Information and Inference: A Journal of the IMA (2018) **7**, 169–181
doi:10.1093/imaiai/iax010
Advance Access Publication on 23 August 2017

When is non-trivial estimation possible for graphons and stochastic block models?‡

AUDRA MCMILLAN†
Department of Mathematics, University of Michigan, 530 Church St, Ann Arbor, MI 48109, USA
†Corresponding author: amcm@umich.edu

AND

ADAM SMITH
Department of Computer Science and Engineering, Pennsylvania State University, 342 Information Sciences and Technology Building, University Park, PA 16802, USA
asmith@cse.psu.edu

[Received on 29 May 2016; accepted on 29 May 2017]

Block graphons (also called stochastic block models) are an important and widely studied class of models for random networks. We provide a lower bound on the accuracy of estimators for block graphons with a large number of blocks. We show that, given only the number k of blocks and an upper bound ρ on the values (connection probabilities) of the graphon, every estimator incurs error $\Omega\left(\min\left(\rho, \sqrt{\frac{\rho k^2}{n^2}}\right)\right)$ in the δ_2 metric with constant probability for at least some graphons. In particular, our bound rules out any non-trivial estimation (that is, with δ_2 error substantially less than ρ) when $k \geq n\sqrt{\rho}$. Combined with previous upper and lower bounds, our results characterize, up to logarithmic terms, the accuracy of graphon estimation in the δ_2 metric. A similar lower bound to ours was obtained independently by Klopp *et al.*

Keywords: network data; graphons; stochastic block models; minimax rates.

1. Introduction

Networks and graphs arise as natural modelling tools in many areas of science. In many settings, particularly in social networks, networks display some type of community structure. In these settings, one may consider the nodes of the graph as belonging to one of k communities, and two nodes are connected with a probability that depends on the communities they belong to. This type of structure is captured in the *k-block graphon* model, also known as stochastic block models. The more communities we allow in the model (or 'types' of nodes we consider), the richer the model becomes and the better we can hope to describe the real world. One can think of a general *graphon* model as an ∞-block graphon, where each node is given a label in $[0, 1]$ rather than $\{1, \ldots, k\}$.

Given an observed network, graphon estimation is the problem of recovering the graphon model from which the graph was drawn. In this article, we are concerned with the fundamental limits of graphon estimation for block graphons. That is, given an n-node network that was generated from a k-block graphon, how accurately can you recover the graphon? We consider the 'non-parametric' setting, where

‡A preliminary version of this work appears as ArXiv report arXiv:1604.01871 [math.ST]. This work was done while the author was visiting Pennsylvania State University.

k may depend on n. Our lower bounds apply even to estimation algorithms that know the true number of blocks k, though this quantity typically needs to be estimated.

In many real-world networks, the average degree of the network is small compared with the number of nodes in the network. Graphons whose expected average degree is linear in n are called dense, while graphons whose expected average degree is sublinear in n are referred to as sparse. In this work, we prove a new lower bound for graphon estimation for sparse networks. In particular, our results rule out *non-trivial* estimation for very sparse networks (roughly, where $\rho = O(k^2/n^2)$). An estimator is non-trivial if its expected error is significantly better than an estimator that ignores the input and always outputs the same model. It follows from recent work [1–3] that non-trivial estimation is impossible when $\rho = O(1/n)$. Ours is the first lower bound that rules out non-trivial graphon estimation for large k. Previous work by Klopp *et al.* [4] provides other lower bounds on graphon estimation that are tight in several regimes. In recent work [5] that is concurrent to ours, the same authors provide a similar bound to the one presented here.

Block graphon models were introduced by Hoff *et al.* [6] under the name latent position graphs. Graphons play an important role in the theory of graph limits (see [7] for a survey) and the connection between the graph model, and convergent graph sequences has been studied in both the dense and the sparse settings [8–11]. Estimation for stochastic block models with a fixed number of blocks was introduced by Bickel and Chen [12], while the first estimation of the general model was proposed by Bickel *et al.* [13]. Since then, many graphon estimation methods, with an array of assumptions on the graphon, have been proposed [14–24]. Gao *et al.* [19] provide the best known upper bounds in the dense setting, while Wolfe and Olhede [23], Borgs *et al.* [25] and Klopp *et al.* [4] give upper bounds for the sparse case.

1.1 *Graphons*

DEFINITION 1 (Bounded graphons and W-random graphs) A (bounded) *graphon* W is a symmetric, measurable function $W : [0,1]^2 \to [0,1]$. Here, symmetric means that $W(x,y) = W(y,x)$ for all $(x,y) \in [0,1]^2$.

For any integer n, a graphon W defines a distribution on graphs on n vertices as follows: Firstly, select n labels $\ell_1, \ldots, \ell_n$ uniformly and independently from $[0,1]$ and form an $n \times n$ matrix H, where $H_{ij} = W(\ell_i, \ell_j)$. We obtain an unlabelled, undirected graph G by connecting the ith and jth nodes with probability H_{ij} independently for each (i,j). The resulting random variable is called a *W-random graph* and denoted $G_n(W)$.

For $\rho \geq 0$, we say a graphon is ρ-bounded if W takes values in $[0,\rho]$ (that is, $\|W\|_\infty \leq \rho$).

We denote the set of graphs with n nodes by $\mathcal{G}_n$, the set of graphons by $\mathcal{W}$ and the set of ρ-bounded graphons by $\mathcal{W}_\rho$. If W is ρ-bounded, then the expected number of edges in $G_n(W)$ is at most $\rho\binom{n}{2} = O(\rho n^2)$. In the case that ρ depends on n and $\lim_{n\to\infty} \rho \to 0$, we obtain a sparse graphon.

We consider the estimation problem: given parameters n and ρ, as well as a graph $G \sim G_n(W)$ generated from an unknown ρ-bounded graphon W, how well can we estimate W?

A natural goal is to design estimators that produce a graphon $\hat{W}$ that is close to W in a metric such as L_2. This is not possible, because there are many graphons that are far apart in L_2, but that generate the same probability distribution on graphs. If there exists a measure-preserving map $\phi : [0,1] \to [0,1]$, such that $W(\phi(x),\phi(y)) = W'(x,y)$ for all $x, y \in [0,1]$, then $G_n(W)$ and $G_n(W')$ are identically distributed. The converse is true if we instead only require $W(\phi(x),\phi(y)) = W'(x,y)$ almost everywhere. Thus, we wish to say that $\hat{W}$ approaches the *class* of graphons that generate $G_n(W)$. To this end, we use the

following metric on the set of graphons,

$$\delta_2(W, W') = \inf_{\substack{\phi:[0,1]\to[0,1] \\ \text{measure-preserving}}} \left\| W_\phi - W' \right\|_2, \tag{1.1}$$

where $W_\phi(x, y) = W(\phi(x), \phi(y))$ and ϕ ranges over all measurable, measure-preserving maps. Two graphons W and W' generate the same probability distribution on the set of graphs if and only if $\delta_2(W, W') = 0$ (see Lovász [7], for example).

Existing upper bounds for graphon estimation are based on algorithms that produce graphons of a particular form, namely *block graphons*, also called *stochastic block models* (even when it is not known that the true graphon is a block graphon).

DEFINITION 2 (k-block graphon (stochastic block models)) For $k \in \mathbb{N}$, a graphon is a *k-block graphon* if there exists a partition of $[0, 1]$ into k measurable sets $I_1, \ldots, I_k$, such that W is constant on $I_i \times I_j$ for all i and j.

We can associate a graphon of this form to every square matrix. Given a $k \times k$ symmetric matrix M, let $W[M]$ denote the k-block graphon with blocks $I_i = \left(\frac{i-1}{k}, \frac{i}{k}\right]$ that takes the value M_{ij} on $I_i \times I_j$.

1.2 *Main result*

We are concerned with the problem of estimating a graphon, W, given a graph sampled from $G_n(W)$. A graphon estimator is a function $\hat{W} : \mathcal{G}_n \to \mathcal{W}$ that takes as input an n node graph, that is generated according to W, and attempts to output a graphon that is close to W. The main contribution of this article is the development of the lower bound

$$\inf_{\hat{W}} \sup_{W} \mathop{\mathbb{E}}_{G\sim G_n(W)} [\delta_2(\hat{W}(G), W)] \geq \Omega\left(\min\left(\rho, \sqrt{\frac{\rho k^2}{n^2}}\right)\right). \tag{1.2}$$

Combined with previous work, we can give the following lower bound on the error of graphon estimators.

THEOREM 3 For any positive integer $2 \leq k \leq n$ and $0 < \rho \leq 1$,

$$\inf_{\hat{W}} \sup_{W} \mathop{\mathbb{E}}_{G\sim G_n(W)} \left[\delta_2(\hat{W}(G), W)\right] \geq \Omega\left(\min\left(\rho, \rho\sqrt[4]{\frac{k}{n}} + \sqrt{\frac{\rho k^2}{n^2}}\right)\right), \tag{1.3}$$

where $\inf_{\hat{W}}$ is the infimum over all estimators $\hat{W} : G_n \to \mathcal{G}$ and $\sup_W$ is the supremum over all k-block, ρ-bounded graphons. If ρn is non-decreasing and there exists a constant $c > 0$, such that $\rho n \geq c$, then

$$\inf_{\hat{W}} \sup_{W} \mathop{\mathbb{E}}_{G\sim G_n(W)} \left[\delta_2(\hat{W}(G), W)\right] \geq \Omega\left(\min\left(\rho, \rho\sqrt[4]{\frac{k}{n}} + \sqrt{\frac{\rho k^2}{n^2}} + \sqrt{\frac{\rho}{n}}\right)\right).$$

Note that k and ρ may depend on n. That is, the theorem holds if we consider sequences ρ_n and k_n. Our result improves on previously known results when $\rho = O\left(\left(\frac{k}{n}\right)^{3/2}\right)$—that is, when the graphs

produced by the graphon are sparse. The upper bound

$$\inf_{\hat{W}} \sup_{W} \mathop{\mathbb{E}}_{G\sim G_n(W)} \left[\delta_2(\hat{W}(G), W)\right] \leq O\left(\min\left(\rho,\ \rho\sqrt[4]{\frac{k}{n}} + \sqrt{\frac{\rho k^2}{n^2}} + \sqrt{\frac{\rho \log k}{n}}\right)\right) \tag{1.4}$$

by Klopp *et al.* [4] implies that our lower bound is almost tight. In particular, if k is constant and ρ is within the designated range then the lower bound in Theorem 3 is tight.

When $\rho = O\left(\frac{k^2}{n^2}\right)$, Theorem 3 implies that the error is $\Omega(\rho)$, which is the error achieved by the trivial estimator $\hat{W} = 0$. That is, in the sparse setting, the trivial estimator achieves the optimal error. To the authors' knowledge, this is the first result that completely rules out non-trivial estimation in the case where k is large. Recent concurrent work ([5]) provides similar bounds.

The bound

$$\inf_{\hat{W}} \sup_{W} \mathop{\mathbb{E}}_{G\sim G_n(W)} [\delta_2(\hat{W}, W)] \geq \Omega\left(\rho\sqrt[4]{\frac{k}{n}}\right) \tag{1.5}$$

is due to previous work of Klopp *et al.* [4] and the bound

$$\inf_{\hat{W}} \sup_{W} \mathop{\mathbb{E}}_{G\sim G_n(W)} [\delta_2(\hat{W}, W)] \geq \Omega\left(\min\left(\rho, \sqrt{\frac{\rho}{n}}\right)\right) \tag{1.6}$$

for constant k follows from the results of Mossel *et al.* [1] and Banerjee [26]. We give details on how to derive (1.6) from their results in the Appendix.

1.3 *Techniques: combinatorial lower bounds for δ_p*

Our proof of the main theorem will involve Fano's lemma. As such, during the course of the proof, we will need to lower bound the packing number, with respect to δ_2, of a large set of k-block graphons. While easily upper bounded, little is known about lower bounds on δ_2. To the authors' knowledge, this work gives the first lower bound for the packing number of $\mathcal{W}_\rho$ with respect to δ_2. We will also give a combinatorial lower bound for the δ_2 metric that is easier to handle than the metric itself.

To understand our technical contributions, it helps to first understand a problem related to graphon estimation, namely that of estimating the matrix of probabilities H. Existing algorithms for graphon estimation are generally analysed in two phases: firstly, one shows that the estimator $\hat{W}$ is close to the matrix H (in an appropriate version of the δ_2 metric) and then uses (high probability) bounds on $\delta_2(W, W[H])$ to conclude that $\hat{W}$ is close to W. Klopp *et al.* [4] show tight upper and lower bounds on estimation of H. One can think of our lower bound as showing that the lower bounds on estimation of H can be transferred to the problem of estimating W.

The main technical difficulty lies in showing that a given pair of matrices A, B lead to graphons that are far apart in the δ_2 metric. Even if A, B are far apart in, say, ℓ_2, they may lead to graphons that are close in δ_2. For consistency with the graphon formalism, we normalize the ℓ_2 metric on $k \times k$ matrices, so that it agrees with the L_2 metric on the corresponding graphons. For a $k \times k$ matrix A,

$$\|A\|_2 \stackrel{\text{def}}{=} \left(\frac{1}{k^2}\sum_{i,j\in[k]} A_{ij}^2\right)^{1/2} = \|W[A]\|_2. \tag{1.7}$$

As an example of the discrepancy between the ℓ_2 and δ_2 metrics, consider the matrices

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \qquad \text{and} \qquad B = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The matrices A and B have positive distance in the ℓ_2 metric, $\|A - B\|_2 = \frac{2}{3}$, but $\delta_2(W[A], W[B]) = 0$.

One can get an *upper bound* on $\delta_2(W[A], W[B])$ by restricting attention in the definition of δ_2 to functions ϕ that permute the blocks I_i. This leads to the following metric on $k \times k$ matrices, which minimizes over permutations of the rows and columns of one of the matrices:

$$\hat{\delta}_2(A, B) \stackrel{\text{def}}{=} \min_{\sigma \in \mathcal{S}_k} \|A_\sigma - B\|_2, \tag{1.8}$$

where A_σ is the matrix with entries $(A_\sigma)_{ij} = A_{\sigma(i),\sigma(j)}$. This metric arises in other work (e.g. [7]), and it is well known that

$$\delta_2(W[A], W[B]) \leq \hat{\delta}_2(A, B). \tag{1.9}$$

To prove lower bounds, we consider a new metric on matrices, in which we allow the rows and columns to be permuted separately. Specifically, let

$$\hat{\hat{\delta}}_2(A, B) \stackrel{\text{def}}{=} \min_{\sigma,\tau \in \mathcal{S}_k} \left\|A_{\sigma,\tau} - B\right\|_2, \tag{1.10}$$

where $A_{\sigma,\tau}$ is the $k \times k$ matrix with entries $(A_{\sigma,\tau})_{ij} = A_{\sigma(i),\tau(j)}$.

LEMMA 1 (Lower bound for δ_2) For every two $k \times k$ matrices A, B,

$$\hat{\hat{\delta}}_2(A, B) \leq \delta_2(W[A], W[B]). \tag{1.11}$$

Because $\hat{\hat{\delta}}_2$ is defined 'combinatorially' (that is, it involves minimization over a discrete set of size about $2^{2k \ln k}$, instead of over all measure-preserving injections), it is fairly easy to lower bound $\hat{\hat{\delta}}_2(A, B)$ for random matrices A, B using the union bound.

In particular, it allows us to give bounds on the packing number of $\mathcal{W}_\rho$ with respect to the δ_2 metric. If (Ω, d) is a metric space, $\epsilon > 0$ and $T \subset \Omega$, then we define the ϵ-packing number of T to be the largest number of disjoint balls of radius ϵ that can fit in T, denoted by $\mathcal{M}(\epsilon, T, d)$. The following proposition will be proved after the proof of Theorem 3.

PROPOSITION 1 There exists $C > 0$ such that the $C\rho$-packing number of $\mathcal{W}_\rho$, equipped with δ_2, is $2^{\Omega(k^2)}$, that is $\mathcal{M}(C\rho, \mathcal{W}_\rho, \delta_2) = 2^{\Omega(k^2)}$.

Finally, we note that these techniques extend directly to the δ_p metric, for $p \in [1, \infty]$. That is, we may define δ_p, $\hat{\delta}_p$ and $\hat{\hat{\delta}}_p$ analogously to the definitions above and obtain the bounds

$$\hat{\hat{\delta}}_p(A, B) \leq \delta_p(W[A], W[B]) \leq \hat{\delta}_p(A, B), \tag{1.12}$$

along with similar lower bounds on the packing number.

1.4 *Related work*

Work on graphon estimation falls broadly into two categories: estimating the matrix H and estimating the graphon W. When estimating H, the aim is to produce a matrix that is close in the ℓ_2 metric to the true matrix of probabilities H that was used to generate the graph G. When estimating the graphon, our aim is to minimize the δ_2 distance between the estimate and the true underlying graphon W that was used to generate G.

Gao *et al.* [19] studied the problem of estimating the matrix of probabilities H, given an instance chosen from W when $\rho = 1$. They proved the following minimax rate for this problem when W is a k-block graphon:

$$\inf_{\hat{M}} \sup_{H} \mathop{\mathbb{E}}_{G\sim G_n(H)} \left[\frac{1}{n^2} \left\| \hat{M}(G) - H \right\|_2 \right] \asymp \sqrt{\frac{k^2}{n^2} + \frac{\log k}{n}}, \tag{1.13}$$

where the infinimum is over all estimators $\hat{M}$ from G_n to the set of symmetric $n \times n$ matrices, the supremum is over all probability matrices H generated from k-block graphons. Klopp *et al.* [4] extended this result to the sparse case, proving that for all $k \leq n$ and $0 < \rho \leq 1$,

$$\inf_{\hat{M}} \sup_{H} \mathop{\mathbb{E}}_{G\sim G_N(H)} \left[\frac{1}{n^2} \left\| \hat{M}(G) - H \right\|_2 \right] \geq \Omega \left(\min \left(\sqrt{\rho \left(\frac{k^2}{n^2} + \frac{\log k}{n} \right)}, \rho \right) \right), \tag{1.14}$$

where the supremum is over all probability matrices H generated from k-block, ρ-bounded graphons.

Klopp *et al.* [4, Corollary 3] also studied the problem of estimating the graphon W. They proved that equation (1.4) holds for any k-block, ρ-bounded graphon, W, with $k \leq n$. They also exhibited the first lower bound (known to us) for graphon estimation using the δ_2 metric. They proved that equation (1.5) holds for $\rho > 0$ and $k \leq n$.

The related problems of distinguishing a graphon with $k > 1$ from an Erdös–Rényi model with the same average degree (called the distinguishability problem) and reconstructing the communities of a given network (called the reconstruction problem) have also been widely studied. This problem is closely related to the problem of estimating H. Recent work by Mossel *et al.* [1] and Neeman and Netrapalli [3] establish conditions under which a k-block graphon is mutually contiguous to the Erdös–Rényi model with the same average degree. Contiguity essentially implies that no test could ever definitely determine which of the two graphons a given sample came from. There is a large body of work on algorithmic and statistical problems in this area, and we have only cited work that is directly relevant here.

2. Lower bound for the δ_2 metric

As mentioned earlier, the main technical contribution of this article is lower bounding the δ_2 metric by the more combinatorial $\hat{\delta}_2$ metric. In this section, we will prove the inequality given in Lemma 1.

PROPOSITION 2 Let W, W' be k-block graphons with blocks $I_i = \left[\frac{i-1}{k}, \frac{i}{k}\right)$ and $\pi : [0,1] \to [0,1]$ be a measure-preserving map. Then there exists a probability distribution $\mathbb{P}$ on $\mathcal{S}_k$ such that

$$\left\| W_\pi - W' \right\|_2^2 = \mathop{\mathbb{E}}_{\sigma,\tau\sim\mathbb{P}} \left[\left\| W_{\sigma,\tau} - W' \right\|_2^2 \right], \tag{2.1}$$

where the expectation is taken over σ, τ selected independently according to $\mathbb{P}$.

Proof. Let $a_i = \frac{i-1}{k}$ and $p_{ij} = \mu(I_i \cap \pi^{-1}(I_j))$. Now, consider a $k \times k$ matrix P with $P_{ij} = kp_{ij}$. Noting that $\sum_{j=1}^{k} p_{ij} = \mu(I_i) = 1/k$ and $\sum_{i=1}^{k} p_{ij} = \mu(\pi^{-1}(I_j)) = 1/k$, we can see that P is doubly stochastic, that is, the rows and columns of P sum to 1. Berkhoff's theorem states that any doubly stochastic matrix can be written as a convex combination of permutation matrices. Therefore, we have a probability distribution $\mathbb{P}$ on $\mathcal{S}_k$ such that $P = \sum_{\sigma \in \mathcal{S}_k} \mathbb{P}(\sigma)\sigma$ and $\sum_{\sigma \in \mathcal{S}_k} \mathbb{P}(\sigma) = 1$ and

$$\mathbb{P}(\sigma(i) = j) = \sum \{\mathbb{P}(\sigma) \mid \sigma(i) = j\} = P_{ij} = kp_{ij}. \tag{2.2}$$

Taking expectations over σ, τ selected independently from $\mathbb{P}$,

$$\begin{aligned}
\mathbb{E}\left[\left\|W_{\sigma,\tau} - W'\right\|_2^2\right] &= \sum_{\sigma,\tau} \mathbb{P}(\sigma)\mathbb{P}(\tau) \sum_{i,j} \frac{1}{k^2} (W(a_{\sigma(i)}, a_{\tau(j)}) - W'(a_i, a_j))^2 \\
&= \sum_{i,i',j,j'} \frac{1}{k^2} \mathbb{P}(\sigma(i) = i')\mathbb{P}(\tau(j) = j')(W(a_i, a_j) - W'(a_{i'}, a_{j'}))^2 \\
&= \sum_{i,i',j,j'} p_{ii'}p_{jj'}(W(a_i, a_j) - W'(a_{i'}, a_{j'}))^2 \\
&= \left\|W_\pi - W'\right\|_2^2.
\end{aligned}$$

□

Proof of Lemma 1. Proposition 2 implies that for all measure-preserving maps $\pi : [0, 1] \to [0, 1]$ and matrices A and B we have

$$\|W[A]_\pi - W[B]\|_2 \geq \inf_{\sigma,\tau \in \mathcal{S}_k} \left\|W[A]_{\sigma,\tau} - W[B]\right\|_2 = \inf_{\sigma,\tau \in \mathcal{S}_k} \left\|A_{\sigma,\tau} - B\right\|_2 = \hat{\hat{\delta}}_2(A, B). \tag{2.3}$$

Since this is true for any π, we have $\delta_2(W[A], W[B]) \geq \hat{\hat{\delta}}_2(A, B)$. □

3. Proof of main theorem

To prove the main theorem, we will use Fano's lemma to find a constant that lower bounds the probability that the estimation exceeds $\min\left(\rho, \sqrt{\frac{\rho k^2}{n^2}}\right)$, which then implies the appropriate lower bound on the expected δ_2 error. To that end, we aim to find a large set, T, of k-block graphons whose Kullback–Leibler (KL) diameter and ϵ-packing number with respect to δ_2 with $\epsilon = \min\left(\rho, \sqrt{\frac{\rho k^2}{n^2}}\right)$ can be bounded. Our proof is inspired by that of Gao *et al.* [19].

Suppose p, q are probability distributions on the same space. Then the (KL) divergence of p and q is defined by $D(p\|q) = \int \left(\log \frac{dp}{dq}\right) dp$. For a collection T of probability distributions, the KL diameter is defined by

$$\mathrm{d}_{\mathrm{KL}}(T) = \sup_{p,q \in T} D(p\|q). \tag{3.1}$$

The following version of Fano's lemma is found in [27]. Recall that $\log \mathcal{M}(\epsilon, T, d)$ is the size of the largest ϵ-packing of T (with distance d).

LEMMA 2 (Fano's inequality) Let (Ω, d) be a metric space and $\{\mathbb{P}_\theta \,|\, \theta \in \Omega\}$ be a collection of probability measures. For any totally bounded $T \subset \Omega$ and $\epsilon > 0$,

$$\inf_{\hat{\theta}} \sup_{\theta \in \Omega} \mathbb{P}_\theta \left(\mathrm{d}^2(\hat{\theta}(X), \theta) \geq \frac{\epsilon^2}{4} \right) \geq 1 - \frac{\mathrm{d}_{\mathrm{KL}}(P_T) + 1}{\log \mathcal{M}(\epsilon, T, d)}, \tag{3.2}$$

where the infimum is over all estimators and $P_T = \{\mathbb{P}_\theta \mid \theta \in T\}$.

The following lemma gives us a way to easily upper bound the KL divergence between the distributions induced by two different graphons.

LEMMA 3 For any graphons $\frac{1}{4} \leq W, W' \leq \frac{3}{4}$, we have

$$D(G_n(W) \| G_n(W')) \leq 8n^2 \left\| W - W' \right\|_2^2. \tag{3.3}$$

Proof. Let T be a variable denoting the choice of labels, so

$$\mathbb{P}_{G_n(W)}(G) = \int_{\ell \in [0,1]^n} \mathbb{P}_T(\ell) \mathbb{P}_{G_n(W)}(G | T = \ell) \, \mathrm{d}\ell. \tag{3.4}$$

Now,

$$\begin{aligned} D(G_n(W) \| G_n(W')) &= \sum_{G \in G_n} \mathbb{P}_{G_n(W)}(G) \ln \left(\frac{\mathbb{P}_{G_n(W)}(G)}{\mathbb{P}_{G_n(W')}(G)} \right) \\ &\leq \sum_{G \in G_n} \int_{\ell \in [0,1]^n} \mathbb{P}_T(\ell) \mathbb{P}_{G_n(W)}(G | T = \ell) \ln \left(\frac{\mathbb{P}_{G_n(W)}(G | T = \ell)}{\mathbb{P}_{G_n(W')}(G | T = \ell)} \right) \mathrm{d}\ell \\ &= \int_{\ell \in [0,1]^n} \mathbb{P}_T(\ell) D(\mathbb{P}_{G_n(W)}(\cdot | T = \ell)) \| \mathbb{P}_{G_n(W')}(\cdot | T = \ell) \, \mathrm{d}\ell, \end{aligned}$$

where the inequality follows from the log-integral inequality [28, Theorem 30.12.4]. Now, the probability density function of T is the constant function 1, so it follows from Gao *et al.* [19, Proposition 4.2] that

$$\begin{aligned} D(G_n(W) \| G_n(W')) &\leq 8 \int_{\ell \in [0,1]^n} \sum_{i,j=1}^{n} (W(\ell_i, \ell_j) - W'(\ell_i, \ell_j))^2 \, \mathrm{d}\ell \\ &= 8 \sum_{i,j=1}^{n} \int_{\ell \in [0,1]^n} (W(\ell_i, \ell_j) - W'(\ell_i, \ell_j))^2 \, \mathrm{d}\ell \\ &\leq 8n^2 \int_{[0,1]^2} (W(x, y) - W'(x, y))^2 \, \mathrm{d}x \, \mathrm{d}y \\ &= 8n^2 \left\| W - W' \right\|_2^2. \end{aligned}$$

□

Recall that we are aiming to define a large set of k-block matrices that are close in KL divergence, but that are ϵ-far apart with respect to δ_2 (with $\epsilon = \min(\rho, \sqrt{\frac{\rho k^2}{n^2}})$). The following lemma shows that there exists a large set of matrices that are pairwise far in Hamming distance, even after every possible permutation of the rows and columns. We will use this in the proof of Theorem 3 to define a large class of k-block graphons that are pairwise far in the $\hat{\hat{\delta}}_2$ metric and hence the δ_2 metric. This gives us a bound on packing number.

LEMMA 4 There exists a set S of symmetric $k \times k$ binary matrices such that $|S| = 2^{\Omega(k^2)}$ and, for every $B, B' \in S$ and $\sigma, \tau \in \mathcal{S}_k$, we have $\mathrm{Ham}(B_{\sigma,\tau}, B') = \Omega(k^2)$.

Proof. Fix permutations σ and τ, and consider two randomly chosen symmetric binary matrices B, B'. For $i \le j$, let $X_{ij} = 1$ if $B_{\sigma(i),\tau(j)} = B'_{i,j}$ and 0 otherwise so X_{ij} is a Bernoulli random variable with $\mathbb{E}[X_{ij}] = \frac{1}{2}$. Thus, by a Chernoff bound,

$$\mathbb{P}\left(\mathrm{Ham}(B_{\sigma,\tau}, B') < \frac{k^2}{6}\right) = \mathbb{P}\left(\sum_{i\le j} X_{ij} \le \frac{k^2}{6}\right) \le e^{\frac{-2\left(\frac{k^2}{6} - \frac{1}{2}\binom{k}{2}\right)^2}{\binom{k}{2}}}. \tag{3.5}$$

Therefore, for randomly chosen B, B',

$$\mathbb{P}\left(\exists \sigma, \tau \text{ s.t. } \mathrm{Ham}(B_{\sigma,\tau}, B') < \frac{k^2}{6}\right) \le e^{\frac{-2\left(\frac{k^2}{6} - \frac{1}{2}\binom{k}{2}\right)^2}{\binom{k}{2}}} (k!)^2 = 2^{-\Omega(k^2)}. \tag{3.6}$$

For a constant $c > 0$, consider the process that selects 2^{ck^2} binary matrices $\{B_i\}_i$ uniformly at random uniformly at random. The probability that all pairs are at Hamming distance at least $k^2/6$ is at least $1 - 2^{2ck^2} 2^{-\Omega(k^2)}$. Selecting c sufficiently small, we get that at least one such set S exists. □

We are not aware of an explicit construction of a large family of matrices that are far apart in $\hat{\hat{\delta}}_2$ metric; we leave such a construction as an open problem.

We now proceed to the proof of Theorem 3. We will use Lemma 4 to define a set T with packing number $2^{\Omega(k^2)}$. The elements of T are all close in $\|\cdot\|_\infty$ norm, so using Lemma 3 we get a bound on the KL diameter. We then directly apply these bounds via Fano's lemma.

THEOREM 4 For any positive integer $k \le n$ and $0 < \rho \le 1$,

$$\inf_{\hat{W}} \sup_{W} \mathop{\mathbb{E}}_{G\sim G_n(W)} \left[\delta_2(\hat{W}(G), W)\right] \ge \Omega\left(\min\left(\rho, \sqrt{\frac{\rho k^2}{n^2}}\right)\right), \tag{3.7}$$

where $\inf_{\hat{W}}$ is the infimum over all estimators $\hat{W} : G_n \to \mathcal{G}$ and $\sup_W$ is the supremum over all k-block, ρ-bounded graphons.

Proof. Let S be a set satisfying the conditions of Lemma 4 and let $\eta = \min(1, \frac{k}{n\sqrt{\rho}})$. For $B \in S$, define

$$Q_B = \rho\left[\frac{1}{2}\mathbf{1} + c\eta(2B - \mathbf{1})\right], \tag{3.8}$$

where $\mathbf{1}$ is the all 1's matrix and c is some constant that we will choose later. That is, $(Q_B)_{ij} = \rho[\frac{1}{2} + c\eta]$ if $B_{ij} = 1$ and $(Q_B)_{ij} = \rho[\frac{1}{2} - c\eta]$ if $B_{ij} = 0$. Let $T = \{W[Q_B] \mid B \in S\}$. Using Lemma 3, we conclude that for all $W, W' \in T$, we have

$$D(G_n(W) \| G_n(W')) \leq 8n^2(2c\rho\eta)^2 \leq 32c^2k^2\rho, \tag{3.9}$$

so $\mathrm{d}_{\mathrm{KL}}(T) = O(c^2k^2\rho)$.

Let $B, B' \in S$ and suppose $\sigma, \tau \in \mathcal{S}_k$. By construction,

$$\left\|(W[Q_B])_{\sigma,\tau} - W[Q_{B'}]\right\|_2^2 \geq \frac{1}{k^2}\mathrm{Ham}(B_{\sigma,\tau}, B')(2\rho c\eta)^2 = \Omega(c^2\rho^2\eta^2). \tag{3.10}$$

Thus by Corollary 2,

$$\delta_2(W[Q_B], W[Q_{B'}]) \geq \hat{\delta}_2(W[Q_B], W[Q_{B'}]) \geq \Omega(c\rho\eta). \tag{3.11}$$

Therefore, there exists $D > 0$ such that if $\epsilon = D\rho c\eta = D\min\left(c\rho, \frac{ck\sqrt{\rho}}{n}\right)$, we have $\log \mathcal{M}(\epsilon, T, \delta_2) = \Omega(k^2)$. Fano's lemma implies

$$\inf_{\hat{W}} \sup_{W} \Pr\left(\delta_2(\hat{W}, W) \geq \frac{\epsilon}{2}\right) \geq 1 - \frac{O(c^2k^2\rho) + 1}{\Omega(k^2)}. \tag{3.12}$$

We can choose c small enough that the right-hand side is larger than a fixed constant for all k and n. By Markov's inequality, we have

$$\inf_{\hat{W}} \sup_{W} \mathbb{E}\left[\delta_2(\hat{W}, W)\right] = \Omega(\epsilon) = \Omega\left(\min\left(\rho, \sqrt{\rho\frac{k^2}{n^2}}\right)\right). \tag{3.13}$$

□

Proof of Proposition 1. During the course of the proof of Theorem 3, we construct $2^{\Omega(k^2)}$ graphons in $\mathcal{W}_\rho$ that are pairwise at least $\Omega(\rho c\eta)$ apart in the δ_2 distance for any $c > 0$ such that $|c\eta| \leq \frac{1}{2}$. Therefore, for some $C > 0$, the $C\rho$-packing number of $\mathcal{W}_\rho$ is at least $2^{\Omega(k^2)}$. □

Acknowledgements

We are grateful for helpful conversations with Christian Borgs, Jennifer Chayes, Olga Klopp and Alexandre Tsybakov.

Funding

Alfred P. Sloan Foundation; a Google Faculty Research Award; and the National Science Foundation award (IIS-1447700).

References

1. Mossel, E., Neeman, J. & Sly, A. (2014) Reconstruction and estimation in the planted partition model. *Probab. Theory Related Fields*, **162**, 431–461.
2. Mossel, E., Neeman, J. & Sly, A. (2015) Consistency thresholds for the planted bisection model. *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC '15* (R. Rubinfeld ed.). New York, NY, USA: ACM, pp. 69–75.
3. Banks, J., Moore, C., Neeman, J. & Netrapalli, P. (2016) Information-theoretic thresholds for community detection in sparse networks. *Proceedings of the 29th Conference on Learning Theory, COLT* (V. Feldman, A. Rakhlin & O. Shamir eds). New York, USA: JMLR, pp. 383–416.
4. Klopp, O., Tsybakov, A. B. & Verzelen, N. (2015) Oracle inequalities for network models and sparse graphon estimation (version 1). *arXiv:1507.04118v1*.
5. Klopp, O., Tsybakov, A. B. & Verzelen, N. (2017) Oracle inequalities for network models and sparse graphon estimation (version 3). *Ann. Statist.*, **45**, 316–354.
6. Hoff, P. D., Raftery, A. E. & Handcock, M. S. (2002) Latent space approaches to social network analysis. *J. Amer. Stat. Assoc.*, **97**, 1090–1098.
7. Lovász, L. (2012) Large networks and graph limits. *Amer. Math. Soc. Colloq. Publ.*, **60**, 1–475.
8. Borgs, C., Chayes, J. T., Cohn, H. & Zhao, Y. (2014) An L^p theory of sparse graph convergence I: limits, sparse random graph models, and power law distributions. *arXiv:1401.2906*.
9. Borgs, C., Chayes, J. T., Cohn, H. & Zhao, Y. (2014) An L^p theory of sparse graph convergence II: LD convergence, quotients, and right convergence. *arXiv:1408.0744*.
10. Borgs, C., Chayes, J. T., Lovász, L., Sós, V. & Vesztergombi, K. (2006) Counting graph homomorphisms. *Topics in Discrete Mathematics* (M. Klazar, J. Kratochvil, M. Loebl, J. Matousek, R. Thomas & P. Valtr eds). Berlin, Heidelberg: Springer, pp. 315–371.
11. Borgs, C., Chayes, J. T., Lovász, L., Sós, V. & Vesztergombi, K. (2008) Convergent graph sequences I: subgraph frequencies, metric properties, and testing. *Adv. Math.*, **219**, 1801–1851.
12. Bickel, P. J. & Chen, A. (2009) A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA*, **106**, 21068–21073.
13. Bickel, P. J., Chen, A. & Levina, E. (2011) The method of moments and degree distributions for network models. *Ann. Statist.*, **39**, 2280–2301.
14. Abbe, E., Bandeira, A. S. & Hall, G. (2016) Exact recovery in the stochastic block model. *IEEE Trans. Inf. Theory*, **62**, 471–487.
15. Abbe, E. & Sandon, C. (2015) Recovering communities in the general stochastic block model without knowing the parameters. *Advances in Neural Information Processing Systems 28* (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama & R. Garnett eds). Montreal, Canada: MIT Press, pp. 676–684.
16. Airoldi, E. M., Costa, T. B. & Chan, S. H. (2013) Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. *Advances in Neural Information Processing Systems 26* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger eds). Lake Tahoe, USA: Curran Associates, Inc., pp. 692–700.
17. Chan, S. H. & Airoldi, E. M. (2014) A consistent histogram estimator for exchangeable graph models. *J. Mach. Learn. Res. Workshop Conf. Proc.*, **32**, 208–216.
18. Chatterjee, S. (2015) Matrix estimation by universal singular value thresholding. *Ann. Statist.*, **43**, 177–214.
19. Gao, C., Lu, Y. & Zhou, H. H. (2015) Rate-optimal graphon estimation. *Ann. Statist.*, **43**, 2624–2652.
20. Latouche, P. & Robin, S. (2016) Variational Bayes model averaging for graphon functions and motif frequencies inference in W-graph models. *Stat. Comput.*, **26**, 1173–1185.

21. Lloyd, J. R., Orbanz, P., Ghahramani, Z. & Roy, D. M. (2012) Random function priors for exchangeable arrays with applications to graphs and relational data. (F. Pereira, C. J. C. Burges, L. Bottou & K. Q. Weinberger eds). Lake Tahoe, USA: Curran Associates, Inc., pp. 998–1006.

22. Tang, M., Sussman, D. L. & Priebe, C. E. (2013) Universally consistent vertex classification for latent positions graphs. *Ann. Statist.*, **41**, 1406–1430.

23. Wolfe, P. & Olhede, S. C. (2013) Nonparametric graphon estimation. *arXiv:1309.5936*.

24. Yang, J. J., Han, Q. & Airoldi, E. M. (2014) Nonparametric estimation and testing of exchangeable graph models. *Proceedings of 17th International Conference on Artificial Intelligence and Statistics* (S. Kaski & J. Corander eds). Reykjavik, Iceland: PMLR, pp. 1060–1067.

25. Borgs, C., Chayes, J. T. & Smith, A. (2015) Private graphon estimation for sparse graphs. *Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS'15* (C. Cortes, D. D. Lee, M. Sugiyama & R. Garnett eds). Cambridge, MA: MIT Press, pp. 1369–1377.

26. Banerjee, D. (2016) Contiguity results for planted partition models: the dense case. *arXiv:1609.02854v1*.

27. Yu, B. (1997) Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam* (D. Pollard, E. Torgersen & G. Yang eds). New York: Springer, pp. 423–435.

28. Lapidoth, A. (2017) *A Foundation in Digital Communication*. Cambridge, UK: Cambridge University Press.

Appendix. Proof of equation (1.6)

We show here how to derive the lower bound in (1.6) from the results of Mossel *et al.* [1] and Banerjee [26].

Let $(\Omega_n, \mathcal{F}_n)$ be a sequence of measurable spaces, each equipped with two probability measures, $\mathbb{P}_n$ and $\mathbb{Q}_n$. We say $\mathbb{P}_n$ and $\mathbb{Q}_n$ are mutually contiguous if for any sequence of events A_n, we have $\lim_{n\to\infty} \mathbb{P}_n(A_n) \to 0$ if and only if $\lim_{n\to\infty} \mathbb{Q}_n(A_n) \to 0$.

Lemma A1 Let W_1 be a k-block graphon generated by a matrix with diagonal entries, p, and off-diagonal entries, q. Let W_2 be the Erdös–Rényi model with the same expected degree as W_1. If W_1 and W_2 are mutually contiguous then

$$\inf_{\hat{W}} \sup_{W} \mathop{\mathbb{E}}_{G\sim G_n(W)} \left[\delta_2(\hat{W}(G), W)\right] \geq \Omega\left(\frac{|p-q|}{\sqrt{k}}\right), \tag{A.1}$$

where $\inf_{\hat{W}}$ is the infimum over all estimators $\hat{W} : G_n \to \mathcal{G}$ and $\sup_W$ is the supremum over all k-block, $\max(p, q)$-bounded graphons.

Proof. Let $\epsilon = |p-q|$, so $\delta_2(W_1, W_2) = \sqrt{\frac{(k-1)^2}{k^3}\epsilon^2 + \frac{(k-1)}{k^3}\epsilon^2} \geq C\frac{\epsilon}{\sqrt{k}}$, for some constant C. Suppose, for sake of contradiction, that there exists $\hat{W}$ such that

$$\sup_{W\in\mathcal{W}_\rho} \mathbb{E}_{G\sim G_n(W)}\left[\delta_2(\hat{W}(G), W)\right]$$

is not $\Omega\left(\frac{\epsilon}{\sqrt{k}}\right)$. Then there exists a subsequence $\{n_t\}_{t\in\mathbb{N}}$ such that

$$\sup_{W\in\mathcal{W}_\rho} \mathbb{E}_{G\sim G_{n_t}(W)}\left[\delta_2(\hat{W}(G), W)\right] \leq \frac{C}{2t}\frac{\epsilon}{\sqrt{k}}$$

for all $t \in \mathbb{N}$.

The above inequality, combined with Markov's inequality, implies

$$\lim_{t\to\infty} Pr_{G\sim G_{n_t}(W_1)}\left[\delta_2(\hat{W}(G), W_1) \geq \frac{C}{2}\frac{\epsilon}{\sqrt{k}}\right] \to 0 \tag{A.2}$$

and

$$\lim_{t\to\infty} Pr_{G\sim G_{n_t}(W_2)}\left[\delta_2(\hat{W}(G), W_2) \geq \frac{C}{2}\frac{\epsilon}{\sqrt{k}}\right] \to 0. \tag{A.3}$$

By equation (A.3) and the contiguity of W_1 and W_2,

$$\lim_{t\to\infty} Pr_{G\sim G_{n_t}(W_1)}\left[\delta_2(\hat{W}(G), W_2) \geq \frac{C}{2}\frac{\epsilon}{\sqrt{k}}\right] \to 0. \tag{A.4}$$

Therefore, equations (A.2) and (A.4) imply that for large enough n, there exists a graph G such that $\delta_2(\hat{W}(G), W_1) < \frac{C}{2}\frac{\epsilon}{\sqrt{k}}$ and $\delta_2(\hat{W}(G), W_2) < \frac{C}{2}\frac{\epsilon}{\sqrt{k}}$, which implies that $\delta_2(W_1, W_2) < C\frac{\epsilon}{\sqrt{k}}$, which is a contradiction. □

There are many results in the literature exploring when block graphons are contiguous with the corresponding Erdös–Rényi model. The following table summarizes some of the known results in this area and translates them into lower bounds on the graphon estimation problem via Lemma A1. Let $\rho = \max(p, q)$.

Condition on p and q for contiguity to hold	Parameter regime	Lower bound on estimation error	Citation
$n(p-q)^2 \leq 2(p+q)$	$p = a/n, q = b/n$ for constants $a, b, k = 2$	$\Omega\left(\min\left(\rho, \sqrt{\frac{\rho}{n}}\right)\right)$	[1]
$n(p-q)^2 \leq 2(p+q)$	$\rho n \to \infty, \rho n = o(n), k = 2$	$\Omega\left(\min\left(\rho, \sqrt{\frac{\rho}{n}}\right)\right)$	[26]
$\frac{n^2(p-q)^2(k-1)}{p+(k+1)q} \leq 2\log(k-1)$	$p = a/n, q = b/n$ for constants a, b	$\Omega\left(\min\left(\frac{\rho}{\sqrt{k}}, \sqrt{\frac{\rho \log k}{n}}\right)\right)$	[3]