Approximate Maximum Rank Aggregation: Beyond the Worst-Case

Yan Hong Yao Alvin
National University of Singapore, Singapore

Diptarka Chakraborty
National University of Singapore, Singapore

Abstract

The fundamental task of *rank aggregation* is to combine multiple rankings on a group of candidates into a single ranking to mitigate biases inherent in individual input rankings. This task has a myriad of applications, such as in social choice theory, collaborative filtering, web search, statistics, databases, sports, and admission systems. One popular version of this task, *maximum rank aggregation* (or the *center ranking* problem), aims to find a ranking (not necessarily from the input set) that minimizes the maximum distance to the input rankings. However, even for four input rankings, this problem is NP-hard (Dwork et al., WWW'01, and Biedl et al., Discrete Math.'09), and only a (folklore) polynomial-time 2-approximation algorithm is known for finding an optimal aggregate ranking under the commonly used *Kendall-tau distance* metric. Achieving a better approximation factor in polynomial time, ideally, a polynomial time approximation scheme (PTAS), is one of the major challenges.

This paper presents significant progress in solving this problem by considering the *Mallows model*, a classical probabilistic model. Our proposed algorithm outputs an $(1+\varepsilon)$-approximate aggregate ranking for any $\varepsilon > 0$, with high probability, as long as the input rankings come from a Mallows model, even in a streaming fashion. Furthermore, the same approximation guarantee is achieved even in the presence of *outliers*, presumably a more challenging task.

2012 ACM Subject Classification Theory of computation → Probabilistic computation; Theory of computation → Facility location and clustering; Theory of computation → Theory and algorithms for application domains

Keywords and phrases Rank Aggregation, Center Problem, Mallows Model, Approximation Algorithms, Clustering with Outliers

Digital Object Identifier 10.4230/LIPIcs.FSTTCS.2023.12

Funding This work received funding from an MoE AcRF Tier 2 grant (MOE-T2EP20221-0009) and Google South & South-East Asia Research Award.

1 Introduction

Ranking a set of candidates or a list of alternatives is ubiquitous in diverse fields ranging from social choice theory [7] to information retrieval [23]. Given a set of different, potentially conflicting rankings of a list of candidates, a fundamental task is to combine them to obtain a consensus ranking that best represents the preferences in the individual rankings. This process, known as *rank aggregation*, has applications in many areas, including voting systems, sports, databases, web search engines, collaborative filtering, machine translation, and statistics [19, 20, 41, 22]. One of the primary objectives is to mitigate apparent biases in individual rankings, such as search engine spam, strategic manipulation of web pages to achieve an undeserved high rank, or biases towards candidates/alternatives based on gender, ethnicity, and so on.

Rank aggregation is often formulated as an optimization problem. One popular version involves finding a ranking (not necessarily from the input set) that minimizes the maximum distance to the input rankings, known as the *maximum rank aggregation* problem in the

literature. The optimization task considered here is the well-known *center* problem, with the underlying domain being the set of rankings/permutations. The general center problem dates back to the nineteenth century [44] and has been studied for various metric spaces, including Euclidean (both constant [36] and high dimension [4, 46]), Hamming [21, 30, 34], the edit metric [37], Jaccard distance [9], rankings [3, 5, 40, 13], and others. The hardness of the center problem varies with the metric space considered.

Several distance metrics have been explored to quantify the dissimilarity between two rankings or permutations. The *Kendall-tau distance* [27, 17, 26, 47, 48], which counts the number of inversions between two rankings (equivalently, the number of transpositions required by the *bubble sort* algorithm), is perhaps the most widely used metric. It is known to satisfy several desirable properties, such as neutrality, consistency, and the extended Condorcet property [26, 47]. In this paper, we focus on the maximum rank aggregation problem under the Kendall-tau metric, as is common in most of the rank aggregation literature [19, 1, 28, 3, 5]. Unfortunately, the problem is known to be NP-hard, even when there are only four input rankings [19, 5]. For most practical applications, computing an approximate maximum aggregate ranking suffices. A simple (folklore) algorithm that outputs one of the input rankings can achieve a 2-approximation factor, which is the case for any arbitrary metric space. However, finding a polynomial-time algorithm that can achieve a better than 2-approximation factor remains a significant open challenge. The ultimate goal is to develop a polynomial-time approximation scheme (PTAS) that can provide a $(1+\varepsilon)$-approximation algorithm for any $\varepsilon > 0$.

Given the difficulty of the maximum rank aggregation problem, it is natural to consider this problem in the context of probabilistic models (i.e., when the input is generated from some probabilistic model). Various probabilistic ranking models have been proposed in the literature to capture preference or choice behavior, such as Bradley-Terry model [6], Luce model [33], Plackett-Luce model [39], and Cayley-Mallows model [16], along with several ranking algorithms and statistical analysis techniques [25, 43, 24]. Mallows model [35], a subclass of the Bradley-Terry model for noisy rankings, is a more tractable probabilistic ranking model. The model is defined by an underlying ranking π and a *dispersion* parameter $\phi \in (0, 1)$, and the probability of observing a ranking σ decreases exponentially with respect to the Kendall-tau distance between π and σ and the dispersion parameter ϕ. Being a problem of utter importance, the rank aggregation problem under the Mallows model has already sought attention from the researchers [8, 10, 45], albeit mostly on minimizing the sum of distances (the *median* objective) instead of minimizing the maximum distance (that we consider in this paper), where the latter task is considered to be more challenging than the first one in the context of aggregating rankings.

1.1 Our Contribution

In this paper, we study the maximum rank aggregation problem when the input is sampled from the Mallows model. So far, nothing better than the folklore 2-approximation algorithm is known, even under this probabilistic ranking model. We make significant progress by providing an $(1+\varepsilon)$-approximation algorithm.

▶ **Theorem 1.** *Consider a $\phi \in (0, 1)$, $\varepsilon > 0$ and a ranking π on n candidates. There is a polynomial-time algorithm that, given a set S of size at least $\frac{50}{(1-\phi)^2}\left(\log\frac{9}{\varepsilon\phi} + \log\frac{1}{1-\phi}\right)$ drawn from a Mallows model $M(\pi, \phi)$, outputs a permutation π_c that is a $(1+\varepsilon)$-approximate center (maximum aggregate) ranking with probability at least $2/3$.*

Although, for simplicity, we state the success probability to be 2/3, a slightly careful look into our analysis[1] reveals that we can actually attain a success probability of at least $1 - e^{-\Omega\left((1-\phi)^2|S|\right)}$. Our algorithm runs in $O(n^3 \log n)$ time, where the $O(\cdot)$ notation hides a constant that depends on $1/\varepsilon$ and $1/(1-\phi)$. The running time can further be reduced to $O(n \log^2 n)$ when the size of the input set is at least $\Omega(\log n)$. We also remark that as such, our algorithm does not need to know the parameter ϕ as long as the input set size satisfies the required minimum bound of $\frac{50}{(1-\phi)^2}\left(\log \frac{9}{\varepsilon\phi} + \log \frac{1}{1-\phi}\right)$. It is worth highlighting that when ϕ is close to 1, the Mallows model defines a distribution close to the uniform distribution. A set of rankings drawn from such a (close to) uniform distribution has little similarity; thus, aggregating them into a representative ranking has little importance. Therefore, the problem is more interesting when ϕ is away from 1, where our algorithm only needs a constant (dependent on parameters ε, ϕ) number of samples. We pose the problem of showing whether so many samples are essential to achieve a $(1+\varepsilon)$-approximation as an interesting open problem.

Our result consists of two parts. In the first part, we prove that when the input set is drawn from a Mallows model $M(\pi, \phi)$, the underlying ranking π gives an additive approximation of $\frac{n}{1-\phi} e^{-\frac{s(1-\phi)^2}{50}}$, where the input set is of size s (Theorem 5). Such an equivalence between the hidden ranking in the Mallows model and an approximate center ranking was not known before. We exploit the relation between the Mallows model and the *Repeated Insertion model* to show our result. This allows us to argue that the distances of rankings in the input set to π are almost the same (Lemma 7). To show the additive approximation guarantee, we consider an arbitrary optimal center ranking σ and then view it as being obtained from π by performing a sequence of swap operations of adjacent symbols (as in bubble sort). Ideally, one would like to show that such swaps would not reduce the (center) objective value by much. However, that is not true. We argue that such swaps can be categorized into two types. In the first one, on a pair of symbols involved in the swap, π and the majority of input rankings agree about their relative ordering. We argue that performing swaps on them cannot significantly improve the (center) objective value. Roughly speaking, the decrease in the distance caused by these swaps "almost equally" spreads out over all the input rankings. On the other hand, the second category constitutes all the pairs of symbols involved in swaps for which π and the majority of input rankings do not agree about their relative ordering. Although these swaps can improve the objective value, we argue that only a few such pairs exist (Lemma 9). Consequently, we derive that the objective value obtained by π is up to a small additive factor of the optimum objective value (i.e., the objective value obtained by σ).

Next, we show how to reconstruct a ranking π_c that is close (in the Kendall-tau distance) to π, without even knowing the parameter ϕ as long as there are sufficiently many input samples (Theorem 10). For that purpose, we build a *tournament graph* by looking into the pairwise relative ordering of the candidates in the majority of the input rankings and then use a PTAS by [28, 42] to solve the *feedback arc set* problem on this instance of tournament graph. Then, we remove the edge set returned by that PTAS from the constructed tournament graph and, finally, use the topological ordering of the resulting acyclic graph. The main difficulty lies in the analysis, where we show that this algorithm indeed reconstructs π up to some small distance. Interestingly, we only need a constant (depending on the dispersion parameter ϕ and the closeness requirement) number of input rankings for the reconstruction.

[1] The success probability of Lemma 9 can easily be improved to $1 - e^{-\Omega\left((1-\phi)^2|S|\right)}$ by adjusting certain parameters (and bounds) in the proof, and as a consequence, a similar success probability holds for both Theorem 5 and Theorem 10, and thus also for Theorem 1.

It is worth highlighting that the problem of approximate reconstruction of π (for the Mallows model) has been studied before [14]. However, our reconstruction result significantly differs from that of [14], primarily because [14] requires $\Omega(\log n)$ samples (with $\Omega(\cdot)$ hiding a factor that depends on $1/(1-\phi)$) while achieving a slightly different approximation guarantee, and thus the reconstruction result of [14] does not help us proving Theorem 1.

Since we only need constantly many inputs for the reconstruction, we can implement our algorithm in a streaming manner (when the input rankings arrive one by one in a streaming fashion) only using $O(n)$ space (note, the input size is $O(n|S|)$) by using a standard sampling technique. Furthermore, our algorithm can be extended to a more general setting (see Section 6) when the input set contains a small number of *outliers* chosen by the adversary (similar to that in [32]).

1.2 Related Works

In addition to the center objective that is the main focus of this paper, the rank aggregation problem has also been studied under the median objective function [19, 1]. This variant seeks to find a ranking (not necessarily from the input set) that minimizes the sum of distances to the input rankings. Under the Kendall-tau metric, the median ranking problem is also NP-hard [19], but it has a PTAS [28]. However, finding a similar PTAS or proving an impossibility result for the center variant remains a significant challenge. While the rank aggregation problem has been studied under Spearman's rho distance [19] or Ulam distance [11, 13, 12], Kendall-tau distance is the most prevalent for aggregating ranks.

The difficulty of the rank aggregation task has led researchers to explore data-driven approaches, such as using various probabilistic models [38, 8, 29]. The Mallows model is a classic probabilistic model for rankings, and [8] presented an algorithm that can find an optimal median of rankings drawn from a Mallows model in polynomial time. Other work has focused on estimating parameters and reconstructing the hidden permutation in more general mixtures of Mallows models [14, 2, 31, 15] or in the presence of outliers chosen by a malicious adversary [32]. [11] investigated a different probabilistic model for finding an approximate median ranking (although under the Ulam metric), but none of these results have yielded a better approximate center ranking than the folklore algorithm under the Mallows model.

2 Preliminaries

Let $[n]$ denote the set $\{1, 2, ..., n\}$. Let S_n denote the set of all rankings/permutations over $[n]$. We use the terms ranking and permutation interchangeably in this paper. Given a ranking π, let the notation $\pi(i)$ refer to the symbol at rank i in π. Let the notation $\pi^{-1}(a)$ refer to the rank of the symbol a in π. Given a pair of symbols a, b, we say $a \prec_\pi b$ if $\pi^{-1}(a) < \pi^{-1}(b)$. Throughout this paper, we will use the variables i, j when referring to ranks and a, b when referring to symbols. Also, note that all the logarithms in this paper are base e.

We consider the distance between the two rankings as their Kendall-tau distance.

▶ **Definition 2** (Kendall-tau Distance). *Given two rankings π, σ, the Kendall-tau distance, denoted by $d(\pi, \sigma)$, is the total number of ordered pairs $(a, b) \in [n] \times [n]$ such that $a \prec_\pi b$ and $b \prec_\sigma a$.*

Suppose we have some pair of symbols (a, b) and two rankings π and σ such that $a \prec_\pi b$ but $b \prec_\sigma a$. We say that the pair (a, b) is inverted in relative order in σ compared to π.

Center and c-approximate center

Given a set $S \subseteq S_n$ and ranking $\sigma \in S_n$, we refer to the center objective value of σ with respect to S as $\texttt{Obj}(S, \sigma) := \max_{\pi_r \in S} d(\sigma, \pi_r)$. Then an (optimal) center of a set $S \subseteq S_n$ is a ranking $\pi^* \in S_n$ that minimizes $\texttt{Obj}(S, \sigma)$, i.e., $\pi^* = \arg\min_{\sigma \in S_n} \texttt{Obj}(S, \sigma)$. We refer to $\texttt{Obj}(S, \pi^*)$ as $\mathsf{OPT}(S)$, or simply OPT when S is clear from the context. We call a ranking $\tilde{\pi}$ a *c-approximate center* (for some $c \geq 1$) for the set S iff $\texttt{Obj}(S, \tilde{\pi}) \leq c \cdot \mathsf{OPT}(S)$.

Mallows model

The Mallows model defines a distribution over rankings, where the probability of a ranking being sampled is dependent on its distance to a reference ranking.

▶ **Definition 3** (Mallows model). *A* Mallows model*, denoted by $M(\pi, \phi)$, on n elements is defined by two parameters, a ranking π and a dispersion parameter $\phi \in (0, 1)$.*

The probability that a ranking σ is sampled by the Mallows model is

$$\Pr[\sigma \mid \pi, \phi] = \frac{1}{Z(\pi, \phi)} \phi^{d(\pi, \sigma)},$$

where $Z(\pi, \phi) = \sum_{\rho \in S_n} \phi^{d(\pi, \rho)}$ is the normalization constant.

In the literature, sometimes the sampling probability is defined as $\frac{1}{Z(\pi,\phi)} \exp(-\phi \cdot d(\pi, \sigma))$ while allowing any $\phi > 0$. However, since both the definitions are equivalent, throughout this paper, we follow Definition 3. Observe, as $\phi \to 1$, the distribution induced by the Mallows model tends towards the uniform distribution over rankings. Also, the normalization constant $Z(\pi, \phi)$ is, in fact, independent of the ranking π (when considering the Kendall-tau distance, which is the case in this paper).

An alternative and mathematically equivalent way of formulating the Mallows model using the Kendall-tau distance is by using a more general *Repeated Insertion Model* [18]. This definition helps to establish many interesting properties of the sampling distribution.

▶ **Definition 4** (Repeated Insertion model). *Given a ranking π and a dispersion parameter $\phi \in (0, 1)$, the* Repeated Insertion model *creates a new ranking σ by the following iterative process: First, initialize σ to be empty. Then for each $i = 1, 2, \cdots, n$, take the element at $\pi(i)$ and insert it at rank j in σ with probability $\frac{\phi^{i-j}}{1+\phi+\cdots+\phi^{i-1}}$.*

3 Approximating a Center

In this section, we show that the hidden ranking π of the Mallows model is a "good" approximation to an (optimal) center.

▶ **Theorem 5.** *Consider a $\phi \in (0, 1)$ and a ranking $\pi \in S_n$. Let S be a set of rankings sampled from the Mallows model $M(\pi, \phi)$, of size s. Then*

$$\texttt{Obj}(S, \pi) \leq \mathsf{OPT}(S) + \frac{n}{1-\phi} e^{-\frac{s(1-\phi)^2}{50}} + o(n)$$

with probability at least $9/10 - 2s/n$.

We first provide a high-level overview of the proof. First, for all $\pi_r \in S$, $d(\pi_r, \pi)$ are the same up to some small factor (see Lemma 7). Next, let σ be an (arbitrary optimal) center of S. Now, we can view σ as being obtained from π by performing a sequence of swap

operations of adjacent symbols (as in bubble sort). Each such pair of symbols fall into two categories. For a particular such pair, either the majority of rankings in S agree with π on the pairwise (relative) order of that pair or the majority of rankings in S disagree with π on that pair. We argue that for all the pairs for which the majority of rankings agree, performing swaps on them cannot improve the objective value by much (roughly speaking, the decrease in the distance caused by these swaps "almost equally" spreads out over all the input rankings). On the other hand, if the majority of rankings disagree, we can improve the objective value by that particular swap; however, we show that there cannot be too many such pairs of symbols (see Lemma 9). Putting these together, we show that the objective value of σ cannot be much better than that of π.

Let us now start proving formally by first providing an upper and lower bound on the distance of any sampled ranking from π. For this section, fix

$$k = \frac{2}{\log(1/\phi)}.$$

Now, we bound the number of inversions created at each step of the Repeated Insertion Model.

▶ **Lemma 6.** *The probability that any step of the Repeated Insertion model creates at least $k \log n$ inversions compared to π is at most $\frac{1}{n}$.*

Proof. Recall that from the definition of the Repeated Insertion model (Definition 4), the probability that m inversions are created at the i-th step (for any $i \in [n]$) is equal to $\frac{\phi^m}{1+\phi+\ldots+\phi^{i-1}}$.

Therefore, the probability that at least $k \log n$ inversions are created is

$$\begin{aligned}
\Pr[\text{At least } k\log n \text{ inversions are created}] &= \frac{\phi^{k\log n} + \phi^{k\log n+1} + \ldots + \phi^{i-1}}{1+\phi+\ldots+\phi^{i-1}} \\
&= \phi^{k\log n}\frac{1+\phi+\ldots+\phi^{i-1-k\log n}}{1+\phi+\ldots+\phi^{i-1}} \\
&\leq \phi^{k\log n} = \frac{1}{n^2}
\end{aligned}$$

where the last equality follows since we set $k = \frac{2}{\log(1/\phi)}$. Taking a union bound over n steps gives the desired probability bound. ◀

Now, we prove a concentration bound on the distance of any sampled ranking from π, which we show by first bounding the expected number of inversions created by any step of the Repeated Insertion model.

▶ **Lemma 7.** *Consider a π_r drawn from $M(\pi, \phi)$. With probability at least $1 - \frac{2}{n}$,*

$$(1-\lambda)\frac{\phi}{1-\phi}n \leq d(\pi, \pi_r) \leq (1+\lambda)\frac{\phi}{1-\phi}n,$$

for some $\lambda = o(1)$.

Proof. We first consider the inversions caused by any step i of the Repeated Insertion model, where $j > k \log n$. Let the number of inversions at step j (where $j > k \log n$) in the model be represented by a random variable X_j. Let $X = \sum_{j=k\log n+1}^{n} X_j$. Let us now calculate $\mathbb{E}[X_j]$ for each $j > k \log n$.

By the definition of the Repeated Insertion model, the probability that at step j we have m inversions created is equal to $\frac{\phi^m}{1+\phi+...+\phi^{j-1}}$. Therefore the expected number of inversions caused by $\pi(j)$ with symbols $\pi(i)$ for all $i < j$ is equal to

$$\frac{1}{1+\phi+...+\phi^j}\sum_{i=0}^{j-1} i\phi^i = \frac{\phi L}{(1-\phi)L + j\phi^{j-1}}, \tag{1}$$

where $L = 1 + 2\phi + 3\phi^2 + ... + (j-1)\phi^{j-2}$.

Then by the linearity of expectation,

$$\mathbb{E}[X] = \sum_{j=k\log n+1}^{n} \mathbb{E}[X_j] = \left(\frac{\phi}{1-\phi} - \delta\right) n$$

where $\delta = O(1/n^2)$. The last equality follows since for $j > k\log n$ and $k = \frac{2}{\log(1/\phi)}$, $j\phi^{j-1} = O\left(\frac{1}{n^2}\right)$.

Let us now consider the good event $\mathcal{G}_1$ that none of the steps of the Repeated Insertion model creates more than $k\log n$ inversions. By Lemma 6, $\Pr[\bar{\mathcal{G}_1}] \leq 1/n$. From now on, we assume that the event $\mathcal{G}_1$ happens. Then the random variables X_j's take on values only in the range $[0, k\log n]$. Thus, by the standard Hoeffding bound,

$$\Pr\left[|X - \mathbb{E}[X]| \geq k\sqrt{n(\log n)^3}\right] \leq 2\exp\left(-2\log n\right).$$

That means given the event $\mathcal{G}_1$,

$$\Pr\left[X \in (1 \pm \lambda')\left(\frac{\phi}{1-\phi} - \delta\right) n\right] \geq 1 - \frac{2}{n^2}$$

where $\lambda' = o(1)$.

It is straightforward to see that the first $k\log n$ steps can create at most $k^2(\log n)^2$ inversions. Recall that $d(\pi, \pi_r)$ is equal to the total number of inversions created by all the steps of the Repeated Insertion Model while sampling π_r. Therefore,

$$(1-\lambda')\left(\frac{\phi}{1-\phi} - \delta\right) n \leq d(\pi, \pi_r) \leq (1+\lambda')\left(\frac{\phi}{1-\phi} - \delta\right) n + k^2(\log n)^2$$

with probability at least $1 - 2/n$ (since $\Pr[\bar{\mathcal{G}_1}] \leq 1/n$). Since $\delta = O(\frac{1}{n^2})$, the lemma follows. ◀

Next, we consider the set S of rankings sampled from the Mallows model $M(\pi, \phi)$. We show that there are not too many pairs of symbols (a, b) such that $a \prec_\pi b$, but the majority of rankings $\pi_r \in S$ have $b \prec_{\pi_r} a$. Let us start with a known result on the probability that a given pair is inverted in a sampled ranking compared to π.

▶ **Lemma 8** ([14]). *Let us consider a $j \in [1, n]$ and an $i < j$. The probability that a ranking π_r drawn from $M(\pi, \phi)$ has $\pi(j) \prec_{\pi_r} \pi(i)$ is equal to $\frac{\phi^k(\phi^{k+1} - (k+1)\phi + k)}{(1-\phi^{k+1})(1-\phi^k)}$, where $k = j - i$.*

For a set S drawn from $M(\pi, \phi)$, let us now consider the set P_S of pairs (a, b) that are inverted in relative order in at least $\frac{|S|}{2}$ rankings in S compared to π. More formally,

$$P_S := \{(a, b) \in [n] \times [n] \mid a \prec_\pi b \text{ but } b \prec_{\pi_r} a \text{ for at least } |S|/2 \text{ many } \pi_r \in S\}.$$

We show that the size of P_S would not be "too large" with a probability at least 9/10.

▶ **Lemma 9.** *Consider $\phi \in (0,1)$ and a set S drawn from $M(\pi,\phi)$, of size s. Then with a probability at least 9/10,*

$$|P_S| \leq \frac{n}{1-\phi}\exp\left(-\frac{s(1-\phi)^2}{50}\right).$$

Proof. Let H be a random variable whose value is equal to $|P_S|$.

First, we fix a specific value of $j \in [1,n]$. Let us consider an $i \in [0, j-1]$ and let $k = j - i$. Let I_k be a random variable whose value is equal to the number of rankings $\pi_r \in S$ such that $\pi(j) \prec_{\pi_r} \pi(i)$. Then it follows from Lemma 8 that

$$\begin{aligned}
\mu := \mathbb{E}[I_k] &= s \cdot \frac{\phi^k(\phi^{k+1} - (k+1)\phi + k)}{(1-\phi^{k+1})(1-\phi^k)} \\
&= s \cdot \phi^k \cdot \frac{k(1-\phi) - \phi(1-\phi^k)}{(1-\phi^{k+1})(1-\phi^k)} \\
&\leq s \cdot \phi^k \cdot \frac{k-\phi}{(1-\phi^{k+1})} \\
&\leq \frac{sk\phi^k}{1-\phi}.
\end{aligned}$$

Applying the standard Chernoff bound,

$$\begin{aligned}
\Pr\left[I_k \geq \frac{s}{2}\right] \leq 2\exp\left(\frac{-(\frac{s}{2\mu}-1)^2\mu}{3}\right) &\leq 2\exp\left(\frac{-(\frac{s}{3\mu})^2\mu}{3}\right) \\
&\leq 2\exp\left(-\frac{s^2}{27\mu}\right) \\
&\leq 2\exp\left(-\frac{s}{27}\cdot\frac{1-\phi}{k\phi^k}\right) \qquad \left(\text{as } \mu \leq \frac{sk\phi^k}{1-\phi}\right).
\end{aligned}$$

Let J_k be a random variable that takes on 1 if $I_k \geq \frac{s}{2}$; and 0 otherwise. Let $H_j = \sum_{k=1}^{j-1} J_k$. Then by the linearity of expectation,

$$\mathbb{E}[H_j] \leq \sum_{k=1}^{j-1} 2\exp\left(-\frac{s}{27}\cdot\frac{1-\phi}{k\phi^k}\right) = \sum_{k=1}^{j-1} 2\exp\left(-\frac{s(1-\phi)}{27}\right)^{\frac{1}{k}\phi^{-k}}. \tag{2}$$

Let $\theta := 1 - \phi$. We have that for any k,

$$\frac{1}{k}\phi^{-k} = \frac{1}{k(1-\theta)^k} \geq \frac{1}{k}(1+k\theta) \geq \theta. \tag{3}$$

Let $K = \frac{4}{\theta}\log(4/\theta)$. Then, it is not hard to see that

$$\text{for all } k \geq K, \quad \frac{1}{k}\phi^{-k} \geq k. \tag{4}$$

For brevity, let $t := \exp\left(-\frac{s(1-\phi)}{27}\right)$. Then we rewrite Equation 2 as follows

$$\begin{aligned}
\mathbb{E}[H_j] &\leq \sum_{k=1}^{K-1} 2t^{\frac{1}{k}\phi^{-k}} + \sum_{k=K}^{j-1} 2t^{\frac{1}{k}\phi^{-k}} \\
&\leq \sum_{k=1}^{K-1} 2t^{\theta} + \sum_{k=K}^{j-1} 2t^{k} && \text{(by Equation 3 and Equation 4)} \\
&\leq 2Kt^{\theta} + 2t^{K}\frac{1}{1-t} \\
&\leq 2K\exp\left(-\frac{\theta^2 s}{27}\right) + 2\exp\left(-\frac{\theta s K}{27}\right)\frac{1}{1-\exp\left(-\frac{s\theta}{27}\right)} \\
&\leq \frac{12}{\theta}\log\frac{4}{\theta}\exp\left(-\frac{\theta^2 s}{27}\right) && \left(\text{by setting } K = \frac{4}{\theta}\log(4/\theta)\right) \\
&\leq \frac{1}{\theta}\exp\left(-\frac{\theta^2 s}{40}\right).
\end{aligned}$$

Since $H = \sum_j H_j$, by the linearity of expectation, we have

$$\mathbb{E}[H] \leq \frac{n}{\theta}\exp\left(-\theta^2 s/40\right).$$

Then by Markov's inequality,

$$\Pr\left[H \geq \frac{10n}{\theta}\exp\left(-\theta^2 s/40\right)\right] \leq 1/10$$

and the lemma follows. ◀

We are now ready to prove Theorem 5.

Proof of Theorem 5. Let σ be an (arbitrary optimal) center of the set S (where $|S| = s$). Let us consider a sequence of swaps (or inversion) of adjacent pair(s) of symbols to convert the ranking π into the ranking σ. (For example, while performing bubble sort on π with σ as the total order of the symbols, we get a sequence of swaps of consecutive symbols.) Further, let this sequence of swaps converts $\pi \to \pi^1 \to \pi^2 \to \cdots \to \sigma$. Next, we categorize each such pair of symbols into two cases.

The first case is when the pair of symbols (a, b) is inverted in relative order in at most half the rankings of S compared to π (i.e., $(a, b) \notin P_S$). Suppose the swap of adjacent pair (a, b) is performed in the intermediate step while converting $\pi^j \to \pi^{j+1}$ (for some j). Observe, for at most $s/2$ rankings $\pi_r \in S$, $d(\pi^{j+1}, \pi_r) = d(\pi^j, \pi_r) - 1$, whereas for all other rankings $\pi_f \in S$, $d(\pi^{j+1}, \pi_f) = d(\pi^j, \pi_f) + 1$. Hence

$$\sum_{\pi_i \in S} d(\pi^{j+1}, \pi_i) \geq \sum_{\pi_i \in S} d(\pi^j, \pi_i). \tag{5}$$

The second case is when the pair of symbols (a, b) is inverted in relative order in more than half the rankings of S compared to π (i.e., $(a, b) \in P_S$). Again, suppose the swap of adjacent pair (a, b) is performed in the intermediate step while converting $\pi^j \to \pi^{j+1}$ (for some j). Since in the worst case, all the rankings in S have a distance of one less to π^{j+1} than to π^j,

$$\sum_{\pi_i \in S} d(\pi^{j+1}, \pi_i) \geq \sum_{\pi_i \in S} d(\pi^j, \pi_i) - s. \tag{6}$$

Let us now assume that the events of Lemma 7 and Lemma 9 occur. More specifically, let us consider the good event $\mathcal{G}$ that

- For all $\pi_r \in S$, $(1-\lambda)\frac{\phi}{1-\phi}n \le d(\pi,\pi_r) \le (1+\lambda)\frac{\phi}{1-\phi}n$, for some $\lambda = o(1)$, and
- $|P_S| \le \frac{n}{1-\phi}\exp\left(-\frac{s(1-\phi)^2}{50}\right)$.

By a union bound, $\Pr[\bar{\mathcal{G}}] \le 1/10 + 2s/n$. From now on, let us assume that $\mathcal{G}$ occurs.

In the process to transforming $\pi \to \pi^1 \to \pi^2 \to \cdots \to \sigma$, by repeatedly applying Equation 5 and Equation 6, we get that

$$d(\pi_i,\sigma) \ge \left(\sum_{\pi_i \in S} d(\pi_i,\pi) - \frac{sn}{1-\phi}e^{\frac{-s(1-\phi)^2}{50}}\right). \tag{7}$$

Hence, we derive that

$$\begin{aligned}
\mathsf{OPT}(S) = \mathtt{Obj}(S,\sigma) &\ge \frac{1}{s}\sum_{\pi_i\in S} d(\pi_i,\sigma) && \text{(by averaging)}\\
&\ge \frac{1}{s}\left(\sum_{\pi_i\in S} d(\pi_i,\pi) - \frac{sn}{1-\phi}e^{\frac{-s(1-\phi)^2}{50}}\right) && \text{(by Equation 7)}\\
&\ge \frac{1}{s}\sum_{\pi_i\in S} d(\pi_i,\pi) - \frac{n}{1-\phi}e^{\frac{-s(1-\phi)^2}{50}}\\
&\ge \min_{\pi_i\in S} d(\pi_i,S) - \frac{n}{1-\phi}e^{\frac{-s(1-\phi)^2}{50}}\\
&\ge (1-\lambda)\frac{\phi}{1-\phi}n - \frac{n}{1-\phi}e^{\frac{-s(1-\phi)^2}{50}} && \text{(assuming the event } \mathcal{G}).
\end{aligned}$$

Recall that assuming the event $\mathcal{G}$, we have that $\mathtt{Obj}(S,\pi) \le (1+\lambda)\frac{\phi}{1-\phi}n$. Therefore,

$$\begin{aligned}
\mathtt{Obj}(S,\pi) &\le \mathsf{OPT}(S) + 2\lambda\frac{\phi}{1-\phi}n + \frac{n}{1-\phi}e^{\frac{-s(1-\phi)^2}{50}}\\
&\le \mathsf{OPT}(S) + \frac{n}{1-\phi}e^{\frac{-s(1-\phi)^2}{50}} + o(n) && (\text{as } \lambda = o(1))
\end{aligned}$$

which concludes the proof. ◀

4 Reconstructing the Hidden Ranking

In this section, we discuss an algorithm that, given samples from a Mallows model, approximately reconstructs the hidden ranking/permutation π in polynomial time.

▶ **Theorem 10.** *Consider $\phi \in (0,1)$, $\alpha > 0$ and a $\pi \in S_n$. There exists a polynomial-time algorithm that, given a set S of size $s \ge \frac{50}{(1-\phi)^2}\left(\log\frac{3}{\alpha} + \log\frac{1}{1-\phi}\right)$ drawn from $M(\pi,\phi)$, computes a ranking π_c such that $d(\pi_c,\pi) \le \alpha n$ with probability at least 9/10.*

It is important to note that the number of samples required in the above theorem is only a constant (that depends on the dispersion parameter ϕ and the closeness parameter α). It is worth mentioning that if the size of the input set is $\Omega(\log n)$, then it is easy to recover the hidden ranking exactly. We now prove Theorem 10.

Description of the algorithm

We start by describing the algorithm. We first build a graph G on the vertex set $[n]$. Add a directed edge from vertex a to vertex b if at least half the rankings in S rank symbol a before b, i.e., for at least $s/2$ samples $\pi_r \in S$, $a \prec_{\pi_r} b$. Observe the graph obtained is a *tournament*

Algorithm 1 ApproxReconstruct.

Input: $S \subset S_n$ of size s
Output: A ranking π_c over $[n]$
$G \leftarrow ([n], E)$ where $E = \{(a, b) \mid a \text{ is ranked before } b \text{ in at least } \frac{s}{2} \text{ rankings}\}$.
Run the PTAS of [28] to get a subset F of edges.
G' is obtained by removing the edge set F from G.
$\pi_c \leftarrow$ topological ordering of G'.
return π_c.

graph[2]. However, it may contain a cycle. Next, we want to remove as few edges as possible from G to make it acyclic, which is the well-known *feedback arc set* problem. Kenyon-Mathieu and Schudy [28, 42] designed a PTAS for the feedback arc set problem over tournament graphs. We run that PTAS to obtain a $(1+\varepsilon)$-approximation of the optimal feedback arc set and remove those edges to form a directed acyclic graph G'. Finally, we obtain a topological order on the vertices of G' and return this as ranking π_c. (See Algorithm 1.)

Approximation guarantee on the reconstructed ranking

We call an edge (b, a) in G *bad* if $a \prec_\pi b$. Let B be the set of all the bad edges. Recall, we define P_S to be the set of pairs (a, b) that are inverted in relative order in at least $\frac{|S|}{2}$ rankings in S compared to π. By the construction of the graph G, the pairs of symbols in P_S essentially constitute the bad edge set B. Thus $|B| = |P_S|$.

So by Lemma 9 we have that the size of B is at most $\frac{n}{1-\phi} \exp\left(-\frac{s(1-\phi)^2}{50}\right)$ with probability at least 9/10. For $s \geq \frac{50}{(1-\phi)^2}\left(\log\frac{3}{\alpha} + \log\frac{1}{1-\phi}\right)$, we have that

$$\begin{aligned} |B| &\leq \frac{n}{1-\phi} \exp\left(-\log\frac{3}{\alpha} - \log\frac{1}{1-\phi}\right) \\ &\leq \frac{\alpha}{3} n. \end{aligned} \tag{8}$$

It is straightforward to observe that removing the set B of bad edges from G makes it acyclic. So $|B|$ (which is at most $\alpha n/3$ by Equation 8) serves as an upper bound on the size of an optimal feedback arc set of G. Since by applying the PTAS of [28], we get the subset F of edges to remove from G to make it acyclic, $|F| \leq (1+\varepsilon)\alpha n/3$, for any $\varepsilon > 0$.

Since π_c is the resulting ranking from taking a topological ordering of the resulting directed acyclic graph $G' = G - F$,

$$d(\pi_c, \pi) \leq |B| + |F| \leq (2+\varepsilon) \cdot \frac{\alpha n}{3}$$

which completes the proof of Theorem 10.

▶ Remark 11. Note the running time of constructing the graph and finding a topological ordering is $O(sn^2)$. However, since for the reconstruction, $s = \mathsf{poly}(\log(1/\alpha), 1/(1-\phi))$ samples suffice, the running time is $O(n^2)$, where $O(\cdot)$ hides the constant term that depends (polynomially) on $1/\alpha$ and $1/(1-\phi)$. This is dominated by the running time of the PTAS for the feedback arc set on tournaments from [28, 42], which is $O\left(\frac{n^3 \log n}{\varepsilon}\right) + n2^{O(\varepsilon^{-6})}$, where ε comes from the $(1+\varepsilon)$-factor of PTAS. We can significantly improve the running time of

[2] A directed graph is called a *tournament* if there is a (directed) edge between every pair of vertices.

this algorithm by increasing the input size required by a constant factor, and then instead of running the PTAS algorithm, we can run a randomized algorithm by [1] that finds a 3-approximation of the optimal feedback arc set and runs in only $O(n \log n)$ time.

Exact reconstruction of the hidden ranking

In Theorem 10, we provide an approximate reconstruction algorithm using only constantly many samples. It is worth mentioning that a polynomial time exact reconstruction algorithm is already known due to [14]; however, that requires $\Omega(\log n)$ samples. We also want to mention that as long as the sampled set is of size at least $\frac{12(1+\phi)^2}{(1-\phi)^2} \log n$, for every pair of symbols (a, b), their relative ordering in at least half of the samples will be exactly the same as that in π with high probability (with probability at least $1 - 1/\mathsf{poly}(n)$). Hence, by using any standard comparison-based sorting algorithm, we can reconstruct π exactly in time $O(n \log^2 n)$. Although this simple approach is quite standard and well-known, just for the sake of completeness, we provide detailed proof of this simple argument in Appendix B.

5 Completing the proof of Theorem 1

We first show a lower bound of $\mathsf{OPT}(S)$ for any set S drawn from a Mallows model.

▶ **Lemma 12.** *Consider a set S (of size at least two) drawn from $M(\pi, \phi)$. Then, with probability at least $1 - \frac{6}{n}$, $\mathsf{OPT}(S) \geq \phi n - o(n)$.*

Proof. Let $\pi_1, \pi_2 \in S$. Let I be the set of pairs that are inverted in relative order in both π_1 and π_2 compared to π.

First, take some fixed $j \in [1, n]$ and consider the step j in the Repeated Insertion model. Let B_j be a random variable that takes on a value equal to the size of the set

$$\{i \mid i < j, \pi(j) \prec_{\pi_1} \pi(i), \pi(j) \prec_{\pi_2} \pi(i)\}.$$

Let C_j be a random variable that takes on a value t if

- $\pi(j)$ is placed at rank $j - t$ in π_1 (i.e., creating t inversions), and
- $\pi(j)$ is placed at some rank $j - r$ for $t \leq r \leq j - 1$ in π_2 (i.e., creating at least t inversions).

Observe that if $|\{i \mid i < j, \pi(j) \prec_{\pi_1} \pi(i), \pi(j) \prec_{\pi_2} \pi(i)\}| = t$, then the insertion of the symbol at rank j must have created at least t inversions in both π_1 and π_2. Therefore for any $t > 0$, $\Pr[B_j \geq t] \leq \Pr[C_j \geq t]$.

Now, let $B = \sum_{j=1}^{n} B_j$ and $C = \sum_{j=1}^{n} C_j$. Observe that the size of the set I is equal to B. Recall that each step of the Repeated Insertion model is independent of each other.

Consider $\Pr[C_j = m]$ for some $j \in [1, n]$ and $m \in [0, j-1]$. As sampling both rankings is independent, by Definition 4, we get that

$$\Pr[C_j = m] = \frac{\phi^m}{1 + \phi + \ldots + \phi^{j-1}} \cdot \frac{\phi^m + \ldots + \phi^{j-1}}{1 + \phi + \ldots + \phi^{j-1}}.$$

Thus, we deduce that

$$\begin{aligned}
\mathbb{E}[C_j] &= \sum_{m=1}^{j-1} m \cdot \frac{\phi^m}{1 + \phi + \ldots + \phi^{j-1}} \cdot \frac{\phi^m + \ldots + \phi^{j-1}}{1 + \phi + \ldots + \phi^{j-1}} \\
&\leq \sum_{m=1}^{j-1} m \cdot \phi^m \frac{\phi^m}{1 + \phi + \ldots + \phi^{j-1}}
\end{aligned}$$

$$\leq \frac{\phi}{1+\phi+...+\phi^{j-1}} \sum_{m=1}^{j-1} m\phi^m$$
$$\leq \phi \cdot \frac{\phi + 2\phi^2 ... + (j-1)\phi^{j-1}}{1+\phi+...+\phi^{j-1}}$$
$$\leq \phi \cdot \frac{\phi(1+2\phi+...+(j-1)\phi^{j-2})}{(1-\phi)(1+2\phi+...+(j-1)\phi^{j-2}) + i\phi^{j-1}}$$
$$\leq \frac{\phi^2}{1-\phi}.$$

By the linearity of expectation, $\mathbb{E}[C] \leq \frac{\phi^2}{1-\phi} n$.

Let us now assume that the events of Lemma 6 and Lemma 7 occur. More specifically, let us consider the good event $\mathcal{G}_2$ that

- None of the steps of the Repeated Insertion model creates more than $k \log n$ inversions, and
- For $r = 1, 2$, $(1-\lambda)\frac{\phi}{1-\phi} n \leq d(\pi, \pi_r) \leq (1+\lambda)\frac{\phi}{1-\phi} n$, for some $\lambda = o(1)$.

By a union bound of the failure probability of Lemma 6 and Lemma 7, $\Pr[\bar{\mathcal{G}_2}] \leq 5/n$. From now on, let us assume that $\mathcal{G}_2$ occurs.

So the random variables C_j's take on values in the range $[0, k \log n]$. Thus, by applying the standard Hoeffding bound,

$$\Pr\left[C - \mathbb{E}[C] \geq k\sqrt{n}(\log n)^{3/2}\right] \leq \exp(-2\log n). \tag{9}$$

So we have that given the event $\mathcal{G}_2$ occurs, with a probability at least $1 - 1/n^2$,

$$|I| \leq (1+\gamma)\frac{\phi^2}{1-\phi} n, \text{ for some } \gamma = o(1). \tag{10}$$

Let $\pi^*_{1,2}$ be an (arbitrary optimal) center of π_1 and π_2. Then by the triangle inequality,

$$d(\pi_1, \pi^*_{1,2}) \geq \frac{1}{2} d(\pi_1, \pi_2).$$

Also, by definition, $d(\pi_1, \pi_2) = d(\pi_1, \pi) + d(\pi_2, \pi) - 2|I|$. Further, observe that $\mathsf{OPT}(S) \geq d(\pi_1, \pi^*_{1,2})$ (as each of the other rankings in S can only increase the objective value).

Therefore,

$$\begin{aligned}
\mathsf{OPT}(S) &\geq \frac{d(\pi_1, \pi) + d(\pi_2, \pi) - 2|I|}{2} \\
&\geq (1-\lambda)\frac{\phi}{1-\phi} n - |I| && \text{(assuming the event } \mathcal{G}_2\text{)} \\
&\geq (1-\lambda)\frac{\phi}{1-\phi} n - (1+\gamma)\frac{\phi^2}{(1-\phi)} n && \text{(by Equation 10)} \\
&\geq \frac{\phi}{1-\phi} n(1 - \lambda - (1+\gamma)\phi) && (11) \\
&\geq \phi n - o(n) && \text{(as } \lambda \text{ and } \gamma \text{ are } o(1)\text{).} \quad (12)
\end{aligned}$$

The success probability follows from the failure probability of Equation 10 and $\Pr[\bar{\mathcal{G}_2}] \leq 5/n$. ◀

Proof of Theorem 1. Set $\alpha = \frac{\varepsilon\phi}{3}$. Recall that by Theorem 10, we find a permutation π_c such that $d(\pi_c, \pi) \leq \alpha n$. Next, recall that from Theorem 5 we have that given s samples,

$$\begin{aligned}\texttt{Obj}(S,\pi) &\leq \mathsf{OPT}(S) + \frac{n}{1-\phi} e^{-\frac{s(1-\phi)^2}{50}} + o(n) \\ &\leq \mathsf{OPT}(S) + \frac{\alpha}{3} n + o(n) \end{aligned} \tag{13}$$

where the last inequality follows for $s \geq \frac{50}{(1-\phi)^2}\left(\log\frac{3}{\alpha} + \log\frac{1}{1-\phi}\right)$.

So, finally, we get that

$$\begin{aligned}\texttt{Obj}(S,\pi_c) &\leq \texttt{Obj}(S,\pi) + d(\pi,\pi_c) && \text{(by the triangle inequality)} \\ &\leq \mathsf{OPT}(S) + \frac{4}{3}\alpha n + o(n) && \text{(by Equation 13)} \\ &\leq \mathsf{OPT}(S) + \frac{\varepsilon\phi}{2} n && \text{(for } \alpha = \varepsilon\phi/3\text{)} \\ &\leq \mathsf{OPT}(S) + \varepsilon \cdot \mathsf{OPT}(S) && \text{(by Lemma 12)} \\ &\leq (1+\varepsilon)\mathsf{OPT}(S). \end{aligned}$$

Since the failure probabilities of Theorem 5, Theorem 10 and Lemma 12 are $1/10 + 2s/n$, $1/10$ and $6/n$ respectively, by a union bound, we have that all three simultaneously hold with probability at least $2/3$, for large enough n (since we can assume that $s = \frac{50}{(1-\phi)^2}\left(\log\frac{3}{\alpha} + \log\frac{1}{1-\phi}\right)$).

It follows from the discussion in Remark 11 that the running time of our algorithm is $O(n^3 \log n)$, where the $O(\cdot)$ notation hides a constant that depends on $1/\varepsilon$ and $1/(1-\phi)$. The running time can further be reduced to $O(n \log^2 n)$ when the size of the input set is at least $\Omega(\log n)$. ◀

6 Extension to Input with Outliers

Our algorithm can be extended to the scenario when a small number of *outliers* is also present in the input set. In the context of the clustering (or center) problem, outliers are the data points that are not part of the intended cluster. In the case of the probabilistic generating model, outliers can be considered a set of adversarially chosen points (that do not come from the generating model). Formally, we consider the following *Strong Contamination Model*. In this model, a set of s samples are first generated from a Mallows model $M(\pi, \phi)$ (for some $\pi \in S_n$ and $\phi \in (0,1)$). Then, an arbitrary subset of size at most δs (for some $\delta \in [0,1)$) is replaced with an arbitrary set of rankings in S_n. We call this resulting set (π, ϕ, δ)-*corrupted sample*. This model was previously considered in the literature (e.g., see [32]).

Given a set S of rankings, the δ-*maximum rank aggregation with outliers* problem seeks to find a ranking that minimizes the maximum distance to a subset $S' \subseteq S$ of size at least $(1-\delta)|S|$. We show that we can find a $(1+\varepsilon)$-approximate δ -maximum rank aggregation with outliers for (π, ϕ, δ)-corrupted sample sets.

▶ **Theorem 13.** *Consider a $\phi \in (0,1)$, $\varepsilon > 0$, $\delta \in (0, 0.5)$ and a ranking π on n candidates. There is a polynomial-time algorithm that, given a (π, ϕ, δ)-corrupted sample set S of size at least* $\mathsf{poly}\left(1/\varepsilon, 1/(1-\phi), 1/(0.5-\delta)\right) \log n$*, outputs a ranking π_c that is a $(1+\varepsilon)$-approximation to the δ-maximum rank aggregation with outliers problem with probability at least $2/3$.*

It is worth noting that the running time mentioned in the above theorem is $n^{O(1)}$ where $O(\cdot)$ notation hides a constant that depends (polynomially) on $\log(1/\varepsilon)$, $1/(1-\phi)$ and $1/(0.5-\delta)$. The brief idea for this extension is to first sample $O(\log n)$ rankings from the input uniformly at random. Then, we try out all possible subsets of this sample set and apply the algorithm from Theorem 10 on each of these subsets. We treat all of these computed rankings as potential centers. Finally, output the best one (that minimizes the objective value) among all these potential centers. We argue that the output will be a $(1+\varepsilon)$-approximation of the optimal solution. The key (though simple) observation is that it suffices to use only $O(\log n)$ samples in Theorem 1, and also, in polynomial time, we can try out all possible subsets. We defer the details to Appendix A.

7 Conclusion

This paper focuses on the maximum rank aggregation, also known as the center problem, concerning the Kendall-tau metric on inputs generated from the well-established Mallows model. Although a folklore algorithm can compute a 2-approximate center in any metric space, no better approximation guarantee is known for the Kendall-tau metric despite its significant importance in various applications. Therefore, finding an $(1+\varepsilon)$-approximation algorithm for this problem remains an essential open problem. This paper proposes a data-driven algorithm that provides a $(1+\varepsilon)$-approximation algorithm for any $\varepsilon > 0$, given that the input set is sampled from a Mallows model with dispersion parameter $\phi \in (0,1)$, and the number of input rankings is at least $\mathsf{poly}(\log(1/\varepsilon), 1/(1-\phi))$. The proposed algorithm can also be extended to datasets containing a few adversarially chosen outliers.

An apparent open direction is to obtain similar results without any assumptions about the input set. Even breaking below the 2-factor in the worst-case would be an exciting direction. Another interesting open problem is to show whether the number of samples required to achieve $(1+\varepsilon)$-approximation in Theorem 1 is tight. Also, developing a similar result for other probabilistic generating models for rankings, as in [29], would be interesting.

References

1. Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)*, 55(5):1–27, 2008.
2. Pranjal Awasthi, Avrim Blum, Or Sheffet, and Aravindan Vijayaraghavan. Learning mixtures of ranking models. *Advances in Neural Information Processing Systems*, 27, 2014.
3. Christian Bachmaier, Franz J. Brandenburg, Andreas Gleißner, and Andreas Hofmeier. On the hardness of maximum rank aggregation problems. *Journal of Discrete Algorithms*, 31:2–13, 2015. 24th International Workshop on Combinatorial Algorithms (IWOCA 2013).
4. Mihai Bādoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *Proceedings of the thiry-fourth annual ACM Symposium on Theory of Computing*, pages 250–257, 2002.
5. Therese Biedl, Franz J Brandenburg, and Xiaotie Deng. On the complexity of crossings in permutations. *Discrete Mathematics*, 309(7):1813–1823, 2009.
6. Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.
7. Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. *Handbook of computational social choice*. Cambridge University Press, 2016.
8. Mark Braverman and Elchanan Mossel. Sorting from noisy information. *arXiv preprint*, 2009. `arXiv:0910.1191`.

9 Marc Bury and Chris Schwiegelshohn. On finding the jaccard center. In *44th International Colloquium on Automata, Languages, and Programming (ICALP 2017)*. Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, 2017.

10 Ioannis Caragiannis, Ariel D Procaccia, and Nisarg Shah. When do noisy votes reveal the truth? *ACM Transactions on Economics and Computation (TEAC)*, 4(3):1–30, 2016.

11 Diptarka Chakraborty, Debarati Das, and Robert Krauthgamer. Approximating the median under the ulam metric. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 761–775. SIAM, 2021.

12 Diptarka Chakraborty, Debarati Das, and Robert Krauthgamer. Clustering permutations: New techniques with streaming applications. In Yael Tauman Kalai, editor, *14th Innovations in Theoretical Computer Science Conference, ITCS 2023, January 10-13, 2023, MIT, Cambridge, Massachusetts, USA*, volume 251 of *LIPIcs*, pages 31:1–31:24. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023.

13 Diptarka Chakraborty, Kshitij Gajjar, and Agastya Vibhuti Jha. Approximating the center ranking under ulam. In *41st IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2021)*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021.

14 Flavio Chierichetti, Anirban Dasgupta, Ravi Kumar, and Silvio Lattanzi. On reconstructing a hidden permutation. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*. Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, 2014. doi:10.4230/LIPIcs.APPROX-RANDOM.2014.604.

15 Fabien Collas and Ekhine Irurozki. Concentric mixtures of mallows models for top-k rankings: sampling and identifiability. In *International Conference on Machine Learning*, pages 2079–2088. PMLR, 2021.

16 Anindya De, Ryan O'Donnell, and Rocco Servedio. Learning sparse mixtures of rankings from noisy information. *arXiv preprint*, 2018. arXiv:1811.01216.

17 Persi Diaconis and R. L. Graham. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(2):262–268, 1977. URL: http://www.jstor.org/stable/2984804.

18 Jean-Paul Doignon, Aleksandar Pekeč, and Michel Regenwetter. The repeated insertion model for rankings: Missing link between two subset choice models. *Psychometrika*, 69(1):33–54, 2004.

19 Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the Tenth International World Wide Web Conference, WWW 10*, pages 613–622, 2001. doi:10.1145/371920.372165.

20 Ronald Fagin, Ravi Kumar, and D. Sivakumar. Efficient similarity search and classification via rank aggregation. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, San Diego, California, USA, June 9-12, 2003*, pages 301–312, 2003.

21 Moti Frances and Ami Litman. On covering problems of codes. *Theory of Computing Systems*, 30(2):113–119, 1997.

22 David F Gleich and Lek-heng Lim. Rank aggregation via nuclear norm minimization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 60–68, 2011.

23 Donna Harman. Ranking algorithms. In William B. Frakes and Ricardo A. Baeza-Yates, editors, *Information Retrieval: Data Structures & Algorithms*, pages 363–392. Prentice-Hall, 1992.

24 Kenneth Hung and William Fithian. Rank verification for exponential families. *The Annals of Statistics*, 47(2):758–782, 2019.

25 David R Hunter. Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32(1):384–406, 2004.

26 John G Kemeny. Mathematics without numbers. *Daedalus*, 88(4):577–591, 1959.

27 Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

28 Claire Kenyon-Mathieu and Warren Schudy. How to rank with few errors. In *Proceedings of the thirty-ninth annual ACM Symposium on Theory of Computing*, pages 95–103, 2007.

29 Ashish Khetan and Sewoong Oh. Data-driven rank breaking for efficient rank aggregation. In *International Conference on Machine Learning*, pages 89–98. PMLR, 2016.

30 Ming Li, Bin Ma, and Lusheng Wang. On the closest string and substring problems. *Journal of the ACM (JACM)*, 49(2):157–171, March 2002.

31 Allen Liu and Ankur Moitra. Efficiently learning mixtures of mallows models. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 627–638. IEEE, 2018.

32 Allen Liu and Ankur Moitra. Robust voting rules from algorithmic robust statistics. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3471–3512. SIAM, 2023.

33 R Duncan Luce. Individual choice behavior, 1959.

34 Bin Ma and Xiaoming Sun. More efficient algorithms for closest string and substring problems. *SIAM Journal on Computing*, 39(4):1432–1443, 2010.

35 Colin L Mallows. Non-null ranking models. i. *Biometrika*, 44(1/2):114–130, 1957.

36 Nimrod Megiddo. Linear programming in linear time when the dimension is fixed. *Journal of the ACM (JACM)*, 31(1):114–127, 1984.

37 François Nicolas and Eric Rivals. Complexities of the centre and median string problems. In *Combinatorial Pattern Matching, 14th Annual Symposium, CPM 2003, Morelia, Michocán, Mexico, June 25-27, 2003, Proceedings*, pages 315–327, 2003.

38 Vasyl Pihur, Susmita Datta, and Somnath Datta. Weighted rank aggregation of cluster validation measures: a monte carlo cross-entropy approach. *Bioinformatics*, 23(13):1607–1615, 2007.

39 Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

40 V Yu Popov. Multiple genome rearrangement by swaps and by element duplications. *Theoretical computer science*, 385(1-3):115–126, 2007.

41 Antti-Veikko Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie Dorr. Combining outputs from multiple machine translation systems. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 228–235, 2007.

42 Warren Schudy. *Approximation schemes for inferring rankings and clusterings from pairwise data*. Ph.D. diss., Brown University, 2012.

43 Nihar B Shah and Martin J Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *The Journal of Machine Learning Research*, 18(1):7246–7283, 2017.

44 James Joseph Sylvester. A question in the geometry of situation. *Quarterly Journal of Pure and Applied Mathematics*, 1(1):79–80, 1857.

45 Wenpin Tang. Mallows ranking models: maximum likelihood estimate and regeneration. In *International Conference on Machine Learning*, pages 6125–6134. PMLR, 2019.

46 E Alper Yildirim. Two algorithms for the minimum enclosing ball problem. *SIAM Journal on Optimization*, 19(3):1368–1391, 2008.

47 H Peyton Young. Condorcet's theory of voting. *American Political science review*, 82(4):1231–1244, 1988.

48 H Peyton Young and Arthur Levenglick. A consistent extension of condorcet's election principle. *SIAM Journal on Applied Mathematics*, 35(2):285–300, 1978.

A Detailed Description of Extension to Input with Outliers

We start with formally defining the input (sample) generating model and the problem statement.

▶ **Definition 14** (Strong Contamination Model). *Consider any $\phi, \delta \in (0,1)$ and a ranking $\pi \in S_n$. A set of samples S of size s is said to be a* (π, ϕ, δ)-corrupted sample set *if it is generated as follows: First, sample s rankings from $M(\pi, \phi)$. Then, a malicious adversary observes these s rankings and replaces up to δs of them with any arbitrary rankings it chooses, and then returns the resulting set of s rankings in an arbitrary order.*

Now, let us formally define the δ-maximum rank aggregation with outliers problem. Given a set S of rankings, the *δ-maximum rank aggregation with outliers* problem seeks to find a ranking (not necessarily from S) that minimizes the maximum distance to a subset $S' \subseteq S$ of size at least $(1-\delta)|S|$.

Given a ranking ρ and set S, we define the objective value of ρ on S for the δ-maximum rank aggregation with outliers problem as

$$\texttt{Obj}_\delta(S, \rho) := \min_{S' \subseteq S : |S'| \geq (1-\delta)|S|} \max_{\pi_r \in S'} d(\pi_r, \rho).$$

Recall, $\texttt{Obj}(S', \rho) = \max_{\pi_r \in S'} d(\pi_r, \rho)$. So we can simply say

$$\texttt{Obj}_\delta(S, \rho) = \min_{S' \subseteq S : |S'| \geq (1-\delta)|S|} \texttt{Obj}(S', \rho).$$

Let us denote the optimum objective value for the δ-maximum rank aggregation with outliers problem for a set S by $\mathsf{OPT}_\delta(S)$. We call a ranking $\tilde{\pi}$ a *c-approximate δ-maximum aggregate ranking with outliers* (for some $c \geq 1$) for the set S iff $\texttt{Obj}_\delta(S, \tilde{\pi}) \leq c \cdot \mathsf{OPT}_\delta(S)$, i.e., there exists a subset $T \subseteq S$ with $|T| \geq (1-\delta)|S|$ such that $\texttt{Obj}(T, \tilde{\pi}) \leq c \cdot \mathsf{OPT}_\delta(S)$. Note when $\delta = 0$, the definition coincides with that without outliers. More specifically, we have $\texttt{Obj}_0(S, \rho) = \texttt{Obj}(S, \rho)$ and $\mathsf{OPT}_0(S) = \mathsf{OPT}(S)$.

▶ **Theorem 13.** *Consider a $\phi \in (0,1)$, $\varepsilon > 0$, $\delta \in (0, 0.5)$ and a ranking π on n candidates. There is a polynomial-time algorithm that, given a (π, ϕ, δ)-corrupted sample set S of size at least $\mathsf{poly}\left(1/\varepsilon, 1/(1-\phi), 1/(0.5-\delta)\right) \log n$, outputs a ranking π_c that is a $(1+\varepsilon)$-approximation to the δ-maximum rank aggregation with outliers problem with probability at least $2/3$.*

Before proving the above theorem, let us introduce a few notations and show a few lemmas that will help us in proving the above theorem. Let $G \subseteq S$ be the set of rankings sampled from the Mallows model $M(\pi, \phi)$.

▶ **Lemma 15.** *Consider a $\phi \in (0,1)$, $\alpha > 0$, $\delta \in (0, 0.5)$, a ranking $\pi \in S_n$. There is a polynomial-time algorithm that, given a (π, ϕ, δ)-corrupted sample set S of size at least $\mathsf{poly}(\log(1/\alpha), 1/(1-\phi), 1/(0.5-\delta)\}) \log n$, constructs a set of rankings P of size $\mathsf{poly}(n)$ such that there exists $\pi_c \in P$ satisfying $d(\pi_c, \pi) \leq \alpha n$ with probability at least 9/10.*

Proof. Let $s = |S|$. Set $p = \frac{2s' \log n}{s}$ where s' is the minimum required size of the input set in Theorem 10, multiplied with the factor $1/(0.5-\delta)^2$.

Now, sample each ranking in S independently uniformly at random with probability p, and let this sample set be T. It is easy to see that $\mathbb{E}[|T|] = ps = 2s' \log n$. By applying the standard Chernoff bound,

$$\Pr\left[|T| \geq 5s' \log n\right] \leq \exp\left(-\frac{9s' \log n}{7}\right) \leq n^{-2}.$$

So with high probability, we have that $|T| \leq 5s' \log n$. Further,

$$\mathbb{E}[|G \cap T|] \geq (1-\delta)ps = 2(1-\delta)s' \log n.$$

Again, by applying the standard Chernoff bound,

$$\Pr\left[|G \cap T| \leq s' \log n\right] \leq \exp\left(-\frac{(0.5-\delta)^2 s' \log n}{(1-\delta)}\right) \leq n^{-2}$$

where the last inequality holds since $s' \geq 2/(0.5-\delta)^2$. So we get with high probability $|G \cap T| \geq s' \log n$. Taking a union bound, we have that with a high probability,

$$|T| \leq 5s' \log n \text{ and } |G \cap T| \geq s' \log n.$$

From now on, we continue the proof by assuming the event that $|T| \leq 5s' \log n$ and $|G \cap T| \geq s' \log n$.

Next, we exhaustively try all possible subsets of the set T and run our reconstruction algorithm from Theorem 10 on each such subset, and finally return the set of all these rankings.

Among all these possible subsets, one of them contains only rankings from the set G and is of size at least $s' \log n$. Using this set, the reconstruction algorithm of Theorem 10 reconstructs a ranking π_c such that $d(\pi_c, \pi) \leq \alpha n$.

Now, we have that $s' = \frac{1}{(0.5-\delta)^2} \cdot \frac{50}{(1-\phi)^2}\left(\log \frac{3}{\alpha} + \log \frac{1}{1-\phi}\right)$. Since $|T| \leq 5s' \log n$, we will have at most $n^{5s'} = n^{\mathsf{poly}(\log(1/\alpha), 1/(1-\phi), 1/(0.5-\delta))}$ many subsets. By Remark 11, the reconstruction algorithm will require at most $O(s'n^3 \log n)$ time for each subset, and thus the lemma follows. ◀

▶ **Lemma 16.** *Consider $\phi \in (0,1)$, $\delta \in (0, 0.5)$, a ranking $\pi \in S_n$. Let S be a (π, ϕ, δ)-corrupted sample set (of size at least $2/(1-2\delta)$). Then $\mathsf{OPT}_\delta(S) \geq \phi n - o(n)$ with a high probability.*

Proof. To prove the above lemma, it suffices to show that for any $S' \subseteq S$ where $|S'| \geq (1-\delta)|S|$, $\mathsf{OPT}(S') \geq \phi n - o(n)$.

By definition, the set $S' \cap G$ only contains the rankings drawn from $M(\pi, \phi)$ and is of size at least two. Then by Lemma 12, $\mathsf{OPT}(S' \cap G) \geq \phi n - o(n)$ (with a high probability). By definition, $\mathsf{OPT}(S') \geq \mathsf{OPT}(S' \cap G)$ (since each of the rankings in $S' \setminus G$ can only increase the value of $\mathsf{OPT}(S')$). So we have $\mathsf{OPT}(S') \geq \phi n - o(n)$. ◀

We are now ready to finish the proof of Theorem 13.

Proof of Theorem 13. Given as input a (π, ϕ, δ)-corrupted sample set S, we first apply the algorithm of Lemma 15 to obtain a set of rankings P. We then simply return a ranking $\tilde{\pi} \in P$ that achieves the minimum objective value, i.e., $\tilde{\pi} := \arg\min_{\rho \in P} \mathtt{Obj}_\delta(S, \rho)$. Observe that we can compute $\mathtt{Obj}_\delta(S, \rho)$ (for each $\rho \in P$) naïvely in polynomial-time because $\mathtt{Obj}_\delta(S, \rho) = \mathtt{Obj}(S', \rho)$, where $S' \subseteq S$ is the set of $(1-\delta)|S|$ many nearest rankings (of S) to ρ.

Next, we argue that $\tilde{\pi}$ is a $(1+\varepsilon)$-approximate δ-maximum aggregate ranking with outliers, i.e., $\mathtt{Obj}_\delta(S, \tilde{\pi}) \leq (1+\varepsilon)\mathsf{OPT}_\delta(S)$. We know that there exists some set G of size at least $(1-\delta)|S|$ such that all the rankings in G are drawn from $M(\pi, \phi)$. By Lemma 7, we can assume that for every $\pi_i \in G$, $d(\pi_i, \pi) \leq (1+\lambda)\frac{\phi}{1-\phi} n$ for some $\lambda = o(1)$. Therefore we know that $\mathsf{OPT}(G) \leq (1+\lambda)\frac{\phi}{1-\phi} n$.

Now, consider any $S' \subseteq S$ such that $|S'| \geq (1-\delta)|S|$. Then we must have $|S' \cap G| \geq (1-2\delta)|S|$. Further, from Theorem 5 we have that for the set $S' \cap G$,

$$\mathsf{OPT}(S' \cap G) \geq (1-\lambda)\frac{\phi}{1-\phi}n - \frac{n}{1-\phi}e^{-\frac{(1-2\delta)|S|(1-\phi)^2}{50}}.$$

As argued in the proof of Lemma 16, $\mathsf{OPT}(S') \geq \mathsf{OPT}(S' \cap G)$. Therefore, we have that

$$\begin{aligned}\mathsf{OPT}(G) &\leq (1+\lambda)\frac{\phi}{1-\phi}n \\ &\leq \mathsf{OPT}(S') + 2\lambda\frac{\phi}{1-\phi}n + \frac{n}{1-\phi}e^{-\frac{(1-2\delta)|S|(1-\phi)^2}{50}} \\ &\leq \mathsf{OPT}(S') + \frac{n}{1-\phi}e^{-\frac{(1-2\delta)|S|(1-\phi)^2}{50}} + o(n).\end{aligned}$$

Since the above is true for all $S' \subseteq S$ of size at least $(1-\delta)|S|$, we derive that

$$\mathsf{OPT}(G) \leq \mathsf{OPT}_\delta(S) + \frac{n}{1-\phi}e^{-\frac{(1-2\delta)|S|(1-\phi)^2}{50}} + o(n).$$

By Theorem 5 (considering G as the sample set), we have

$$\begin{aligned}\mathtt{Obj}(G,\pi) &\leq \mathsf{OPT}(G) + \frac{n}{1-\phi}e^{-\frac{|G|(1-\phi)^2}{50}} + o(n) \\ &\leq \mathsf{OPT}_\delta(S) + \frac{n}{1-\phi}e^{-\frac{(1-2\delta)|S|(1-\phi)^2}{50}} + \frac{n}{1-\phi}e^{-\frac{|G|(1-\phi)^2}{50}} + o(n).\end{aligned}$$

Now, set $\alpha = \frac{\varepsilon\phi}{4}$. By Lemma 15, there exists $\pi_c \in P$ such that $d(\pi_c, \pi) \leq \alpha n$. By Lemma 16, we have that $\mathsf{OPT}_\delta(S) \geq \phi n - o(n)$. Consider $|S| \geq \frac{50}{(0.5-\delta)^2(1-\phi)^2}\left(\log\frac{3}{\alpha} + \log\frac{1}{1-\phi}\right)$. By the triangle inequality, we have

$$\begin{aligned}\mathtt{Obj}(G,\pi_c) &\leq \mathsf{OPT}_\delta(S) + 2\alpha n + o(n) \\ &\leq \mathsf{OPT}_\delta(S) + 3\alpha n \\ &\leq (1+\varepsilon)\mathsf{OPT}_\delta(S).\end{aligned}$$

Since $|G| \geq (1-\delta)|S|$, by the construction,

$$\mathtt{Obj}_\delta(S,\tilde{\pi}) \leq \mathtt{Obj}(G,\pi_c) \leq (1+\varepsilon)\mathsf{OPT}_\delta(S).$$

The running time of the algorithm is dominated by the time required to construct the set of rankings P. Therefore the algorithm requires $n^{\mathsf{poly}(\log(1/\varepsilon),1/(1-\phi),1/(0.5-\delta))}$ time. ◀

B A Simple Exact Reconstruction of the Hidden Ranking

From the previous section, we see that if the number of sampled rankings is large enough, we expect that with a large probability, taking the majority will give the correct order of most pairs of symbols in π. We now prove that exact reconstruction is possible with high probability if the sampled set is of size at least $\frac{12(1+\phi)^2}{(1-\phi)^2}\log n$. Further, we describe a simpler algorithm for the reconstruction if this condition is met.

▶ **Theorem 17.** *Consider a $\phi \in (0,1)$ and a $\pi \in S_n$. There exists a deterministic algorithm that, given a set S of size at least $\frac{12(1+\phi)^2}{(1-\phi)^2}\log n$ drawn from $M(\pi,\phi)$, computes π with probability at least $1-\frac{1}{n}$ in time $O(n\log^2 n)$.*

The algorithm constructs a new ranking using any comparison-based algorithm on $[n]$ and returns this as the hidden ranking. When comparing two symbols a and b, we set $a \prec b$ if at least half the sampled rankings rank a before b and set $a \succ b$ otherwise.

We first need to prove an upper bound on the probability that a ranking drawn from the Mallows model has a given pair of symbols inverted in relative order compared to π.

▶ **Lemma 18.** *Consider a ranking ρ drawn from a Mallows model $M(\pi, \phi)$. The probability that for a given pair of symbols (a, b) that $a \prec_\rho b$ and $b \prec_\pi a$ is at most $\frac{\phi}{1+\phi}$.*

Proof. Recall that from Lemma 8, for some $i < j$ and some sampled ranking π_r, we have $\pi(j) \prec_{\pi_r} \pi(i)$ with probability $\frac{\phi^k(\phi^{k+1}-(k+1)\phi+k)}{(1-\phi^k)(1-\phi^{k+1})}$ where $k = j - i$.

This value is smallest at $k = 1$. Substituting $k = 1$, we get the probability of an inversion is at most $\frac{\phi}{1+\phi}$. ◀

We now want to prove that with high probability, for every pair of symbols, the majority of rankings in S have that pair of symbols in the same relative order as π. If this holds, then following the majority when deciding the relative order of each pair of symbols will exactly reconstruct π.

▶ **Lemma 19.** *Let S be a set of at least $\frac{12(1+\phi)^2}{(1-\phi)^2} \log n$ rankings drawn from $M(\pi, \phi)$. Then we have that with probability at least $1 - \frac{1}{n}$, for every pair of symbols (a, b), if $a \prec_\pi b$, then*

$$|\{\pi_r \mid a \prec_{\pi_i} b, \pi_r \in S\}| \geq \frac{|S|}{2}.$$

Proof. Let $s = |S|$ and take any pair of symbols $a, b \in [n]$.

Given a ranking $\pi_r \in S$, let X_r be an indicator random variable that takes on 1 if π_r has an inversion in relative order of a and b compared to π. By Lemma 18, we have that $\mathbb{E}[X_r] \leq \frac{\phi}{1+\phi}$.

Let $X = \sum_{r=1}^{s} X_r$. Then by the linearity of expectation, $\mathbb{E}[X] \leq s\frac{\phi}{1+\phi}$. Applying the standard Hoeffding bound,

$$\begin{aligned}
\Pr\left[X - \mathbb{E}[X] > \left(\frac{1}{2} - \frac{\phi}{1+\phi}\right)s\right] &\leq \exp\left(\frac{-2\left(\left(\frac{1}{2} - \frac{\phi}{1+\phi}\right)s\right)^2}{s}\right) \\
&\leq \exp\left(-2\left(\frac{1}{2} - \frac{\phi}{1+\phi}\right)^2 \frac{12(1+\phi)^2}{(1-\phi)^2}\log n\right) \\
&\leq \exp\left(-3\log n\right) \\
&\leq n^{-3}.
\end{aligned}$$

As there are $n(n-1)/2$ pairs of symbols, taking a union bound over all such pairs gives that with probability at most $1/n$, there exists at least one pair in different pairwise order compared to π. Therefore, with probability at least $1 - 1/n$, we have that every pair of symbols has the same relative order in at least half of the rankings in S and in π. ◀

Finally, the proof of Theorem 17 follows directly from Lemma 19.

Proof of Theorem 17. Consider the execution of the comparison-based sorting algorithm. Consider each time the algorithm compares some pair of symbols (a, b). If $a \prec_\pi b$ and $|\{\pi_i | a \prec_{\pi_i} b\}| \geq \frac{|S|}{2}$, then it will order this pair of symbols identically to π. If this holds for every pair it compares, then it will produce the ranking π at the end of the execution. From Lemma 19, with probability at least $1 - 1/n$, this holds for every pair of symbols, and so the algorithm correctly returns π. ◀