Matrix Estimation, Latent Variable Model and Collaborative Filtering[*][†]

Devavrat Shah

Massachusetts Institute of Technology, Cambridge, USA
devavrat@mit.edu

Abstract

Estimating a matrix based on partial, noisy observations is prevalent in variety of modern applications with recommendation system being a prototypical example. The non-parametric latent variable model provides canonical representation for such matrix data when the underlying distribution satisfies "exchangeability" with graphons and stochastic block model being recent examples of interest. Collaborative filtering has been a successfully utilized heuristic in practice since the dawn of e- commerce. In this extended abstract, we will argue that collaborative filtering (and its variants) solve matrix estimation for a generic latent variable model with near optimal sample complexity.

1998 ACM Subject Classification Probability and Statistics, Non-parametric Statistics

Keywords and phrases Matrix Estimation, Graphon Estimation, Collaborative Filtering

Digital Object Identifier 10.4230/LIPIcs.FSTTCS.2017.4

Category Invited Paper

1 Introduction

We consider the question of sparse matrix estimation (or completion) with noisy observations. As a prototype for such a problem, consider a noisy observation of a social network where observed interactions are signals of true underlying connections. We might want to predict the probability that two users would choose to connect if recommended by the platform, e.g. LinkedIn. As a second example, consider a recommendation system where we observe movie ratings provided by users, and we may want to predict the probability distribution over ratings for specific movie-user pairs. A popular collaborative filtering approach suggests using "similarities" between pairs of users to estimate the probability of a connection being formed or a movie being liked. Traditionally, the similarities between pair of users in a social network is computed by comparing the set of their friends or in the context of movie recommendation, by comparing commonly rated movies. In the sparse setting, however most pairs of users have no common friends, or most pairs of users have no commonly rated movies; thus there is insufficient data to compute the traditional similarity metrics.

In this work, the primary interest is to understand how well does such a simple, intuitive approach to compute similarities between pair of users for matrix estimation work. In the

[*] This work was partially supported by DARPA project W911NF-16-1-0551 and NSF project CMMI-1462158 .

[†] This extended abstract of talk at FSTTCS 2017 is based on two works: Lee, Li, Shah and Song (2016) [27] and Borgs, Chayes, Lee and Shah (2017) [7].

37th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2017).
Editors: Satya Lokam and R. Ramanujam; Article No. 4; pp. 4:1–4:8
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

process, we provide a way to extend the notion of the similarities utilized in practice by incorporating information within a larger radius neighborhood rather than restricting only to immediate neighbors. We establish that it achieves best-known sample complexity which matches well known, conjectured lower bound for a special instance of the generic problem, the mixed membership stochastic block model.

1.1 Model, Problem Statement

The question discussed above can be mathematically formulated as a matrix estimation problem. Let F be an $n \times n$ matrix which we would like to estimate, and let Z be a noisy signal of matrix F such that $\mathbb{E}[Z] = F$. The available data is denoted by $(\mathcal{E}, M)$, where $\mathcal{E} \subset [n] \times [n]$ denotes the subset of indices for which data is observed, and M is the $n \times n$ symmetric data matrix[1] where $M(u,v) = Z(u,v)$ for $(u,v) \in \mathcal{E}$, and $M(u,v) = 0$ for $(u,v) \notin \mathcal{E}$. We can equivalently represent the data with an undirected weighted graph $\mathcal{G}$ with vertex set $[n]$, edge set $\mathcal{E}$, and edge weights given by M. We shall use graph and matrix notations in an interchangeable manner. Given the data $(\mathcal{E}, M)$, we would like to estimate the original matrix F. We assume a uniform sampling model, where each entry is observed with probability p independently of all other entries.

We shall assume that each $u \in [n]$ is associated to a latent variable $\alpha_u \in \mathcal{X}_1$, which is drawn independently across indices $[n]$ as per distribution P_1 over a bounded compact space $\mathcal{X}_1$. We shall assume that the expected data matrix can be described by the latent function f, i.e. $F(u,v) = f(\alpha_u, \alpha_v)$, where $f : \mathcal{X}_1 \times \mathcal{X}_1 \to \mathbb{R}$ is a symmetric function. We note that such a structural assumption or the so-called Latent Variable Model is a canonical representation for exchangeable arrays as shown by Aldous and Hoover [5, 22, 6]. For each observation, we assume that $\mathbb{E}[Z(u,v)] = F(u,v)$, $Z(u,v)$ is bounded and $\{Z(u,v)\}_{1 \le u < v \le n}$ are independent conditioned on the node latent variables.

The goal is to find smallest p, as a function of n and structural properties of f, so that there exists an algorithm that can produce $\hat{F}$, an estimate of matrix F, so that the Mean-Squared-Error (MSE) between $\hat{F}$ and F, $\frac{1}{n^2}\sum_{u,v \in [n]}(\hat{F}(u,v) - F(u,v))^2$, converges to 0 as $n \to \infty$.

1.2 Related Works

The matrix estimation problem introduced above, as special cases, includes problems from different areas of literature: matrix completion popularized in the context of recommendation systems, graphon estimation arising from the asymptotic theory of graphs, and community detection using the stochastic block model or its generalization known as the mixed membership stochastic block model. We shall discuss key representative results here. We discuss the scaling of the sample complexity with respect to d (model complexity, usually rank) and n for polynomial time algorithms, including results for both mean squared error convergence, exact recovery in the noiseless setting, and convergence with high probability in the noisy setting.

In the context of matrix completion, there has been much progress under the low-rank assumption and additive noise model. Most theoretically founded methods are based on spectral decompositions or minimizing a loss function with respect to spectral constraints [23, 24, 14, 15, 32, 30, 18, 17, 16].

[1] The asymmetric variation of this question can be casted as symmetric version. See [7] for a detailed discussion.

Most of the results in matrix completion require additive noise models, which do not extend to setting when the observations are binary or quantized. The Universal Singular Value Thresholding (USVT) estimator [16] is able to handle general bounded noise, although it requires a few log factors more in its sample complexity compared to what is conjectured optimal and what our result achieves for low-rank matrices.

There is also a significant amount of literature which looks at the estimation problem when the data matrix is binary, also known as 1-bit matrix completion, stochastic block model (SBM) parameter estimation, or graphon estimation. The latter two terms are found within the context of community detection and network analysis, as the binary data matrix can alternatively be interpreted as the adjacency matrix of a graph – which are symmetric, by definition. Under the SBM, each vertex is associated to one of d community types, and the probability of an edge is a function of the community types of both endpoints. Estimating the $n \times n$ parameter matrix becomes an instance of matrix estimation. In SBM, the expected matrix is at most rank d due to its block structure. Precise thresholds for cluster detection (better than random) and estimation have been established by [1, 2, 3]. Our work, both algorithmically and technically, draws insight from this sequence of works, extending the analysis to a broader class of generative models through the design of an iterative algorithm, and improving the technical results with precise MSE bounds.

The mixed membership stochastic block model (MMSBM) allows each vertex to be associated to a length d vector, which represents its weighted membership in each of the d communities. The probability of an edge is a function of the weighted community memberships vectors of both endpoints, resulting in an expected matrix with rank at most d. Recent work by [33] provides an algorithm for weak detection for MMSBM with sample complexity d^2n, when the community membership vectors are sparse and evenly weighted. They provide partial results to support a conjecture that d^2n is a computational lower bound, separated by a gap of d from the information theoretic lower bound of dn. Our result achieves close to this conjectured lower bound, with a sample complexity of $\omega(d^5n)$ in order to guarantee consistency, which is much stronger than weak detection, in a much more generic setting.

Graphon estimation extends SBM and MMSBM to the generic Latent Variable Model where the probability of an edge can be any measurable function f of real-valued types (or latent variables) associated to each endpoint. Graphons were first defined as the limiting object of a sequence of large dense graphs [13, 19, 29], with recent work extending the theory to sparse graphs [11, 12, 10, 34]. In the graphon estimation problem, we would like to estimate the function f given an instance of a graph generated from the graphon associated to f.

[20, 25] provide minimax optimal rates for graphon estimation; however a majority of the proposed estimators are not computable in polynomial time, since they require optimizing over an exponentially large space (e.g. least squares or maximum likelihood) [35, 9, 8, 20, 25]. [9] provided a polynomial time method based on degree sorting in the special case when the expected degree function is monotonic. To our knowledge, existing positive results for sparse graphon estimation require either strong monotonicity assumptions [9], or rank constraints as assumed in the SBM, the 1-bit matrix completion, and in this work.

We call special attention to the similarity based methods which are able to bypass the rank constraints, relying instead on smoothness properties of the latent function f (e.g. Lipschitz) [36] as well as this work [27, 7]. They hinge upon computing similarities between rows or columns by comparing commonly observed entries. Similarity based methods, also known in the literature as collaborative filtering, have been successfully employed across many large scale industry applications (Netflix, Amazon, Youtube) due to its simplicity and scalability [21, 28, 26, 31]; however the theoretical results have been relatively sparse.

These recent results suggest that the practical success of these methods across a variety of applications may be due to its ability to capture local structure.

A key limitation of this approach that this work overcomes is ability to deal with sparsity. In particular, [36] works when $p = 1$ (or all data is observed). Our result requires for $p = \omega(n^{-1/2})$ for any Lipschitz function instead, and $p = \omega(d^5 n^{-1})$ when the underlying model is low-rank with rank being d.

2 Algorithm

The algorithm that we propose uses the concept of local approximation, first determining which datapoints are similar in value, and then computing neighborhood averages for the final estimate. All similarity-based collaborative filtering methods have the following basic format:

1. Compute distances between pairs of vertices, e.g.,

$$\texttt{dist}(u, a) \approx \textstyle\int_{\mathcal{X}_1} (f(\alpha_u, t) - f(\alpha_a, t))^2 dP_1(t). \tag{1}$$

2. Form estimate by averaging over "nearby" datapoints,

$$\hat{F}(u, v) = \tfrac{1}{|\mathcal{E}_{uv}|} \textstyle\sum_{(a,b)\in\mathcal{E}_{uv}} M(a, b), \tag{2}$$

where $\mathcal{E}_{uv} := \{(a, b) \in \mathcal{E} \ s.t. \ \texttt{dist}(u, a) < \xi(n), \texttt{dist}(v, b) < \xi(n)\}$.

We will choose the threshold $\xi(n)$ depending on `dist` in order to guarantee that it is small enough to drive the bias to zero, ensuring the included datapoints are close in value, yet large enough to reduce the variance, ensuring $|\mathcal{E}_{uv}|$ diverges. In what follows, we describe two methods to compute distances. The first method uses immediate neighbors to estimate distance while the second utilizes far-away neighbors to estimate distance. Therefore, the first method works well when we have denser sampling, $p = \omega(n^{-1/2})$ while the second works for much sparse regime.

2.1 Distance using Immediate Neighbors

Like classical collaborative filtering using in practice, we approximate the L_2 distance of (1) by using variants of the finite sample approximation,

$$\texttt{dist}_0(u, a) = \tfrac{1}{|\mathcal{O}_{ua}|} \textstyle\sum_{y \in \mathcal{O}_{ua}} (F(u, y) - F(a, y))^2, \tag{3}$$

where $y \in \mathcal{O}_{ua}$ iff $(u, y) \in \mathcal{E}$ and $(a, y) \in \mathcal{E}$ [4, 36, 27]. This approach works well when $p = \omega(n^{-1/2})$. However, for much sparser setting (i.e. $p = o(n^{-1/2})$) with high probability, $\mathcal{O}_{ua} = \emptyset$ for almost all pairs (u, a), such that this distance cannot be computed. This requires us to utilize far-way neighbors.

2.2 Distance using Far-away Neighbors

Some Notations. We shall assume that f has finite spectrum with rank d when regarded as an integral operator, i.e. for any $\alpha_u, \alpha_v \in \mathcal{X}_1$,

$$f(\alpha_u, \alpha_v) = \sum_{k=1}^{d} \lambda_k q_k(\alpha_u) q_k(\alpha_v),$$

where $\lambda_k \in \mathbb{R}$ for $1 \leq k \leq d$, q_k are orthonormal ℓ_2 functions for $1 \leq k \leq d$ such that

$$\int_{\mathcal{X}_1} q_k(y)^2 dP_1(y) = 1 \text{ and } \int_{\mathcal{X}_1} q_k(y)q_h(y)dP_1(y) = 0 \text{ for } k \neq h.$$

We assume that there exists some B_q such that $\sup_{y\in[0,1]} |q_k(y)| \leq B_q$ for all k.

Let Λ denote the $d \times d$ diagonal matrix with $\{\lambda_k\}_{k\in[d]}$ as the diagonal entries, and let Q denote the $d \times n$ matrix where $Q(k,u) = q_k(\alpha_u)$. Since Q is a random matrix depending on the sampled α, it is not guaranteed to be an orthonormal matrix (even though q_k are orthonormal functions). By definition, it follows that $F = Q^T \Lambda Q$. Let $d' \leq d$ be the number of distinct valued eigenvalues amongst $\lambda_k, 1 \leq k \leq d$. Let $\tilde{\Lambda}$ denote the $d \times d'$ matrix where $\tilde{\Lambda}(a,b) = \lambda_a^{b-1}$.

Intuition. To that end, visualize the data via a graph with edge set $\mathcal{E}$, then (3) corresponds to comparing common neighbors of vertices u and a. A natural extension when u and a have no common neighbors, is to instead compare the r-hop neighbors of u and a, i.e. vertices y which are at distance exactly r from both u and a. We compare the product of weights along edges in the path from u to y and a to y respectively, which in expectation approximates

$$\begin{aligned}
&\textstyle\int_{\mathcal{X}_1^{r-1}} f(\alpha_u, t_1)(\prod_{s=1}^{r-2} f(t_s, t_{s+1}))f(t_{r-1}, \alpha_y) d\prod_{i\in[r-1]} P_1(t_i) \\
&\quad = \textstyle\sum_k \lambda_k^r q_k(\alpha_u) q_k(\alpha_y) \\
&\quad = e_u^T Q^T \Lambda^r Q e_y. \qquad (4)
\end{aligned}$$

We choose a large enough r such that there are sufficiently many "common" vertices y which have paths to both u and a, guaranteeing that our distance can be computed from a sparse dataset.

Definition of Distance. We first expand local neighborhoods of radius r around each vertex. Let $\mathcal{S}_{u,s}$ denote the set of vertices which are at distance s from vertex u in the graph defined by edge set[2] $\mathcal{E}$. Specifically, $i \in \mathcal{S}_{u,s}$ if the shortest path in $\mathcal{G} = ([n], \mathcal{E})$ from u to i has a length of s. Let $\mathcal{B}_{u,s}$ denote the set of vertices which are at distance at most s from vertex u in the graph defined by $\mathcal{E}$, i.e. $\mathcal{B}_{u,s} = \cup_{t=1}^{s} \mathcal{S}_{u,t}$. Let $\mathcal{T}_u$ denote a breadth-first tree in $\mathcal{G}$ rooted at vertex u. The breadth-first property ensures that the length of the path from u to i within $\mathcal{T}_u$ is equal to the length of the shortest path from u to i in $\mathcal{G}$. If there is more than one valid breadth-first tree rooted at u, choose one uniformly at random. Let $N_{u,r} \in [0,1]^n$ denote the following vector with support on the boundary of the r-radius neighborhood of vertex u (we also call $N_{u,r}$ the neighborhood boundary):

$$N_{u,r}(i) = \begin{cases} \prod_{(a,b)\in \text{path}_{\mathcal{T}_u}(u,i)} M(a,b) & \text{if } i \in \mathcal{S}_{u,r}, \\ 0 & \text{if } i \notin S_{u,r}, \end{cases}$$

where $\text{path}_{\mathcal{T}_u}(u,i)$ denotes the set of edges along the path from u to i in the tree $\mathcal{T}_u$. The sparsity of $N_{u,r}(i)$ is equal to $|\mathcal{S}_{u,r}|$, and the value of the coordinate $N_{u,r}(i)$ is equal to the product of weights along the path from u to i. Let $\tilde{N}_{u,r}$ denote the normalized neighborhood boundary such that $\tilde{N}_{u,r} = N_{u,r}/|\mathcal{S}_{u,r}|$. We will choose radius $r = \frac{6\ln(1/p)}{8\ln(pn)}$.

[2] For establishing correctness of algorithm, the edges are divided into three random subsets. See [7] for details. Here, to keep exposition simple, we will ignore this technical aspect. We conjecture that similar results hold for this variant of the algorithm as well.

1. For each pair (u, v), compute $\texttt{dist}_1(u, v)$ according to

$$\left(\tfrac{1-p}{p}\right)\left(\tilde{N}_{u,r} - \tilde{N}_{v,r}\right)^T M\left(\tilde{N}_{u,r+1} - \tilde{N}_{v,r+1}\right).$$

2. For each pair (u, v), compute distance according to

$$\texttt{dist}_2(u, v) = \textstyle\sum_{i\in[d']} z_i \Delta_{uv}(r, i),$$

where $\Delta_{uv}(r, i)$ is defined as

$$\left(\tfrac{1-p}{p}\right)\left(\tilde{N}_{u,r} - \tilde{N}_{v,r}\right)^T M\left(\tilde{N}_{u,r+i} - \tilde{N}_{v,r+i}\right),$$

and $z \in \mathbb{R}^{d'}$ is a vector that satisfies $\Lambda^{2r+2}\tilde{\Lambda}z = \Lambda^2\mathbf{1}$. z always exists and is unique because $\tilde{\Lambda}$ is a Vandermonde matrix (see below where both Λ and $\tilde{\Lambda}$ are defined) and $\Lambda^{-2r}\mathbf{1}$ lies within the span of its columns.

3 Main Results

We state results for both type of distances: using immediate neighbors and using far-away neighbors.

3.1 Distance Using Immediate Neighbors

The distance defined using immediate neighbors (cf. (3)) was analyzed in [27]. It proves that a similarity based collaborative filtering-style algorithm provides a consistent estimator for matrix completion under the additive noise model with generic function as long as the latent function is Lipschitz, not just low rank; however, it requires $\tilde{O}(n^{3/2})$ samples. We refer a reader to see [27] precise statement of the theorem.

3.2 Distance Using Immediate Neighbors

The distance defined using far-away neighbors (cf. $\texttt{dist}_1$ and $\texttt{dist}_2$) was analyzed in [7]. It establishes that the expected squared error of the estimate computed in (2) using $\texttt{dist}_1$ converges to zero with n for $p = \omega(n^{-1+\epsilon})$ for some $\epsilon > 0$, i.e. p must be polynomially larger than n^{-1}. On the other hand, the expected squared error of the estimate computed in (2) using $\texttt{dist}_2$ conveges to zero for $p = \omega(d^5 n^{-1})$.

It should be noted that computing $\texttt{dist}_1$ does not require knowledge of the spectrum of f. But, computing $\texttt{dist}_2$ requires full knowledge of the eigenvalues $(\lambda_1 \ldots \lambda_d)$ to compute the vector z. It seems plausible that the technique employed by [2] could be used to design a modified algorithm which does not need to have prior knowledge of the spectrum. They achieve this for the stochastic block model case by bootstrapping the algorithm with a method which estimates the spectrum first and then computes pairwise distances with the estimated eigenvalues.

Acknowledgements. I want to thank my co-authors Christian Borgs (Microsoft Research New England), Jennifer Chayes (Microsoft Research New England), Christina Lee (MIT / Microsoft Research New England / Cornell), Yihua Li (MIT / Microsoft) and Dogyoon Song (MIT). The bulk of the work reported in this extended abstract of talk is part of PhD Thesis of Christina Lee.

References

1 Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 670–688. IEEE, 2015.

2 Emmanuel Abbe and Colin Sandon. Recovering communities in the general stochastic block model without knowing the parameters. In *Advances in neural information processing systems*, 2015.

3 Emmanuel Abbe and Colin Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *Advances in neural information processing systems*, 2016.

4 Edo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.

5 D.J. Aldous. Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.*, 11:581–598, 1981.

6 Tim Austin. Exchangeable random arrays. *Technical Report, Notes for IAS workshop.*, 2012.

7 Christian Borgs, Jennifer Chayes, Christina E. Lee, and Devavrat Shah. Thy friend is my friend: Iterative collaborating filtering for sparse graphon estimation. In *Advances in Neural Information Processing Systems 30*, 2017.

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

9 Christian Borgs, Jennifer T Chayes, Henry Cohn, and Shirshendu Ganguly. Consistent nonparametric estimation for heavy-tailed sparse graphs. *arXiv preprint arXiv:1508.06675*, 2015.

10 Christian Borgs, Jennifer T Chayes, Henry Cohn, and Nina Holden. Sparse exchangeable graphs and their limits via graphon processes. *arXiv preprint arXiv:1601.07134*, 2016.

11 Christian Borgs, Jennifer T Chayes, Henry Cohn, and Yufei Zhao. An L^p theory of sparse graph convergence I: limits, sparse random graph models, and power law distributions. *arXiv preprint arXiv:1401.2906*, 2014.

12 Christian Borgs, Jennifer T Chayes, Henry Cohn, and Yufei Zhao. An L^p theory of sparse graph convergence II: Ld convergence, quotients, and right convergence. *arXiv preprint arXiv:1408.0744*, 2014.

13 Christian Borgs, Jennifer T Chayes, László Lovász, Vera T Sós, and Katalin Vesztergombi. Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing. *Advances in Mathematics*, 219(6):1801–1851, 2008.

14 Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2009.

15 Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

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

17 Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

18 Mark A Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

19 Persi Diaconis and Svante Janson. Graph limits and exchangeable random graphs. *Rendiconti di Matematica*, VII(28):33–61, 2008.

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

21 David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 1992.

22 D.N. Hoover. Row-column exchangeability and a generalized model for probability. In *Exchangeability in Probability and Statistics (Rome, 1981)*, pages 281–291, 1981.

23 Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

24 Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(Jul):2057–2078, 2010.

25 Olga Klopp, Alexandre B Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *To appear in Annals of Statistics*, 2015.

26 Yehuda Koren and Robert Bell. Advances in collaborative filtering. In *Recommender Systems Handbook*, pages 145–186. Springer US, 2011.

27 Christina E. Lee, Yihua Li, Devavrat Shah, and Dogyoon Song. Blind regression: Nonparametric regression for latent variable models via collaborative filtering. In *Advances in Neural Information Processing Systems 29*, pages 2155–2163, 2016.

28 Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

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

30 Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

31 Xia Ning, Christian Desrosiers, and George Karypis. *Recommender Systems Handbook*, chapter A Comprehensive Survey of Neighborhood-Based Recommendation Methods, pages 37–76. Springer US, 2015.

32 Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

33 David Steurer and Sam Hopkins. Bayesian estimation from few samples: community detection and related problems. 2017.

34 Victor Veitch and Daniel M Roy. The class of random graphs arising from exchangeable random measures. *arXiv preprint arXiv:1512.03099*, 2015.

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

36 Yuan Zhang, Elizaveta Levina, and Ji Zhu. Estimating network edge probabilities by neighborhood smoothing. *arXiv preprint arXiv:1509.08588*, 2015.