Two-Sample Testing on Pairwise Comparison Data and the Role of Modeling Assumptions

Charvi Rastogi[1], Sivaraman Balakrishnan[2], Nihar Shah[1,3], Aarti Singh[1]
Machine Learning Department[1], Department of Statistics[2], Computer Science Department[3]
Carnegie Mellon University
Email: {crastogi@cs, siva@stat, nihars@cs, aarti@cs}.cmu.edu

***Abstract*—A number of applications require two-sample testing of pairwise comparison data. For instance, in crowdsourcing, there is a long-standing question of whether comparison data provided by people is distributed similar to ratings-converted-to-comparisons. Other examples include sports data analysis and peer grading. In this paper, we design a two-sample test for pairwise comparison data. We establish an upper bound on the sample complexity required to correctly distinguish between the distributions of the two sets of samples. Our test requires essentially no assumptions on the distributions. We then prove complementary information-theoretic lower bounds showing that our results are tight (in the minimax sense) up to constant factors. We also investigate the role of modeling assumptions by proving information-theoretic lower bounds for a range of pairwise comparison models (WST, MST, SST, parameter-based such as BTL and Thurstone).**

I. Introduction

Data in the form of pairwise comparisons arises in a wide variety of settings. For instance, when eliciting data from people (say, in crowdsourcing), there is a long-standing debate over two forms of data collection: asking people to compare pairs of items or asking people to provide numeric scores to the items. A natural question here is whether people implicitly generate pairwise comparisons using a fundamentally different mechanism than first forming numeric scores and then converting them to a comparison. Thus, we are interested in testing if the data obtained from pairwise comparisons is distributed in a manner similar to if the numeric scores were converted to pairwise comparisons [21], [22]. Or, for instance, in sports and online games, a match between two players or two teams is a pairwise comparison between them [10], [11]. Here again arises a natural question of whether the relative performance of the teams has changed significantly across a certain period of time (e.g., to design an appropriate rating system [5]). A third example is peer grading where students are asked to compare pairs of homeworks [14], [21]. A question of interest here is whether a certain group of students (female/senior/...) grade very differently as compared to another group (male/junior/...) [24].

Each of the aforementioned problems involves two-sample testing. With this motivation, in this paper we consider the problem of two-sample testing on pairwise comparison data. Specifically, consider a collection of items (e.g., teams in a sports league). The data we consider comprises comparisons between pairs of these items, where the outcome of a comparison involves one of the items beating the other. In the two-sample testing problem, we have access to two sets of such pairwise comparisons, obtained from two different sources (e.g., the current season in a sports league forming one set of pairwise comparisons and the previous season forming a second set). The goal is to test whether the underlying distributions (winning probabilities) in the two sets of data are identical or (significantly) different.

Contributions. We now outline the contributions of this paper, also summarized in Table I.

- We present a test for two-sample testing on pairwise comparison data and associated upper bounds on its sample complexity. Our test makes essentially no assumptions on the outcome probabilities of the pairwise comparisons.
- We prove information-theoretic lower bounds on the critical testing radius for this problem. Our bounds show that our test is minimax optimal for this problem.
- Finally, we investigate the role of modeling assumptions. We show that our test is minimax optimal under WST and MST models. We prove an information-theoretic lower bound under the SST and parameter-based models. We also provide a computational lower bound for the SST model with single observation per pair of items, which matches the sample complexity bound attained by our test.

Related literature. The problem of two-sample testing of pairwise comparisons is at the intersection of two rich areas of research – two-sample testing and analyzing pair-

978-1-7281-6432-8/20/$31.00 ©2020 IEEE

Model ($\mathcal{M}$)	Upper Bound	Lower Bound	Computational Lower Bound
Model-free	$\epsilon^2_{\mathcal{M}} \le c\frac{\mathbb{I}(k>1)}{kd}$ (Thm. 1)	$\epsilon^2_{\mathcal{M}} > c\frac{\mathbb{I}(k>1)}{kd}$ (Prop. 1)	$\epsilon^2_{\mathcal{M}} > c\frac{\mathbb{I}(k>1)}{kd}$
WST and MST	$\epsilon^2_{\mathcal{M}} \le c\frac{1}{kd}$	$\epsilon^2_{\mathcal{M}} > c\frac{1}{kd}$ (Thm. 2)	$\epsilon^2_{\mathcal{M}} > c\frac{1}{kd}$
SST	$\epsilon^2_{\mathcal{M}} \le c\frac{1}{kd}$	$\epsilon^2_{\mathcal{M}} > c\frac{1}{kd^{3/2}}$	for $k=1$, $\epsilon^2_{\mathcal{M}} > \frac{c}{kd}$ (Thm. 4)
Parameter-based	$\epsilon^2_{\mathcal{M}} \le c\frac{1}{kd}$	$\epsilon^2_{\mathcal{M}} > c\frac{1}{kd^{3/2}}$ (Thm. 3)	$\epsilon^2_{\mathcal{M}} > c\frac{1}{kd^{3/2}}$

TABLE I: In this table, we summarize our results for the two-sample testing problem in (1) for different pairwise comparison models. Here, d denotes the number of items, and we obtain k samples (comparisons) per pair of items from each of the two populations. The probability distributions of the outcomes are governed by the matrices P and Q. In this work, we provide upper and lower bounds on the critical testing radius $\epsilon_{\mathcal{M}}$, defined in (3). The upper bound in Theorem 1 is due to the test in Algorithm 1 which is computationally efficient. We note that the constant c varies from result to result.

wise comparison data. The problem of two-sample testing has a long history in statistics [15, and references therein]. Several recent works have studied the minimax rate of two-sample testing for high-dimensional multinomials [2], [27], [28], and testing for sparseness in regression [4], [7]. We build on some of these ideas in our work. We also note the paper [17] which proposes a kernel-based two-sample test for distributions over permutations (i.e. distributions over complete rankings and not pairwise comparisons).

The analysis of pairwise comparison data goes back to the seminal work [26] and subsequently [3] and [16]. A number of papers [6], [18], [20], [22, and references therein] analyze parameter-based models such as the BTL and the Thurstone models. Here the goal is usually to estimate the parameters of the model or the underlying ranking of the items. Of particular interest is [1] which suggests some simple statistics to test for change in the performance of sports teams over time, and leaves designing principled tests as an open problem. To this end, we provide a two-sample test without any assumptions and with rigorous guarantees, and also use it subsequently to conduct such a test on real-world data. Some recent papers [9], [23], [25] focus on the role of the modeling assumptions in estimation and ranking from pairwise comparisons. We study the role of modeling assumptions for two-sample testing and prove performance guarantees for some pairwise comparison models.

II. Problem Setting

Our focus in this paper is on the two-sample testing problem where the two sets of samples come from two potentially different populations. Specifically, consider a collection of d items. The two sets of samples comprise outcomes of comparisons between various pairs of these items. In the first set of samples, the outcomes are governed by an unknown matrix $P \in [0,1]^{d\times d}$. The $(i,j)^{\text{th}}$ entry of matrix P is denoted as p_{ij}, and any comparison between items i and j results in i beating j with probability p_{ij}, independent of all else. We assume there are no ties. Analogously, the second set of samples comprises outcomes of pairwise comparisons between the d items governed by a (possibly different) unknown matrix $Q \in [0,1]^{d\times d}$, wherein item i beats item j with probability q_{ij}, the $(i,j)^{\text{th}}$ entry of matrix Q. For any pair (i,j) of items, we let k^p_{ij} and k^q_{ij} denote the number of times a pair of items (i,j) is compared in the first and second set of samples respectively. Finally, we let X_{ij} denote the number of times item $i \in [d]$ beats item $j \in [d]$ in the first set of samples, and let Y_{ij} denote the analogous quantity in the second set of samples. It follows that X_{ij} and Y_{ij} are Binomial random variables independently distributed as $X_{ij} \sim \text{Bin}(k^p_{ij}, p_{ij})$ and $Y_{ij} \sim \text{Bin}(k^q_{ij}, q_{ij})$. We adopt the convention $X_{ij} = 0$ when $k^p_{ij} = 0$, and $Y_{ij} = 0$ when $k^q_{ij} = 0$, and $k^p_{ii} = k^q_{ii} = 0$.

Hypothesis test. Consider any class $\mathcal{M}$ of pairwise-comparison probability matrices, and any given parameter $\epsilon > 0$. Then the goal is to test the hypotheses

$$\begin{aligned} &H_0 : P = Q \\ &H_1 : \frac{1}{d}|\!|\!|P-Q|\!|\!|_{\text{F}} \ge \epsilon, \end{aligned} \tag{1}$$

where $P, Q \in \mathcal{M}$.

A. Hypothesis Testing and Risk

We now provide a brief background on hypothesis tests. Consider the hypothesis testing problem defined in (1). We define a test ϕ as $\phi : \{k^p_{ij}, k^q_{ij}, X_{ij}, Y_{ij}\}_{(i,j)\in[d]^2} \rightarrow \{0,1\}$. Let $\mathbb{P}_0$ and $\mathbb{P}_1$ denote the distribution of the input variables under the null and under the alternate respectively. Let $\mathcal{M}_0$ and $\mathcal{M}_1$ denote the set of matrix pairs (P, Q) that satisfy the null condition and the alternate condition in (1) respectively. Then, we define the minimax risk as

$$\mathcal{R}_{\mathcal{M}} = \inf_{\phi}\{ \sup_{(P,Q)\in\mathcal{M}_0} \mathbb{P}_0(\phi = 1) + \sup_{(P,Q)\in\mathcal{M}_1} \mathbb{P}_1(\phi = 0)\}, \tag{2}$$

where the infimum is over all $\{0, 1\}-$valued tests ϕ. The corresponding critical radius is the smallest value ϵ for which a hypothesis test has non-trivial power. Formally, we define the critical radius as

$$\epsilon_{\mathcal{M}} = \inf\{\epsilon : \mathcal{R}_{\mathcal{M}} \leq 1/3\}. \tag{3}$$

The constant $1/3$ is arbitrary; we could use any specified constant in $(0, 1)$.[1] In this paper, we focus on providing tight bounds on the critical testing radius.

Some pairwise comparison models in the literature. A model for the pairwise comparison probabilities is a set of matrices in $[0, 1]^{d\times d}$. In the context of our problem setting, assuming a model means that the matrices P and Q are guaranteed to be drawn from this set. In this paper, the proposed test and the associated guarantees do not make any assumptions on the pairwise comparison probability matrices P and Q, that is, we allow P and Q to be any arbitrary matrices in $[0, 1]^{d\times d}$. However, there are a number of popular models in the literature on pairwise comparisons, and we provide a brief overview of them here. In what follows, we let $M \in [0, 1]^{d\times d}$ denote a generic pairwise comparison probability matrix and the models impose conditions on the matrix M.

Parameter-based models: A parameter-based model is associated with some known, non-decreasing function $f : \mathbb{R} \to [0, 1]$ such that $f(\theta) = 1 - f(-\theta) \quad \forall\ \theta \in \mathbb{R}$. We refer to any such function f as being "valid". The parameter-based model associated to a given valid function f is given by

$$M_{ij} = f(w_i - w_j) \quad \text{for all pairs} \quad (i, j), \tag{4}$$

for some unknown vector $w \in \mathbb{R}^d$ that represents the notional qualities of the d items. It is typically assumed that the vector w satisfies the conditions $\sum_{i\in[d]} w_i = 0$ and that $\|w\|_\infty$ is bounded above by a known constant.

Bradley-Terry-Luce (BTL) model: This is a specific parameter-based model with $f(\theta) = \dfrac{1}{1 + e^{-\theta}}$.

Thurstone model: This is a specific parameter-based model with $f(\theta) = \Phi(\theta)$, where Φ is the Gaussian CDF.

Strong stochastic transitivity (SST): The model assumes that the set of items $[d]$ is endowed with an unknown total ordering π, where $\pi(i) < \pi(j)$ implies that item i is preferred to item j. A matrix $M \in [0, 1]^{d\times d}$ is said to follow the SST model if it satisfies $M_{ij} = 1 - M_{ji}$ for every pair $i, j \in [d]$ and the condition

$$\begin{aligned} &M_{i\ell} \geq M_{j\ell} \quad \text{for every} \quad i, j \in [d] \\ &\qquad \text{such that } \pi(i) < \pi(j) \text{ and for every } \ell \in [d]. \end{aligned} \tag{5}$$

Moderate stochastic transitivity (MST): The model assumes that the set of items $[d]$ is endowed with an unknown total ordering π. A matrix $M \in [0, 1]^{d\times d}$ is said to follow the MST model if it satisfies $M_{ij} = 1 - M_{ji}$ for every pair $i, j \in [d]$ and the condition

$$\begin{aligned} &M_{i\ell} \geq \min\{M_{ij}, M_{j\ell}\} \quad \text{for every} \quad i, j, \ell \in [d] \\ &\qquad\qquad \text{such that } \pi(i) < \pi(j) < \pi(\ell). \end{aligned} \tag{6}$$

Weak stochastic transitivity (WST): The model assumes that the set of items $[d]$ is endowed with an unknown total ordering π. A matrix $M \in [0, 1]^{d\times d}$ is said to follow the WST model if it satisfies $M_{ij} = 1 - M_{ji}$ for every pair $i, j \in [d]$ and the condition

$$M_{ij} \geq \frac{1}{2} \quad \text{for every} \quad i, j \in [d] \quad \text{such that } \pi(i) < \pi(j).$$

Model hierarchy: Note that there is a structured hierarchy between these models, that is, {BTL, Thurstone} ⊂ parameter-based ⊂ SST ⊂ MST ⊂ WST ⊂ model-free.

III. Main Results

We now present the main theoretical results of this paper.

A. Test and guarantees

Our first result provides an algorithm for the two-sample testing problem (1), and associated upper bounds on its critical radius. Importantly, we do not make any modeling assumptions on the probability matrices P and Q. We consider a random-design setup wherein for every pair of items (i, j), the sample sizes k^p_{ij}, k^q_{ij} are drawn iid from some distribution $\mathcal{D}$ supported over non-negative integers. Let μ and σ denote the mean and standard deviation of distribution $\mathcal{D}$ respectively, and let $p_1 := \Pr_{Z\sim\mathcal{D}}(Z = 1)$. We assume that $\mathcal{D}$ has a finite mean and

$$\mu \geq c_1 p_1; \quad \mu \geq c_2\sigma, \tag{8}$$

for some constants $c_1 > 1$ and $c_2 > 1$. Many commonly occurring distributions obey these properties, for instance, Binomial distribution, Poisson distribution, geometric distribution and discrete uniform distribution.

[1] Our designed test can guarantee any desired error level, as discussed in the sequel.

Input : Samples X_{ij}, Y_{ij} denoting the number of times item i beat item j in the observed k^p_{ij}, k^q_{ij} pairwise comparisons from populations denoted by probability matrices P, Q respectively

Test Statistic :

$$T = \sum_{i=1}^{d} \sum_{j=1}^{d} \mathbb{I}_{ij} \frac{k^q_{ij}(k^q_{ij}-1)(X^2_{ij}-X_{ij}) + k^p_{ij}(k^p_{ij}-1)(Y^2_{ij}-Y_{ij}) - 2(k^p_{ij}-1)(k^q_{ij}-1)X_{ij}Y_{ij}}{(k^p_{ij}-1)(k^q_{ij}-1)(k^p_{ij}+k^q_{ij})} \tag{7}$$

where $\mathbb{I}_{ij} = \mathbb{I}(k^p_{ij} > 1) \times \mathbb{I}(k^q_{ij} > 1)$.

Output : If $T \geq 11d$, then reject the null.

Algorithm 1: Model-free two-sample test with pairwise comparisons

Our test is presented in Algorithm 1. The following theorem characterizes the performance of this test, thereby establishing an upper bound on the critical radius of this two-sample testing problem in a random-design setting.

Theorem 1. *Consider the testing problem in* (1) *with $\mathcal{M}$ as the class of all pairwise probability matrices. Suppose the number of comparisons between the two populations k^p_{ij}, k^q_{ij} are drawn iid from some distribution $\mathcal{D}$ that satisfies* (8) *(for all $i \neq j$ in the asymmetric setting and all $i < j$ in the symmetric setting). There is a constant $c > 0$ such that if $\epsilon^2 \geq \dfrac{c}{\mu d}$, then the sum of Type I error and Type II error of Algorithm 1 is at most $\frac{1}{3}$.*

As a consequence of Theorem 1, to control the probability of error, we get a sufficient condition on the (per pair) sample complexity as $k \geq \frac{c}{d\epsilon^2}$. Theorem 1 provides an upper bound of $\frac{1}{3}$ on probability of error, and this number is closely tied to the threshold used in the test above. More generally, for any specified constant $\nu \in (0, 1)$, the test achieves a probability of error at most ν by setting the threshold as $d\sqrt{\frac{24(2-\nu)}{\nu}}$, with the same order of sample complexity as in Theorem 1. Moreover, if the sample complexity is increased by some factor R, then running Algorithm 1 on R independent instances of the data and taking the majority answer results in exponential $(\exp(-2R))$ decrease in probability of error. Finally, we note that a sharper but non-explicit threshold can be obtained using the permutation test method [8] to control the Type I error.

Proof Sketch for Theorem 1. The test statistic T in (7) is designed to ensure that $\mathbb{E}_{H_0}[T] = 0$ for any values of $\{k^p_{ij}, k^q_{ij}\}_{1 \leq i,j \leq d}$. We lower bound the expected value of T under the alternate hypothesis as $\mathbb{E}_{H_1}[T] \geq c\mu d^2 \epsilon^2$. Next, we prove upper bounds for $\text{Var}[T]$ under both the null and the alternate. This allows us to choose a threshold value of $11d$. Finally, using Chebyshev's inequality, we obtain the claimed upper bound on the sample complexity with guarantees on Type I and Type II error.

B. Information-theoretic converse results and the role of modeling assumptions

In this section we look at the role of modeling assumptions on the pairwise comparison probability matrices in the two-sample testing problem in (1).

Information-Theoretic Lower Bound for MST Class. Having established an upper bound on the rate of two-sample testing without modeling assumptions on the pairwise comparison probability matrices P, Q, we show matching lower bounds that hold under the MST class.

Theorem 2. *Consider the testing problem in* (1) *with $\mathcal{M}$ as the class of matrices described by the MST model. Suppose we have k comparisons for each pair (i, j) from each population. There exists a constant $c > 0$, such that the critical radius $\epsilon_{\mathcal{M}}$ is lower bounded as $\epsilon^2_{\mathcal{M}} > \dfrac{c}{kd}$.*

This lower bound matches the bound derived for Algorithm 1 in Theorem 1, thereby establishing the information-theoretic minimax optimality of our algorithm (up to constant factors) under MST, WST modeling assumptions in addition to the model-free setting. We provide a proof sketch for Theorem 2 in Section III-B.

Necessity of $\mu > p_1$. Recall that the upper bound derived in Theorem 1 under the model-free setting holds under the assumption that $\mu \geq c_1 p_1$ with $c_1 > 1$, as stated in (8). We now state a negative result for the case $\mu \leq p_1$ (which implies that $k^p_{ij}, k^q_{ij} \leq 1 \, \forall \, (i, j)$).

Proposition 1. *Consider the testing problem in* (1) *with $\mathcal{M}$ as the class of all pairwise probability matrices. Suppose we have at most one comparison for each pair (i, j) from each population (for all $i \neq j$ in the asymmetric setting and all $i < j$ in the symmetric setting). Then, the critical radius in* (3) *does not exist, that is, for any value of ϵ, the minimax risk is at least $\frac{1}{2}$.*

If $k^p_{ij} = k^q_{ij} \leq 1 \, \forall \, (i, j)$, then at best we have first order information of each entry of P and Q, that is, one has access to only $\Pr(X_{ij} = 1), \Pr(Y_{ij} = 1), \Pr(X_{ij} =$

$1, Y_{ij} = 1)$ for each pair (i, j). This observation leads to an example where the null and the alternate cannot be distinguished from each other by any test.

Information-Theoretic Lower Bound for Parameter-Based Class. We now prove a lower bound for our two-sample testing problem (1) wherein the probability matrices follow the parameter-based model.

Theorem 3. *Consider the testing problem in* (1) *with* $\mathcal{M}$ *as the parameter-based class of probability matrices. Suppose we have* k *comparisons for each pair* (i, j) *from each population. There exists a constant* $c > 0$*, such that the critical radius* $\epsilon_{\mathcal{M}}$ *is lower bounded as* $\epsilon^2_{\mathcal{M}} > \frac{c}{kd^{3/2}}$.

This lower bound also applies to probability matrices in the SST class described in (5). We provide a brief proof sketch in Section III-B.

Computational Lower Bound for SST Class. Given the gap between Theorem 1 and Theorem 3, it is natural to wonder whether there is another polynomial-time testing algorithm for testing under the SST and/or parameter-based modeling assumption. We answer this question in the negative, for the SST model and single observation setup ($k = 1$), conditionally on the average-case hardness of the planted clique problem [12], [13]. In informal terms, the planted clique conjecture asserts that there is no polynomial-time algorithm that can detect the presence of a planted clique of size $\kappa = o(\sqrt{d})$ in an Erdős-Rényi random graph with d nodes. We construct SST matrices that are similar to matrices in the planted clique problem. Then, as a direct consequence of the planted clique conjecture, we have the following result.

Theorem 4. *Consider the testing problem in* (1) *with* $\mathcal{M}$ *as the class of matrices described by the SST model. Suppose the planted clique conjecture holds. Suppose we have one comparison for each pair* (i, j) *from each population. Then there exists a constant* $c > 0$ *such that for polynomial-time testing algorithms the critical radius* $\epsilon_{\mathcal{M}}$ *is lower bounded as* $\epsilon^2_{\mathcal{M}} > \frac{c}{d}$.

Thus, for $k = 1$, the computational lower bound on the testing rate for the SST model matches the rate derived for Algorithm 1 (up to constant factors).

Proof sketches for Theorem 2 and Theorem 3: To prove the information-theoretic lower bound under the different modeling assumptions, we construct a null and alternate belonging to the corresponding class of probability matrices. The bulk of our technical effort is devoted to upper bounding the χ^2 divergence between the probability measure under the null and the alternate. We then invoke Le Cam's lower bound for testing to obtain a lower bound on the minimax risk which gives us the information-theoretic lower bound. We now look at the constructions for the two modeling assumptions.

Lower bound construction for MST class. We construct a null and alternate such that under the null $P = Q = [\frac{1}{2}]^{d\times d}$ and under the alternate $P = [\frac{1}{2}]^{d\times d}$ and $Q \sim \text{Unif}(\Theta)$ with $\frac{1}{d}|||P - Q|||_{\text{F}} = \epsilon$. Here Θ is a set of matrices following the MST model, in which the upper right quadrant has exactly one entry equal to $\frac{1}{2} + \eta$ in each row and each column and the remaining entries above the diagonal are $\frac{1}{2}$. The entries below the diagonal follow from skew symmetry.

Lower bound construction for parameter-based class. The construction is same as the construction given above except we define a different set Θ of probability matrices. According to the parameter-based model, the matrices P and Q depend on the vectors $w_p \in \mathbb{R}^d$ and $w_q \in \mathbb{R}^d$ respectively. Now, for simplicity in this sketch, suppose that d is even. We set $w_p = [0, \cdots, 0]$, which fixes $p_{ij} = \frac{1}{2} \; \forall \, (i, j)$. For the alternate, consider a collection of vectors, w_Q, each with half the entries as δ and the other half as $-\delta$, thereby ensuring that $\sum_{i\in[d]} w_i = 0$. We set δ to ensure that each of the probability matrices induced by this collection of vectors obey $\frac{1}{d}|||P-Q|||_{\text{F}} = \epsilon$. We then consider the setting where Q is chosen uniformly at random from the set of pairwise comparison probability matrices induced by the collection of values of w_Q.

IV. Discussion

In an extended version of this paper, we use our test to understand two important empirical questions based on real-world datasets. In a setting where crowdsourcing workers provide ratings and comparisons over a set of objects, our test concludes there is a statistically significant difference (p = 0.003) between comparisons and ratings-converted-to-comparisons given by people. As a second experiment, we test for difference in relative performance of teams in consecutive seasons of the European football leagues and our test does not find any significant difference ($p = 0.97$).

On the theoretical front, there is a gap between the testing rate of our algorithm and our information-theoretic lower bound for the SST and parameter-based models, and closing this gap is an open problem of interest. In the future, our work may also help address open problems of two-sample testing pertaining to more general aspects of data from people such as partial or total rankings [17], function evaluations [19], and strategic behavior [29].

Acknowledgments: This work was supported in part by NSF grants 1763734, 1755656 and 1942124.

References

[1] D. Aldous. Elo ratings and the sports model: A neglected topic in applied probability? *Statistical Science*, 32(4):616–629, 2017.

[2] S. Balakrishnan and L. Wasserman. Hypothesis testing for high-dimensional multinomials: A selective review. *Ann. Appl. Stat.*, 12(2):727–749, 2018.

[3] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, pages 324–345, 1952.

[4] A. Carpentier, O. Collier, L. Comminges, A. B. Tsybakov, and Y. Wang. Minimax rate of testing in sparse linear regression. *arXiv:1804.06494*, 2018.

[5] M. Cattelan, C. Varin, and D. Firth. Dynamic Bradley–Terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1):135–150, 2013.

[6] Y. Chen and C. Suh. Spectral MLE: Top-k rank aggregation from pairwise comparisons. In *ICML*, pages 371–380, 2015.

[7] O. Collier, L. Comminges, and A. B. Tsybakov. Minimax estimation of linear and quadratic functionals on sparsity classes. *The Annals of Statistics*, 45(3):923–958, 2017.

[8] P. Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer Series in Statistics. Springer, 2013.

[9] R. Heckel, N. B. Shah, K. Ramchandran, and M. J. Wainwright. Active ranking from pairwise comparisons and when parametric assumptions don't help. *arxiv:1606.08842*, 2016.

[10] R. Herbrich, T. Minka, and T. Graepel. Trueskill: A Bayesian skill rating system. In *Advances in neural information processing systems*, 2007.

[11] L. M. Hvattum and H. Arntzen. Using Elo ratings for match result prediction in association football. *International Journal of forecasting*, 26(3), 2010.

[12] M. Jerrum. Large cliques elude the metropolis process. *Random Struct. Algorithms*, 1992.

[13] L. Kučera. Expected complexity of graph partitioning problems. *Discrete Appl. Math.*, 57, 1995.

[14] A. Lamon, D. Comroe, P. Fader, D. McCarthy, R. Ditto, and D. Huesman. Making WHOOPPEE: A collaborative approach to creating the modern student peer assessment ecosystem. In *EDUCAUSE*, 2016.

[15] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, third edition, 2005.

[16] R. D. Luce. *Individual choice behavior: A theoretical analysis*. New York: Wiley, 1959.

[17] H. Mania, A. Ramdas, M. J. Wainwright, M. I. Jordan, and B. Recht. On kernel methods for covariates that are rankings. *Electronic Journal of Statistics*, 12(2):2537–2577, 2018.

[18] S. Negahban, S. Oh, and D. Shah. Iterative ranking from pairwise comparisons. In *Advances in neural information processing systems*, 2012.

[19] R. Noothigattu, N. B. Shah, and A. D. Procaccia. Loss functions, axioms, and peer review. *arXiv preprint arXiv:1808.09057*, 2018.

[20] A. Rajkumar and S. Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *ICML*, pages 118–126, 2014.

[21] K. Raman and T. Joachims. Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1037–1046, 2014.

[22] N. B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *JMLR*, 17(1), 2016.

[23] N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *IEEE Transactions on Information Theory*, 63(2):934–959, 2017.

[24] N. B. Shah, B. Tabibian, K. Muandet, I. Guyon, and U. Von Luxburg. Design and analysis of the NIPS 2016 review process. *JMLR*, 19, 2018.

[25] N. B. Shah and M. J. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *JMLR*, 18(199):1–38, 2018.

[26] L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273, 1927.

[27] G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

[28] P. Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6), 2011.

[29] Y. Xu, H. Zhao, X. Shi, and N. B. Shah. On strategyproof conference peer review. In *IJCAI*, 2019.