Journal of Machine Learning Research 18 (2018) 1-38 Submitted 4/16; Revised 11/17; Published 4/18

Simple, Robust and Optimal Ranking from Pairwise Comparisons

Nihar B. Shah NIHARS@CS.CMU.EDU
Machine Learning Department and Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Martin J. Wainwright WAINWRIG@BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences and Department of Statistics
University of California
Berkeley, CA 94720, USA

Editor: Sujay Sanghavi

Abstract

We consider data in the form of pairwise comparisons of n items, with the goal of identifying the top k items for some value of $k < n$, or alternatively, recovering a ranking of all the items. We analyze the Borda counting algorithm that ranks the items in order of the number of pairwise comparisons won, and show it has three attractive features: (a) it is an optimal method achieving the information-theoretic limits up to constant factors; (b) it is robust in that its optimality holds without imposing conditions on the underlying matrix of pairwise-comparison probabilities, in contrast to some prior work that applies only to the BTL parametric model; and (c) its computational efficiency leads to speed-ups of several orders of magnitude. We address the problem of exact recovery, and for the top-k recovery problem we also extend our results to obtain sharp guarantees for approximate recovery under the Hamming distortion metric, and more generally, to any arbitrary error requirement that satisfies a simple and natural monotonicity condition. In doing so, we introduce a general framework that allows us to treat a variety of problems in the literature in an unified manner.

Keywords: Pairwise comparisons, Ranking, Set recovery, Approximate recovery, Borda count, Permutation-based models, Occam's razor

1. Introduction

Ranking problems involve a collection of n items, and some unknown underlying total ordering of these items. In many applications, one may observe noisy comparisons between various pairs of items. Examples include matches between football teams in tournament play; consumer's preference ratings in marketing; and certain types of voting systems in politics. Given a set of such noisy comparisons between items, it is often of interest to find the true underlying ordering of all n items, or more generally, given some given positive integer $k \leq n$, to find the subset of k most highly rated items. These two problems are the focus of this paper.

There is a substantial literature on the problem of finding approximate rankings based on noisy pairwise comparisons. A number of papers (e.g., Kenyon-Mathieu and Schudy, 2007;

Braverman and Mossel, 2008; Eriksson, 2013) consider models in which the probability of a pairwise comparison agreeing with the underlying order is identical across all pairs. These results break down when, for one or more pairs, the probability of agreeing with the underlying ranking is close to or exactly equal to $\frac{1}{2}$. Another set of papers (Hunter, 2004; Negahban et al., 2012; Hajek et al., 2014; Soufiani et al., 2014; Shah et al., 2016a) work using parametric models of pairwise comparisons, and address the problem of recovering the parameters associated to every individual item. A more recent line of work (Chatterjee, 2014; Shah et al., 2017a, 2016d) studies a more general class of models based on the notion of strong stochastic transitivity (SST), and derives conditions on recovering the pairwise comparison probabilities themselves. However, it remains unclear whether or not these results can directly extend to tight bounds for the problem of recovery of the top k items. Another line of work (Jagabathula and Shah, 2008; Mitliagkas et al., 2011; Ammar and Shah, 2012; Ding et al., 2015) focuses on mixture models, in which every pairwise comparison is associated to a certain individual making the comparison, and it is assumed that the preferences across individuals can be described by a low-dimensional model.

Most related to our work are the papers by Wauthier et al. (2013); Rajkumar and Agarwal (2014); Rajkumar et al. (2015), and Chen and Suh (2015), which we briefly discuss here and in a more detailed manner in the sequel. Wauthier et al. (2013) analyze a weighted counting algorithm to recover approximate rankings; their analysis applies to a specific model in which the pairwise comparison between any pair of items remains faithful to their relative positions in the true ranking with a probability common across all pairs. They consider recovery of an approximate ranking under Kendall's tau and maximum displacement metrics, but do not provide results on exact recovery. As the analysis of this paper shows, their bounds are quite loose: more precisely, their results are tight only when there are a total of at least $\Theta(n^2)$ comparisons. Two other papers (Rajkumar and Agarwal, 2014; Rajkumar et al., 2015) consider ranking under several models and several metrics. In the part that is common with our setting, they show that the counting algorithm is consistent in terms of recovering the full ranking, which automatically implies consistency in exactly recovering the top k items. They obtain upper bounds on the sample complexity in terms of a separation threshold that is identical to a parameter Δ_k defined subsequently in this paper (see Section 3). However, as our analysis shows, their bounds are loose by at least an order of magnitude. They also assume a certain high-SNR condition on the probabilities, an assumption that is not imposed in our analysis.

Finally, in very recent work on this problem, Chen and Suh (2015) proposed an algorithm called the Spectral MLE for exact recovery of the top k items. They showed that, if the pairwise observations are assumed to drawn according to the Bradley-Terry-Luce (BTL) parametric model (Bradley and Terry, 1952; Luce, 1959), the Spectral MLE algorithm recovers the k items correctly with high probability under certain regularity conditions. In addition, they also show, via matching lower bounds, that their regularity conditions are tight up to constant factors. While these guarantees are attractive, it is natural to ask how such an algorithm behaves when the data is *not* drawn from the BTL model. In real-world instances of pairwise ranking data, it is often found that parametric models, such as the BTL model and its variants, fail to provide accurate fits (for instance, see the papers Davidson and Marschak, 1959; McLaughlin and Luce, 1965; Tversky, 1972; Ballinger and Wilcox, 1997 and references therein).

With this context, the main contribution of this paper is to analyze a classical counting-based method for ranking, often called the Borda count method (de Borda, 1781), and to show that it is optimal and robust. Our analysis does not require that the data-generating mechanism follow either the BTL or other parametric assumptions, nor other regularity conditions such as stochastic transitivity. We show that the Borda counting algorithm has the following properties:

- Simplicity: The algorithm is simple, as it just orders the items by the number of pairwise comparisons won. As we will subsequently see, the execution time of this counting algorithm is several orders of magnitude lower as compared to prior work on ranking from noisy pairwise comparisons.
- Optimality: We derive conditions under which the counting algorithm achieves the stated goals, and by means of matching information-theoretic lower bounds, show that these conditions are tight.
- Robustness: The guarantees that we prove do not require any assumptions on the pairwise-comparison probabilities, and the counting algorithm performs well for various classes of data sets. In contrast, we find that the spectral MLE algorithm performs poorly when the data is not drawn from the BTL model.

In doing so, we consider three different instantiations of the problem of set-based recovery:

(i) Recovering the top k items perfectly;

(ii) Recovering the top k items allowing for a certain Hamming error tolerance; and

(iii) A more general recovery problem for set families that satisfy a natural "set-monotonicity" condition. In order to tackle this general problem, we introduce a general framework that allows us to treat a variety of problems in the literature in an unified manner.

The remainder of this paper is organized as follows. We begin in Section 2 with background and a more precise formulation of the problem. Section 3 presents our main theoretical results on top-k recovery under various requirements. Section 4 provides the results of experiments on both simulated and real-world data sets. We provide all proofs in Section 5. The paper concludes with a discussion in Section 6.

2. Background and problem formulation

In this section, we provide a more formal statement of the problem along with background on various types of ranking models.

2.1 Problem statement

Given an integer $n \geq 2$, we consider a collection of n items, indexed by the set $[n] := \{1, \ldots, n\}$. For each pair $i \neq j$, we let M_{ij} denote the probability that item i wins the comparison with item j. We assume that each comparison necessarily results in one winner, meaning that

$$M_{ij} + M_{ji} = 1, \qquad \text{and} \quad M_{ii} = \frac{1}{2}, \tag{1}$$

where we set the diagonal as $\frac{1}{2}$ for concreteness.

For any item $i \in [n]$, we define an associated score τ_i as

$$\tau_i(M) := \frac{1}{n} \sum_{j=1}^{n} M_{ij}. \tag{2}$$

In words, the score $\tau_i(M)$ of any item $i \in [n]$ corresponds to the probability that item i beats an item chosen uniformly at random from all n items. In the sequel, we will use the shorthand τ_i for the score of any item i, and drop the dependence on M from the notation wherever the value of M is clear from context.

Given a set of noisy pairwise comparisons, our goals are (a) to recover the k items with the maximum values of their scores; and (b) to recover the full ordering of all the items as defined by the score vector. The notion of ranking items via their scores (2) generalizes the explicit rankings under popular models in the literature. Indeed, as we discuss shortly, most models of pairwise comparisons considered in the literature either implicitly or explicitly assume that the items are ranked according to their scores. Note that neither the scores $\{\tau_i\}_{i \in [n]}$ nor the matrix $M := \{M_{ij}\}_{i,j \in [n]}$ of probabilities are assumed to be known.

More concretely, we consider a random-design observation model defined as follows. Each pair is associated with a random number of noisy comparisons, following a binomial distribution with parameters (r, p), where $r \geq 1$ is the number of trials and $p \in (0, 1]$ is the probability of making a comparison on any given trial. Thus, each pair (i, j) is associated with a binomial random variable with parameters (r, p) that governs the number of comparisons between the pair of items. We assume that the observation sequences for different pairs are independent. Note that in the special case $p = 1$, this random binomial model reduces to the case in which we observe exactly r observations of each pair; in the special case $r = 1$, the set of pairs compared form an (n, p) Erdős-Rényi random graph.

In this paper, we begin in Section 3.1 by analyzing the problem of exact recovery. More precisely, for a given matrix M of pairwise probabilities, suppose that we let $\mathcal{S}_k^*$ denote the (unknown) set of k items with the largest values of their respective scores, assumed to be unique for concreteness.

Given noisy observations specified by the pairwise probabilities M, our goal is to establish conditions under which there exists some algorithm $\widehat{\mathcal{S}}_k$ that identifies k items based on the outcomes of various comparisons such that the probability $\mathbb{P}_M(\widehat{\mathcal{S}}_k = \mathcal{S}_k^*)$ is very close to one. In the case of recovering the full ranking, our goal is to identify conditions which ensure that the probability $\mathbb{P}_M\big(\bigcap_{k \in [n]} (\widehat{\mathcal{S}}_k = \mathcal{S}_k^*)\big)$ is close to one.

In Section 3.2, we consider the problem of recovering a set of k items that approximates $\mathcal{S}_k^*$ with a minimal Hamming error. For any two subsets of the set $[n]$, we define their Hamming distance D_{H}, also referred to as their Hamming error, to be the number of items that belong to exactly one of the two sets—that is

$$D_{\mathrm{H}}(A, B) = \mathrm{card}\Big(\{A \cup B\} \backslash \{A \cap B\}\Big). \tag{3}$$

For a given user-defined tolerance parameter $h \geq 0$, we derive conditions that ensure that $D_{\mathrm{H}}(\widehat{\mathcal{S}}_k, \mathcal{S}_k^*) \leq 2h$ with high probability.

Finally, we generalize our results to the problem of satisfying any a general class of requirements on set families. These requirement are specified in terms of which k-sized

subsets of the items are allowed, and is required to satisfy only one natural condition, that of set-monotonicity, meaning that replacing an item in an allowed set with a higher rank item should also be allowed. See Section 3.3 for more details on this general framework.

2.2 A range of pairwise comparison models

Our work makes minimal assumptions on the the pairwise comparison probabilities. Our model is based on a "permutation-based" approach, and is described below (see Shah et al., 2017a, 2016d,c, 2017b for other uses of permutation-based models and Shah, 2017, Chapter 1 and Part 1 for a general treatment). In order to put our work in context of the literature, we also briefly review some standard models used for pairwise comparison data – all of these models form special cases of our general model.

Our model: We assume that any requirements or metrics for recovery of a partial or total order of the items are governed by the scores of the items defined in equation (2). In other words, any item i is considered as ranked higher than any item j when their scores satisfy $\tau_i > \tau_j$. We make no other assumptions on the probabilities $\{M_{ij}\}_{i,j\in[n]}$. In what follows, we show that several other popular classes of models arise as special cases of our model.

Parametric models: A broad class of parametric models, including the Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Luce, 1959) as a special case, are based on assuming the existence of "quality" parameter $w_i \in \mathbb{R}$ for each item $i \in [n]$, and requiring that the probability of an item beating another is a specific function of the difference between their values. In the BTL model, the probability M_{ij} that i beats j is given by the logistic model

$$M_{ij} = \frac{1}{1 + e^{-(w_i - w_j)}}. \tag{4a}$$

More generally, parametric models assume that the pairwise comparison probabilities take the form

$$M_{ij} = F(w_i - w_j), \tag{4b}$$

where $F : \mathbb{R} \to [0, 1]$ is some strictly increasing cumulative distribution function. The function F is typically assumed to be known. By construction, any parametric model has the following property: if $w_i > w_j$ for some pair of items (i, j), then we are also guaranteed that $M_{i\ell} > M_{j\ell}$ for every item ℓ. As a consequence, we are guaranteed that $\tau_i > \tau_j$, which implies that ordering of the items in terms of their quality vector $w \in \mathbb{R}^n$ is identical to their ordering in terms of the score vector $\tau \in \mathbb{R}^n$. Consequently, if the data is actually drawn from a parametric model, then recovering the top k items according to their scores is the same as recovering the top k items according their respective quality parameters.

Strong Stochastic Transitivity (SST) class: The class of strong stochastic transitivity (SST) models is a superset of parametric models (Shah et al., 2017a). It does not assume the existence of a quality vector, nor does it assume any specific form of the probabilities as in equation (4a). Instead, the SST class is defined by assuming the existence of a total ordering of the n items, and imposing the inequality constraints $M_{i\ell} \geq M_{j\ell}$ for every pair of items (i, j) in which i is ranked above j in the ordering, and every item ℓ. One can verify

that an ordering by the scores $\{\tau_i\}_{i\in[n]}$ of the items lead to an ordering of the items that is consistent with that defined by the SST class.

Thus, we see that in a broad class of models for pairwise ranking, the total ordering defined by the score vector (2) coincides with the underlying ordering used to define the models. In this paper, we analyze the performance of a counting algorithm, essentially without imposing any modeling conditions on the family of pairwise probabilities. The next three sections establish theoretical guarantees on the recovery of the top k items under various requirements.

2.3 Borda counting algorithm

The analysis of this paper focuses on a simple counting-based algorithm, often called the Borda count method (de Borda, 1781). We employ this method here for the setting of pairwise comparisons, noting that the Borda count method more generally also supports comparisons between more than two items.

More precisely, for each distinct $i, j \in [n]$ and every integer $\ell \in [r]$, let $Y_{ij}^{\ell} \in \{-1, 0, +1\}$ represent the outcome of the ℓ^{th} comparison between the pair i and j, defined as

$$Y_{ij}^{\ell} = \begin{cases} 0 & \text{no comparison between } (i,j) \text{ in trial } \ell \\ +1 & \text{if comparison is made and item } i \text{ beats } j \\ -1 & \text{if comparison is made and item } j \text{ beats } i. \end{cases} \tag{5}$$

Note that this definition ensures that $Y_{ij}^{\ell} = -Y_{ji}^{\ell}$. For each $i \in [n]$, the quantity

$$N_i := \sum_{j\in[n]} \sum_{\ell\in[r]} \mathbf{1}\{Y_{ij}^{\ell} = 1\} \tag{6}$$

corresponds to the number of pairwise comparisons won by item i. Here we use $\mathbf{1}\{\cdot\}$ to denote the indicator function that takes the value 1 if its argument is true, and the value 0 otherwise. For each integer k, the vector $\{N_i\}_{i=1}^n$ of number of pairwise wins defines a k-sized subset

$$\widetilde{\mathcal{S}}_k = \Big\{ i \in [n] \mid N_i \text{ is among the } k \text{ highest number of pairwise wins} \Big\}, \tag{7}$$

corresponding to the set of k items with the largest values of N_i. In other words, the set $\widetilde{\mathcal{S}}_k$ corresponds to the rank statistics of the top k-items in the pairwise win ordering. (If there are any ties, we resolve them by choosing the indices with the smallest value of i.)

3. Main results

In this section, we present our main theoretical results on top-k recovery under the three settings described earlier. Note that the three settings are ordered in terms of increasing generality, with the advantage that the least general setting leads to the simplest form of theoretical claim. We also discuss optimal exact recovery of the full ranking.

3.1 Thresholds for exact recovery of the top k items

We begin with the goal of exactly recovering the k top-ranked items. As one might expect, the difficulty of this problem turns out to depend on the degree of separation between the top k items and the remaining items. More precisely, let us use (k) and $(k+1)$ to denote the indices of the items that are ranked k^{th} and $(k+1)^{th}$ respectively. With this notation, the *k-separation threshold* Δ_k is given by

$$\Delta_k(M) := \tau_{(k)}(M) - \tau_{(k+1)}(M) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} M_{(k)i}}_{\text{Term (i)}} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} M_{(k+1)i}}_{\text{Term (ii)}}. \tag{8}$$

In words, the quantity $\Delta_k(M)$ is the difference between (i) the probability of item (k) beating another item chosen uniformly at random, and (ii) the same probability for item $(k+1)$.

As shown by the following theorem, success or failure in recovering the top k entries is determined by the size of $\Delta_k(M)$ relative to the number of items n, observation probability p and number of repetitions r. In particular, consider the family of matrices

$$\mathcal{F}_k(\alpha; n, p, r) := \left\{ M \in [0,1]^{n\times n} \mid M + M^T = 11^T, \text{ and } \Delta_k(M) \geq \alpha\sqrt{\frac{\log n}{npr}} \right\}. \tag{9}$$

To simplify notation, we often adopt $\mathcal{F}_k(\alpha)$ as a convenient shorthand for this set, where its dependence on (n, p, r) should be understood implicitly.

With this notation, the achievable result in part (a) of the following theorem is based on the estimator that returns the set $\widetilde{\mathcal{S}}_k$ of the the k items defined by the number of pairwise comparisons won, as defined in equation (7). On the other hand, the lower bound in part (b) applies to *any estimator*, meaning any measurable function of the observations.

Theorem 1 *(a) Consider any $n \geq 2$, $r \geq 1$ and $p \in (0, 1]$. Then if $\alpha \geq 8$, the set $\widetilde{\mathcal{S}}_k$ of top k items (7) given by the Borda counting algorithm satisfies*

$$\sup_{M\in\mathcal{F}_k(\alpha)} \mathbb{P}_M\big[\widetilde{\mathcal{S}}_k \neq \mathcal{S}_k^*\big] \leq \frac{1}{n^{14}}. \tag{10a}$$

(b) Conversely, suppose that $n \geq 7$ and $p \geq \frac{\log n}{2nr}$. Then for any $\alpha \leq \frac{1}{7}$, the error probability of any *estimator $\widehat{\mathcal{S}}_k$ is lower bounded as*

$$\sup_{M\in\mathcal{F}_k(\alpha)} \mathbb{P}_M\big[\widehat{\mathcal{S}}_k \neq \mathcal{S}_k^*\big] \geq \frac{1}{7}. \tag{10b}$$

Remarks: First, it is important to note that the negative result in part (b) holds even if the supremum is further restricted to a particular parametric sub-class of $\mathcal{F}_k(\alpha)$, such as the pairwise comparison matrices generated by the BTL model, or by the SST model. The proof for the lower bound constructs a packing set of possible pairwise comparison probabilities

and then applies Fano's inequality. The construction ensures that every element of the packing set also lies in the parametric and SST models. The packing set is based on a generalization of a construction introduced by Chen and Suh (2015) for the BTL model, which we adapt to the general definition (8) of the separation threshold Δ_k.

Second, we note that in the regime $p < \frac{\log n}{2nr}$, standard results from random graph theory (Erdős and Rényi, 1960) can be used to show that there are at least $\sqrt{n}$ items (in expectation) that are never compared to any other item. Of course, estimating the rank is impossible in this pathological case, so we omit it from consideration.

Third, the two parts of the theorem in conjunction show that the counting algorithm is essentially optimal. The only room for improvement is in the difference between inequality $\alpha \geq 8$ in the achievable result, and $\alpha \leq \frac{1}{7}$ in the lower bound.

Theorem 1 can also be used to derive guarantees for recovery of other functions of the underlying ranking. Here we consider the problem of identifying the ranking of all n items, which we denote by the permutation π^*. In this case, we require that each of the separations $\{\Delta_j\}_{j=1}^{n-1}$ are suitably lower bounded: more precisely, we study models M that belong to the intersection $\cap_{j=1}^{n-1}\mathcal{F}_j(\alpha)$.

Theorem 2 *Consider any $n \geq 2$, $r \geq 1$ and $p \in (0, 1]$. Let $\widetilde{\pi}$ be the permutation of the items specified by the Borda counting algorithm in order of the number of pairwise comparisons won. Then for any $\alpha \geq 8$, we have*

$$\sup_{M \in \cap_{j=1}^{n-1}\mathcal{F}_j(\alpha)} \mathbb{P}_M\big[\widetilde{\pi} \neq \pi^*\big] \leq \frac{1}{n^{13}}.$$

Conversely, if $n \geq 9$, then the separation condition on $\{\Delta_j\}_{j=1}^{n-1}$ that defines the set $\cap_{j=1}^{n-1}\mathcal{F}_j(\alpha)$ is unimprovable beyond constant factors.

The upper bound of Theorem 2 follows from the equivalence of the correct recovery of the ranking with the recovery of the top k items for every value of $k \in [n]$. The proof of the lower bound requires a markedly different set of arguments; the proof does not follow from Theorem 1(b) since for any given value of k a condition of the form $\min_{j \in [n-1]} \Delta_j \leq \alpha$ in general does not imply $\Delta_k \leq \alpha$ which would otherwise be required to use Theorem 1(b).

Detailed comparison to related work: In the remainder of this subsection, we make a detailed comparison to the related works (Wauthier et al., 2013; Rajkumar and Agarwal, 2014; Rajkumar et al., 2015; Chen and Suh, 2015) that were briefly discussed in Section 1.

Wauthier et al. (2013) analyze a weighted counting algorithm for approximate recovery of rankings; they work under a model in which $M_{ij} = \frac{1}{2} + \gamma$ whenever item i is ranked above item j in an assumed underlying ordering. Here the parameter $\gamma \in (0, \frac{1}{2}]$ is independent of (i, j), and as a consequence, the best ranked item is assumed to be as likely to beat the worst item as it is to beat the second ranked item, for instance. They analyze approximate ranking under Kendall tau and maximum displacement metrics. In order to have a displacement upper bounded by by some $\delta > 0$, their bounds require the order of $\frac{n^5}{\delta^2\gamma^2}$ pairwise comparisons. In comparison, our model is more general in that we do not impose the γ-condition on the pairwise probabilities When specialized to the γ-model, the

quantities $\{\Delta_j\}_{j=1}^n$ in our analysis takes the form $\Delta_j = \frac{2\gamma}{n}$, and Theorem 2 shows that $\frac{n \log n}{\min_{j \in [n]} \Delta_j^2} = \frac{n^3 \log n}{\gamma^2}$ observations are sufficient to recover the exact total ordering. Thus, for any constant δ, Theorem 2 guarantees exact recovery with a sample complexity that is a multiplicative factor of order $\frac{n^2}{\log n}$ smaller than that established by Wauthier et al. (2013).

The two papers by Rajkumar and Agarwal (2014) and Rajkumar et al. (2015) consider ranking under several models and several metrics. For the subset of their models common with our setting—namely, Bradley-Terry-Luce and the so-called low noise models—they show that the counting algorithm is consistent in terms of recovering the full ranking or the top subset of items. The guarantees are obtained under a low-noise assumption: namely, that the probability of any item i beating j is at least $\frac{1}{2} + \gamma$ whenever item i is ranked higher than item j in an assumed underlying ordering. Their guarantees are based on a sample size of at least $\frac{\log n}{\gamma^2 \mu^2}$, where μ is a parameter lower bounded as $\mu \geq \frac{1}{n^2}$. Once again, our setting allows for the parameter γ to be arbitrarily close to zero, and furthermore as one can see from the discussion above, our bounds are much stronger. Moreover, while Rajkumar et al. focus on upper bounds alone, we also prove matching lower bounds on sample complexity showing that our results are unimprovable beyond constant factors. It should be noted that Rajkumar et al. also provide results for other types of ranking problems that lie outside the problem class treated in the current paper.

Most recently, Chen and Suh (2015) consider a random-design setting and show that if the pairwise observations are assumed to drawn according to the Bradley-Terry-Luce (BTL) parametric model (4a), then their proposed Spectral MLE algorithm recovers the k items correctly with high probability when a certain separation condition on the parameters $\{w_i\}_{i=1}^n$ of the BTL model is satisfied. Their random-design setting is similar to ours except that they first choose a set of pairs of items with each pair chosen with probability p, and then make r comparisons between the two items in every chosen pair. We believe our random-design setting is more natural; the two are identical when $r = 1$. Chen and Suh also show, via matching lower bounds, that the separation condition they derive for the BTL model is tight up to constant factors. In real-world instances of pairwise ranking data, it is often found that parametric models, such as the BTL model and its variants, fail to provide accurate fits (Davidson and Marschak, 1959; McLaughlin and Luce, 1965; Tversky, 1972; Ballinger and Wilcox, 1997). Our results make no such assumptions on the noise, and furthermore, our notion of the ordering of the items in terms of their scores (2) strictly generalizes the notion of the ordering with respect to the BTL parameters. In empirical evaluations presented subsequently, we see that the counting algorithm is significantly more robust to various kinds of noise, and takes several orders of magnitude lesser time to compute.

Finally, in addition to the notion of exact recovery considered so far, in the next two subsections we also derive tight guarantees for the Hamming error metric and more general metrics inspired by the requirements of many relevant applications (Ilyas et al., 2008; Michel et al., 2005; Babcock and Olston, 2003; Metwally et al., 2005; Kimelfeld and Sagiv, 2006; Fagin et al., 2003).

3.2 Approximate recovery under Hamming error

In the previous section, we analyzed performance in terms of exactly recovering the top-k subset. Although exact recovery is suitable for some applications (e.g., a setting with high stakes, in which any single error has a large price), there are other settings in which it may be acceptable to return a subset that is "close" to the correct k-ranked subset. In this section, we analyze this problem of approximate recovery when closeness is measured under the Hamming error. More precisely, for a given threshold $h \in [0, k)$, suppose that our goal is to output a set k-sized set $\widehat{\mathcal{S}}_k$ such that its Hamming distance to the set $\mathcal{S}_k^*$ of the true top k items, as defined in equation (3), is bounded as

$$D_{\mathrm{H}}(\widehat{\mathcal{S}}_k, \mathcal{S}_k^*) \leq 2h. \tag{11}$$

Our goal is to establish conditions under which it is possible (or impossible) to return an estimate $\widehat{\mathcal{S}}_k$ satisfying the bound (11) with high probability.[1]

As before, we use $(1), \ldots, (n)$ to denote the permutation of the n items in decreasing order of their scores. With this notation, the following quantity plays a central role in our analysis:

$$\Delta_{k,\,h}(M) := \tau_{(k-h)}(M) - \tau_{(k+h+1)}(M). \tag{12a}$$

The quantity $\Delta_{k,\,h}$ measures the difference between the scores associated to the items which are h positions on either side of our desired boundary between the k^{th} and $(k+1)^{\text{th}}$ items. Observe that $\Delta_{k,\,h}$ is a generalization of the quantity Δ_k defined previously in equation (8); and the quantity Δ_k corresponds to $\Delta_{k,\,h}$ with $h = 0$. We then define a generalization of the family $\mathcal{F}_k(\alpha; n, p, r)$, namely

$$\mathcal{F}_{k,h}(\alpha; n, p, r) := \left\{ M \in [0,1]^{n \times n} \;\mid\; M + M^T = 11^T, \text{ and } \Delta_{k,\,h} \geq \alpha \sqrt{\frac{\log n}{npr}} \right\}. \tag{12b}$$

As before, we adopt the shorthand $\mathcal{F}_{k,h}(\alpha)$, with the dependence on (n, p, r) being understood implicitly.

Theorem 3 *(a) Consider any $n \geq 2$, $r \geq 1$ and $p \in (0, 1]$. Then if $\alpha \geq 8$, the set $\widetilde{\mathcal{S}}_k$ of top k items* (7) *given by the Borda counting algorithm satisfies*

$$\sup_{M \in \mathcal{F}_{k,h}(\alpha)} \mathbb{P}_M\big[D_{\mathrm{H}}(\widetilde{\mathcal{S}}_k, \mathcal{S}_k^*) > 2h\big] \leq \frac{1}{n^{14}}. \tag{13a}$$

(b) Conversely, in the regime $p \geq \frac{\log n}{2nr}$ and for given constants $\nu_1, \nu_2 \in (0, 1)$, suppose that $2h \leq \frac{1}{1+\nu_2} \min\{n^{1-\nu_1}, k, n-k\}$. Then for any $\alpha \leq \frac{\sqrt{\nu_1 \nu_2}}{14}$, any estimator $\widehat{\mathcal{S}}_k$ has error at least

$$\sup_{M \in \mathcal{F}_{k,h}(\alpha)} \mathbb{P}_M\big[D_{\mathrm{H}}(\widehat{\mathcal{S}}_k, \mathcal{S}_k^*) > 2h\big] \geq \frac{1}{7}, \tag{13b}$$

for all n larger than a constant $c(\nu_1, \nu_2)$.

1. The requirement $h < k$ is sensible because if $h \geq k$, the problem is trivial: any two k-sized sets $\widehat{\mathcal{S}}_k$ and $\mathcal{S}_k^*$ satisfy the bound $D_{\mathrm{H}}(\widehat{\mathcal{S}}_k, \mathcal{S}_k^*) \leq 2k \leq 2h$.

This result is similar to that of Theorem 1, except that the relaxation of the exact recovery condition allows for a less constrained definition of the separation threshold $\Delta_{k,h}$. As with Theorem 1, the lower bound in part (b) applies even if probability matrix M is restricted to lie in a parametric model (such as the BTL model), or the more general SST class. The counting algorithm is thus optimal for estimation under the relaxed Hamming metric as well.

The proof of the upper bound involves a transformation of the Hamming error requirement into one of exact recovery requirement, and then transforming the result of Theorem 1(a) to that required here via certain algebraic arguments. The lower bound is significantly more intricate: we carefully design a packing set using a coding-theoretic result due to Levenshtein (1971), which we then employ in Fano's inequality.

Finally, it is worth making a few comments about the constants appearing in these claims. We can weaken the lower bound on Δ_k required in Theorem 3(a) at the expense of a lower probability of success; for instance, if we instead require that $\alpha \geq 4$, then the probability of error is guaranteed to be at most n^{-2}. Subsequently in the paper, we provide the results of simulations with $n = 500$ items and $\alpha = 4$. On the other hand, in Theorem 3(b), if we impose the stronger upper bound $\alpha = \mathcal{O}(1/\sqrt{h \log n})$, then we can remove the condition $h \leq n^{1-\nu_1}$.

3.3 An abstract form of k-set recovery

In earlier sections, we investigated recovery of the top k items either exactly or under a Hamming error. Exact recovery may be quite strict for certain applications, whereas the property of Hamming error allowing for a few of the top k items to be replaced by *arbitrary* items may be undesirable. Indeed, many applications have requirements that go beyond these metrics; for instance, see the papers Ilyas et al. (2008); Michel et al. (2005); Babcock and Olston (2003); Metwally et al. (2005); Kimelfeld and Sagiv (2006); Fagin et al. (2003) and references therein for some examples. In this section, we generalize the notion of exact or Hamming-error recovery in order to accommodate a fairly general class of requirements.

Both the exact and approximate Hamming recovery settings require the estimator to output a set of k items that are either exactly or approximately equal to the true set of top k items. When is the estimate deemed successful? One way to think about the problem is as follows. The specified requirement of exact or approximate Hamming recovery is associated to a set of k-sized subsets of the n possible ranks. The estimator is deemed successful if the true ranks of the chosen k items equals one of these subsets. In our notion of generalized recovery, we refer to such sets as *allowed sets*. For example, in the case $k = 3$, we might say that the set $\{1, 4, 10\}$ is allowed, meaning that an output consisting of the items that are ranked "first", "fourth" and "tenth" in the ground truth is considered correct.

In more generality, let $\mathfrak{S}$ denote a family of k-sized subsets of $[n]$, which we refer to as *family of allowed sets.* Notice that any allowed set is defined by the *positions* of the items in the true ordering and not the items themselves.[2] Once some true underlying ordering of the n items is fixed, each element of the family $\mathfrak{S}$ then specifies a set of the items themselves. We use these two interpretations depending on the context — the definition in terms of

2. In case of two or more items with identical scores, the choice of any of these items is considered valid.

positions to specify the requirements, and the definition in terms of the items to evaluate an estimator for a given underlying probability matrix M.

We let $\mathcal{S}_k^\dagger$ denote a k-set estimate, meaning a function that given a set of observations as input, returns a k-sized subset of $[n]$ as output.

Definition 4 ($\mathfrak{S}$-respecting estimators) *For any family $\mathfrak{S}$ of allowed sets, a k-set estimate $\mathcal{S}_k^\dagger$* respects *its structure if the set of k positions of the items in $\mathcal{S}_k^\dagger$ belongs to the set family $\mathfrak{S}$.*

Our goal is to determine conditions on the set family $\mathfrak{S}$ under which there exist estimators $\mathcal{S}_k^\dagger$ that respect its structure. In order to illustrate this definition, let us return to the examples treated thus far.

Example 1 (Exact and approximate Hamming recovery) *The requirement of exact recovery of the top k items has $\mathfrak{S}$ consisting of exactly one set, the set of the top k positions $\mathfrak{S} = \{[k]\}$. In the case of recovery with a Hamming error at most $2h$, the set $\mathfrak{S}$ of all allowed sets consists all k-sized subsets of $[n]$ that contain at least $(k-h)$ positions in the top k positions. For instance, in the case $h = 1$, $k = 2$ and $n = 4$, we have*

$$\mathfrak{S} = \Big\{\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\}\Big\}.$$

Apart from these two requirements, there are several other requirements for top-k recovery popular in the literature (Carmel et al., 2001; Fagin et al., 2003; Babcock and Olston, 2003; Michel et al., 2005; Metwally et al., 2005; Kimelfeld and Sagiv, 2006; Ilyas et al., 2008). Let us illustrate them with another example:

Example 2 *Let $\pi^* : [n] \to [n]$ denote the true underlying ordering of the n items. The following are four popular requirements on the set $\mathcal{S}_k^\dagger$ for top-k identification, with respect to the true permutation π^*, for a pre-specified parameter $\epsilon \geq 0$.*

(i) All items in the set $\mathcal{S}_k^\dagger$ must be contained contained within the top $(1+\epsilon)k$ entries:

$$\max_{i \in \mathcal{S}_k^\dagger} \pi^*(i) \leq (1+\epsilon)k. \tag{14a}$$

(ii) The rank of any item in the set $\mathcal{S}_k^\dagger$ must lie within a multiplicative factor $(1+\epsilon)$ of the rank of any item not in the set $\mathcal{S}_k^\dagger$:

$$\max_{i \in \mathcal{S}_k^\dagger} \pi^*(i) \leq (1+\epsilon) \min_{j \in [n] \setminus \mathcal{S}_k^\dagger} \pi(j). \tag{14b}$$

(iii) The rank of any item in the set $\mathcal{S}_k^\dagger$ must lie within an additive factor ϵ of the rank of any item not in the set $\mathcal{S}_k^\dagger$:

$$\max_{i \in \mathcal{S}_k^\dagger} \pi^*(i) \leq \min_{j \in [n] \setminus \mathcal{S}_k^\dagger} \pi^*(j) + \epsilon. \tag{14c}$$

(iv) The sum of the ranks of the items in the set $\mathcal{S}_k^\dagger$ must be contained within a factor $(1+\epsilon)$ of the sums of ranks of the top k entries:

$$\sum_{i \in \mathcal{S}_k^\dagger} \pi^*(i) \leq (1+\epsilon)\frac{1}{2}k(k+1). \tag{14d}$$

Note that each of these requirements reduces to the exact recovery requirement when $\epsilon = 0$. Moreover, each of these requirements can be rephrased in terms of families of allowed sets. For instance, if we focus on requirement (i), then any k-sized subset of the top $(1+\epsilon)k$ positions is an allowed set.

In this paper, we derive conditions that govern k-set recovery for allowed set systems that satisfy a natural "monotonicity" condition. Informally, the monotonicity condition requires that the set of k items resulting from replacing an item in an allowed set with a higher ranked item must also be an allowed set. More precisely, for any set $\{t_1, \ldots, t_k\} \subseteq [n]$, let $\Lambda(\{t_1, \ldots, t_k\}) \subseteq 2^{[n]}$ be the set defined by all of its monotone transformations—that is

$$\Lambda(\{t_1, \ldots, t_k\}) := \Big\{ \{t'_1, \ldots, t'_k\} \subseteq [n] \mid t'_j \leq t_j \text{ for every } j \in [k] \Big\}.$$

Using this notation, we have the following:

Definition 5 (Monotonic set systems) *The set $\mathfrak{S}$ of allowed sets is a* monotonic set system *if*

$$\Lambda(T) \subseteq \mathfrak{S} \quad \textit{for every } T \in \mathfrak{S}. \tag{15}$$

One can verify that condition (15) is satisfied by the settings of exact and Hamming-error recovery, as discussed in Example 1. The condition is also satisfied by all four requirements discussed in Example 2.

Our next result establishes conditions under which one can (or cannot) produce an estimator that respects an allowed set requirement. In order to state the result, we recall the score $\tau_i(M) := \frac{1}{n}\sum_{j=1}^n M_{ij}$, as previously defined in equation (2) for each $i \in [n]$. For notational convenience, we also define $\tau_i(M) := -\infty$ for every $i > n$ and every M. Consider any monotonic family of allowed sets $\mathfrak{S}$, and for some integer $\beta \geq 1$, let $T^1, \ldots, T^\beta \in \mathfrak{S}$ such that $\mathfrak{S} = \bigcup_{b \in [\beta]} \Lambda(T^b)$. For every $b \in [\beta]$, let $t_1^b < \cdots < t_k^b$ denote the entries of T^b. We then define the critical threshold based on the scores:

$$\Delta_{\mathfrak{S}}(M) := \max_{b \in [\beta]} \min_{j \in [k]} (\tau_{(j)}(M) - \tau_{(k+t_j^b - j+1)}(M)). \tag{16}$$

The term $\Delta_{\mathfrak{S}}$ is a further generalization of the quantities Δ_k and $\Delta_{k,h}$ defined in earlier sections. For example, for the exact recovery setting we have $\beta = 1$ and $T^1 = \{1, \ldots, k\}$, and after some algebraic simplifications of (16), we obtain that the critical threshold $\Delta_{\mathfrak{S}}$ reduces exactly to the threshold Δ_k defined earlier in (8). As a second example, the setting allowing a Hamming error at most $2h$ can be described with the choice $\beta = 1$ and $T^1 = \{h+1, \ldots, k,\ n-h+1, \ldots, n\}$. Some algebraic simplifications of (16) reduce $\Delta_{\mathfrak{S}}$ to the threshold $\Delta_{k,h}$ defined in (12a). As an example with $\beta > 1$, consider requirement (ii)

in Example 2. For simplicity of exposition, assume that $\epsilon > \frac{1}{k-1}$ and $n \geq 2k(1+\epsilon)$. For this requirement, we have $\beta = (k - \lceil \frac{k}{1+\epsilon} \rceil)$, and for every $b \in \{\lceil \frac{k}{1+\epsilon} \rceil + 1, \ldots, k\}$, we have $T^{b-\lceil \frac{k}{1+\epsilon} \rceil} = \{1, \ldots, b-1, \lceil (1+\epsilon)b - (k-b) \rceil, \ldots, \lceil (1+\epsilon)b \rceil\}$. Then some algebraic simplifications of equation (16) yield that the critical threshold for this requirement is given by $\Delta_{\mathfrak{S}}(M) = \max_{b \in \{\lceil \frac{k}{1+\epsilon} \rceil + 1, \ldots, k\}} \min\{\tau_{(b-1)}(M) - \tau_{(k+1)}(M), \tau_{(k)}(M) - \tau_{(\lceil (1+\epsilon)b \rceil + 1)}(M)\}$.

With the definition of the critical threshold $\Delta_{\mathfrak{S}}$ in place, we now define a generalization $\mathcal{F}_{\mathfrak{S}}(\cdot)$ of the families $\mathcal{F}_k(\cdot)$ and $\mathcal{F}_{k,h}(\cdot)$ as

$$\mathcal{F}_{\mathfrak{S}}(\alpha; n, p, r) := \left\{ M \in [0,1]^{n \times n} \mid M + M^T = 11^T \text{ and } \Delta_{\mathfrak{S}}(M) \geq \alpha \sqrt{\frac{\log n}{npr}} \right\}. \tag{17}$$

As before, we use the shorthand $\mathcal{F}_{\mathfrak{S}}(\alpha)$, with the dependence on (n, p, r) being understood implicitly.

Theorem 6 *Consider any allowed set requirement specified by a monotonic set class $\mathfrak{S}$.*

(a) For any $\alpha \geq 8$, the set $\widetilde{\mathcal{S}}_k$ of top k items (7) given by the Borda counting algorithm satisfies

$$\sup_{M \in \mathcal{F}_{\mathfrak{S}}(\alpha)} \mathbb{P}_M\big[\widetilde{\mathcal{S}}_k \notin \mathfrak{S}\big] \leq \frac{1}{n^{13}}.$$

(b) Conversely, in the regime $p \geq \frac{\log n}{2nr}$, and for given constants $\mu_1 \in (0,1), \mu_2 \in (\frac{3}{4}, 1]$, suppose that $\max_{b \in [\beta]} t^b_{\lceil \mu_2 k \rceil} \leq \frac{n}{2}$ and $8(1-\mu_2)k \leq n^{1-\mu_1}$. Then for any α smaller than a constant $c_u(\mu_1, \mu_2) > 0$, any estimator $\widehat{\mathcal{S}}_k$ has error at least

$$\sup_{M \in \mathcal{F}_{\mathfrak{S}}(\alpha)} \mathbb{P}_M\big[\widetilde{\mathcal{S}}_k \notin \mathfrak{S}\big] \geq \frac{1}{15}, \tag{18}$$

for all n larger than a constant $c_0(\mu_1, \mu_2)$.

A few remarks on the bounds are in order. For the lower bound, first, it continues to hold even if the probability matrix M is restricted to follow a parametric model such as BTL or restricted to lie in the SST class. Second, in terms of the threshold for α, the lower bound holds with $c_u(\mu_1, \mu_2) = \frac{1}{15}\sqrt{\mu_1 \min\big\{\frac{1}{4(1-\mu_2)-1}, \frac{1}{2}\big\}}$. Third, it is worth noting that one must necessarily impose some conditions for the lower bound, along the lines of those required in Theorem 6(b) for the allowed sets to be "interesting" enough. As a concrete illustration of the necessity of this condition, consider the requirement defined by the parameters $b = 1$, $k = 1$ and $\mathfrak{S} = \Lambda(\{n - \sqrt{n}\})$. For $\mu_1 = \mu_2 = \frac{9}{10}$, this requirement satisfies the condition $8(1-\mu_2)k \leq n^{1-\mu_1}$ but violates the condition $t_{\lceil \mu_2 k \rceil} \leq \frac{n}{2}$. Now, a selection of $k = 1$ item made uniformly at random (independent of the data) satisfies this allowed set requirement with probability $1 - \frac{1}{\sqrt{n}}$. Given the success of such a random selection algorithm in this parameter regime, we see that the lower bounds therefore cannot be universal, but must require some conditions on the allowed sets.

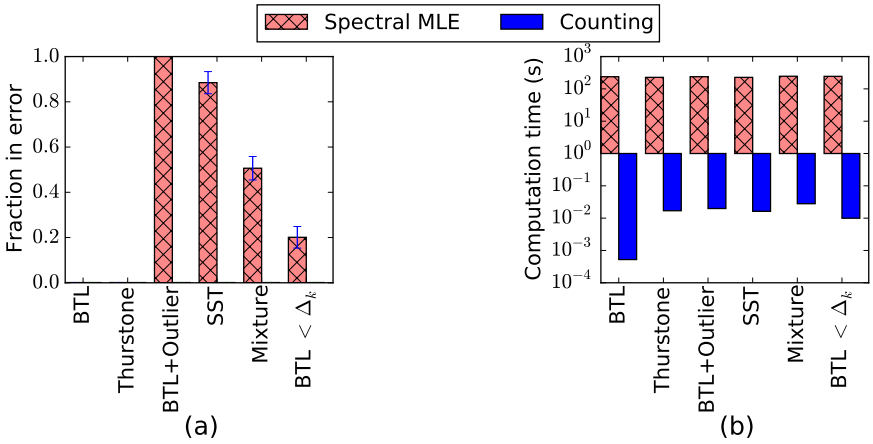

Figure 1. Simulation results comparing Spectral MLE and the counting algorithm in terms of error rates for exact recovery of the top k items, and computation time. (a) Histogram of fraction of instances where the algorithm failed to recover the k items correctly, with each bar being the average value across 50 trials. The counting algorithm has 0% error across all problems, while the spectral MLE is accurate for parametric models (BTL, Thurstone), but not very accurate for other models. (b) Histogram plots of the maximum computation time taken by the counting algorithm and the minimum computation time taken by Spectral MLE across all trials. Even though this maximum-to-minimum comparison is unfair to the counting algorithm, it involves five or more orders of magnitude less computation.

4. Simulations and experiments

In this section, we empirically evaluate the performance of the counting algorithm and compare it with the Spectral MLE algorithm via simulations on synthetic data, as well as experiments using datasets from the Amazon Mechanical Turk crowdsourcing platform.

4.1 Simulated data

We begin with simulations using synthetically generated data with $n = 500$ items and observation probability $p = 1$, and with pairwise comparison models ranging over six possible types. Panel (a) in Figure 1 provides a histogram plot of the associated error rates (with a bar for each one of these six models) in recovering the $k = n/4 = 125$ items for the counting algorithm versus the Spectral MLE algorithm. Each bar corresponds to the average over 50 trials. Panel (b) compares the CPU times of the two algorithms. The value of α (and in turn, the value of r) in the first five models is as derived in Section 3.1. In more detail, the six model types are given by:

(I) *Bradley-Terry-Luce (BTL) model:* Recall that the theoretical guarantees for the Spectral MLE algorithm (Chen and Suh, 2015) are applicable to data that is generated from the BTL model (4a), and as guaranteed, the Spectral MLE algorithm gives a 100% accuracy under this model. The counting algorithm also obtains a 100% accuracy, but importantly, the counting algorithm requires a computational time that is five orders of magnitude lower than that of Spectral MLE.

(II) *Thurstone model:* The Thurstone model (Thurstone, 1927) is another parametric model, with the function F in equation (4b) set as the cumulative distribution function of the standard Gaussian distribution. Both Spectral MLE and the counting algorithm gave 100% accuracy under this model.

(III) *BTL model with one (non-transitive) outlier:* This model is identical to BTL, with one modification. Comparisons among $(n-1)$ of the items follow the BTL model as before, but the remaining item always beats the first $\frac{n}{4}$ items and always loses to each of the other items. We see that the counting algorithm continues to achieve an accuracy of 100% as guaranteed by Theorem 1. The departure from the BTL model however prevents the Spectral MLE algorithm from identifying the top k items.

(IV) *Strong stochastic transitivity (SST) model:* We simulate the "independent diagonals" construction of Shah et al. (2017a) in the SST class. Spectral MLE is often unsuccessful in recovering the top k items, while the counting algorithm always succeeds.

(V) *Mixture of BTL models:* Consider two sets of people with opposing preferences. The first set of people have a certain ordering of the items in their mind and their preferences follow a BTL model under this ordering. The second set of people have the opposite ordering, and their preferences also follow a BTL model under this opposite ordering. The overall preference probabilities is a mixture between these two sets of people. In the simulations, we observe that the counting algorithm is always successful while the Spectral MLE method often fails.

(VI) *BTL with violation of separation condition:* We simulate the BTL model, but with a choice of parameter r small enough that the value of α is about one-tenth of its recommended value in Section 3.1. We observe that the counting algorithm continues to incur a lower error than the Spectral MLE algorithm, thereby demonstrating its robustness.

To summarize, the performance of the two algorithms can be contrasted in the following way. When our stated lower bounds on α are satisfied, then consistent with our theoretical claims, the Borda counting algorithm succeeds irrespective of the form of the pairwise probability distributions. The Spectral MLE algorithm performs well when the pairwise comparison probabilities are faithful to parametric models, but is often unsuccessful otherwise. Even when the condition on α is violated, the performance of the counting algorithm remains superior to that of the Spectral MLE.[3] In terms of computational complexity, for every instance we simulated, the counting algorithm took several orders of magnitude less time as compared to Spectral MLE.

Simulations with adversarial, imbalanced choice of pairs: The theoretical results in the earlier sections addressed a random design setting where the pairs to be compared are chosen at random in a homogeneous manner. While such a random design setting is widespread in various applications such as crowdsourcing and others, and is also the focus of a bulk of past literature on related topics, it is also of interest to understand situations where the

3. Note that part (b) of Theorem 1 is a minimax converse meaning that it appeals to the worst case scenario.

comparisons may be imbalanced. With this goal, we new present simulations that contrast the behavior of the counting algorithm in a random-design setting with an adversarial-design setting.

In this set of simulations, we consider the problem of recovering the top item (that is, $k = 1$). Moreover, we adopt a parametric model in which every item $i \in [n]$ is assumed to be governed by a parameter $w_i^* \in [0, 1]$, and the probability of any item i beating item j is set as $M_{ij} = \frac{1+w_i^*-w_j^*}{2}$. As before, the number of times any pair of items (i, j) is compared is drawn as a binomial distribution with parameters r and p. In the standard random-design setting studied throughout the remainder of this paper, the choice of the number of comparisons is unrelated to the choice of the parameters w^*. However in the adversarial setting, we make w^* adversarially misaligned to the number of comparisons between various pairs. In particular, the parameters associated to the items are chosen as follows for the random and the adversarial settings:

1. Random: $w_1^* = 1$ and $w_2^* = 0.9$ are the top two items. For every $i \in \{3, \ldots, n\}$, we draw w_i^* uniformly at random from the set $\{0.1, 0.7\}$.

2. Adversarial: $w_1^* = 1$ and $w_2^* = 0.9$ are the top two items. For every $i \in \{3, \ldots, n\}$, we set $w_i^* = 0.7$ if item i is compared more often to item 1 than to item 2, set $w_i^* = 0.1$ if item i is compared more often to item 2 than to item 1, and draw w_i^* uniformly at random from the set $\{0.1, 0.7\}$ otherwise.

The results of applying the counting algorithm are shown in Figure 2. From these simulations we observe that the simple counting algorithm is indeed sensitive to the imbalanced choice of pairs to be compared (that is, when the top item is compared more often to higher ranked items and the second item is compared more often to lower ranked items). Designing algorithms for ranking from pairwise comparisons that can optimally handle such imbalanced, adversarial-design settings is left as a problem for future work.

4.2 Experiments on data from Amazon Mechanical Turk

In this section, we describe experiments on real world datasets collected from the Amazon Mechanical Turk (`mturk.com`) commercial crowdsourcing platform.

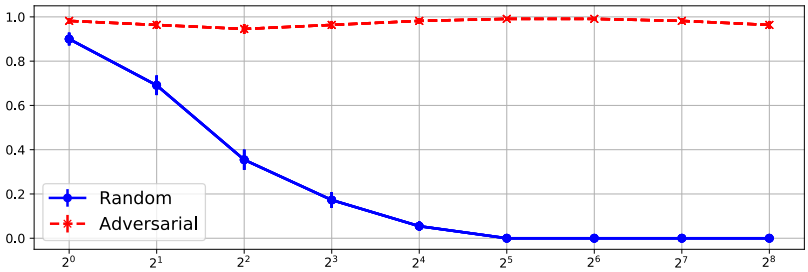

Figure 2. Performance of the counting algorithm for top $k = 1$ recovery when the pairs to be compared are chosen randomly and when they are chosen adversarially to create an imbalanced setting.

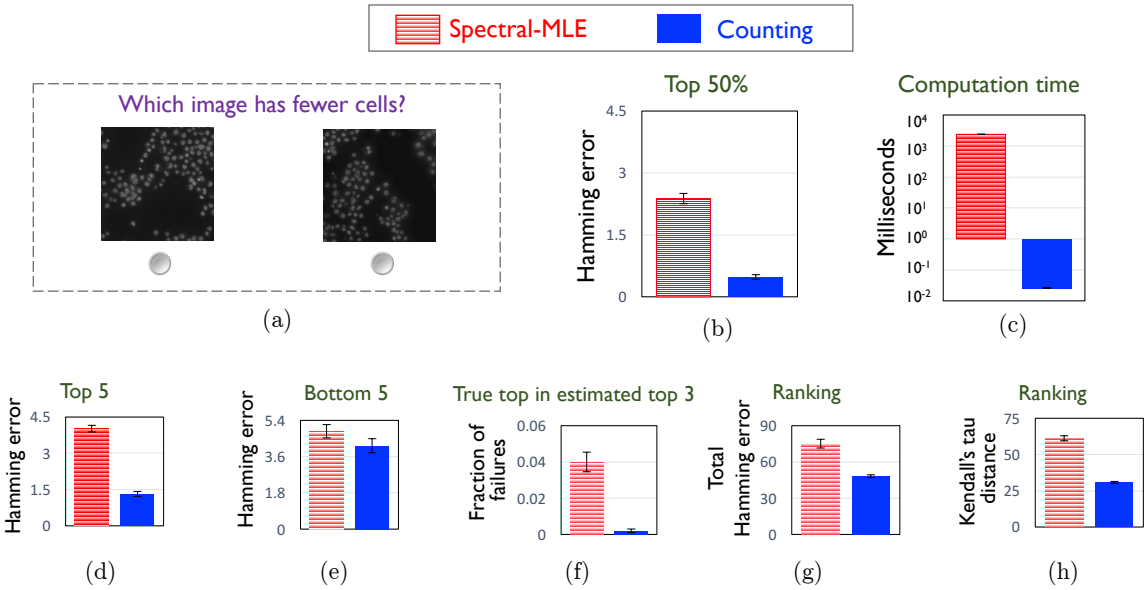

Figure 3. An illustration of the cell counting experiment (panel a) and results comparing Spectral MLE and the counting algorithm in terms of accuracy and computation time (panels b–h).

4.2.1 Experiment on counting cells

We begin with an experiment on counting (biological) cells in images.

Data. We employed a dataset of 23 images, each comprising several (biological) cells. The images of the cells and the ground truth counts of the numbers of cells in each image were obtained from the dataset collected by Carpenter et al. (2006).

On the Amazon Mechanical Turk crowdsourcing platform, we recruited a total of 64 workers and showed multiple pairs of such images to each worker – see Figure 3(a) for an illustrative example. For each pair of images, the worker was asked to select the image that the worker considered to have fewer cells. For each worker, the pairs were chosen by permuting the 23 images uniformly at random and asking for a comparison between the first and second images, between the third and fourth images and so on, for a total of 11 pairs of images per worker. In the raw data, 9.8% of the responses provided by the workers were erroneous. In the raw data we also observed that unsurprisingly, the number of errors increase as the actual cell counts in the pair of images come closer. The interface seen by the workers as well as the raw data obtained from Amazon Mechanical Turk is available on the website of the first author.

The goal of any algorithm is to take this set of noisy pairwise comparisons from the workers and estimate the images with the fewest cells. Such estimates are useful to detect various conditions, for instance, a low count of red blood cells in images of human cells indicates anemia. To this end, we executed the Spectral MLE algorithm (Chen and Suh, 2015) and the Borda counting algorithm on the set of pairwise comparisons obtained from the workers.

Results. We compared the performance of the two algorithms on a variety of metrics. In what follows, we subsample the responses with $p = 0.5$, that is, for each response for each question, we keep the response independently with probability 0.5 and discard it otherwise. We execute the two algorithms on this subsampled data. We repeat this process for 100 trials and plot the mean of the metric under consideration along with error bars representing the standard error of the mean.

We first consider recovering the set of top $k = \frac{n}{2}$ items. The natural metric of error here is the Hamming error, that is, the number of images that are misclassified. For this objective, while Spectral MLE does quite well, counting incurs a significantly lower Hamming error – see Figure 3(b). As one may expect, the count estimator also requires a much lower computation time – see Figure 3(c) for a comparison. In Figures 3(d) to (h), we see that counting also performs quite well for other exact or approximate requirements of top k or ranking recovery.

4.2.2 Data from earlier experiments on Amazon Mechanical Turk

We now describe three additional experiments using data collected from Amazon Mechanical Turk in our past work Shah et al. (2016a).

Data. In order to evaluate the accuracy of the algorithms under consideration, we require datasets consisting of pairwise comparisons in which the questions can be associated with an objective and verifiable ground truth. To this end, we used the "cardinal versus ordinal" dataset from our past work Shah et al. (2016a); three of the experiments performed in that paper are suitable for the evaluations here—namely, ones in which each question has a ground truth, and the pairs of items are chosen uniformly at random. The three experiments tested the workers' general knowledge, audio, and visual understanding, and the respective tasks involved: (i) identifying the pair of cities with a greater geographical distance, (ii) identifying the higher frequency key of a piano, and (iii) identifying spelling mistakes in a paragraph of text. The number of items n in the three experiments were 16, 10 and 8 respectively. The total number of pairwise comparisons were 408, 265 and 184 respectively. The fraction of pairwise comparisons whose outcomes were incorrect (as compared to the ground truth) in the raw data are 17%, 20% and 40% respectively.

Results. We compared the performance of the counting algorithm with that of the Spectral MLE algorithm. For each value of a "subsampling probability" $q \in \{0.1, 0.2, \ldots, 1.0\}$, we subsampled a fraction q of the data and executed both algorithms on this subsampled data. We evaluated the performance of the algorithms on their ability to recover the top $k = \lceil \frac{n}{4} \rceil$ items under the Hamming error metric.

Figure 4 shows the results of the experiments. Each point in the plots is an average across 100 trials. Observe that the counting algorithm consistently outperforms Spectral MLE. (We think that the erratic fluctuations in the spelling mistakes data are a consequence of a high noise and a relatively small problem size.) Moreover, the Spectral MLE algorithm required about 5 orders of magnitude more computation time (not shown in the figure) as compared to counting. Thus the counting algorithm performs well on simulated as well as real data. It outperforms Spectral MLE not only when the number of items is large (as in the simulations) but also when the problem sizes are small as seen in these experiments.

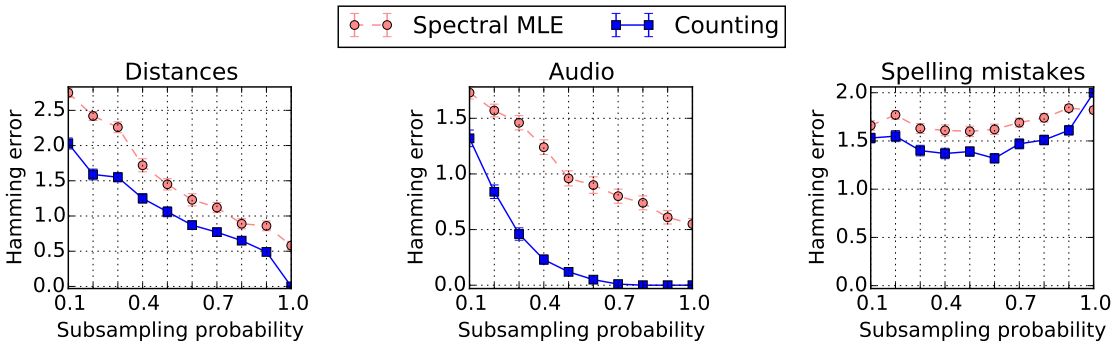

Figure 4. Evaluation of Spectral MLE and the counting algorithm on three datasets (from left to right: Distances, Audio, Spelling mistakes) from Amazon Mechanical Turk in terms of the error rates for top k-subset recovery. The three panels plot the Hamming error when recovering the top k items in the three datasets when a q^{th} fraction of the total data is used, for various values of subsampling probability $q \in (0, 1]$.

5. Proofs

We now turn to the proofs of our main results. We continue to use the notation $[i]$ to denote the set $\{1, \ldots, i\}$ for any integer $i \geq 1$. We ignore floor and ceiling conditions unless critical to the proof. All logarithms are taken to the base e.

Our lower bounds are based on a standard form of Fano's inequality (Cover and Thomas, 2012; Tsybakov, 2008) for lower bounding the probability of error in an L-ary hypothesis testing problem. We state a version here for future reference. For some integer $L \geq 2$, fix some collection of distributions $\{\mathbb{P}^1, \ldots, \mathbb{P}^L\}$. Suppose that we observe a random variable Y that is obtained by first sampling an index A uniformly at random from $[L] = \{1, \ldots, L\}$, and then drawing $Y \sim \mathbb{P}^A$. (As a result, the variable Y is marginally distributed according to the mixture distribution $\bar{\mathbb{P}} = \frac{1}{L} \sum_{a=1}^{L} \mathbb{P}^a$.) Given the observation Y, our goal is to "decode" the value of A, corresponding to the index of the underlying mixture component. Using $\mathcal{Y}$ to denote the sample space associated with the observation Y, Fano's inequality asserts that any test function $\phi : \mathcal{Y} \to [L]$ for this problem has error probability lower bounded as

$$\mathbb{P}[\phi(Y) \neq A] \geq 1 - \frac{I(Y; A) + \log 2}{\log L},$$

where $I(Y; A)$ denotes the mutual information between Y and A. A standard convexity argument for the mutual information yields the weaker bound

$$\mathbb{P}[\phi(Y) \neq A] \geq 1 - \frac{\max\limits_{a,b \in [L]} D_{\mathrm{KL}}(\mathbb{P}^a \| \mathbb{P}^b) + \log 2}{\log L}, \tag{19}$$

We make use of this weakened form of Fano's inequality in several proofs.

5.1 Proof of Theorem 1

We begin with the proof of Theorem 1, dividing our argument into two parts.

5.1.1 Proof of part (a)

For any pair of items (i,j), let us encode the outcomes of the r trials by an i.i.d. sequence $V_{ij}^{(\ell)} = [X_{ij}^{(\ell)} \quad X_{ji}^{(\ell)}]^T$ of random vectors, indexed by $\ell \in [r]$. Each random vector follows the distribution

$$\mathbb{P}\big[x_{ij}^{(\ell)},\, x_{ji}^{(\ell)}\big] = \begin{cases} 1-p & \text{if } (x_{ij}^{(\ell)},\, x_{ji}^{(\ell)}) = (0,0) \\ pM_{ij} & \text{if } (x_{ij}^{(\ell)},\, x_{ji}^{(\ell)}) = (1,0) \\ p(1-M_{ij}) & \text{if } (x_{ij}^{(\ell)},\, x_{ji}^{(\ell)}) = (0,1) \\ 0 & \text{otherwise.} \end{cases}$$

With this encoding, the variable $W_a := \sum_{\ell\in[r]} \sum_{z\in[n]\setminus\{a\}} X_{aj}^{(r)}$ encodes the number of wins for item a.

Consider any item $a \in \mathcal{S}_k^*$ which ranks among the top k in the true underlying ordering, and any item $b \in [n]\setminus\mathcal{S}_k^*$ which ranks outside the top k. We claim that with high probability, item a will win more pairwise comparisons than item b. More precisely, let $\mathcal{E}_{ba}$ denote the event that item b wins at least as many pairwise comparisons than a. We claim that

$$\mathbb{P}(\mathcal{E}_{ba}) \overset{(i)}{\leq} \exp\left(-\frac{\frac{1}{2}(rpn\Delta_k)^2}{rpn(2-\Delta_k)+\frac{2}{3}rpn\Delta_k}\right) \overset{(ii)}{\leq} \frac{1}{n^{16}}. \tag{20}$$

Given this bound, the probability that the counting algorithm will rank item b above a is no more than n^{-16}. Applying the union bound over all pairs of items $a \in \mathcal{S}_k^*$ and $b \in [n]\setminus\mathcal{S}_k^*$ yields $\mathbb{P}\big[\widetilde{\mathcal{S}}_k \neq \mathcal{S}_k^*\big] \leq n^{-14}$ as claimed.

We note that inequality (ii) in equation (20) follows from inequality (i) combined with the condition on Δ_k that arises by setting $\alpha \geq 8$ as assumed in the hypothesis of the theorem. Thus, it remains to prove inequality (i) in equation (20). By definition of $\mathcal{E}_{ba}$, we have

$$\mathbb{P}(\mathcal{E}_{ba}) = \mathbb{P}\Big(\underbrace{\sum_{\ell\in[r]}\sum_{z\in[n]\setminus\{b\}} X_{bz}^{(\ell)}}_{W_b} - \underbrace{\sum_{\ell\in[r]}\sum_{z\in[n]\setminus\{a\}} X_{az}^{(\ell)}}_{W_a} \geq 0\Big). \tag{21}$$

It is convenient to recenter the random variables. For every $\ell \in [r]$ and $z \in [n]\setminus\{a,b\}$, define the zero-mean random variables

$$\overline{X}_{az}^{(\ell)} = X_{az}^{(\ell)} - \mathbb{E}[X_{az}^{(\ell)}] = X_{az}^{(\ell)} - pM_{az} \;\text{ and }\; \overline{X}_{bz}^{(\ell)} = X_{bz}^{(\ell)} - \mathbb{E}[X_{bz}^{(\ell)}] = X_{bz}^{(\ell)} - pM_{bz}.$$

Also, let

$$\overline{X}_{ab}^{(\ell)} = (X_{ab}^{(\ell)} - X_{ba}^{(\ell)}) - \mathbb{E}[X_{ab}^{(\ell)} - X_{ba}^{(\ell)}] \;=\; (X_{ab}^{(\ell)} - X_{ba}^{(\ell)}) - (pM_{ab} - pM_{ba}).$$

We then have

$$\mathbb{P}(\mathcal{E}_{ba}) = \mathbb{P}\Bigg(\sum_{\ell\in[r]}\Big(\sum_{z\in[n]\setminus\{a,b\}} \overline{X}_{bz}^{(\ell)} - \sum_{z\in[n]\setminus\{a,b\}} \overline{X}_{az}^{(\ell)} - \overline{X}_{ab}^{(\ell)}\Big) \geq rp\sum_{z\in[n]}\Big(M_{az} - M_{bz}\Big)\Bigg).$$

Since $a \in \mathcal{S}_k^*$ and $b \in [n] \backslash \mathcal{S}_k^*$, from the definition of Δ_k, we have $n\Delta_k \leq \sum_{z \in [n]} (M_{az} - M_{bz})$, and consequently

$$\mathbb{P}\left(\mathcal{E}_{ba}\right) \leq \mathbb{P}\left(\sum_{\ell \in [r]} \Big(\sum_{z \in [n] \backslash \{a,b\}} \overline{X}_{bz}^{(\ell)} - \sum_{z \in [n] \backslash \{a,b\}} \overline{X}_{az}^{(\ell)} - \overline{X}_{ab}^{(\ell)} \Big) \geq rpn\Delta_k \right). \tag{22}$$

By construction, all the random variables in the above inequality are zero-mean, mutually independent, and bounded in absolute value by 2. These properties alone would allow us to obtain a tail bound by Hoeffding's inequality; however, in order to obtain the stated result (20), we need the more refined result afforded by Bernstein's inequality (e.g., Boucheron et al., 2013). In order to derive a bound of Bernstein type, the only remaining step is to bound the second moments of the random variables at hand. Some straightforward calculations yield

$$\mathbb{E}[(-\overline{X}_{az}^{(\ell)})^2] \leq pM_{az}, \qquad \mathbb{E}[(\overline{X}_{bz}^{(\ell)})^2] \leq pM_{bz}, \quad \text{and} \quad \mathbb{E}[(\overline{X}_{ab}^{(\ell)})^2] \leq pM_{ab} + pM_{ba}.$$

It follows that

$$\begin{aligned}
\sum_{z \in [n] \backslash \{a,b\}} \mathbb{E}[(-\overline{X}_{az}^{(\ell)})^2] + \sum_{z \in [n] \backslash \{a,b\}} \mathbb{E}[(\overline{X}_{bz}^{(\ell)})^2] + \mathbb{E}[(\overline{X}_{ab}^{(\ell)})^2] & \\
&\leq p \left(\sum_{z \in [n] \backslash \{a,b\}} (M_{az} + M_{bz}) + M_{ab} + M_{ba} \right) \\
&\overset{(iii)}{\leq} p \left(2 \sum_{z \in [n]} M_{az} - n\Delta_k \right) \\
&\overset{(iv)}{<} pn(2 - \Delta_k),
\end{aligned}$$

where the inequality (iii) follows from the definition of Δ_k, and step (iv) follows because $M_{az} \leq 1$ for every z and $M_{aa} = \frac{1}{2}$. Applying the Bernstein inequality now yields the stated bound (20)(i).

5.1.2 Proof of part (b)

We prove the claim by constructing a packing set that satisfies our general requirements as well as also lies within the SST model and all parametric models, and subsequently using the packing set in an application of Fano's inequality. We then carefully bound the Kullback-Leibler divergence between the probability distributions on the outcomes induced by any pair of elements in the packing set in order to obtain a tractable bound.

In more detail, the symmetry of the problem allows us to assume, without loss of generality, that $k \leq \frac{n}{2}$. We first construct an ensemble of $n - k + 1$ different problems, and considering the problem of distinguishing between them. For each $a \in \{k, \ldots, n\}$, let us define the k-sized subset $\mathcal{S}^*[a] := \{1, \ldots, k-1\} \cup \{a\}$, and the associated matrix of pairwise

probabilities

$$M_{ij}^a := \begin{cases} \frac{1}{2} & \text{if } i,j \in \mathcal{S}^*[a], \text{ or } i,j \notin \mathcal{S}^*[a] \\ \frac{1}{2}+\delta & \text{if } i \in \mathcal{S}^*[a] \text{ and } j \notin \mathcal{S}^*[a] \\ \frac{1}{2}-\delta & \text{if } i \notin \mathcal{S}^*[a] \text{ and } j \in \mathcal{S}^*[a], \end{cases}$$

where $\delta \in (0, \frac{1}{2})$ is a parameter to be chosen. We use $\mathbb{P}^a$ to denote probabilities taken under pairwise comparisons drawn according to the model M^a.

One can verify that the construction above falls in the intersection of parametric models and the SST model. In the parametric case, this construction amounts to having the parameters associated to every item in $\mathcal{S}_k^*$ to have the same value, and those associated to every item in $[n] \backslash \mathcal{S}_k^*$ to have the same value. Also observe that for every such distribution $\mathbb{P}^a$, the associated k-separation threshold is $\Delta_k = \delta$.

Any given set of observations can be described by the collection of random variables $Y = \{Y_{ij}^{(\ell)}, j > i \in [n], \ell \in [r]\}$. When the true underlying model is $\mathbb{P}^a$, the random variable $Y_{ij}^{(\ell)}$ follows the distribution

$$Y_{ij}^{(\ell)} = \begin{cases} 0 & \text{with probability } 1-p \\ i & \text{with probability } pM_{ij}^a \\ j & \text{with probability } p(1-M_{ij}^a). \end{cases}$$

The random variables $\{Y_{ij}^{(\ell)}\}_{i,j \in [n], i<j, \ell \in [r]}$ are mutually independent, and the distribution $\mathbb{P}^a$ is a product distribution across pairs $\{i > j\}$ and repetitions $\ell \in [r]$.

Let $A \in \{k, \ldots, n\}$ follow a uniform distribution over the index set, and suppose that given $A = a$, our observations Y has components drawn according to the model $\mathbb{P}^a$. Consequently, the marginal distribution of Y is the mixture distribution $\frac{1}{n-k+1} \sum_{a=k}^n \mathbb{P}^a$ over all $(n-k+1)$ models. Based on observing Y, our goal is to recover the correct index $A = a$ of the underlying model, which is equivalent to recovering the planted subset $\mathcal{S}^*[a]$. We use the Fano bound (19) to lower bound the error bound associated with any test for this problem. In order to apply Fano's inequality, the following result provides control over the Kullback-Leibler divergence between any pair of probabilities involved.

Lemma 7 *For any distinct pair $a, b \in \{k, \ldots, n\}$, we have*

$$D_{\mathrm{KL}}(\mathbb{P}^a \| \mathbb{P}^b) \leq \frac{2npr}{\frac{1}{4\delta^2}-1}. \tag{23}$$

See the end of this section for the proof of this claim.

Given this bound on the Kullback-Leibler divergence, Fano's inequality (19) implies that any estimator ϕ of A has error probability lower bounded as

$$\mathbb{P}[\phi(Y) \neq A] \geq 1 - \frac{\frac{2npr}{\frac{1}{4\delta^2}-1} + \log 2}{\log(n-k+1)} \;\geq\; \frac{1}{7}.$$

Here the final inequality holds whenever $\delta \leq \frac{1}{7}\sqrt{\frac{\log n}{npr}}$, $p \geq \frac{\log n}{2nr}$, $n \geq 7$ and $k \leq \frac{n}{2}$. The condition $p \geq \frac{\log n}{2nr}$ also ensures that $\delta < \frac{1}{2}$ thereby ensuring that our construction is valid. It only remains to prove Lemma 7.

5.1.3 Proof of Lemma 7

Since the distributions $\mathbb{P}^a$ and $\mathbb{P}^b$ are formed by components that are independent across edges $i > j$ and repetitions $\ell \in [r]$, we have

$$D_{\mathrm{KL}}(\mathbb{P}^a \| \mathbb{P}^b) = \sum_{\ell \in [r]} \sum_{1 \leq i < j \leq n} D_{\mathrm{KL}}(\mathbb{P}^a(X_{ij}^{(\ell)}) \| \mathbb{P}^b(X_{ij}^{(\ell)})) \; = \; r \sum_{1 \leq i < j \leq n} D_{\mathrm{KL}}(\mathbb{P}^a(X_{ij}^{(1)}) \| \mathbb{P}^b(X_{ij}^{(1)})),$$

where the second equality follows since the r trials are all independent and identically distributed.

We now evaluate each individual term in right hand side of the above equation. Consider any $i, j \in [n]$. We divide our analysis into three disjoint cases:
Case I: Suppose that $i, j \in [n] \backslash \{a, b\}$. The distribution of $X_{ij}^{(1)}$ is identical across the distributions $\mathbb{P}^a$ and $\mathbb{P}^b$. As a result, we find that

$$D_{\mathrm{KL}}(\mathbb{P}^a(X_{ij}^{(1)}) \| \mathbb{P}^b(X_{ij}^{(1)})) = 0.$$

Case II: Suppose that $i = a, \; j \in [n] \backslash \{a, b\}$ or $i = b, \; j \in [n] \backslash \{a, b\}$. We then have

$$D_{\mathrm{KL}}(\mathbb{P}^a(X_{ij}^{(1)}) \| \mathbb{P}^b(X_{ij}^{(1)})) \leq p \frac{\delta^2}{(\frac{1}{2} - \delta)(\frac{1}{2} + \delta)}.$$

Case III: Suppose that $i = a, \; j = b$. We then have

$$D_{\mathrm{KL}}(\mathbb{P}^a(X_{ij}^{(1)}) \| \mathbb{P}^b(X_{ij}^{(1)})) \leq p \frac{(2\delta)^2}{(\frac{1}{2} - \delta)(\frac{1}{2} + \delta)}.$$

Combining the bounds from all three cases, we find that the KL divergence is upper bounded as

$$\frac{1}{r} D_{\mathrm{KL}}(\mathbb{P}^a \| \mathbb{P}^b) \leq 2(n-2)p \frac{\delta^2}{(\frac{1}{2} - \delta)(\frac{1}{2} + \delta)} + p \frac{(2\delta)^2}{(\frac{1}{2} - \delta)(\frac{1}{2} + \delta)}.$$

Some simple algebraic manipulations yield the claimed result.

5.2 Proof of Theorem 2

We now turn to the proof of Theorem 2. Beginning with the claim of sufficiency, it is easy to see that the ranking is correctly recovered whenever the top k items are correctly recovered for every value of $k \in [n]$. Consequently, one can apply the union bound to (10a) over all values of $k \in [n]$ and this gives the desired upper bound.

Now turning to the claim of necessity, we first introduce some notation to aid in subsequent discussion. Defining the parameter $\Delta_0 := \min_{j \in [n-1]} (\tau_{(j)} - \tau_{(j+1)})$, we have shown that the lower bound

$$\Delta_0 \geq 8 \sqrt{\frac{\log n}{npr}}$$

is sufficient to guarantee exact recovery of the full ranking. Further, one must also have

$$\Delta_0 \leq \frac{1}{n-1}\sum_{j=1}^{n-1}(\tau_{(j)} - \tau_{(j+1)}) = \frac{1}{n-1}(\tau_{(1)} - \tau_{(n)}) \leq \frac{1}{n-1}.$$

Here we show that this pair of requirements is jointly tight up to constant factors, meaning that for any value of Δ_0 satisfying $\Delta_0 \leq \frac{1}{9}\sqrt{\frac{\log n}{npr}}$ and $\Delta_0 \leq \frac{1}{9}\frac{1}{n-1}$, there are instances where recovery of the underlying ranking fails with probability at least $\frac{1}{70}$ for any estimator.

Consider the following ensemble of $(n-1)$ different problems, indexed by $a \in [n-1]$. For every value of $a \in [n-1]$, define a permutation π^a of the n items as

$$\pi^a(i) = \begin{cases} i+1 & \text{if } i = a \\ i-1 & \text{if } i = a+1 \\ i & \text{otherwise.} \end{cases}$$

In words, the permutation π^a equals the identity permutation except for the swapping of items a and $(a+1)$. Define an associated matrix of pairwise-comparison probabilities M^a as

$$M^a_{ij} = \frac{1}{2} - (\pi^a(i) - \pi^a(j))\Delta_0,$$

and $M^a_{ji} = 1 - M^a_{ij}$. Let $\mathbb{P}^a$ denote the probabilities taken under pairwise comparisons drawn according to the model M^a. The condition $\Delta_0 \leq \frac{1}{9}\frac{1}{n-1}$ ensures that this construction is a valid probability distribution. One can then compute that under distribution $\mathbb{P}^a$, the score τ^a_i of any item i equals

$$\tau^a_i = \frac{1}{2} - \big(\pi^a(i) - \frac{n+1}{2}\big)\Delta_0.$$

One can also verify that for any $a \in [n-1]$, and any $i \in [n-1]$, we have

$$\tau^a_{\pi^a(i)} - \tau^a_{\pi^a(i+1)} = \Delta_0,$$

where we have used the fact that $\pi^a(\pi^a(i)) = i$. The requirement imposed by the hypothesis is thus satisfied.

We now use Fano's inequality (19) obtain the claimed lower bound. In order to apply this result, we first obtain an upper bound on the Kullback-Leibler divergence between the probability distributions of the observed data under any pair of problems constructed above.

Lemma 8 *For any distinct pair $a, b \in [n-1]$, we have*

$$D_{\mathrm{KL}}(\mathbb{P}^a \| \mathbb{P}^b) \leq 50npr\Delta_0^2.$$

See the end of this section for the proof of this claim.

Given this bound on the Kullback-Leibler divergence, the Fano bound (19) implies that any method ϕ for identifying the true ranking has error probability

$$\mathbb{P}[\phi(Y) \neq A] \geq 1 - \frac{50npr\Delta_0^2 + \log 2}{\log(n-1)} \geq \frac{1}{70},$$

where the final inequality holds whenever $\Delta_0 \leq \frac{1}{9}\sqrt{\frac{\log n}{npr}}$ and $n \geq 9$.

The only remaining detail is the proof of Lemma 8.

5.2.1 Proof of Lemma 8

Since the distributions $\mathbb{P}^a$ and $\mathbb{P}^b$ are formed by components that are independent across edges $i > j$ and repetitions $\ell \in [r]$, we have

$$D_{\mathrm{KL}}(\mathbb{P}^a \| \mathbb{P}^b) = \sum_{\ell \in [r]} \sum_{1 \leq i < j \leq n} D_{\mathrm{KL}}(\mathbb{P}^a(X_{ij}^{(\ell)}) \| \mathbb{P}^b(X_{ij}^{(\ell)})) \;=\; r \sum_{1 \leq i < j \leq n} D_{\mathrm{KL}}(\mathbb{P}^a(X_{ij}^{(1)}) \| \mathbb{P}^b(X_{ij}^{(1)})),$$

where the second equality follows since the r trials are all independent and identically distributed.

We now evaluate each individual term in right hand side of the above equation. Consider any $i, j \in [n]$. We divide our analysis into three disjoint cases:

Case I: Suppose that $i, j \in [n] \backslash \{a, a+1, b, b+1\}$. The distribution of $X_{ij}^{(1)}$ is identical across the distributions $\mathbb{P}^a$ and $\mathbb{P}^b$. As a result, we find that

$$D_{\mathrm{KL}}(\mathbb{P}^a(X_{ij}^{(1)}) \| \mathbb{P}^b(X_{ij}^{(1)})) = 0.$$

Case II: Alternatively, suppose $i \in \{a, a+1, b, b+1\}$ and $j \in [n] \backslash \{a, a+1, b, b+1\}$ or if $j \in \{a, a+1, b, b+1\}$ and $i \in [n] \backslash \{a, a+1, b, b+1\}$. Then we have

$$D_{\mathrm{KL}}(\mathbb{P}^a(X_{ij}^{(1)}) \| \mathbb{P}^b(X_{ij}^{(1)})) \leq 5p\Delta_0^2,$$

where we have used the fact that $\mathbb{P}^a(X_{ij}^{(1)})$ and $\mathbb{P}^b(X_{ij}^{(1)})$ both take values in $[\frac{7}{18}, \frac{11}{18}]$ since $\Delta_0 \leq \frac{1}{9}\frac{1}{n-1}$.

Case III: Otherwise, suppose that both $i, j \in \{a, a+1, b, b+1\}$. Then we have

$$D_{\mathrm{KL}}(\mathbb{P}^a(X_{ij}^{(1)}) \| \mathbb{P}^b(X_{ij}^{(1)})) \leq 20p\Delta_0^2.$$

Combining the bounds from the three cases, we find that the KL divergence is upper bounded as

$$\frac{1}{r} D_{\mathrm{KL}}(\mathbb{P}^a \| \mathbb{P}^b) \leq 40(n-4)p\Delta_0^2 + 240p\Delta_0^2 \leq 50np\Delta_0^2,$$

where we have used the assumption $n \geq 9$ to obtain the final inequality.

5.3 Proof of Theorem 3

We now turn to the proof of Theorem 3, beginning with part (a).

5.3.1 Proof of part (a)

Without loss of generality, we can assume that the true underlying ranking is the identity ranking, that is, item i is ranked at position i for every $i \in [n]$. Given the the lower bound $\alpha \geq 8$ is satisfied, Theorem 1 ensures that with probability at least $1 - n^{-16}$, the counting estimator $\widetilde{\mathcal{S}}_k$ ranks every item in $\{1, \ldots, k-h\}$ higher than every item in the set $\{k+h+1, \ldots, n\}$. Thus, we are guaranteed that either $\widetilde{\mathcal{S}}_k \subseteq [k+h]$ and/or $[k-h] \subseteq \widetilde{\mathcal{S}}_k$. One can verify either case leads to $|\widetilde{\mathcal{S}}_k \cap [k]| \geq k-h$, thereby proving the claimed result.

5.3.2 Proof of part (b)

At a higher level, the crux of this proof is the construction of a packing set of pairwise comparison probability matrices, where every element of the set is also guaranteed to lie in the parametric classes and the SST class. The packing set is constructed via a careful application of a coding theoretic result due to Levenshtein (1971), such that the pairwise Kullback-Leibler divergence is small but the pairwise Hamming error is large enough, and that the packing set is also large enough. An application of Fano's inequality and some algebra yields the claimed result.

In more detail, we assume without loss of generality that $k \leq \frac{n}{2}$. (Otherwise, one can equivalently study the problem of recovering the last k items.) Since the case $h = 0$ is already covered by Theorem 1(b), we may also assume that $h \geq 1$.

The proof involves construction of $L \geq 1$ sets of probability matrices $\{M^a\}_{a \in [L]}$ of the pairwise comparisons with the following two properties:

(i) For every $a \in [L]$, let $S_k^a \subseteq [n]$ denote the set of the top k items under the a^{th} set of distributions. Then for every k-sized set $S \in [n]$,

$$\sum_{a=1}^{L} \mathbf{1}\{D_{\mathrm{H}}(S, S_k^a) \leq 2h\} \leq 1.$$

(ii) If the underlying distribution a is chosen uniformly at random from this set of L distributions, then any estimator that attempts to identify the underlying distribution $a \in [L]$ errs with probability at least $\frac{1}{7}$.

Now consider any estimator $\widehat{\mathcal{S}}_k$ for identifying the top k items $\mathcal{S}_k^*$. Given property (i), whenever the estimator is successful under the Hamming error requirement $D_{\mathrm{H}}(\widehat{\mathcal{S}}_k, \mathcal{S}_k^*) \leq 2h$, it must be able to uniquely identify the index $a \in [L]$ of the underlying distribution of pairwise comparison probabilities. However, property (ii) mandates that any estimator for identifying the underlying distribution errs with a probability at least $\frac{1}{7}$. Assuming that such sets of probability distributions satisfying these two properties exist, putting these results together yields the claimed result.

We now proceed to construct probability distributions satisfying the two aforementioned properties. Consider any positive number Δ_0 satisfying the upper bound

$$\Delta_0 \leq \frac{1}{14}\sqrt{\frac{\nu_1\nu_2 \log n}{npr}}. \tag{24}$$

The L matrices $\{M^a\}_{a\in[L]}$ of probability distributions we construct differ only in a permutation of their rows and columns, and modulo this permutation, have identical values. In other words, these L distributions differ only in the identities of the n items and the values of the pairwise-comparison probabilities $M^a_{(i)(j)}$ among the ordered sequence of the n items are identical across all distributions $a \in [L]$.

For any ordering $(1), \ldots, (n)$ of the n items, for every $a \in [L]$, set

$$M^a_{(i)(j)} = \begin{cases} \frac{1}{2} + \Delta_0 & \text{if } i \in [k] \text{ and } j \notin [k] \\ \frac{1}{2} - \Delta_0 & \text{if } i \notin [k] \text{ and } j \in [k] \\ \frac{1}{2} & \text{otherwise.} \end{cases} \tag{25}$$

Note that the upper bound (24) on Δ_0, coupled with the assumption $p \geq \sqrt{\frac{\log n}{2nr}}$, ensures that $\Delta_0 < \frac{1}{3}$ and hence that our definition (25) leads to a valid set of probabilities. Given this construction, the scores of the n items are $\tau_{(1)} = \cdots = \tau_{(k)} = \tau_{(k+1)} + \Delta_0 = \cdots = \tau_{(n)} + \Delta_0$. The bound (24) ensures that the condition $\alpha \leq \frac{\sqrt{\nu_1\nu_2}}{14}$ required by the hypothesis of the theorem is satisfied.

It remains to specify the ordering of the n items in each set of probability distributions. This specification relies on the following lemma, that in turn uses a coding-theoretic result due to Levenshtein (1971). It applies in the regime $2h \leq \frac{1}{1+\nu_2}\min\{n^{1-\nu_1}, k, n-k\}$ for some constants $\nu_1 \in (0,1)$ and $\nu_2 \in (0,1)$, and when n is larger than a (ν_1, ν_2)-dependent constant. For any pair of binary vectors b, b' of the same length, we define the Hamming error as $D_{\mathrm{H}}(b, b') = \sum_i \mathbf{1}\{b_i \neq b'_i\}$. We also let $\mathbf{0}$ denote the all-zero vector.

Lemma 9 *Under the previously given conditions, there exists a subset $\{b^1, \ldots, b^L\} \subseteq \{0,1\}^{n/2}$ with cardinality $L \geq e^{\frac{9}{10}\nu_1\nu_2 h \log n}$, such that*

$$D_{\mathrm{H}}(b^j, \mathbf{0}) = 2(1+\nu_2)h, \quad \textit{and} \quad D_{\mathrm{H}}(b^j, b^\ell) > 4h \quad \textit{for all } j \neq \ell \in [L].$$

We prove this lemma at the end of this section. Given this lemma, we now complete the proof of the theorem. Map the $\frac{n}{2}$ items $\{\frac{n}{2}+1, \ldots, n\}$ to the $\frac{n}{2}$ bits in each of the strings given by Lemma 9. For each $\ell \in [e^{\frac{9}{10}\nu_1\nu_2 h \log n}]$, let B_ℓ denote the $2(1+\nu_2)h$-sized subset of $\{\frac{n}{2}+1, \ldots, n\}$ corresponding to the $2(1+\nu_2)h$ positions equaling 1 in the ℓ^{th} string. Also define sets $A_\ell = \{1, \ldots, k - 2(1+\nu_2)h\}$ and $C_\ell = [n] \backslash (A_\ell \cup B_\ell)$. We note that this construction is valid since $2h \leq \frac{1}{1+\nu_2}k$.

We now construct $L = e^{\frac{9}{10}\nu_1\nu_2 h \log n}$ sets of pairwise comparison probability distributions $M^1, \ldots, M^L$ and show that these sets satisfy the two required properties. As mentioned earlier, each matrix of comparison-probabilities M^ℓ takes values as given in (25), but differs in the underlying ordering of the n items. In particular, associate the set $\ell \in [L]$ of distributions to any ordering of the n items that ranks every item in A_ℓ higher than every

item in B_ℓ, and every item in B_ℓ in turn higher than every item in C_ℓ. Then for any ℓ, the set of top k items is given by $A_\ell \cup B_\ell$. From the guarantees provided by Lemma 9, for any distinct $\ell, m \in [L]$, we have $D_{\mathrm{H}}(A_\ell \cup B_\ell, A_m \cup B_m) \geq 4h+1$. This construction consequently satisfies the first required property.

We now show that the construction also satisfies the second property: namely, it is difficult to identify the true index. We do so using Fano's inequality (19), for which we denote the probability distribution of the observations due to any matrix M^ℓ, $\ell \in [L]$, as $\mathbb{P}^\ell$.

We first derive an upper bound on the Kullback-Leibler divergence between any two distributions $\mathbb{P}^\ell$ and $\mathbb{P}^m$ of the observations. Observe that $[M^\ell]_{ij} \neq [M^m]_{ij}$ only if $i \in B_\ell \cup B_m$ or $j \in B_\ell \cup B_m$. In this case, we have $D_{\mathrm{KL}}([M^\ell]_{ij} \| [M^m]_{ij}) \leq \frac{4\Delta_0^2}{\frac{1}{4}-\Delta_0^2}$. Since both sets B_ℓ and B_m have a cardinality of $2(1+\nu_2)h$, aggregating over all possible observations across all pairs, we obtain that

$$D_{\mathrm{KL}}(\mathbb{P}^\ell \| \mathbb{P}^m) \leq 4(1+\nu_2)hnpr \frac{4\Delta_0^2}{\frac{1}{4}-\Delta_0^2}. \tag{26}$$

In the regime $p \geq \frac{\log n}{2nr}$ and $\Delta_0 \leq \frac{1}{14}\sqrt{\frac{\nu_1\nu_2 \log n}{npr}}$, we have $\Delta_0 \leq \frac{1}{14\sqrt{2}}$. Substituting the inequality $\Delta_0 \leq \frac{1}{14}\sqrt{\frac{\nu_1 \log n}{npr}}$ in the numerator and $\frac{1}{4} - \Delta_0^2 \geq \frac{1}{4} - \left(\frac{1}{14\sqrt{2}}\right)^2$ in the denominator of the right hand side of the bound (26), we find that

$$D_{\mathrm{KL}}(\mathbb{P}^\ell \| \mathbb{P}^m) \leq \frac{3}{4}\nu_1\nu_2 h \log n.$$

Now suppose that we draw Y from some distribution chosen uniformly at random from $\{\mathbb{P}^1, \ldots, \mathbb{P}^L\}$. Applying Fano's inequality (19) ensures that any test ϕ for estimating the index A of the chosen distribution must have error probability lower bounded as

$$\mathbb{P}\big[\phi(Y) \neq A\big] \geq \left(1 - \frac{\frac{3}{4}\nu_1\nu_2 h \log n + \log 2}{\frac{9}{10}\,\nu_1\nu_2 h \log n}\right) \geq \frac{1}{7}.$$

Here the final inequality holds as long as n is larger than some universal constant.

5.3.3 Proof of Lemma 9

We divide the proof into two cases depending on the value of h.

Case I: $h \geq \frac{1}{2\nu_1\nu_2}$: Let L denote the number of binary strings of length m_0 such that each has a Hamming weight w_0 and each pair has a Hamming distance at least d_0. It is known (Levenshtein, 1971; Jiang and Vardy, 2004) that L can be lower bounded as:

$$L \geq \frac{\binom{m_0}{w_0}}{\sum_{i=0}^{\lfloor \frac{d_0-1}{2} \rfloor} \binom{w_0}{j}\binom{m_0-w_0}{j}} \geq \frac{\left(\frac{m_0}{w_0}\right)^{w_0}}{\frac{d_0+1}{2}\left(\frac{ew_0}{\min\{d_0,w_0\}/2}\right)^{\min\{d_0,w_0\}/2}\left(\frac{em_0}{\min\{d_0,m_0\}/2}\right)^{\min\{d_0,m_0\}/2}}.$$

Note that for the setting at hand, we have $m_0 = \frac{n}{2}$, $w_0 = 2(1+\nu_2)h$ and $d_0 = 4h+1$. Since $\nu_1 \in (0,1)$ and $\nu_2 \in (0,1)$, we have the chain of inequalities

$$w_0 < d_0 \leq 4n^{1-\nu_1} \overset{(i)}{<} \frac{n}{2} = m_0,$$

where the inequality (i) holds when n is large enough. These relations allow for the simplification:

$$\log L \geq \log \left\{ \frac{\left(\frac{m_0}{w_0}\right)^{w_0}}{\frac{d_0+1}{2}\left(\frac{ew_0}{w_0/2}\right)^{w_0/2}\left(\frac{em_0}{d_0/2}\right)^{d_0/2}} \right\}$$
$$= (w_0 - d_0/2)\log m_0 - w_0 \log w_0 + \frac{d_0}{2}\log d_0 - \frac{d_0+w_0}{2}\log(2e) - \log((d_0+1)/2).$$

Substituting the values of w_0, d_0 and m_0 and then simplifying yields

$$\begin{aligned}\log L &\geq (2\nu_2 h - \frac{1}{2})\log\frac{n}{2} - 2(1+\nu_2)h\log(2(1+\nu_2)h) + (2h+\frac{1}{2})\log(4h+1) \\ &\quad - (((3+\nu_2)h) + \frac{1}{2})\log(2e) - \log(2h+1) \\ &\geq (2\nu_2 h - \frac{1}{2})\log\frac{n}{2} - 2\nu_2 h\log(2(1+\nu_2)h) - c_1' h,\end{aligned}$$

where c_1' is a constant whose value depends only on (ν_1, ν_2). In the regime $\frac{1}{\nu_1\nu_2} \leq 2h \leq \frac{n^{1-\nu_1}}{1+\nu_2}$, some algebraic manipulations then yield

$$\log L \geq (2\nu_1\nu_2 h - \frac{1}{2})\log\frac{n}{2} - c_2' h \geq \nu_1\nu_2 h(\log n - \log 2 - c_3') \geq \frac{9}{10}\nu_1\nu_2 h\log n,$$

where the final inequality holds when n is large enough, and where c_2' and c_3' are (ν_1, ν_2)-dependent positive constants.

Case II: $h < \frac{1}{2\nu_1\nu_2}$ Consider a partition of the $\frac{n}{2}$ bits into $\frac{n}{4(1+\nu_2)h}$ sets of size $2(1+\nu_2)h$ each. Define an associated set of $\frac{n}{4(1+\nu_2)h}$ binary strings, each of length $\frac{n}{2}$, with the i^{th} string having ones in the positions corresponding to the i^{th} set in the partition and zeros elsewhere. Then each of these strings have a Hamming weight of $2(1+\nu_2)h$, and every pair has a Hamming distance at least $4(1+\nu_2)h > 4h$. The total number of such strings equals

$$\exp\left(\log\frac{n}{4(1+\nu_2)h}\right) \overset{(i)}{\geq} \exp\left(\log n - \log(\frac{2(1+\nu_2)}{\nu_1\nu_2})\right) \overset{(ii)}{\geq} \exp\left(\frac{9}{10}\log n\right) \overset{(iii)}{>} \exp\left(1.8\nu_1\nu_2 h\log n\right),$$

where the inequalities (i) and (iii) are a result of operating in the regime $h < \frac{1}{2\nu_1\nu_2}$ and the inequality (ii) assumes that n is greater than a (ν_1, ν_2)-dependent constant.

5.4 Proof of Theorem 6

We now turn to the proof of Theorem 6.

5.4.1 Proof of part (a)

For every $i \in [n]$, let (i) denote the item ranked i according to their latent scores, as defined in equation (2). Recall from the proof of Theorem 1 that for any $u < v \in [n]$, the condition

$$\tau_{(u)} - \tau_{(v)} \geq 8\sqrt{\frac{\log n}{npr}}$$

ensures that with probability at least $1 - n^{-14}$, every item in the set $\{(1), \ldots, (u)\}$ wins more comparisons than every item in the set $\{(v), \ldots, (n)\}$. Consequently, if the set $\widetilde{\mathcal{S}}_k$ contains any item in $\{(v), \ldots, (n)\}$, then it must contain the entire set $\{(1), \ldots, (u)\}$. In other words, at least one of the following must be true: either $\{(1), \ldots, (u)\} \subseteq \widetilde{\mathcal{S}}_k$ or $\widetilde{\mathcal{S}}_k \subseteq \{(1), \ldots, (v-1)\}$. Consequently, in the regime $v = k + t - u + 1$ for any $1 \le u \le k$ and $u \le t \le n$, we have that

$$|\widetilde{\mathcal{S}}_k \cap \{(1), \ldots, (t)\}| \ge u. \tag{27}$$

Now consider any $b \in [\beta]$ that satisfies the condition

$$\min_{j \in [k]} (\tau_{(j)} - \tau_{(k+t_j^b - j + 1)}) \ge 8\sqrt{\frac{\log n}{npr}}.$$

For any $j \in [k]$, setting $u = j$ and $v = (k + t_j^b - j + 1)$ in (27), and applying the union bound over all values of $j \in [k]$ yields that

$$|\widetilde{\mathcal{S}}_k \cap \{(1), \ldots, (t_j^b)\}| \ge j \quad \text{for every } j \in [k],$$

with probability at least $1 - n^{-13}$. Consequently, we have that

$$\mathbb{P}\big(\widetilde{\mathcal{S}}_k \in \Lambda(T_b)\big) \ge 1 - n^{-13},$$

completing the proof of the claim.

5.4.2 Proof of part (b)

In the regime $t_{\mu_2 k}^b \le \frac{n}{2}$ for every $b \in [\beta]$, it suffices to show that any estimator $\widehat{\mathcal{S}}_k$ will incur an error lower bounded as

$$\mathbb{P}\big(|\widehat{\mathcal{S}}_k \cap \{(1), \ldots, (n/2)\}| < \mu_2 k\big) \ge \frac{1}{15},$$

where (i) denotes the item ranked i according to their latent scores according to equation (2).

Our proof relies on the result and proof of the Hamming error case analyzed in Theorem 3(b). To this end, let us set the parameter h of Theorem 3(b) as $h = 2(1 - \mu_2)k$. We claim that this value of h lies in the regime $h \le \frac{1}{2(1+\nu_2)} \min\{k, n-k, n^{1-\nu_1}\}$ for some values $\nu_1 \in (0,1)$ and $\nu_2 \in (0,1)$, as required by Theorem 3(b). This claim follows from the fact that

$$h = 2(1 - \mu_2)k \le \frac{1}{2(1+\nu_2)} k,$$

for $\nu_2 = \min\{\frac{1}{4(1-\mu_2)} - 1, \frac{1}{2}\} \in (0,1)$. Furthermore,

$$h = 2(1 - \mu_2)k \overset{(i)}{\le} \frac{n^{1-\mu_1}}{4} \overset{(ii)}{\le} \frac{1}{2(1+\nu_2)} n^{1-\nu_1}$$

for $\nu_1 = \frac{9}{10}\mu_1 \in (0,1)$, where (i) is a result of our assumption $8(1 - \mu_2)k \le n^{1-\mu_1}$ and (ii) holds when n is large enough. This assumption also implies that $k \le n - k$ for a large enough value of n. We have now verified operation in the regime required by Theorem 3(b).

The construction in the proof of Theorem 3 is based on setting

$$\tau_{(1)} = \cdots \tau_{(k)} = \tau_{(k+1)} + \Delta_0 = \cdots = \tau_{(n)} + \Delta_0,$$

for any real number Δ_0 in the interval $\Big(0,\ \frac{1}{14}\sqrt{\frac{\nu_1\nu_2 \log n}{npr}}\ \Big]$. This condition is also satisfied in our construction due to the assumed upper bound $\alpha \leq \frac{1}{15}\sqrt{\mu_1 \min\big\{\frac{1}{4(1-\mu_2)-1}, \frac{1}{2}\big\}}$. Consequently, the result of Theorem 3(b) implies that in this setting, any estimator $\widehat{\mathcal{S}}_k$ will incur a Hamming error greater than $h = 2(1-\mu_2)k$ with probability at least $\frac{1}{7}$, or equivalently,

$$\mathbb{P}\big(|\widehat{\mathcal{S}}_k \cap \{(1), \ldots, (k)\}| < (2\mu_2 - 1)k\big) \geq \frac{1}{7}.$$

Under this event, the estimator $\widehat{\mathcal{S}}_k$ contains at most $(2\mu_2 - 1)k - 1$ items from the set of top k items. In order to ensure it gets at least $\mu_2 k$ items from $\{(1), \ldots, (n/2)\}$, the remaining $2(1-\mu_2)k+1$ chosen items must have at least $(1-\mu_2)k+1$ items from $\{(k+1), \ldots, (n/2)\}$. However, in the construction, items $(k+1), \ldots, (n)$ are indistinguishable from each other, and hence by symmetry these $2(1-\mu_2)k+1$ chosen items must contain at least $(1-\mu_2)k+1$ items from the set $\{(n/2+1), \ldots, (n)\}$ with probability at least $\frac{1}{2}$. Putting these arguments together, we obtain that under this construction, any estimator $\widehat{\mathcal{S}}_k$ has error probability lower bounded as

$$\mathbb{P}\big(|\widehat{\mathcal{S}}_k \cap \{(1), \ldots, (n/2)\}| < \mu_2 k\big) \geq \frac{1}{14}. \tag{28}$$

It remains to deal with a subtle technicality. The construction above involves items $(k+1), \ldots, (n)$ with identical scores. Recall that in the definition of the user-defined requirement, in case of multiple items with identical scores, we considered the choice of either of such items as valid. The following lemma helps overcome this issue.

Lemma 10 *Consider any two $(n \times n)$ matrices M^a and M^b of pairwise probabilities such that*

$$\max_{(i,j)\in[n]^2} |[M^a]_{ij} - [M^b]_{ij}| \leq \epsilon, \tag{29a}$$

for some $\epsilon \in [0,1]$. Then for any k-sized sets of items $T_1, \ldots, T_\beta \subseteq [n]$*, and any estimator $\widehat{\mathcal{S}}_k$, we have*

$$|\ \mathbb{P}_{M^a}(\widehat{\mathcal{S}}_k \in \{T_1, \ldots, T_\beta\}) - \mathbb{P}_{M^b}(\widehat{\mathcal{S}}_k \in \{T_1, \ldots, T_\beta\})\ | \leq 6^{n^2 r}\epsilon. \tag{29b}$$

See Section 5.4.3 for the proof of this claim.

Now consider an $(n \times n)$ pairwise probability matrix M' whose entries take values

$$M'_{(i)(j)} = \begin{cases} \frac{1}{2} + \Delta_0 + \epsilon & \text{if } i \in [k] \text{ and } j \in [n]\backslash[n/2] \\ \frac{1}{2} + \Delta_0 & \text{if } i \in [k] \text{ and } j \in [n/2]\backslash[k] \\ \frac{1}{2} + \epsilon & \text{if } i \in [n/2]\backslash[k] \text{ and } j \in [n]\backslash[n/2] \\ \frac{1}{2} & \text{otherwise,} \end{cases}$$

and $M'_{ji} = 1 - M'_{ij}$, whenever $i \leq j$. Set $\epsilon = 7^{-n^2 r}$.

One can verify that under the probability matrix M', the scores of the n items satisfy the relations

$$\tau_{(1)} = \cdots = \tau_{(k)} = \tau_{(k+1)} + \Delta_0 = \cdots = \tau_{(n/2)} + \Delta_0 = \tau_{(n/2+1)} + \Delta_0 + \epsilon = \cdots = \tau_{(n)} + \Delta_0 + \epsilon.$$

The set of items $\{(1), \ldots, (n/2)\}$ are thus explicitly distinguished from the items $\{(n/2 + 1), \ldots, (n)\}$. We now call upon Lemma 10 with $M^a = M'$, and M^b as the matrix of probabilities constructed in the proof of Theorem 3, where both sets have the same ordering of the items. This assignment is valid given that $\Delta_0 < \frac{1}{3}$ and $\epsilon = 7^{-n^2 r}$. Lemma 10 then implies that any estimator that is $\mathfrak{S}$-respecting with probability at least $1 - \frac{1}{15}$ under M^b must also be $\mathfrak{S}$-respecting with probability at least $1 - \frac{1}{14.5}$ under M^a. But by equation (28), the latter condition is impossible, which implies our claimed lower bound.

5.4.3 Proof of Lemma 10

Let $\mathbb{P}^a$ and $\mathbb{P}^b$ denote the probabilities induced by the matrices M^a and M^b respectively. Consider any fixed observation $Y_1 \subseteq \{0, 1, \phi\}^{n \times n \times r}$, where ϕ denotes the absence of an observation. Let $\mathbb{P}^a(Y = Y_1)$ and $\mathbb{P}^b(Y = Y_1)$ denote the probabilities of observing Y_1 under $\mathbb{P}^a$ and $\mathbb{P}^b$, respectively. Given the bounds (29a), some algebra leads to

$$\begin{aligned}
| \, \mathbb{P}^a(Y = Y_1) - \mathbb{P}^b(Y = Y_1) \, | &\leq \max_{u \in [0, 1-\epsilon]^{n^2 r}} \Big(\prod_{i=1}^{n^2 r} (u_i + \epsilon) - \prod_{i=1}^{n^2 r} u_i \Big) \\
&\leq \max_{u \in [0, 1-\epsilon]^{n^2 r}} \Big(u_{n^2 r} \big(\prod_{i=1}^{n^2 r - 1} (u_i + \epsilon) - \prod_{i=1}^{n^2 r - 1} u_i \big) + \epsilon \Big) \\
&\vdots \\
&\leq n^2 r \epsilon. \qquad (30)
\end{aligned}$$

Now consider any estimator $\widehat{\mathcal{S}}_k$, which is permitted to be randomized. Let $L \leq 3^{n^2 r}$ denote the total number of possible values of the observation Y, and let $\{Y_1, \ldots, Y_L\} = \{0, 1, \phi\}^{n \times n \times r}$ denote the set of all possible valid values of the observation. For each $i \in [L]$, let $q_i \in [0, 1]$ denote the probability that the estimator $\widehat{\mathcal{S}}_k$ succeeds in satisfying the given requirement when the data observed equals Y_i. (Recall that the given requirement is in terms of the actual items and not their positions.) Then we have

$$\begin{aligned}
\big| \mathbb{P}^1(\widehat{\mathcal{S}}_k \in \{T_1, \ldots, T_\beta\}) - \mathbb{P}^2(\widehat{\mathcal{S}}_k \in \{T_1, \ldots, T_\beta\}) \big| &= \big| \sum_{i=1}^{L} \mathbb{P}^1(Y = Y_i) q_i - \sum_{i=1}^{L} \mathbb{P}^2(Y = Y_i) q_i \big| \\
&\leq \sum_{i=1}^{L} | \, \mathbb{P}^1(Y = Y_i) - \mathbb{P}^2(Y = Y_i) \, | \, q_i \\
&\overset{(i)}{\leq} \sum_{i=1}^{L} n^2 r \epsilon q_i \overset{(ii)}{\leq} 6^{n^2 r} \epsilon,
\end{aligned}$$

as claimed, where step (i) follows from our earlier bound (30) and step (ii) uses the bounds $L \leq 3^{n^2 r}$ and $n^2 r \leq 2^{n^2 r}$.

6. Discussion

In this paper, we analyzed the problem of recovering the k most highly ranked items based on observing noisy comparisons. We proved that an algorithm that simply selects the items that win the maximum number of comparisons is, up to constant factors, an information-theoretically optimal procedure. Our results also extend to recovering the entire ranking of the items. The results of this paper thus underscore the philosophy of *Occam's razor* that the simplest answer is often correct.

Empirical evaluations reveal the superior performance of the counting algorithm that we analyzed through our "permutation-based" approach as compared to the Spectral MLE algorithm. The intuition is that Spectral MLE is too tied to the restrictive parameter-based model and the model mismatch in this crowdsourcing data causes the high amount of error. On the other hand, the robustness and accuracy guarantees of the count estimator due to our permutation-based approach carry over to practice. More generally, parameter-based models are popular in many applications in part because they are quite intuitive to write down, and in part because they are sometimes analytically more tractable. However, instead if one were to consider rich enough models like permutation-based models then they may yield a broader perspective and richer insights into the problem that can lead to superior results (Shah, 2017, Chapter 1, Part 1).

There are number of open questions suggested by our work. The notion of allowed sets introduced in this paper apply to recovery of k-sized subsets of the items; such a formulation and associated results may apply to recovery of partial or total orderings of the items. The observation model considered here is based on a random number of observations for all pairs of comparisons. It would be interesting to extend our results to cases in which only specific subsets of pairs are observed, and particularly when these pairs are chosen adversarially. A parallel line of literature (e.g., Kaufmann and Kalyanakrishnan, 2013; Busa-Fekete et al., 2013; Jamieson et al., 2015; Heckel et al., 2016) studies settings in which the pairs to be compared can be chosen sequentially in a data-dependent manner, but to the best of our knowledge, this line of literature considers only the metric of exact recovery of the top k items. It is of interest to investigate the Hamming and allowed set recovery problems in such an active setting. Finally, it will also be useful to obtain analogous results for ranking problems where the identity of the person making the comparison is known and influences the outcomes of the comparison, for instance, in applications of peer-grading (Shah et al., 2013; Song et al., 2017) and peer-reviews (Shah et al., 2017c).

Acknowledgments

We would like to thank the AE Sujay Sanghavi for the efficient handling of the paper and the anonymous reviewers for their valuable comments. This work was partially supported by National Science Foundation Grant NSF-DMS-1612948; DOD Advanced Research Projects Agency W911NF-16-1-0552 and Office of Naval Research grant DOD ONR-N00014. In addition, NBS was supported in part by a Microsoft Research PhD fellowship.

References

Ammar, A. and Shah, D. Efficient rank aggregation using partial data. In *ACM SIGMETRICS Performance Evaluation Review*, 2012.

Babcock, B. and Olston, C. Distributed top-k monitoring. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, 2003.

Ballinger, T. P. and Wilcox, N. Decisions, error and heterogeneity. *The Economic Journal*, 107(443):1090–1105, 1997.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Bradley, R. and Terry, M. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 1952.

Braverman, M. and Mossel, E. Noisy sorting without resampling. In *Proc. ACM-SIAM symposium on Discrete algorithms*, 2008.

Busa-Fekete, R., Szorenyi, B., Cheng, W., Weng, P., and Hüllermeier, E. Top-k selection based on adaptive sampling of noisy preferences. In *International Conference on Machine Learning*, 2013.

Carmel, D., Cohen, D., Fagin, R., Farchi, E., Herscovici, M., Maarek, Y. S., and Soffer, A. Static index pruning for information retrieval systems. In *ACM SIGIR conference on Research and development in information retrieval*, 2001.

Carpenter, A., Jones, T., Lamprecht, M., Clarke, C., Kang, I., Friman, O., Guertin, D., Chang, J., Lindquist, R., Moffat, J., and Golland, P. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7(10):R100, 2006.

Chatterjee, S. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2014.

Chen, Y. and Suh, C. Spectral MLE: Top-k rank aggregation from pairwise comparisons. In *International Conference on Machine Learning*, 2015.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2012.

Davidson, D. and Marschak, J. Experimental tests of a stochastic decision theory. *Measurement: Definitions and theories*, 1959.

de Borda, J. C. Mémoire sur les élections au scrutin. *Histoire de l\'Academie Royale des Sciences*, 1781.

Ding, W., Ishwar, P., and Saligrama, V. A topic modeling approach to ranking. In *Conference on Artificial Intelligence and Statistics*, 2015.

Erdős, P. and Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960.

Eriksson, B. Learning to top-k search using pairwise comparisons. In *Conference on Artificial Intelligence and Statistics*, 2013.

Fagin, R., Lotem, A., and Naor, M. Optimal aggregation algorithms for middleware. *Journal of computer and system sciences*, 66(4):614–656, 2003.

Hajek, B., Oh, S., and Xu, J. Minimax-optimal inference from partial rankings. In *Advances in Neural Information Processing Systems*, 2014.

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

Hunter, D. MM algorithms for generalized Bradley-Terry models. *Annals of Statistics*, 2004.

Ilyas, I. F., Beskales, G., and Soliman, M. A. A survey of top-k query processing techniques in relational database systems. *ACM Computing Surveys*, 2008.

Jagabathula, S. and Shah, D. Inferring rankings under constrained sensing. In *Advances in Neural Information Processing Systems*, 2008.

Jamieson, K., Katariya, S., Deshpande, A., and Nowak, R. Sparse dueling bandits. *arXiv preprint arXiv:1502.00133*, 2015.

Jiang, T. and Vardy, A. Asymptotic improvement of the Gilbert-Varshamov bound on the size of binary codes. *IEEE Transactions on Information Theory*, 2004.

Kaufmann, E. and Kalyanakrishnan, S. Information complexity in bandit subset selection. In *Conference on Learning Theory*, 2013.

Kenyon-Mathieu, C. and Schudy, W. How to rank with few errors. In *Symposium on Theory of computing (STOC)*. ACM, 2007.

Kimelfeld, B. and Sagiv, Y. Finding and approximating top-k answers in keyword proximity search. In *Symposium on Principles of database systems*, 2006.

Levenshtein, V. I. Upper-bound estimates for fixed-weight codes. *Problemy Peredachi Informatsii*, 7(4):3–12, 1971.

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

McLaughlin, D. H. and Luce, R. D. Stochastic transitivity and cancellation of preferences between bitter-sweet solutions. *Psychonomic Science*, 1965.

Metwally, A., Agrawal, D., and El Abbadi, A. Efficient computation of frequent and top-k elements in data streams. In *Database Theory-ICDT*. 2005.

Michel, S., Triantafillou, P., and Weikum, G. Klee: A framework for distributed top-k query algorithms. In *International conference on Very large data bases*, 2005.

Mitliagkas, I., Gopalan, A., Caramanis, C., and Vishwanath, S. User rankings from comparisons: Learning permutations in high dimensions. In *Allerton Conference on Communication, Control, and Computing*, 2011.

Negahban, S., Oh, S., and Shah, D. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*, 2012.

Rajkumar, A. and Agarwal, S. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *International Conference on Machine Learning*, 2014.

Rajkumar, A., Ghoshal, S., Lim, L.-H., and Agarwal, S. Ranking from stochastic pairwise preferences: Recovering Condorcet winners and tournament solution sets at the top. In *International Conference on Machine Learning*, 2015.

Shah, N. B. *Learning From People*. PhD thesis, EECS Department, University of California, Berkeley, 2017.

Shah, N. B., Bradley, J. K., Parekh, A., Wainwright, M., and Ramchandran, K. A case for ordinal peer-evaluation in MOOCs. In *NIPS Workshop on Data Driven Education*, December 2013.

Shah, N. B., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K., and Wainwright, M. J. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal on Machine Learning Research*, 2016a.

Shah, N. B., Balakrishnan, S., and Wainwright, M. J. A permutation-based model for crowd labeling: optimal estimation and robustness. *arXiv:1606.09632*, 2016c.

Shah, N. B., Balakrishnan, S., and Wainwright, M. J. Feeling the Bern: Adaptive estimators for Bernoulli probabilities of pairwise comparisons. In *International Symposium on Information Theory*, 2016d.

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

Shah, N. B., Balakrishnan, S., and Wainwright, M. J. Low permutation-rank matrices: Structural properties and noisy completion. *arXiv preprint arXiv:1709.00127*, 2017b.

Shah, N. B., Tabibian, B., Muandet, K., Guyon, I., and von Luxburg, U. Design and analysis of the nips 2016 review process. *arXiv preprint arXiv:1708.09794*, 2017c.

Song, Y., Yifan, G., and Edward, G. An exploratory study of reliability of ranking vs. rating in peer assessment. In *ICALT*, volume 2017, 2017.

Soufiani, H., Parkes, D., and Xia, L. Computing parametric ranking models via rank-breaking. In *International Conference on Machine Learning*, 2014.

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

Tsybakov, A. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. 2008.

Tversky, A. Elimination by aspects: A theory of choice. *Psychological review*, 79(4):281, 1972.

Wauthier, F., Jordan, M., and Jojic, N. Efficient ranking from pairwise comparisons. In *International Conference on Machine Learning*, 2013.