Rates of Convergence of Spectral Methods for Graphon Estimation

Jiaming Xu [1]

Abstract

This paper studies the problem of estimating the graphon function – a generative mechanism for a class of random graphs that are useful approximations to real networks. Specifically, a graph of n vertices is generated such that each pair of two vertices i and j are connected independently with probability $\rho_n \times f(x_i, x_j)$, where x_i is the unknown d-dimensional label of vertex i, f is an unknown symmetric function, and ρ_n, assumed to be $\Omega(\log n/n)$, is a scaling parameter characterizing the graph sparsity. The task is to estimate graphon f given the graph. Recent studies have identified the minimax optimal estimation error rate for $d = 1$. However, there exists a wide gap between the known error rates of polynomial-time estimators and the minimax optimal error rate. We improve on the previously known error rates of polynomial-time estimators, by analyzing a spectral method, namely universal singular value thresholding (USVT) algorithm. When f belongs to either Hölder or Sobolev space with smoothness index α, we show the error rates of USVT are at most $(n\rho)^{-2\alpha/(2\alpha+d)}$. These error rates approach the minimax optimal error rate $\log(n\rho)/(n\rho)$ proved in prior work for $d = 1$, as α increases, i.e., f becomes smoother. Furthermore, when f is analytic with infinitely many times differentiability, we show the error rate of USVT is at most $\log^d(n\rho)/(n\rho)$. When f is a step function which corresponds to the stochastic block model with k blocks for some k, the error rate of USVT is at most $k/(n\rho)$, which is larger than the minimax optimal error rate by at most a multiplicative factor $k/\log k$. This coincides with the computational gap observed in community detection. A key ingredient of our analysis is to derive the eigenvalue decaying rate of the edge probability matrix using piecewise polynomial approximations of the graphon function f.

1. Introduction

Many modern systems and datasets can be represented as networks with vertices denoting the objects and edges (possibly weighted or labelled) encoding their interactions. Examples include online social networks such as Facebook friendship network, biological networks such as protein-protein interaction networks, and recommender systems such as movie rating datasets. A key task in network analysis is to estimate the underlying network generating mechanism, i.e., how the edges are formed in a network. It is useful for many important applications such as studying network evolution over time (Pensky, 2016), predicting missing links in networks (Miller et al., 2009; Airoldi et al., 2013; Gao et al., 2016), learning hidden user prefererences in recommender systems (Song et al., 2016), and correcting errors in crowd-sourcing systems (Lee & Shah, 2017). However, in practice we usually observe only a very small fraction of edge connections in these networks, which obscures the underlying network generating mechanism. For example, around 80% of the molecular interactions in cells of Yeast (Yu et al., 2008) are unknown. In Netflix movie dataset, about 99% of movie ratings are missing and the observed ratings are noisy.

In this paper, we are interested in understanding when and how the underlying network generating mechanism can be efficiently inferred from a partial observation of a network. We assume the network is generated according to the graphon model (Lovász & Szegedy, 2006). Concretely, given n vertices, the edges are generated independently, connecting each pair of two distinct vertices i and j with a probability

$$M_{ij} = f(x_i, x_j), \tag{1}$$

where $x_i \in \mathcal{X}$ is the latent feature vector of vertex i that captures various characteristics of vertex i; $f : \mathcal{X} \times \mathcal{X} \to [0, 1]$ is a symmetric function called graphon. We assume no self loop and set $M_{ii} = 0$ for $1 \le i \le n$. We further assume the feature vectors x_i's are drawn i.i.d. from a measurable space $\mathcal{X}$ according to a probability distribution μ.

Graphon model captures a key characteristic of real networks, that is, the edge connections are dependent on latent

[1]The Fuqua School of Business, Duke University, Durham, NC, USA.. Correspondence to: J. Xu <jiaming.xu868@duke.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

features of vertices rather than specific vertex identities. Graphon model was originally developed as a limit of a sequence of graphs with growing sizes (Lovász, 2012), and has been applied to various network analysis problems ranging from testing graph properties to counting homomorphisms to charactering distances between two graphs (Lovász, 2012; Borgs et al., 2008; 2012) to detecting communities (Bickel & Chen, 2009). Graphon model encompasses many existing network models as special cases. Setting f to be a constant p, it gives rise to Erdős-Rényi random graphs, where each edge is formed independently with probability p. In the case where f is a step function or $\mathcal{X}$ is a discrete set, the model specializes to the stochastic block model (Holland et al., 1983), where each vertex belongs to a community, and the edge probability between i and j depends only on which communities they are in. If $\mathcal{X}$ is a Euclidean space of dimension d and $f(x_i, x_j)$ is a function of the Euclidean distance $\|x_i - x_j\|$, then the grahon model reduces to the latent space model (Hoff et al., 2002; Handcock et al., 2007).

To capture a partial observation of the network, we assume every edge is observed independently with probability $\rho \in [0, 1]$, where $\rho = \rho_n$ may converge to 0 as $n \to \infty$. Given the resulting observed graph, the problem of interest is to estimate the underlying network generating mechanism – the graphon f. However, without observing x_i's, there is no way to uniquely identify f. To overcome this identifiability issue, we follow the prior work (Gao et al., 2015) and consider estimating f under the expected empirical loss[1]:

$$\frac{1}{n^2}\mathbb{E}\left[\sum_{i,j\in[n]}\left(\widehat{f}(x_i, x_j) - f(x_i, x_j)\right)^2\right].$$

This is equivalent to estimating the edge probability matrix M under the mean squared error (Gao et al., 2015):

$$\text{MSE}(\widehat{M}) \triangleq \frac{1}{n^2}\mathbb{E}\left[\left\|\widehat{M} - M\right\|_F^2\right], \tag{2}$$

where $\widehat{M}_{ij} = \widehat{f}(x_i, x_j)$. The fundamental estimation limits are phrased in terms of the minimax mean-squared error:

$$R_n^* \triangleq \inf_{\widehat{M}} \sup_{f\in\mathcal{F}} \sup_{\mu\in\mathcal{P}} \text{MSE}(\widehat{M}),$$

where $\mathcal{F}$ denotes a set of admissible graphon functions f, and $\mathcal{P}$ denotes the set of all possible probability measures on $\mathcal{X}$. The minimax estimation error depends on the smoothness of graphon f, the structure of latent space $(\mathcal{X}, \mu)$, and the observation probability ρ.

There is a recent surge of interest in graphon estimation. Various procedures have been proposed and analyzed (Gao et al., 2015; Klopp et al., 2015; Gao et al., 2016; Wolfe & Olhede, 2013; Airoldi et al., 2013; Yang et al., 2014; Chan & Airoldi, 2014; Cai et al., 2014; Zhang et al., 2015; Borgs et al., 2015a; Klopp & Verzelen, 2017; Borgs et al., 2017). A recent line of work (Gao et al., 2015; Klopp et al., 2015; Gao et al., 2016) has characterized the minimax error rate in certain special cases. In particular, for stochastic block model with k blocks, it is shown that the minimax error rate is $\frac{k^2}{n^2\rho} + \frac{\log k}{n\rho}$. For fully observed graphons with f being Hölder smooth on $\mathcal{X} = [0, 1]$ and $\rho = 1$, the minimax error rate turns out be $n^{-1}\log k + n^{-2\alpha/(\alpha+1)}$, where α is the smoothness index of f. This result was extended by (Klopp et al., 2015; Gao et al., 2016) to sparse regimes where $\rho \to 0$ as $n \to \infty$.

From a computational perspective, the problem appears to be much harder and far less well-understood. In the special case where f is α-Hölder smooth on $\mathcal{X} = [0, 1]$, a universal singular value thresholding (USVT) algorithm is shown in (Chatterjee, 2015) to achieve an error rate of $n^{-1/3}\rho^{-1/2}$. However, this performance guarantee is far from the minimax optimal rate $\log(n\rho)/(n\rho)$. A similar spectral method is shown in (Xu et al., 2014) to achieve a vanishing MSE when $n\rho \gg \log n$ but without an explicit characterization of the rate of the convergence. The nearest-neighbor based approach is analyzed in (Song et al., 2016) under a stringent assumption $n\rho \gg \sqrt{n}$. A simple degree sorting algorithm (Borgs et al., 2015b) is shown to achieve an error rate of $(\log(n\rho)/(n\rho))^{\alpha/(4\alpha+d)}$ for $\alpha \in (0, 1]$ under the restrictive assumption that $\int_0^1 f(x, y)\mathrm{d}y$ is strictly monotone in x.

In summary, despite the recent significant effort devoted to developing fundamental limits and efficient algorithms for graphon estimation, an understanding of the statistical and computational aspects of graphon estimation is still lacking. In particular, there is a wide gap between the known performance bounds of polynomial-time estimator and the minimax optimal estimation rate. This raises a fundamental question:

Is there a polynomial-time estimator that is guaranteed to achieve the minimax optimal rate?

In this paper, we provide a partial answer to this question by analyzing the universal singular value thresholding (USVT) algorithm proposed by Chatterjee (Chatterjee, 2015). The universal singular value thresholding is a simple and versatile method for structured matrix estimation and has been applied to a variety of different problems such as rank-

[1] By the definition of the expected empirical loss, we only need to estimate the edge probabilities $\{f(x_i, x_j)\}$. An alternative way to overcome the identifiablity issue is to consider estimating f up to weak isometry under the expected integral loss (Wolfe & Olhede, 2013). As shown in (Klopp et al., 2015)[Section 3], the results of estimating f under the expected empirical loss can be readily extended to estimating f up to weak isometry under the expected integral loss by including an extra agnostic error term due to the discretization of $\mathcal{X}$ by x_i's.

ing (Shah et al., 2016). It truncates the singular values of A at a threshold slightly above the spectral norm $\|A - \mathbb{E}[A]\|$, and estimates M by a properly rescaled A after truncation. It is computationally efficient, however, its performance guarantee established in (Chatterjee, 2015) requires the total number of observed edges to be much larger than $n^{(2d+2)/(d+2)}$ to attain a vanishing MSE. In contrast, our improved performance bound shows that the total number of observed edges only needs to be a constant factor larger than $n \log n$, irrespective of the latent space dimension d.

More formally, by assuming the average vertex degree is at least logarithmic in n, i.e., $n\rho = \Omega(\log n)$, and $\mathcal{X}$ is a compact subset in $\mathbb{R}^d$, the mean-squared error rate of USVT is shown to be upper bounded by $(n\rho)^{-2\alpha/(2\alpha+d)}$, when f belongs to either α-smooth Hölder function class $\mathcal{H}(\alpha, L)$ or α-smooth Sobolev space $\mathcal{S}(\alpha, L)$. This convergence rate of USVT is approaching the minimax optimal rate $\log(n\rho)/(n\rho)$ provided in (Gao et al., 2015) for $d = 1$, as f becomes smoother, i.e., α increases. In fact, we show that if f is analytic with infinitely many times differentiability[2], then the error rate is upper bounded by $\log^d(n\rho)/(n\rho)$.

In the special case where f is a step fuction or $\mathcal{X}$ is a discrete set, then the graphon model specializes to the stochastic block model with k blocks for some k. In this case, the error rate of USVT is shown to be $k/(n\rho)$, which is larger than the optimal minimax rate by at most a multiplicative factor $k/\log k$. This factor coincides with the ratio of the Kesten-Stigum threshold and information-theoretic threshold in community detection (Banks et al., 2016; Abbe & Sandon, 2015; Banks et al., 2018). Based on compelling but non-rigorous statistical physics arguments, it is believed that no polynomial-time algorithms are able to detect the communities between the KS-threshold and IT-threshold (Moore, 2017). This coincidence indicates that $k/(n\rho)$ may be the optimal estimation rate among all polynomial-time algorithms, and the minimax optimal rate may not be attainable in polynomial-time. During the preparation of this manuscript, we became aware of an earlier arXiv preprint (Klopp & Verzelen, 2017)[Proposition 4] which also derives the error rate of $k/(n\rho)$.

Our proof incorporates three interesting ingredients. One is a characterization of the estimation error of USVT in terms of the tail of eigenvalues of M, and the spectral norm of the noise perturbation $\|A - \mathbb{E}[A|M]\|$, see e.g., (Shah et al., 2016)[Lemma 3]. The second one is a high-probability upper bound on $\|A - \mathbb{E}[M|A]\|$ obtained from matrix concentration inequalities initially developed by (Feige & Ofek, 2005). The last but most important one is a characterization of the tail of eigenvalues of M using piecewise polynomial approximations of f, which were originally used to study the spectrum of integral operators defined by f (Birman & Solomyak, 1967; 1977). The piecewise *constant* approximations of f have appeared in the previous work on graphon estimation (Chatterjee, 2015; Gao et al., 2015; Klopp et al., 2015), and are sufficient for the purpose of deriving minimax estimation rates because the smoothness of f beyond $\alpha = 1$ does not improve the rates. However, piecewise degree-$\lfloor \alpha \rfloor$ polynomial approximations are needed for showing USVT to achieve a faster converging rate as α increases.

[2]The minimax lower bound in (Gao et al., 2015)[Appendix A.1] is only established for the α-smooth Hölder function class for any fixed α and $d = 1$. It is an open question whether the error rate of $\log^d(n\rho)/(n\rho)$ is minimax-optimal for analytic graphons.

Algorithm 1 Universal Singular Value Thresholding (USVT) (Chatterjee, 2015)

Input: $A \in \mathbb{R}^{n \times n}$, $\rho \in [0,1]$ and a threshold $\tau > 0$.

Let $A = \sum_{i=1}^n s_i u_i v_i^\top$ be its singular value decomposition with $s_1 \geq s_2 \geq \cdots \geq s_n$.

Let S be the set of "thresholded" singular values: $S = \{i : s_i \geq \tau\}$.

Let

$$\widehat{A} = \sum_{i \in S} s_i u_i v_i^\top$$

and $\widetilde{M} = \widehat{A}/\rho$.

Output a matrix $\widehat{M} \in [0,1]^{n \times n}$ such that $\widehat{M}_{ii} = 0$ for all $i \in [n]$, and for $1 \leq i < j \leq n$, $\widehat{M}_{ij} = \widehat{M}_{ji}$ and

$$\widehat{M}_{ij} = \begin{cases} \widetilde{M}_{ij}, & \text{if } \widetilde{M}_{ij} \in [0,1] \\ 1, & \text{if } \widetilde{M}_{ij} > 1 \\ 0, & \text{if } \widetilde{M}_{ij} < 0. \end{cases}$$

Notation For a vector $x \in \mathbb{R}^d$, let $\|x\|_2$ denote its ℓ_2 norm and $\|x\|_\infty = \max_{1 \leq i \leq d} |x_i|$ denote its ℓ-infinity norm. For any matrix M, let $\|M\|$ denote its spectral norm and $\|M\|_F$ denote its Frobenius norm. For any positive integer n, let $[n] = \{1, \ldots, n\}$. For any positive constant α, let $\lfloor \alpha \rfloor$ denotes the largest integer strictly smaller than α. For two real numbers α and β, let $\alpha \wedge \beta = \min\{\alpha, \beta\}$ and $\alpha \vee \beta = \max\{\alpha, \beta\}$. If $\kappa = (\kappa_1, \ldots, \kappa_d)$ is a multi-index with $\kappa_i \in \mathbb{N}$, then $|\kappa| = \sum_{i=1}^d \kappa_i$, $\kappa! = \prod_{i=1}^d \kappa_i!$, and $x^\kappa = \prod_{i=1}^d x_i^{\kappa_i}$ for a vector $x \in \mathbb{R}^d$.

2. Main results

Let A denote the adjacency matrix of the observed graph with $A_{ii} = 0$ by convention. Then conditional on $x = (x_1, \ldots, x_n)$, for $1 \leq i < j \leq n$, $A_{ij} = A_{ji}$ are independently distributed as $\mathrm{Bern}(\rho M_{ij})$. In particular, $\mathbb{E}[A|M] = \rho M$.

To describe our main results, we first recall the *universal singular value thresholding* (USVT) algorithm (Chatterjee,

2015) as stated in Algorithm 1. Note that according to the graphon model (1), the edge probability matrix M may not be of low-rank. Nevertheless, it is possible that the singular values of M, or equivalently magnitudes of eigenvalues, drop off fast enough and as a consequence M is approximately low-rank. If this is indeed the case, then a natural idea to estimate M is via low-rank approximations of A. In particular, USVT truncates the singular values of A at a proper threshold τ, and estimates M by the rescaled A after truncation.

Note that Algorithm 1 applies hard-thresholding to the singular values of A. Alternatively, we can use soft-thresholding (Koltchinskii et al., 2011) and let $\widehat{A} = \sum_{i \in S}(s_i - \tau)u_i v_i^\top$. Our main results with the hard-thresholding also apply to the soft-thresholding. As argued in (Chatterjee, 2015), the cut-off threshold τ is chosen to be slightly above $\|A - \mathbb{E}[A|M]\|$, so that noise is suppressed and signals corresponding to large singular values of $\mathbb{E}[A|M]$ are maintained. Since conditional on M, A is a random matrix with independent entries bounded in $[0,1]$ of variance at most ρ, it is expected that $\|A - \mathbb{E}[A|M]\| \lesssim \sqrt{n\rho}$ with high probability, in view of standard matrix concentration inequalities. This turns out to be true if the observed graph is not too sparse, i.e., there exists a positive constant C such that

$$n\rho \geq C \log n. \tag{3}$$

However, when the observed graph is sparse with $n\rho = o(\log n)$, due to the existence of high-degree vertices, $\|A - \mathbb{E}[A|M]\| \gg \sqrt{n\rho}$ with high probability (Krivelevich & Sudakov, 2003).

Motivated by the discussion above, we focus on the relatively sparse regime where (3) holds, and set $\tau = c_0\sqrt{n\rho}$ for a positive large constant c_0, whose value depends on the constant C in (3). It is known that with high probability,

$$\|A - \mathbb{E}[A|M]\| \leq \kappa\sqrt{n\rho},$$

where

$$\kappa = \begin{cases} 4 + o(1) & n\rho = \omega(\log n) \\ 2 + o(1) & n\rho = \omega(\log^4 n) \end{cases}, \tag{4}$$

see, e.g., (Hajek et al., 2016)[Lemma 30]. Hence, the constant c_0 can be set to be a universal constant strictly larger than 4 in the case of $n\rho \gg \log(n)$ and 2 in the case of $n\rho \gg \log^4(n)$. Notably, in these cases, the cut-off threshold τ is universal, independent of the underlying graphon f.

Next, we present our main results without proofs. The excluded details can be found in the full paper (Xu, 2017). Our first result provides an upper bound to the estimation error of USVT. A similar result without explicit constants is proved in (Shah et al., 2016)[Lemma 3], which improves on the previous result in (Chatterjee, 2015)[Lemma 3.5]. Another similar result with slightly different constants is proved in (Koltchinskii et al., 2011)[Theorem 1] for soft singular value thresholding and in (Klopp et al., 2011)[Theorem 2] for hard singular value thresholding.

Theorem 1. *Consider the relatively sparse regime where (3) holds. For all $c > 0$ there exists a positive constant κ such that if $\tau = (1+\delta)\kappa\sqrt{n\rho}$ for a fixed constant $\delta > 0$, then conditional on M, with probability at least $1 - n^{-c}$,*

$$\frac{1}{n^2}\|\widehat{M} - M\|_F^2 \leq 16(1+\delta)^2 \min_{0 \leq r \leq n}\left(\frac{\kappa^2 r}{n\rho} + \frac{1}{n^2\delta^2}\sum_{i \geq r+1}\lambda_i^2(M)\right).$$

It further follows that $\mathrm{MSE}(\widehat{M})$ is bounded by the same error term as above plus the failing probability n^{-c}.

Theorem 1 gives an upper bound to the estimation error of USVT in terms of the tail of eigenvalues of M and the observation probability ρ. The upper bound invovles minimization of a sum of two terms over integers $0 \leq r \leq n$: the first term $r/(n\rho)$ can be viewed as the estimation error for a rank-r matrix; the second term $n^{-2}\sum_{i \geq r+1}\lambda_i^2(M)$ is the tail of eigenvalues of M and charaterizes the approximation error of M by the best rank-r matrix. The optimal r is chosen to achieve the best trade-off between the estimation error and the approximaiton error. Moreover, a lighter tail of eigenvalues of M implies a faster convergence rate of the estimation error. To characterize different tails of eigenvalues of M, we introduce the following definitions of polynomial and super-polynomial decays.

Definition 1 (Polynomial decay)**.** *We say the eigenvalues of M asymptotically satisfy a polynomial decay with rate $\beta > 0$ if for all integers $0 \leq r \leq n-1$,*

$$\frac{1}{n^2}\sum_{i \geq r+1}\mathbb{E}\left[\lambda_i^2(M)\right] \leq c_0 r^{-\beta} + c_1 n^{-1},$$

where c_0 and c_1 are two constants independent of n and r.

Definition 2 (Super-polynomial decay)**.** *We say the eigenvalues of M asymptotically satisfy a super-polynomial decay with rate $\alpha > 0$ if for all integers $0 \leq r \leq n-1$,*

$$\frac{1}{n^2}\sum_{i \geq r+1}\mathbb{E}\left[\lambda_i^2(M)\right] \leq c_0 e^{-c_2 r^\alpha} + c_1 n^{-1},$$

where c_0, c_1, c_2 are constants independent of n and r.

We remark that in the above two definitions, we allow for a residual term $c_1 n^{-1}$, which is responsible for the contribution of diagonal entries of M. According to Theorem 1, this residual term only induces an additional n^{-1} error in the upper bound to MSE and will not affect our main results. The following corollary readily follows from Theorem 1 by choosing the optimal r according to the decay rates of eigenvalues of M.

Corollary 1. *Consider the relatively sparse regime where (3) holds and suppose the eigenvalues of M satisfy a polynomial decay with rate $\beta > 0$. Then there exists a positive constant $\kappa > 0$ such that if $\tau = (1+\delta)\kappa\sqrt{n\rho}$ for a fixed constant $\delta > 0$,*

$$\mathrm{MSE}(\widehat{M}) \le c'(n\rho)^{-\frac{\beta}{\beta+1}}.$$

If instead the eigenvalues of M satisfy a super-polynomial decay with rates $\alpha > 0$, then

$$\mathrm{MSE}(\widehat{M}) \le c' \frac{(\log(n\rho))^{1/\alpha}}{n\rho},$$

where c' is a positive constant independent of n.

Proof. The first conclusion follows from Theorem 1 by choosing $c = 1$ and $r = \lfloor (n\rho)^{1/(\beta+1)} \rfloor$ and the second one follows by choosing $c = 1$ and $r = \lfloor (\log(n\rho)/c_2)^{1/\alpha} \rfloor$. □

Next we specialize our general results in different settings by deriving the decay rates of eigenvalues of M.

2.1. Stochastic block model

We first present results on the rate of convergence in the stochastic block model setting, where $x_i \in \{1, 2, \ldots, k\}$ indicating which community that vertex i belongs to. In this case, M_{ij} only depends on the communities of vertex i and vertex j, and M has rank at most k.

Theorem 2. *Assume (3) holds under the stochastic block model with k blocks. There exists a positive constant $\kappa > 0$ such that if $\tau = (1+\delta)\kappa\sqrt{n\rho}$ for some fixed constant $\delta > 0$, then*

$$\mathrm{MSE}(\widehat{M}) \le c'' \left[\frac{k}{n\rho} \wedge 1\right].$$

where c'' is a positive constant depending on κ and δ.

Proof. Under the stochastic block model, M is of rank at most k. Thus $\lambda_i(M) = 0$ for all $i \ge k+1$. Moreover, since $M_{ij} \in [0,1]$, it follows that $\sum_{i=1}^k \lambda_i^2(M) = \|M\|_F^2 \le n^2$. Applying Theorem 1 with $r = 0$ and $r = k$ yields the desired result. □

Theorem 2 shows that the convergence rate of MSE of USVT is at most $\frac{k}{n\rho} \wedge 1$, while the previous result in (Chatterjee, 2015) establishes that the convergence rate is at most $\sqrt{k/n}$ for $\rho = 1$. During the preparation of this manuscript, we became aware of an earlier arXiv preprint (Klopp & Verzelen, 2017)[Proposition 4] which also proves the error rate of $k/(n\rho)$.

The minimax optimal rate derived in (Klopp et al., 2015; Gao et al., 2016) is $\left(\frac{k^2}{n^2\rho} + \frac{\log k}{n\rho}\right) \wedge 1$. Hence, the error rate of USVT is larger than the minimax optimal rate by at most a multiplicative factor of $k/\log k$, which resembles the computational gap observed in community detection (Banks et al., 2016; Abbe & Sandon, 2015) and the related high-dimensional statistical inference problems discussed in (Banks et al., 2018). In particular, it is shown in (Banks et al., 2016; Abbe & Sandon, 2015) that estimation better than randomly guessing is attainable efficiently by spectral methods when above the Kesten-Stigum threshold, while it is information-theoretically possible even strictly below the KS threshold by a multiplicative factor $k/\log k$ for large k. In between the KS threshold and information-theoretic threshold, non-trivial estimation is information-theoretically possible but believed to require exponential time. The same conclusion also holds for exact community recovery as shown in (Chen & Xu, 2014). Due to this coincidence, it is tempting to believe that $\frac{k}{n\rho} \wedge 1$ might be the optimal estimation rate among all polynomial-time algorithms; however, we do not have a proof.

2.2. Smooth graphon

Next we proceed to the smooth graphon setting. We assume $\mathcal{X} = [0,1)^d$ for simplicity. [3]. There are various notions to characterize the smoothness of graphon. In this paper, we focus on the following two notions, which are widely adopted in the non-parametric regression literature (Tsybakov, 2008). Given a function $g : \mathcal{X} \to \mathbb{R}$ and a multi-index κ, let

$$\nabla_\kappa g(x) = \frac{\partial^{|\kappa|} g(x)}{(\partial x)^\kappa} \tag{5}$$

denote its partial derivative whenever it exists.

Definition 3 (Hölder class). *Let α and L be two positive numbers. The Hölder class $\mathcal{H}(\alpha, L)$ on $\mathcal{X}$ is defined as the set of functions $g : \mathcal{X} \to \mathbb{R}$ whose partial derivatives satisfy*[4]

$$\sum_{\kappa : |\kappa| = \lfloor \alpha \rfloor} \frac{1}{\kappa!} \left| \nabla_\kappa g(x) - \nabla_\kappa g(x') \right| \le L \|x - x'\|_\infty^{\alpha - \lfloor \alpha \rfloor}. \tag{6}$$

[3] If $\mathcal{X}$ is a compact set in $\mathbb{R}^d$, then there exists a positive constant a such that $\mathcal{X} \subset [-a, a)^d$. Hence, the general compact set case can be reduced to $\mathcal{X} = [0,1)^d$ by a proper scaling.

[4] Changing the infinity-norm to a different norm (e.g. ℓ_2 norm) only changes L (by a factor may depending on d and α) and thus will not affect rates of convergence.

Note that if $\alpha \in (0,1]$, then (6) is equivalent to the Lip-α condition:

$$|g(x) - g(x')| \leq L\|x - x'\|_\infty^\alpha. \tag{7}$$

One can also measure the smoothness with respect to the underlying measure μ. This leads to the consideration of Sobolev space (Leoni, 2009). For ease of exposition, we assume μ is the Lebesgue measure. The main results can be extended to more general Borel measures.

Definition 4 (Sobolev space). *Let α and L be two positive numbers. The Sobolev space $S(\alpha, L)$ on $(\mathcal{X}, \mu)$ is defined as the set of functions $g : \mathcal{X} \to \mathbb{R}$ whose partial derivatives*[5] *satsify*

$$\sum_{\kappa:|\kappa|=\alpha} \int_{\mathcal{X}} \|\nabla_\kappa g(x)\|_2^2 \, \mathrm{d}x \leq L^2, \quad \textit{for integral } \alpha,$$

and for non-integral α,

$$\sum_{\kappa:|\kappa|=\lfloor\alpha\rfloor} \int_{\mathcal{X}\times\mathcal{X}} \frac{\|\nabla_\kappa g(x) - \nabla_\kappa g(y)\|_2^2}{\|x-y\|_2^{2(\alpha-\lfloor\alpha\rfloor)+d}} \, \mathrm{d}x\mathrm{d}y \leq L^2.$$

Note that the graphon $f(x, y)$ is a bi-variate function. We treat it as a function of x for every fixed y, and introduce the following two conditions on f.

Condition 1 (Hölder condition on f). *There exist two positive numbers α and L such that $f(\cdot, y) \in \mathcal{H}(\alpha, L)$ for every $y \in \mathcal{X}$.*

Condition 2 (Sobolev condition on f). *There exist two positive numbers α and L such that $f(\cdot, y) \in \mathcal{S}(\alpha, L(y))$ for every y, where $L(y) : \mathcal{X} \to \mathbb{R}$ satisfies that $\int_{\mathcal{X}} L^2(y)\mathrm{d}y \leq L^2$.*

The following key result shows that the eigenvalues of M drop off to zero in a polynomial rate depending on the smoothness index α of f.

Proposition 1. *Suppose that f satisfies either Condition 1 or Condition 2. Then there exists a constant $C = C(\alpha, L, d)$ only depending on α, L, and d such that for all integers $0 \leq r \leq n-1$,*

$$\frac{1}{n^2} \sum_{i \geq r+1} \mathbb{E}\left[\lambda_i^2(M)\right] \leq C(\alpha, L, d)\left(n^{-1} + r^{-2\alpha/d}\right).$$

Remark 1. *In the special case where f is Hölder smooth with $\alpha = 1$, Proposition 1 has been proved in (Chatterjee, 2015). In particular, it is shown in (Chatterjee, 2015) that f can be well-approximated by a piecewise constant function. As a consequence, M can be approximated by a rank-r block matrix with r^2 blocks, and the entry-wise approximation error in the squared Frobenius norm is shown to be approximately $r^{-2\alpha/d}$. The same idea can be readily extended to the case $\alpha \in [0,1]$. However, piecewise constant approximations of f no longer suffice for $\alpha > 1$, because Hölder smoothness condition (6) no longer implies Lip-α condition (7). In fact (7) with $\alpha > 1$ will imply that $f \equiv C$ for some constant C. Instead, we show that f can be well approximated by piecewise polynomials of degree $\lfloor\alpha\rfloor$.*

By combining Proposition 1 with Corollary 1, we immediately get the following result on the convergence rate of the estimation error of USVT.

Theorem 3. *Under the graphon estimation model, assume (3) holds, and f satisfies either Condition 1 or Condition 2. There exists a positive constant κ such that if $\tau = (1+\delta)\kappa\sqrt{n\rho}$ for some fixed constant $\delta > 0$, then*

$$\mathrm{MSE}(\widehat{M}) \leq c''(n\rho)^{-\frac{2\alpha}{2\alpha+d}},$$

where c'' is a positive constant independent of n.

Theorem 3 implies that if f is infinitely many times differentiable, then the MSE of USVT converges to zero faster than $(n\rho)^{-1+\epsilon}$ for an arbitrarily small constant $\epsilon > 0$. In fact, we can prove a sharper result when f is analytic, i.e., f is infinitely differentiable and its Taylor series expansion around any point in its domain converges to the function in some neighborhood of the point. One concerete example of analytic function which appears in the study of matrix completion is $f(x, y) = 1/(1+\exp(-\langle x, y\rangle))$ (Ganti et al., 2015).

Theorem 4. *Under the graphon estimation model, suppose there there exists positive constants a and b such that for all multi-indices κ and all $y \in \mathcal{X}$*

$$\sup_{x\in\mathcal{X}} \frac{\partial^{|\kappa|} f(x,y)}{(\partial x)^\kappa} \leq b a^{|\kappa|} \kappa!. \tag{8}$$

There exists positive constants c_0 and c_1 only depending on a, b, d such that for all integers $0 \leq r \leq n-1$,

$$\frac{1}{n^2} \sum_{i \geq r+1} \lambda_i^2(M) \leq c_1\left(n^{-1} + \exp\left(-c_0 r^{1/d}\right)\right). \tag{9}$$

Moreover, assume (3) holds. Then there exists positive constants c', c'' such that if $\tau = c''\sqrt{n\rho}$,

$$\mathrm{MSE}(\widehat{M}) \leq c' \frac{\log^d(n\rho)}{n\rho}.$$

We remark that for a fixed $y \in \mathcal{X}$, (8) is a sufficient and necessary condition for $f(\cdot, y)$ being analytic (Komatsu, 1960). Note that (9) implies the eigenvalues of M has a super-polynomial decay with rate $\alpha = 1/d$. Its proof is

[5] More generally, the Sobolev space is defined when only weak derivatives exist (Leoni, 2009).

based on approximating $f(\cdot, y)$ using its Taylor series truncated at degree $\ell \asymp r^{1/d}$. When $d = 1$, the eigenvalues of M decays to zero exponentially fast in r; such an exponential decay can be also proved via Chebyshev polynomial approximation of f as shown in (Little & Reade, 1984).

2.2.1. Comparison to minimax optimal rates

We compare the rates of convergence of USVT for estimating Hölder smooth graphons to the minimax optimal rates when the dimension of latent feature space $d = 1$ (Gao et al., 2015; Klopp et al., 2015; Gao et al., 2016):

$$R_n^* \asymp \begin{cases} 1, & n\rho = O(1) \\ \frac{\log(n\rho)}{n\rho}, & \omega(1) \le n\rho \le n^\alpha (\log n)^{\alpha+1} \\ (n^2\rho)^{-\frac{\alpha}{\alpha+1}}, & n\rho \ge n^\alpha(\log n)^{\alpha+1} \end{cases}. \tag{10}$$

Thus, as graphon gets smoother, i.e., α increases, the upper bound to the rate of convergence of USVT $(n\rho)^{-2\alpha/(2\alpha+1)}$ approaches the minimax optimal rate $\log(n\rho)/(n\rho)$.

2.3. Connections to spectrum of integral operators

We state a useful result, connecting the eigenvalues of M to the spectrum of an integral operator defined in terms of f. This allows us to translate existing results on the decay rates of eigenvalues of integral operators to those of M. Define an operator $\mathcal{T} : L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu)$ as

$$(\mathcal{T}g)(x) \triangleq \int_{\mathcal{X}} f(x,y)g(y)\mu(\mathrm{d}y), \ \forall g \in L^2(\mathcal{X}, \mu), \tag{11}$$

where f acts as a kernal function. Hence, M can be also viewed as a kernal matrix. We assume that the graphon f is square-integrable, i.e., $\int_{\mathcal{X}\times\mathcal{X}} f^2(x,y)\mu(\mathrm{d}x)\mu(\mathrm{d}y) < \infty$. In this case, the operator $\mathcal{T}$ is known as Hilbert-Schmidt integral operator, which is compact. Therefore it admits a discrete spectrum with finite multiplicity of all of its nonzero eigenvalues (see e.g. (Kato, 1966; Koltchinskii, 1998; von Luxburg et al., 2005)). Moreover, any of its eigenfunctions is continuous on $\mathcal{X}$. Denote the eigenvalues of operator $\mathcal{T}$ sorted in decreasing order by $|\lambda_1(\mathcal{T})| \ge |\lambda_2(\mathcal{T})| \ge \cdots$ and its corresponding eigenfunctions with unit $L^2(\mathcal{X}, \mu)$ norm by $\phi_1, \phi_2, \cdots$. By the definition of λ_k and ϕ_k, we have

$$\int_{\mathcal{X}\times\mathcal{X}} \left(f(x,y) - \sum_{k=1}^m \lambda_k(\mathcal{T})\phi_k(x)\phi_k(y) \right)^2 \mu(\mathrm{d}x)\mu(\mathrm{d}y) \to 0, \text{ as } m \to \infty \tag{12}$$

see, e.g., (Kato, 1966)[Chapter Five, Section 2.4].

The following theorem upper bounds the tail of eigenvalues of M in expectation using the tail of eigenvalues of $\mathcal{T}$. Previous results in (Koltchinskii & Giné, 2000) provide similar upper bounds to the ℓ_2 distance between the ordered eigenvalues of M and those of $\mathcal{T}$.

Theorem 5. *For any integer $r \ge 0$,*

$$\frac{1}{n^2}\sum_{k>r} \mathbb{E}\left[\lambda_k^2(M)\right] \le \sum_{k>r}^{\infty} \lambda_k^2(\mathcal{T}) + \frac{1}{n}\sum_{k,\ell=1}^{r} \lambda_k(\mathcal{T})\lambda_\ell(\mathcal{T})\mathbb{E}\left[\phi_k^2(x_1)\phi_\ell^2(x_1)\right]. \tag{13}$$

The second term on the right hand side of (13) is responsible for the contribution of the diagonal entries of M. When $\mathbb{E}\left[\phi_k^2(x_1)\phi_\ell^2(x_1)\right]$ is bounded and $\sum_{k=1}^\infty \lambda_k(\mathcal{T}) < \infty$, this second term is on the order of n^{-1}.

It is well known that if the kernel function f is smoother, the eigenvalues of $\mathcal{T}$ drops to zero faster. There is vast literature on estimating the decay rates of the eigenvalues of $\mathcal{T}$ in terms of the smoothness conditions of f, see, e.g., (Krein, 1965; Birman & Solomyak, 1977; König, 2013; Delgado & Ruzhansky, 2014). Theorem 5 allows us to translate those existing results on the decay rates of eigenvalues of $\mathcal{T}$ to those of M, as illustrated by examples in Section 3.

3. Numerical examples

We provide numerical results on synthetic datasets, which corroborate our theoretical results. Additional numerical results on stochastic block models can be found in the full paper (Xu, 2017). We assume the sparsity level ρ is known and set the threshold $\tau = 2.01\sqrt{n\rho}$ throughout the experiments. In the case where ρ is unknown, one can apply cross-validation procedure to adaptively choose the sparsity level ρ as shown in (Gao et al., 2016). We first apply USVT with input (A, τ, ρ), and then output the estimator $\widehat{M}$, and finally calculate the MSE error MSE$(\widehat{M})$.

3.1. Translation invariant graphon

For some $a > 0$, let $h : [-a, a] \to \mathbb{R}$ denote an even function, i.e., $h(x) = h(-x)$. Let us extends its domain to the real line by the periodic extension such that $h(x + 2ka) = h(x)$ for all $x \in [-a, a]$ and integers $k \in \mathbb{Z}$. By construction h has a period $2a$. Using this function, we can define a translation-invariant graphon on the product space $[-a, a] \times [-a, a]$ via $f(x, y) = h(x - y)$. Since h is even, it follows that f is symmetric. Then the integral operator $\mathcal{T}$ defined in (11) reduces to: for all $x \in [-a, a]$,

$$(\mathcal{T}g)(x) = \frac{1}{2a}\int_{-a}^{a} h(x-y)g(y)\mathrm{d}y = \frac{1}{2a}(h * g)(x),$$

where $*$ denotes the convolution. Hence, we can explicitly determine the eigenvalues of $\mathcal{T}$ via Fourier analysis. In particular, let $\widehat{h}[k]$ denote the Fourier coefficients:

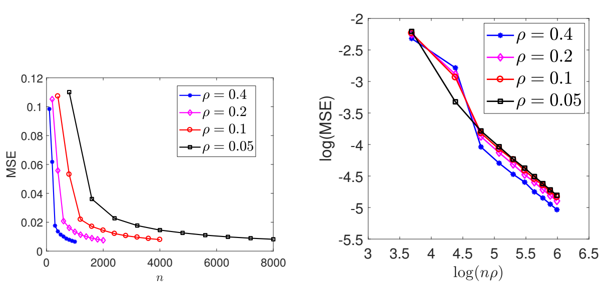
$$\widehat{h}[k] = \frac{1}{2a} \int_{-a}^{a} h(x) e^{-j\pi k x/a} \mathrm{d}x,$$

where throughout this section j denotes the imaginary part such that $j^2 = -1$, Since h is even, it follows that $\widehat{h}[k]$'s are real and $\widehat{h}[k] = \widehat{h}[-k]$. Fourier analysis entails a one-to-one correspondence between eigenvalues of $\mathcal{T}$ and Fourier coefficients of h: $\lambda_k(\mathcal{T}) = \widehat{h}[k]$.

We specify $h : [-1, 1] \to \mathbb{R}$ as $h(x) = |x|$ and simulate the graphon model with $f(x, y) = h(x - y)$ for $x, y \in [-1, 1]$ and the underlying measure μ being uniform over $[-1, 1]$. Since $h(x) = |x|$, the Fourier coefficients can be explicitly computed as $\lambda_k(\mathcal{T}) = \widehat{h}[k] = 2\sin^2(\pi k/2)/(\pi^2 k^2)$ with eigenfunctions given by $\{\cos(\pi k x)\}_{k=0}^{\infty}$ and $\sin(\pi k x)\}_{k=1}^{\infty}$. It follows from Theorem 5 that the eigenvalues of M satisfy

$$\frac{1}{n^2} \sum_{i \geq r+1} \mathbb{E}\left[\lambda_i^2(M)\right] \leq O(n^{-1}) + O(r^{-3}) \tag{14}$$

uniformly over all integers $r \geq 0$. Therefore, our theory predicts that the MSE of USVT converges to zero at least in a rate of $(n\rho)^{-3/4}$. The simulation results for varying observation probabilities are depicted in Fig. 1. Panel (a) shows the MSE converges to 0 as the number of vertices n increases. In Panel (b), we rescale the x-axis to $\log(n\rho)$ and the y-axis to the log of MSE. The curves for different ρ align well with each other after the rescaling and decrease linearly with a slope of approximately 0.8, which is close to $3/4$ as predicted by our theory.



(a) (b)

Figure 1. The MSE error of USVT estimator under the translation invariant graphon $f(x, y) = |x - y|$. Panel (a): MSE versus the number of vertices n; Panel (b): The log of MSE versus $\log(n\rho)$. Each point represents the average of MSE over 10 independent runs.

3.2. Sobolev graphon

In this section, we simulate the graphon model with $\mathcal{X} = [0, 1]$ and μ being the uniform measure and $f(x, y) = \min\{x, y\}$. Then $\nabla_x f(x, y) = \mathbf{1}_{\{x \leq y\}}$ and $\nabla_y f(x, y) = \mathbf{1}_{\{y \leq x\}}$. Moreover, $|f(x, y) - f(x', y')| \leq |x - x'| + |y - y'|$. However, the second-order weak derivatives of f do not exist. Therefore, f is Sobolev smooth with $\alpha = 1$. In this case, one can get a bound on the eigenvalue decay rate tighter than Proposition 1 by directly computing $\lambda_k(\mathcal{T})$ and invoking Theorem 5. Note that

$$(\mathcal{T} g)(x) = \int_0^x y g(y) \mathrm{d}y + x \int_x^1 g(y) \mathrm{d}y.$$

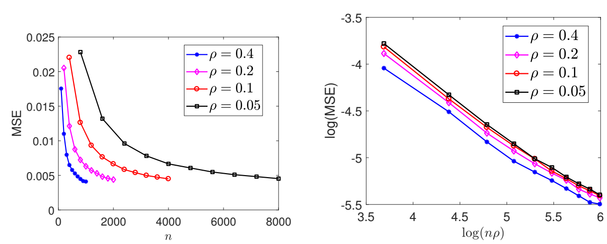
Suppose ϕ is an eigenfunction of $\mathcal{T}$ with eigenvalue λ. Then

$$\int_0^x y\phi(y)\mathrm{d}y + x \int_x^1 \phi(y)\mathrm{d}y = \lambda\phi(x).$$

It follows that $\phi(0) = 1$ and $\lambda\phi'(x) = \int_x^1 \phi(y)\mathrm{d}y$. It further implies that $\phi'(1) = 0$ and $\lambda\phi'' + \phi = 0$. Therefore, the eigenfunction and eigenvalue pairs are given by

$$\phi_k(x) = \sin\frac{(2k-1)\pi x}{2}, \text{ and } \lambda_k(\mathcal{T}) = \left(\frac{2}{(2k-1)\pi}\right)^2.$$

It follows from Theorem 5 that the eigenvalues of M satisfy (14) uniformly over all integers $r \geq 0$. Therefore, our theory predicts that the MSE of USVT converges to zero in a rate of $(n\rho)^{-3/4}$. The simulation results for varying observation probabilities are depicted in Fig. 2. The curves in Panel (b) for different ρ align well with each other after the rescaling and decrease linearly with a slope of approximately 0.7, which is close to $3/4$ as predicted by our theory.



(a) (b)

Figure 2. The MSE error of USVT estimator under the first-order sobolev graphon $f(x, y) = \min\{x, y\}$. Panel (a): MSE versus the number of vertices n; Panel (b): The log of MSE versus $\log(n\rho)$. Each point represents the average of MSE over 10 independent runs.

4. Conclusions and future work

In this paper, we establish upper bounds to the graphon estimation error of USVT when the average vertex degree is at least logarithmic in n. Our results can be extended to the case of bounded average degrees by first trimming the high-degree vertices (Feige & Ofek, 2005) and then applying USVT. We leave this extension as future work. Another fundamental and open question is whether the minimax optimal rate can be achieved in polynomial-time.

References

Abbe, E. and Sandon, C. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. arXiv 1512.09080, Dec 2015.

Airoldi, E. M., Costa, T. B., and Chan, S. H. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems 26*, pp. 692–700, 2013.

Banks, J., Moore, C., Neeman, J., and Netrapalli, P. Information-theoretic thresholds for community detection in sparse networks. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, NY, June 23-26 2016*, pp. 383–416, 2016.

Banks, J., Moore, C., Vershynin, R., Verzelen, N., and Xu, J. Information-theoretic bounds and phase transitions in clustering, sparse pca, and submatrix localization. *IEEE Transactions on Information Theory*, PP(99):1–1, 2018.

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

Birman, M. S. and Solomyak, M. Z. Piecewise-polynomial approximations of functions of the classes W_p^α. *Mathematics of the USSR-Sbornik*, 2(3):295, 1967.

Birman, M. S. and Solomyak, M. Z. Estimates of singular numbers of integral operators. *Russian Mathematical Surveys*, 32(1):15, 1977.

Borgs, C., Chayes, J. T., Lovász, L., Sós, V. T., and Vesztergombi, K. Convergent sequences of dense graphs i: Subgraph frequencies, metric properties and testing. *Advances in Mathematics*, 219(6):1801–1851, 2008.

Borgs, C., Chayes, J. T., Lovász, L., Sós, V. T., and Vesztergombi, K. Convergent sequences of dense graphs ii. multiway cuts and statistical physics. *Annals of Mathematics*, 176(1):151–219, 2012.

Borgs, C., Chayes, J., and Smith, A. Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems*, pp. 1369–1377, 2015a.

Borgs, C., Chayes, J. T., Cohn, H., and Ganguly, S. Consistent nonparametric estimation for heavy-tailed sparse graphs. *arXiv preprint arXiv:1508.06675*, 2015b.

Borgs, C., Chayes, J., Lee, C. E., and Shah, D. Thy friend is my friend: Iterative collaborative filtering for sparse matrix estimation. In *Advances in Neural Information Processing Systems*, pp. 4718–4729, 2017.

Cai, D., Ackerman, N., and Freer, C. An iterative step-function estimator for graphons. *arXiv preprint arXiv:1412.2129*, 2014.

Chan, S. and Airoldi, E. A consistent histogram estimator for exchangeable graph models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 208–216, 2014.

Chatterjee, S. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

Chen, Y. and Xu, J. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. In *Proceedings of ICML 2014 (Also arXiv:1402.1267)*, Feb 2014.

Delgado, J. and Ruzhansky, M. Schatten classes on compact manifolds: kernel conditions. *Journal of Functional Analysis*, 267(3):772–798, 2014.

Feige, U. and Ofek, E. Spectral techniques applied to sparse random graphs. *Random Struct. Algorithms*, 27(2):251–275, Sept. 2005. ISSN 1042-9832. doi: 10.1002/rsa.v27:2. URL http://dx.doi.org/10.1002/rsa.v27:2.

Ganti, R., Balzano, L., and Willett, R. Matrix completion under monotonic single index models. In *Proceedings of Neural Information Processing Systems (NIPS)*, Dec. 2015.

Gao, C., Lu, Y., Zhou, H. H., et al. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.

Gao, C., Lu, Y., Ma, Z., and Zhou, H. H. Optimal estimation and completion of matrices with biclustering structures. *Journal of Machine Learning Research*, 17(161):1–29, 2016.

Hajek, B., Wu, Y., and Xu, J. Semidefinite programs for exact recovery of a hidden community. In *Proceedings of Conference on Learning Theory (COLT)*, pp. 1051–1095, New York, NY, Jun 2016. arXiv:1602.06410.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170 (2):301–354, 2007.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97:1090+, December 2002.

Holland, P. W., Laskey, K. B., and Leinhardt, S. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

Kato, T. *Perturbation Theory for Linear Operators*. Springer, Berlin, 1966.

Klopp, O. and Verzelen, N. Optimal graphon estimation in cut distance. *arXiv preprint arXiv:1703.05101*, 2017.

Klopp, O., Tsybakov, A. B., and Verzelen, N. Oracle inequalities for network models and sparse graphon estimation. *arXiv preprint arXiv:1507.04118*, 2015.

Klopp, O. et al. Rank penalized estimators for high-dimensional matrices. *Electronic Journal of Statistics*, 5: 1161–1183, 2011.

Koltchinskii, V. and Giné, E. Random matrix approximation of spectra of integral operators. *Bernoulli*, pp. 113–167, 2000.

Koltchinskii, V., Lounici, K., Tsybakov, A. B., et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

Koltchinskii, V. I. Asymptotics of spectral projections of some random matrices approximating integral operators. *Progress in Probability*, 1998.

Komatsu, H. A characterization of real analytic functions. *Proceedings of the Japan Academy*, 36(3):90–93, 1960.

König, H. *Eigenvalue distribution of compact operators*, volume 16. Birkhäuser, 2013.

Krein, I. G. Y. M. Introduction to the theory of linear nonselfadjoint operators in hilbert space. *American Mathematical Society*, 1965.

Krivelevich, M. and Sudakov, B. The largest eigenvalue of sparse random graphs. *Combinatorics, Probability and Computing*, 12(1):61–72, 2003.

Lee, C. E. and Shah, D. Unifying framework for crowd-sourcing via graphon estimation. *arXiv preprint arXiv:1703.08085*, 2017.

Leoni, G. *A first course in Sobolev spaces*, volume 105. American Mathematical Society Providence, RI, 2009.

Little, G. and Reade, J. Eigenvalues of analytic kernels. *SIAM journal on mathematical analysis*, 15(1):133–136, 1984.

Lovász, L. *Large networks and graph limits*, volume 60. American Mathematical Society Providence, 2012.

Lovász, L. and Szegedy, B. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957, 2006.

Miller, K., Jordan, M. I., and Griffiths, T. L. Nonparametric latent feature models for link prediction. In *Advances in neural information processing systems*, pp. 1276–1284, 2009.

Moore, C. The computer science and physics of community detection: landscapes, phase transitions, and hardness. *arXiv preprint arXiv:1702.00467*, 2017.

Pensky, M. Dynamic network models and graphon estimation. *arXiv preprint arXiv:1607.00673*, 2016.

Shah, N., Balakrishnan, S., Guntuboyina, A., and Wainwright, M. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. In *International Conference on Machine Learning*, pp. 11–20, 2016.

Song, D., Lee, C. E., Li, Y., and Shah, D. Blind regression: Nonparametric regression for latent variable models via collaborative filtering. In *Advances in Neural Information Processing Systems*, pp. 2155–2163, 2016.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

von Luxburg, U., Bousquet, O., and Belkin, M. On the convergence of spectral clustering on random samples: the normalized case. *NIPS*, 2005.

Wolfe, P. J. and Olhede, S. C. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.

Xu, J. Rates of convergence of spectral methods for graphon estimation. arXiv 1709.03183, Sept. 2017.

Xu, J., Massoulié, L., and Lelarge, M. Edge label inference in generalized stochastic block models: from spectral theory to impossibility results. In *COLT*, pp. 903–920, 2014.

Yang, J., Han, C., and Airoldi, E. Nonparametric estimation and testing of exchangeable graph models. In *Artificial Intelligence and Statistics*, pp. 1060–1067, 2014.

Yu, H., Braun, P., Yıldırım, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., et al. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, 2008.

Zhang, Y., Levina, E., and Zhu, J. Estimating network edge probabilities by neighborhood smoothing. *arXiv preprint arXiv:1509.08588*, 2015.