Submitted to the Annals of Statistics

CONSISTENT NONPARAMETRIC ESTIMATION FOR HEAVY-TAILED SPARSE GRAPHS

By Christian Borgs[†], Jennifer T. Chayes[†], Henry Cohn[‡], and Shirshendu Ganguly[*,†]

University of California, Berkeley[†] *and Microsoft Research*[‡]

We study graphons as a nonparametric generalization of stochastic block models, and show how to obtain compactly represented estimators for sparse networks in this framework. In contrast to previous work, we relax the usual boundedness assumption for the generating graphon and instead assume only integrability, so that we can handle networks that have long tails in their degree distributions. We also relax the usual assumption that the graphon is defined on the unit interval, to allow latent position graphs based on more general spaces.

We analyze three algorithms. The first is a least squares algorithm, which gives a consistent estimator for all square-integrable graphons, with errors expressed in terms of the best possible stochastic block model approximation. Next, we analyze an algorithm based on the cut norm, which works for all integrable graphons. Finally, we show that clustering based on degrees works whenever the underlying degree distribution is atomless.

CONTENTS

1. Introduction. Motivated by real-world technological, social, and biological networks, the study of large networks has become increasingly important. Much work in the statistics and machine learning communities has

[*] Supported by an internship at Microsoft Research New England.

MSC 2010 subject classifications: Primary 62G20; secondary 62H30, 05C80

Keywords and phrases: sparse networks, estimation, graphons

focused on the questions of modeling and estimation for these networks. In this paper, we analyze three algorithms for estimating the structure of a sparse network. Our results apply to a substantially larger class of network models than those analyzed in previous papers. Specifically, our techniques handle networks with heavy-tailed degree distributions.

1.1. *Stochastic block models and W-random graphs.*

Many previous papers have described large networks in terms of parametric models, one of the most popular being the stochastic block model, introduced in [41]. These models can be characterized by a vector of probabilities $\mathbf{p} = (p_i)$ on a finite set of communities and a matrix $B = (\beta_{ij})$ of "affinities." Given these parameters, one then generates a graph on n labeled nodes by assigning a community to each vertex, independently at random according to the probability distribution $\mathbf{p}$, and then connecting vertices belonging to communities i and j with probability β_{ij}. Such a model is often considered a reasonable approximation of a social network characterized by a limited number of communities.

More recently, motivated by extremely large networks, researchers have begun to consider nonparametric stochastic block models, for which there is a continuous family of communities, i.e., for which the $k \times k$ matrix of edge probabilities is replaced by a two-dimensional function. The nonparametric models we study in this paper are usually referred to as W-random graphs or latent position graphs. In the most general setup, such a model is defined in terms of a probability space $(\Omega, \mathcal{F}, \pi)$ (the space of latent positions or features) and a *graphon* W over $(\Omega, \mathcal{F}, \pi)$, defined as an integrable, nonnegative function on $\Omega \times \Omega$ that is symmetric in the sense that $W(x, y) = W(y, x)$ for all $x, y \in \Omega$. To generate a graph on n nodes, one then chooses n "positions" $x_1, \dots, x_n$ i.i.d. at random from $(\Omega, \mathcal{F}, \pi)$ and, conditioned on these, chooses edges independently, with the probability of an edge between vertices i and j given by $W(x_i, x_j)$. The resulting graph is called a *W-random graph.*

As originally proposed in [40], the space of latent positions Ω comes equipped with a metric and the probability of connection is a function of distance, but the more general setting we have described is commonly studied. Note that in the dense setting, this model is quite natural, since it can be shown [42, 6, 32] that if a random graph G is the restriction of (an ergodic component of) an infinite, exchangeable random graph, then G must be an instance of a W-random graph for some function W with values in $[0, 1]$. Due to this connection, W-random graph models are often called exchangeable graph models.

1.2. *Dense and sparse graphs.* To model sparse graphs in this nonparametric setup, one uses connection probabilities that are given by a symmetric function W times a *target density* ρ, leading to the model of "inhomogeneous random graphs" defined in [11], with nodes i and j now being connected with probability $\min\{1, \rho W(x_i, x_j)\}$. The resulting graph is called a W-random graph with target density ρ and will be denoted by $G_n(\rho W)$.

For both dense and sparse graphs, this kind of model is related to the theory of convergent graph sequences [17, 18, 12, 19, 15, 14]. In the setting of dense graph limits, W-random graphs were first explicitly proposed in [51], although they can be implicitly traced back to the much earlier work of [42] and [6] mentioned above. The term "graphon" originated in [18].

While for dense graphs one only needs to consider bounded graphons, this boundedness assumption is not very natural for sparse graphs. Indeed, suppose W is a bounded graphon, and let n tend to infinity with $\rho = \rho_n$ chosen so that $\rho_n n \to \infty$ as well. Then one can check that the nonzero degrees in $G_n(\rho_n W)$ are of the same order of magnitude, in the sense that for each $\varepsilon > 0$, there exists a constant $c > 0$ such that with probability $1 - o(1)$, at least a $1 - \varepsilon$ fraction of the nonzero degrees in $G_n(\rho_n W)$ are within a factor of c of $\rho_n n$. This behavior is natural for dense graphs, but it is a serious restriction for sparse graphs. Instead, many real-world networks have long-tailed degree distributions. For applications, one would therefore want to consider unbounded graphons W.

1.3. *Estimation and previous literature.* How can we estimate a graphon W given a sample G of a W-random graph? This problem encapsulates the idea of inferring the underlying structure in a random network.

For the special case where W is a stochastic block model, the estimation problem is closely related to the problem of graph partitioning and has been intensely studied in the literature [62, 41, 34], using methods that range from maximum likelihood estimates [61] and Gibbs sampling [58] or simulated annealing [43] to spectral clustering [13, 53, 31, 30, 25] and tensor algebra [8]. Proving consistency of these methods is often not hard in the dense regime, but it becomes more difficult for sparse graphs. See, for example, [48, 49] for a proof of consistency for spectral clustering when the average degree is as small as $\log n$, and [2, 3] for an effective algorithm that is provably consistent as long as the average degree diverges.

Estimating graphons that are not block models is more challenging. This problem is implicit in [44], but the first explicit discussion of the nonparametric problem we are aware of was given in [9], even though the actual consistency proof there is still limited to stochastic block models with a fixed number of blocks. The restriction to a fixed number of blocks was relaxed in [56] and [29]. The full nonparametric model was studied in [10], under the assumption that none of the eigenfunctions of the operator associated with the kernel W is orthogonal to the constant function 1 and the eigenvalues are distinct.

Many further papers have been written on graphon estimation, including [56, 26, 50, 5, 7, 35, 46, 52, 55, 59, 63, 1, 21, 22, 28, 64, 37, 38, 54, 60, 65, 66, 23, 27, 39, 20, 45]. Each paper makes different assumptions about the density and the underlying graphon. Strong results are known for dense graphs: [23] shows how to approximate arbitrary measurable graphons W with values in $[0, 1]$ given a dense W-random graph, and [37] attains an optimal rate for least squares estimators of both stochastic block models and Hölder-continuous graphons from a dense graph. For sparse graphs, [63] proves convergence of a maximum likelihood estimator under the assumption that W is bounded, bounded away from zero, and Hölder-continuous. Most recently, [20] introduces a modified version of the least squares algorithm that optimizes over block models with bounded L^∞ norm; this algorithm achieves consistency for arbitrary bounded graphons and arbitrary densities, as long as the average degree diverges with the number of vertices. The same paper also gives a differentially private version of the least squares algorithm which works again for arbitrary bounded graphons, now requiring that the average degree must grow at least like the logarithm of the number of vertices. Independently, [45] proposes and analyzes the modified (nonprivate) algorithm and proves matching upper and lower bounds for the rates achieved by this algorithm.

But more important than some of the technical assumptions used by previous authors is the fact that *all* the previous results we are aware of require W to be *bounded*. As pointed out above, this assumption, while natural for dense graphs, rules out most degree distributions observed in real-world networks. Our goal here is to remove this assumption.

1.4. *Identifiability.* Before summarizing our contributions, we need to discuss the fact that in general, W cannot be uniquely determined from the observation of even the full sequence $(G_n)_{n\geq 1}$, a problem called the identifiability issue in the literature; see, for example, [9, 22]. To discuss this, consider two probability spaces $(\Omega, \mathcal{F}, \pi)$ and $(\Omega', \mathcal{F}', \pi')$, a measure-preserving map $\phi\colon \Omega' \to \Omega$, and a graphon W over $(\Omega, \mathcal{F}, \pi)$. Define the pullback of W to $(\Omega', \mathcal{F}', \pi')$ to be the graphon W^ϕ defined by $W^\phi(x, y) = W(\phi(x), \phi(y))$. Then the sequences of random graphs generated from two graphons W and W' have the same distribution if $W' = W^\phi$. While it was stated in some of the early literature on graphon estimation that the converse is true as well, that turns out to be false; see, for example, Example A.3 below for a counterexample. To formulate the correct statement, we define W and W' to be *equivalent* if there exists a third graphon U over a probability space $(\Omega'', \mathcal{F}'', \pi'')$ and two measure-preserving maps $\phi\colon \Omega \to \Omega''$ and $\phi'\colon \Omega' \to \Omega''$ such that $W = U^{\phi}$ and $W' = U^{\phi'}$ almost everywhere. With this definition, we are now ready to characterize the full extent to which W is not identifiable:

THEOREM 1.1. *Let W and W' be graphons over the probability spaces $(\Omega, \mathcal{F}, \pi)$ and $(\Omega', \mathcal{F}', \pi')$, respectively, and assume that $n\rho_n \to \infty$ and $\rho_n \to 0$. Then the random graphs $G_n(\rho_n W)$ and $G_n(\rho_n W')$ are identically distributed for all n if and only if W and W' are equivalent.*

The analogue of this theorem for the dense case (where $\rho_n = 1$ and W and W' take values in $[0, 1]$) follows from the results of [16] by a simple argument involving subgraph counts or the results of [32] if we assume that both graphons are defined over $[0, 1]$. But for the sparse case and general integrable (rather than bounded) graphons this is a new result; see Remark 3.11(i) in Section 3.5 for the proof. In view of Theorem 1.1, both the feature space $(\Omega, \mathcal{F}, \pi)$ and the graphon W are unobservable in general, and even if we fix the feature space there is no "canonical graphon" an estimation procedure can output. In light of these facts, the natural way of dealing with the identification problem is to admit that there is nothing canonical about any particular representative W, and to define consistency as consistency with respect to a metric between equivalence classes, rather than between graphons themselves.

1.5. *Goals.* In this paper, we follow the spirit of [63] and define consistency with respect to a metric on equivalence classes of graphons, but in contrast to [63], we allow for more general spaces than just the uniform distribution over the unit interval since more general feature spaces are more natural from an application point of view (see Remark 3.7 below). To define our notion of distance, we recall that a *coupling* between probability measures π and π' is a measure ν on the product space such that the projections of ν to the two coordinates are equal to π and π', respectively. Given $p \geq 1$ and two L^p graphons W over $(\Omega, \mathcal{F}, \pi)$ and W' over (Ω', π') (i.e., graphons such that $\int_{\Omega} |W|^p \, d\pi < \infty$ and $\int_{\Omega'} |W'|^p \, d\pi' < \infty$), we then define the distance $\delta_p(W, W')$ by

$$\delta_p(W, W') = \inf_{\nu} \left(\int \left| W(x, y) - W'(x', y') \right|^p d\nu(x, x') \, d\nu(y, y') \right)^{1/p}, \tag{1.1}$$

where the infimum is over all couplings ν of π and π'. Note that this distance is a version of the Wasserstein p-distance.

Having defined a metric on equivalence classes of graphons, we can now formulate the estimation problem considered in this paper: *Given a single instance* of a W-random graph defined on an unobserved probability space $(\Omega, \mathcal{F}, \pi)$, find an algorithm that (a) outputs an estimator $\widehat{W}$ such that $\widehat{W}$ has a *concise representation* whose size grows only slowly with n; (b) estimates W consistently *assuming just integrability conditions*; (c) works for *arbitrary target densities*, as long as the graph is not too sparse (say has divergent average degree); and (d) runs in *polynomial time.*

While efficiency (property (d)) is clearly important for practical applications, our main focus in this paper will be the fundamental problem of consistent estimation under as few restrictions on W as possible, i.e., algorithms achieving properties (a)–(c). Indeed, none of the three algorithms we study in this paper achieves all four properties. Two of them achieve (a)–(c), and hence solve the desired problem of consistent estimation, but do not run in polynomial time. The third achieves (a), (c), and (d), and hence is efficient, but requires an additional condition to ensure consistency.

Focusing on approximation under a given metric is a useful abstraction, but it can obscure one issue: in practice it is generally not enough just to know that $\delta_p(W, \widehat{W})$ is small. In addition, we would like to find an explicit coupling ν between W and $\widehat{W}$ in (1.1), not necessarily achieving the exact infimum but at least providing a good bound for $\delta_p(W, \widehat{W})$. In principle one could imagine algorithms without this property, but the estimators we analyze in this paper all produce explicit couplings.

1.6. *Organization.* The rest of the paper is organized as follows. In Sections 2 and 3, we state our main results and place them in the context of the theory of graphons. Sections 4 through 6 outline the proofs of our main theorems. We conclude the body of the paper with Section 7, which examines how our bounds behave given a greater degree of regularity than we assume elsewhere in the paper (namely, Hölder continuity). A more detailed treatment of our theory is given in the appendices, which can be found in the supplementary material. Appendices A through C provide a thorough account of measure-theoretic technicalities and various estimates for graphons and degree distributions. Appendices D and E fill in the details of the proofs of our main theorems. Appendix F proves bounds for the special case of Hölder-continuous graphons. Appendix G analyzes several examples of network models with power-law degree distributions, and shows exactly how our theorems apply. Finally, Appendix H derives some concentration bounds we use in our proofs.

2. Summary of results.

In this paper, our estimator $\widehat{W}$ will be given in terms of a block model, with a number of blocks that grows slowly with the number of vertices of the input graph. Given this framework, it is natural to compare the performance of our algorithm to that of the best possible block model in a suitable class. Here we consider the class $\mathcal{B}_{\geq\kappa} = \{(\mathbf{p}, B) : \min_i p_i \geq \kappa\}$ of all block models with minimal block size at least κ. For an approximation outputting a block model in $\mathcal{B}_{\geq\kappa}$, the best error we could achieve is

$$\varepsilon^{(p)}_{\geq\kappa}(W) = \inf_{W' \in \mathcal{B}_{\geq\kappa}} \delta_p(W, W'). \tag{2.1}$$

We often refer to this benchmark as an *oracle error*, since it is the best an oracle with access to the unknown W could do. Our goal is to prove *oracle inequalities* that bound the estimation error in terms of the oracle error, as well as a few additional terms that account for variance and the visibility of heavy tails at finite scale.

When establishing the estimation error for W, we first prove a bound on the estimation error for the intermediate matrix $Q_n = Q_n(\rho W)$ with entries

$$(Q_n)_{ij} = \min\{1, \rho W(x_i, x_j)\} \quad \text{if } i \neq j \tag{2.2}$$

and $(Q_n)_{ii} = 0$. The estimation error for Q_n will be expressed in terms of an oracle error for Q_n plus a concentration error stemming from the fact that, even after conditioning on Q_n, the observed graph G_n is random; see Theorems 4.1 and 5.1 below. In a second step, we then prove consistency for the original estimation error, given bounds that estimate the difference between $\widehat{W}$ and W. Note that part of the literature stops at the first step, effectively avoiding the identifiability issue discussed above.

In this paper, we consider three algorithms for producing a block model approximation to W from a single instance of a W-random graph G: two inefficient ones and one whose running time is polynomial in n.

1. The well-known *least squares algorithm*, which has been analyzed under various additional assumptions on W, until recently [20, 45] not even covering arbitrary bounded graphons. Here we will prove consistency of this algorithm in the metric δ_2 for arbitrary L^2 graphons.
2. A *least cut norm algorithm*, which we prove to be consistent under the cut norm for arbitrary L^1 graphons. The cut norm is defined below.
3. A *degree sorting algorithm*, which we show is consistent whenever the degree distribution of W is atomless. (Graphons with this property

are equivalent to graphons over $[0,1]$ such that $W_x := \int_0^1 W(x,y)\,dy$ is strictly monotone in x.) This algorithm runs in polynomial time.

To state our results, we need a few definitions. As usual, $[n]$ denotes the set $\{1,\dots,n\}$. Given an $n \times n$ matrix A, we use $\|A\|_p$ to denote its L^p norm, defined by $\|A\|_p^p = \frac{1}{n^2}\sum_{i,j}|A_{ij}|^p$. Given a graph G on $[n]$, we use $A(G)$ to denote the adjacency matrix of G, and $\rho(G) = \|A(G)\|_1$ to denote its density. We identify partitions of $[n]$ into k classes (some of which can be empty) with maps $\tau\colon [n] \to [k]$, where $V_i = V_i(\tau) = \tau^{-1}(\{i\})$ is the i^{th} class of the partition. Given such a map and a $k \times k$ matrix B, we will use B^τ for the $n \times n$ matrix with entries $(B^\tau)_{ij} = B_{\tau(i)\tau(j)}$. Finally, for an $n \times n$ matrix A, we use A_τ to denote the matrix where for each $(x,y) \in V_i \times V_i$, the matrix element A_{xy} is replaced by the average over $V_i \times V_j$, and A/τ to denote the $k \times k$ matrix of block averages

$$(A/\tau)_{ij} = \frac{1}{|V_i|\,|V_j|}\sum_{(u,v)\in V_i \times V_j} A_{uv},$$

defined to be 0 if either V_i or V_j is empty; note that the two are related by $A_\tau = (A/\tau)^\tau$.

Throughout this paper, we will assume that the graph is sparse (in the sense that $\rho \to 0$), but that it has divergent average degree (i.e., we assume that $n\rho \to \infty$). Under these assumptions we will prove the following results.

2.1. *Least squares estimation.* Given an input graph G on n vertices and a parameter $\kappa \in (0,1]$ such that $\kappa n \geq 1$, let

$$(\hat{\tau}, \hat{B}) \in \operatorname*{argmin}_{\tau,B} \|A(G) - B^\tau\|_2, \tag{2.3}$$

where the optimization is over all $k \times k$ matrices B and all partitions $\tau\colon [n] \to [k]$ such that all non-empty classes of τ have size at least $\lfloor \kappa n \rfloor$, with k chosen so that it can accommodate all such partitions, say $k = \lceil \frac{n}{\lfloor n\kappa \rfloor} \rceil$. Setting $\hat{p}_i = \frac{1}{n}|V_i(\hat{\tau})|$ to be the relative size of the i^{th} partition class of $\hat{\tau}$, the least squares algorithm then outputs the block model $\widehat{W} = (\hat{\mathbf{p}}, \hat{B})$. Note that the above minimization problem is slightly helped by the fact that we minimize the L^2 norm. For a given τ, the minimizer $\hat{B}$ can therefore be obtained by averaging $A(G)$ over the classes of τ, showing that $\hat{B}$ is of the form $A(G)/\tau$. Nevertheless the algorithm is inefficient, since we still need to minimize over partitions $\tau\colon [n] \to [k]$.

Our main result concerning this algorithm is that if G is a W-random graph at target density ρ and $W \in L^2$, then the algorithm is consistent in the sense that $\delta_2\left(\frac{1}{\rho}\widehat{W}, \frac{1}{\|W\|_1}W\right) \to 0$ with probability 1 as $n \to \infty$, as long as $\kappa \to 0$ and $\kappa^{-2}\log(1/\kappa) = o(n\rho)$. To give a quantitative error bound, we define

$$\mathrm{tail}_\rho^{(p)}(W) := \|W - \min\{W, \rho^{-1}\}\|_p,$$

a quantity which measures the difference between W and $\frac{1}{\rho}\min\{1, \rho W\}$ and tends to 0 as $\rho \to 0$ provided $W \in L^p$. Note that in addition to the oracle error $\varepsilon^{(2)}_{\geq\kappa}(W)$, an error term reflecting the difference between W and $\frac{1}{\rho}\min\{1, \rho W\}$ is unavoidable, since the parts of W that are larger than $1/\rho$ are not reflected in the distribution of $G_n(\rho W)$.

THEOREM 2.1. *Let W be an L^2 graphon, normalized so that $\|W\|_1 = 1$, and let $\widehat{W} = (\hat{\mathbf{p}}, \hat{B})$ be the output of the least squares algorithm (2.3) for a W-random graph G on n vertices with target density ρ.*

(i) If $\kappa \in (n^{-1}, 1]$ and $\frac{1+\log(1/\kappa)}{\kappa^2} = O(\rho n)$, then

$$\delta_2\Big(\frac{1}{\rho}\widehat{W}, W\Big) = O_p\left(\varepsilon^{(2)}_{\geq\kappa}(W) + \sqrt[4]{\frac{1+\log(1/\kappa)}{\kappa^2\rho n}} + \sqrt[4]{\frac{\log n}{\kappa n}} + \mathrm{tail}_\rho^{(2)}(W)\right).$$

(ii) If $\kappa \in (0,1]$ is fixed and $\rho = \rho_n$ is such that $\rho_n \to 0$ and $n\rho_n \to \infty$, then

$$\delta_2\Big(\frac{1}{\rho}\widehat{W}, W\Big) \to \varepsilon^{(2)}_{\geq\kappa}(W) \quad \textit{with probability } 1.$$

(iii) If $\rho = \rho_n$ and $\kappa = \kappa_n$ are such that $\rho_n \to 0$, $n\rho_n \to \infty$, $\kappa_n \to 0$, and $\kappa_n^{-2}\log(1/\kappa_n) = o(n\rho_n)$ as $n \to \infty$, then

$$\delta_2\Big(\frac{1}{\rho}\widehat{W}, W\Big) \to 0 \quad \textit{with probability } 1.$$

The proof is given in Section 4 and Appendix D.

The conditions on ρ_n in Theorem 2.1 are very natural: they simply say that the graph is sparse but the average degree tends to infinity. The conditions on κ_n are an artifact of our proof techniques, but they have a reasonable interpretation. For example, if we disregard the logarithmic factor, $\kappa_n^{-2}\log(1/\kappa_n) = o(n\rho_n)$ says that the number of parts in the partition must be asymptotically smaller than the square root of the average degree.

The four error terms in part (i) arise for different reasons. First, when estimating the L^2 distance between the matrix of probabilities Q_n and the estimator $\widehat{W}$, one encounters an oracle error for Q_n and a concentration error, the latter being the second error term. Second, one encounters an additional error when bounding the oracle error for Q_n in terms of the oracle error for W. Since Q_n is random, this involves another concentration error, which is the third term. Finally, we need to estimate the δ_2 distance between W and $\frac{1}{\rho}Q_n$, which involves bounding both the distance between W and $\frac{1}{\rho}\min\{1, \rho W\}$, and the distance between $\min\{1, \rho W\}$ and Q_n. It turns out that the latter error can be absorbed in the other terms present above, while the former leads to the term $\operatorname{tail}_\rho^{(2)}(W)$.

Note that the term $\sqrt[4]{\frac{1+\log(1/\kappa)}{\kappa^2\rho n}}$ in the oracle inequality is larger than the next term $\sqrt[4]{\frac{\log n}{\kappa n}}$ when $\rho \le 1/\log n$. We have included both terms to handle the case in which ρ is large enough that the latter term dominates, but $\sqrt[4]{\frac{1+\log(1/\kappa)}{\kappa^2\rho n}}$ should be viewed as the primary term.

We expect these bounds can be improved. From our perspective, their purpose is to give concrete meaning to asymptotic consistency by providing specific guarantees. In the case of bounded graphons, the optimal convergence rate is known [37, 45], and it is better than what can be deduced from our theorem. For comparison, when $\rho = n^{-1/2}$, $\kappa = n^{-1/6}$, and W is a bounded graphon, Proposition 2.1 of [45] implies that

$$\delta_2\Big(\frac{1}{\rho}\widehat{W}, W\Big) = O_p\bigg(\varepsilon_{\ge\kappa}^{(2)}(W) + \frac{\sqrt{\log n}}{n^{1/4}}\bigg),$$

while Theorem 2.1 implies that

$$\delta_2\Big(\frac{1}{\rho}\widehat{W}, W\Big) = O_p\bigg(\varepsilon_{\ge\kappa}^{(2)}(W) + \frac{\sqrt[4]{\log n}}{n^{1/24}}\bigg).$$

Obtaining a tight estimate for unbounded graphons, and thus heavy-tailed graphs, will likely require further development of these techniques.

For general graphons, our results do not give explicit error bounds, since all we know is that $\varepsilon_{\ge\kappa}^{(2)}(W)$ and $\operatorname{tail}_\rho^{(2)}(W)$ tend to 0 as $\kappa \to 0$ and $\rho \to 0$. But in many applications, one has additional information on the generating graphon, for example, that it is actually a stochastic block model with a fixed number of classes, in which case both $\varepsilon_{\ge\kappa}^{(2)}(W)$ and $\operatorname{tail}_\rho^{(2)}(W)$ become identically zero once κ and ρ are small enough, leaving us only with the explicit terms in the above bound.

Another class of examples consists of α-Hölder-continuous graphons over $\mathbb{R}^d$ equipped with a probability measure that decays fast enough to make the function $|x|^\beta$ integrable. This class encompasses many models of latent position spaces used in practice. When W is α-Hölder-continuous and $|x|^\beta$ is integrable with $\alpha \in (0,1]$ and $\beta > 2\alpha$, we prove that $\varepsilon_{\ge\kappa}^{(2)}(W) = O(\kappa^{\alpha'})$

and $\mathrm{tail}_\rho^{(2)}(W) = O(\rho^{\beta'})$ for some $\alpha', \beta' > 0$, with $\alpha' = \alpha/d$ and $\beta' = \infty$ in the simple case of the uniform distribution over a box of the form $[-R, R]^d$. See Propositions 7.1 and 7.2 below.

This scaling behavior for the oracle error and tail bounds is typical. We have stated the oracle inequality in full generality, but when the graphon is sufficiently well behaved to estimate the oracle error and tail bounds, one can balance the error terms and derive the scaling rate for κ that optimizes these bounds. For example, suppose the error bound is

$$O_p\left(\kappa^{\alpha'} + \sqrt[4]{\frac{1+\log(1/\kappa)}{\kappa^2\rho n}} + \sqrt[4]{\frac{\log n}{\kappa n}} + \rho^{\beta'}\right).$$

Choosing κ proportional to $\left(\frac{\log(\rho n)}{\rho n}\right)^{\frac{1}{4\alpha'+2}}$ optimizes this bound (assuming $n\rho \to \infty$ as $n \to \infty$) and yields an error bound of

$$O_p\left(\left(\frac{\log(\rho n)}{\rho n}\right)^{\frac{\alpha'}{4\alpha'+2}} + \rho^{\beta'}\right),$$

which becomes $O_p\left(\left(\frac{\log(\rho n)}{\rho n}\right)^{\frac{\alpha}{4\alpha+2d}}\right)$ in the case of an α-Hölder-continuous graphon over $[-R, R]^d$ equipped with the uniform distribution.

2.2. *Cut norm estimation for general L^1 graphons.* To give an explicit description of the least cut norm algorithm, we need the notion of the cut norm, first introduced in [36]. For an $n \times n$ matrix A, it is defined as

$$\|A\|_\square = \max_{S,T\subseteq[n]} \frac{1}{n^2}\Big|\sum_{(i,j)\in S\times T} A_{ij}\Big|. \tag{2.4}$$

One way to define the least cut norm algorithm would be to output a block model defined in terms of the minimizer of $\|A(G) - B^\tau\|_\square$. But since we now need to minimize the cut norm rather than an L^2 norm, this would involve yet another optimization problem to find the best matrix B for each distribution τ. To circumvent this issue, we always obtain B by averaging. In other words, we calculate

$$\hat{\tau} \in \operatorname*{argmin}_{\tau} \|A(G) - (A(G))_\tau\|_\square, \tag{2.5}$$

where the argmin is again over partitions $\tau\colon [n] \to [k]$ such that every nonempty partition class has size at least $\lfloor \kappa n \rfloor$. The least cut norm algorithm then outputs the block average corresponding to $\hat{\tau}$; i.e., it outputs the block model $\widehat{W} = (\hat{\mathbf{p}}, \hat{B})$ where $\hat{p}_i$ is again the relative size of the i^{th} partition class of $\hat{\tau}$ and $\hat{B} = A(G)/\hat{\tau}$.

We will show that the least cut norm algorithm is consistent in the cut metric $\delta_\square$ on graphons, defined similar to δ_p, except that now we use the cut norm instead of the L^p norm $\|\cdot\|_p$; see (3.3) below for the precise definition. More precisely, we will show that a.s., the error in the $\delta_\square$ distance tends to zero for a W-random graph G if $\kappa \to 0$ in such a way that $\kappa^{-1} = o(\frac{n}{\log n})$. In addition to consistency, we will again show a quantitative bound, this time involving the oracle error and tail bound in the L^1 norm.[1]

Theorem 2.2. *Let W be an L^1 graphon, normalized so that $\|W\|_1 = 1$, and let $\widehat{W} = (\hat{p}, \hat{B})$ be the output of the least cut norm algorithm (2.5).*

(i) If $\kappa \in [\frac{\log n}{n}, 1]$, then

$$\delta_\square\Big(\frac{1}{\rho}\widehat{W}, W\Big) = O_p\left(\varepsilon^{(1)}_{\geq\kappa}(W) + \sqrt{\frac{1}{\rho n}} + \sqrt{\frac{\log n}{\kappa n}} + \operatorname{tail}^{(1)}_\rho(W)\right).$$

(ii) If $\kappa \in (0,1]$ is fixed and $\rho = \rho_n$ is such that $\rho_n \to 0$ and $n\rho_n \to \infty$, then

$$\limsup_{n\to\infty} \delta_\square\Big(\frac{1}{\rho}\widehat{W}, W\Big) \leq 2\varepsilon^{(1)}_{\geq\kappa}(W) \quad \textit{with probability } 1.$$

(iii) If $\rho = \rho_n$ and $\kappa = \kappa_n$ are such that $\rho_n \to 0$, $n\rho_n \to \infty$, $\kappa_n \to 0$, and $\frac{\log n}{n\kappa_n} \to 0$, then

$$\delta_\square\Big(\frac{1}{\rho}\widehat{W}, W\Big) \to 0 \quad \textit{with probability } 1.$$

The proof is given in Section 5 and Appendix E.

The four error terms in part (i) have the same explanation as those for the least squares algorithm: the oracle error for W, a concentration error appearing when estimating the cut norm error with respect to Q_n, a concentration error stemming from the random nature of the oracle error for Q_n, and a tail bound stemming from the fact that for unbounded graphons, the matrix Q_n generating G_n involves a truncation of the entries that are larger than 1. For Hölder-continuous graphons over $\mathbb{R}^d$ we can again give explicit error bounds of the form $\varepsilon^{(1)}_{\geq\kappa}(W) = O(\kappa^{\alpha'})$ and $\operatorname{tail}^{(1)}_\rho(W) = O(\rho^{\beta'})$; see Propositions 7.1 and 7.2 below.

[1] For analyzing the optimal convergence rate, it would be natural to use the cut norm for the oracle and tail bounds. We use the L^1 norm for two reasons: it fits naturally with our proof techniques, and it suffices to obtain asymptotic consistency in Theorem 2.2(iii).

2.3. *Graphon estimation via degree sorting.* The last algorithm we consider in this paper is the degree sorting algorithm, which proceeds as follows. Given a degree G on n vertices with vertex degrees $d_1, \dots, d_n$, we sort the vertices by choosing a permutation σ of $[n]$ such that

$$d_{\sigma(1)} \geq d_{\sigma(n)} \geq \cdots \geq d_{\sigma(n)}.$$

To separate the sorted vertices into k classes of nearly equal size, we choose integers $0 = n_0 < n_1 < \cdots < n_k = n$ such that

$$\left| n_i - \frac{in}{k} \right| < 1,$$

and we define $\tau\colon [n] \to [k]$ by $\tau(j) = i$ if $n_{i-1} < \sigma(j) \leq n_i$. Thus, τ groups the vertices into k classes, sorted by degree. The output of the algorithm is the block model $\widehat{W} = (\hat{p}, \hat{B})$ with $\hat{p}_i = 1/k$ and $\hat{B} = A(G)/\tau$. In other words, we simply cluster vertices with similar degrees and then average over these clusters.

This algorithm has the advantage of being very efficient, but it has no hope of working unless the degrees suffice to distinguish between the vertices. More precisely, we need the limiting distribution of normalized degrees to be atomless (i.e., there should not exist a nonzero fraction of the vertices with nearly the same degree). Note that this hypothesis is not satisfied by stochastic block models, because vertices in the same block cannot be distinguished by their degrees, but it holds generically.

If G is a W-random graph, then we can express the limiting degree distribution as $n \to \infty$ in terms of W; more precisely, it is just given by the distribution function D_W of the random variable $W_x = \int W(x,y)\, d\pi(y)$, where y is chosen according to π; see Appendix B for the proof. The next theorem states consistency of the degree sorting algorithm under the condition that the degree distribution of W is atomless.

Theorem 2.3. *Let W be a graphon whose degree distribution function $D_W\colon [0,\infty) \to [0,1]$ is continuous, let G_n be a W-random graph on n vertices with target density ρ_n, and let $\widehat{W}_n$ be the output of the degree sorting algorithm with k_n parts and input G_n.*

Suppose $\rho_n \to 0$, $n\rho_n \to \infty$, $k_n \to \infty$, $\log k_n = o(n\rho_n)$, and $k_n = o\big(n\sqrt{\rho_n}\big)$ as $n \to \infty$. Then $\rho_n^{-1}\widehat{W}_n$ converges a.s. to W under δ_1.

The proof is given in Section 6.

2.4. *Graphs with power-law degree distribution.* To give an example of how unbounded graphons can be analyzed, we consider two simple models for graphs with power-law degree distributions. Both are generated by graphons over $[0,1]$, with the first one given by $W(x,y) = \frac{1}{2}(g(x)+g(y))$, where $g(x) = (1-\alpha)(1-x)^{-\alpha}$ for some $\alpha \in (0,1)$, and the second one given by $W(x,y) = g(x)g(y)$. Both can be seen to have a degree distribution with density function $f(\lambda) = \Theta\bigl(\lambda^{-(1+1/\alpha)}\bigr)$, i.e., a power-law degree distribution with exponent $1+\frac{1}{\alpha}$. Both graphons are in L^p as long as $1 \le p < \frac{1}{\alpha}$.

It turns out that the first graphon can be expressed as an equivalent Hölder-continuous graphon over $\mathbb{R}^d$ equipped with a heavy-tailed distribution, while this is not possible for the second; see Appendix G for details. But both fit into our general theory, implying consistency for all three algorithms without any additional work, and both allow for explicit bounds similar to the ones obtained for Hölder-continuous graphons, even though only one of them can actually be expressed as a Hölder-continuous graphon. See Lemma G.1 for the precise estimates.

2.5. *Comparison with related results.* As discussed above, our primary contribution in this paper is to analyze the case of unbounded graphons, thus removing the restriction to networks in which all the degrees are of the same order. We also formulate our results over general probability spaces, which increases their applicability. (One can always pass to an equivalent graphon over $[0,1]$, but standardizing the underlying space prevents taking advantage of any smoothness or regularity the graphon possesses, because these properties are not invariant under equivalence.)

Least squares estimation is of course not a novel idea. Gao, Lu, and Zhou [37] proved consistency of least squares estimation based on sparse graphs obtained from bounded graphons satisfying a Hölder condition, and Wolfe and Olhede proved consistency under a Hölder condition and boundedness away from zero in an updated version of [63] that has not yet, as of this writing, been circulated publicly. Borgs, Chayes, and Smith [20] and Klopp, Tsybakov, and Verzelen [45] proved consistency for bounded graphons with no Hölder conditions or additional assumptions, but they did not handle the unbounded case. Our paper thus completes the analysis of this important algorithm, by proving consistency even when the underlying graphon is unbounded.

Bounded graphons are automatically square-integrable, but that is not necessarily true for unbounded graphons. Least squares estimation is an appropriate technique only for L^2 graphons, and we propose least cut norm estimation as a substitute that is applicable to arbitrary graphons.

Exact optimization is computationally inefficient for both the least squares and the least cut norm algorithms. Thus, our consistency results should be viewed not as a proposal that exact optimization should be carried out in practice for large networks, but rather as a benchmark for approximate or heuristic optimization.

Degree sorting has the advantage of efficiency, although it works only for graphons whose degrees are sufficiently well distributed. The idea of clustering vertices according to degree has a long history (see, for example, [33]), as well as connections with the theory of random graphs with a given degree sequence [47, 24]. Degree sorting has recently been analyzed as a graphon estimation algorithm by Chan and Airoldi [22]. They showed that their sorting and smoothing algorithm is consistent for dense graphs under two-sided Lipschitz conditions on the degrees of the underlying graphon. Our analysis accommodates sparse graphs and even unbounded graphons, while avoiding these Lipschitz conditions.

3. Graphons, identifiability, and graph convergence.

3.1. *Notation.* We will continue to use the notation from the previous section. Specifically, $[n]$ denotes the set $\{1, \dots, n\}$,

$$\|A\|_p^p = \frac{1}{n^2} \sum_{i,j} |A_{ij}|^p$$

for an $n \times n$ matrix A, and $A(G)$ denotes the adjacency matrix of a graph G on $[n]$. We identify partitions of $[n]$ into k classes (some of which can be empty) with maps $\tau \colon [n] \to [k]$, where $V_i = V_i(\tau) = \tau^{-1}(\{i\})$ is the i^{th} class of the partition. Given such a map and a $k \times k$ matrix B, we will use B^τ for the $n \times n$ matrix with entries $(B^\tau)_{ij} = B_{\tau(i)\tau(j)}$. Finally, for an $n \times n$ matrix A, we use A_τ to denote the matrix where for each $(x, y) \in V_i \times V_i$, the matrix element A_{xy} is replaced by the average over $V_i \times V_j$, and A/τ to denote the $k \times k$ matrix of block averages

$$(A/\tau)_{ij} = \frac{1}{|V_i|\,|V_j|} \sum_{(u,v) \in V_i \times V_j} A_{uv},$$

defined to be 0 if either V_i or V_j is empty; note that the two are related by $A_\tau = (A/\tau)^\tau$.

As usual, we use S_n to denote the set of permutations on $[n]$. We usually assume that $n \geq 2$ to avoid trivial counterexamples to assertions about graphs or matrices. The *density* of a nonnegative $n \times n$ matrix H is defined as $\rho(H) = \frac{1}{n^2}\sum_{i,j} H_{ij}$, and the *density* $\rho(G)$ of a graph G is defined as the density of its adjacency matrix. We use λ to denote the standard Lebesgue measure on $[0,1]$ (or, when we do not expect this to create confusion, the Lebesgue measure on $[0,1]^2$). We use Δ_k to denote the simplex of probability measures on $[k]$, i.e., $\Delta_k = \{\mathbf{p} = (p_i) \in [0,1]^k : \sum_i p_i = 1\}$. The notation O_p means big-O in probability: if X and Y are random variables, then $X = O_p(Y)$ means for each $\varepsilon > 0$, there exists an M such that $|X| \le M|Y|$ with probability at least $1-\varepsilon$. Finally, we use the abbreviation a.s. for "almost surely" or "almost sure" and i.i.d. for "independent and identically distributed."

We will also consider general probability spaces $(\Omega, \mathcal{F}, \pi)$, where $\mathcal{F}$ is a σ-algebra on Ω and π is a probability measure on Ω with respect to $\mathcal{F}$. As usual, a map $\phi\colon (\Omega, \mathcal{F}, \pi) \to (\Omega', \mathcal{F}', \pi')$ is called *measure preserving* if for all $F' \in \mathcal{F}'$, $\phi^{-1}(F') \in \mathcal{F}$ and $\pi(\phi^{-1}(F')) = \pi'(F')$. We call such a map an *isomorphism* if it is a bijection and its inverse is measure preserving as well, and an *isomorphism modulo* 0 if, after removing sets of measure zero from Ω and Ω', it becomes an isomorphism between the resulting probability spaces.

In addition to the distance δ_p, we also consider the (in general larger) distance $\hat{\delta}_p(A,B)$ between two $n \times n$ matrices A, B, defined as

$$\hat{\delta}_p(A,B) = \min_{\sigma \in S_n} \|A^\sigma - B\|_p. \tag{3.1}$$

Note that by definition, $\hat{\delta}_p(A,B)$ is a distance invariant under relabeling; i.e., it is a distance on equivalence classes of $n \times n$ matrices with respect to relabeling of the "vertices" in $[n]$. We will need a similar version of the cut distance $\|A - B\|_\square$. It is defined as

$$\hat{\delta}_\square(A,B) = \min_{\sigma \in S_n} \|A^\sigma - B\|_\square, \tag{3.2}$$

where $\|\cdot\|_\square$ is defined in (2.4).

As pointed out in Section 1, for any practical application it is not enough merely to obtain a close approximation to a matrix under a metric such as $\hat{\delta}_p$ or $\hat{\delta}_\square$. Instead, it is important to obtain the relabeling σ as well. All our algorithms have this property: they do not simply produce good approximations in the abstract, but also explicit relabelings.

Note also that the L^2 norm is related to a scalar product $\langle\cdot,\cdot\rangle$ via $\|A\|_2^2 = \langle A, A\rangle$, with the scalar product between two $n \times n$ matrices A, B defined as

$$\langle A, B\rangle = \frac{1}{n^2}\sum_{i,j \in [n]} A_{ij}B_{ij}.$$

3.2. *Graphons and the cut metric.* Given a probability space $(\Omega, \mathcal{F}, \pi)$, a measurable function $W \colon \Omega\times\Omega \to \mathbb{R}$ is called *symmetric* if $W(x, y) = W(y, x)$ for all $x, y \in \Omega$. We call such a function a *graphon* if it takes nonnegative values and $\|W\|_1 < \infty$, where as usual, the L^p norm of a function $f \colon \Omega\times\Omega \to \mathbb{R}$ is defined by $\|f\|_p^p = \int_{\Omega\times\Omega} |f(x, y)|^p \, d\pi(x)\, d\pi(y)$. We call W an *L^p graphon* if $\|W\|_p < \infty$, and we say that W is *normalized* if $\|W\|_1 = 1$.

We will refer to W as a graphon over $(\Omega, \mathcal{F}, \pi)$, or often just as a graphon over Ω when the σ-algebra $\mathcal{F}$ and the probability measure π are clear from the context. For example, when we say that W is a graphon over $[0, 1]$, we mean that W is a graphon over $[0, 1]$ equipped with the Borel σ-algebra and the uniform measure, unless stated otherwise.

Note that graphs are special cases of graphons: given a graph G with vertex set V and adjacency matrix A, we view it as a graphon on V by equipping V with the uniform distribution and choosing $W(u, v)$ to be A_{uv}. We can also embed graphs into graphons over $[0, 1]$ by first dividing $[0, 1]$ into n adjacent intervals $I_1, \dots, I_n$ of length $1/n$ and then setting the graphon equal to A_{uv} on $I_u \times I_v$. The resulting graphon is called the *empirical graphon corresponding to G* and will be denoted by $\mathsf{W}[G]$. Note that $\mathsf{W}[G]$ is a pull-back of the graph G considered as a graphon on $[n]$ (under the map ϕ sending $I_k \subseteq [0, 1]$ to the point $k \in [n]$), so in particular G and $\mathsf{W}[G]$ are equivalent as graphons.

In addition to the L^p norm of a graphon W, we will also use the cut norm $\|W\|_\square$, defined as

$$\|W\|_\square = \sup_{S,T\subseteq\Omega} \left| \int_{S\times T} W(x, y)\, d\pi(x)\, d\pi(y) \right|,$$

where the supremum is over measurable subsets of Ω (i.e., elements of $\mathcal{F}$). The corresponding metric is defined for a pair of graphons W and W' on two probability spaces $(\Omega, \mathcal{F}, \pi)$ and $(\Omega', \mathcal{F}', \pi')$ by

$$\begin{aligned} &\delta_\square(W, W') = \\ &\qquad \inf_\nu \sup_{S,T\subseteq\Omega\times\Omega'} \left| \int_{S\times T} \Big(W(x, y) - W'(x', y') \Big)\, d\nu(x, x')\, d\nu(y, y') \right|, \end{aligned} \tag{3.3}$$

where the infimum is over couplings ν of the two measures π and π' and the supremum is over measurable subsets of $\Omega \times \Omega'$. Because graphs are special cases of graphons, this in particular defines a distance between a graph and an arbitrary graphon.

Remark 3.1. (i) We will often consider graphons over $[0, 1]$ (with the Borel σ-algebra unless otherwise specified). For such graphons, both the cut distance $\delta_\square$ and the L^p distance δ_p can be defined in a simpler way. Specifically,

$$\begin{aligned} \delta_p(W, W') &= \inf_\Phi \|W^\Phi - W'\|_p \qquad \text{and} \\ \delta_\square(W, W') &= \inf_\Phi \|W^\phi - W'\|_\square, \end{aligned} \tag{3.4}$$

where the infima over Φ are over isomorphisms from $[0, 1]$ to itself. In fact, this simpler definition is equivalent to the definitions (1.1) and (3.3) for many spaces used in practice, as long as they are atomless. See Lemma A.7 in Appendix A for the precise setting. This lemma also shows that for many spaces of interest, the infima in the expressions (1.1) and (3.3) are actually minima.

(ii) When comparing a finite graph G to a graphon W over $[0, 1]$, we will sometimes use an extension of the definition (3.1). It is defined as

$$\hat{\delta}_p(A, W) = \min_{\sigma \in S_n} \|\mathsf{W}[A^\sigma] - W\|_p, \tag{3.5}$$

where $(A^\sigma)_{ij} = A_{\sigma(i)\sigma(j)}$ and $\mathsf{W}[\cdot]$ is the empirical graphon defined above.

3.3. *Examples of W-random graphs.* Recall the definitions of a W-random graph at target density ρ, denoted by $G_n = G_n(\rho W)$, from Section 1 and the definition of the matrix $Q_n = Q_n(\rho W)$ from Section 2. Considering Q_n as a weighted graph on n vertices, we often call it a *weighted W-random graph* at target density ρ. Before giving a few examples, we note that for $n \geq 2$, the expected densities of the graphs Q_n and G_n are $\|\min\{1, \rho W\}\|_1$, which is $(1 + o(1))\rho\|W\|_1$ provided $\rho = \rho_n \to 0$ as $n \to \infty$. That is why we call ρ the target density for Q_n and G_n.

Example 3.2 (Stochastic block model on k blocks). Let $\Omega = [k]$, and let the probability distribution π on Ω be given by a vector $\mathbf{p} = (p_1, \ldots, p_k) \in \Delta_k$. Setting $W(i, j) = \beta_{ij}$ for some symmetric matrix $B = (\beta_{ij})$ of nonnegative numbers then describes the standard stochastic block model with parameters $(\mathbf{p}, B)$. We denote the set of all block models on k blocks by $\mathcal{B}_k$ and use $\mathcal{B}$ to denote the union $\mathcal{B} = \bigcup_{k \geq 1} \mathcal{B}_k$. For $\kappa \in (0, 1/2]$, we use $\mathcal{B}_{\geq\kappa}$ to denote all block models $(\mathbf{p}, B)$ such that $p_i \geq \kappa$ for all i.

Alternatively, we can use the uniform distribution over the interval $[0, 1]$ as our probability space. Then we define $\widetilde{W}$ by first partitioning $[0, 1]$ into k adjacent intervals of lengths $p_1, \ldots, p_k$, and then setting $\widetilde{W}$ equal to β_{ij} on $I_i \times I_j$. Note that the random graphs generated by W and $\widetilde{W}$ are equal in distribution. We denote the graphon $\widetilde{W}$ by $\mathsf{W}[\mathbf{p}, B]$, or by $\mathsf{W}[B]$ if all the probabilities p_i are equal.

Note that the output of our three algorithms are block models, in fact, block models whose block sizes are all a multiple of $1/n$.

Example 3.3 (Mixed membership stochastic block model). To express the mixed membership block model of [4] as a W-random graph, we define Ω to be the k dimensional simplex Δ_k and equip it with a Dirichlet distribution with some parameters $\alpha = (\alpha_1, \dots, \alpha_k)$. In other words, the probability density at $(p_1, \dots, p_k)$ is proportional to $\prod_i p_i^{\alpha_i - 1}$. Given a symmetric matrix (β_{ij}) of nonnegative numbers, we then define

$$W(\mathbf{p}, \mathbf{p}') = \sum_{i,j} \beta_{ij} p_i p'_j.$$

As in the stochastic block model, β_{ij} describes the affinity between communities i and j, but now each vertex is assigned a probability distribution $\mathbf{p}$ over the set of communities (rather than being assigned a single community).

3.4. *Equivalence and identifiability.* In this section, we discuss the notion of equivalence introduced in the context of in Theorem 1.1. We start with the following remark.

Remark 3.4. As claimed in the introduction, the metric (1.1) is indeed a distance on equivalence classes; in other words, $\delta_p(W, W') = 0$ if W and W' are equivalent. To see this, let ϕ and ϕ' be measure preserving transformations such that a.s., $W = U^{\phi}$ and $W' = U^{\phi'}$ for some graphon U over $(\Omega'', \mathcal{F}'', \pi'')$. Define a coupling $d\nu(x, x'')$ of π'' and π by choosing $x \in \Omega$ according to π and then setting $x'' = \phi(x)$. Using this coupling, it is easy to see that $\delta_p(U, W) = 0$. Similarly, $\delta_p(U, W') = 0$, which together with the triangle inequality proves the claim.

Theorem 3.5. *Let W be a graphon over an arbitrary probability space $(\Omega, \mathcal{F}, \pi)$. Then there exists an equivalent graphon over $[0, 1]$ equipped with the uniform distribution.*

The theorem follows easily from the results of [16]. See Appendix A, where we also show that every graphon is equivalent to a twin-free graphon (Theorem A.5).

Our next theorem gives a different characterization of equivalence in terms of the metrics δ_p and $\delta_\square$.

Theorem 3.6. *Let $p \geq 1$, and let W and W' be L^p graphons over two arbitrary probability spaces. Then the following statements are equivalent:*

(i) $\delta_\square(W, W') = 0$*;*
(ii) $\delta_p(W, W') = 0$*;*
(iii) W *and* W' *are equivalent.*

The theorem follows again from the results of [16], even though the details are a little more involved than for the previous theorem and in particular make use of the fact that the infimum in (3.3) is actually a minimum if the underlying space is the unit interval. See Appendix A for the proof.

Remark 3.7. In a purely measure-theoretic study of W-random graphs, we could restrict our attention to graphons over the unit interval without any loss of generality, since by Theorem 3.5, every integrable graphon W is equivalent to a graphon W' defined over $[0, 1]$. However, when W is given in an application, it is often a continuous function over a higher dimensional space, and while W' leads to the same distribution of W-random graphs, the transformation from W to W' ruins continuity, which is often needed to prove good approximation bounds. For applications, the general setup is therefore more natural.

3.5. *Relation to graph convergence.* As mentioned before, W-random graphs arise very naturally as nonparametric models when considering a given graph as a finite subgraph of an infinite, exchangeable array, at least in the dense setting. Indeed, as the works of Hoover [42] and Aldous [6] show, any graph which is an induced subgraph of an infinite, exchangeable array can be modeled as a W-random graph for some (possibly random) graphon W.

A different window into the theory of W-random graphs is given by the theory of graph convergence. Here one asks when a sequence of graphs G_n should be considered convergent. Motivated by extremal combinatorics, one way to address this question is to define a sequence of graphs to be convergent if the number of subgraphs isomorphic to a given graph H converges for every finite graph H, once suitably normalized. It turns out that in the dense setting, this notion is equivalent to many other natural notions of graph convergence that are relevant in computer science, statistical physics, and other fields [17, 18, 19].

One of these equivalent notions is convergence in metric, defined in terms of the cut metric (3.3). We say that a sequence of dense graphs *converges to a graphon* W *in metric* if $\delta_\square(G_n, W) \to 0$ as $n \to \infty$. Note that the limit W is not unique, since two graphons W and W' that are equivalent have distance $\delta_\square(W, W') \le \delta_1(W, W') = 0$. The results of [16] imply that this is the only ambiguity: if W and W' are such that $\delta_\square(G_n, W) \to 0$ and $\delta_\square(G_n, W') \to 0$, then W and W' are equivalent.

Given this notion of convergence, one may ask whether all sequences of graph G_n have a limit, or whether they at least have a subsequence which converges in the metric $\delta_\square$. For dense graphs, the answer to this question is yes and was given in [51], where it was shown that every sequence of dense graphs has a subsequence that is a Cauchy sequence in the metric $\delta_\square$, and that every Cauchy sequence converges to a graphon over $[0, 1]$.

Thus the results of [51] completely parallel the results on exchangeable arrays of [42, 6]: given an ergodic component of an infinite, exchangeable graph, one can find a graphon over $[0, 1]$ that generates this array, and given an arbitrary sequence of (random or non-random) dense graphs, one can find a subsequence and a graphon over $[0, 1]$ such that the subsequence converges to that graphon. In both cases, the graphon is identifiable only up to equivalence. Finally, combining [51] with [16], we know that if the sequence of graphs happens to be a sequence of W-random graphs, then it converges a.s., and the generating graphon W is a representative from the equivalence class of limits.

The net result of this theory is that a convergent sequence of dense networks behaves like a sequence of W-random graphs for some graphon W and can thus be viewed as W-quasi-random graphs. Having established this connection between W-random graphs and W-quasi-random graphs in the dense setting, one might ask whether it can be extended to a convergence theory for sparse graph sequences. The answer turns out to be yes, provided we modify the definition of convergence in metric appropriately. To this end we define, for an arbitrary graph G with adjacency matrix $A(G)$ and a constant $c \in \mathbb{R}$, the graph cG to be the weighted graph with adjacency matrix $cA(G)$.

Definition 3.8. Let W be a graphon over an arbitrary probability space. A sequence of graphs G_n *converges to W in metric* if

$$\delta_\square\Big(\frac{1}{\rho(G_n)} G_n, W\Big) \to 0 \qquad \text{as } n \to \infty.$$

In this case, we call G_n a *W-quasi-random sequence with target density* $\rho(G_n)\|W\|_1$.

Remark 3.9. This definition is an extension of the one given in [15] for graphons W over $[0, 1]$. There, as in the earlier literature on graph convergence for dense graphs, the distance between a graph G and a graphon W was defined as the distance between W and the embedding $\mathsf{W}[G]$ of

G into the space of graphons over $[0,1]$. In our setting, this embedding is not needed, since the cut distance (3.3) is defined on equivalence classes of graphons, and G and its embedding $\mathsf{W}[G]$ are equivalent.

Given the above definition of convergence for sparse graphs, one might ask whether this notion is again equivalent to other notions of convergence, and whether sparse W-random graphs converge again to the generating graphon. The answer to both questions is yes, with one exception: convergence of subgraph counts is no longer equivalent to convergence in metric. But all other notions of convergence proved to be equivalent for dense graphs in [19] remain equivalent in the sparse setting, as shown in [14]. It is also again true that a sequence of W-random graphs converges to the generating graphon. This is the content of the following theorem.

Theorem 3.10. *Let $G_n = G_n(\rho_n W)$ where W is a normalized graphon over an arbitrary probability space, and $\rho_n \to 0$ in such a way that $n\rho_n \to \infty$. Then a.s. $\rho(G_n)/\rho_n \to 1$ and*

$$\delta_\square\Big(\frac{1}{\rho(G_n)}G_n, W\Big) \to 0.$$

Proof. For graphons over $[0,1]$, this theorem was established in [15]. The general case follows from observing that by Theorem 3.5, we can find a graphon over $[0,1]$ that is equivalent to W. Since equivalent graphons lead to identically distributed random graphs, this proves the claim. □

Remark 3.11. (i) Theorems 3.10 and 3.6 immediately imply Theorem 1.1. Indeed, let $G_n = G_n(\rho_n W)$ and $G'_n = G_n(\rho_n W')$. By Theorem 3.10, $\delta_\square(\frac{1}{\rho_n}G_n, W) \to 0$ and $\delta_\square(\frac{1}{\rho_n}G'_n, W') \to 0$, and hence $\delta_\square(W, W') = 0$ if G_n and G'_n are identically distributed. By Theorem 3.6, this implies that W and W' are equivalent. Since, on the other hand, G_n and G'_n are clearly identically distributed if W and W' are equivalent, this proves Theorem 1.1.

(ii) Theorem 3.10 has interesting consequences for graphon estimation. Assume that an algorithm produces an estimator $\widehat{W}$ for the generating graphon W which is close in δ_p for $p \geq 1$. These distances dominate the invariant L^1 distance δ_1, which in turn dominates the cut distance $\delta_\square$. Combined with the results from [14] which state that many other notions of convergence are equivalent to convergence in metric (see Theorem 2.10), we obtain that consistent approximation for W leads to consistent approximations for various quantities of interest, such as minimal energies of graphical models defined on G_n (see Proposition 5.11 in [14], which actually gives a quantitative bound in terms of the cut distance) or collections of cuts in G_n (see Lemma 5.10 in [14], which again gives a quantitative bound). By Theorem B.1 below, we also get good approximations for the empirical distributions of the degrees of G_n.

4. Least squares estimation. In this section, we present the main steps in the proof of Theorem 2.1. They are based on the following two observations. First, for any map $\tau\colon [n] \to [k]$ and any $k \times k$ matrix B,

$$\|A(G) - B^\tau\|_2^2 = \|A(G)\|_2^2 - 2\langle A(G), B^\tau\rangle + \|B^\tau\|_2^2.$$

Therefore, the argmin of the left side is the argmax of $2\langle A(G), B^\tau\rangle - \|B^\tau\|_2^2$. Second, conditioned on the weighted W-random graph $Q = Q_n(\rho_n W)$,

$$\mathbb{E}\Big[2\langle A(G), B^\tau\rangle - \|B^\tau\|_2^2\Big] = 2\langle Q, B^\tau\rangle - \|B^\tau\|_2^2.$$

Up to errors stemming from imperfect concentration, we therefore expect that the argmin $(\hat{B}, \hat{\tau})$ from (2.3) is a maximizer for $2\langle Q, B^\tau\rangle - \|B^\tau\|_2^2$, and hence a minimizer for $\|Q - B^\tau\|_2$. Thus, we would expect that, again up to issues of concentration, the L^2 error is bounded by a term $\hat{\varepsilon}^{(2)}_{\geq\kappa}(Q)$ defined as follows. For an arbitrary $n \times n$ matrix H, we set

$$\hat{\varepsilon}^{(2)}_{\geq\kappa}(H) = \min_{B \in \mathcal{A}_{n,\geq\kappa}} \|H - B\|_2,$$

where $\mathcal{A}_{n,\geq\kappa}$ is the set of all $n \times n$ matrices made up of constant blocks of size at least $\lfloor \kappa n \rfloor$.

For bounded graphons, this strategy was implemented in [20], leading to a proof of consistency for all bounded graphons W and a differentially private algorithm achieving the same goal under slightly less general conditions (requiring ρn to grow at least like $\log n$). For the case of general L^2 graphons, the above motivation still lies behind our proof, but the actual implementation proceeds along slightly different lines, and combines elements of the (sparse graph) strategy of [20] with elements of the (dense graph) strategy developed in [37].

The resulting estimates are stated in Theorem 4.1, which bounds the L^2 difference between the output of the algorithm (2.3) and the matrix Q in terms of $\hat{\varepsilon}^{(2)}_{\geq\kappa}(Q)$ and an error term representing errors from imperfect concentration. To obtain Theorem 2.1 from Theorem 4.1, we will need to transform an estimate on the L^2 error with respect to Q into an L^2 error with respect to W, and we will want to express the result in terms of $\varepsilon^{(2)}_{\geq\kappa}(W)$

instead of $\hat{\varepsilon}^{(2)}_{\geq\kappa}(Q)$. This leads to two extra error terms, the last two terms in the bound of statement (i) in Theorem 2.1; see Appendix D for details.

To state Theorem 4.1 formally, we note that the output of the least squares algorithm consists of block models of the form $\widehat{W} = (\hat{p}, \widehat{B})$, where $\hat{p}$ is an integer multiple of $1/n$. As a consequence, it can equivalently be represented by an $n \times n$ matrix $M_n(\widehat{W}) \in \mathcal{A}_{n,\geq\kappa}$, where $M_n(\widehat{W})$ is the block matrix with entries $\widehat{B}_{ij}$ and block sizes $p_i n \times p_j n$.

Theorem 4.1. *Let W be an L^2 graphon, normalized so that $\|W\|_1 = 1$, let $0 < \rho, \kappa \leq 1$ and $n \in \mathbb{N}$, let $G = G_n(\rho W)$ and $Q = Q_n(\rho W)$, and let $\widehat{W} = (\hat{p}, \hat{B})$ be the output of the least squares algorithm (2.3) with input G. If $n\kappa > 1$ and $\frac{1+\log(1/\kappa)}{\kappa^2} = O(\rho n)$, then*

$$\hat{\delta}_2\Big(M_n(\widehat{W}), Q\Big) \leq \hat{\varepsilon}^{(2)}_{\geq\kappa}(Q) + O_p\left(\rho\sqrt[4]{\frac{1+\log(1/\kappa)}{\kappa^2\rho n}}\right),$$

where the constant implicit in the O_p symbol depends on the L^2 norm of W.

If $\rho = \rho_n$ is such that $n\rho_n \to \infty$ and $\rho_n \to 0$, then almost surely, for n large enough and all κ with $n\kappa > 1$ and $\frac{1+\log(1/\kappa)}{\kappa^2} = O(\rho n)$,

$$\hat{\delta}_2\Big(M_n(\widehat{W}), Q\Big) \leq \hat{\varepsilon}^{(2)}_{\geq\kappa}(Q) + O\left(\rho\sqrt[4]{\frac{1+\log(1/\kappa)}{\kappa^2\rho n}}\right),$$

where again the constant implicit in the big-O symbol depends on the L^2 norm of W.

Proof. Let $\widehat{M} = M_n(\widehat{W})$, $A = A(G)$, and $k = \lceil \frac{n}{\lfloor\kappa n\rfloor} \rceil$. As a first step, we will prove that

$$\hat{\delta}_2\Big(\widehat{M}, Q\Big) \leq \hat{\varepsilon}^{(2)}_{\geq\kappa}(Q) + 2k\varepsilon + 2\sqrt{k\varepsilon\|Q\|_2}, \tag{4.1}$$

where

$$\varepsilon = \max_{\tau\colon [n]\to[k]} \|A_\tau - Q_\tau\|_1.$$

To prove (4.1) we note that $\widehat{M} = M_n(\widehat{W})$ is a minimizer of $\|A - M\|_2$ over all $M \in \mathcal{A}_{n,\geq\kappa}$. As a consequence,

$$-2\langle A, \widehat{M}\rangle + \|\widehat{M}\|_2^2 \leq -2\langle A, M\rangle + \|M\|_2^2$$

for all $M \in \mathcal{A}_{n,\geq\kappa}$, which in turn implies that

$$\begin{aligned}\hat{\delta}_2\left(\widehat{M}, Q\right)^2 &\leq \left\|\widehat{M} - Q\right\|_2^2 \\ &\leq \|M\|_2^2 - 2\left\langle \widehat{M}, Q\right\rangle + \left\|Q\right\|_2^2 + 2\left\langle \widehat{M} - M, A\right\rangle \\ &= \left\|M - Q\right\|_2^2 + 2\left\langle \widehat{M} - M, A - Q\right\rangle.\end{aligned}$$

Since $M, \widehat{M} \in \mathcal{A}_{n,\geq\kappa}$, we know that there are partitions $\tau, \hat{\tau} \colon [n] \to [k]$ such that $M = M_\tau$, $\widehat{M} = \widehat{M}_{\hat{\tau}}$, and all non-empty classes of τ and $\hat{\tau}$ have size at least $\lfloor \kappa n \rfloor$. As a consequence,

$$|\langle M, A - Q\rangle| = |\langle M, (A - Q)_\tau\rangle| \leq \|M\|_\infty \|(A - Q)_\tau\|_1 \leq \varepsilon \|M\|_\infty.$$

Furthermore,

$$\|M\|_\infty \leq \frac{n}{\lfloor \kappa n \rfloor} \|M\|_2 \leq k\|M\|_2,$$

because M is an $n \times n$ block matrix such that each block contains at least $\lfloor \kappa n \rfloor^2$ elements (and thus $n^2\|M\|_2^2 = \sum_{i,j} M_{i,j}^2 \geq \lfloor \kappa n \rfloor^2 \|M\|_\infty^2$). It follows that

$$|\langle M, A - Q\rangle| \leq k\varepsilon\|M\|_2.$$

Bounding $|\langle \widehat{M}, A - Q\rangle|$ in the same way, we find that

$$\hat{\delta}_2\left(\widehat{M}, Q\right)^2 \leq \left\|M - Q\right\|_2^2 + 2k\varepsilon(\|M\|_2 + \|\widehat{M}\|_2).$$

Bounding $\|\widehat{M}\|_2 = \hat{\delta}_2(0, \widehat{M}) \leq \|Q\|_2 + \hat{\delta}_2\left(\widehat{M}, Q\right)$ and $\|M\|_2 \leq \|Q\|_2 + \|M - Q\|_2$, a small calculation then shows that

$$\left(\hat{\delta}_2\left(\widehat{M}, Q\right) - k\varepsilon\right)^2 \leq \left(\left\|M - Q\right\|_2 + k\varepsilon\right)^2 + 4k\varepsilon\|Q\|_2.$$

Choosing M in such a way that $\hat{\varepsilon}^{(2)}_{\geq\kappa}(Q) = \|M - Q\|_2$, this proves (4.1).

For all $\tau \colon [n] \to [k]$, we have $\mathbb{E}[A_\tau \mid Q] = Q_\tau$. Using this fact and a concentration argument, one can show that conditioned on Q, with probability at least $1 - e^{-n}$

$$\varepsilon \leq 8\sqrt{\rho(Q)\left(\frac{1 + \log k}{n} + \frac{k^2}{n^2}\right)}, \tag{4.2}$$

whenever $\rho(Q)n \geq 1$; see Lemma H.2 in Appendix H. The lemma also gives a bound on the expectation, implying in particular that conditioned on Q,

$$\varepsilon = O_p\left(\sqrt{\rho(Q)\left(\frac{1 + \log k}{n} + \frac{k^2}{n^2}\right)}\right),$$

whether or not the condition $\rho(Q)n \geq 1$ holds. (Specifically, the probability that ε exceeds this bound by a factor of M can be no larger than $1/M$, or else the expectation would be too large.)

Since $\mathbb{E}[\rho(Q)] \leq \rho \|W\|_1 = \rho$ and $\mathbb{E}[\|Q\|_2^2] \leq \rho^2 \|W\|_2^2$, this proves that

$$\begin{aligned} 2k\varepsilon + 4\sqrt{k\varepsilon \|Q\|_2} &= O_p\left(\rho \sqrt{\frac{k^2(1+\log k)}{\rho n} + \frac{k^4}{\rho n^2}}\right) \\ &\quad + O_p\left(\rho \sqrt[4]{\frac{k^2(1+\log k)}{\rho n} + \frac{k^4}{\rho n^2}}\right), \end{aligned}$$

with the constant implicit in the O_p symbol depending on $\|W\|_2$. To transform this bound into the bound in the statement of the theorem, we observe that for $\kappa = 1$, $k = \lceil \frac{n}{\lfloor \kappa n \rfloor} \rceil$ is equal to $\frac{1}{\kappa}$, while for $\kappa < 1$, the assumption $n\kappa > 1$ implies that $k = \lceil \frac{n}{\lfloor \kappa n \rfloor} \rceil \leq \frac{3}{2\kappa}$. In either case,

$$\frac{k^2(1+\log k)}{n} = O\left(\frac{1+\log(1/\kappa)}{\kappa^2 n}\right)$$

and

$$\frac{k^4}{n^2} = O\left(\frac{1}{\kappa^4 n^2}\right) = O\left(\left(\frac{1+\log(1/\kappa)}{\kappa^2 n}\right)^2\right) = O\left(\frac{1+\log(1/\kappa)}{\kappa^2 n}\right),$$

where in the last step we used the fact that the assumption $\frac{1+\log(1/\kappa)}{\kappa^2} = O(\rho n)$ implies that $\frac{1+\log(1/\kappa)}{\kappa^2 n} = O(1)$. Thus,

$$\begin{aligned} 2k\varepsilon + 4\sqrt{k\varepsilon \|Q\|_2} &= O_p\left(\rho \sqrt{\frac{1+\log(1/\kappa)}{\kappa^2 \rho n}} + \rho \sqrt[4]{\frac{1+\log(1/\kappa)}{\kappa^2 \rho n}}\right) \\ &= O_p\left(\rho \sqrt[4]{\frac{1+\log(1/\kappa)}{\kappa^2 \rho n}}\right), \end{aligned}$$

because $\frac{1+\log(1/\kappa)}{\kappa^2 \rho n} = O(1)$. This completes the proof of the bound in probability.

To prove the a.s. statement, we note that by Lemma C.6 in Appendix C, $\rho(Q_n)/\rho_n \to 1$, which together with the hypothesis that $n\rho_n \to \infty$ implies that almost surely, $n\rho(Q_n) \geq 1$ holds for sufficiently large n, which allows us to use the bound (4.2). By a simple union bound, this bound holds for all $k \leq n$ with probability at least $1 - ne^{-n}$. Since the failure probability is summable, we conclude that there exists a random n_0 (depending on W and the sequence ρ_n, but not on k or κ) such that the bound (4.2) holds for all $n \geq n_0$ and all $k \leq n$. Combined with the fact that by the law of large numbers for U-statistics (see Lemma C.3 in Appendix C), $\frac{1}{\rho_n}\|Q\|_2 \to \|W\|_2$ a.s. as $n \to \infty$, we obtain the almost sure statement of the theorem. □

As noted above, see Appendix D for the derivation of Theorem 2.1 from Theorem 4.1.

5. Least cut norm estimation. In this section, we outline the main steps in the proof of Theorem 2.2. The proof relies again on a concentration argument, this time starting from the observation that for all $S, T \subseteq [n]$,

$$\mathbb{E}\Big[\sum_{(x,y)\in S\times T} A_{xy}(G)\Big] = \sum_{(x,y)\in S\times T} Q_{xy}. \tag{5.1}$$

Therefore, up to issues of concentration, minimizing the cut distance between $A(G)$ and a block model in

$$\mathcal{B}_{\geq\kappa,n} := \{(\mathbf{p}, B) \in \mathcal{B} : \text{for all } i,\ p_i n \in \mathbb{Z} \text{ and } p_i n \geq \lfloor n\kappa \rfloor\}.$$

is the same as minimizing the cut distance between Q and a block model in $\mathcal{B}_{\geq\kappa,n}$. In other words, up to issues of concentration, one might hope that the distance between Q and the output $\widehat{W}$ of the algorithm (2.5) is just $\hat{\varepsilon}_{\geq\kappa,\square}(Q)$, where for an arbitrary $n \times n$ matrix H,

$$\hat{\varepsilon}_{\geq\kappa,\square}(H) = \min_{B\in\mathcal{A}_{n,\geq\kappa}} \|H - B\|_\square.$$

It turns out that we lose a factor of two with respect to this optimum, due to the fact that in (2.5), we optimize over all block matrices of the form $A(G)_\tau$, rather than all block matrices that are constant on the blocks determined by τ. While these two minimizations are equivalent in the least squares case, they are not here, leading to the loss of a factor of two. (At the cost of an even slower algorithm, this could be cured by redefining the algorithm (2.5) to optimize over all block matrices that are constant on the blocks determined by τ.)

The following theorem states our approximation guarantees with respect to Q. Theorem 2.2 follows from it in essentially the same way as Theorem 2.1 follows from Theorem 4.1; see Appendix E for details. To state the theorem, we recall the definition (3.2) of the distance $\hat{\delta}_\square$.

THEOREM 5.1. *Let W be a normalized L^1 graphon, let $0 < \rho \le 1$ and $n \in \mathbb{N}$, and let $G = G_n(\rho W)$ and $Q = Q_n(\rho W)$. If $\kappa \in (n^{-1}, 1]$ and $\widehat{W} = (\hat{p}, \hat{B})$ is the output of the least cut norm algorithm (2.5) with input G, then*

$$\hat{\delta}_\square\Big(M_n(\widehat{W}), Q\Big) \le 2\hat{\varepsilon}_{\ge\kappa,\square}(Q) + O_p\Bigg(\rho\sqrt{\frac{1}{\rho n}}\Bigg).$$

If $\rho = \rho_n$ is such that $n\rho_n \to \infty$ and $\rho_n \to 0$, then almost surely, for n large enough and all $\kappa \in (n^{-1}, 1]$,

$$\hat{\delta}_\square\Big(M_n(\widehat{W}), Q\Big) \le 2\hat{\varepsilon}_{\ge\kappa,\square}(Q) + O\Bigg(\rho\sqrt{\frac{1}{\rho n}}\Bigg).$$

We will prove the theorem in Appendix E.

6. Degree sorting. To analyze the degree sorting algorithm, it is useful to study the distribution function D_W of the marginal

$$W_x = \int W(x,y)\, d\pi(y),$$

where y is chosen according to π:

$$D_W(\lambda) = \pi(\{x : W_x \le \lambda\}). \tag{6.1}$$

We start with the observation that D_W is continuous if and only if the degree distribution of W is atomless. Graphons with this property have a useful characterization as graphons over $[0,1]$:

LEMMA 6.1. *The degree distribution function D_W of a graphon W is continuous if and only if W is equivalent to a graphon U over $[0,1]$ whose degrees U_x are strictly decreasing in x.*

PROOF. Every graphon W is equivalent to a graphon U over $[0,1]$ by Theorem 3.5, and via monotone rearrangement we can furthermore assume that U_x is weakly decreasing in x (see [57] for a thorough discussion of the measure-theoretic technicalities). Furthermore, the degree distribution of U is atomless if and only if $x \mapsto U_x$ is non-constant on every set of positive measure. In other words, D_U is continuous if and only if U_x is strictly decreasing in x. □

If W is a graphon over $(\Omega, \mathcal{F}, \pi)$ and $\mathcal{P}$ is a partition of Ω into finitely many measurable pieces, then $W_\mathcal{P}$ denotes the step function defined by

$$W_\mathcal{P}(x,y) = \frac{1}{\pi(I)\pi(J)} \int_{I\times J} W(u,v)\, d\pi(u)\, d\pi(v)$$

whenever x is in the part I of $\mathcal{P}$ and y is in the part J. (This is not well defined for parts of measure zero, but they can be ignored.) We will need the following sufficient condition for when averaging over partitions converges under the L^1 norm.

LEMMA 6.2. *Let W be an L^1 graphon over $[0,1]$, and let $\mathcal{P}_1, \mathcal{P}_2, \dots$ be partitions of $[0,1]$ into finitely many measurable pieces. Let $p_{n,\varepsilon}$ be the probability that independent random elements $x, y \in [0,1]$ satisfy $|x-y| \geq \varepsilon$, conditioned on x and y lying in the same part of $\mathcal{P}_n$. If*

$$\lim_{n\to\infty} p_{n,\varepsilon} = 0$$

for each $\varepsilon > 0$, then

$$\lim_{n\to\infty} ||W_{\mathcal{P}_n} - W||_1 = 0.$$

See Appendix C.3 for the proof of Lemma 6.2.

PROOF OF THEOREM 2.3. By Lemma 6.1, we can assume that W is a graphon over $[0,1]$ for which the degrees W_x are strictly decreasing in x.

Let $I_{i,n} = [(i-1)/n, i/n]$, so that $I_{1,n}, I_{2,n}, \dots, I_{n,n}$ form a partition of $[0,1]$ (up to the measure-zero set of their endpoints, which we will ignore). We will assume the vertices of G_n are ordered so that the corresponding sample points in $[0,1]$ satisfy $x_1 < x_2 < \cdots < x_n$, and we view G_n as a graphon over $[0,1]$ via the blocks $I_{i,n}$ and this vertex ordering.

Let $d_1, \dots, d_n$ be the vertex degrees, and set $\bar{d} = (d_1 + \cdots + d_n)/n$. Recall that the degree sorting algorithm works as follows. We choose a permutation σ of $[n]$ such that

$$d_{\sigma(1)} \geq d_{\sigma(n)} \geq \cdots \geq d_{\sigma(n)}$$

and integers $0 = n_0 < n_1 < \cdots < n_k = n$ such that

$$\left| n_i - \frac{in}{k} \right| < 1.$$

Then we define $\tau\colon [n] \to [k]$ by $\tau(j) = i$ if $n_{i-1} < \sigma(j) \leq n_i$. The output of the algorithm is the block model $\widehat{W} = (\hat{p}, \hat{B})$ with $\hat{p}_i = 1/k$ and $\hat{B} = A(G)/\tau$.

Let $V_1, \dots, V_k$ be the preimages of $1, \dots, k$ under τ, and set

$$J_i = \bigcup_{j \in V_i} I_{j,n}.$$

Then $J_1, \dots, J_k$ form a partition $\mathcal{P}_n$ of $[0,1]$, and $\widehat{W}_n$ is equivalent to $(G_n)_{\mathcal{P}_n}$. (Recall that we view G_n as a graphon over $[0,1]$.) We wish to prove that

$$\delta_1\big(\rho_n^{-1}(G_n)_{\mathcal{P}_n}, W\big) \to 0.$$

In fact, we will prove that $\left\|\rho_n^{-1}(G_n)_{\mathcal{P}_n} - W\right\|_1 \to 0$, given our ordering of the vertices of G_n.

We will use the notation established in previous sections, such as Q_n for the weighted random graph used to generate G_n. As shown in Lemma C.6 in Appendix C.2, a.s. $\rho(Q_n)/\rho_n \to 1$ and $\|\rho_n^{-1}Q_n - W\|_1 \to 0$.

We begin with the inequality

$$\begin{aligned}\|\rho_n^{-1}(G_n)_{\mathcal{P}_n} - W\|_1 \le &\left\|\rho_n^{-1}(G_n)_{\mathcal{P}_n} - \rho_n^{-1}(Q_n)_{\mathcal{P}_n}\right\|_1 + \left\|\rho_n^{-1}Q_n - W\right\|_1 + \\ &\left\|\rho_n^{-1}(Q_n)_{\mathcal{P}_n} - \rho_n^{-1}Q_n\right\|_1.\end{aligned}$$

The second term on the right tends to zero a.s. For the first term, we have

$$\left\|\rho_n^{-1}(G_n)_{\mathcal{P}_n} - \rho_n^{-1}(Q_n)_{\mathcal{P}_n}\right\|_1 = \rho_n^{-1}\left\|(G_n)_{\mathcal{P}_n} - (Q_n)_{\mathcal{P}_n}\right\|_1.$$

Using Lemma H.2 and $\rho(Q_n)/\rho_n \to 1$ a.s., we can bound $\left\|(G_n)_{\mathcal{P}_n} - (Q_n)_{\mathcal{P}_n}\right\|_1$ by $O\left(\sqrt{\rho\left(\frac{1+\log k}{n} + \frac{k^2}{n^2}\right)}\right)$ a.s., and thus the hypotheses $\log k_n = o(n\rho_n)$ and $k_n = o\big(n\sqrt{\rho_n}\big)$ imply that

$$\left\|\rho_n^{-1}(G_n)_{\mathcal{P}_n} - \rho_n^{-1}(Q_n)_{\mathcal{P}_n}\right\|_1 \to 0.$$

All that remains is to handle the third term, $\left\|\rho_n^{-1}(Q_n)_{\mathcal{P}_n} - \rho_n^{-1}Q_n\right\|_1$. Because $\left\|\rho_n^{-1}Q_n - W\right\|_1 \to 0$, it will suffice to show that $\left\|W_{\mathcal{P}_n} - W\right\|_1 \to 0$. We will do so using Lemma 6.2.

Fix $\varepsilon > 0$, and let $p_{n,\varepsilon}$ be the probability that independent random elements $x, y \in [0,1]$ satisfy $|x-y| \ge \varepsilon$, conditioned on x and y lying in the same part of $\mathcal{P}_n$. By contrast, let $p'_{n,\varepsilon}$ be the probability that $|x-y| \ge \varepsilon$ and both points lie in the same part of $\mathcal{P}_n$, without the conditioning. Because each part J_i of $\mathcal{P}_n$ satisfies $\lambda(J_i) = (1+o(1))/k_n$, proving that $p_{n,\varepsilon} \to 0$ is equivalent to proving that $k_n p'_{n,\varepsilon} \to 0$. Thus, to apply Lemma 6.2, we must show that $k_n p'_{n,\varepsilon} \to 0$.

Instead of analyzing the points x and y, it will be convenient to consider the intervals $I_{\ell,n}$ and $I_{m,n}$ containing them. We will use the bound

(6.2)
$$\begin{aligned}p'_{n,\varepsilon} &\le \Pr_{\ell,m\in[n]}\big(\tau(\ell) = \tau(m) \text{ and } \max\{|x-y| : x \in I_{\ell,n}, y \in I_{m,n}\} \ge \varepsilon\big)\\ &= \Pr_{\ell,m\in[n]}\big(\tau(\ell) = \tau(m) \text{ and } |\ell/n - m/n| \ge \varepsilon - 1/n\big),\end{aligned}$$

where of course $\Pr_{\ell,m\in[n]}$ denotes the probability if ℓ and m are chosen uniformly at random from $[n]$.

To analyze these probabilities, we need to bound how close the degrees in G_n are to those in W. Convergence of degree distributions is analyzed in Appendix B, and Lemma B.2 provides suitable bounds. To apply this lemma, we must quantify how quickly the degrees in W change as a function of distance. Let

$$\delta = \inf_{|x-y| \geq \varepsilon/4 - 1/n} |W_x - W_y|.$$

Because $x \mapsto W_x$ is strictly decreasing, $\delta > 0$. Call an element $i \in [n]$ *good* if the normalized degree $d_i/\bar{d}$ is within $\delta/3$ of W_x for some $x \in I_{i,n}$. Taking $U = \rho_n^{-1} G_n$ in Lemma B.2 shows that the fraction of bad elements is at most

$$\frac{2}{\delta/3} \|\rho_n^{-1} G_n - W\|_\square,$$

which tends to zero as $n \to \infty$. If i and j are good and $|i/n - j/n| \geq \varepsilon/4$, then

$$\left| \frac{d_i}{\bar{d}} - \frac{d_j}{\bar{d}} \right| \geq \delta/3.$$

It follows that if i and j are good and $|i/n - j/n| \geq 3\varepsilon/4$, then at least the middle $\lfloor n\varepsilon/4 \rfloor$ vertices between i and j have degrees strictly between d_i and d_j. When n is large enough, this is much larger than the number of vertices in any part of $\mathcal{P}_n$. In particular, if n is large enough then good i and j with $|i/n - j/n| \geq 3\varepsilon/4$ cannot possibly end up in the same part after the degrees are sorted.

Thus, by (6.2),

$$\begin{aligned} p'_{n,\varepsilon} &\leq \Pr_{\ell,m \in [n]} \big(\ell \text{ or } m \text{ is bad and } \tau(\ell) = \tau(m)\big) \\ &\leq 2 \Pr_{\ell,m \in [n]} \big(\ell \text{ is bad and } \tau(\ell) = \tau(m)\big) \\ &\leq 2 \Pr_{m \in [n]} \big(\ell \text{ is bad}\big) \max_i \lambda(J_i) \\ &\leq \frac{4}{\delta/3} \|\rho_n^{-1} G_n - W\|_\square \frac{1 + o(1)}{k_n}. \end{aligned}$$

It now follows from $\|\rho_n^{-1} G_n - W\|_\square \to 0$ that $k_n p'_{n,\varepsilon} \to 0$, as desired. □

7. Hölder-continuous graphons.

In this section, we discuss the least squares and the least cut norm algorithms for the case of Hölder-continuous graphons. As discussed in the introduction, our approach allows us to reduce this to the analysis of the two error terms $\text{tail}_\rho^{(p)}(W)$ and $\varepsilon_{\geq \kappa}^{(p)}(W)$ for $p = 2$ and $p = 1$, respectively, which reduces the analysis to pure approximation theory.

Throughout this section, we consider graphons W over $\mathbb{R}^d$ (equipped with the standard Borel σ-algebra and some probability measure π) that are α-Hölder-continuous for some $\alpha \in (0,1]$, i.e., graphons W for which there exists a constant C such that

$$|W(x,y) - W(x',y)| \leq C|x-x'|_\infty^\alpha \quad \text{for all } x, x', y \in \mathbb{R}^d,$$

with $|\cdot|_\infty$ denoting the L^∞ distance on $\mathbb{R}^d$ (note that we only require this for one of the two coordinates of W, since for the other one it follows from the fact that W is symmetric). We denote the set of graphons obeying this bound by $\mathcal{H}_{C,\alpha}$. If we restrict ourselves to graphons on a subset Λ of $\mathbb{R}^d$, we use the notation $\mathcal{H}_{C,\alpha}(\Lambda)$.

Our first proposition concerns the case when the support of the underlying measure π is compact, in which case we may assume without loss of generality that π is a measure on $\Lambda_R = [-R,R]^d$ for some $R \in [0,\infty)$. Note that many examples of W-random graphs considered in the statistics and machine learning literature fit into this setting, e.g., the mixed membership block model of [4]. Note also that while these models can be mapped onto W-random graphs over $[0,1]$ with the uniform distribution by a measure-preserving map, such a map will typically not do this in a continuous way. So if one wants to use continuity properties of the generating graphon W, one has to analyze it on the original space on which it was defined, not on $[0,1]$.

Proposition 7.1. *Let $d \geq 1$, $R \in [0,\infty)$, $\alpha \in (0,1]$, and $C < \infty$, let π be a probability measure on $\Lambda_R \subseteq \mathbb{R}^d$, and let W be a normalized graphon in $\mathcal{H}_{C,\alpha}(\Lambda_R)$. Then there exists a constant D depending only on R, C, and α such that the following hold:*

(i) We have $\|W\|_\infty \leq D$. So in particular

$$\operatorname{tail}_\rho^{(p)}(W) = 0 \qquad \text{if } \rho \leq \frac{1}{D}.$$

(ii) For $p \geq 1$ and $\kappa > 0$,

$$\varepsilon_{\geq\kappa}^{(p)}(W) \leq 4D\kappa^{\alpha'}, \tag{7.1}$$

where $\alpha' = \frac{\alpha}{p\alpha + d}$. If π is the uniform measure, then the bound (7.1) holds for $\alpha' = \alpha/d$.

This proposition is proved in Appendix F. It generalizes Proposition 2.1 from [37], which is the case when $d = 1$ and π is the uniform measure.

(However, we do not obtain tight bounds on convergence here, because Theorem 2.1 is not tight.)

In many applications, the underlying measure on the latent position space Ω does not have compact support. Gaussians are a noteworthy case, as are distributions with heavier tails (such as Student distributions). Another reason to consider measures without compact support comes from the desire to model graphs with power-law degree distributions. As discussed already in Section 1.2, bounded graphons do not allow for power-law degree distributions, showing in particular that Hölder-continuous graphons over $\mathbb{R}^d$ equipped with a measure with compact support do not lead to graphs that exhibit power-law degree distributions. For measures with non-compact support, this reasoning no longer applies, and as shown in Appendix G, there are indeed Hölder-continuous graphons over $\mathbb{R}^d$ that generate graphs with power-law degree distributions. For all these reasons, we aim for a generalization of Proposition 7.1 to measures whose supports are not necessarily compact.

Since we want graphons to be integrable (in fact, for the least squares algorithm to be consistent, we need them to be square integrable) we will restrict ourselves to probability distributions π over $\mathbb{R}^d$ in

$$\mathcal{M}_\beta = \left\{ \pi \,\middle|\, \int_{\mathbb{R}^d} |x|_\infty^\beta \, d\pi(x) < \infty \right\},$$

where $\beta > 0$ is a parameter which we will choose to be at least α (or at least 2α when we want to guarantee that the graphons in $\mathcal{H}_{C,\alpha}$ are in L^2).

PROPOSITION 7.2. *Let $d \geq 1$ and $\beta \geq \alpha > 0$, let $\pi \in \mathcal{M}_\beta$, and let W be an α-Hölder-continuous graphon over $\mathbb{R}^d$ equipped with the probability distribution π, normalized in such a way that $\|W\|_1 = 1$. If $1 \leq p < \beta/\alpha$ and $\kappa \leq 1/2$, then*

$$\varepsilon_{\geq\kappa}^{(p)}(W) = O\big(\kappa^{\alpha'}\big) \qquad \text{and} \qquad \mathrm{tail}_\rho^{(p)}(W) = O\big(\rho^{\beta'}\big),$$

where $\beta' = \frac{\beta}{p\alpha} - 1$ and $\alpha' = \frac{\alpha}{p\alpha+d}\frac{\beta'}{1+\beta'}$, and the constants implicit in the big-O symbols depend on the distribution π and the constants α, β, p, and C.

This proposition is also proved in Appendix F.

Acknowledgments. We thank David Choi, Sofia Olhede, and Patrick Wolfe for initially introducing us to applications of graphons in machine learning of networks, and, in particular, to the problem of graphon estimation. We are indebted to Sofia Olhede and Patrick Wolfe for numerous helpful discussions in the early stages of this work, to Alessandro Rinaldo for providing valuable feedback on our paper, and to the anonymous referees for their many helpful comments.

References.

[1] Abbe, E., Bandeira, A. S. and Hall, G. (2016). Exact recovery in the stochastic block model. *IEEE Trans. Inform. Theory* **62** 471–487.

[2] Abbe, E. and Sandon, C. (2015). Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. In *56th Annual Symposium on Foundations of Computer Science* 670–688.

[3] Abbe, E. and Sandon, C. (2015). Recovering communities in the general stochastic block model without knowing the parameters. In *Advances in Neural Information Processing Systems 28* 676–684.

[4] Airoldi, E. M., Blei, D. M., Fienberg, S. E. and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

[5] Airoldi, E. M., Costa, T. B. and Chan, S. H. (2013). Stochastic blockmodel approximation of a graphon: theory and consistent estimation. In *Advances in Neural Information Processing Systems 26* 692–700. Extended version with proofs available at arXiv:1311.1731.

[6] Aldous, D. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivar. Anal.* **11** 581–598.

[7] Amini, A. A., Chen, A., Bickel, P. J. and Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* **41** 2097–2122.

[8] Anandkumar, A., Ge, R., Hsu, D. and Kakade, S. M. (2014). A tensor approach to learning mixed membership community models. *J. Mach. Learn. Res.* **15** 2239–2312.

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

[10] Bickel, P. J., Chen, A. and Levina, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist.* **39** 2280–2301.

[11] Bollobás, B., Janson, S. and Riordan, O. (2007). The phase transition in inhomogeneous random graphs. *Random Struct. Algorithms* **31** 3–122.

[12] Bollobás, B. and Riordan, O. (2009). Metrics for sparse graphs. In *Surveys in combinatorics 2009* (S. Huczynska, J. D. Mitchell and C. M. Roney-Dougal, eds.). *London Math. Soc. Lecture Note Ser.* **365** 211–287. Cambridge University Press.

[13] Boppana, R. B. (1987). Eigenvalues and graph bisection: an average-case analysis. In *28th Annual Symposium on Foundations of Computer Science* 280–285.

[14] Borgs, C., Chayes, J. T., Cohn, H. and Zhao, Y. (2018). An L^p theory of sparse graph convergence II: LD convergence, quotients and right convergence. *Ann. Probab.* **46** 337–396.

[15] Borgs, C., Chayes, J. T., Cohn, H. and Zhao, Y. (2019). An L^p theory of sparse graph convergence I: Limits, sparse random graph models, and power law distributions. *Trans. Amer. Math. Soc.* **372** 3019–3062.

[16] Borgs, C., Chayes, J. and Lovász, L. (2010). Moments of two-variable functions and the uniqueness of graph limits. *Geom. Funct. Anal.* **19** 1597–1619.

[17] Borgs, C., Chayes, J. T., Lovász, L., Sós, V. and Vesztergombi, K. (2006). Counting graph homomorphisms. In *Topics in discrete mathematics* (M. Klazar, J. Kratochvíl, M. Loebl, J. Matoušek, R. Thomas and P. Valtr, eds.) 315–371. Springer.

[18] Borgs, C., Chayes, J. T., Lovász, L., Sós, V. and Vesztergombi, K. (2008). Convergent graph sequences I: subgraph frequencies, metric properties, and testing. *Advances in Math.* **219** 1801–1851.

[19] Borgs, C., Chayes, J. T., Lovász, L., Sós, V. and Vesztergombi, K. (2012). Convergent graph sequences II: multiway cuts and statistical physics. *Ann. of Math.* **176** 151–219.

[20] Borgs, C., Chayes, J. T. and Smith, A. (2015). Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems 28* 1369–1377. Extended version with proofs available at arXiv:1506.06162.

[21] Cai, D., Ackerman, N. and Freer, C. (2014). An iterative step-function estimator for graphons. *Preprint, arXiv:1412.2129.*

[22] Chan, S. H. and Airoldi, E. M. (2014). A consistent histogram estimator for exchangeable graph models. In *Proceedings of the 31st International Conference on Machine Learning (JMLR Workshop and Conference Proceedings Volume 32)* 208–216.

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

[24] Chatterjee, S., Diaconis, P. and Sly, A. (2011). Random graphs with a given degree sequence. *Ann. Appl. Probab.* **21** 1400–1435.

[25] Chaudhuri, K., Chung, F. and Tsiatas, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. In *Proceedings of the 25th Annual Conference on Learning Theory (JMLR Workshop and Conference Proceedings Volume 23)* 35.1–35.23.

[26] Chen, Y., Sanghavi, S. and Xu, H. (2012). Clustering sparse graphs. In *Advances in Neural Information Processing Systems 25* 2204–2212.

[27] Chin, P., Rao, A. and Vu, V. (2015). Stochastic block model and community detection in sparse graphs: a spectral algorithm with optimal rate of recovery. In *Proceedings of the 28th Conference on Learning Theory (JMLR Workshop and Conference Proceedings Volume 40)* 391–423.

[28] Choi, D. S. and Wolfe, P. J. (2014). Co-clustering separately exchangeable network data. *Ann. Statist.* **42** 29–63.

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

[30] Coja-Oghlan, A. (2010). Graph partitioning via adaptive spectral techniques. *Combin. Probab. Comput.* **19** 227–284.

[31] Dasgupta, A., Hopcroft, J., Kannan, R. and Mitra, P. (2006). Spectral clustering by recursive partitioning. In *Algorithms – ESA 2006* (Y. Azar and T. Erlebach, eds.). *Lecture Notes in Computer Science* **4168** 256–267. Springer.

[32] Diaconis, P. and Janson, S. (2008). Graph limits and exchangeable random graphs. *Rend. Mat. Appl. (7)* **28** 33–61.

[33] Dyer, M. E. and Frieze, A. M. (1989). The solution of some random NP-hard problems in polynomial expected time. *J. Algorithms* **10** 451–489.

[34] Fienberg, S. E., Meyer, M. M. and Wasserman, S. S. (1985). Statistical analysis of multiple sociometric relations. *J. Amer. Statist. Assoc.* **80** 51–67.

[35] Fishkind, D. E., Sussman, D. L., Tang, M., Vogelstein, J. T. and Priebe, C. E. (2013). Consistent adjacency-spectral partitioning for the stochastic block model

when the model parameters are unknown. *SIAM J. Matrix Anal. Appl.* **34** 23–39.

[36] Frieze, A. and Kannan, R. (1999). Quick approximation to matrices and applications. *Combinatorica* **19** 175–220.

[37] Gao, C., Lu, Y. and Zhou, H. H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* **43** 2624–2652.

[38] Hajek, B., Wu, Y. and Xu, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Trans. Inform. Theory* **62** 2788–2797.

[39] Hajek, B., Wu, Y. and Xu, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming: extensions. *IEEE Trans. Inform. Theory* **62** 5918–5937.

[40] Hoff, P. D., Raftery, A. E. and Handcock, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098.

[41] Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983). Stochastic blockmodels: first steps. *Social Networks* **5** 109–137.

[42] Hoover, D. (1979). Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ.*

[43] Jerrum, M. and Sorkin, G. B. (1998). The Metropolis algorithm for graph bisection. *Discrete Appl. Math.* **82** 155–175.

[44] Kallenberg, O. (1999). Multivariate sampling and the estimation problem for exchangeable arrays. *J. Theoret. Probab.* **12** 859–883.

[45] Klopp, O., Tsybakov, A. B. and Verzelen, N. (2017). Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.* **45** 316–354.

[46] Latouche, P. and Robin, S. (2016). Variational Bayes model averaging for graphon functions and motif frequencies inference in W-graph models. *Stat. Comput.* **26** 1173–1185.

[47] Lauritzen, S. L. (2003). Rasch models with exchangeable rows and columns. In *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 215–232. Oxford University Press.

[48] Lei, J. and Rinaldo, A. (2014). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237.

[49] Lei, J. and Zhu, L. (2017). Generic sample splitting for refined community recovery in degree corrected stochastic block models. *Statist. Sinica* **27** 1639–1659.

[50] Lloyd, J. R., Orbanz, P., Ghahramani, Z. and Roy, D. M. (2012). Random function priors for exchangeable arrays with applications to graphs and relational data. In *Advances in Neural Information Processing Systems 25* **25** 1007–1015.

[51] Lovász, L. and Szegedy, B. (2006). Limits of dense graph sequences. *J. Combin. Theory Ser. B* **96** 933–957.

[52] Massoulié, L. (2014). Community detection thresholds and the weak Ramanujan property. In *2014 ACM Symposium on Theory of Computing* 694–703.

[53] McSherry, F. (2001). Spectral partitioning of random graphs. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science* 529–537.

[54] Olhede, S. C. and Wolfe, P. J. (2014). Network histograms and universality of blockmodel approximation. *Proc. Natl. Acad. Sci. USA* **111** 14722–14727.

[55] Qin, T. and Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems 26* 3120–3128.

[56] Rohe, K., Chatterjee, S. and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915.

[57] Ryff, J. V. (1970). Measure preserving transformations and rearrangements. *J. Math. Anal. Appl.* **31** 449–458.

[58] Snijders, T. A. B. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification* **14** 75–100.

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

[60] Vu, V. (2018). A simple SVD algorithm for finding hidden partitions. *Combin. Probab. Comput.* **27** 124–140.

[61] Wang, Y. J. and Wong, G. Y. (1987). Stochastic blockmodels for directed graphs. *J. Amer. Statist. Assoc.* **82** 8–19.

[62] White, H. C., Boorman, S. A. and Breiger, R. L. (1976). Social structure from multiple networks. I. Blockmodels of roles and positions. *Am. J. Sociol.* **80** 730–780.

[63] Wolfe, P. J. and Olhede, S. C. (2013). Nonparametric graphon estimation. *Preprint, arXiv:1309.5936.*

[64] Yang, J. J., Han, Q. and Airoldi, E. M. (2014). Nonparametric estimation and testing of exchangeable graph models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (JMLR Workshop and Conference Proceedings Volume 33)* 1060–1067.

[65] Yun, S.-Y. and Proutiere, A. (2014). Accurate community detection in the stochastic block model via spectral algorithms. *Preprint, arXiv:1412.7335.*

[66] Yun, S.-Y. and Proutiere, A. (2014). Community detection via random and adaptive sampling. In *Proceedings of the 27th Conference on Learning Theory (JMLR Workshop and Conference Proceedings Volume 35)* 138–175.

C. Borgs
J. T. Chayes
S. Ganguly
University of California, Berkeley
Berkeley, CA 94720, USA
E-mail: borgs@berkeley.edu
jchayes@berkeley.edu
sganguly@berkeley.edu

H. Cohn
Microsoft Research
One Memorial Drive
Cambridge, MA 02142, USA
E-mail: cohn@microsoft.com