The Annals of Statistics
2020, Vol. 48, No. 2, 1072–1097
https://doi.org/10.1214/19-AOS1838
© Institute of Mathematical Statistics, 2020

WORST-CASE VERSUS AVERAGE-CASE DESIGN FOR ESTIMATION FROM PARTIAL PAIRWISE COMPARISONS

BY ASHWIN PANANJADY[1,*], CHENG MAO[2], VIDYA MUTHUKUMAR[1,**], MARTIN J. WAINWRIGHT[1,†] AND THOMAS A. COURTADE[1,‡]

[1] *Departments of EECS and Statistics, University of California Berkeley,* [*] *ashwinpm@berkeley.edu;* [**] *vidya.muthukumar@berkeley.edu;* [†] *wainwrig@berkeley.edu;* [‡] *courtade@berkeley.edu*
[2] *Department of Statistics and Data Science, Yale University,* *cheng.mao@yale.edu*

Pairwise comparison data arises in many domains, including tournament rankings, web search and preference elicitation. Given noisy comparisons of a fixed subset of pairs of items, we study the problem of estimating the underlying comparison probabilities under the assumption of strong stochastic transitivity (SST). We also consider the noisy sorting subclass of the SST model. We show that when the assignment of items to the topology is arbitrary, these permutation-based models, unlike their parametric counterparts, do not admit consistent estimation for most comparison topologies used in practice. We then demonstrate that consistent estimation is possible when the assignment of items to the topology is randomized, thus establishing a dichotomy between worst-case and average-case designs. We propose two computationally efficient estimators in the average-case setting and analyze their risk, showing that it depends on the comparison topology only through the degree sequence of the topology. We also provide explicit classes of graphs for which the rates achieved by these estimators are optimal. Our results are corroborated by simulations on multiple comparison topologies.

1. Introduction. The problems of ranking and estimation from ordinal data arise in many applications, including web search and information retrieval [15], crowdsourcing [11], tournament play [22], social choice theory [7] and recommender systems [2]. The ubiquity of such datasets stems from the relative ease with which ordinal data can be obtained, and from the empirical observation that using pairwise comparisons as a means of data elicitation can lower the noise level in the observations [4, 52].

Given that the number of items n to be compared can be very large, it is often difficult or impossible to obtain comparisons between all $\binom{n}{2}$ pairs of items. A subset of pairs to compare, which defines the *comparison topology*, must therefore be chosen. Fixed comparison topologies can arise in problem domains where there are artificial constraints imposed on the comparisons that can be made. For example, in league stage tournament formats in sports, passive pairwise comparisons are used, and artificial constraints on comparisons exist; for example, there are more games played within a conference than between conferences in all four major sports leagues in the United States. Different comparison topologies can also arise as a property of the elicitation process. For instance, in movie rating or recommender systems, pairwise comparisons between items of the same genre are typically more abundant than comparisons between items of dissimilar genres [35, 51]; as a specific example, consider the UT Zappos online shopping dataset [56], in which the comparison topology has different intercategory and intra-category edge densities. Another example is the increasing use of ordinal data in peer-grading in massive open online courses [49]: here, comparisons

Received October 2017; revised October 2018.
MSC2010 subject classifications. 62J15, 62F07.
Key words and phrases. Pairwise comparisons, strong stochastic transitivity, structured matrix completion.

are passive since all students must be graded at once, and not necessarily made at random due to preexisting biases in selecting graders [41].

Fixed comparison topologies are also important in rank breaking [19, 26], and more generally in matrix completion based on structured observations [27, 42]. For these reasons, it is of interest to study statistical properties of ranking models and performance of ranking algorithms based on fixed comparison topologies.

An important problem in ranking is the design of accurate models for capturing uncertainty in pairwise comparisons. Given a collection of n items, the outcomes of pairwise comparisons are completely characterized by the $n \times n$ matrix of comparison probabilities; a comparison probability refers to the probability that item i beats item j in a comparison between them. Various models have been proposed for such matrices. The most classical models, among them the Bradley–Terry–Luce [5, 28] and Thurstone models [53], assign a quality vector to the set of items, and assign pairwise probabilities by applying a cumulative distribution function to the difference of qualities associated to the pair. There is now a relatively large body of work on methods for ranking in such parametric models (e.g., see the papers [13, 19, 35, 45] as well as references therein). By contrast, relatively less attention has been paid to a richer class of models proposed decades ago in the sociology literature [16, 34], which impose a milder set of constraints on the pairwise comparison matrix. Rather than positing a quality vector, these models impose constraints that are typically given in terms of a latent permutation that rearranges the matrix into a specified form, and hence can be referred to as *permutation-based* models. Two such models that have been recently analyzed are those of strong stochastic transitivity [46], as well as the special case of noisy sorting [6]. The strong stochastic transitivity (SST) model, in particular, has been shown to offer significant robustness guarantees and to provide a good fit to many existing datasets [1], and this flexibility has driven recent interest in understanding its statistical properties. Also, perhaps surprisingly, past work has shown that this additional flexibility comes at only a small price when one has access to all possible pairwise comparisons, or more generally, to comparisons chosen at random [46]; in particular, the rates of estimation in these SST models differ from those in parametric models by only logarithmic factors in the number of items. On a related note, permutation-based models have also recently been shown to be useful in other settings like crowd-labeling [48], statistical seriation [17] and to a variety of regression problems [39, 44].

Given pairwise comparison data from one of these models, the problem of estimating the comparison probabilities has applications in inferring customer preferences in recommender systems, advertisement placement and sports, and is the main focus of this paper.

Our contributions. Our goal is to estimate the matrix of comparison probabilities for fixed comparison topologies, studying both the noisy sorting and SST classes of matrices. Focusing first on the worst-case setting in which the assignment of items to the topology may be arbitrary, we show in Theorem 1 that consistent estimation is impossible for many natural comparison topologies. This result stands in sharp contrast to parametric models, and may be interpreted as a "no free lunch" theorem: although it is possible to estimate SST models at rates comparable to parametric models when given a full set of observations [46], the setting of fixed comparison topologies is problematic for the SST class. This can be viewed as a price to be paid for the additional robustness afforded by the SST model.

Seeing as such a worst-case design may be too strong for permutation-based models, we turn to an average-case setting in which the items are assigned to a fixed graph topology in a randomized fashion. Under such an observation model, we propose and analyze two efficient estimators: Theorems 2 and 4 show that consistent estimation is possible under commonly used comparison topologies; it is important to note that Theorem 4 for the SST class holds for a slightly different notion of average case design. Moreover, the error rates of these estimators depend only on the degree sequence of the comparison topology, and are shown to be unimprovable for a class of graphs that we explicitly characterize in Theorem 3.

Our results therefore establish a sharp distinction between average-case and worst-case design when using fixed comparison topologies in permutation-based models. Such a phenomenon arises from the difference between the Bayes risk under a uniform prior on the ranking versus the minimax risk, and may also be worth studying for other ranking models.

Related work. The literature on ranking and estimation from pairwise comparisons is vast, and we refer the reader to some surveys [8, 18, 32] and references therein for a more detailed overview. Estimation from pairwise comparisons has been analyzed under various metrics like top-k ranking [12, 13, 24, 50] and comparison probability or parameter estimation [19, 45, 46]. There have been studies of these problems under active [23, 33], passive [35, 43] and collaborative settings [36, 40], and also for fixed as well as random comparison topologies [46, 55]. Here, we focus on the subset of papers that are most relevant to the work described here.

The problem of comparison probability estimation has been analyzed for both nonparametric and parametric models. In the former case, Braverman and Mossel [6] propose an algorithm the noisy sorting model that runs in polynomial-time, and show that it is minimax optimal for the complete graph topology. Mao et al. [31], in work carried out after the submission of this manuscript, provide a more efficient algorithm that is also minimax optimal for the random comparison topology. For the SST model class, Shah et al. [46] analyze the case where the comparison topology is random, and provide minimax rates and efficient algorithms for this setting. Two salient features of this work are most relevant to to our development. First, as mentioned before, the minimax rates of estimation for the random comparison topology differ from those for the smaller class of parametric models only by logarithmic factors in the number of items. Second, the computationally efficient algorithms proposed for the random comparison topology [10, 46, 47] are not minimax optimal for this setting, and this has remained a challenging open problem that has only recently seen some progress [29, 30]. It is important to state here that while obtaining minimax optimal and efficient algorithms for the random graph topology is acknowledged to be a challenging problem, the focus of this paper is on studying general comparison topologies.

Comparison probability estimation has also been analyzed under a fixed topology for parametric models by Hajek et al. [19] and Shah et al. [45]. Both papers analyze the worst-case design setting in which the assignment of items to the topology may be arbitrary, and derive bounds on the minimax risk of parameter (or equivalently, comparison probability) estimation. While their characterizations are not sharp in general, the rates are shown to depend on the spectrum of the Laplacian matrix of the topology. We point out an interesting consequence of both results: in the parametric model, provided that the comparison graph G is connected, the maximum likelihood solution, in the limit of infinite samples for each graph edge, allows for exact recovery of the quality vector, and hence the matrix of comparison probabilities. We will see that this property no longer holds for the SST models considered in this paper: there are comparison topologies and SST matrices for which it is impossible to recover the full matrix even given an infinite amount of data per graph edge. It is also worth mentioning that the top-k ranking problem has been analyzed for parametric models under fixed design assumptions [24], and here as well, asymptotic consistency is observed for connected comparison topologies.

Notation. Here, we summarize some notation used throughout the remainder of this paper. We use n to denote the number of items, and adopt the shorthand $[n] := \{1, 2, \ldots, n\}$.

We use $\mathrm{Ber}(p)$ to denote a Bernoulli random variable with success probability p. For two sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n \lesssim b_n$ if there is a universal constant C such that $a_n \leq Cb_n$ for all $n \geq 1$. The relation $a_n \gtrsim b_n$ is defined analogously, and we write $a_n \asymp b_n$ if the relations $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold simultaneously. We use c, c_1, c_2 to denote universal constants that may change from line to line. Given a matrix $M \in \mathbb{R}^{n\times n}$, its ith row is denoted by M_i. For a graph G with edge set E, let $M(G)$ denote the entries of the matrix M restricted to the edge set of G, and let $\|M\|_E^2 = \sum_{(i,j)\in E} M_{ij}^2$. For a matrix $M \in \mathbb{R}^{n\times n}$ and a permutation $\pi : [n] \to [n]$, we use the shorthand $\pi(M) = \Pi M \Pi^\top$, where Π represents the row permutation matrix corresponding to the permutation π. We let id denote the identity permutation. The Kendall tau distance [25] between permutations π and π' is given by $\mathsf{KT}(\pi, \pi') := \sum_{i,j\in[n]} \mathbf{1}\{\pi(i) < \pi(j), \pi'(i) > \pi'(j)\}$.

We let $\alpha(G)$ denote the size of the largest independent set of a graph G; note that an independent set is defined as any subset of vertices that have no edges among them. Define a biclique of a graph as two disjoint subsets of its vertices V_1 and V_2 such that $(u, v) \in E(G)$ for all $u \in V_1$ and $v \in V_2$. Define the biclique number $\beta(G)$ as the maximum number of edges in any such biclique, given by $\max_{V_1,V_2 \text{ biclique}} |V_1||V_2|$. Let d_v denote the degree of vertex v.

2. Background and problem formulation.

Consider a collection of $n \geq 2$ items that obey a total ordering or ranking determined by a permutation $\pi^* : [n] \to [n]$. More precisely, item $i \in [n]$ is preferred to item $j \in [n]$ in the underlying ranking if and only if $\pi^*(i) < \pi^*(j)$. We are interested in observations arising from stochastic pairwise comparisons between items. We denote the matrix of underlying comparison probabilities by $M^* \in [0, 1]^{n\times n}$, with $M^*_{ij} = \Pr\{i \succ j\}$ representing the probability that item i beats item j in a comparison; we set $M^*_{ii} = 1/2$ by convention.

Each item i is associated with a *score* τ_i^* that measures the probability that item i beats another item chosen uniformly at random, that is,

$$\tau_i^* := [\tau^*(M^*)]_i := \frac{1}{n-1}\sum_{j\neq i} M^*_{ij}. \tag{1}$$

Arranging the scores in descending order naturally yields a ranking of items. In fact, for the models we define below, the ranking given by the scores is consistent with the ranking given by π^*, that is, $\tau_i^* \geq \tau_j^*$ if $\pi^*(i) < \pi^*(j)$. The converse also holds if the scores are distinct.

2.1. *Pairwise comparison models.*

Let us consider a permutation-based model for the comparison matrix M^* defined by the property of *strong stochastic transitivity* [16, 34], or the SST property for short. In particular, a matrix M^* of pairwise comparison probabilities is said to obey the SST property if for items i, j and k in the total ordering such that $\pi^*(i) < \pi^*(j) < \pi^*(k)$, it holds that $\Pr(i \succ k) \geq \Pr(i \succ j)$. Recalling that $\pi(M)$ denotes the matrix obtained from M by permuting its rows and columns according to the permutation π, the SST matrix class can be equivalently defined in terms of permutations applied to the class $\mathbb{C}_{\mathsf{BISO}}$ of bivariate isotonic matrices as

$$\mathbb{C}_{\mathsf{SST}} := \bigcup_{\pi} \pi(\mathbb{C}_{\mathsf{BISO}}) = \bigcup_{\pi} \{\pi(M) : M \in \mathbb{C}_{\mathsf{BISO}}\}. \tag{2}$$

Here, the class $\mathbb{C}_{\mathsf{BISO}}$ of bivariate isotonic matrices is given by

$$\{M \in [0, 1]^{n\times n} : M + M^\top = \mathbf{e}\mathbf{e}^\top \text{ and}$$
$$M \text{ has nondecreasing rows and nonincreasing columns}\},$$

where $\mathbf{e} \in \mathbb{R}^n$ denotes a vector of all ones.

As shown by Shah et al. [46], the SST class is substantially larger than commonly used class of *parametric* models, in which each item i is associated with a scalar parameter w_i, and the probability that item i beats item j is given by $F(w_i - w_j)$, where $F : \mathbb{R} \mapsto [0, 1]$ is smooth and monotone.

A special case of the SST model that we also study is the *noisy sorting* model [6], in which the all underlying probabilities are described with a single parameter $\lambda \in [0, 1/2]$. The matrix $M_{\mathsf{NS}}(\pi, \lambda) \in [0, 1]^{n\times n}$ has entries

$$[M_{\mathsf{NS}}(\pi, \lambda)]_{ij} = 1/2 + \lambda \cdot \mathsf{sgn}(\pi(j) - \pi(i)),$$

and the noisy sorting classes are given by

$$\mathbb{C}_{\mathsf{NS}}(\lambda) := \bigcup_{\pi} \{M_{\mathsf{NS}}(\pi, \lambda)\}, \quad \text{and} \quad \mathbb{C}_{\mathsf{NS}} := \bigcup_{\lambda \in [0,1/2]} \mathbb{C}_{\mathsf{NS}}(\lambda). \tag{3}$$

Here, $\mathsf{sgn}(x)$ is the sign operator, with the convention that $\mathsf{sgn}(0) = 0$. In words, the noisy sorting class models the case where the probability $\Pr\{i \succ j\}$ depends only on the parameter λ and whether $\pi^*(i) < \pi^*(j)$. Although the noisy sorting subclass is a very special case of the SST class, it cannot be represented by any parametric class with a smooth function F (apart from the degenerate case $\lambda^* = 1/2$), and so captures some essential difficulties of learning in the SST class.

We now turn to describing the observation models that we consider in this paper.

2.2. *Partial observation models.* Our goal is to provide guarantees on estimating the underlying comparison matrix M^* when the comparison topology is fixed. In other words, we would like to provide such a guarantee when we have access to a subset of pairwise comparisons. Given such a subset $\mathcal{E} \subseteq \binom{[n]}{2}$ of compared items, our observations obey the probabilistic model

$$Y_{ij} = \begin{cases} \mathrm{Ber}(M^*_{ij}) & \text{for } (i, j) \in \mathcal{E}, \text{ independently} \\ \star & \text{otherwise,} \end{cases} \tag{4}$$

where $\star$ indicates a missing observation. We set the diagonal entries of Y equal to $1/2$, and also specify that $Y_{ji} = 1 - Y_{ij}$ for $j > i$ where $(i, j) \in \mathcal{E}$. We consider two different instantiations of the set $\mathcal{E}$ given the comparison topology $G(V, E)$, which is a graph on n vertices.

2.2.1. *Worst-case setting.* In this setting, we assume that the assignment of items to vertices of the comparison graph G is arbitrary. In particular, we consider the set of fixed comparisons given by $\mathcal{E} = E$. Our goal is to provide uniform guarantees in the metric $\|\widehat{M} - M^*\|_F^2$ over all matrices M^* in our model class given this restricted set of observations.

This setting is of the worst-case type, since the adversary is allowed to choose the underlying matrix with knowledge of the edge set E. Providing guarantees against such an adversary is known to be possible for parametric models [19, 45]. However, as we show in Section 3.1, such a guarantee is impossible to obtain even over the the noisy sorting subclass of the full SST class. Consequently, the latter parts of our analysis apply to a less rigid, average-case setting.

2.2.2. *Average-case setting.* In this setting, we assume that the assignment of items to vertices of the comparison graph G is random. Consequently, the set of comparisons $\mathcal{E}$ is given by the edge set of the randomly relabeled graph. Recalling the notation $\pi(M)$ that we used to denote the matrix M with rows and column permuted by the permutation π, we can also describe our observations as follows. Given a fixed comparison graph G having adjacency matrix A, define the random observation matrix $\mathcal{O} = \sigma(A)$ for a permutation

$\sigma : [n] \to [n]$ chosen uniformly at random. Then the set of observed entries is given by $\mathcal{E} = \{(i, j) : i > j \text{ and} \mathcal{O}_{ij} = 1\}$. Such a setting is reasonable when the graph topology is constrained, but we are still given the freedom to assign items to vertices of the comparison graph.

Our guarantees in the one sample setting with the observation matrix $\mathcal{O}$ can be seen as a form of Bayes risk, where given a fixed observation pattern E (consisting of the entries of the comparison matrix Y determined by the adjacency matrix A of the graph G, with A_{ij} representing the indicator that entry Y_{ij} is observed), we want to estimate a matrix M^* under a uniform Bayesian prior on the ranking π^*. Studying this average-case setting is well motivated, since given fixed comparisons between a set of items, there is no reason to assume a priori that the underlying ranking is generated adversarially.

In this paper, we theoretically analyze an extension of such an observation model in which we have two random designs, consisting of two random observation matrices $\mathcal{O}^{(1)} = \sigma_1(A)$ and $\mathcal{O}^{(2)} = \sigma_2(A)$, chosen with independent, random permutations σ_1 and σ_2, and where the set of observed comparisons is given by their union $\mathcal{E} = \{(i, j) : i > j \text{ and } \mathcal{O}^{(1)}_{ij} + \mathcal{O}^{(2)}_{ij} \geq 1\}$. Our main motivation for studying such an observation model is theoretical: having two random designs allows us to split our samples in the most natural way, and simplifies the analysis of one of our algorithms by exploiting the statistical independence between the observations used in distinct steps.[1] Having said this, our proposed estimator can be implemented for one random observation as well, and we show empirically that the estimation error rates are very similar.

We are now ready to state the goal of the paper. We address the problems of recovering the ranking π^* and estimating the matrix M^* in the Frobenius norm. More precisely, given the observation matrix $Y = Y(E)$ (where the set E is random in the average-case observation model), we would like to output a matrix $\widehat{M}$ that is a function of Y, and for which good control on the squared Frobenius norm error $\|\widehat{M} - M^*\|_F^2$ can be guaranteed.

3. Main results. In this section, we state our main results and discuss some of their consequences. Proofs are deferred to Section 5.

3.1. *Worst-case design*: *Minimax bounds*. In the worst-case setting of Section 2.2.1, the performance of an estimator is measured in terms of the normalized minimax error

$$\mathcal{M}(G, \mathbb{C}) = \inf_{\widehat{M} = f(Y(G))} \sup_{M^* \in \mathbb{C}} \mathbb{E}\left[\frac{1}{n^2} \|\widehat{M} - M^*\|_F^2\right],$$

where the expectation is taken over the randomness in the observations Y as well as any randomness in the estimator, and $\mathbb{C} \in \{\mathbb{C}_{\mathsf{SST}}, \mathbb{C}_{\mathsf{NS}}\}$ represents the model class. Our first result shows that for many comparison topologies, the minimax risk is prohibitively large even for the noisy sorting model.

THEOREM 1. *For any graph G*, *the diameter of the set consistent with observations on the edges of G is lower bounded as*

$$\sup_{\substack{M_1, M_2 \in \mathbb{C}_{\mathsf{NS}} \\ M_1(G) = M_2(G)}} \|M_1 - M_2\|_F^2 \geq \alpha(G)\big(\alpha(G) - 1\big) \vee \beta\big(G^c\big). \tag{5a}$$

Consequently, *the minimax risk of the noisy sorting model is lower bounded as*

$$\mathcal{M}(G, \mathbb{C}_{\mathsf{NS}}) \gtrsim \frac{1}{n^2}\big[\alpha(G)\big(\alpha(G) - 1\big) \vee \beta\big(G^c\big) \vee n\big]. \tag{5b}$$

[1] For the special case of the noisy sorting model, we obtain our theoretical result for one random observation—we handle the dependence between the two steps of the algorithm directly.

Note that via the inclusion $\mathbb{C}_{\mathsf{NS}} \subset \mathbb{C}_{\mathsf{SST}}$, Theorem 1 also implies the same lower bound (5b) on the risk $\mathcal{M}(G, \mathbb{C}_{\mathsf{SST}})$. In addition to these bounds, the lower bounds for estimation in parametric models, known from past work [45], carry over directly to the SST model, since parametric models are subclasses of the SST class.

Inequality (5a) of Theorem 1 is approximation-theoretic in nature: it is a statement purely about the size of the set of matrices consistent with observations on the graph. It may be interpreted as the worst-case error in the infinite sample limit of observations on G: for graphs with large independent sets or complement bicliques, it is impossible to distinguish two matrices that are separated in the Frobenius norm from observations just on the edges of the graph. The intuition behind the result is illustrated by a simple example: take an n-vertex star graph with the central vertex ranked highest in the underlying ordering. In the noisy sorting model, an infinite number of samples effectively allows us to estimate (exactly) the probability with which the top item beats the other items, and hence the parameter λ. However, we obtain no information about the partial orders between items ranked second through last, only that the central item is ranked first. In contrast, for parametric models (which are a special subset of permutation-based models), knowing the exact probabilities with which the top item beats the remaining items allows us to then infer the associated quality vector up to a global shift, since this amounts to inverting a well-behaved generalized linear model (see the papers [19, 45] for details). The intuition from the star graph can be extended to show that connected graphs with large independent sets or with large bicliques in their complements (e.g., complete bipartite and barbell graphs) result in large worst-case approximation errors. In particular, when we have $\alpha(G) \vee \beta(G) \gtrsim n^2$, the lower bound evaluates to an order 1 quantity, and a matching upper bound is trivially achieved by any reasonable estimator.

Since inequality (5a) is purely an approximation-theoretic statement, it does not capture the uncertainty due to noise, and thus can be a loose characterization of the minimax risk for some graphs, with the complete graph being one example. The minimax risk bound (5b) combines the aforementioned approximation theoretic bound with a lower bound due to noisy observations on the complete graph, which is of the order $1/n$. For graphs with small independent sets and bicliques, the estimation-theoretic lower bound is larger, and sharp in the case of the complete graph topology.

In Section 4, we specialize Theorem 1 to some commonly used graph topologies, showing that for many of them, the approximation error is lower bounded by a constant even for graphs admitting consistent parametric estimation. Seeing as the minimax error in the worst-case setting can be prohibitively large, we now turn to evaluating practical estimators in the random observation models of Section 2.2.2.

3.2. *Average-case design.*

We first present our results for noisy sorting matrix estimation, and then for SST matrix estimation.

3.2.1. *Noisy sorting matrix estimation.*

In the average-case setting described in Section 2.2.2, we measure the performance of an estimator using the risk

$$\sup_{M^* \in \mathbb{C}} \mathbb{E}_{\mathcal{O},Y} \frac{1}{n^2} \| \widehat{M} - M^* \|_F^2.$$

It is important to note that the expectation is taken over both the comparison noise, as well as the random observation pattern $\mathcal{O}$ (or equivalently, the underlying random permutation σ assigning items to vertices). We propose the Average-Sort-Project estimator (ASP for short) for matrix estimation in this metric, which is a natural generalization of the Borda count estimator [10, 47]. It consists of three steps, described below for the noisy sorting model:

(I) *Averaging step:* Compute the average $\widehat{\tau}_i = [\widehat{\tau}(Y)]_i = \frac{\sum_{j \neq i} Y_{ij} \mathcal{O}_{ij}}{\sum_{j \neq i} \mathcal{O}_{ij}}$, corresponding to the fraction of comparisons won by item i.

(II) *Sorting step:* Choose the permutation $\widehat{\pi}_{\mathsf{ASP}}$ such that the sequence $\{\widehat{\tau}_{\widehat{\pi}_{\mathsf{ASP}}^{-1}(i)}\}_{i=1}^n$ is decreasing in i, with ties broken arbitrarily.

(III) *Projection step:* Find the maximum likelihood estimate $\widehat{\lambda}$ by treating $\widehat{\pi}_{\mathsf{ASP}}$ as the true permutation that sorts items in decreasing order. Output the matrix $\widehat{M}_{\mathsf{ASP}} := M_{\mathsf{NS}}(\widehat{\pi}_{\mathsf{ASP}}, \widehat{\lambda})$.

The average-sort-project algorithm, like a broad class of other algorithms in this space, decouples the estimation of the unknown permutation and the unknown parameter λ. As mentioned in Section 2, the true score vector τ^* is consistent with the true permutation π^*; step (I) simply computes an unbiased estimate $\widehat{\tau}$ of the true scores, and treating this as the underlying truth, step (II) computes a permutation estimate $\widehat{\pi}_{\mathsf{ASP}}$. This should be intuitive, since items that are ranked higher should win a larger fraction of their comparisons than lower-ranked items. In step (III), we focus on estimating the unknown parameter λ, treating our estimate of the permutation from step (2) as the ground truth. Specifically, we assume that when viewed along the permutation $\widehat{\pi}_{\mathsf{ASP}}$ on both dimensions, the underlying matrix has constant upper and lower triangular portions. Thus, the MLE of the quantity $1/2 + \lambda^*$ is given by averaging the upper triangular portion of the sorted matrix. The estimator runs in time $O(n^2)$, which is linear in the input size.

We now state an upper bound on the mean-squared Frobenius error achievable using the ASP estimator. It involves the degree sequence $\{d_v\}_{v \in V}$ of a graph G without isolated vertices, meaning that $d_v \geq 1$ for all $v \in V$. Also recall our notation $\mathcal{O}$ to denote the random observation process.

THEOREM 2. *For any graph $G = (V, E)$ without isolated vertices and any matrix $M^* \in \mathbb{C}_{\mathsf{NS}}(\lambda^*)$, we have*

$$\mathbb{E}_{\mathcal{O},Y}\left[\frac{1}{n^2}\|\widehat{M}_{\mathsf{ASP}} - M^*\|_F^2\right] \lesssim \frac{1}{|E|} + \frac{n \log n}{|E|^2} + \frac{\lambda^*}{n} \sum_{v \in V} \frac{1}{\sqrt{d_v}}, \quad \textit{and} \tag{6a}$$

$$\mathbb{E}_{\mathcal{O},Y}\left[\mathsf{KT}(\pi^*, \widehat{\pi}_{\mathsf{ASP}})\right] \lesssim \frac{n}{\lambda^*} \sum_{v \in V} \frac{1}{\sqrt{d_v}}. \tag{6b}$$

A few comments are in order. First, while the results are stated in expectation, a high probability bound can be proved for permutation estimation—namely

$$\Pr_{\mathcal{O},Y}\left\{\mathsf{KT}(\pi^*, \widehat{\pi}_{\mathsf{ASP}}) \gtrsim \frac{n\sqrt{\log n}}{\lambda^*} \sum_{v \in V} \frac{1}{\sqrt{d_v}}\right\} \leq n^{-10}.$$

Second, it can be verified that $\frac{1}{|E|} + \frac{n \log n}{|E|^2} \lesssim \frac{1}{n} \sum_{v \in V} \frac{1}{\sqrt{d_v}}$, so that taking a supremum over the parameter $\lambda^* \in [0, 1/2]$ guarantees that the mean-squared Frobenius error is upper bounded as $O(\frac{1}{n} \sum_{v \in V} \frac{1}{\sqrt{d_v}})$, uniformly over the entire noisy sorting class $\mathbb{C}_{\mathsf{NS}}$. Third, it is also interesting to note the dependence of the bounds on the noise parameter λ^* of the noisy sorting model. The "high-noise" regime $\lambda^* \approx 0$ is a good one for estimating the underlying matrix, since the true matrix M^* is largely unaffected by errors in estimating the true permutation. However, as captured by equation (6b), the permutation estimation problem is more challenging in this regime.

Let us also provide some intuition for the terms appearing in bound (6a). Ignoring the second term for the moment, which should be thought of as a slack term that allows the application of union bounds, the first and third terms have an operational interpretation. As will be made clear in the proof, the Frobenius error decomposes into two terms: the first term $1/|E|$ represents the error due to estimation of the scalar parameter λ, since we have an effective sample size $|E|$. The quantity $\frac{1}{n}\sum_{v\in V}\frac{1}{\sqrt{d_v}}$ should be thought of as the average deviation of the scores τ_i^*, and enters our bounds as a proxy for the error of permutation estimation.

The bound (6a) can be specialized to the complete graph K_n and the Erdős–Rényi random graph with edge probability p to obtain the rates $1/\sqrt{n}$ and $1/\sqrt{np}$, respectively, for estimation in the mean-squared Frobenius norm. While we eliminate logarithmic factors that were present in previous work [46], these rates are strictly suboptimal for these graphs, since the minimax rates scale as $1/n$ and $1/(np)$, respectively; both are achieved by the global MLE [46]. This is unsurprising, since the ASP estimator is based on Borda counts, which provide permutation estimates that are too noisy for minimax-optimal estimation in the full observation case.[2] It is important to note, however, that the MLE for estimating permutation-based models is computationally intractable. Until very recently, Borda count algorithms (along with their spectral counterparts) were the best performing polynomial time algorithms uniformly over the SST class, even for the special case of full observations. Recently, work by a subset of the current authors [30] has shown that the Borda counts can be carefully denoised to obtain faster rates of the order $n^{-3/4}$.

Interestingly, we can show that Borda count algorithms are, in fact, minimax optimal in the *average-case, fixed design setting* for certain comparison topologies. This is because some topologies—for instance, those composed of disjoint cliques—provide very little information about many partial orders. Although refining the count estimate within the cliques has its benefits, the error is driven by permutation estimation *across* the cliques, which is information-theoretically limited (see Theorem 3). The Borda counts are thus good enough to achieve the globally minimax-optimal rate for these comparison topologies.

We require some additional notation to state this precisely. Fix constants $C_1 = 10^{-2}$ and $C_2 = 10^2$ and two sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$ of (strictly) positive scalars. For each $n \geq 1$, define the family of graphs

$$\mathcal{G}_n(a_n, b_n) := \Biggl\{G(V, E) \text{ is connected} : |V| = n, C_1 a_n \leq |E| \leq C_2 a_n, \text{ and} \\ C_1 b_n \leq \sum_{v\in V}\frac{1}{\sqrt{d_v}} \leq C_2 b_n\Biggr\}.$$

As noted in Section 2.2.2, the average-case design observation model is equivalent to choosing the matrix M^* from a random ensemble with the permutation π^* chosen uniformly at random, and observing fixed pairwise comparisons. Such a viewpoint is useful in order to state our lower bound. Expectations are taken over the randomness of both π^* and the Bernoulli observation noise.

THEOREM 3.

(a) *Let $M^* = M_{\mathsf{NS}}(\pi^*, 1/4)$, where the permutation π^* is chosen uniformly at random on n items. For any pair of sequences of positive numbers $(\{a_n\}_{n\geq 1}, \{b_n\}_{n\geq 1})$ such that the set $\mathcal{G}_n(a_n, b_n)$ is nonempty for every $n \geq 1$, and for any estimators $(\widehat{M}, \widehat{\pi})$ that are measurable*

[2] Interestingly, in the parametric case, Borda count estimators similar to the ASP estimator are still minimax-optimal (up to log factors) in estimating *the comparison matrix itself* with full observations (but suboptimal in estimating the parameters underlying the model). This phenomenon is due to the fact that parametric models afford smoothness that is not present in the more general SST case; see Chatterjee and Mukherjee [10] for more precise statements of these claims.

functions of the observations on G, we have

$$\sup_{G \in \mathcal{G}_n(a_n,b_n)} \mathbb{E}\left[\frac{1}{n^2}\|\widehat{M} - M^*\|_F^2\right] \gtrsim \frac{b_n}{n}, \quad \textit{and} \quad \sup_{G \in \mathcal{G}_n(a_n,b_n)} \mathbb{E}\big[\mathsf{KT}(\pi^*, \widehat{\pi})\big] \gtrsim n b_n.$$

(b) *For any graph G, let $M^* = M_{\mathsf{NS}}(\pi^*, c\sqrt{n/|E|})$, with the permutation π^* chosen uniformly at random and the constant c chosen sufficiently small. Then for any estimators $(\widehat{M}, \widehat{\pi})$ that are measurable functions of the observations on G, we have*

$$\mathbb{E}\left[\frac{1}{n^2}\|\widehat{M} - M^*\|_F^2\right] \gtrsim \frac{n}{|E|}.$$

Parts (a) and (b) of the lower bound may be interpreted respectively as the approximation error caused by having observations only on a subset of edges, and the estimation error arising from the Bernoulli observation noise. Note that part (b) applies to every graph, and is particularly noteworthy for sparse graphs. In particular, in the regime in which the graph has bounded average degree, it shows that the inconsistency exhibited by the ASP estimator is unavoidable for any estimator. A more detailed discussion for specific graphs may be found in Section 4.

Although part (a) of the theorem is stated for a supremum over graphs, we actually prove a stronger result that explicitly characterizes the class of graphs that attain these lower bounds. As an example, given the sequences $a_n = n^2$ and $b_n = \sqrt{n}$, we show that the ASP estimator is information-theoretically optimal for the sequence of graphs consisting of two disjoint cliques $K_{n/2} \cup K_{n/2}$, which can be verified to lie within the class $\mathcal{G}(a_n, b_n)$. As alluded to before, our lower bound applies more generally to graphs in which the comparison topology provides very little information about many partial orders, so that it is difficult to estimate these partial orders better than with a random guess. The ASP algorithm is therefore able to attain the minimax-optimal rate for these comparison topologies.

Having addressed the noisy sorting model, a natural question is whether a variant of the ASP estimator is applicable even to matrix estimation under the SST model. In particular, the ASP estimator for the SST model would replace step (III), as stated, by a maximum likelihood estimate over bivariate isotonic matrices using the entries on the edges that we observe. While such an estimator is reasonable to implement, analyzing it is a challenging problem due to structural dependencies introduced by the comparison topology. In order to manage these dependencies, we establish some structural properties of SST matrices, and leverage these to propose a different estimator that exploits these properties to attain provable rates of estimation.

3.2.2. *SST matrix estimation.*

For the more general SST model, the same permutation estimator $\widehat{\pi}_{\mathsf{ASP}}$ in fact achieves the optimal rate for estimating the true ranking with error measured by $\|\widehat{\pi}_{\mathsf{ASP}}(M^*) - M^*\|_F^2$, as stated in Theorem 4 below. However, estimating the underlying matrix of probabilities M^* requires more effort.

For SST matrix estimation, we rely on a structural property of SST matrices that we establish. In particular, we show (in Lemma 6) that every SST matrix is close in the squared Frobenius norm to a matrix that is constant on suitably chosen blocks. Consequently, we relate the problem of SST matrix estimation to estimating these blockwise constant matrices, thus drastically reducing the number of parameters to be estimated. Such a blockwise approximation can be chosen in many ways: our choice is guided by the goal of balancing the estimation error of the aforementioned problem and the approximation error between the SST matrix and its blockwise approximation. Our algorithm relies on a certain block-averaging subroutine, which we describe next.

Block-average subroutine. Given a partially observed and noisy SST matrix, the goal of this subroutine is to output a blockwise constant matrix estimate that is close to the true matrix, so that the approximation error is minimized. In order to do so, we apply two natural steps. First, we identify the configuration of the blocks on the matrix by the following simple intuition: grouping "similar" items into the same block should result in small approximation error, and a good indicator of similarity between items is the similarity between their scores. Once such a configuration of blocks has been obtained, we then average our noisy observations over these blocks, thus controlling the estimation error of the problem.

Let us now make both of these steps precise. For any vector $v \in [0, 1]^n$, fix some value $t \in (0, 1)$ and define a block partition $\mathsf{bl}_t(v)$ of v as

$$[\mathsf{bl}_t(v)]_i = \{j \in [n] : v_j \in [(i-1)t, it)\}. \tag{7}$$

We are particularly interested in obtaining such a block partition from score estimates $\widehat{\tau}(Y)$ of a partially observed matrix Y. In this case, the blocking vector $\mathsf{bl}_t(\widehat{\tau}(Y))$ contains a partition of indices such that the row sums of the matrix within each block of the partition differ by at most a scalar t.

Any partition bl of the index set $[n]$ induces a natural partition of the set $[n] \times [n]$ partitioned according to bl along both dimensions. Let us denote such a partition by $\mathcal{B}_{\mathsf{bl}}$. For every pair $(i, j) \in [n] \times [n]$, we denote the *block* in the partition $\mathcal{B}_{\mathsf{bl}}$ that contains the index-pair (i, j) by $B_{\mathsf{bl}}(i, j) \subseteq [n] \times [n]$. Note that by definition, we have the inclusion $B_{\mathsf{bl}}(i, j) \in \mathcal{B}_{\mathsf{bl}}$. Now given a partially observed matrix $Y' \in [0, 1]^{n \times n}$, we use $E \subseteq [n] \times [n]$ to denote the (symmetric) set of edges where we observe entries of Y', and define the blocked version of Y' on the set E as

$$[\mathsf{B}(Y', \mathsf{bl})]_{ij} = \begin{cases} \dfrac{1}{|B_{\mathsf{bl}}(i, j) \cap E|} \displaystyle\sum_{(k,\ell) \in B_{\mathsf{bl}}(i,j) \cap E} Y'_{k\ell} & \text{if } B_{\mathsf{bl}}(i, j) \cap E \neq \phi, \\ 1/2 & \text{otherwise.} \end{cases} \tag{8}$$

In words, this defines a projection of the matrix Y' onto the set of blockwise constant matrices, by averaging the entries of Y' over each block using only the observed set of entries E.

We have now established the notation required to precisely describe the subroutine. It takes a tuple of parameters (Y, Y') as input, and outputs a blockwise constant matrix $\mathsf{BA}(Y, Y')$. Let $S = \frac{1}{n} \sum_{v \in V} d_v^{-1/2}$.

(I) *Block:* Compute a block partition $\mathsf{bl} = \mathsf{bl}_S(\widehat{\tau}(Y))$.

(II) *Average:* Average the matrix Y' according to the blocks computed in step (I), as $\mathsf{B}(Y', \mathsf{bl})$. Return the estimate $\mathsf{BA}(Y, Y') = \mathsf{B}(Y', \mathsf{bl})$.

Note that as defined, the blocking and averaging steps can be carried out for two *different* observation matrices Y and Y'. The various steps of this subroutine and examples of our notation are illustrated in Figure 1 for the special case $Y = Y'$.

BAP *algorithm.* With the block-average subroutine in place, we are now ready to describe the entire algorithm. The only missing detail is that the estimate returned at the end of the block-average subroutine may not lie within the SST class, and so we need to define a suitable projection onto it. This leads to the block-average-project (BAP) algorithm, which, as before, takes as input the tuple (Y, Y'), and outputs a matrix $\widehat{M}_{\mathsf{BAP}}(Y, Y') \in \mathbb{C}_{\mathsf{SST}}$.

(I) *Block-Average:* Compute the block averaged matrix $\widetilde{M} = \mathsf{BA}(Y, Y')$.

(II) *ASP:* Using the observation Y, compute a permutation estimate $\widehat{\pi}_{\mathsf{ASP}}$ as in step (II) of the ASP estimator.

(III) *Projection:* Project $\widetilde{M}$ onto the space $\widehat{\pi}_{\mathsf{ASP}}(\mathbb{C}_{\mathsf{BISO}}) = \{\widehat{\pi}_{\mathsf{ASP}}(M) : M \in \mathbb{C}_{\mathsf{BISO}}\}$, to obtain the estimator $\widehat{M}_{\mathsf{BAP}}(Y, Y')$.

$$\underset{Y}{\begin{bmatrix} & 0 & & 1 & 0 & 1 & 0 \\ 1 & & 1 & 0 & 0 & 0 & 0 \\ & 0 & & 1 & & & 0 \\ 0 & 1 & 0 & & 0 & 0 & 1 \\ 1 & 1 & & 1 & & 0 & \\ 0 & 1 & & 1 & 1 & & \\ 1 & 1 & 1 & 0 & & & \end{bmatrix}} \qquad \underset{\mathsf{bl}_S(Y) = \{\{1,2,3,4\},\{5,6,7\}\}}{\left[\begin{array}{cccc:ccc} & 0 & & 1 & 0 & 1 & 0 \\ 1 & & 1 & 0 & 0 & 0 & 0 \\ & 0 & & 1 & & & 0 \\ 0 & 1 & 0 & & 0 & 0 & 1 \\ \hdashline 1 & 1 & & 1 & & 0 & \\ 0 & 1 & & 1 & 1 & & \\ 1 & 1 & 1 & 0 & & & \end{array}\right]} \qquad \underset{\mathsf{BA}(Y,Y,E)}{\begin{bmatrix} .5 & .5 & .5 & .5 & .2 & .2 & .2 \\ .5 & .5 & .5 & .5 & .2 & .2 & .2 \\ .5 & .5 & .5 & .5 & .2 & .2 & .2 \\ .5 & .5 & .5 & .5 & .2 & .2 & .2 \\ .8 & .8 & .8 & .8 & .5 & .5 & .5 \\ .8 & .8 & .8 & .8 & .5 & .5 & .5 \\ .8 & .8 & .8 & .8 & .5 & .5 & .5 \end{bmatrix}}$$

FIG. 1. *Illustration of the block-average subroutine for the special case where $Y' = Y$, and for a particular realization of the edge set E. Here, we have $n = 7$, as the degree of a vertex is given by the number of entries observed in the corresponding row; the degree functional is given by $S = \frac{1}{n}\sum_{v \in V} d_v^{-1/2} = 0.45$. The dotted lines separate the blocks induced by the partition $\mathsf{bl}_S(\widehat{\tau}(Y))$, and while it is the case in this example, the blocks are not contiguous in general. Let us also illustrate some notation: in this example, it is also worth noting that $B_{\mathsf{bl}}(1, 2) = \{1, 2, 3, 4\} \times \{5, 6, 7\}$.*

While the estimator can in principle be implemented in the average-case setting of the previous section by setting $Y' = Y$, and we can obtain good permutation estimates in this setting (see equation (9a) to follow), our analysis of the block-average subroutine for comparison matrix estimation requires an independent tuple of samples (Y, Y'). Toward this end, recall the average-case setting with multiple random designs, as described in Section 2.2.2, in which the comparison topology is fixed ahead of time, but one can collect two sets of observations $Y^{(1)}$ and $Y^{(2)}$ by assigning items to the vertices of the underlying graph at random. Notice that observation $Y^{(i)}$ is obtained as a result of observing the matrix M^* (with noise) on the random subset of entries given by the nonzero entries in the matrix $\mathcal{O}^{(i)}$. The BAP estimator produces the following rate for this observation model.

THEOREM 4. *Consider any graph G without isolated vertices and any matrix $M^* \in \mathbb{C}_{\mathsf{SST}}$. For the permutation estimator $\widehat{\pi}_{\mathsf{ASP}}$ computed using $Y^{(1)}$, we have*

$$\text{(9a)} \qquad \mathbb{E}\left[\frac{1}{n^2}\|\widehat{\pi}_{\mathsf{ASP}}(M^*) - M^*\|_F^2\right] \lesssim \frac{1}{n}\sum_{v \in V}\frac{1}{\sqrt{d_v}},$$

where the expectation is taken over one sample of observations. For the matrix estimator $\widehat{M}_{\mathsf{BAP}}(Y^{(1)}, Y^{(2)})$, we have

$$\text{(9b)} \qquad \mathbb{E}\left[\frac{1}{n^2}\|\widehat{M}_{\mathsf{BAP}}(Y^{(1)}, Y^{(2)}) - M^*\|_F^2\right] \lesssim \frac{1}{n}\sum_{v \in V}\frac{1}{\sqrt{d_v}},$$

where the expectation is taken over both samples of observations.

In the simulations of Section 4, we see that for a large variety of graphs, using a single sample $\mathcal{O}^{(1)}$ enjoys similar performance to using two independent samples $\mathcal{O}^{(1)}$ and $\mathcal{O}^{(2)}$. Consequently, we make the following conjecture for the case where this single sample is given by Y and the block-average subroutine performs both steps on the matrix Y.

CONJECTURE 1. *For any graph G without isolated vertices and any matrix $M^* \in \mathbb{C}_{\mathsf{SST}}$, we have*

$$\mathbb{E}\left[\frac{1}{n^2}\|\widehat{M}_{\mathsf{BAP}}(Y, Y) - M^*\|_F^2\right] \lesssim \frac{1}{n}\sum_{v \in V}\frac{1}{\sqrt{d_v}}.$$

It is also important to note that while we were able to show a bound on permutation recovery in the Kendall tau distance for estimation in the noisy sorting case, such a result is not immediately obtainable uniformly over the SST class. For SST matrices satisfying a certain separation condition between the items, guarantees in the Frobenius norm can be turned into those that hold for permutation recovery in the Kendall tau distance (see, e.g., Appendix A.1 of Shah et al. [46]). However, without an appropriate separation condition, such a guarantee cannot be immediately obtained: for example, there is no way to obtain a permutation with a low Kendall tau error when a few items are near identical.

4. Dependence on graph topologies. In this section, we discuss implications of our results for some comparison topologies. Let us focus first on the worst-case design setting, and the lower bound of Theorem 1. For any graph with bounded average degree (including the star and path graphs) as well as for complete bipartite graphs, one can verify that we have $\alpha(G) \asymp n$, so $\mathcal{M}(G, \mathbb{C}_{\mathsf{NS}}) \asymp 1$. If the graph is a union of disjoint cliques $K_{n/2} \cup K_{n/2}$ (or having a constant number of edges across the cliques, like a barbell graph), then we see that $\beta(G^c) \asymp n^2$, so $\mathcal{M}(G, \mathbb{C}_{\mathsf{NS}}) \asymp 1$. Thus, our theory yields pessimistic results for many practically motivated comparison topologies under worst-case designs, even though all the connected graphs above admit consistent estimation for parametric models—the complete bipartite graph, for instance, admits optimal rates of estimation—as the number of samples grows. In the average case setting of Section 2.2.2, Theorems 2, 3 and 4 characterize the mean-squared Frobenius norm errors of the corresponding estimators (up to constants) as $\mathcal{D}(G) := \frac{1}{n} \sum_{v \in V} \frac{1}{\sqrt{d_v}}$.

The SST model has been validated extensively on real data in past work (see, e.g., Ballinger and Wilcox [1]). In order to illustrate our results for the average-case setting, we present the results of simulations on data generated synthetically from two special cases of the SST model. We fix $\pi^* = \mathsf{id}$ without loss of generality, and generate the ground truth comparison matrix M^* in one of two ways:

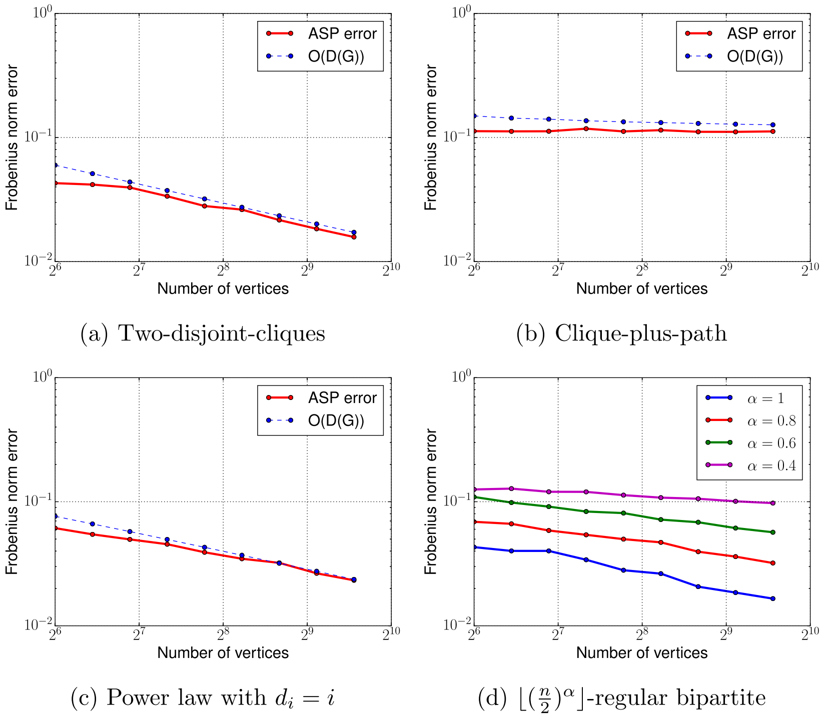
(i) Noisy sorting with high SNR: We set $M^* = M_{\mathsf{NS}}(\mathsf{id}, 0.4)$. This matrix is used in the simulations illustrated in Figure 2.

(ii) SST with independent bands: We first set $M^*_{ii} = 1/2$ for every i. Entries on the diagonal band immediately above the diagonal (i.e., $M^*_{i,i+1}$ for $i \in [n-1]$) are chosen i.i.d. and uniformly at random from the set $[1/2, 1]$. The band above is then chosen uniformly at random from the allowable set, where every entry is constrained to be upper bounded by 1 and lower bounded by the entries to its left and below. We also set $M^*_{ij} = 1 - M^*_{ji}$ to fill the lower triangle of the matrix. This matrix is used in the simulations illustrated in Figure 3.

For each graph G with adjacency matrix A, the data is generated from ground truth by observing independent Bernoulli comparisons under the observation process $\mathcal{O} = \sigma(A)$, for a randomly generated permutation σ. For the SST model, we also generate data from two independent random observations $\mathcal{O}^{(1)}$ and $\mathcal{O}^{(2)}$ as required by the BAP estimator; however, we also simulate the behavior of the estimator for one sample $\mathcal{O}^{(1)}$ and show that it closely tracks that of the two-sample estimator.

Recall that the estimation error rate was dictated by the degree functional $\mathcal{D}(G)$. While our graphs were chosen to illustrate scalings of $\mathcal{D}(G)$, some variants of these graphs also naturally arise as comparison topologies.

(a) *Two-disjoint-clique graph:* For this graph $K_{n/2} \cup K_{n/2}$, we have $d_v = \frac{n}{2} - 1$ for every $v \in V$, and simple calculations yield $\mathcal{D}(G) \asymp \frac{1}{\sqrt{n}}$. It is interesting to note that this graph has unfavorable guarantees for parametric estimation under the adversarial model, because it is disconnected (and thus has a Laplacian with zero spectral gap). We observe that this spectral property does not play a role in our analysis of the ASP or BAP estimator under the average-case observation model, and this behavior is corroborated by our simulations. Although we do not show it here, a similar behavior is observed for the stochastic block model, a practically



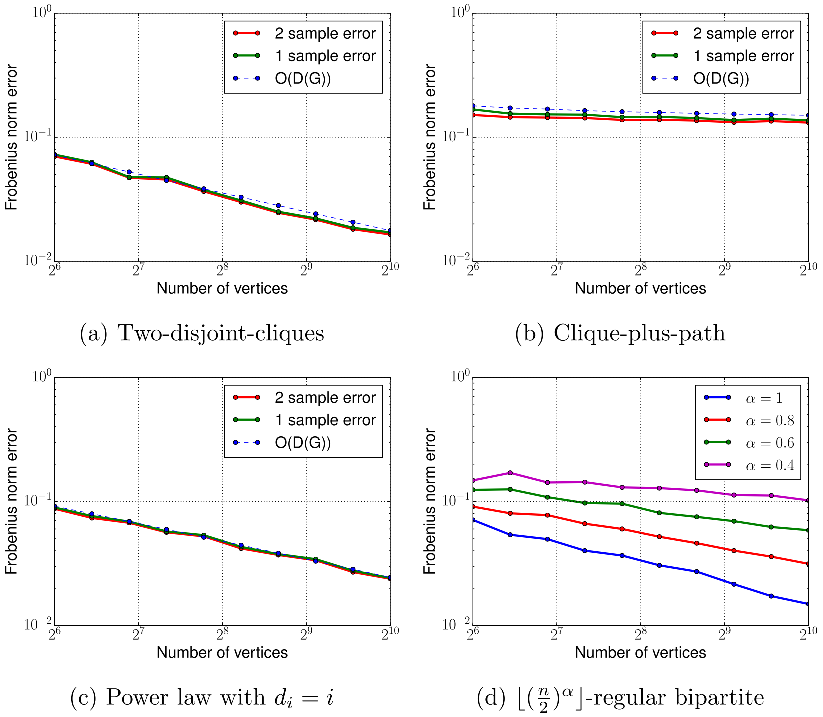
(a) Two-disjoint-cliques

(b) Clique-plus-path

(c) Power law with $d_i = i$

(d) $\lfloor (\frac{n}{2})^\alpha \rfloor$-regular bipartite

FIG. 2. *Normalized Frobenius norm error* $\frac{1}{n^2}\|\widehat{M}_{\mathsf{ASP}} - M^*\|_F^2$ *with data generated using the noisy sorting model* $M^* = M_{\mathsf{NS}}(\mathsf{id}, 0.4)$, *averaged over* 10 *trials.*



(a) Two-disjoint-cliques

(b) Clique-plus-path

(c) Power law with $d_i = i$

(d) $\lfloor (\frac{n}{2})^\alpha \rfloor$-regular bipartite

FIG. 3. *Normalized Frobenius norm error* $\frac{1}{n^2}\|\widehat{M}_{\mathsf{BAP}} - M^*\|_F^2$ *with data generated using the SST model with independent bands*, *averaged over* 10 *trials*, *plotted for one and two samples.*

motivated comparison topology when there are genres present among the items, which is a relaxation of the two-clique case allowing for sparser "communities" instead of cliques, and edges between the communities.

(b) *Clique-plus-path graph:* The nodes are partitioned into two sets of $n/2$ nodes each. The graph contains an edge between every two nodes in the first set, and a path starting from one of the nodes in the first set and chaining the other $n/2$ nodes. This is an example of a graph construction that has many ($\asymp n^2$) edges, but is unfavorable for noisy sorting or SST estimation. Simple calculations show that the degree functional is dominated by the constant degree terms and we obtain $\mathcal{D}(G) \asymp 1$.

(c) *Power law graph:* We consider the special power law graph [3] with degree sequence $d_i = i$ for $1 \leq i \leq n$, and construct it using the Havel–Hakimi algorithm [20, 21]. For this graph, we have a disparate degree sequence, but $\mathcal{D}(G) \asymp \frac{1}{\sqrt{n}}$, and the simulated estimators are consistent.

(d) $\lfloor (n/2)^\alpha \rfloor$*-regular bipartite graphs:* A final powerful illustration of our theoretical guarantees is provided by a regular bipartite graph construction in which the nodes are partitioned into two sets of $n/2$ nodes each, and each node in one set is (deterministically) connected to $\lfloor (n/2)^\alpha \rfloor$ nodes in the other set. This results in the degree sequence $d_v = \lfloor (n/2)^\alpha \rfloor$ for all $v \in V$, and the degree functional evaluates to $\mathcal{D}(G) \asymp n^{-\alpha/2}$. The value of α thus determines the scaling of the estimation error for the ASP estimator in the noisy sorting case, as well as the BAP estimator in the SST case, as seen from the slopes of the corresponding plots.

Some other graphs that were considered in parametric model environments [45], such as the star, cycle, path and hypercube graphs, turn out to be unfavorable for permutation-based models even in the average-case setting, as corroborated by the lower bound of Theorem 3, part (b).

5. Proofs.

In this section, we provide the proofs of our main results. We assume throughout that $n \geq 2$, and use c, c' to denote universal constants that may change from line to line. We defer the proofs of technical lemmas to the Supplementary Material [38].

5.1. *Proof of Theorem* 1.

Here, we focus on establishing the bound (5b) from bound (5a). The proof of the claimed lower bound (5a) is deferred to Appendix A of the Supplementary Material [38].

First, the estimation-theoretic lower bound of order $1/n$ in inequality (5b) holds for the noisy sorting model even if we have complete observations [31, 46]. Therefore, we focus on the other two terms that result from approximations.

For each fixed graph G, define the quantity

$$\mathcal{A}(G) := \sup_{\substack{M, M' \in \mathbb{C}_{\mathsf{NS}} \\ M(G) = M'(G)}} \frac{1}{n^2} \sum_{(i,j) \notin E} (M_{ij} - M'_{ij})^2$$

corresponding to the diameter quantity that is lower bounded in equation (5a). We claim that the minimax risk is lower bounded in terms of $\mathcal{A}(G)$ as

$$\inf_{\widehat{M} = f(Y(G))} \sup_{M^* \in \mathbb{C}_{\mathsf{NS}}} \mathbb{E}\left[\frac{1}{n^2} \|\widehat{M} - M^*\|_F^2\right] \geq \frac{1}{4} \mathcal{A}(G). \tag{10}$$

In order to verify this claim, consider the two matrices $M^1, M^2 \in \mathbb{C}_{\mathsf{NS}}$ that attain the supremum in the definition of $\mathcal{A}(G)$; note that such matrices exist due to the compactness of the space and the continuity of the squared loss. By construction, these two matrices satisfy the properties

$$M^1(G) = M^2(G), \quad \text{and} \quad \sum_{(i,j)\notin E} \left(M^1_{ij} - M^2_{ij}\right)^2 = n^2 \mathcal{A}(G).$$

We can now reduce the problem to one of testing between the two matrices M^1 and M^2, with the distribution of observations being identical for both alternatives. Consequently, any procedure can do no better than to make a random guess between the two, so we have

$$\inf_{\widehat{M}} \sup_{M^* \in \mathbb{C}_{\mathsf{NS}}} \mathbb{E}\big[\|\widehat{M} - M^*\|_F^2\big] \ge \frac{1}{4} \sum_{(i,j)\notin E} \left(M^1_{ij} - M^2_{ij}\right)^2,$$

which proves claim (10).

5.2. *Proof of Theorem* 2. Without loss of generality, reindexing as necessary, we may assume that the true permutation π^* is the identity id, thereby ensuring that $M^* = M_{\mathsf{NS}}(\mathsf{id}, \lambda^*)$. We begin by applying the triangle inequality to upper bound the error as a sum of two terms:

$$\frac{1}{2}\|\widehat{M}_{\mathsf{ASP}} - M^*\|_F^2 \le \underbrace{\|\widehat{M}_{\mathsf{ASP}} - \widehat{\pi}_{\mathsf{ASP}}(M^*)\|_F^2}_{\text{estimation error}} + \underbrace{\|\widehat{\pi}_{\mathsf{ASP}}(M^*) - M^*\|_F^2}_{\text{approximation error}}.$$

Lemma B.1 in the Supplementary Material [38] bounds the approximation error as

$$\mathbb{E}\big[\|\widehat{\pi}_{\mathsf{ASP}}(M^*) - M^*\|_F^2\big] \le cn \sum_{v\in V} \frac{1}{\sqrt{d_v}}.$$

We now turn to the estimation error term, which evaluates to $n^2(\widehat{\lambda} - \lambda^*)^2$, with $\widehat{\lambda}$ representing the MLE of λ^* conditional on $\widehat{\pi}$ being the correct permutation. Abusing notation slightly, let E denote the random set of observed edges, where we use the convention that each ordered pair $(i, j) \in E$ obeys $i < j$. Define the set

$$I_\pi(E) = \big\{(i, j) \in E \mid i < j, \pi(i) > \pi(j)\big\},$$

corresponding to the set of inversions that are also observed on the edge set E. With this notation in place, the MLE takes the form

$$\begin{aligned} 1/2 + \widehat{\lambda} &= \frac{1}{|E|}\Bigg(\sum_{(i,j)\in E\setminus I_{\widehat{\pi}_{\mathsf{ASP}}}(E)} Y_{ij} + \sum_{(i,j)\in I_{\widehat{\pi}_{\mathsf{ASP}}}(E)} (1 - Y_{ij})\Bigg) \\ &= \frac{1}{|E|}\Bigg(\sum_{(i,j)\in E} Y_{ij} + \sum_{(i,j)\in I_{\widehat{\pi}_{\mathsf{ASP}}}(E)} (1 - 2Y_{ij})\Bigg) \\ &= 1/2 + \lambda^* + \frac{1}{|E|}\Bigg(\sum_{(i,j)\in E} W_{ij}\Bigg) + \frac{1}{|E|}\Bigg(\sum_{(i,j)\in I_{\widehat{\pi}_{\mathsf{ASP}}}(E)} -2\lambda^* - 2W_{ij}\Bigg), \end{aligned}$$

where we have written $Y_{ij} = M^*_{ij} + W_{ij}$. Consequently, the error obeys

$$\begin{aligned} &(\widehat{\lambda} - \lambda^*)^2 \\ &\quad \le \frac{3}{|E|^2}\Bigg(\sum_{(i,j)\in E} W_{ij}\Bigg)^2 + \frac{12}{|E|^2}(\lambda^*)^2 |I_{\widehat{\pi}_{\mathsf{ASP}}}(E)|^2 + \frac{12}{|E|^2}\Bigg(\sum_{(i,j)\in I_{\widehat{\pi}_{\mathsf{ASP}}}(E)} W_{ij}\Bigg)^2 \\ &\quad \overset{\text{(i)}}{\le} \underbrace{\frac{3}{|E|^2}\Bigg(\sum_{(i,j)\in E} W_{ij}\Bigg)^2}_{T_1} + \underbrace{\frac{12}{|E|}(\lambda^*)^2 |I_{\widehat{\pi}_{\mathsf{ASP}}}(E)|}_{T_2} + \underbrace{\frac{12}{|E|^2}\Bigg(\sum_{(i,j)\in I_{\widehat{\pi}_{\mathsf{ASP}}}(E)} W_{ij}\Bigg)^2}_{T_3}, \end{aligned}$$

where step (i) follows since $|I_{\widehat{\pi}_{\mathsf{ASP}}}(E)| \le |E|$ pointwise. We now bound each of the terms T_1, T_2 and T_3 separately.

Bounding T_1. First, by standard subexponential tail bounds [54], and noting that $W_{ij} \in [-1, 1]$, we have $\mathbb{E}[T_1] \le \frac{3}{|E|}$, and $\Pr\{T_1 \ge \frac{6}{|E|}\} \le e^{-|E|}$.

Bounding T_2. Writing

$$\begin{aligned}\frac{|E|}{12(\lambda^*)^2}\mathbb{E}[T_2] &= \mathbb{E}\big[\big|I_{\widehat{\pi}_{\mathsf{ASP}}}(E)\big|\big] \\ &= \sum_{i<j}\sum_{(u,v)\in E} \Pr\big[\sigma(i)=u, \sigma(j)=v\big] \\ &\quad \times \Pr\big[\widehat{\pi}_{\mathsf{ASP}}(i) > \widehat{\pi}_{\mathsf{ASP}}(j) | \sigma(i)=u, \sigma(j)=v\big] \\ &= \sum_{(u,v)\in E}\sum_{i<j}\frac{1}{n(n-1)}\Pr\big[\widehat{\pi}_{\mathsf{ASP}}(i) > \widehat{\pi}_{\mathsf{ASP}}(j) | \sigma(i)=u, \sigma(j)=v\big],\end{aligned}$$

and the following lemma provides a bound on this quantity. We remark that most of the technical effort of the proof is devoted to proving this lemma, and we do so in the Supplementary Material [38].

LEMMA 1. *For any pair of vertices $u \ne v$, we have*

$$\begin{aligned}&\sum_{i<j}\frac{1}{n(n-1)}\Pr\big[\widehat{\pi}_{\mathsf{ASP}}(i) > \widehat{\pi}_{\mathsf{ASP}}(j) | \sigma(i)=u, \sigma(j)=v\big] \\ &\quad \le \frac{c}{\lambda^*}\left(\frac{1}{\sqrt{d_u}} + \frac{1}{\sqrt{d_v}}\right).\end{aligned} \tag{11}$$

Using Lemma 1 in conjunction with our previous bounds yields

$$\mathbb{E}[T_2] \le c\frac{\lambda^*}{|E|}\sum_{(u,v)\in E}\left(\frac{1}{\sqrt{d_u}} + \frac{1}{\sqrt{d_v}}\right) = c\lambda^*\frac{\sum_{u\in V}\sqrt{d_u}}{\sum_{u\in V} d_u}, \tag{12}$$

where the equality follows since each term $\frac{1}{\sqrt{d_u}}$ appears d_u times in the sum over all edges, and $2|E| = \sum_{u\in V} d_u$. Let $\{d_{(u)}\}_{u=1}^n$ represent the sequence of vertex degrees sorted in ascending order. An application of Lemma B.2 in the Supplementary Material [38] with $a_u = d_{(u)}$ and $b_u = \frac{1}{\sqrt{d_{(u)}}}$ for $u \in [n]$ yields

$$\sum_{u\in V}\sqrt{d_u} \le \frac{1}{n}\left(\sum_{u\in V} d_u\right)\left(\sum_{u\in V}\frac{1}{\sqrt{d_u}}\right).$$

Together with equation (12), we find that

$$\mathbb{E}[T_2] \le \frac{c\lambda^*}{n}\sum_{u\in V}\frac{1}{\sqrt{d_u}}.$$

Bounding T_3. Note that this step is nontrivial, since the noise terms W_{ij} for $(i, j) \in I_{\widehat{\pi}_{\mathsf{ASP}}}(E)$ depend on and are coupled through the data-dependent quantity $\widehat{\pi}_{\mathsf{ASP}}$. In order to circumvent this tricky dependency, consider some *fixed* permutation π, and let $T_3^\pi = (\sum_{(i,j)\in I_\pi(E)} W_{ij})^2$. Note that T_3^π has two sources of randomness: randomness in the edge set E and randomness in observations. Since the observations $\{W_{ij}\}$ are independent and bounded and $|I_\pi(E)| \le |E|$, the term $\sum_{(i,j)\in I_\pi(E)} W_{ij}$ is sub-Gaussian with parameter at most $\sqrt{|E|}$. Since the squares of sub-Gaussian variables have subexponential tails [54], we then have the uniform subexponential tail bound

$$\Pr\{T_3^\pi \ge |E| + \delta\} \le e^{-c\delta}. \tag{13}$$

Notice that for any $\alpha \in \mathbb{R}$, the inequality $T_3 \ge \alpha$ implies that the inequality $\frac{12}{|E|^2} T_3^\pi \ge \alpha$ holds for some *fixed* permutation π. Taking a union bound over all $n! \le e^{n \log n}$ fixed permutations, and setting $\delta = cn \log n$ for a constant $c > 1$ yields

$$\Pr\left\{T_3 \ge \frac{12}{|E|} + c\frac{n \log n}{|E|^2}\right\} \le \exp\{n \log n - cn \log n\} \le \exp\{-c' n \log n\}.$$

Noticing, in addition, that $T_3 \le 1$ pointwise, we obtain

$$\begin{aligned}\mathbb{E}[T_3] &\le \Pr\left\{T_3 \ge \frac{12}{|E|} + c\frac{n \log n}{|E|^2}\right\} \\ &\quad + \left(1 - \Pr\left\{T_3 \ge \frac{12}{|E|} + c\frac{n \log n}{|E|^2}\right\}\right)\left(\frac{12}{|E|} + c\frac{n \log n}{|E|^2}\right) \\ &\le \exp\{-c' n \log n\} + \frac{12}{|E|} + c\frac{n \log n}{|E|^2} \\ &\le c'\left(\frac{1}{|E|} + \frac{n \log n}{|E|^2}\right).\end{aligned}$$

Combining the pieces proves the claimed bound on the expectation.

5.3. *Proof of Theorem* 3. We sketch the proof of part (a) in the main text; the proof of part (b) may be found in the Supplementary Material [38]. The proof of part (a) is based on the following lemmas.

LEMMA 2. *Consider a matrix of the form $M^* = M_{\mathsf{NS}}(\pi^*, 1/4)$ where the permutation π^* is chosen uniformly at random. For any graph $G = K_1 \cup K_2 \cup \ldots$ composed of multiple disjoint cliques with the number of vertices bounded as $C \le |K_i| \le n/5$ for all i, and for any estimators $(\widehat{M}, \widehat{\pi})$ that are measurable functions of the observations on G, we have*

$$\mathbb{E}\left[\frac{1}{n^2}\|\widehat{M} - M^*\|_F^2\right] \ge \frac{c_2}{n} \sum_{v\in V} \frac{1}{\sqrt{d_v}}, \quad \textit{and}$$
$$\mathbb{E}[\mathsf{KT}(\pi^*, \widehat{\pi})] \ge c_2 n \sum_{v\in V} \frac{1}{\sqrt{d_v}}. \tag{14}$$

LEMMA 3. *Given any graph G with degree sequence $\{d_v\}_{v\in V}$, there exists a graph G' consisting of multiple disjoint cliques with degree sequence $\{d'_v\}_{v\in V}$ such that*

$$|E| \asymp |E'| \quad \textit{and} \quad \sum_{v\in V} \frac{1}{\sqrt{d_v}} \asymp \sum_{v\in V} \frac{1}{\sqrt{d'_v}}. \tag{15}$$

Part (a) of the theorem follows by combining these two lemmas, so that it suffices to prove each of the lemmas individually. We provide the proof of Lemma 2 below, and defer the proof of Lemma 3 to the Supplementary Material [38].

Proof of Lemma 2. Our result is structural, and proved for permutation recovery in the Kendall tau metric. The bound for matrix recovery follows as a corollary. Assume we are given a graph on n vertices consisting of k disjoint cliques of sizes $n_1, \dots, n_k$. Let $N_0 = 0$ and $N_j = \sum_{i=1}^{j} n_i$ for $j \in [k]$. Without loss of generality, we let the jth clique consist of the set of vertices V_j indexed by $\{N_{j-1} + 1, \dots, N_j\}$. By assumption, each n_j is upper bounded by $n/5$ and lower bounded by a universal constant.

Note that any estimator can only use the observations to construct the correct partial order within each clique, but not across cliques. We denote the induced partial order of a permutation π on the clique V_j by the permutation $\pi_j : [n_j] \to [n_j]$. (As an example, the identity permutation $\pi = \mathsf{id}$ would yield $\pi_j = \mathsf{id}$ on $[n_j]$ for all $j \in [k]$.) We claim that there exists a coupling of two marginally uniform random permutations $(\pi^*, \pi^\#)$ on the set $[n]$ such that

$$\mathbb{E}[\mathsf{KT}(\pi^*, \pi^\#)] \geq cn \sum_{j=1}^{k} \sqrt{n_j} = cn \sum_{v \in V} \frac{1}{\sqrt{d_v}}, \quad \text{and} \tag{16a}$$

$$\pi_j^* = \pi_j^\# \quad \text{for all } j \in [k]. \tag{16b}$$

Taking such a coupling as given for the moment, let us establish Lemma 2.

Let $\mathbb{E}[\cdot \mid \pi^*]$ denote the expectation over the observations conditional on π^*. Given a pair of permutations $(\pi^*, \pi^\#)$ satisfying the above assumption, we view them as two hypotheses of the latent permutation. Then for any estimator $\widehat{\pi}$, the Neyman–Pearson lemma [37] guarantees that

$$\mathbb{E}[\mathsf{KT}(\widehat{\pi}, \pi^*) \mid \pi^*] + \mathbb{E}[\mathsf{KT}(\widehat{\pi}, \pi^\#) \mid \pi^\#] \geq \mathsf{KT}(\pi^\#, \pi^*)$$

for each instance of $(\pi^*, \pi^\#)$, because the observations are identical for π^* and $\pi^\#$. Taking expectation over $(\pi^*, \pi^\#)$, we obtain the bound

$$2\mathbb{E}[\mathsf{KT}(\widehat{\pi}, \pi^*)] \geq \mathbb{E}[\mathsf{KT}(\pi^*, \pi^\#)] \geq cn \sum_{v \in V} \frac{1}{\sqrt{d_v}},$$

since both π^* and $\pi^\#$ are marginally uniform. It remains to construct the claimed coupling.

Constructing the coupling. The construction is done as follows. First, two permutations π^* and $\tilde{\pi}$ on the set $[n]$ are generated uniformly at random and independently. Second, we sort the permutation $\tilde{\pi}$ on each clique according to π^*, and denote the resulting permutation by $\pi^\#$. Then the permutations π^* and $\pi^\#$ are marginally uniform and have common induced partial orders on the cliques, which we denote by $\{\pi_j^* : j \in [k]\}$.

With some extra notation, we can define the sorting step more formally for the interested reader. For a set of partial orders on the cliques $\{\pi_j : j \in [k]\}$, we define a special permutation that effectively orders vertices within each clique V_j according to its corresponding partial order π_j, but does not permute any vertices across cliques. We denote this special permutation by $\pi_{\mathsf{par}}(\{\pi_j : j \in [k]\})$. For every clique V_j, we consider the permutation $\pi_{\mathsf{sort},j} := \pi_j^* \circ (\tilde{\pi}_j)^{-1}$. Now we define the sorting step to generate $\pi^\#$ by

$$\pi^\# := \pi_{\mathsf{par}}(\{\pi_{\mathsf{sort},j} : j \in [k]\}) \circ \tilde{\pi}.$$

By construction, such a coupling obeys condition (16b); it remains to evaluate the expected Kendall tau distance between these coupled permutations in order to establish condition (16a). By the tower property, we have

$$\mathbb{E}[\mathsf{KT}(\pi^*, \pi^\#)] = \mathbb{E}[\mathbb{E}[\mathsf{KT}(\pi^*, \pi^\#) \mid \{\pi_j^* : j \in [k]\}]].$$

The inner expectation can be simplified as follows. Precomposing permutations π^* and $\pi^\#$ with any permutation does not change the Kendall tau distance between them, so we have

$$\mathbb{E}[\mathsf{KT}(\pi^*, \pi^\#) \mid \{\pi_j^* : j \in [k]\}] = \mathbb{E}[\mathsf{KT}(\pi, \pi')],$$

where the permutations π and π' are drawn independently and uniformly at random from the set of permutations that are increasing on every clique. That is, for every clique V_j and every two vertices $i_1, i_2 \in V_j$, we have $\pi(i_1) < \pi(i_2)$ and $\pi'(i_1) < \pi'(i_2)$. To understand why π and π' can be chosen independently, note that the only dependency between the original permutations π^* and $\pi^\#$ is through the common induced partial orders $\{\pi_j^* : j \in [k]\}$. By conditioning and precomposing, we removed that dependency.

We now turn to computing the quantity $\mathbb{E}[\mathsf{KT}(\pi, \pi')]$. It is well known [14] that $2\mathsf{KT}(\pi, \pi') \geq \|\pi - \pi'\|_1$. This fact together with Jensen's inequality implies that

$$\begin{aligned} 2\mathbb{E}[\mathsf{KT}(\pi, \pi')] &\geq \sum_{i=1}^n \mathbb{E}[|\pi(i) - \pi'(i)|] \\ &\geq \sum_{i=1}^n \mathbb{E}[|\mathbb{E}[\pi(i) - \pi'(i) \mid \pi]|] \qquad (17) \\ &= \sum_{i=1}^n \mathbb{E}[|\pi(i) - \mathbb{E}[\pi'(i)]|] = \mathbb{E}[\|\pi - \mathbb{E}[\pi]\|_1]. \end{aligned}$$

It therefore suffices to lower bound the quantity $\mathbb{E}[\|\pi - \mathbb{E}[\pi]\|_1]$.

Fix any $i \in [n]$. Then i is ℓth smallest index in the jth clique for some $j \in [k]$ and $\ell \in [n_j]$, or succinctly, $i = N_{j-1} + \ell$. If we view π^{-1} as random draws from the n items, then $\pi(i)$ is equal to the the number of draws needed to get the ℓth smallest element of V_j. Denoting $\mathbb{E}[\pi(i)]$ by μ, we have

$$\begin{aligned} \mu &= \ell + \mathbb{E}\Big[\sum_{r : \sigma(r) \notin V_j} \mathbf{1}\{r \text{ is drawn before } i\}\Big] \\ &= \ell + (n - n_j)\frac{\ell}{n_j + 1} = \ell\frac{n+1}{n_j+1}, \end{aligned}$$

since the probability that an item not in V_j is drawn before the ℓth smallest element of V_j is $\ell/(n_j + 1)$. Furthermore, $\pi(i) = s$ if and only if $\ell - 1$ elements of V_j are selected in the first $s - 1$ draws and the sth draw is from V_j, so

$$\Pr\{\pi(i) = s\} = \binom{n_j}{\ell - 1}\binom{n - n_j}{s - \ell}\binom{n}{s-1}^{-1}\frac{n_j - \ell + 1}{n - s + 1}. \qquad (18)$$

For any $0 \leq m \leq n/\sqrt{n_j}$, equation (18) and Lemma D.1 together yield

$$\mathbb{E}[|\pi(i) - \mu|] \geq m \Pr\{|\pi(i) - \mu| \geq m\} \geq m[1 - c(2m+1)\sqrt{n_j}/n]$$

by Markov's inequality. Choosing $m = \frac{n}{6c\sqrt{n_j}}$ yields

$$\mathbb{E}[|\pi(i) - \mu|] \geq c_2 n/\sqrt{n_j}$$

for some positive constant c_2. Summing over ℓ in the given range and putting together the pieces establishes that condition (16a) holds.

5.4. *Proof of Theorem* 4. The first part of the theorem is the bound on $\|\widehat{\pi}_{\mathsf{ASP}}(M^*) - M^*\|_F^2$ for the permutation estimator $\widehat{\pi}_{\mathsf{ASP}}$, and follows immediately from Lemma B.1 in the Supplementary Material [38].

We now concentrate on the second part of the theorem. Recall that we have two samples of independent observations, and for the rest of the proof, it is helpful to think of the observation model in its linearized form. In particular, we have two random edge sets E_1 and E_2 and the observation matrices $Y^{(i)} := M^* + W^{(i)}$ for each $i \in \{1, 2\}$. For a set $B \subseteq [n] \times [n]$, recall that we use the notation $\|M\|_B^2 := \sum_{(i,j)\in B} M_{ij}^2$.

Recall the definition of the estimator $\widetilde{M} = \mathsf{BA}(Y^{(1)}, Y^{(2)}) = \mathsf{B}(Y^{(2)}, \mathsf{bl})$ where $\mathsf{bl} = \mathsf{bl}_S(\tau(Y^{(1)}))$. Then the estimator $\widehat{M}_{\mathsf{BAP}}$ is obtained by projecting the matrix $\widetilde{M}$ onto the space $\widehat{\pi}_{\mathsf{ASP}}(\mathbb{C}_{\mathsf{BISO}}) = \{\widehat{\pi}_{\mathsf{ASP}}(M) : M \in \mathbb{C}_{\mathsf{BISO}}\}$. Our first step is to show that it suffices to consider the estimator $\widetilde{M}$, because the projection to obtain $\widehat{M}_{\mathsf{BAP}}$ is nonexpansive with respect to the Frobenius norm.

More precisely, the triangle inequality yields

$$\begin{aligned} \|\widehat{M}_{\mathsf{BAP}} - M^*\|_F^2 &\le 2\|\widehat{M}_{\mathsf{BAP}} - \widehat{\pi}_{\mathsf{ASP}}(M^*)\|_F^2 + 2\|M^* - \widehat{\pi}_{\mathsf{ASP}}(M^*)\|_F^2 \\ &\overset{\text{(i)}}{\le} 2\|\widetilde{M} - \widehat{\pi}_{\mathsf{ASP}}(M^*)\|_F^2 + 2\|M^* - \widehat{\pi}_{\mathsf{ASP}}(M^*)\|_F^2 \qquad (19)\\ &\le 4\|\widetilde{M} - M^*\|_F^2 + 6\|M^* - \widehat{\pi}_{\mathsf{ASP}}(M^*)\|_F^2, \end{aligned}$$

where step (i) follows from the nonexpansiveness of the projection operator on a convex set. The second term in inequality (19) is already bounded as desired, so it remains to consider the first term.

Toward that end, we split the error into the approximation error, measuring how well the matrix M^* is approximated by a block matrix, and the estimation error, measuring how well this block approximation is estimated by our estimator $\widetilde{M}$. Namely, we again apply the triangle inequality to write

$$\|\widetilde{M} - M^*\|_F^2 \le 2\|M^* - \mathsf{B}(M^*, \mathsf{bl})\|_F^2 + 2\|\widetilde{M} - \mathsf{B}(M^*, \mathsf{bl})\|_F^2. \qquad (20)$$

Bounding the estimation error $\mathbb{E}[\|\widetilde{M} - \mathsf{B}(M^*, \mathsf{bl})\|_F^2]$. Before diving into the details, it is worth noting that this is the only bound that requires the two sample assumption. In particular, once the blocks have been obtained, Lemma 4 relies on the fact that the remaining samples are uniformly distributed across each blocks, so that sample averages estimate the population block-average. Lemma 5 then shows that there are sufficiently many samples seen on each block, so that the variance of such an averaging estimator is bounded as desired. We expect that the proof of Conjecture 1 will involve showing analogues of these lemmas when the samples used in the two steps are coupled.

Let us now provide the required bound in the two sample setting. We write the error as a sum of errors on individual blocks

$$\|\widetilde{M} - \mathsf{B}(M^*, \mathsf{bl})\|_F^2 = \sum_{B \in \mathcal{B}_{\mathsf{bl}}} \|\widetilde{M} - \mathsf{B}(M^*, \mathsf{bl})\|_B^2.$$

Note that it is sufficient to consider off diagonal blocks in the sum, since both $\widetilde{M}$ and $\mathsf{B}(M^*, \mathsf{bl})$ have all entries equal to 1/2 in the diagonal blocks due to the skew symmetry property $M^* + (M^*)^\top = \mathbf{e}\mathbf{e}^\top$. Considering each block separately, we now split the analysis into two cases.

Case 1: $B \cap E_2 = \phi$. Because the entries of the error matrix are bounded within $[-1, 1]$, we have $\|\widetilde{M} - \mathsf{B}(M^*, \mathsf{bl})\|_B^2 \le |B|$.

Case 2: $B \cap E_2 \neq \phi$. Since both $\widetilde{M}$ and $\mathsf{B}(M^*, \mathsf{bl})$ are constant on each block, we have

$$\begin{aligned}
&\|\widetilde{M} - \mathsf{B}(M^*, \mathsf{bl})\|_B^2 \\
&\quad = \frac{|B|}{|B \cap E_2|} \|\widetilde{M} - \mathsf{B}(M^*, \mathsf{bl})\|_{B\cap E_2}^2 \\
&\quad = \frac{|B|}{|B \cap E_2|} \|\mathsf{B}(M^* + W^{(2)}, \mathsf{bl}, E_2) - \mathsf{B}(M^*, \mathsf{bl})\|_{B\cap E_2}^2 \\
&\quad \leq 2\frac{|B|}{|B \cap E_2|} (\|\mathsf{B}(M^* + W^{(2)}, \mathsf{bl}, E_2) \\
&\qquad - \mathsf{B}(\mathsf{B}(M^*, \mathsf{bl}) + W^{(2)}, \mathsf{bl}, E_2)\|_{B\cap E_2}^2 \\
&\qquad + \|\mathsf{B}(\mathsf{B}(M^*, \mathsf{bl}) + W^{(2)}, \mathsf{bl}, E_2) - \mathsf{B}(M^*, \mathsf{bl})\|_{B\cap E_2}^2).
\end{aligned} \tag{21}$$

Before proceeding further, we require the following technical lemma.

LEMMA 4. *For any block B and tuple $(i, j) \in B$, we have*

$$\Pr\{(i, j) \in E_2 \mid |B \cap E_2| = k\} = \frac{k}{|B|}.$$

See the Supplementary Material [38] for a proof of this claim. This key lemma hinges on the independence between the block B which is determined using the first sample, and the edge set E_2 from the second sample. In the case where only one sample is available, the exact statement does not hold.

Let us now handle each term on the RHS of inequality (21) separately. First, by nonexpansiveness of the projection operation defined by equation (8), we have

$$\begin{aligned}
&\|\mathsf{B}(M^* + W^{(2)}, \mathsf{bl}, E_2) - \mathsf{B}(\mathsf{B}(M^*, \mathsf{bl}) + W^{(2)}, \mathsf{bl}, E_2)\|_{B\cap E_2}^2 \\
&\quad \leq \|M^* - \mathsf{B}(M^*, \mathsf{bl})\|_{B\cap E_2}^2.
\end{aligned} \tag{22}$$

Now taking expectation over the randomness in E_2 (which, crucially, is independent of the randomness in bl), we have

$$\begin{aligned}
&\mathbb{E}_{E_2}[\|M^* - \mathsf{B}(M^*, \mathsf{bl})\|_{B\cap E_2}^2 | |B \cap E_2| = k] \\
&\quad = \sum_{(i,j)\in B} \Pr\{(i, j) \in E_2 | |B \cap E_2| = k\} \cdot [M^* - \mathsf{B}(M^*, \mathsf{bl})]_{ij}^2 \\
&\quad \overset{\text{(ii)}}{=} \sum_{(i,j)\in B} \frac{k}{|B|} [M^* - \mathsf{B}(M^*, \mathsf{bl})]_{ij}^2 \\
&\quad = \frac{k}{|B|} \|M^* - \mathsf{B}(M^*, \mathsf{bl})\|_B^2,
\end{aligned} \tag{23}$$

where step (ii) follows from Lemma 4.

Next, we turn to the second term on the RHS of inequality (21). Notice that $[W^{(2)}]_{ij}$ for $(i, j) \in E_2$ is independent and bounded within the interval $[-1, 1]$. Also, notice that the entries of the matrix $\mathsf{B}(M^*, \mathsf{bl})$ are constant on the set of indices $B \cap E_2$. Consequently, given $|B \cap E_2|$ noisy samples of a constant parameter, we can achieve the estimation rate

$$\mathbb{E}_{W^{(2)}}[\|\mathsf{B}(\mathsf{B}(M^*, \mathsf{bl}) + W^{(2)}, \mathsf{bl}, E_2) - \mathsf{B}(M^*, \mathsf{bl})\|_{B\cap E_2}^2] \leq 1. \tag{24}$$

It follows from equations (21), (22), (23) and (24) that

$$\mathbb{E}[\|\widetilde{M} - \mathsf{B}(M^*, \mathsf{bl})\|_B^2] \le 2\mathbb{E}\bigg[\frac{|B|}{|B \cap E_2|}\bigg] + 2\mathbb{E}[\|M^* - \mathsf{B}(M^*, \mathsf{bl})\|_B^2].$$

Combining the two cases and summing over the blocks, we obtain

$$\begin{aligned}\mathbb{E}[\|\widetilde{M} - \mathsf{B}(M^*, \mathsf{bl})\|_F^2] \le 2 \sum_{B \in \mathcal{B}_{\mathsf{bl}}} \mathbb{E}\bigg[\frac{|B|}{|B \cap E_2| \vee 1}\bigg] \\ + 2\mathbb{E}[\|M^* - \mathsf{B}(M^*, \mathsf{bl})\|_F^2].\end{aligned} \tag{25}$$

The first term can be bounded using the following lemma, which is proved in the Supplementary Material [38].

LEMMA 5. *With $S = \sum_{v \in V} 1/\sqrt{d_v}$ and for the partition $\mathsf{bl} = \mathsf{bl}_S(Y^{(1)})$, we have*

$$\mathbb{E}_{E_2}\bigg[\sum_{B \in \mathcal{B}_{\mathsf{bl}}} \frac{|B|}{|B \cap E_2| \vee 1}\bigg] \le nS.$$

Thus, we have bounded the estimation error as

$$\|\widetilde{M} - \mathsf{B}(M^*, \mathsf{bl})\|_F^2 \le 2nS + 2\mathbb{E}[\|M^* - \mathsf{B}(M^*, \mathsf{bl})\|_F^2].$$

Notice that the second term above is identical to the approximation error; therefore, in order to complete the proof, it suffices to bound the second term of inequality (20).

Bounding the approximation error $\mathbb{E}[\|M^* - \mathsf{B}(M^*, \mathsf{bl})\|_F^2]$. Let us use $X_i^\top$ to denote the ith row of a matrix X. Given a matrix M^* and a partition bl of $[n]$, define $\mathsf{R}(M^*, \mathsf{bl})$ by averaging the rows of M^* over bl, that is,

$$[\mathsf{R}(M^*, \mathsf{bl})]_i^\top = \frac{1}{|\mathsf{bl}(i)|} \sum_{j \in \mathsf{bl}(i)} (M^*)_j^\top,$$

where $\mathsf{bl}(i)$ is the set of indices in the partition bl where i lies in. Note that by definition, we have

$$\mathsf{B}(M^*, \mathsf{bl}) = \mathsf{R}(\mathsf{R}(M^*, \mathsf{bl})^\top)^\top.$$

The following lemma, required for the approximation bound, is a generalization of a result of Chatterjee [9] to the noisy, two-dimensional setting. Its proof is postponed to the Supplementary Material [38].

LEMMA 6. *Given any matrix $X \in [0, 1]^{n \times n}$ with monotone columns, let $\tau(X)_i = \frac{1}{n} \sum_{j=1}^n X_{ij}$. For any score vector $\widehat{\tau} \in [0, 1]^n$, and any value $t \in [0, n]$, we have*

$$\|X - \mathsf{R}(X, \mathsf{bl}_t(\widehat{\tau}))\|_F^2 \le nt + 2n\|\widehat{\tau} - \tau(X)\|_1.$$

Let us now focus on the expression we wish to bound. We have

$$\begin{aligned}&\|M^* - \mathsf{B}(M^*, \mathsf{bl})\|_F^2 \\ &\quad \le 2\|M^* - \mathsf{R}(M^*, \mathsf{bl})\|_F^2 + 2\|\mathsf{R}(M^*, \mathsf{bl}) - \mathsf{B}(M^*, \mathsf{bl})\|_F^2 \\ &\quad = 2\|M^* - \mathsf{R}(M^*, \mathsf{bl})\|_F^2 + 2\|\mathsf{R}(M^*, \mathsf{bl})^\top - \mathsf{R}(\mathsf{R}(M^*, \mathsf{bl})^\top, \mathsf{bl})\|_F^2.\end{aligned}$$

Recall that $\mathsf{bl} = \mathsf{bl}_S(\widehat{\tau}(Y^{(1)}))$. Applying Lemma 6 to both terms, we obtain[3]

$$\|M^* - \mathsf{B}(M^*, \mathsf{bl})\|_F^2 \leq 2nS + 4n\|\widehat{\tau}(Y^{(1)}) - \tau^*\|_1,$$

because both M^* and $\mathsf{R}(M^*, \mathsf{bl})^\top$ have normalized row sums equal to τ^*. Lemma B.1 in the Supplementary Material [38] yields the desired bound on the quantity $\|\widehat{\tau}(Y^{(1)}) - \tau^*\|_1$ in expectation. This together with equations (20) and (25) completes the proof of Theorem 4 with the choice $S = \sum_{v \in V} 1/\sqrt{d_v}$.

6. Discussion. In this paper, we studied the problem of estimating the comparison probabilities from noisy pairwise comparisons under worst-case and average-case design assumptions. We exhibited a dichotomy between worst-case and average-case models for permutation-based models, which suggests that a similar distinction may exist even for their parametric counterparts. Our bounds leave a few interesting questions unresolved: Is there a sharp characterization of the diameter $\mathcal{A}(G)$ quantifying the worst-case approximation error of a comparison topology G? The Borda count estimator, a variant of which we analyzed, is known to achieve a suboptimal rate for random comparison topologies; the estimator of Mao et al. [31] achieves the optimal rate over the noisy sorting class. What is the analog of such an estimator in the average-case setting with partial pairwise comparisons? Is there a computational lower bound to show that our estimators are the best possible polynomial-time algorithms for SST matrix estimation?

Acknowledgments. The first and the fourth authors were supported by NSF Grant DMS-1612948, DOD Advanced Research Projects Agency Grant W911NF-16-1-0552, and DOD Office of Naval Research grant N00014-18-1-2640.

The first and the fifth authors were supported by NSF Grants CCF-1528132, CCF-1704967 and CCF-0939370.

The second author was supported by NSF CAREER Grant DMS-1541099 and was visiting the Simons Institute for the Theory of Computing when part of this work was done.

The third author was supported by NSF Grants AST-144078 and ECCS-1343398.

SUPPLEMENTARY MATERIAL

Supplement to "Worst-case versus average-case design for estimation from partial pairwise comparisons" (DOI: 10.1214/19-AOS1838SUPP; .pdf). Due to space constraints, we have relegated the technical details of remaining proofs to the supplement [38]. The supplement also contains a section characterizing the minimax denoising error under partial observations.

REFERENCES

[1] BALLINGER, T. P. and WILCOX, N. T. (1997). Decisions, error and heterogeneity. *Econ. J.* **107** 1090–1105.

[2] BALTRUNAS, L., MAKCINSKAS, T. and RICCI, F. (2010). Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems. RecSys '10* 119–126. ACM, New York.

[3] BARABÁSI, A.-L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512. MR2091634 https://doi.org/10.1126/science.286.5439.509

[4] BARNETT, W. (2003). The modern theory of consumer behavior: Ordinal or cardinal? *Q. J. Austrian Econ.* **6** 41–65.

[5] 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 https://doi.org/10.2307/2334029

[3] Lemma 6 can be applied in this setting because the diagonal entries of M^* and Y are all equal to $1/2$ and hence inconsequential.

[6] BRAVERMAN, M. and MOSSEL, E. (2008). Noisy sorting without resampling. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 268–276. ACM, New York. MR2485312

[7] CAPLIN, A. and NALEBUFF, B. (1991). Aggregation and social choice: A mean voter theorem. *Econometrica* **59** 1–23. MR1085582 https://doi.org/10.2307/2938238

[8] CATTELAN, M. (2012). Models for paired comparison data: A review with emphasis on dependent data. *Statist. Sci.* **27** 412–433. MR3012434 https://doi.org/10.1214/12-STS396

[9] CHATTERJEE, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist.* **43** 177–214. MR3285604 https://doi.org/10.1214/14-AOS1272

[10] CHATTERJEE, S. and MUKHERJEE, S. (2019). Estimation in tournaments and graphs under monotonicity constraints. *IEEE Trans. Inform. Theory* **65** 3525-3539. https://doi.org/10.1109/TIT.2019.2893911

[11] CHEN, X., BENNETT, P. N., COLLINS-THOMPSON, K. and HORVITZ, E. (2013). Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining* 193–202. ACM.

[12] CHEN, X., GOPI, S., MAO, J. and SCHNEIDER, J. (2017). Competitive analysis of the top-K ranking problem. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms* 1245–1264. SIAM, Philadelphia, PA. MR3627810 https://doi.org/10.1137/1.9781611974782.81

[13] CHEN, Y. and SUH, C. (2015). Spectral MLE: Top-k rank aggregation from pairwise comparisons. In *International Conference on Machine Learning* 371–380.

[14] DIACONIS, P. and GRAHAM, R. L. (1977). Spearman's footrule as a measure of disarray. *J. Roy. Statist. Soc. Ser. B* **39** 262–268. MR0652736

[15] DWORK, C., KUMAR, R., NAOR, M. and SIVAKUMAR, D. (2001). Rank aggregation methods for the web. In *Proceedings of the* 10*th International Conference on World Wide Web* 613–622. ACM, New York.

[16] FISHBURN, P. C. (1973). Binary choice probabilities: On the varieties of stochastic transitivity. *J. Math. Psych.* **10** 327–352. MR0327330 https://doi.org/10.1016/0022-2496(73)90021-7

[17] FLAMMARION, N., MAO, C. and RIGOLLET, P. (2019). Optimal rates of statistical seriation. *Bernoulli* **25** 623–653. MR3892331 https://doi.org/10.3150/17-bej1000

[18] FLIGNER, M. A. and VERDUCCI, J. S., eds. (1993). *Probability Models and Statistical Analyses for Ranking Data. Lecture Notes in Statistics* **80**. Springer, New York. MR1237197 https://doi.org/10.1007/978-1-4612-2738-0

[19] HAJEK, B., OH, S. and XU, J. (2014). Minimax-optimal inference from partial rankings. In *Advances in Neural Information Processing Systems* 1475–1483.

[20] HAKIMI, S. L. (1962). On realizability of a set of integers as degrees of the vertices of a linear graph. I. *J. Soc. Indust. Appl. Math.* **10** 496–506. MR0148049

[21] HAVEL, V. (1955). A remark on the existence of finite graphs. *Čas. Pěst. Mat.* **80** 477–480.

[22] HERBRICH, R., MINKA, T. and GRAEPEL, T. (2006). Trueskill™: A Bayesian skill rating system. In *Proceedings of the* 19*th International Conference on Neural Information Processing Systems* 569–576. MIT Press.

[23] JAMIESON, K. G. and NOWAK, R. D. (2011). Active ranking using pairwise comparisons. In *Advances in Neural Information Processing Systems* 2240–2248.

[24] JANG, M., KIM, S., SUH, C., and OH, S. (2017). Optimal sample complexity of m-wise data for top-k ranking. In *Advances in Neural Information Processing Systems* 1686–1696.

[25] KENDALL, M. G. (1948). Rank correlation methods. Oxford Univerosty Press, New York.

[26] KHETAN, A. and OH, S. (2016). Data-driven rank breaking for efficient rank aggregation. *J. Mach. Learn. Res.* **17** 193. MR3567461

[27] KIRÁLY, F. J., THERAN, L. and TOMIOKA, R. (2015). The algebraic combinatorial approach for low-rank matrix completion. *J. Mach. Learn. Res.* **16** 1391–1436. MR3417786

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

[29] MAO, C., PANANJADY, A. and WAINWRIGHT, M. J. (2018). Breaking the $1/\sqrt{n}$ barrier: Faster rates for permutation-based models in polynomial time. In *Proceedings of the* 31*st Conference on Learning Theory* (S. Bubeck, V. Perchet and P. Rigollet, eds.). *Proceedings of Machine Learning Research* **75** 2037–2042.

[30] MAO, C., PANANJADY, A. and WAINWRIGHT, M. J. (2019+). Towards optimal estimation of bivariate isotonic matrices with unknown permutations. *Ann. Statist.* To appear.

[31] MAO, C., WEED, J. and RIGOLLET, P. (2018). Minimax rates and efficient algorithms for noisy sorting. In *Algorithmic Learning Theory* 2018. *Proc. Mach. Learn. Res.* (*PMLR*) **83** 821–847. Proceedings of Machine Learning Research PMLR. MR3857331

[32] MARDEN, J. I. (1995). *Analyzing and Modeling Rank Data. Monographs on Statistics and Applied Probability* **64**. CRC Press, London. MR1346107

[33] MAYSTRE, L. and GROSSGLAUSER, M. (2017). Just sort it! A simple and effective approach to active preference learning. In *International Conference on Machine Learning* 2344–2353.

[34] MCLAUGHLIN, D. H. and LUCE, R. D. (1965). Stochastic transitivity and cancellation of preferences between bitter-sweet solutions. *Psychol. Sci.* **2** 1–12. 89–90.

[35] NEGAHBAN, S., OH, S. and SHAH, D. (2017). Rank centrality: Ranking from pairwise comparisons. *Oper. Res.* **65** 266–287. MR3613103 https://doi.org/10.1287/opre.2016.1534

[36] NEGAHBAN, S., OH, S., THEKUMPARAMPIL, K. K. and XU, J. (2018). Learning from comparisons and choices. *J. Mach. Learn. Res.* **19** 40. MR3862447

[37] NEYMAN, J. and PEARSON, E. S. (1966). Joint statistical papers. Univ. California.

[38] PANANJADY, A., MAO, C., MUTHUKUMAR, V., WAINWRIGHT, M. J. and COURTADE, T. A. (2020). Supplement to "Worst-case versus Average-case Design for Estimation from Partial Pairwise Comparisons." https://doi.org/10.1214/19-AOS1838SUPP.

[39] PANANJADY, A., WAINWRIGHT, M. J. and COURTADE, T. A. (2017). Denoising linear models with permuted data. In 2017 *IEEE International Symposium on Information Theory* (*ISIT*) 446–450. IEEE.

[40] PARK, D., NEEMAN, J., ZHANG, J., SANGHAVI, S. and DHILLON, I. (2015). Preference completion: Large-scale collaborative ranking from pairwise comparisons. In *International Conference on Machine Learning* 1907–1916.

[41] PIECH, C., HUANG, J., CHEN, Z., DO, C., NG, A. and KOLLER, D. (2013). Tuned models of peer assessment in MOOCs. In *Proceedings of the* 6*th International Conference on Educational Data Mining* 153–160.

[42] PIMENTEL-ALARCÓN, D. L., BOSTON, N. and NOWAK, R. D. (2016). A characterization of deterministic sampling patterns for low-rank matrix completion. *IEEE J. Sel. Top. Signal Process.* **10** 623–636.

[43] RAJKUMAR, A. and AGARWAL, S. (2016). When can we rank well from comparisons of $O(n\log(n))$ non-actively chosen pairs? In 29*th COLT* **49** 1376–1401.

[44] RIGOLLET, P. and WEED, J. (2019+). Uncoupled isotonic regression via minimum Wasserstein deconvolution. *Information and Inference*: *A Journal of the IMA* To appear.

[45] SHAH, N. B., BALAKRISHNAN, S., BRADLEY, J., PAREKH, A., RAMCHANDRAN, K. and WAINWRIGHT, M. J. (2016). Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *J. Mach. Learn. Res.* **17** 58. MR3504618

[46] SHAH, N. B., BALAKRISHNAN, S., GUNTUBOYINA, A. and WAINWRIGHT, M. J. (2017). Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *IEEE Trans. Inform. Theory* **63** 934–959. MR3604649 https://doi.org/10.1109/TIT.2016.2634418

[47] SHAH, N. B., BALAKRISHNAN, S. and WAINWRIGHT, M. J. (2016). Feeling the Bern: Adaptive estimators for Bernoulli probabilities of pairwise comparisons. In *Information Theory* (*ISIT*), 2016 *IEEE International Symposium on* 1153–1157. IEEE.

[48] SHAH, N. B., BALAKRISHNAN, S. and WAINWRIGHT, M. J. (2016). A permutation-based model for crowd labeling: Optimal estimation and robustness. ArXiv preprint. Available at arXiv:1606.09632.

[49] SHAH, N. B., BRADLEY, J. K., PAREKH, A., WAINWRIGHT, M. and RAMCHANDRAN, K. (2013). A case for ordinal peer-evaluation in MOOCs. In *NIPS Workshop on Data Driven Education*.

[50] SHAH, N. B. and WAINWRIGHT, M. J. (2017). Simple, robust and optimal ranking from pairwise comparisons. *J. Mach. Learn. Res.* **18** 199. MR3827087

[51] SREBRO, N. and SALAKHUTDINOV, R. R. (2010). Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *Advances in Neural Information Processing Systems* 2056–2064.

[52] STEWART, N., BROWN, G. D. A. and CHATER, N. (2005). Absolute identification by relative judgment. *Psychol. Rev.* **112** 881–911. https://doi.org/10.1037/0033-295X.112.4.881

[53] THURSTONE, L. L. (1927). A law of comparative judgment. *Psychol. Rev.* **34** 273.

[54] WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics*: *A Non-asymptotic Viewpoint*. Cambridge Univ. Press, Cambridge, UK.

[55] WAUTHIER, F., JORDAN, M. and JOJIC, N. (2013). Efficient ranking from pairwise comparisons. In *International Conference on Machine Learning* 109–117.

[56] YU, A. and GRAUMAN, K. (2014). Fine-grained visual comparisons with local learning. In *Computer Vision and Pattern Recognition* (*CVPR*) 192–199.