AN ITERATIVE STEP-FUNCTION ESTIMATOR FOR GRAPHONS

By Diana Cai†‡, Nathanael Ackerman‡, and Cameron Freer†⋆

†Gamalon, ‡Harvard University, and ⋆Massachusetts Institute of Technology

Exchangeable graphs arise via a sampling procedure from the measurable functions known as graphons. A natural estimation problem is how well we can recover a graphon given a single graph sampled from it. One general framework for estimating a graphon uses stepfunctions obtained by partitioning the nodes of the graph according to some clustering algorithm. We propose an iterative step-function estimator (ISFE) that, given an initial partition, iteratively clusters nodes based on their edge densities with respect to the previous iteration’s partition. We demonstrate the performance of ISFE with respect to various clustering algorithms, and in comparison with other graphon estimation techniques.

1. Introduction. Latent variable models of graphs can be used to model hidden structure in large networks and have been applied to a variety of problems such as community detection (Girvan and Newman, 2002; Newman and Girvan, 2004) and link prediction (Miller et al., 2009). Furthermore, many graphs are naturally modeled as exchangeable when the nodes have no particular ordering (Hoff, 2007). Examples of exchangeable graph models include the stochastic block model (Holland et al., 1983) and its extensions (Kemp et al., 2006), latent feature models (Miller et al., 2009; Palla et al., 2012), and latent distance models (Hoff et al., 2002).

Several key inference problems in exchangeable graph models can be formulated in terms of estimating symmetric measurable functions $W: [0,1]^2 \rightarrow [0,1]$, known as graphons. There is a natural sampling procedure that produces an exchangeable (undirected) random graph from a graphon $W$ by first sampling a countably infinite set of independent uniform random variables $\{U_i\}_{i \in \mathbb{N}}$, and then sampling an edge between every pair of distinct vertices $i$ and $j$ according to an independent Bernoulli random variable with weight $W(U_i, U_j)$. In the case where the graphon is constant or piecewise constant with a finite number of pieces, this procedure recovers the standard notions of Erdős–Rényi graphs and stochastic block models, respectively. But this procedure is much more general; indeed, Aldous (1981) and Hoover (1979) showed, via what can be viewed as a higher-dimensional analogue of de Finetti’s theorem, that the distribution of any exchangeable graph arises from a mixture of such sampling procedures from graphons.

Graphon estimation has been studied in two contexts: (1) graphon function estimation (Choi and Wolfe, 2014; Wolfe and Olhede, 2013), where we are concerned with inverting the entire sampling procedure to recover a measurable function from a single sampled graph, and (2) graphon value estimation (Airoldi et al., 2013; Chan and Airoldi, 2014; Chatterjee, 2014; Gao et al., 2014), where we are interested in inverting just the second step of the sampling procedure, to obtain estimates of the latent values $W(U_i, U_j)$ from a single graph (or several (Airoldi et al., 2013)) sampled using the sequence $\{U_i\}_{i \in \mathbb{N}}$. This is somewhat analogous to the distinction in nonparametric function estimation between MISE and MSE risk, although a key complication here is that the values $U_i$ are latent. In this paper, we mainly address the graphon value estimation problem, which can be

Keywords and phrases: graphon, exchangeable graph, histogram estimator, network model, cut distance
seen as finding the underlying structure of a sample from a graphon, although we briefly address
the function estimation problem.

Graphons are well-approximated by (finite) step-functions in the cut distance (Frieze and Kan-
nan, 1999a,b; Lovász, 2012), a notion of distance between graphs that extends to graphons, which
we describe in Section 2. Although the topology on graphons induced by the cut distance is coarser
than that induced by $L^2$ (as used in MSE and MISE risk), two graphons are close in the cut distance
precisely when their random samples differ by a small number of edges (after reordering). Hence
it is natural to consider graphon estimators that produce step-functions; this has been extensively
studied with the stochastic block model.

A standard approach to approximating graphons using step-functions is to first partition the
vertices of the sampled graph $G$ in some way and then return the step-function graphon determined
by that partition along with the average edge densities in $G$ between classes of the partition. In
this way, every clustering algorithm can be seen to induce a graphon (function or value) estimation
procedure, as we describe in Section 3.1. While many clustering algorithms thereby give rise to
reasonably tractable graphon estimators, one challenge is to produce clustering algorithms that
induce good estimators. In this paper, we introduce a method, motivated by the cut distance, that
takes a vertex partition and produces another partition that yields an improved graphon estimate.
By iterating this method, even better estimates can be obtained. We describe and analyze the
graphon estimator that results from this iterative procedure applied to the result of a clustering
algorithm. Our preliminary results of the empirical performance of this estimation procedure are
promising, as we describe, while a theory to explain its performance remains to be examined.

1.1. Contributions. We propose a graphon estimation procedure, iterative step-function esti-
mation (ISFE), based on iteratively improving a clustering of the vertices of the sampled graph.
At each stage, a new partition of the vertices is formed by considering the average edge densities
between each vertex and each of the classes of the existing partition: vertices are placed in the
same class (of the new partition) when their edge densities are similar with respect to the classes
of the existing partition. By applying this method iteratively to an initial partition, one obtains a
partition whose edge densities in the sampled graph approximate the block structure implicit in
the original graphon via Frieze–Kannan weak regularity.

The initial partition may be obtained using an existing clustering algorithm — or by starting
with some simpler base case, e.g., with the trivial partition (having one class that contains all
vertices), the discrete partition (having one class for each vertex), or a random assignment — in
which cases the iterated method itself can be considered as a new clustering method. By forming
the step-function graphon induced by the partition, this iterative procedure constitutes a graphon
estimator. We describe ISFE in detail in Section 3.

In Section 4, we demonstrate our graphon estimation method on data sampled from several
graphons, such as a continuous gradient and finite and infinite block model graphons. We also
compare ISFE to several other graphon estimation methods and analyze the conditions in which
the behavior of those methods differ substantially from ours. Finally, we apply our method to three
real-world network datasets: the NIPS co-authorship dataset, the arXiv astro-physics co-authorship
dataset, and the epinions network dataset.

2. Background and related work. Throughout this paper, graphs are undirected and simple;
often we consider sequences of graphs that are dense, in that a graph with $n$ vertices has $\Omega(n^2)$
edges. For all natural numbers $n \in \mathbb{N}$, we define a graph on $[n]$ to be a graph with set of vertices
$[n] := \{0, \ldots, n-1\}$. In this case, its adjacency matrix is the $\{0,1\}$-valued $n \times n$ matrix $(G_{ij})_{i,j \in [n]}$,
where $G_{ij} = 1$ iff $G$ has an edge between vertices $i$ and $j$. Graphs on $\mathbb{N}$, and their adjacency matrices, are defined similarly. We write $x \overset{d}{=} y$ when two random variables $x$ and $y$ are equal in distribution, and abbreviate *almost surely* and *almost every* by a.s. and a.e., respectively.

2.1. Graphons. For detailed background on graphons and the relationship between graphons and exchangeable graphs, see the book by Lovász (2012) and surveys by Diaconis and Janson (2008) and Austin (2008). Here we briefly present the key facts that we will use.

A random graph $G$ on $\mathbb{N}$ is **exchangeable** when its distribution is invariant under arbitrary permutations of $\mathbb{N}$. In particular, if such a graph is not a.s. empty, then the marginal probability of an edge between any two vertices is positive.

**Definition 2.1.** A random $\{0,1\}$-valued array $(A_{ij})_{i,j \in \mathbb{N}}$ is (jointly) **exchangeable** when

$$(A_{ij}) \overset{d}{=} (A_{\sigma(i),\sigma(j)})$$

for every permutation $\sigma$ of $\mathbb{N}$.

Note that a random graph on $\mathbb{N}$ is exchangeable precisely when its adjacency matrix is jointly exchangeable. We now define a sampling procedure that produces exchangeable graphs.

**Definition 2.2.** A **graphon** $W$ is a symmetric measurable function $[0,1]^2 \rightarrow [0,1]$.

A graphon can be thought of as a continuum-sized, edge-weighted graph. We can sample from a graphon in the following way.

**Definition 2.3.** Let $W$ be a graphon. The $W$-random graph on $\mathbb{N}$, written $G(\mathbb{N},W)$, has adjacency matrix $(G_{ij})_{i,j \in \mathbb{N}}$ given by the following sampling procedure:

$$U_i \overset{\text{iid}}{\sim} \text{Uniform}[0,1]$$

$$G_{ij} \mid U_i, U_j \overset{\text{iid}}{\sim} \text{Bernoulli}(W(U_i,U_j)) \text{ for } i < j$$

For $n \in \mathbb{N}$, the random graph $G(n,W)$ on $[n]$ is formed similarly.

Every $W$-random graph is exchangeable, as is any mixture of $W$-random graphs. Conversely, the following statement is implied by the Aldous–Hoover theorem, a two-dimensional generalization of de Finetti’s theorem characterizing exchangeable sequences as mixtures of i.i.d. sequences.

**Theorem 2.4 (Aldous (1981); Hoover (1979)).** Suppose $G$ is an exchangeable graph on $\mathbb{N}$. Then $G$ can be written as the mixture of $W$-random graphs $G(\mathbb{N},W)$ for some probability measure on graphons $W$.

The Aldous–Hoover representation has since been extended to higher dimensions, more general spaces of random variables, and weaker notions of symmetry; for a detailed presentation, see Kallenberg (2005).

Since every exchangeable graph is a mixture of graphon sampling procedures, many network models can be described in this way (Hoff, 2007). The stochastic block model (Holland et al., 1983) is such an example, as explored further by Bickel and Chen (2009) and others; it plays a special role as one of the simplest models that can approximate arbitrary graphon sampling procedures,
as we will describe below. Some Bayesian nonparametric models, including the eigenmodel (Hoff, 2007), Mondrian process graph model (Roy and Teh, 2008), and random function model (Lloyd et al., 2012) were built knowing the Aldous–Hoover representation. Furthermore, many other such models are naturally expressed in terms of a distribution on graphons (Lloyd et al., 2012; Orbanz and Roy, 2014), including the infinite relational model (IRM) (Kemp et al., 2006) the latent feature relational model (LFRM) (Miller et al., 2009), and the infinite latent attribute model (ILA) (Palla et al., 2012).

Two different graphons can give rise to the same distribution on graphs. For example, modifying a graphon on a measure zero subset does not change the distribution on graphs. Moreover, applying a measure-preserving transformation to the unit interval, before sampling the graphon, leaves the distribution on graphs unchanged. The following is a consequence of Proposition 7.10 and Equation (10.3) of Lovász (2012).

**Proposition 2.5.** Let $\varphi : [0,1] \to [0,1]$ be a measure-preserving transformation i.e., a map such that $\varphi(U)$ is uniformly distributed for $U \sim \text{Uniform}[0,1]$. Then the graphon $W^\varphi$ defined by $W^\varphi(x,y) = W(\varphi(x),\varphi(y))$ is weakly isomorphic to $W$, in the sense that $G(N,W) \overset{d}{=} G(N,W^\varphi)$.

Thus, the graphon from which an exchangeable graph is sampled is non-identifiable; see Orbanz and Roy (2014, §III.D) for details. Hence the appropriate object to estimate is a graphon up to weak equivalence. Such measure-preserving transformations are essentially the only freedom allowed.

**Theorem 2.6 (Hoover (1979)).** If $W_0$ and $W_1$ are weakly isomorphic, then there are measure-preserving transformations $\varphi_0, \varphi_1$ and a graphon $V$ such that $W_0^{\varphi_0} = W_1^{\varphi_1} = V$ a.e.

As a result of Theorem 2.6, when considering the problem of estimating a graphon, we only ask to recover the graphon up to a measure-preserving transformation; this is analogous to a key aspect of the definition of cut distance between graphons, which we describe below.

2.2. The graphon estimation problems. Given a graph with adjacency matrix $(G_{ij})$ sampled according to Equation (2), there are two natural ways one may seek to invert this sampling procedure. Here we consider two distinct graphon estimation problems that correspond to inverting one or both of the sampling steps. The “graphon value estimation problem” aims to invert the second step of the sampling procedure, and hence can be thought of as finding the local underlying structure of a graph sampled from a graphon (without concluding anything about the graphon at any location that wasn’t involved in the sample). Suppose we have sampled the $W$-random graph $G(n,W)$ using $\{U_i\}_{i \in [n]}$ as in Equation (2). Graphon value estimation consists of giving an estimator $\hat{M} := (\hat{M}_{ij})_{i,j \in [n]}$ for the matrix $M := (M_{ij})_{i,j \in [n]}$ where each $M_{ij} := W(U_i,U_j)$. One measure of success for the graphon value estimation problem is given by the mean squared error:

$$
\text{MSE}(\hat{M}) := \mathbb{E} \left( \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (M_{ij} - \hat{M}_{ij})^2 \right),
$$

as used by Chatterjee (2014) and Gao et al. (2014) (see also Airoldi et al. (2013) and Chan and Airoldi (2014)). Whereas MSE in nonparametric function estimation is typically with respect to particular points of the domain (see, e.g., Tsybakov (2009, §1.2.1)), here the random sequence $\{U_i\}_{i \in [n]}$ is latent, and so we take the expectation also with respect to the randomness in the terms $U_i$ (and hence in the terms $M_{ij}$), following Chatterjee (2014, §2.6).
The “graphon function estimation problem” aims to invert the entire sampling procedure to recover a graphon (i.e., symmetric measurable function). A notion of success for the graphon function estimator problem, used by Wolfe and Olhede (2013), Choi and Wolfe (2014), and Olhede and Wolfe (2014), is given by the mean integrated squared error for an estimator \( \hat{W} \) of a graphon \( W \):

\[
\text{MISE}(\hat{W}) := \mathbb{E} \inf_{\varphi} \| W - \hat{W}^{\varphi} \|_2 = \mathbb{E} \inf_{\varphi} \int_{[0,1]^2} (W(x, y) - \hat{W}^{\varphi}(x, y))^2 \, dx \, dy,
\]

where \( \varphi \) ranges over measure-preserving transformations of \([0,1]\). However, as we describe below in Section 2.3, there are graphons \( W \) and \( V \) such that the random graphs \( G(\mathbb{N}, W) \) and \( G(\mathbb{N}, V) \) are close in distribution, but \( W \) and \( V \) are far in the above \( L^2 \) norm (up to measure-preserving transformations). An alternative global notion of success for the function estimation problem is to use the distribution of such random graphs directly (Kallenberg, 1999), or to use the cut distance, which we describe below.

The distinction between these two problems is analogous to the typical distinction between MSE and MISE in nonparametric function estimation (Tsypakov, 2009); see also the two estimation problems in Yang et al. (2014, §2.1).

In general, it is impossible to recover a measurable function from its values at a countable number of points. However, if we assume that the measurable function has specific structure (e.g., is continuous, Lipschitz, a step-function, etc.), then it may become possible. As a result, many graphon estimation methods, which we describe below, require the graphon to have a representation of a certain form. However, the problem of recovering a real-valued function from its values at a random set of inputs, under various assumptions on the function, may be treated separately from the estimation of these values. We therefore largely restrict our attention to the graphon value estimation problem, although we briefly address the question of bounds on the function estimation problem that are implied by value estimation.

### 2.3. Step-function estimators for graphons

A step-function graphon can be associated with any finite graph along with a partition of its vertices. Our presentation largely follows §7.1 and §9.2 of Lovász (2012), with slightly different notation.

A graphon \( V \) is called a **step-function** when there is a partition \( \mathcal{S} = \{S_0, \ldots, S_{k-1}\} \) of \([0,1]\) into finitely many measurable pieces, called **steps**, such that \( V \) is constant on each set \( S_i \times S_j \). Suppose \( H \) is a vertex-weighted, edge-weighted graph on \([n]\), with vertex weights \( \alpha_i \) and edge-weights \( \beta_{ij} \) for \( i, j \in [n] \). Then the **step-function graphon** \( W_H \) associated with \( H \) is defined by \( W_H(x, y) = \beta_{ij} \) for \( x \in J_i \) and \( y \in J_j \), where the steps \( J_0, \ldots, J_{n-1} \) are a partition of \([0,1]\) into consecutive intervals of size \( \frac{\alpha_i}{\sum_{i=1}^{n} \alpha_i} \) for \( i \in [n] \). (Consider an unweighted finite graph \( G \) to be the weighted graph with vertex weights \( \alpha_i = 1 \) and edge weights \( \beta_{ij} = G_{ij} \).)

Given a graph \( G \) on \([n]\) and vertex sets \( X, Y \subseteq [n] \), write \( c_G(X, Y) := \sum_{i \in X} \sum_{j \in Y} G_{ij} \) for the number of edges across the cut \((X, Y)\). Then the **edge density** in \( G \) between \( X \) and \( Y \) is defined to be

\[
e_G(X, Y) := \frac{c_G(X, Y)}{|X||Y|};
\]

when \( X \) and \( Y \) are disjoint, this quantity is the fraction of possible edges between \( X \) and \( Y \) that \( G \) contains.

Now suppose \( G \) is a graph on \([n]\), and \( \mathcal{P} = \{P_0, \ldots, P_{k-1}\} \) is a partition of the vertices of \( G \) into \( k \) classes. The **quotient graph** \( G/\mathcal{P} \) is defined to be the weighted graph on \([k]\) with respective
vertex weights $|P_i|/n$ and edge weights $e_G(P_i, P_j)$. In our step-function estimator, we will routinely pass from a sampled graph $G$ and a partition $\mathcal{P}$ of its vertex set to the graphon $W_{G/\mathcal{P}}$ formed from the quotient $G/\mathcal{P}$.

One may similarly define the step-function graphon $V_S$ of $V$ with respect to a measurable partition $S = \{S_0, \ldots, S_{k-1}\}$ of $[0,1]$ as the step-function graphon of the weighted graph with each vertex weight $\alpha_i$ equal to the measure of $S_i$ and edge weight $\beta_{ij} = \int_{S_i \times S_j} V(x, y) \, dx \, dy$.

We next describe the cut metric, which defines a notion of distance between two graphs or graphons, following §8.1 and §8.2 of Lovász (2012). We begin with finite graphs on the same vertex set.

**Definition 2.7**. Let $F, G$ be two graphs on $[n]$. The cut metric between $F$ and $G$ is given by

$$d_{\square}(F, G) := \max_{S, T \subseteq [n]} \frac{c_F(S, T) - c_G(S, T)}{n^2}.$$

Note that the denominator of Equation (5) is $n^2$ regardless of the size of $S$ and $T$; having large distance between $F$ and $G$ in the cut metric implies that there is some large vertex set on which their respective edge densities differ.

The cut distance $\delta_{\square}$ between two graphs on different vertex sets of the same size $n$ is then defined to be the minimum of $d_{\square}$ over all relabelings of $F$ and $G$ by $[n]$. While the cut distance can be extended to arbitrary finite weighted graphs on different vertex sets, these definitions are rather technical, and so we instead define the analogous quantity for graphons, from which one may inherit the corresponding notions via step-function graphons.

**Definition 2.8**. Let $W, V$ be graphons. The cut metric between $W$ and $V$ is given by

$$d_{\square}(W, V) := \sup_{S, T \subseteq [0,1]} \left| \int_{S \times T} (W(x, y) - V(x, y)) \, dx \, dy \right|,$$

where $S$ and $T$ range over measurable subsets. The cut distance between $W$ and $V$ is defined to be

$$\delta_{\square}(W, V) := \inf_{\varphi} d_{\square}(W, V^\varphi),$$

where $\varphi$ is a measure-preserving transformation of $[0,1]$.

Note that the cut distance is only a pseudometric, as it is zero for weakly isomorphic graphons.

The Frieze–Kannan weak regularity lemma (Frieze and Kannan, 1999a,b) implies that every graphon is well-approximated in the cut distance by its step-functions; moreover, the quality of such an approximation is bounded by the size of the partition, uniformly in the choice of graphon.

**Lemma 2.9** (Weak Regularity (Lovász, 2012, Lemma 9.3)). For every $k \geq 1$ and any graphon $W$, there is a partition $\mathcal{P}$ of the vertices of $W$ into $k$ classes such that

$$\delta_{\square}(W_{W/\mathcal{P}}) \leq \frac{2}{\sqrt{\log k}}.$$
An analogous statement holds for graphons (Lovász, 2012, Corollary 9.13). Step-functions also approximate graphons arbitrarily well in $L^1$ distance (or $L^2$, as in MSE or MISE), but the convergence is not uniform in the choice of graphon (Lovász, 2012, Proposition 9.8).

In graphon estimation by step-functions, we are given a finite graph sampled from a graphon, and need to choose the number of steps by which to approximate the graphon. Under $L^2$ risk, the number of such steps could be arbitrarily large depending on the particular graphon. But moreover, this number can vary wildly even among graphons that give rise to similar distributions — which therefore must be close in the cut distance (Lovász, 2012, Lemma 10.31). For example (Janson, 2013, Example 10.11), this can be seen by considering a constant function graphon $W$ (whose samples are Erdős–Rényi graphs) and the step-function graphon $V_k$ induced by a graph of size $k$ sampled from $W$. Their samples $G(n,W)$ and $G(n,V_k)$ have similar distributions, and indeed $\delta(W,V_k) = O(1/\sqrt{k})$, even though the $L^1$ distance between $W$ and $V_k$ is roughly 1/2 regardless of $k$. For this reason, it may be more appropriate to consider risk based on the cut distance, rather than the $L^2$-based MISE, for the function estimation problem for arbitrary graphons.

On the other hand, both the cut metric and $L^1$ can be close for step-functions (Lovász, 2012, Equation (8.15)). Hence even in $L^1$ (or $L^2$), it can be more reasonable to approximate a step-function graphon (as opposed to an arbitrary graphon) by step-function graphons.

### 2.4. Graphon estimation methods.

The first study of graphon estimation was by Kallenberg (1999) in the more general context of exchangeable arrays. This work predates the development of the theory of graphons; for details, see Orbanz and Roy (2014, §V).

A number of graphon estimators have been proposed in recent years. Here we mention several that are most closely related to our approach. The stochastic block model approximation (SBA) (Airoldi et al., 2013) requires multiple samples on the same vertex set, but is similar to our approach in some respects, as it partitions the vertex set according to the $L^2$ metric on their edge vectors (in essence, the vector of average edge densities with respect to the discrete partition). Sorting and smoothing (SAS) (Chan and Airoldi, 2014) takes a different approach to providing a computationally tractable estimator, and requires the graphon to have absolutely continuous degree distribution.

Several estimators are based on spectral methods, including Universal Singular Value Thresholding (USVT) (Chatterjee, 2014). Rather than estimating a specific cluster and using this to define a step-function, Amini and Levina (2014) first estimate a co-cluster matrix and then obtain a graphon estimate from this matrix by using eigenvalue truncation and $k$-means.

Other recent work in graphon estimation has focused on minimax optimality, histogram bin width, estimation using moments, or consequences of the graphon satisfying certain Lipschitz or Hölder conditions (Bickel and Chen, 2009; Bickel et al., 2011; Choi and Wolfe, 2014; Olhede and Wolfe, 2014; Wolfe and Olhede, 2013; Yang et al., 2014).

The estimation problem for latent space models can also be seen as graphon estimation, as such models are equivalent to graphon sampling procedures for graphons having nicer properties than mere measurability (Chatterjee, 2014, §2.4).

Many of the above graphon estimators are formulated in the setting of bipartite graphs and separate exchangeability, where the distribution is invariant under separate permutations of the rows and columns. For notational simplicity, we focus on the case of arbitrary undirected graphs, whose adjacency matrices are symmetric, and for which joint exchangeability is the appropriate notion, but many of our results have straightforward analogues for bipartite graphs.

### 3. Iterative Step-Function Estimation.

We have seen how a partition of a finite graph’s vertex set induces a step-function graphon. Many clustering algorithms can produce such partitions,
and thereby give rise to graphon estimators. Here we propose an approach to iteratively improving such estimates by forming a new partition whose classes contain vertices that have similar edge densities with respect to the old partition.

3.1. Graphon estimation via clustering. As described in Section 2.3, given a partition $\mathcal{P}$ of a finite graph $G$, one may form the step-function $W_{G/\mathcal{P}}$. Such a partition $\mathcal{P}$ may be formed by clustering the nodes using some general clustering method, such as $k$-means (MacQueen, 1967), hierarchical agglomerative clustering (Ward, Jr., 1963), random assignment, or a trivial clustering by assigning all vertices to the same class. In Figure 1, we display the result of estimating a graphon according to several particular choices of clustering algorithms.

Within the graphon estimator literature, several techniques produce step-functions, but the analysis has generally focused on the choice of partition size (Olhede and Wolfe, 2014), the convergence rates for optimal partitions (Choi and Wolfe, 2014; Gao et al., 2014; Wolfe and Olhede, 2013), or else the technique requires multiple observations (Airoldi et al., 2013). Here we aim to exploit structural aspects of graphs, such weak regularity (i.e., their uniform approximability in the cut distance), via an algorithm for forming a new partition that improves the step-function estimate $W_{G/\mathcal{P}}$ produced by any given partition $\mathcal{P}$.

![Fig 1: Comparison of the step-function graphons obtained using various clustering algorithms on an infinite relational model graphon. Clustering was performed using the Python package scikit-learn (Pedregosa et al., 2011) clustering defaults for $k$-means, agglomerative clustering, affinity propagation, and spectral clustering, except the number of clusters were set to $k = 15$ for each method using fixed cluster sizes. For (a), ISFE was applied to the trivial partition, where all vertices were initially assigned to a single bin.](image)

3.2. Iterative Step-Function Estimation. Given a finite graph $G$, consider the following graphon estimator procedure: (a) partition the vertices of $G$ according to some clustering algorithm; (b) repeatedly improve this partition by iteratively running Algorithm 1 for $T \geq 0$ iterations; and (c) report the step-function graphon $W_{G/\mathcal{P}}$, where $\mathcal{P}$ is the final partition produced, with its classes sorted according to their average edge densities. In aggregate, we call this procedure Iterative Step-Function Estimation (ISFE).

Let $W$ be a graphon and $n \in \mathbb{N}$, and suppose $G$ is a sample of the $W$-random graph $\mathbb{G}(n,W)$. ISFE on $G$ can be evaluated in terms of MISE by directly comparing the resulting graphon estimate $W_{G/\mathcal{P}}$ to $W$. We now describe how MSE is evaluated in a synthetic run where we retain the history of how $\mathcal{P}$ was formed from the values $(W(U_i,U_j))_{i,j \in [n]}$. Let $k$ be the number of classes in the partition $\mathcal{P}$. Implicit in the run of ISFE is a map $p : [n] \rightarrow [k]$ sending each vertex of $G$ to the index of its class in $\mathcal{P}$. Then MSE is evaluated by comparing $(W_{G/\mathcal{P}}(\frac{2p(i)+1}{2n}, \frac{2p(j)+1}{2n}))_{i,j \in [n]}$ with
\((W(U_i, U_j))_{i, j \in [n]}\). In other words, a regular grid \((\frac{2^{\ell+1}}{2n}, \frac{2m+1}{2n})_{\ell, m \in [n]}\) of \(n \times n\) points within \([0, 1]^2\) is chosen as a set of representatives of the piecewise constant regions of \(W_G / \mathcal{P}\), in some order that corresponds to how the vertices of \(G\) were rearranged into the partition \(\mathcal{P}\).

Algorithm 1 Iterative Step-Function Estimation Algorithm

1: **Input:** graph \(G\), initial partition \(\mathcal{P}^{(\text{old})}\), minimum number of classes \(\ell\), decay \(d\)
2: **Output:** new partition \(\mathcal{P}^{(\text{new})}\)
3: Initialize \(Q = \{Q_0\}, Q_0 = \{0\}, c_0 = 0, \epsilon = 1, B = 1\).
4: while number of classes \(B < \ell\) do
5:   for vertices \(i = 0, \ldots, n - 1\) do
6:     Compute edge densities vector \(e_i := \{e_G(P_j, \{i\})\}_j\) for all \(j = 0, \ldots, |\mathcal{P}^{(\text{old})}| - 1\).
7:     Compute min distance between vector \(e_i\) and some vector \(e_{c_j}\), where \(j = 0, \ldots, |\mathcal{P}^{(\text{old})}| - 1\):
8:       \(j^* := \arg \min_j d_j(e_i, e_{c_j}) = \arg \min_j \sum_r |e(\{i\}, \mathcal{P}_r^{(\text{old})}) - e(\{c_j\}, \mathcal{P}_r^{(\text{old})})|\)
9:     if min distance \(d_j(e_i, e_{c_j}) < \epsilon\) then
10:        Add vertex \(i\) to existing class \(Q_j\).
11:     else
12:        Create new class \(Q_B\) with \(Q_B = \{i\}\) and centroid \(c_B = i\).
13:        Add class \(Q_B\) to partition \(Q\).
14:        \(B \leftarrow B + 1\)
15:   end if
16: end for
17: \(\mathcal{P}^{(\text{new})} \leftarrow Q\)

Fig 2: Iterations of ISFE on a 200 vertex sample from a stochastic block model graphon with \(p = 0.5, q = (0.7, 0.3)\), beginning with a random partition into 6 classes.

As discussed in Section 2.3, by the weak regularity lemma for graphons, every graphon can be approximated to arbitrary accuracy in cut distance by a step-function, whose number of steps depends on the desired accuracy and not the graphon. ISFE seeks to take advantage of this structure. In an iteration of ISFE, each vertex is grouped with other vertices that are similar in average edge density with respect to the input partition. Each subsequent iteration seeks to improve upon the previous partition, by using it as the basis for the next round of density calculations.

Figure 2 demonstrates how ISFE extracts structure over the course of several iterations, beginning with a random partition (iteration 0), in which each vertex was independently placed into one of 6 classes uniformly at random. (For details on the SBM parameters, see Section 3.3 below.) Slight discrepancies in the edge densities between classes in the random partition of iteration 0 are amplified in iterations 1 and 2. The substantial correlations between the classes of the partition obtained in iteration 2 and the true block structure allow iteration 3 to produce a partition each of whose classes is largely from a single block. This is refined slightly in iteration 4.
3.3. Examples of ISFE. We now present three examples of the behavior of ISFE on certain classes of graphons.

Finite graphs. If one clusters the vertices of a sampled graph discretely, by assigning one class for each vertex, this typically induces a poor graphon estimator, as the reported graphon is merely the step-function of the sample. On the other hand, if we perform even a single iteration of ISFE on such a partition, we obtain an estimator that clusters vertices according to the Hamming distances between their edge vectors. In the case where the original graphon is the step-function of some finite graph on \( n \) vertices, ISFE following such a partition can recover the original graphon exactly in the sense of MSE, so long as the requested partition size is at least \( n \). (Estimation with respect to MISE is limited only by the extent to which the sampled graph distorts the proportion of vertices arising from each vertex of the original.)

We present an example in Figure 3. In this example, we form the step-function graphon \( W_G \) of a graph \( G \) with 7 vertices; its adjacency matrix can be seen via the black and white squares in Figure 3a. ISFE is run on a 70 vertex sample (Figure 3b) for a single iteration starting from the partition with every vertex in its own class. This single iteration amounts to forming a partition based on Hamming distance on the vector of edges for each vertex (i.e., rows of the adjacency matrix); so long as the requested number of bins is at least 7, the original structure will be recovered. The resulting graphon from ISFE in Figure 3d and the original step-function graphon in Figure 3a are not weakly isomorphic, because some of the 7 steps of the original resulted in slightly more or fewer than 10 vertices in the sample, but they are very close in the cut distance and, after rearranging by a measure-preserving transformation, in \( L^1 \). Note that sorting by degree (Figure 3c) garbles the structure among those regions corresponding to distinct vertices of the original having the same degree.

Stochastic block model. The stochastic block model (SBM) has been extensively studied from many perspectives; for a survey of some of the statistical literature as it relates to the graphon estimation problems, see Chatterjee (2014, §2.2).

In the stochastic block model, we assume there are \( k \) classes. We define the SBM graphon for \( k = 2 \) as follows: given parameters \( p \in [0,1], q = (q_0, q_1) \in [0,1]^2 \), we partition \([0,1]\) into two pieces \( P_0, P_1 \) of length \( p \) and \( 1 - p \), where \( p \in [0,1] \). The value of the graphon is constant on \( P_i \times P_i \) with value \( q_0 \) and constant on \( P_i \times P_{1-i} \) with value \( q_1 \), for \( i = 0, 1 \).

We show the result of ISFE on a graph sampled from an SBM graphon in Figure 4a for \( p = 0.5, q = (0.7, 0.3) \). In the first column, we display a plot of the graphon. We sample a 200 vertex graph from this SBM (column 2) and run ISFE (starting with the trivial partition) on the sample...
with $\ell = 8$ for $T = 5$ iterations (column 3). We show in the fourth column the original sample rearranged according to the step-function estimate produced by ISFE. The last column is the step-function estimate discretized on a $200 \times 200$ grid and sorted according to increasing $U_i$ value. (On a real dataset, where there are no such values $U_i$, it is not possible to produce this reordering.) The graphon in this final column is more easily visually compared to the original graphon in column 1, whereas the original estimate in column 3 shows more directly the partition found by ISFE.

![Graphon Estimation Examples](image)

(a) Stochastic block model

(b) Infinite relational model

Fig 4: Examples of graphon estimation using ISFE. Columns: (1) the original graphon; (2) a 200 vertex random sample from the original graphon; (3) ISFE-estimated graphon for $\ell = 8$ (SBM), 15 (IRM), $T = 4$; (4) the random sample reordered according to the ISFE estimate; (5) ISFE estimate rearranged by increasing $U_i$. ISFE was applied to the trivial partition, i.e., all vertices were initially in a single class.

Infinite Relational Model. The infinite relational model (IRM) (Kemp et al., 2006) is a non-parametric extension of the SBM, where the (infinite) partition is generated by a Chinese restaurant process with concentration parameter $\alpha$. For a description of the distribution on graphons implicit in this model, see Orbanz and Roy (2014, Example IV.1) and Lloyd et al. (2012, §4). For each class of the partition, the graphon is constant with value sampled from a beta distribution with parameters $a, b$.

We show the result of ISFE on a graph sampled from an IRM graphon with $\alpha = 3, a = 3, b = 2.9$ in Figure 4b. The five columns are analogous to those described above for Figure 4a.

4. Results.

4.1. Synthetic data. We examine synthetic data sampled from several graphons: (1) a gradient given by the function $W(x, y) = \frac{(1-x) + (1-y)}{2}$; (2) an SBM graphon with $p = 0.5, q = (0.7, 0.3)$; (3) an SBM graphon with $p = 0.3, q = (0.7, 0.3)$; and (4) an IRM graphon with $\alpha = 3, a = 3, b = 2.9$. In Figure 5, we display the results of ISFE and two other estimators on a 200 vertex sample from each of the graphons. The first column displays the graphon, the second column is the result of ISFE on the sample (starting with the trivial partition), the third column shows the result of SAS (Chan and Airoldi, 2014), and the last column shows the result of USVT (Chatterjee, 2014).
We evaluate all estimators using the mean squared error (MSE) given in Equation (3). For all graphon estimator images, we re-sort the result by increasing $U_i$ so that the result can be visually compared to the original graphon. While MSE is directly a measure of similarity between this re-sorted estimate and the original graphon (evaluated at certain points), in some cases better upper-bounds on the MISE may be obtained from other rearrangements of the estimate. Indeed, the smoothed estimates obtained by SAS and USVT for the gradient lead to a considerably smaller upper-bound on the MISE than their displayed re-sorted estimates.

For ISFE, the parameter $\ell$ was set to the value that minimized the MSE over a grid of values $\ell = 2, 4, 6, 8, 10, 12, 15, 30$; the gradient graphon used $\ell = 20$, the SBM graphon with $p = 0.5$ used $\ell = 8$, the SBM graphon with $p = 0.3$ used $\ell = 10$, and the IRM graphon used $\ell = 15$. For SAS, we modified the implementation by minimizing the total variation distance using a general-purpose constrained optimization method instead of using the alternating direction method of multipliers. For USVT, we follow Chan and Airoldi (2014) and sort the sample by degree before applying USVT.

SAS and USVT, which involve sorting the sample by degree, perform reasonably not only for graphons with monotonizable degree distribution, such as gradients (as in Figure 5a), for which SAS was explicitly designed, but also ones that are monotonizable up to some partition (as in Figures 5c and 5d). However, when the degree distribution is constant over regions with different structure (as in Figure 5b), SAS and USVT fail to discern this structure. In contrast, ISFE is able to recover much of the structure after a small number of iterations, even when it begins with no structural information, i.e., the trivial partition.

4.2. Real-world datasets. We examine three real-world social network datasets, considering a denser subgraph constructed by taking the top $K$ highest-degree vertices and the edges between them, for reasons we describe below. We randomized the order of the vertices for each graph before running the ISFE algorithm, which we present in Figure 6.

1. NIPS co-authorship dataset (Globerson et al., 2007): This dataset is an undirected network of co-authorships in the NIPS conference from Proceedings 1–12, with 2,037 vertices and 1,740 edges. We choose $K = 234$ for the denser subset, which has been studied in other work (Miller et al., 2009; Palla et al., 2012). For the ISFE parameters, we set $T = 8, \ell = 95$, initializing it with a 90 cluster $k$-means partition.

2. ca-AstroPh co-authorship dataset (Newman, 2001): This dataset is an undirected network of co-authorships between scientists posting pre-prints on the Astrophysics E-Print Archive between Jan 1, 1995 and December 31, 1999 with 18,772 vertices and 396,160 edges. We choose $K = 1000$, and set the ISFE parameters to $T = 8, \ell = 160$, initializing it with a 150 cluster partition from $k$-means.

3. Epinions dataset (Richardson et al., 2003): This dataset is a who-trusts-whom network of Epinions.com with 75,879 vertices, 508,837 edges. We work with the undirected graph obtained by symmetrizing the original undirected graph, choose $K = 1000$, and set $T = 8, \ell = 40$, initializing it with a 35 cluster partition from $k$-means.

Many real-world networks, such as those arising from co-authorship, social interactions, etc., are not well-modeled as exchangeable graphs, as they tend to exhibit power-law degree distributions, “small-world” phenomena such as short path lengths, and other properties that generally hold only for sparse sequences of graphs (having $O(n^2)$ edges among $n$ vertices, which is not possible for non-empty exchangeable graphs). For a detailed discussion, see Orbanz and Roy (2014, §VII).

One common approach to modeling sparse graphs using graphons is the Bollobás–Janson–Riordan model (Bollobás et al., 2007), where edges are independently deleted from an exchangeable
Fig 5: Comparison of graphon estimation methods ISFE, SAS, and USVT, along with their mean squared errors, on a gradient graphon, two stochastic block model graphons, and an infinite relational model graphon. The first column shows the original graphon. The other columns show ISFE (beginning with the trivial partition), SAS, and USVT, respectively; we display all estimator images rearranged by increasing $U_i$.

graph to achieve the desired edge density. Although this process will not achieve many of the above real-world phenomena (Orbanz and Roy, 2014, Example VII.4), the behavior of graphon estimators on graphs sampled in this way has been considered (Bickel et al., 2011; Wolfe and Olhede, 2013).

The question of how to model sparse graphs remains an important open problem. To demonstrate our graphon estimator, here we have sidestepped these issues to some extent by considering a denser
Fig 6: ISFE results on real-world datasets for NIPS co-authorship, Astrophysics arXiv co-authorship, and epinions trust network. Columns: (1) A denser subset of the original graph; (2) estimated ISFE graphon; (3) adjacency matrix rearranged according to ISFE estimate.

subgraph of the original network, which is presumably modeled more accurately than the original as a sample from a graphon.

5. Discussion. While any clustering algorithm naturally induces a graphon estimator, we have described and shown some of the improvements that may be obtained by grouping vertices according to their average edge densities with respect to the clusters. The fact that such improvements are possible is unsurprising for graphs admitting block structure (although there are many possible refinements and further analysis of the algorithm we describe). Indeed, arbitrary graphons are well-approximated by step-function graphons in the cut metric, and step-functions graphons (which arise as models in the stochastic block model and elsewhere) are well-approximated by such in $L^2$. A key problem is to devise graphon estimators that further leverage this structure.
Acknowledgments. The authors would like to thank Daniel Roy for detailed comments on a draft, and James Lloyd and Peter Orbanz for helpful conversations.

This material is based upon work supported by the United States Air Force and the Defense Advanced Research Projects Agency (DARPA) under Contract Numbers FA8750-14-C-0001 and FA8750-14-2-0004. Work by C.F. was also made possible through the support of Army Research Office grant number W911NF-13-1-0212 and a grant from Google. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force, Army, or DARPA.

References.
E. M. Airoldi, T. B. Costa, and S. H. Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In Adv. Neural Inform. Process. Syst. (NIPS) 26, pages 692–700, 2013.
D. J. Aldous. Representations for partially exchangeable arrays of random variables. J. Multivariate Anal., 11(4):581–598, 1981.
A. A. Amini and E. Levina. On semidefinite relaxations for the block model. ArXiv e-print 1406.5647, 2014.
T. Austin. On exchangeable random variables and the statistics of large graphs and hypergraphs. Probab. Surv., 5:80–145, 2008.
P. J. Bickel and A. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. Proc. Natl. Acad. Sci. USA, 106(50):21068–21073, 2009.
P. J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. Ann. Statist., 39(5):2280–2301, 2011.
B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. Random Structures Algorithms, 31(1):3–122, 2007.
S. H. Chan and E. Airoldi. A consistent histogram estimator for exchangeable graph models. In Proc. 31st Int. Conf. Mach. Learn. (ICML), pages 208–216, 2014.
S. Chatterjee. Matrix estimation by universal singular value thresholding. ArXiv e-print 1212.1247v6, 2014.
P. Diaoconis and S. Janson. Graph limits and exchangeable random graphs. Rend. Mat. Appl. (7), 28(1):80–145, 2008.
A. Frieze and R. Kannan. A simple algorithm for constructing Szemerédi’s regularity partition. Electron. J. Combin., 6:R17, 7 pp., 1999a.
A. Frieze and R. Kannan. Quick approximation to matrices and applications. Combinatorica, 19(2):175–220, 1999b.
C. Gao, Y. Lu, and H. H. Zhou. Rate-optimal graphon estimation. ArXiv e-print 1410.5837, 2014.
M. Girvan and M. E. J. Newman. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA, 99(12):7821–7826, 2002.
A. Globerson, G. Chechik, F. Pereira, and N. Tishby. Euclidean embedding of co-occurrence data. J. Mach. Learn. Res., 8:2265–2295, 2007.
P. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In Adv. Neural Inform. Process. Syst. (NIPS) 20, pages 657–664, 2007.
P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. J. Amer. Statist. Assoc., 97(460):1090–1098, 2002.
P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: first steps. Social Networks, 5(2):109–137, 1983.
D. N. Hoover. Relations on probability spaces and arrays of random variables. Preprint, Institute for Advanced Study, Princeton, NJ, 1979.
S. Janson. Graphons, cut norm and distance, couplings and rearrangements, volume 4 of New York J. Math. Monographs. Univ. at Albany, State Univ of New York, Albany, NY, 2013.
O. Kallenberg. Multivariate sampling and the estimation problem for exchangeable arrays. J. Theoret. Probab., 12(3):859–883, 1999.
O. Kallenberg. Probabilistic symmetries and invariance principles. Probab. Appl. Springer, New York, 2005.
C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In Proc. 21st Nat. Conf. Artificial Intelligence (AAAI-06), 2006.
J. R. Lloyd, P. Orbanz, Z. Ghahramani, and D. M. Roy. Random function priors for exchangeable arrays with applications to graphs and relational data. In Adv. Neural Inform. Process. Syst. (NIPS) 25, pages 1007–1015, 2012.
L. Lovász. Large networks and graph limits, volume 60 of American Math. Soc. Colloq. Publ. American Math. Soc., Providence, RI, 2012.
J. MacQueen. Some methods for classification and analysis of multivariate observations. In Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), pages Vol. I: Statistics, pp. 281–297. Univ. California Press, Berkeley, Calif., 1967.

K. T. Miller, T. L. Griffiths, and M. I. Jordan. Nonparametric latent feature models for link prediction. In Adv. Neural Inform. Process. Syst. (NIPS) 22, pages 1276–1284, 2009.

M. E. J. Newman. The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA, 98(2):404–409, 2001.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. Phys. Rev. E, 69:026113, 2004.

S. C. Olhede and P. J. Wolfe. Network histograms and universality of blockmodel approximation. Proc. Natl. Acad. Sci. USA, 2014.

P. Orbanz and D. M. Roy. Bayesian models of graphs, arrays and other exchangeable random structures. IEEE Trans. Pattern Anal. Mach. Intell., 36, 2014.

K. Palla, D. A. Knowles, and Z. Ghahramani. An infinite latent attribute model for network data. In Proc. 29th Int. Conf. Mach. Learn. (ICML), 2012.

F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res., 12:2825–2830, 2011.

M. Richardson, R. Agrawal, and P. Domingos. Trust management for the semantic web. In The Semantic Web - ISWC 2003, volume 2870 of Lecture Notes Comput. Sci., pages 351–368. Springer, 2003.

D. M. Roy and Y. W. Teh. The Mondrian process. In Adv. Neural Inform. Process. Syst. (NIPS) 21, pages 1377–1384, 2008.

A. B. Tsybakov. Introduction to nonparametric estimation. Springer Ser. Statist. Springer, New York, 2009.

J. H. Ward, Jr. Hierarchical grouping to optimize an objective function. J. Amer. Statist. Assoc., 58:236–244, 1963.

P. J. Wolfe and S. C. Olhede. Nonparametric graphon estimation. ArXiv e-print 1309.5936, 2013.

J. Yang, C. Han, and E. Airoldi. Nonparametric estimation and testing of exchangeable graph models. In Proc. 17th Int. Conf. Artificial Intelligence Statist. (AISTATS), pages 1060–1067, 2014.