The Annals of Statistics
2015, Vol. 43, No. 1, 177–214
DOI: 10.1214/14-AOS1272

MATRIX ESTIMATION BY UNIVERSAL SINGULAR VALUE THRESHOLDING

BY SOURAV CHATTERJEE[1]

Stanford University

Consider the problem of estimating the entries of a large matrix, when the observed entries are noisy versions of a small random fraction of the original entries. This problem has received widespread attention in recent times, especially after the pioneering works of Emmanuel Candès and collaborators. This paper introduces a simple estimation procedure, called Universal Singular Value Thresholding (USVT), that works for any matrix that has "a little bit of structure." Surprisingly, this simple estimator achieves the minimax error rate up to a constant factor. The method is applied to solve problems related to low rank matrix estimation, blockmodels, distance matrix completion, latent space models, positive definite matrix completion, graphon estimation and generalized Bradley–Terry models for pairwise comparison.

1. Introduction. Consider a statistical estimation problem where the unknown parameter is not a single value or vector, but an $m \times n$ matrix M. Given an estimator $\hat{M}$, one choice for a measure of the error in estimation is the mean-squared error, defined as

$$\text{MSE}(\hat{M}) := \mathbb{E}\left[\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(\hat{m}_{ij} - m_{ij})^2\right]. \tag{1}$$

Here, $\hat{m}_{ij}$ and m_{ij} denote the (i, j)th elements of $\hat{M}$ and M, respectively. If we have a sequence of such problems, and M_n and $\hat{M}_n$ denote the parameter and the estimator in the nth problem, then by usual statistical terminology we may say that the sequence of estimators $\hat{M}_n$ is consistent if

$$\lim_{n\to\infty} \text{MSE}(\hat{M}_n) = 0.$$

The problem of estimating the entries of a large matrix from incomplete and/or noisy entries has received widespread attention ever since the proliferation of large data sets. Early work using spectral analysis was done by a number of authors in the engineering literature, for example, by Azar et al. [10] and Achlioptas and

Received September 2013; revised September 2014.
[1]Supported in part by NSF Grant DMS-10-05312.
MSC2010 subject classifications. Primary 62F12, 62G05; secondary 05C99, 60B20.
Key words and phrases. Matrix completion, matrix estimation, sochastic blockmodel, latent space model, distance matrix, covariance matrix, singular value decomposition, low rank matrices, graphons.

McSherry [1]. This was followed by a sizable body of work on spectral methods, the main pointers to which may be found in the important recent papers of Keshavan, Montanari and Oh [62, 63]. Nonspectral methods also appeared, for example, in [83].

In a different direction, statisticians have worked on matrix completion problems under a variety of modeling assumptions. Possibly the earliest works are due to Fazel [45] and Rudelson and Vershynin [87]. The emergence of compressed sensing [27, 43] has led to an explosion in activity in the field of matrix estimation and completion, beginning with the work of Candès and Recht [26]. The pioneering works of Emmanuel Candès and his collaborators [24–26, 28] introduced the technique of matrix completion by minimizing the nuclear norm under convex constraints, which is a convex optimization problem tractable by standard algorithms. This method has the advantage of *exactly*, rather than approximately, recovering the entries of the matrix when a suitable low rank assumption is satisfied, together with a certain other assumption called "incoherence."

Since the publication of [26], a number of statistics papers have attacked the matrix completion problem from various angles. Some notable examples are [38, 64, 65, 73, 76, 84]. In a different direction, a paper that seems to have a close bearing on the analytical aspects of this paper is a manuscript of Oliveira [79].

In addition to the theoretical advances, a large number of algorithms for matrix completion and estimation have emerged. The main ones are nicely summarized and compared in [73].

The purpose of this paper is to introduce a new estimator that is capable of solving a variety of matrix estimation problems that are not tractable by existing tools (at least in a mathematically provable sense). The estimator and its properties are described in this introductory section. Section 2 focuses on applications, which include applications to low rank matrices, stochastic blockmodels, distance matrices, latent space models, positive definite matrices, graphons and generalized Bradley–Terry models. All proofs are in Section 3. An expanded version (version 5) of the paper containing more theorems, examples and simulation results is available on arXiv at the URL: http://arxiv.org/pdf/1212.1247v5.pdf.

For interesting new developments that appeared after the first draft of this paper was posted on arXiv, see [35, 47, 75]. Further references and citations are given in subsequent sections.

1.1. *The setup.* Suppose that we have a $m \times n$ matrix M, where $m \leq n$ and the entries of M are bounded by 1 in absolute value. Let X be a matrix whose elements are independent random variables, and $\mathbb{E}(x_{ij}) = m_{ij}$ for all i and j [where, as usual, x_{ij} and m_{ij} denote the (i, j)th entries of X and M, resp.]. Assume that the entries of X are also bounded by 1 in absolute value, with probability one. A matrix such as X will henceforth be called a "data matrix with mean M." The matrix M will sometimes be called the "parameter matrix." Let p be a real number belonging

to the interval [0, 1]. Suppose that each entry of X is observed with probability p, and unobserved with probability $1 - p$, independently of the other entries.

The above model will henceforth be referred to as the "asymmetric model." The "symmetric model" is defined in a similar manner: Take any n and let M be a symmetric matrix of order n, whose entries are bounded by 1 in absolute value. Let X be a symmetric random matrix of order n whose elements on and above the diagonal are independent, and $\mathbb{E}(x_{ij}) = m_{ij}$ for all $1 \le i \le j \le n$. As before, assume that the entries of X are almost surely bounded by 1 in absolute value. Take any $p \in [0, 1]$ and suppose that each entry of X on and above the diagonal is observed with probability p, and unobserved with probability $1 - p$, independently of the other entries.

Similarly, one can define the "skew-symmetric model," where the difference $X - M$ is skew-symmetric, with independence on and above the diagonal as in the symmetric model. This model is used for analyzing the nonparametric Bradley–Terry model in Section 2.7.

1.2. *The USVT estimator.* In the above models, we construct an estimator $\hat{M}$ of M based on the observed entries of X along the following steps. Tentatively, I call this the Universal Singular Value Thresholding (USVT) algorithm.

1. For each i, j, let $y_{ij} = x_{ij}$ if x_{ij} is observed, and let $y_{ij} = 0$ if x_{ij} is unobserved. Let Y be the matrix whose (i, j)th entry is y_{ij}.
2. Let $Y = \sum_{i=1}^{m} s_i u_i v_i^T$ be the singular value decomposition of Y. (In the symmetric and skew-symmetric models, $m = n$.)
3. Let $\hat{p}$ be the proportion of observed values of X. In the symmetric and skew-symmetric models, let $\hat{p}$ be the proportion of observed values on and above the diagonal.
4. Choose a small positive number $\eta \in (0, 1)$ and let S be the set of "thresholded singular values," defined as

$$S := \{i : s_i \ge (2 + \eta)\sqrt{n\hat{p}}\}.$$

[Note: (a) In simulations, the method described below seemed to work even if η was taken to be exactly equal to zero; but the mathematical proof that I have requires η to be positive. In practice, one may choose η a priori to be some arbitrary small positive number, say, 0.01; but a data-dependent choice is not allowed. (b) If it is known that $\mathrm{Var}(x_{ij}) \le \sigma^2$ for all i, j, where σ is a known constant ≤ 1, then the threshold $(2 + \eta)\sqrt{n\hat{p}}$ may be improved to $(2 + \eta)\sqrt{n\hat{q}}$, where $\hat{q} := \hat{p}\sigma^2 + \hat{p}(1 - \hat{p})(1 - \sigma^2)$.]

5. Define

$$W := \frac{1}{\hat{p}} \sum_{i \in S} s_i u_i v_i^T.$$

6. Let w_{ij} denote the (i, j)th element of W. Define

$$\hat{m}_{ij} := \begin{cases} w_{ij}, & \text{if } -1 \le w_{ij} \le 1, \\ 1, & \text{if } w_{ij} > 1, \\ -1, & \text{if } w_{ij} < -1. \end{cases}$$

7. Let $\hat{M}$ be the matrix whose (i, j)th entry is $\hat{m}_{ij}$.

8. If the entries of M and X are known to belong to an interval $[a, b]$ instead of $[-1, 1]$, then subtract $(a+b)/2$ from each entry of X and divide by $(b-a)/2$, so that the entries are forced to lie in $[-1, 1]$, then apply the above procedure, and finally multiply the end-result by $(b-a)/2$ and add $(a+b)/2$ to get the estimate of M.

9. If $m > n$, then one should work with M^T and X^T instead of M and X, so that the number of rows is forced to be $\le$ the number of columns.

1.3. *Main result.* Recall that the *nuclear norm* of M, written $\|M\|_*$, is defined as the sum of the singular values of M. Recall also the definition (1) of the mean squared error of a matrix estimator. The following theorem gives an error bound for the estimator $\hat{M}$ in terms of the nuclear norm of M. This is the main result of this paper.

THEOREM 1.1. *Let $\hat{M}$ and M be as above. Let* $\mathrm{MSE}(\hat{M})$ *be defined as in* (1). *Suppose that $p \ge n^{-1+\varepsilon}$ for some $\varepsilon > 0$. Then*

$$\mathrm{MSE}(\hat{M}) \le C \min\left\{\frac{\|M\|_*}{m\sqrt{np}}, \frac{\|M\|_*^2}{mn}, 1\right\} + C(\varepsilon)e^{-cnp},$$

where C and c are positive constants that depend only on the choice of η and $C(\varepsilon)$ depends only on ε and η. The same result holds for the symmetric and skew-symmetric models, after putting $m = n$.

Moreover, if in the same setting as above, we know that $\mathrm{Var}(x_{ij}) \le \sigma^2$ *for all i, j for some known $\sigma^2 \le 1$, and the threshold is set at $(2+\eta)\sqrt{n\hat{q}}$ (see step* 4 *of the algorithm), the same result holds under the condition that $q \ge n^{-1+\varepsilon}$, where $q := p\sigma^2 + p(1-p)(1-\sigma^2)$. In this case the exponential term in the error changes to $C(\varepsilon)e^{-cnq}$ and the term $\|M\|_*/(m\sqrt{np})$ improves to $\|M\|_*\sqrt{q}/(m\sqrt{np})$.*

Incidentally, the proof shows that the condition $p > n^{-1+\varepsilon}$ may be improved to $p > n^{-1}(\log n)^{6+\varepsilon}$ (see Theorem 3.4), but I prefer to retain the present version for aesthetic reasons, especially considering that it is not a real improvement from any practical point of view.

It should be emphasized that although singular value thresholding has been used in a number of papers on matrix completion and estimation (see, e.g., [1, 10, 24, 62, 63] and references therein), the above algorithm has the unique feature that the threshold is universal. In the literature, it is usually assumed that the

matrix M has a rank r that is known, and uses the value of r while thresholding. The USVT algorithm manages to cut off the singular values at the "correct" level, depending on the structure of the unknown parameter matrix. The adaptiveness of the USVT threshold is somewhat similar in spirit to that of the *SureShrink* algorithm of Donoho and Johnstone [44]. SureShrink performs function estimation by estimating Fourier coefficients in some suitable basis, and then thresholds the coefficients at a threshold that automatically adapts to the smoothness of the unknown function. Analogously, the USVT algorithm computes the eigenvalues of the observed matrix, and then thresholds the eigenvalues at a universal threshold that is automatically adaptive in nature, because it picks out as much "structure" as is available and throws out all the randomness. This point will become more clear from the examples discussed in Section 2.

One limitation of USVT is the requirement that the entries should lie in a bounded interval. One may relax this requirement by assuming, for example, that the errors $x_{ij} - m_{ij}$ are distributed as normal random variables with mean zero and variance σ^2. If σ^2 is known, then I believe that one can modify the USVT algorithm by thresholding at $(2+\eta)\sigma\sqrt{n}$ and obtain the same theorems. The rationale behind this belief is as follows: if A is a large symmetric random matrix whose entries on and above the diagonal are independent, have zero mean, and are bounded by 1 in absolute value, then the spectral norm of A is less than $2+\eta$ with high probability. This is the key ingredient in the proof of Theorem 1.1. But such a result continues to be true, after replacing $2+\eta$ with $(2+\eta)\sigma$, if the entries are normally distributed with mean zero and variance bounded by σ^2. Therefore, it is conceivable that the proof of Theorem 1.1 may be modified to accommodate this altered situation. However, if σ^2 is unknown, I do not know how to proceed. In reality, σ^2 will not be known; this is why I have not worked with the normality assumption. Also, the situation of normally distributed entries but with a large proportion missing, seems to be trickier.

1.4. *Minimax lower bound.* It is not difficult to prove that for an $m \times n$ matrix M with entries bounded by 1 in absolute value, where $m \le n$, the nuclear norm is bounded by $m\sqrt{n}$. Given a number $\delta \in [0, m\sqrt{n}]$, one may take an arbitrary estimator $\tilde{M}$ and try to find the M among all M satisfying $\|M\|_* \le \delta$ for which $\mathrm{MSE}(\tilde{M})$ is maximum. Recall that an estimator that minimizes this maximum error is classically known as a minimax estimator. The following theorem shows that our estimator $\hat{M}$ is minimax up to a constant multiplicative factor and an exponentially small additive discrepancy.

THEOREM 1.2. *Consider the general matrix estimation problem outlined in the beginning of this section. Given any estimator $\tilde{M}$ and any $\delta \in [0, m\sqrt{n}]$, there exists M satisfying $\|M\|_* \le \delta$ and a data matrix X with mean M, such that for this*

M and X, the estimator $\tilde{M}$ satisfies

$$\mathrm{MSE}(\tilde{M}) \geq c \min\left\{\frac{\delta}{m\sqrt{np}}, \frac{\delta^2}{mn}, 1\right\},$$

where c is a positive universal constant. Moreover, if $p < 1/2$ then X and M may be chosen such that $X = M$. The same lower bound holds in the symmetric case and in the skew-symmetric case.

It is worth noting that the exponentially small discrepancy is necessary. For example, if $\delta = 0$, then the minimax error is obviously zero. However, there is still an exponentially small chance that $\hat{M}$ may be nonzero. It is also worth noting that if δ is not too small (e.g., if $\delta > \sqrt{m/p}$), then the exponential discrepancy does not matter, and the combination of Theorems 1.1 and 1.2 gives the correct minimax error up to a universal multiplicative constant.

An examination of the proof of Theorem 1.1 indicates that with slight modifications, one may obtain bounds on tail probabilities instead of an upper bound on the mean squared error. I have retained the present version for aesthetic reasons.

Incidentally, two notable recent papers, namely, Koltchinskii et al. [65] and Davenport et al. [38], have suggested matrix estimation by nuclear norm penalization and proved minimax optimality results that match up to logarithmic factors. Davenport et al. [38], Theorem 3, show (in the notation of our Theorem 1.2) that if the entries of X belong to $\{-1, 1\}$ and if $\delta \geq 4\sqrt{mn}$, then the minimax error is bounded below by a universal constant times $\min\{\delta/(m\sqrt{np}), 1\}$, provided that this quantity is bigger than $\delta^2/(m^2 n)$. This is almost the same as the conclusion of Theorem 1.2, except that it does not cover the case of δ smaller than $4\sqrt{mn}$. Section 3.1 of [38] gives a matrix estimation algorithm based on nuclear norm penalization that achieves this minimax rate up to a logarithmic factor. However, the implementation of this algorithm requires that the user has a reasonable estimate for the nuclear norm of the unknown matrix M, since that is used as the regularization parameter. USVT has no such requirement. Another advantage that USVT has over the algorithm of [38] is that it may be easier to implement, especially for very large matrices, because it does not involve convex optimization.

The estimator of Koltchinskii et al. [65] is also based on nuclear norm penalization: translating to our notation, they estimate M by minimizing $\|X - \hat{M}\|_F^2 + \lambda\|\hat{M}\|_*$ over all $\hat{M}$, where $\|\cdot\|_F$ is Frobenius norm, $\|\cdot\|_*$ is nuclear norm, and λ is a regularization parameter. It is shown in [65] that this problem is actually equivalent to soft singular value thresholding, where the threshold depends on the parameter λ. A conservative choice of λ (albeit with an unspecified constant) and a minimax lower bound that matches the upper bound up to a logarithmic factor are given in [65]. The minimax bound is computed over the set of all matrices with rank less than a given number and, therefore, is not directly comparable to the minimax bound in Theorem 1.2. With a suitable choice of λ—but again with

unspecified constants—the upper bound in [65], Theorem 3, becomes (up to a logarithmic factor) essentially equal to $\|M\|_*/(m\sqrt{np})$. Note that this is the same as the main term in Theorem 1.1. However, if we additionally know that M has low rank, then the upper bound in [65], Theorem 3, becomes substantially better (see Section 2.1).

1.5. *Practical issues and warnings.* I do not consider the USVT algorithm as presented above to be in a form that may implemented "as is." This is mainly for the following reasons:

(a) USVT is minimax optimal only up to a constant factor. In fact, it is very likely that one may be able to build a better estimator by taking into account the ratio m/n, and getting improved bounds when this ratio is small. Although Theorem 1.2 shows that the improvement will be limited to multiplication by a constant factor, such an improvement may be important for practical purposes. The recent paper [47] has explored the issue of attaining the minimax error all the way up to the correct constant.

(b) The number η is a "tuning parameter" for this algorithm, that may be chosen by the implementer. The theorem is valid with any choice of η in the interval $(0, 1)$, although the constants in the error bounds blow up as η tends to zero. I have noticed in simulations that taking $\eta = 0$ works quite well, but I do not know how to prove that. Choosing η to be a small but fixed positive number such as 0.01 is consistent with the requirements of Theorem 1.1 and seemed to give good results in simulations. Choosing η in a data-dependent manner is, however, not covered by Theorem 1.1.

(c) Note that in practice, any data matrix may be centered and scaled so that the entries are forced to lie in the interval $[-1, 1]$. However, if the centering and scaling are done in a data-dependent manner, then the assertion of Theorem 1.1 is no longer guaranteed to be true.

1.6. *An impossibility theorem for error estimates.* Theorem 1.1 gives an upper bound on the mean squared error of $\hat{M}$. The estimate involves the nuclear norm of parameter matrix M. A natural question is: Is it possible to estimate the true MSE of $\hat{M}$ from the data?

A straightforward approach is to use parametric bootstrap. Having estimated M using $\hat{M}$, one may choose a large number K, generate K copies of the data using $\hat{M}$ as the parameter matrix, compute the estimates $\hat{M}^{(i)}$, $i = 1, \dots, K$ for the K simulations, and estimate the MSE of $\hat{M}$ using the bootstrap estimator

$$\widehat{\mathrm{MSE}}_{\mathrm{BS}}(\hat{M}) = \frac{1}{K}\sum_{i=1}^{K} \frac{\|\hat{M}^{(i)} - \hat{M}\|_F^2}{mn}.$$

For the validity of the bootstrap estimate of the MSE, it is essential that the original $\hat{M}$ is an accurate estimate of M. In other words, we need to know a priori that

$\mathrm{MSE}(\hat{M})$ is small to be able to claim that the bootstrap estimator of $\mathrm{MSE}(\hat{M})$ is accurate. Theorem 1.1 implies that if we know that $\|M\|_*$ is small enough from assumptions, this is true.

But is it possible to somehow determine whether $\mathrm{MSE}(\hat{M})$ is small or not from the data, if we do not make any assumption about M to start with? We will now show that it is impossible to do so, not only for the estimator $\hat{M}$ but for any "nontrivial" estimator $\tilde{M}$.

The definition of a nontrivial estimator is as follows. Given a parameter matrix M and a data matrix X satisfying the conditions of Section 1, the trivial estimator of M based on X is simply X itself. We will denote the trivial estimator as $\hat{M}^{\mathrm{Trv}}$. Now suppose we are given some estimator $\tilde{M}$. We will say that the estimator $\tilde{M}$ is nontrivial if there exists a sequence of parameter matrices M_n and data matrices X_n such that

$$\mathrm{MSE}(\hat{M}_n^{\mathrm{Trv}}) \not\to 0 \qquad \text{as } n \to \infty,$$

but $\lim_{n\to\infty} \mathrm{MSE}(\tilde{M}_n) = 0$. In other words, $\tilde{M}_n$ solves a nontrivial estimation problem. The USVT estimator is clearly nontrivial, as demonstrated by the examples from Section 2.

Suppose that we have a nontrivial estimator $\tilde{M}$ and a procedure P that gives an estimate $\widehat{\mathrm{MSE}}_{\mathrm{P}}(\tilde{M})$ of the MSE of $\tilde{M}$. The MSE estimate is computed using only the data. The procedure will be called "good" if the following two conditions hold:

(a) Whenever M_n is a sequence of parameter matrices and X_n is a sequence of data matrices such that $\mathrm{MSE}(\tilde{M}_n)$ tends to zero, the estimate $\widehat{\mathrm{MSE}}_{\mathrm{P}}(\tilde{M}_n)$ also tends to zero in probability.

(b) Whenever M_n is a sequence of parameter matrices and X_n is a sequence of data matrices such that $\mathrm{MSE}(\tilde{M}_n)$ does not tend to zero, $\widehat{\mathrm{MSE}}_{\mathrm{P}}(\tilde{M}_n)$ also does not tend to zero in probability.

In the above setting, the following theorem establishes the impossibility of the existence of a good estimator for the MSE.

THEOREM 1.3. *There cannot exist a good procedure for estimating the mean squared error of a nontrivial estimator.*

2. Applications. Throughout this section, m, n, M, X, p and $\hat{M}$ will be as in Section 1. Just to remind the reader, M is an $m \times n$ matrix where $1 \le m \le n$. The entries of M are assumed to be bounded by 1 in absolute value. The matrix X is a random matrix whose entries are independent, and the (i, j)th element x_{ij} has expected value equal to m_{ij}, the (i, j)th entry of M. Moreover, they satisfy $|x_{ij}| \le 1$ with probability one. In particular, X may be exactly equal to M, with no randomness. Each entry of X is observed with probability p and unobserved with

probability $1-p$, independently of other entries. Occasionally, we will assume the symmetric model, where $m=n$, and the matrices M and X are symmetric. In the special case of the Bradley–Terry model in Section 2.7, we will assume the skew-symmetric model, where $X-M$ is skew-symmetric.

We will now work out various specific cases where Theorem 1.1 gives useful results.

2.1. *Low rank matrices*. Estimating low rank matrices has been the focus of the vast majority of prior work [1, 10, 24–26, 28, 45, 62–65, 73, 76, 83, 87]. Theorem 1.1 works for low rank matrices. The following theorem, which is a simple corollary of Theorem 1.1, shows that $\hat{M}$ is a good estimate whenever the rank of M is small compared to mp (after assuming, as in Theorem 1.1, that $p \geq n^{-1+\varepsilon}$).

THEOREM 2.1. *Suppose that M has rank r. Suppose that $p \geq n^{-1+\varepsilon}$ for some $\varepsilon > 0$. Then*

$$\mathrm{MSE}(\hat{M}) \leq C \min\left\{\sqrt{\frac{r}{mp}}, 1\right\} + C(\varepsilon)e^{-cnp},$$

where C and c depend only on η and $C(\varepsilon)$ depends only on ε and η. Moreover, the same result holds when M and X are symmetric.

The term $1/np$ in the error bound is necessary to take care of the case $r=0$. Even if M is identically zero, the estimator $\hat{M}$ will incur some error due to the (possible) randomness in X.

Let us now inspect how the condition $r \ll mp$ compares with available bounds. In a notable sequence of papers, Keshavan, Montanari and Oh [62, 63] obtain the same condition but only if m and n are comparable and the rank is known. Theorem 2.1, on the other hand, works even for "very rectangular" matrices where $m \ll n$ and the rank is unknown.

Candès and Tao [28] obtain the condition $r \ll mp$ with an extra polylogarithmic term in the error. Moreover, they too require that m and n be comparable, and additionally they need the so-called "incoherence condition". However, as noted before, the incoherence condition allows exact recovery, while our approach only gives approximate recovery.

The recent important work of Davenport et al. [38] gives an estimator with an error bound that is almost the same as that given by Theorem 2.1, but with a complicated optimization algorithm.

Theorem 2.1, however, is probably not an optimal result. It has been shown by Koltchinskii et al. [65], Theorems 3 and 5, that the true minimax error rate for a closely related problem is actually r/mp, up to a logarithmic factor.

The following theorem shows that the condition $r \ll mp$ is necessary for estimating M.

THEOREM 2.2. *Given any estimator $\tilde{M}$, there exists an $m \times n$ matrix M of rank r with entries bounded between -1 and 1, such that when the data is sampled from M,*

$$\mathrm{MSE}(\tilde{M}) \geq C(1-p)^{[m/r]},$$

where C is a positive universal constant and $[m/r]$ is the integer part of m/r.

2.2. *The stochastic blockmodel.* Consider an undirected graph on n vertices. A stochastic blockmodel assumes that the vertices $1, \ldots, n$ are partitioned into k blocks, and the probability that vertex i is connected to vertex j by an edge depends only on the blocks to which i and j belong. As usual, edges are independent of each other. Let M be the matrix whose (i, j)th element is the probability of an edge existing between vertices i and j. The matrix X here is the adjacency matrix of the observed graph. Here, all elements of X are observed, so $p = 1$.

This is commonly known as the stochastic blockmodel. It was introduced by Holland, Laskey and Leinhardt [56] as a simple stochastic model of social networks. It has become one of the most successful and widely used models for community structure in networks, especially after the advent of large data sets.

Early analysis of the stochastic blockmodel was carried out by Snijders and Nowicki [77, 91], who provided consistent parameter estimates when there are exactly two blocks. This was extended to a finite but fixed number of blocks of equal size by Condon and Karp [37]. Bickel and Chen [16] were the first to give consistent estimates for finite number of blocks of unequal size. It was observed by Leskovec et al. [67] that in real data, the number of blocks often seem to grow with the number of nodes. This situation was rigorously analyzed for the first time in Rohe et al. [85], and was followed up shortly thereafter by [17, 34, 36, 74] with more advanced results.

However, all in all, I am not aware of any estimator for the stochastic blockmodel that works whenever the number of blocks is small compared to the number of nodes. The best result till date is in the very recent manuscript of Rohe et al. [86], who prove that a penalized likelihood estimator works whenever k is comparable to n "up to log factors." The following theorem says that the USVT estimator $\hat{M}$ gives a complete solution to the estimation problem in the stochastic blockmodel if $k \ll n$, with no further conditions required. (The method will not work very well for sparse graphs, however; for recent advances on estimation in sparse graphs, see [7].)

THEOREM 2.3. *For a stochastic blockmodel with k blocks,*

$$\mathrm{MSE}(\hat{M}) \leq C\sqrt{\frac{k}{n}},$$

where C is a constant that depends only on our choice of η.

Note that estimating the stochastic blockmodel is a special case of low rank matrix estimation with noise. It is not difficult to prove that the estimation problem is impossible when k is of the same order as n. We will not bother to write down a formal proof.

2.3. *Distance matrices.* Suppose that K is a compact metric space with metric d. Let $x_1, \dots, x_n$ be arbitrary points from K, and let M be the $n \times n$ matrix whose (i, j)th entry is $d(x_i, x_j)$. Such matrices are called "distance matrices". Since K is a compact metric space, the diameter of K with respect to the metric d must be finite. Scaling d by a constant factor, we may assume without loss of generality that the diameter is bounded by 1, so that the entries of M are bounded by 1 as required by Theorem 1.1.

Completing a distance matrix with missing entries has been a popular problem in the engineering and social sciences for a long time; see, for example, [6, 11, 18, 89, 90, 92]. It has become particularly relevant in engineering problems related to sensor networks. It is also an important issue in multidimensional scaling [19]. For some recent theoretical advances, see [60, 78].

In general, distance matrices need not be of low rank. Therefore, much of the literature on matrix estimation and completion does not apply to distance matrices. Surprisingly, Theorem 1.1 gives a complete solution of the distance matrix completion and estimation problem.

THEOREM 2.4. *Suppose that $p \geq n^{-1+\varepsilon}$ for some $\varepsilon > 0$. If M is a distance matrix as above, then*

$$\mathrm{MSE}(\hat{M}) \leq \frac{C(K, d, n)}{\sqrt{p}} + C(\varepsilon) e^{-cnp},$$

where c depends only on η, $C(\varepsilon)$ depends only on ε and η, and $C(K, d, n)$ is a number depending only on K, d, n and η such that

$$\lim_{n \to \infty} C(K, d, n) = 0.$$

The above theorem is not wholly satisfactory, since it does not indicate how fast p can go to zero as $n \to \infty$ so that $\hat{M}$ is still consistent. To understand that, we need to know more about the structure of the space K. The following theorem gives a quantitative estimate.

THEOREM 2.5. *Suppose that for each $\delta > 0$, $N(\delta)$ is a number such that K may be covered by $N(\delta)$ open d-balls of radius δ. Then*

$$\mathrm{MSE}(\hat{M}) \leq C \inf_{\delta > 0} \min\left\{ \frac{\delta + \sqrt{N(\delta/4)/n}}{\sqrt{p}}, 1 \right\} + C(\varepsilon) e^{-cnp},$$

where C and c depend only on η and $C(\varepsilon)$ depends only on ε and η.

To see how Theorem 2.5 may be used, suppose that K is a compact subset of the real line and d is the usual distance on $\mathbb{R}$, scaled by a factor to ensure that the diameter of K is ≤ 1. Then $N(\delta)$ increases like $1/\delta$ as $\delta \to 0$. Consequently, given n, the optimal choice of δ is of the order $n^{-1/3}$, which gives the bound

$$\mathrm{MSE}(\hat{M}) \leq \frac{Cn^{-1/3}}{\sqrt{p}}.$$

(Note that the exponential term need not appear because the main term is bounded below by a positive constant if $p < n^{-2/3}$.) Thus, $\hat{M}$ is a consistent estimate as long as p goes to zero slower than $n^{-2/3}$ as $n \to \infty$.

2.4. *Latent space models.* Suppose that $\beta_1, \ldots, \beta_n$ are vectors belonging to some bounded closed set $K \subseteq \mathbb{R}^k$, where k is some arbitrary but fixed dimension. Let $f : K \to [-1, 1]$ be a continuous function. Let M be the $n \times n$ matrix whose (i, j)th element is $f(\beta_i, \beta_j)$. Then our data matrix X has the form

$$x_{ij} = f(\beta_i, \beta_j) + \varepsilon_{ij},$$

where ε_{ij} are independent errors with zero mean, satisfying the restriction that $|x_{ij}| \leq 1$ almost surely. For example, X may be the adjacency matrix of a random graph where the probability of an edge existing between vertices i and j is $f(\beta_i, \beta_j)$. This is one context where latent space models are widely used, starting with the work of Hoff, Raftery and Handcock [55]. A large body of work applying the latent space approach to real data has grown in the last decade. On the theoretical side, it was observed in [16, 17] that the latent space model arises naturally from an exchangeability assumption due to the Aldous–Hoover theorem [5, 57]. Note that distance matrices and stochastic blockmodels are both special cases of latent space models.

There have been various attempts to estimate parameters in the latent space models (e.g., [4, 53, 55]). Almost all of these approaches rely on heuristic arguments and justification through simulations. The problem is that in addition to the vectors $\beta_1, \ldots, \beta_n$, the function f itself is an unknown parameter. If either β_i's are known, or f is known, the estimation problem is tractable. For example, when $f(x, y)$ is of the form $e^{x+y}/(1 + e^{x+y})$, the problem was solved in [31]. However, when both f and β_i's are unknown, the problem becomes seemingly intractable. In particular, there is an identifiability issue because $f(x, y)$ may be replaced by $h(x, y) := f(g(x), g(y))$ and β_i by $g^{-1}(\beta_i)$ for any invertible function g without altering the model.

In view of the above discussion, it is a rather surprising consequence of Theorem 1.1 that it is possible to estimate the numbers $f(\beta_i, \beta_j)$, $i, j = 1, \ldots, n$ from a single realization of the data matrix, under no additional assumptions than the stated ones.

THEOREM 2.6. *Suppose that* $p \geq n^{-1+\varepsilon}$. *If* M *is as above, then*

$$\mathrm{MSE}(\hat{M}) \leq \frac{C(K, k, f, n)}{\sqrt{p}} + C(\varepsilon)e^{-cnp},$$

where c *depends only on* η, $C(\varepsilon)$ *depends only on* ε *and* η, *and* $C(K, k, f, n)$ *depends only on* K, k, f, n *and* η *such that*

$$\lim_{n\to\infty} C(K, k, f, n) = 0.$$

The problem with Theorem 2.6, just like Theorem 2.4 in Section 2.3, is that it does not give an explicit error bound, which makes it impossible to determine how fast p can go to zero with n so that consistency holds. Again, this is easy to fix by assuming smoothness properties of f and applying Lemma 3.6. As a particular example, suppose that f is Lipschitz with Lipschitz constant L, in the sense that

$$\left|f(x, y) - f\left(x', y'\right)\right| \leq L\left\|x - x'\right\| + L\left\|y - y'\right\|$$

for all $x, y, x', y' \in K$.

THEOREM 2.7. *In the above setting,*

$$\mathrm{MSE}(\hat{M}) \leq C(K, k, L)\frac{n^{-1/(k+2)}}{\sqrt{p}},$$

where $C(K, k, L)$ *is a constant depending only on* K, k, L *and* η.

2.5. *Positive definite matrices.* Assume that $m = n$ and M is positive semi-definite. (In the statistical context, this is the same as saying that M is a covariance matrix. When the diagonal entries are all 1, M is a correlation matrix.)

Completing positive definite matrices with missing entries has received a lot of attention in the linear algebra literature [15, 51, 61], although most of the techniques are applicable only for relatively small matrices or when a sizable fraction of the entries are observed. In the engineering sciences, estimation of covariance matrices from a small subset of observed entries arises in the field of remote sensing (see [25, 26, 28] for brief discussions).

The statistical matrix completion literature cited in Section 1 applies only to low rank positive definite matrices. It is therefore quite a surprise that the completion problem may be solved for *any* positive definite matrix whenever we get to observe a large number of entries from each row.

THEOREM 2.8. *Suppose that* $m = n$ *and* M *is positive semi-definite. Suppose that* $p \geq n^{-1+\varepsilon}$. *Then*

$$\mathrm{MSE}(\hat{M}) \leq \frac{C}{\sqrt{np}} + C(\varepsilon)e^{-cnp},$$

where C *and* c *depend only on* η *and* $C(\varepsilon)$ *depends only on* ε *and* η.

What if p is of order $1/n$ or less? The following theorem shows that it is impossible to estimate M in this situation.

THEOREM 2.9. *Given any estimator $\tilde{M}$, there exists a correlation matrix M such that when the data is sampled from M,*

$$\mathrm{MSE}(\tilde{M}) \geq C(1-p)^n,$$

where C is a positive universal constant.

2.6. *Graphon estimation.* A graphon is a measurable function f from $[0,1]^2$ into $[0,1]$ that satisfies $f(x,y) \equiv f(y,x)$. The term "graphon" was coined by Lovász and coauthors in the growing literature on limits of dense graphs [20–22, 68, 69]. Such functions also arise in the related study of weakly exchangeable random arrays [5, 9, 42, 57]. They have also appeared recently in large deviations [32, 33, 70] and mathematical statistics [30, 81].

In the graph limits literature, graphons arise as limits of graphs with increasing number of nodes. Conversely, graphons are often used to generate random graphs in a natural way. Take any n and let $U_1, \ldots, U_n$ be i.i.d. Uniform$[0,1]$ random variables. Construct a random undirected graph on n vertices by putting an edge between vertices i and j with probability $f(U_i, U_j)$, doing this independently for all $1 \leq i < j \leq n$. This procedure is sometimes called "sampling from a graphon" (see [21], Section 4.4).

The statistical question is the following: Suppose that we have a random graph on n vertices that is sampled from a graphon. Is it possible to estimate the graphon from a single realization of the graph? More precisely, is it possible to accurately estimate the numbers $f(U_i, U_j)$, $1 \leq i < j \leq n$, from a single realization of the random graph? The question is similar to the one investigated in Section 2.3, but the difference is that here we are not allowed to assume any regularity on f except measurability.

Taking things back to our usual setting, let M be the matrix whose (i,j)th element is $f(U_i, U_j)$. Note that unlike our previous examples, M is now random. So the definition of MSE should be modified to take expectation over M as well.

THEOREM 2.10. *In the above setting,*

$$\mathrm{MSE}(\hat{M}) \leq C(f,n),$$

where $C(f,n)$ is a constant depending only on f, n and η, such that

$$\lim_{n\to\infty} C(f,n) = 0.$$

Incidentally, after the first version of this paper was put up on arXiv, several papers (e.g., [95, 96]) on graphon estimation, advocating a number of different techniques and demonstrating applications in the statistical study of networks, have appeared in the literature.

2.7. *Nonparametric Bradley–Terry model.* Suppose there are n teams playing against each other in a tournament. Every team plays against every other team at least once (often, exactly once). Suppose that p_{ij} is the probability that team i wins against team j in a match between i and j. Then $p_{ji} = 1 - p_{ij}$.

The Bradley–Terry model [23], originally proposed by Zermelo [97], assumes that p_{ij} is of the form $a_i/(a_i + a_j)$ for some unknown nonnegative numbers $a_1, \dots, a_n$. It is known how to estimate the parameters $a_1, \dots, a_n$ if we assume that the outcomes of all games are independent—which, in this case, is a reasonable assumption.

The Bradley–Terry model has found great success among practitioners. For an old survey of the literature on the model dating back to 1976, see [40]. Numerous extensions and applications have been proposed, for example, [3, 54, 58, 71, 72, 80, 82]. The monographs of David [39] and Diaconis [41], Chapter 9, explain the statistical foundations of these models. More recently, several authors have proposed to perform Bayesian inference for (generalized) Bradley–Terry models [2, 29, 48–50, 52].

For the basic Bradley–Terry model, it is possible to find the maximum likelihood estimate of the a_i's using a simple iterative procedure [59, 66, 97]. The maximum likelihood estimate was shown to be jointly consistent for all n parameters by Simons and Yao [88].

We now generalize the Bradley–Terry model as follows. Suppose, as before, that p_{ij} is the probability that team i beats team j. Suppose that the teams have a particular ordering in terms of strength that is unknown to the observer. Assume that *if team i is stronger than team j, then $p_{ik} \geq p_{jk}$ for all $k \neq i, j$.* Do not assume anything else about the p_{ij}'s; in particular, do not assume any formula for the p_{ij}'s in terms of hidden parameters. This is what we may call a "nonparametric Bradley–Terry model." Note that the usual Bradley–Terry model is a special case of the nonparametric version.

In the nonparametric Bradley–Terry model, is it possible to estimate all the p_{ij}'s from a tournament where every team plays against every other exactly once? Is it possible to estimate the p_{ij}'s if only a randomly chosen fraction of the games are played? How small can this fraction be, so that accurate estimation is still possible? The following theorem provides some answers.

THEOREM 2.11. *Consider the nonparametric Bradley–Terry model defined above. Let M be the matrix whose (i, j)th entry is p_{ij} if $i \neq j$ and 0 if $i = j$. Let X be the data matrix whose (i, j)th entry is 1 if team i won over team j, 0 if team j won over team i and recorded as missing if team i did not play versus team j. If team i has played against team j multiple times, let the (i, j)th entry of X be the proportion of times that i won over j. (Draws are not allowed.) Let all diagonal entries of X be zero. Given $p \in [0, 1]$, suppose that for each i and j, the game between i and j takes place with probability p and does not take place with*

probability $1-p$, *independent of other games. Let* $\hat{M}$ *be the estimate of* M *based on the data matrix* X. *Then*

$$\mathrm{MSE}(\hat{M}) \le \frac{Cn^{-1/4}}{\sqrt{p}},$$

where C *depends only on our choice of* η. *In particular, the estimation problem is solvable whenever* $p \gg n^{-1/2}$.

A natural question is whether the threshold $p \gg n^{-1/2}$ is sharp. I do not know the answer to this question.

3. Proofs.

3.1. *Proof of Theorem* 1.1 (*Main result*). We need to recall some background material before embarking on the proof of Theorem 1.1.

Matrix norms. Let $A = (a_{ij})_{1\le i\le m, 1\le j\le n}$ be an $m\times n$ real matrix with singular values $\sigma_1,\dots,\sigma_k$, where $k = \min\{m,n\}$. The following matrix norms are widely used in this proof.

The nuclear norm or the trace norm of A is defined as

$$\|A\|_* := \sum_{i=1}^{k} \sigma_i.$$

The Frobenius norm, also called the Hilbert–Schmidt norm, is defined as

$$\|A\|_F := \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2} = \left(\mathrm{Tr}(A^TA)\right)^{1/2} = \left(\sum_{i=1}^{k}\sigma_i^2\right)^{1/2}.$$

By the Cauchy–Schwarz inequality,

$$\|A\|_* \le \sqrt{\mathrm{rank}(A)}\,\|A\|_F. \tag{2}$$

The sup-norm is defined as

$$\|A\|_\infty := \max_{i,j} |a_{ij}|.$$

The spectral norm or the operator norm of A is defined as

$$\|A\| := \max_{1\le i\le k} |\sigma_i|.$$

The spectral norm may be alternatively expressed as

$$\|A\| = \max_{x\in\mathbb{S}^{m-1}, y\in\mathbb{S}^{n-1}} x^TAy, \tag{3}$$

where $\mathbb{S}^{m-1}$ and $\mathbb{S}^{n-1}$ are the Euclidean unit spheres in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. The above representation implies that the spectral norm satisfies the triangle inequality. Consequently, for any two $m \times n$ matrices A and B,

$$\big|\|A\| - \|B\|\big| \le \|A - B\| \le \|A - B\|_F.$$

In particular, the spectral norm is a Lipschitz function of the matrix entries (with Lipschitz constant 1), if the entries are collectively considered as a vector of length mn.

The triangle inequality for the spectral norm also implies that the map $A \mapsto \|A\|$ is convex. Indeed, for any $0 \le t \le 1$,

$$\big\|tA + (1-t)B\big\| \le t\|A\| + (1-t)\|B\|.$$

For more on matrix norms, see [14].

Perturbation of singular values. The following perturbative result from matrix analysis is used several times in this manuscript. Let A and B be two $m \times n$ matrices. Let $k = \min\{m, n\}$. Let $\sigma_1, \dots, \sigma_k$ be the singular values of A in decreasing order and repeated by multiplicities, and let $\tau_1, \dots, \tau_k$ be the singular values of B in decreasing order and repeated by multiplicities. Let $\delta_1, \dots, \delta_k$ be the singular values of $A - B$, in any order but still repeated by multiplicities.

THEOREM 3.1. *For any* $1 \le p < \infty$,

$$\sum_{i=1}^{k} |\sigma_i - \tau_i|^p \le \sum_{i=1}^{k} |\delta_i|^p$$

and

$$\max_{1 \le i \le k} |\sigma_i - \tau_i| \le \max_{1 \le i \le k} |\delta_i|.$$

The above result follows, for example, from a combination of Theorem III.4.4 and Exercise II.1.15 in [14]. It may also be derived as a consequence of Wielandt's minimax principle [14], Section III.3, or Lidskii's theorem [14], Exercise III.4.3. The case $p = 2$ is sometimes called the Hoffman–Wielandt theorem [8], Lemma 2.1.19 and Remark 2.1.20, and the inequality involving the maximum is sometimes called Weyl's perturbation theorem [14], Corollary III.2.6.

Bernstein's inequality. The following inequality is known as "Bernstein's inequality."

THEOREM 3.2. *Suppose that* $X_1, \dots, X_n$ *are independent random variables with zero mean, and* M *is a constant such that* $|X_i| \le M$ *with probability one for each* i. *Let* $S := \sum_{i=1}^{n} X_i$ *and* $v := \mathrm{Var}(S)$. *Then for any* $t \ge 0$,

$$\mathbb{P}\big(|S| \ge t\big) \le 2\exp\bigg(-\frac{3t^2}{6v + 2Mt}\bigg).$$

This inequality was proved by Bernstein [13]. For a discussion of Bernstein's inequality and improvements, see Bennett [12].

Talagrand's concentration inequality. Recall that a median m of a random variable Y is a real number such that $\mathbb{P}(Y \le m) \ge 1/2$ and $\mathbb{P}(Y \ge m) \ge 1/2$. The median may not be unique.

The following concentration inequality is one of the several striking inequalities that are collectively known as "Talagrand's concentration inequalities."

THEOREM 3.3. *Suppose that $f : [-1, 1]^n \to \mathbb{R}$ is a convex Lipschitz function with Lipschitz constant L. Let $X_1, \ldots, X_n$ be independent random variables taking value in $[-1, 1]$. Let $Y := f(X_1, \ldots, X_n)$ and let m be a median of Y. Then for any $t \ge 0$,*

$$\mathbb{P}\big(|Y - m| \ge t\big) \le 4e^{-t^2/16L^2}.$$

For a proof of Theorem 3.3, see [93], Theorem 6.6.

It is easy to modify Theorem 3.3 to have concentration around the mean instead of the median. Just observe that by Theorem 3.3, $\mathbb{E}(Y - m)^2 \le 64L^2$. Since $\mathbb{E}(Y - m)^2 \ge \mathrm{Var}(Y)$, this shows that $\mathrm{Var}(Y) \le 64L^2$. Thus, by Chebychev's inequality,

$$\mathbb{P}\big(|Y - \mathbb{E}(Y)| \ge 16L\big) \le \tfrac{1}{4}.$$

By the definition of a median, this shows that $\mathbb{E}(Y) - 16L \le m \le \mathbb{E}(Y) + 16L$. Together with Theorem 3.3, this implies that for any $t \ge 0$,

$$\mathbb{P}\big(|Y - \mathbb{E}(Y)| \ge 16L + t\big) \le 4e^{-t^2/2L^2}. \tag{4}$$

The above inequality has a number of uses in the proof of Theorem 1.1.

Spectral norms of random matrices. The following bound on spectral norms of random matrices is a crucial ingredient for this paper. The proof follows from a combinatorial argument of Vu [94] (which is itself a refinement of a classical argument of Füredi and Komlós [46]), together with Talagrand's inequality (4).

THEOREM 3.4. *Take any two numbers m and n such that $1 \le m \le n$. Suppose that $A = (a_{ij})_{1 \le i \le m, 1 \le j \le n}$ is a matrix whose entries are independent random variables that satisfy, for some $\sigma^2 \in [0, 1]$,*

$$\mathbb{E}(a_{ij}) = 0, \qquad \mathbb{E}\big(a_{ij}^2\big) \le \sigma^2 \quad \text{and} \quad |a_{ij}| \le 1 \qquad a.s.$$

Suppose that $\sigma^2 \ge n^{-1+\varepsilon}$ for some $\varepsilon > 0$. Then for any $\eta \in (0, 1)$,

$$\mathbb{P}\big(\|A\| \ge (2 + \eta)\sigma\sqrt{n}\big) \le C_1(\varepsilon)e^{-C_2\sigma^2 n},$$

where $C_1(\varepsilon)$ depends only on ε and η and C_2 depends only on η. The same result is true when $m = n$ and A is symmetric or skew-symmetric, with independent entries on and above the diagonal, all other assumptions remaining the same. Lastly, all results remain true if the assumption $\sigma^2 \geq n^{-1+\varepsilon}$ is changed to $\sigma^2 \geq n^{-1}(\log n)^{6+\varepsilon}$.

PROOF. First assume that $m = n$ and A is symmetric. Note that for any even number k,

$$\mathbb{E}\|A\|^k \leq \mathbb{E}\big(\mathrm{Tr}\big(A^k\big)\big) = \sum_{1\leq i_1,\ldots,i_k\leq n} \mathbb{E}(a_{i_1i_2}a_{i_2i_3}\cdots a_{i_{k-1}i_k}a_{i_ki_1}). \tag{5}$$

Consider $i_1, i_2, \ldots, i_{k-1}, i_k, i_1$ as a closed tour of a subset of the vertices of the complete graph on n vertices (with self-edges included). From the given assumptions about the a_{ij}'s, it follows that the term $\mathbb{E}(a_{i_1i_2}a_{i_2i_3}\cdots a_{i_ki_1})$ is zero if there is an edge that is traversed exactly once. Suppose that each edge in the tour is traversed at least twice. Let p be the number of distinct vertices visited by the tour. Then the number of distinct edges traversed by the tour is at least $p-1$. Since $\sigma^2 \leq 1$, $|a_{ij}| \leq 1$, and $\mathbb{E}|a_{ij}|^l \leq \sigma^2$ for any $l \geq 2$, this shows that

$$\big|\mathbb{E}(a_{i_1i_2}a_{i_2i_3}\cdots a_{i_ki_1})\big| \leq \sigma^{2p-2}. \tag{6}$$

Thus, if $W(n,k,p)$ is the number of tours of length k that visit exactly p vertices and traverse each of its edges at least twice, then

$$\mathbb{E}\|A\|^k \leq \sum_{p=1}^{k} \sigma^{2p-2} W(n,k,p). \tag{7}$$

Vu [94], equation (5), proves that if $p > k/2$ then $W(n,k,p) = 0$ and if $p \leq k/2$ then

$$W(n,k,p) \leq n(n-1)\cdots(n-p+1)\binom{k}{2p-2} p^{2(k-2p+2)} 2^{2p-2}.$$

Using this bound, one can proceed as in [94], Section 2, to arrive at the conclusion that if k is largest even number $\leq \sigma^{1/3} n^{1/6}$, then

$$\mathbb{E}\|A\|^k \leq 2n(2\sigma\sqrt{n})^k.$$

Consequently,

$$\mathbb{E}\|A\| \leq \big(\mathbb{E}\|A\|^k\big)^{1/k} \leq (2n)^{1/k} 2\sigma\sqrt{n}.$$

This shows that if $\sigma^2 \geq n^{-1+\varepsilon}$ [or if $\sigma^2 \geq n^{-1}(\log n)^{6+\varepsilon}$], then there is a constant $C(\varepsilon)$ depending only on ε and η such that if $n \geq C(\varepsilon)$ then

$$\mathbb{E}\|A\| \leq (2 + \eta/4)\sigma\sqrt{n}. \tag{8}$$

Since a_{ij} are independent and $|a_{ij}| \leq 1$ almost surely for all i, j, and the spectral norm is a convex Lipschitz function of matrix entries with Lipschitz constant 1

(by the discussion about matrix norms at the beginning of this section), therefore one can apply Talagrand's inequality [Theorem 3.3 and inequality (4)] together with (8) and the assumption that $\sigma^2 \geq n^{-1+\varepsilon}$ to conclude that there is a constant $C(\varepsilon)$ such that if $n \geq C(\varepsilon)$ then

$$\mathbb{P}\big(\|A\| \geq (2+\eta/2)\sigma\sqrt{n}\big) \leq C_1 e^{-C_2\sigma^2 n}, \tag{9}$$

where C_1 and C_2 depend only on η. Replacing C_1 by a large enough constant $C_1(\varepsilon)$, the condition $n \geq C(\varepsilon)$ may be dropped. It is clear from the argument that it goes through in the skew-symmetric case as well.

Let us now drop the assumption of symmetry, but retain the assumption that $m = n$. Let $a'_{ij} := a_{ji}$. Then inequality (5) must be modified to say that for any even k,

$$\begin{aligned}\mathbb{E}\|A\|^k &\leq \mathbb{E}\big(\mathrm{Tr}\big((A^TA)^{k/2}\big)\big)\\ &= \sum_{1\leq i_1,\ldots,i_k\leq n} \mathbb{E}\big(a'_{i_1i_2}a_{i_2i_3}a'_{i_3i_4}a_{i_4i_5}\cdots a'_{i_{k-1}i_k}a_{i_ki_1}\big).\end{aligned}$$

As before, the term inside the sum is zero for any tour that traverses an edge exactly once. (In fact, there are more terms that are zero now; a term may be zero even if a tour traverses all of its edges at least twice.) Similarly, inequalities (6) and (7) continue to hold and, therefore, so does the rest of the argument.

Lastly, consider the case $m < n$. Augment the matrix A by adding an extra $n - m$ rows of zeros to make it an $n \times n$ matrix that satisfies all the conditions of the theorem. Clearly, the new matrix has the same spectral norm as the old one. This completes the proof. □

The key lemma. Suppose that A and B are two $m \times n$ matrices, where $m \leq n$. Let a_{ij} be the (i, j)th entry of A and b_{ij} be the (i, j)th entry of B. It is easy to see from definition that

$$\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(a_{ij}-b_{ij})^2 = \frac{1}{mn}\|A-B\|_F^2 \leq \frac{1}{n}\|A-B\|^2.$$

Thus, if $\|A - B\|$ is small enough, then the entries of A are approximately equal to the entries of B, on average. In other words, the matrix A is an estimate of the matrix B.

The goal of this section is to show that if in addition to the smallness of $\|A-B\|$, we also know that the nuclear norm $\|B\|_*$ is not too large, it is possible to get a better estimate of B based on A.

LEMMA 3.5. *Let $A = \sum_{i=1}^m \sigma_i x_i y_i^T$ be the singular value decomposition of A. Fix any $\delta > 0$ and define*

$$\hat{B} := \sum_{i\,:\,\sigma_i > (1+\delta)\|A-B\|} \sigma_i x_i y_i^T.$$

Then

$$\|\hat{B} - B\|_F \le K(\delta)\big(\|A - B\|\|B\|_*\big)^{1/2},$$

where $K(\delta) = (4 + 2\delta)\sqrt{2/\delta} + \sqrt{2 + \delta}$.

PROOF. Let $B = \sum_{i=1}^{m} \tau_i u_i v_i^T$ be the singular value decomposition of B. Without loss of generality, assume that σ_i's and τ_i's are arranged in decreasing order. Let S be the set of i such that $\sigma_i > (1 + \delta)\|A - B\|$. Define

$$G := \sum_{i \in S} \tau_i u_i v_i^T.$$

Note that by the definition of $\hat{B}$, the largest singular value of $A - \hat{B}$ is bounded above by $(1 + \delta)\|A - B\|$. In other words,

$$\|A - \hat{B}\| \le (1 + \delta)\|A - B\|. \tag{10}$$

On the other hand, by Theorem 3.1,

$$\max_{1 \le i \le m} |\sigma_i - \tau_i| \le \|A - B\|.$$

In particular, for $i \notin S$,

$$\tau_i \le \sigma_i + \|A - B\| \le (2 + \delta)\|A - B\|, \tag{11}$$

and for $i \in S$,

$$\tau_i \ge \sigma_i - \|A - B\| \ge \delta\|A - B\|. \tag{12}$$

By (11),

$$\|B - G\| \le (2 + \delta)\|A - B\|. \tag{13}$$

By (10) and (13), we have

$$\|\hat{B} - G\| \le \|\hat{B} - A\| + \|A - B\| + \|B - G\| \le (4 + 2\delta)\|A - B\|. \tag{14}$$

Since $\hat{B}$ and G both have rank $\le |S|$, the difference $\hat{B} - G$ has rank at most $2|S|$. Using this and (14), we have

$$\|\hat{B} - G\|_F \le \sqrt{2|S|}\|\hat{B} - G\| \le (4 + 2\delta)\sqrt{2|S|}\|A - B\|. \tag{15}$$

Next, observe that by (11),

$$\|B - G\|_F^2 = \sum_{i \notin S} \tau_i^2 \le (2 + \delta)\|A - B\| \sum_{i \notin S} \tau_i \le (2 + \delta)\|A - B\|\|B\|_*. \tag{16}$$

Combining (15) and (16), we have

$$\begin{aligned} \|\hat{B} - B\|_F &\le \|\hat{B} - G\|_F + \|B - G\|_F \\ &\le (4 + 2\delta)\sqrt{2|S|}\|A - B\| + \big((2 + \delta)\|A - B\|\|B\|_*\big)^{1/2}. \end{aligned} \tag{17}$$

Next, note that by (12),

$$\|B\|_* \geq \sum_{i \in S} \tau_i \geq \delta |S| \|A - B\|,$$

and thus

$$|S| \leq \frac{\|B\|_*}{\delta \|A - B\|}. \tag{18}$$

Combining (17) and (18), the proof is complete. □

Finishing the proof of Theorem 1.1. We will prove the theorem only for the asymmetric model. The only difference in the proofs for the symmetric model and the skew-symmetric model is that we need to use the symmetric and skew-symmetric parts of Theorem 3.4 instead of the asymmetric part.

Throughout this proof, $C(\varepsilon)$ will denote any constant that depends only on ε and η, and C and c will denote constants that depend only on η. The values of $C(\varepsilon)$, C and c may change from line to line or even within a line. We will use the fact that $\eta \in (0, 1)$ without mention on many occasions.

Note that for all i and j,

$$\mathbb{E}(y_{ij}) = pm_{ij}$$

and

$$\operatorname{Var}(y_{ij}) \leq \mathbb{E}(y_{ij}^2) = p\mathbb{E}(x_{ij}^2) \leq p. \tag{19}$$

Let $\hat{p}$ be the proportion of observed entries. Define two events E_1 and E_2 as

$$E_1 := \{\|Y - pM\| \leq (2 + \eta/2)\sqrt{np}\},$$
$$E_2 := \{|\hat{p} - p| \leq \eta p/20\}.$$

By Theorem 3.4,

$$\mathbb{P}(E_1) \geq 1 - C(\varepsilon)e^{-cnp}. \tag{20}$$

By Bernstein's inequality (Theorem 3.2), for any $t \geq 0$,

$$\mathbb{P}(|\hat{p} - p| \geq t) \leq 2\exp\biggl(-\frac{3mnt^2}{6p(1-p) + 2t}\biggr).$$

In particular,

$$\mathbb{P}(E_2) \geq 1 - 2e^{-cmnp}. \tag{21}$$

Let δ be defined by the relation

$$(1 + \delta)\|Y - pM\| = (2 + \eta)\sqrt{n\hat{p}}.$$

If E_1 and E_2 both happen, then

$$1 + \delta \geq \frac{(2 + \eta)\sqrt{n\hat{p}}}{(2 + \eta/2)\sqrt{np}} \geq \frac{(2 + \eta)\sqrt{(1 - \eta/20)np}}{(2 + \eta/2)\sqrt{np}} \geq 1 + \eta/5.$$

Let $K(\delta)$ be the constant in the statement of Lemma 3.5. It is easy to see that there is a constant C depending only on η such that if $\delta \geq \eta/5$, then $K(\delta) \leq C\sqrt{1+\delta}$. Therefore, by Lemma 3.5, if E_1 and E_2 both happen, then

$$\begin{aligned} \|\hat{p}W - pM\|_F^2 &\leq C(1+\delta)\|Y - pM\|\|pM\|_* \\ &\leq C\sqrt{n\hat{p}}\|pM\|_* \\ &\leq Cn^{1/2}p^{3/2}\|M\|_*. \end{aligned} \tag{22}$$

By the definition of $\hat{M}$, it is obvious that $|\hat{m}_{ij} - m_{ij}| \leq |w_{ij} - m_{ij}|$ for all i and j. Together with (22), this shows that under $E_1 \cap E_2$,

$$\begin{aligned} p^2\|\hat{M} - M\|_F^2 &\leq p^2\|W - M\|_F^2 \\ &\leq C\hat{p}^2\|W - M\|_F^2 \\ &\leq C\|\hat{p}W - pM\|_F^2 + C(\hat{p} - p)^2\|M\|_F^2 \\ &\leq Cn^{1/2}p^{3/2}\|M\|_* + C(\hat{p} - p)^2 mn. \end{aligned}$$

Note that $\mathbb{E}(\hat{p} - p)^2 = p(1-p)/mn$ and that $\|\hat{M} - M\|_F^2 \leq 4mn$. Thus, by (20) and (21),

$$\begin{aligned} \mathbb{E}\|\hat{M} - M\|_F^2 &\leq Cn^{1/2}p^{-1/2}\|M\|_* + Cp^{-1} + Cmn\big(1 - \mathbb{P}(E_1 \cap E_2)\big) \\ &\leq Cn^{1/2}p^{-1/2}\|M\|_* + Cp^{-1} + C(\varepsilon)mne^{-cnp}. \end{aligned}$$

Dividing throughout by mn, we arrive at the inequality

$$\frac{1}{mn}\mathbb{E}\|\hat{M} - M\|_F^2 \leq \frac{C\|M\|_*}{m\sqrt{np}} + \frac{C}{np} + C(\varepsilon)e^{-cnp}. \tag{23}$$

The next goal is to show that

$$\frac{1}{mn}\mathbb{E}\|\hat{M} - M\|_F^2 \leq \frac{C\|M\|_*^2}{mn} + C(\varepsilon)e^{-cnp}. \tag{24}$$

First, suppose that $\|M\|_* > \eta\sqrt{n/p}/20$. Then

$$\frac{\|M\|_*}{m\sqrt{np}} + \frac{1}{np} \leq \frac{C\|M\|_*^2}{mn},$$

and so (24) follows from (23). Therefore, assume that $\|M\|_* \leq \eta\sqrt{n/p}/20$. Then in particular, $\|M\| \leq \eta\sqrt{n/p}/20$. Therefore, if $E_1 \cap E_2$ happens, then

$$\begin{aligned} \|Y\| &\leq \|Y - pM\| + \|pM\| \\ &\leq (2 + \eta/2 + \eta/20)\sqrt{np} \\ &\leq \frac{(2 + 11\eta/20)\sqrt{n\hat{p}}}{1 - \eta/20} \leq (2 + 13\eta/20)\sqrt{n\hat{p}}. \end{aligned}$$

This implies that there is no singular value of Y that exceeds $(2+\eta)\sqrt{n\hat{p}}$, and therefore $\hat{M}=0$. Consequently,

$$\|\hat{M}-M\|_F^2 = \|M\|_F^2 \le \|M\|_*^2.$$

Thus, if $\|M\|_* \le \eta\sqrt{n/p}/20$, then by (20) and (21),

$$\frac{1}{mn}\mathbb{E}\|\hat{M}-M\|_F^2 \le \frac{\|M\|_*^2}{mn} + C\bigl(1-\mathbb{P}(E_1\cap E_2)\bigr) \le \frac{\|M\|_*^2}{mn} + C(\varepsilon)e^{-cnp}.$$

Combining the above steps and observing that $\mathrm{MSE}(\hat{M}) \le 1$ due to the boundedness of the entries of M and $\hat{M}$, we get

$$\mathrm{MSE}(\hat{M}) \le C\min\biggl\{\frac{\|M\|_*}{m\sqrt{np}} + \frac{1}{np}, \frac{\|M\|_*^2}{mn}, 1\biggr\} + C(\varepsilon)e^{-cnp}.$$

To remove the $1/np$ term, note that if that term indeed matters, then we are in a situation where

$$\frac{\|M\|_*}{m\sqrt{np}} \le \frac{\|M\|_*^2}{mn}.$$

But this inequality, on the other hand, implies that

$$\frac{\|M\|_*}{m\sqrt{np}} \ge \frac{1}{mp} \ge \frac{1}{np}.$$

Therefore, the $1/np$ term can be removed from the above bound. This completes the proof of Theorem 1.1 if no nontrivial bound on $\mathrm{Var}(x_{ij})$ is known.

If $\sigma^2 \le 1$ is a known constant such that $\mathrm{Var}(x_{ij}) \le \sigma^2$ for all i, j, then the estimate (19) may be improved to

$$\mathrm{Var}(y_{ij}) = p\,\mathrm{Var}(x_{ij}) + p(1-p)m_{ij}^2 \le \max_{(a,b)\in R}\bigl(pb + p(1-p)a\bigr),$$

where R is the quadrilateral region

$$\bigl\{(a,b): 0\le a\le 1, 0\le b\le \sigma^2, 0\le a+b\le 1\bigr\}.$$

The maximum must be attained at one of the four vertices of R. An easy verification shows that the maximum is always attained at the vertex $(1-\sigma^2,\sigma^2)$, which gives the upper bound

$$\mathrm{Var}(y_{ij}) \le q := p\sigma^2 + p(1-p)\bigl(1-\sigma^2\bigr).$$

This allows us to replace the threshold $(2+\eta)\sqrt{n\hat{p}}$ by $(2+\eta)\sqrt{n\hat{q}}$, where $\hat{q} = \hat{p}\sigma^2 + \hat{p}(1-\hat{p})(1-\sigma^2)$. As before, we need that $q \ge n^{-1+\varepsilon}$. The rest of the proof goes through with the following modifications: Replace $\sqrt{np}$ by $\sqrt{nq}$ in the definition of E_1, keep E_2 the same, and define an event $E_3 = \{|\hat{q}-q| \le \eta q/20\}$.

By Theorem 3.4, $\mathbb{P}(E_1) \geq 1 - C(\varepsilon)e^{-cnq}$, $\mathbb{P}(E_2) \geq 1 - 2e^{-cmnp} \geq 1 - 2e^{-cmnq}$, and $\mathbb{P}(E_3) \geq 1 - 2e^{-cmnq}$ since $|\hat{q} - q| \leq |\hat{p} - p|$ and $q \geq p(1-p)$ and, therefore,

$$\mathbb{P}(E_3^c) \leq \mathbb{P}(|\hat{p} - p| > \eta q/20) \leq 2\exp\biggl(-\frac{cmnq^2}{6p(1-p) + \eta q/10}\biggr) \leq 2e^{-cmnq}.$$

If $E_1 \cap E_2 \cap E_3$ happens, then the subsequent steps remain the same, but with some suitable modifications that replace the term $\|M\|_*/(m\sqrt{np})$ by the improved term $\|M\|_*\sqrt{q}/(m\sqrt{n}p)$.

3.2. *Proof of Theorem 1.2* (*Minimax optimality*). Throughout this proof, C will denote any positive universal constant, whose value may change from line to line.

Take any $\delta \in [0, m\sqrt{n}]$ and let $\theta := \delta/(m\sqrt{n})$. We will first work out the proof under the assumption that $p < 1/2$. Under this assumption, three situations are considered. First, suppose that

$$\theta/\sqrt{p} \leq 1 \quad \text{and} \quad m\theta\sqrt{p} \geq 1. \tag{25}$$

Let $k := [m\theta\sqrt{p}]$. Clearly, $k \leq m$. Let M be an $m \times n$ random matrix whose first k rows consist of i.i.d. Uniform$[-1, 1]$ random variables, and copy this block $[1/p]$ times. This takes care of $k[1/p]$ rows. [This is okay, since $k/p \leq m\theta/\sqrt{p} \leq m$ by (25).] Declare the remaining rows, if any, to be zero. Then note that M has rank $\leq k \leq m\theta\sqrt{p}$. Therefore, by inequality (2),

$$\|M\|_* \leq (m\theta\sqrt{p})^{1/2}\|M\|_F \leq (m\theta\sqrt{p})^{1/2}(mn\theta/\sqrt{p})^{1/2} = m\sqrt{n}\theta.$$

Let $X = M$. Let D be our data, that is, the observed values of X. One can imagine D as a matrix whose (i, j)th entry is x_{ij} if x_{ij} is observed, and a question mark if x_{ij} is unobserved. For any (i, j) belonging to the nonzero portion of the matrix M, M contains $[1/p]$ copies of m_{ij}. Since the X-value at the location of each copy is observed with probability p, independent of the other copies, and $p < 1/2$, therefore, the chance that none of these copies are observed is bounded below by a positive universal constant. If none of the copies are observed, then the data contains no information about m_{ij}. Using this, it is not difficult to write down a formal argument that shows

$$\mathbb{E}\bigl(\mathrm{Var}(m_{ij}|D)\bigr) \geq C.$$

On the other hand, since $\tilde{m}_{ij}$ is a function of D, the definition of variance implies that

$$\mathbb{E}\bigl((\tilde{m}_{ij} - m_{ij})^2|D\bigr) \geq \mathrm{Var}(m_{ij}|D).$$

Combining the last two displays, we see that

$$\begin{aligned}\mathbb{E}\|\tilde{M} - M\|_F^2 &\geq \sum_{i=1}^{k[1/p]}\sum_{j=1}^{n}\mathbb{E}(\tilde{m}_{ij} - m_{ij})^2 \\ &\geq Ck[1/p]n \geq \frac{Cmn\theta}{\sqrt{p}}.\end{aligned} \tag{26}$$

The argument that led to the above lower bound is a typical example of the classical Bayesian argument for obtaining minimax lower bounds, and will henceforth be referred to as the "standard minimax argument" to avoid repetition of details.

Next, assume that

$$\theta/\sqrt{p} \le 1 \quad \text{and} \quad m\theta\sqrt{p} < 1. \tag{27}$$

Let M be an $m \times n$ matrix whose first row consists of i.i.d. random variables uniformly distributed over the interval $[-m\theta\sqrt{p}, m\theta\sqrt{p}]$, and this row is copied $[1/p]$ times, and all other rows are zero. Then M has rank ≤ 1, and therefore by inequality (2),

$$\|M\|_* \le \|M\|_F \le m\theta\sqrt{p}\sqrt{\frac{n}{p}} = m\theta\sqrt{n}.$$

On the other hand, a standard minimax argument as before implies that for any estimator $\tilde{M}$,

$$\mathbb{E}\|\tilde{M} - M\|_F^2 \ge (m\theta\sqrt{p})^2\frac{n}{p} = nm^2\theta^2.$$

In particular, under (27), there exists M with $\|M\|_* \le \delta$ such that

$$\mathrm{MSE}(\hat{M}) \ge \frac{C\delta^2}{mn}.$$

Finally, suppose that

$$\theta/\sqrt{p} > 1. \tag{28}$$

Let M be an $m \times n$ matrix whose first $[mp]$ rows consist of i.i.d. random variables uniformly distributed over $[-1, 1]$, and this block is copied $[1/p]$ times. Then the rank of M is $\le [mp]$, and so by (28) and (2),

$$\|M\|_* \le \sqrt{mp}\|M\|_F \le m\sqrt{np} \le \theta m\sqrt{n}.$$

Again by a standard minimax argument, it is easy to conclude that for any estimator $\tilde{M}$, there exists M with $\|M\|_* \le \delta$ such that

$$\mathrm{MSE}(\tilde{M}) \ge C.$$

This completes the proof when $p < 1/2$. Next, suppose that $p \ge 1/2$. The only place where the assumption $p < 1/2$ was used previously was for proving that $\mathbb{E}(\mathrm{Var}(m_{ij}|D)) \ge C$. This can be easily taken care of by inserting some randomness into the data matrix X, as follows. First, replace M by $\frac{1}{2}M$ in all three cases above. This retains the condition $\|M\|_* \le \delta$. Given M, let X be the data matrix whose (i, j)th entry x_{ij} is uniformly distributed over the interval $[m_{ij} - 1/2, m_{ij} + 1/2]$, whenever (i, j) is the "main block" of M; and this value of x_{ij} is copied $[1/p]$ times in the appropriate places.

Since the entries of M are now guaranteed to be in $[-1/2, 1/2]$, this ensures that the entries of X are in $[-1, 1]$. Now note that even if x_{ij} or one of its copies is observed, it gives only limited information about m_{ij}. In particular, it is easy to prove that $\mathbb{E}(\mathrm{Var}(m_{ij}|D)) \geq C$ and complete the proof as before.

This complete the proof of Theorem 1.2 for the asymmetric model. For the symmetric model, simply observe that the singular values of any square matrix M are the same as those of the symmetric matrix

$$\begin{bmatrix} 0 & M \\ M^T & 0 \end{bmatrix}$$

with multiplicity doubled. It is now clear how the minimax arguments for the asymmetric model may be carried over to the symmetric case by considering the same Bayesian models for M and working with the corresponding symmetrized matrices. For the skew-symmetric case, replace the M^T by $-M^T$ in the above matrix.

3.3. *Proof of Theorem 1.3* (*Impossibility of error estimation*). Suppose that a good procedure P exists. By the definition of nontriviality of the estimator $\tilde{M}$, there exists a sequence of parameter matrices M_n and data matrices X_n such that

$$\mathrm{MSE}\big(\hat{M}_n^{\mathrm{Trv}}\big) \not\to 0 \qquad \text{as } n \to \infty, \tag{29}$$

but

$$\lim_{n\to\infty} \mathrm{MSE}(\tilde{M}_n) = 0. \tag{30}$$

Then by the definition of goodness,

$$\widehat{\mathrm{MSE}}_{\mathrm{P}}(\tilde{M}_n) \to 0 \qquad \text{in probability as } n \to \infty.$$

Suppose, without loss of generality, that all the data matrices are defined on the same probability space. Then taking a subsequence if necessary, we may assume that in addition to (29) and (30), we also have

$$\mathbb{P}\Big(\lim_{n\to\infty} \widehat{\mathrm{MSE}}_{\mathrm{P}}(\tilde{M}_n) = 0\Big) = 1. \tag{31}$$

Let $M_n' := X_n$ and $X_n' := X_n$ for all n. Consider M_n' as a (random) parameter matrix and X_n' as its data matrix. Given M_n', the expected value of X_n' is M_n'; so it is okay to treat M_n' as a parameter matrix and X_n' as its data matrix. We will denote the estimate of M_n' constructed using X_n' as $\tilde{M}_n'$. Note that since M_n' is random, the mean squared error of $\tilde{M}_n'$ is a random variable.

Now, since the estimator $\tilde{M}_n'$ is computed using the data matrix only, and $X_n' = X_n$, it is clear that $\tilde{M}_n' = \tilde{M}_n$. There is no randomness in $\tilde{M}_n'$ when M_n' is given,

since $X_n' = M_n'$. Thus, if r_n and c_n denote the number of rows and columns of M_n, then

$$\begin{aligned}\mathrm{MSE}(\tilde{M}_n') &= \frac{1}{r_n c_n}\|\tilde{M}_n' - M_n'\|_F^2 \\ &= \frac{1}{r_n c_n}\|\tilde{M}_n - X_n\|_F^2 \\ &= \frac{1}{r_n c_n}\|\tilde{M}_n - \hat{M}_n^{\mathrm{Trv}}\|_F^2 \\ &\geq \frac{1}{2r_n c_n}\|\hat{M}_n^{\mathrm{Trv}} - M_n\|_F^2 - \frac{1}{r_n c_n}\|\tilde{M}_n - M_n\|_F^2,\end{aligned}$$

where the last step follows from the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ and the triangle inequality for the Frobenius norm. Taking expectation on both sides gives

$$\mathbb{E}(\mathrm{MSE}(\tilde{M}_n')) \geq \tfrac{1}{2}\,\mathrm{MSE}(\hat{M}_n^{\mathrm{Trv}}) - \mathrm{MSE}(\tilde{M}_n).$$

Therefore, by (29) and (30),

$$\mathbb{E}(\mathrm{MSE}(\tilde{M}_n')) \not\to 0 \qquad \text{as } n \to \infty.$$

In particular, since mean squared errors are uniformly bounded by 1,

$$\mathbb{P}(\mathrm{MSE}(\tilde{M}_n') \not\to 0 \text{ as } n \to \infty) > 0. \tag{32}$$

Again since $\widehat{\mathrm{MSE}}_{\mathrm{P}}$ is computed using the data matrix only, therefore, for all n,

$$\widehat{\mathrm{MSE}}_{\mathrm{P}}(\tilde{M}_n) = \widehat{\mathrm{MSE}}_{\mathrm{P}}(\tilde{M}_n').$$

Therefore, by (31),

$$\mathbb{P}\Big(\lim_{n\to\infty} \widehat{\mathrm{MSE}}_{\mathrm{P}}(\tilde{M}_n') = 0\Big) = 1. \tag{33}$$

Equations (32) and (33) demonstrate the existence of a sequence of parameter matrices M_n' and data matrices X_n' such that $\mathrm{MSE}(\tilde{M}_n') \not\to 0$ but $\widehat{\mathrm{MSE}}_{\mathrm{P}}(\tilde{M}_n') \to 0$ in probability. This contradicts the goodness of $\widehat{\mathrm{MSE}}_{\mathrm{P}}$.

3.4. *Proof of Theorem 2.1* (*Upper bound for low rank matrix estimation*). Inequality (2) implies that

$$\|M\|_* \leq \sqrt{\mathrm{rank}(M)}\,\|M\|_F \leq \sqrt{rmn}.$$

The result now follows from Theorem 1.1.

3.5. *Proof of Theorem 2.2* (*Lower bound for low rank matrix estimation*). Let M be an $m \times n$ random matrix whose first r rows consist of i.i.d. Uniform$[-1, 1]$ random variables, and copy this block $[m/r]$ times. Declare the remaining rows, if any, to be zero. Then note that M has rank $\le r$.

Let D be our data, that is, the observed values of M. One can imagine D as a matrix whose (i, j)th entry is m_{ij} if m_{ij} is observed, and a question mark if m_{ij} is unobserved. For any (i, j) belonging to the nonzero portion of the matrix M, M contains $[m/r]$ copies of m_{ij}. Since the M-value at the location of each copy is observed with probability p, independent of the other copies, the chance that none of these copies are observed is equal to $(1-p)^{[m/r]}$. If none of the copies are observed, then the data contains no information about m_{ij}. Using this, it is not difficult to write down a formal argument that shows

$$\mathbb{E}\big(\mathrm{Var}(m_{ij}|D)\big) \ge C(1-p)^{[m/r]},$$

where C is some universal constant. On the other hand, since $\tilde{m}_{ij}$ is a function of D, the definition of variance implies that

$$\mathbb{E}\big((\tilde{m}_{ij} - m_{ij})^2|D\big) \ge \mathrm{Var}(m_{ij}|D).$$

Combining the last two displays, we see that

$$\mathbb{E}\|\tilde{M} - M\|_F^2 \ge Cmn(1-p)^{[m/r]}.$$

This completes the proof.

3.6. *Proof of Theorem 2.3* (*Block model estimation*). If two vertices i and j are in the same block, then the ith and jth rows of M are identical. Therefore, M has at most k distinct rows and so the rank of M is $\le k$. An application of Theorem 2.1 completes the proof.

3.7. *Proofs of Theorems 2.4 and 2.5* (*Distance matrix estimation*). The proofs of Theorems 2.4 and 2.5 follow from a more general lemma that will also be useful later for other purposes. Suppose that $S = \{x_1, \dots, x_n\}$ is a finite set and $f : S \times S \to [-1, 1]$ is an arbitrary function. Suppose that for each $\delta > 0$, there exists a partition $\mathcal{P}(\delta)$ of S such that whenever x, y, x', y' are four points in S such that $x, x' \in P$ for some $P \in \mathcal{P}(\delta)$ and $y, y' \in Q$ for some $Q \in \mathcal{P}(\delta)$, then $|f(x, y) - f(x', y')| \le \delta$. Let M be the $n \times n$ matrix whose (i, j)th element is $f(x_i, x_j)$.

LEMMA 3.6. *In the above setting,*

$$\mathrm{MSE}(\hat{M}) \le C \inf_{\delta > 0} \min\left\{\frac{\delta + \sqrt{|\mathcal{P}(\delta)|/n}}{\sqrt{p}}, 1\right\} + C(\varepsilon)e^{-cnp},$$

where C and c depend only on η, and $C(\varepsilon)$ depends only on ε and η.

PROOF. Fix some $\delta > 0$. Let T be a subset of S consisting of exactly one point from each member of $\mathcal{P}(\delta)$. For each $x \in S$, let $p(x)$ be the unique element of T such that x and $p(x)$ belong to the same element of $\mathcal{P}(\delta)$. Let N be the matrix whose (i, j)th element is $f(p(x_i), p(x_j))$. Then

$$\|M - N\|_F^2 = \sum_{i,j=1}^{n} \bigl(f(x_i, x_j) - f\bigl(p(x_i), p(x_j)\bigr)\bigr)^2 \le n^2\delta^2.$$

By the triangle inequality for the nuclear norm, the inequality (2) and the above inequality,

$$\begin{aligned}\|M\|_* &\le \|M - N\|_* + \|N\|_* \\ &\le \sqrt{n}\|M - N\|_F + \|N\|_* \\ &\le n^{3/2}\delta + \|N\|_*.\end{aligned}$$

Now, if x_i and x_j belong to the same element of $\mathcal{P}(\delta)$, then $p(x_i) = p(x_j)$, and hence the ith and jth rows of N are identical. This shows that N has at most $|\mathcal{P}(\delta)|$ distinct rows and, therefore, has rank $\le |\mathcal{P}(\delta)|$. Therefore, by the inequality (2),

$$\|N\|_* \le \sqrt{|\mathcal{P}(\delta)|}\|N\|_F \le \sqrt{|\mathcal{P}(\delta)|}n.$$

The proof is completed by applying Theorem 1.1. □

Using Lemma 3.6, it is easy to prove Theorems 2.4 and 2.5.

PROOF OF THEOREM 2.5. Let all notation be as in Theorem 2.5. To apply Lemma 3.6, let S be the set $\{x_1, \ldots, x_n\}$. From the definition of $N(\delta)$, it is easy to see that there is a partition $\mathcal{P}(\delta)$ of S of size $\le N(\delta/4)$, such that any two points belonging to the same element of the partition are at distance $\le \delta/2$ from each other. Consequently, if $x, x' \in P$ and $y, y' \in Q$ for some $P, Q \in \mathcal{P}(\delta)$, then by the triangle inequality for the metric d,

$$\begin{aligned}\bigl|d(x, y) - d\bigl(x', y'\bigr)\bigr| &\le \bigl|d(x, y) - d\bigl(x', y\bigr)\bigr| + \bigl|d\bigl(x', y\bigr) - d\bigl(x', y'\bigr)\bigr| \\ &\le d\bigl(x, x'\bigr) + d\bigl(y, y'\bigr) \le \delta.\end{aligned}$$

Putting $f = d$ in Lemma 3.6, the proof is complete. □

PROOF OF THEOREM 2.4. Since K is compact, there exists a finite number $N(\delta)$ for each $\delta > 0$ such that K may be covered by $N(\delta)$ open d-balls of radius δ. By Theorem 2.5, this shows that for any sequence δ_n decreasing to zero,

$$\mathrm{MSE}(\hat{M}) \le C \min\left\{\frac{\delta_n + \sqrt{N(\delta_n/4)/n}}{\sqrt{p}}, 1\right\} + C(\varepsilon)e^{-cnp}.$$

To complete the proof, choose δ_n going to zero so slowly that $N(\delta_n/4) = o(n)$ as $n \to \infty$. □

3.8. *Proof of Theorem* 2.6 (*Latent space models*: *General case*). We will apply Lemma 3.6. Let S be the set $\{\beta_1, \ldots, \beta_n\}$. Since f is continuous on K and K is compact, f must be uniformly continuous. This shows that for each $\delta > 0$ we can find a partition $\mathcal{P}(\delta)$ of S satisfying the condition required for Lemma 3.6, such that the size of $\mathcal{P}(\delta)$ may be bounded by a constant $N(\delta)$ depending only on K, k, f and δ. Choosing $\delta_n \to 0$ slowly enough so that $N(\delta_n/4) = o(n)$ and applying Lemma 3.6 completes the proof.

3.9. *Proof of Theorem* 2.7 (*Latent space models*: *Lipschitz functions*). Let $S = \{\beta_1, \ldots, \beta_n\}$. Take any $\delta > 0$. From the Lipschitzness condition, it is easy to see that we can find a partition $\mathcal{P}(\delta)$ of S whose size may be bounded by $C(K, k, L)\delta^{-k}$, where $C(K, k, L)$ depends only on K, k and L. Choosing $\delta = n^{-1/(k+2)}$ and applying Lemma 3.6 completes the proof. Note that the exponential term need not appear since the main term is bounded below by a positive constant if $p < n^{-2/(k+2)}$.

3.10. *Proof of Theorem* 2.8 (*Upper bound for positive definite matrix estimation*). Since M is positive semi-definite, $\|M\|_* = \mathrm{Tr}(M)$. Since the entries of M are bounded by 1, $\mathrm{Tr}(M) \le n$. The proof now follows from an application of Theorem 1.1.

3.11. *Proof of Theorem* 2.9 (*Lower bound for positive definite matrix estimation*). Throughout this proof, C will denote any positive universal constant, whose value may change from line to line.

Let $U_1, \ldots, U_n$ be i.i.d. Uniform[0, 1] random variables. Let M be the random matrix whose (i, j)th element m_{ij} is equal to $U_i U_j$ if $i \neq j$ and 1 if $i = j$. It is easy to verify that M is a correlation matrix. Suppose that we observe each element of M on and above the diagonal with probability p, independent of each other. Let D be our data, represented as follows: D is a matrix whose (i, j)th element is m_{ij} if the element is observed, and a question mark otherwise.

Now, the probability that no element from the ith row and the ith column is observed is exactly equal to $(1 - p)^n$. If we do not observe any element from the ith row and ith column, we have no information about the value of U_i. From this, it is not difficult to write down a formal argument to prove that for any $j \neq i$,

$$\mathrm{Var}\big(m_{ij} | D, (U_k)_{k \neq i}\big) = U_j^2 \,\mathrm{Var}\big(U_i | D, (U_k)_{k \neq i}\big) \ge C(1 - p)^n U_j^2.$$

If $\tilde{M}$ is any estimator, then $\tilde{m}_{ij}$ is a function of D. Therefore, by the above inequality and the definition of variance,

$$\mathbb{E}\big((\tilde{m}_{ij} - m_{ij})^2 | D, (U_k)_{k \neq i}\big) \ge C(1 - p)^n U_j^2,$$

and thus

$$\mathbb{E}(\tilde{m}_{ij} - m_{ij})^2 \ge C(1 - p)^n.$$

Since this is true for all $i \neq j$, the proof is complete.

3.12. *Proof of Theorem 2.10* (*Graphon estimation*). Here, all entries of the adjacency matrix are visible, so $p = 1$. Define a sequence of functions $f_1, f_2, \ldots$ according the following standard construction. For each k, let $\mathcal{P}_k$ be the kth level dyadic partition of $[0,1)^2$, that is, the partition of the unit square into sets of the form $[(i-1)/2^k, i/2^k) \times [(j-1)/2^k, j/2^k)$. Let f_k be the function that is equal to the average value of f within each square of the partition $\mathcal{P}_k$. If $\mathcal{F}_k$ denotes the sigma-algebra of sets generated by the partition $\mathcal{P}_k$, then the sequence f_k is a martingale with respect to the filtration $\mathcal{F}_k$. Moreover, $f_k = \mathbb{E}(f|\mathcal{F}_k)$. Finally, observe that the sequence f_k is uniformly bounded in L^2. Combining all these observations, it is evident that $f_k \to f$ in L^2.

Now fix some $\varepsilon > 0$ and an integer n. Take a large enough $k = k(\varepsilon)$ such that $\|f - f_k\|_{L^2} \le \varepsilon$. Let N be the $n \times n$ matrix whose (i,j)th element is $f_k(U_i, U_j)$. Then

$$\begin{aligned}\mathbb{E}\|M - N\|_F^2 &= \sum_{i,j=1}^{n} \mathbb{E}\big(f(U_i,U_j) - f_k(U_i,U_j)\big)^2 \\ &\le n + n^2 \mathbb{E}\big(f(U_1,U_2) - f_k(U_1,U_2)\big)^2 \\ &= n + n^2 \|f - f_k\|_{L^2}^2 \le n + n^2\varepsilon^2.\end{aligned} \tag{34}$$

Now note that if U_i and U_j belong to the same dyadic interval $[r/2^k, (r+1)/2^k)$, then the ith and jth rows of N are identical. Hence, N has at most 2^k distinct rows, and therefore has rank $\le 2^k$. Therefore, by (2),

$$\|N\|_* \le 2^{k/2}\|N\|_F \le 2^{k/2} n.$$

Consequently, by the inequality (2),

$$\begin{aligned}\|M\|_* &\le \|M - N\|_* + \|N\|_* \\ &\le \sqrt{n}\|M - N\|_F + 2^{k/2} n.\end{aligned} \tag{35}$$

Combining (34) and (35) gives

$$\mathbb{E}\|M\|_* \le \big(2^{k/2} + 1\big)n + n^{3/2}\varepsilon.$$

Choosing a sequence ε_n going to zero so slowly that $2^{k(\varepsilon_n)/2} = o(n^{-1/2})$, we can now apply Theorem 1.1 to complete the proof.

3.13. *Proof of Theorem 2.11* (*Bradley–Terry models*). Throughout the proof C will denote any constant that depends only on η, whose value may change from line to line.

Recall that the definition of the skew-symmetric model stipulates that $X - M$ is skew-symmetric, which is true for the nonparametric Bradley–Terry model. There

is nothing to prove if $p < n^{-2/3}$, so assume that $p \geq n^{-2/3}$. This allows us to drop the exponential term in Theorem 1.1 and conclude that

$$\text{MSE}(\hat{M}) \leq \frac{C\|M\|_*}{n^{3/2}\sqrt{p}}. \tag{36}$$

Let k be an integer less than n, to determined later. For each i, let

$$t_i := \sum_{j=1}^{n} p_{ij}.$$

Note that each t_i belongs to the interval $[0, n]$. For $l = 1, \ldots, k$, let T_l be the set of all i such that $t_i \in [n(l-1)/k, nl/k)$. Additionally, if $t_i = n$, put i in T_k.

For each l, let $r(l)$ be a distinguished element of T_l. For each $1 \leq i, j \leq n$, if $i \in T_l$ and $j \in T_m$, let $n_{ij} := p_{r(l)j}$. Let N be the matrix whose (i, j)th element is n_{ij}. Note that if $i, i' \in T_l$ for some l, then $n_{ij} = n_{i'j}$ for all j. In particular, N has at most k distinct rows and therefore has rank $\leq k$. Thus, by inequality (2),

$$\|N\|_* \leq \sqrt{k}\|N\|_F \leq n\sqrt{k}. \tag{37}$$

Now take any $1 \leq i \leq n$. Suppose that $i \in T_l$. Let $i' = r(l)$. Suppose that team i' is weaker than team i. Then $p_{i'j} \leq p_{ij}$ for all $j \neq i, i'$. Thus,

$$\begin{aligned}\sum_{j=1}^{n}(p_{ij} - n_{ij})^2 &= \sum_{j=1}^{n}(p_{ij} - p_{i'j})^2 \leq \sum_{j=1}^{n}|p_{ij} - p_{i'j}| \\ &= \sum_{j\neq i,i'}(p_{ij} - p_{i'j}) + p_{ii'} + p_{i'i} \\ &\leq t_i - t_{i'} + 2 \leq \frac{n}{k} + 2 \leq \frac{3n}{k}.\end{aligned} \tag{38}$$

Similarly, if team i' is stronger than team i,

$$\sum_{j=1}^{n}(p_{ij} - n_{ij})^2 \leq t_{i'} - t_i + 2 \leq \frac{3n}{k}. \tag{39}$$

By (37), (38), (39) and (2) we have

$$\begin{aligned}\|M\|_* &\leq \|M - N\|_* + \|N\|_* \\ &\leq \sqrt{n}\|M - N\|_F + n\sqrt{k} \\ &\leq \frac{3n^{3/2}}{\sqrt{k}} + n\sqrt{k}.\end{aligned}$$

Choosing $k = [n^{1/2}]$, we get $\|M\|_* \leq Cn^{5/4}$. Combined with (36), this proves the claim.

Acknowledgments. I would like to thank Emmanuel Candès for introducing me to this topic, Andrea Montanari and Peter Bickel for pointing out many relevant references, and Persi Diaconis for helpful advice. Special thanks to Yaniv Plan for pointing out an important mistake in the first draft, and to Philippe Rigollet for correcting an error in Theorem 2.11. I would also like to thank the three anonymous referees for a long list of useful comments.

REFERENCES

[1] ACHLIOPTAS, D. and MCSHERRY, F. (2001). Fast computation of low rank matrix approximations. In *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing* 611–618 (electronic). ACM, New York. MR2120364

[2] ADAMS, E. (2005). Bayesian analysis of linear dominance hierarchies. *Animal Behaviour* **69** 1191–1201.

[3] AGRESTI, A. (1990). *Categorical Data Analysis*. Wiley, New York. MR1044993

[4] AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

[5] ALDOUS, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.* **11** 581–598. MR0637937

[6] ALFAKIH, A. Y., KHANDANI, A. and WOLKOWICZ, H. (1999). Solving Euclidean distance matrix completion problems via semidefinite programming. *Comput. Optim. Appl.* **12** 13–30. MR1704098

[7] AMINI, A. A., CHEN, A., BICKEL, P. J. and LEVINA, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* **41** 2097–2122. MR3127859

[8] ANDERSON, G. W., GUIONNET, A. and ZEITOUNI, O. (2010). *An Introduction to Random Matrices*. *Cambridge Studies in Advanced Mathematics* **118**. Cambridge Univ. Press, Cambridge. MR2760897

[9] AUSTIN, T. (2008). On exchangeable random variables and the statistics of large graphs and hypergraphs. *Probab. Surv.* **5** 80–145. MR2426176

[10] AZAR, Y., FLAT, A., KARLIN, A., MCSHERRY, F. and SALA, J. (2001). Spectral analysis of data. In *Proceedings of the Thirty-third Annual ACM Symposium on Theory of Computing* 619–626. ACM, New York.

[11] BAKONYI, M. and JOHNSON, C. R. (1995). The Euclidean distance matrix completion problem. *SIAM J. Matrix Anal. Appl.* **16** 646–654. MR1321802

[12] BENNETT, G. (1962). Probability inequalities for sums of independent random variables. *J. Amer. Statist. Assoc.* **57** 33–45.

[13] BERNSTEIN, S. (1924). Sur une modification de l'inéqualité de Tchebichef. *Annals Science Institute Sav. Ukraine, Sect. Math. I* (Russian, French summary.) 38–49.

[14] BHATIA, R. (1997). *Matrix Analysis*. *Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662

[15] BHATIA, R. (2007). *Positive Definite Matrices*. Princeton Univ. Press, Princeton, NJ. MR2284176

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

[17] BICKEL, P. J., CHEN, A. and LEVINA, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist.* **39** 2280–2301. MR2906868

[18] BISWAS, P., LIAN, T.-C., WANG, T.-C. and YE, Y. (2006). Semidefinite programming based algorithms for sensor network localization. *ACM Trans. Sen. Netw.* **2** 188–220.

[19] BORG, I. and GROENEN, P. J. F. (2005). *Modern Multidimensional Scaling*: *Theory and Applications*, 2nd ed. Springer, New York. MR2158691

[20] BORGS, C., CHAYES, J., LOVÁSZ, L., SÓS, V. T. and VESZTERGOMBI, K. (2006). Counting graph homomorphisms. In *Topics in Discrete Mathematics*. *Algorithms Combin*. **26** 315–371. Springer, Berlin. MR2249277

[21] BORGS, C., CHAYES, J. T., LOVÁSZ, L., SÓS, V. T. and VESZTERGOMBI, K. (2008). Convergent sequences of dense graphs. I. Subgraph frequencies, metric properties and testing. *Adv. Math*. **219** 1801–1851. MR2455626

[22] BORGS, C., CHAYES, J. T., LOVÁSZ, L., SÓS, V. T. and VESZTERGOMBI, K. (2012). Convergent sequences of dense graphs II. Multiway cuts and statistical physics. *Ann. of Math*. (2) **176** 151–219. MR2925382

[23] BRADLEY, R. A. and TERRY, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39** 324–345. MR0070925

[24] CAI, J.-F., CANDÈS, E. J. and SHEN, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM J. Optim*. **20** 1956–1982. MR2600248

[25] CANDÈS, E. J. and PLAN, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE* **98** 925–936.

[26] CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math*. **9** 717–772. MR2565240

[27] CANDÈS, E. J., ROMBERG, J. and TAO, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory* **52** 489–509. MR2236170

[28] CANDÈS, E. J. and TAO, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory* **56** 2053–2080. MR2723472

[29] CARON, F. and DOUCET, A. (2012). Efficient Bayesian inference for generalized Bradley–Terry models. *J. Comput. Graph. Statist*. **21** 174–196. MR2913362

[30] CHATTERJEE, S. and DIACONIS, P. (2013). Estimating and understanding exponential random graph models. *Ann. Statist*. **41** 2428–2461. MR3127871

[31] CHATTERJEE, S., DIACONIS, P. and SLY, A. (2011). Random graphs with a given degree sequence. *Ann. Appl. Probab*. **21** 1400–1435. MR2857452

[32] CHATTERJEE, S. and VARADHAN, S. R. S. (2011). The large deviation principle for the Erdős–Rényi random graph. *European J. Combin*. **32** 1000–1017. MR2825532

[33] CHATTERJEE, S. and VARADHAN, S. R. S. (2012). Large deviations for random matrices. *Commun. Stoch. Anal*. **6** 1–13. MR2890846

[34] CHAUDHURI, K., CHUNG, F. and TSIATAS, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. *J. Mach. Learn. Res*. **35** 1–23.

[35] CHOI, D. and WOLFE, P. J. (2014). Co-clustering separately exchangeable network data. *Ann. Statist*. **42** 29–63. MR3161460

[36] CHOI, D. S., WOLFE, P. J. and AIROLDI, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99** 273–284. MR2931253

[37] CONDON, A. and KARP, R. M. (2001). Algorithms for graph partitioning on the planted partition model. *Random Structures Algorithms* **18** 116–140. MR1809718

[38] DAVENPORT, M. A., PLAN, Y., VAN DEN BERG, E. and WOOTTERS, M. (2012). 1-bit matrix completion. Preprint. Available at arXiv:1209.3672.

[39] DAVID, H. A. (1988). *The Method of Paired Comparisons*, 2nd ed. *Griffin's Statistical Monographs & Courses* **41**. Oxford Univ. Press, London. MR0947340

[40] DAVIDSON, R. R. and FARQUHAR, P. H. (1976). A bibliography on the method of paired comparisons. *Biometrics* **32** 241–252. MR0408134

[41] DIACONIS, P. (1988). *Group Representations in Probability and Statistics*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **11**. IMS, Hayward, CA. MR0964069

[42] DIACONIS, P. and JANSON, S. (2008). Graph limits and exchangeable random graphs. *Rend. Mat. Appl.* (7) **28** 33–61. MR2463439

[43] DONOHO, D. L. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* **52** 1289–1306. MR2241189

[44] DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224. MR1379464

[45] FAZEL, M. (2002). Matrix rank minimization with applications. Ph.D. thesis, Stanford Univ., Stanford, CA.

[46] FÜREDI, Z. and KOMLÓS, J. (1981). The eigenvalues of random symmetric matrices. *Combinatorica* **1** 233–241. MR0637828

[47] GAVISH, M. and DONOHO, D. L. (2014). The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Trans. Inform. Theory* **60** 5040–5053. MR3245370

[48] GORMLEY, I. C. and MURPHY, T. B. (2008). Exploring voting blocs within the Irish electorate: A mixture modeling approach. *J. Amer. Statist. Assoc.* **103** 1014–1027. MR2528824

[49] GORMLEY, I. C. and MURPHY, T. B. (2009). A grade of membership model for rank data. *Bayesian Anal.* **4** 265–295. MR2507364

[50] GÖRÜR, D., JÄKEL, F. and RASMUSSEN, C. E. (2006). A choice model with infinitely many latent features. In *Proceedings of the* 23*rd Annual International Conference on Machine Learning* 361–368. ACM, New York.

[51] GRONE, R., JOHNSON, C. R., DE SÁ, E. M. and WOLKOWICZ, H. (1984). Positive definite completions of partial Hermitian matrices. *Linear Algebra Appl.* **58** 109–124. MR0739282

[52] GUIVER, J. and SNELSON, E. (2009). Bayesian inference for Plackett–Luce ranking models. In *Proceedings of the* 26*th Annual International Conference on Machine Learning* 377–384. ACM, New York.

[53] HANDCOCK, M. S., RAFTERY, A. E. and TANTRUM, J. M. (2007). Model-based clustering for social networks. *J. Roy. Statist. Soc. Ser. A* **170** 301–354. MR2364300

[54] HASTIE, T. and TIBSHIRANI, R. (1998). Classification by pairwise coupling. *Ann. Statist.* **26** 451–471. MR1626055

[55] HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262

[56] HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 109–137. MR0718088

[57] HOOVER, D. N. (1982). Row-column exchangeability and a generalized model for probability. In *Exchangeability in Probability and Statistics* (*Rome*, 1981) 281–291. North-Holland, Amsterdam. MR0675982

[58] HUANG, T.-K., WENG, R. C. and LIN, C.-J. (2006). Generalized Bradley–Terry models and multi-class probability estimates. *J. Mach. Learn. Res.* **7** 85–115. MR2274363

[59] HUNTER, D. R. (2004). MM algorithms for generalized Bradley–Terry models. *Ann. Statist.* **32** 384–406. MR2051012

[60] JAVANMARD, A. and MONTANARI, A. (2011). Localization from incomplete noisy distance measurements. In 2011 *IEEE International Symposium on Information Theory Proceedings* (*ISIT*) 1584–1588. IEEE, New York.

[61] JOHNSON, C. R. (1990). Matrix completion problems: A survey. In *Matrix Theory and Applications* (*Phoenix*, *AZ*, 1989). *Proc. Sympos. Appl. Math.* **40** 171–198. Amer. Math. Soc., Providence, RI. MR1059486

[62] KESHAVAN, R. H., MONTANARI, A. and OH, S. (2010). Matrix completion from noisy entries. *J. Mach. Learn. Res.* **11** 2057–2078. MR2678022

[63] KESHAVAN, R. H., MONTANARI, A. and OH, S. (2010). Matrix completion from a few entries. *IEEE Trans. Inform. Theory* **56** 2980–2998. MR2683452

[64] KOLTCHINSKII, V. (2011). Von Neumann entropy penalization and low-rank matrix estimation. *Ann. Statist.* **39** 2936–2973. MR3012397

[65] KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869

[66] LANGE, K., HUNTER, D. R. and YANG, I. (2000). Optimization transfer using surrogate objective functions. *J. Comput. Graph. Statist.* **9** 1–59. MR1819865

[67] LESKOVEC, J., LANG, K. J., DASGUPTA, A. and MAHONEY, M. W. (2008). Statistical properties of community structure in large social and information networks. In *Proceeding of the* 17*th International Conference on World Wide Web* 695–704. ACM, Beijing, China.

[68] LOVÁSZ, L. (2012). *Large Networks and Graph Limits*. *American Mathematical Society Colloquium Publications* **60**. Amer. Math. Soc., Providence, RI. MR3012035

[69] LOVÁSZ, L. and SZEGEDY, B. (2006). Limits of dense graph sequences. *J. Combin. Theory Ser. B* **96** 933–957. MR2274085

[70] LUBETZKY, E. and ZHAO, Y. (2012). On replica symmetry of large deviations in random graphs. Preprint. Available at arXiv:1210.7013.

[71] LUCE, R. D. (1959). *Individual Choice Behavior*: *A Theoretical Analysis*. Wiley, New York. MR0108411

[72] LUCE, R. D. (1977). The choice axiom after twenty years. *J. Math. Psych.* **15** 215–233. MR0462675

[73] MAZUMDER, R., HASTIE, T. and TIBSHIRANI, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* **11** 2287–2322. MR2719857

[74] MOSSEL, E., NEEMAN, J. and SLY, A. (2012). Stochastic block models and reconstruction. Preprint. Available at arXiv:1202.1499.

[75] NADAKUDITI, R. R. (2013). OptShrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. Preprint. Available at arXiv:1306.6042.

[76] NEGAHBAN, S. and WAINWRIGHT, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39** 1069–1097. MR2816348

[77] NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.* **96** 1077–1087. MR1947255

[78] OH, S., MONTANARI, A. and KARBASI, A. (2010). Sensor network localization from local connectivity: Performance analysis for the MDS-MAP algorithm. In *Information Theory Workshop* (*ITW*) 1–5. IEEE, New York.

[79] OLIVEIRA, R. I. (2009). Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. Preprint. Available at arXiv:0911.0600.

[80] PLACKETT, R. L. (1975). The analysis of permutations. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **24** 193–202. MR0391338

[81] RADIN, C. and YIN, M. (2013). Phase transitions in exponential random graphs. *Ann. Appl. Probab.* **23** 2458–2471. MR3127941

[82] RAO, P. V. and KUPPER, L. L. (1967). Ties in paired-comparison experiments: A generalization of the Bradley–Terry model. *J. Amer. Statist. Assoc.* **62** 194–204. MR0217963

[83] RENNIE, J. D. and SREBRO, N. (2005). Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the* 22*nd International Conference on Machine Learning* 713–719. ACM, New York.

[84] ROHDE, A. and TSYBAKOV, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39** 887–930. MR2816342

[85] ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856

[86] ROHE, K., QIN, T. and FAN, H. (2012). The highest dimensional Stochastic Blockmodel with a regularized estimator. Preprint. Available at arXiv:1206.2380.

[87] RUDELSON, M. and VERSHYNIN, R. (2007). Sampling from large matrices: An approach through geometric functional analysis. *J. ACM* **54** Art. 21, 19 pp. (electronic). MR2351844

[88] SIMONS, G. and YAO, Y.-C. (1999). Asymptotics when the number of parameters tends to infinity in the Bradley–Terry model for paired comparisons. *Ann. Statist.* **27** 1041–1060. MR1724040

[89] SINGER, A. (2008). A remark on global positioning from local distances. *Proc. Natl. Acad. Sci. USA* **105** 9507–9511. MR2430205

[90] SINGER, A. and CUCURINGU, M. (2009/10). Uniqueness of low-rank matrix completion by rigidity theory. *SIAM J. Matrix Anal. Appl.* **31** 1621–1641. MR2595541

[91] SNIJDERS, T. A. B. and NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification* **14** 75–100. MR1449742

[92] SPENCE, I. and DOMONEY, D. (1974). Single subject incomplete designs for nonmetric multidimensional scaling. *Psychometrika* **39** 469–490.

[93] TALAGRAND, M. (1996). A new look at independence. *Ann. Probab.* **24** 1–34. MR1387624

[94] VU, V. H. (2007). Spectral norm of random matrices. *Combinatorica* **27** 721–736. MR2384414

[95] WOLFE, P. J. and OLHEDE, S. C. (2013). Nonparametric graphon estimation. Preprint. Available at arXiv:1309.5936.

[96] YANG, J. J., HAN, Q. and AIROLDI, E. M. (2014). Nonparametric estimation and testing of exchangeable graph models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*. *Journal of Machine Learning Research*, *Conference and Workshop Proceedings*, *Vol.* 33 1060–1067.

[97] ZERMELO, E. (1929). Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Math. Z.* **29** 436–460. MR1545015

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: souravc@stanford.edu