follows from the proof of (43) in [15] using the trivial inequality

$$\|W_u(x,y) - W_v(\tau(x),\tau(y))\|_2^2 \leq \|W_u(x,y) - W_v(\tau(x),\tau(y))\|_1. \tag{85}$$

The proof of (83) follows the lines of the proof of (44) using that $\|\mathbf{B}\|_2^2 = \|\mathbf{B}\|_1$ for matrices with entries in $\{-1,1\}$. The proof of (84) is identical to the proof of (45) in [15].

References

[1] Peter J Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[2] Peter J Bickel, Aiyou Chen, and Elizaveta Levina. The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):2280–2301, 2011.

[3] Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures Algorithms*, 31(1):3–122, 2007.

[4] C. Borgs, J. T. Chayes, H. Cohn, and S. Ganguly. Consistent nonparametric estimation for heavy-tailed sparse graphs. *ArXiv e-prints*, August 2015.

[5] C. Borgs, J. T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi. Convergent sequences of dense graphs. I. Subgraph frequencies, metric properties and testing. *Adv. Math.*, 219(6):1801–1851, 2008.

[6] C. Borgs, J. T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi. Convergent sequences of dense graphs II. Multiway cuts and statistical physics. *Ann. of Math. (2)*, 176(1):151–219, 2012.

[7] Christian Borgs, Jennifer Chayes, and Adam Smith. Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems*, pages 1369–1377, 2015.

[8] Stanley H. Chan and Edoardo M. Airoldi. A consistent histogram estimator for exchangeable graph models. In *Proceedings of The 31st International Conference on Machine Learning*, pages 208–216, 2014.

[9] S. Chatterjee. Matrix estimation by universal singular value thresholding. *Ann. Statist.*, 43(1):177–214, 2015.

[10] D. S. Choi, P. J. Wolfe, and E. M. Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284, 2012.

[11] Alan Frieze and Ravi Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

[12] Chao Gao, Yu Lu, and Harrison H Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.

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

[14] Svante Janson. *Graphons, cut norm and distance, couplings and rearrangements*, volume 4 of *New York Journal of Mathematics. NYJM Monographs*. State University of New York, University at Albany, Albany, NY, 2013.

[15] O. Klopp, A. B. Tsybakov, and N. Verzelen. Oracle inequalities for network models and sparse graphon estimation. *to appear in The Annals of Statistics*, 2016.

[16] László Lovász. *Large networks and graph limits*, volume 60 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2012.

[17] László Lovász and Balázs Szegedy. Limits of dense graph sequences. *J. Combin. Theory Ser. B*, 96(6):933–957, 2006.

[18] S. J. Szarek. On the best constants in the Khinchin inequality. *Studia Math.*, 58(2):197–208, 1976.

[19] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

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

[21] Jiaming Xu, Laurent Massoulié, and Marc Lelarge. Edge label inference in generalized stochastic block models: from spectral theory to impossibility results. *COLT*, 2014.

[22] Justin Yang, Christina Han, and Edoardo M. Airoldi. Nonparametric estimation and testing of exchangeable graph models. In *AISTATS*, 2014.