Robust Consensus in Ranking Data Analysis: Definitions, Properties and Computational Issues

Ekhine Irurozki, Morgane Goibert, Clément Calauzènes, Stéphan Clémençon

Robust Consensus in Ranking Data Analysis: Definitions, Properties and Computational Issues

Morgane Goibert [1 2] **Clément Calauzènes** [1] **Ekhine Irurozki** [2] **Stephan Clémençon** [2]

Abstract

As the issue of robustness in AI systems becomes vital, statistical learning techniques that are reliable even in presence of partly contaminated data have to be developed. Preference data, in the form of (complete) rankings in the simplest situations, are no exception and the demand for appropriate concepts and tools is all the more pressing given that technologies fed by or producing this type of data (*e.g.* search engines, recommending systems) are now massively deployed. However, the lack of vector space structure for the set of rankings (*i.e.* the symmetric group $\mathfrak{S}_n$) and the complex nature of statistics considered in ranking data analysis make the formulation of robustness objectives in this domain challenging. In this paper, we introduce notions of robustness, together with dedicated statistical methods, for *Consensus Ranking* the flagship problem in ranking data analysis, aiming at summarizing a probability distribution on $\mathfrak{S}_n$ by a *median* ranking. Precisely, we propose specific extensions of the popular concept of *breakdown point*, tailored to consensus ranking, and address the related computational issues. Beyond the theoretical contributions, the relevance of the approach proposed is supported by an experimental study.

1. Introduction

One of the keys to the path of a trustworthy AI is undeniably the design of statistical learning techniques that can resist, to a certain extent, possible corruptions of the training dataset. The analysis of the influence of atypical observations on the outputs of machine-learning algorithms has received increasing interest in the AI literature these last few years and has recently motivated a wide variety of dedicated works (refer to Lugosi & Mendelson (2019); Lerasle et al. (2019) for instance), revisiting in particular seminal concepts in *Robust Statistics* such as the ε-contamination model, where the training dataset is supposedly contaminated by a fraction $\varepsilon \in (0, 1)$ of outliers (Huber, 1964). It is the goal of this paper to investigate the statistical analysis of ranking data from the perspective of robustness. Ranking data are indeed ubiquitous in modern technologies such as search engines or recommending systems and the question of their reliability in presence of corrupted data is a scientific challenge. Given the nature of preference data, observable in the form of permutations (complete rankings, *i.e.* elements of the symmetric group $\mathfrak{S}_n$) in the simplest case, informative statistics based on the latter are far from being simple. This is mainly due to the lack of vector space structure on $\mathfrak{S}_n$ and the impossibility of averaging directly such data. A major problem in ranking data analysis referred to as *Consensus Ranking* or *Ranking Aggregation*, and which the present article focuses on, consists in its simplest formulation in summarizing a ranking distribution (*i.e.* a probability distribution on $\mathfrak{S}_n$) by a *median ranking* (Kemeny, 1959). Even though this problem has a long history in social choice theory, see *e.g.* De Condorcet et al. (1785); de Borda (1781), it has been the subject of much attention within the machine-learning community, see *e.g.* Procaccia & Shah (2016); Jiao et al. (2016) among many others, references being far too numerous to be listed exhaustively. While most documented works concern the issue of computing (approximately) median rankings with theoretical guarantees, this paper studies in contrast the robustness properties of consensus ranking methods by means of a novel approach, extending that developed in Huber & Ronchetti (2009) for multivariate data. We emphasize that this angle is original to the best of our knowledge and distinguishes itself from related results in social choice theory, where median rankings are identified with *voting rules*. In line with these works, the well-known Gibbard-Satterthwaite theorem (Gibbard et al., 1973; Satterthwaite, 1975) states that every reasonable voting rule can be manipulated. We point out that there has been a wide body of research devoted to characterizing the complexity of computing manipulations, NP-hardness result on manipulation being considered as a guarantee for robustness

[1]Criteo AI Lab, Paris, France [2]Télécom Paris, Paris, France. Correspondence to: Morgane Goibert <morgane.goibert@gmail.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

(Bartholdi III et al., 1989; Davies et al., 2011; Brandt et al., 2016). However, beyond-worst-case analysis shows that the problems are easy in practice (Zuckerman et al., 2009). In the present article, we complement these works on the issue of robustness to vote manipulation by investigating how the seminal concept of *breakdown point*, a popular measure of robustness of estimators in multivariate statistical analysis, may apply to consensus ranking. Basically, it can be defined as the proportion of outliers or (possibly deliberately) corrupted observations that can contaminate the data sample without jeopardizing the statistic. As will be shown, one of the main difficulties faced in the context considered here lies in the fact that consensus rankings are often obtained by solving an optimization problem and no closed analytical form for the solutions is available in general. Consequently, the computation of breakdown points of ranking statistics is generally a computational challenge. Our main proposal here consists in relaxing the constraint stipulating that the summary of a ranking distribution should be necessarily represented by a single ranking (*i.e.* a strict order on the set of items indexed by $i \in \{1, \ldots, n\}$), or equivalently by a point mass on $\mathfrak{S}_n$. Instead, we suggest summarizing a ranking distribution by a *bucket ranking* (*i.e.* a weak order on the set $\{1, \ldots, n\}$), the possibility of observing ties in the orderings considered being shown to have crucial advantages regarding robustness.

The paper is organized as follows. In Section 2, basics in ranking aggregation and the notion of breakdown function are introduced, as well as the contributions of our paper. Section 3 focus on robustness, by detailing our theoretical results on the breakdown functions for the classical median, extending this concept to bucket rankings, and providing an optimization algorithm to estimate it in practice. Section 4 is dedicated to the definition of our robust statistic, called the Downward Merge statistic. Finally, experiments are done in Section 5 to highlight the usefulness of our Downward Merge statistic for solving Robust Consensus Ranking tasks.

2. Framework and Problem Statement

We start with a reminder of key concepts in ranking data analysis and *Robust Statistics*. The interested reader can refer to Alvo & Yu (2014); Huber & Ronchetti (2009) for more details. Here and throughout, a ranking over a set of $n \geq 1$ items is represented as a permutation $\sigma \in \mathfrak{S}_n$ where $\mathfrak{S}_n$ is the symmetric group. By convention, the rank r of an item $i \in [n]$ is $r = \sigma(i)$. For any measurable space $\mathcal{X}$, $\mathcal{M}_+^1(\mathcal{X})$ is the set of probability measures on $\mathcal{X}$, $\mathrm{TV}(p, q)$ the total variation distance between p and q in $\mathcal{M}_+^1(\mathcal{X})$.

2.1. Ranking Data and Summary Statistics

The descriptive analysis of probability distributions, or datasets for their empirical counterparts, is a fundamental problem in statistics. For distributions on Euclidean spaces such as $\mathbb{R}^d$, this problem has been widely studied and covered by the literature, with the study of statistics ranging from the simplistic sample mean to more sophisticated data functionals, such as $U/L/R/M$-statistics or depth functions for instance (van der Vaart, 1998).

Defining similar notions for probability distributions on $\mathfrak{S}_n$, the space of rankings, is challenging due to the absence of vector space structure. However, fueled by the recent surge of applications using preference data, such as *e.g.* recommender systems, the statistical analysis of ranking data has recently regained attention and certain classic problems have been revisited, as for instance those related to consensus rankings and their generalization ability (see *e.g.* Korba et al. (2017) and the references therein) or to the extension of depth functions to ranking data (Goibert et al., 2022).

Central tendency or location. Statistics measuring centrality, such as the mean (or the median for univariate distribution), can be seen as barycenters of the sampling observations w.r.t a certain distance. Consensus Ranking / Ranking Aggregation extends this idea to probability distributions on $\mathfrak{S}_n$ (Deza & Deza, 2009). Given a (pseudo-)metric d defined on $\mathfrak{S}_n$ and a distribution $p \in \mathcal{M}_+^1(\mathfrak{S}_n)$, a *ranking median* $\sigma_{p,d}^{\mathrm{med}} \in \mathfrak{S}_n$ can be defined as

$$\sigma_d^{\mathrm{med}}(p) := \operatorname*{argmin}_{\sigma \in \mathfrak{S}_n} \mathbb{E}_{\Sigma \sim p}(d(\sigma, \Sigma)). \tag{1}$$

A well-studied instance of ranking median is the *Kemeny consensus*, which corresponds to the situation where d is the *Kendall Tau* distance: for all $\sigma,\ \nu$ in $\mathfrak{S}_n$,

$$d_\tau(\sigma, \nu) = \frac{2}{n(n-1)} \sum_{i<j} \mathbb{1}_{[\sigma(i)<\sigma(j)]} \mathbb{1}_{[\nu(i)>\nu(j)]} \tag{2}$$

Another common choice is the *Borda count* when d is the *Spearman Rho*, see Appendix A for more details. Moreover, when d is the Kendall tau, Borda is a $O(n \log n)$, 5-approximation of the Kemeny ranking (Caragiannis et al., 2013; Jiao et al., 2016; Coppersmith et al., 2010), which is a NP-hard to compute (Dwork et al., 2001).

More complex statistics based on ranking data. Often, the information carried by a location statistic must be complemented. For instance, a notion of *dispersion* or *shape* is generally key to assessing convergence results or building confidence regions. To this end, the notion of *statistical depth function* has been developed for multivariate data (in Euclidean spaces) (see (Zuo & Serfling, 2000) and the references therein) and recently adapted to ranking, refer to (Goibert et al., 2022). However, as more complex statistics are more likely to exhibit robustness issues, we focus on simple statistics estimating location for ranking distribution.

2.2. Robust Statistics

To evaluate the robustness of a statistic, the notion of *breakdown function* has been introduced in the seminal work of (Huber, 1964). Informally, the breakdown function for a statistic T on a distribution p measures the minimal attack budget required for an adversarial distribution to change the outcome of the statistic T by an amount at least $\delta > 0$.

Definition 2.1. (BREAKDOWN FUNCTION) Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, $p \in \mathcal{M}_+^1(\mathcal{X})$, $T : \mathcal{M}_+^1(\mathcal{X}) \to \mathcal{Y}$ a measurable function and d a metric on $\mathcal{Y}$. For any level $\delta \geq 0$, the breakdown function of the functional T at p is

$$\varepsilon^\star_{d,p,T}(\delta) = \inf \left\{ \varepsilon > 0 \,\middle|\, \sup_{q:\mathrm{TV}(p,q)\leq\varepsilon} d(T(p), T(q)) \geq \delta \right\}.$$

In the traditional case $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, the level δ is generally set to $+\infty$ and the budget required is referred to as *breakdown point*. In the extreme case, when T is the identity and $\delta = 0^+$, $\varepsilon^\star$ quantifies the budget of attack under which *identifiability* of the distribution is possible (which requires the additional knowledge that p belongs to some family).

Application to Ranking Data. In Agarwal et al. (2020) such a study on identifiability is provided for the Bradley-Terry-Luce (Bradley & Terry, 1952; Luce, 1959) model under a budget constraint on pairwise marginals rather than the Total Variation, and Jin et al. (2018) on the Heterogeneous Thurstone Models (Thurstone, 1927). However, summary statistics, such as a central tendency, are generally harder to break than the full distribution itself, so the breakdown function provides a finer quantification of robustness than the identifiability of the distribution. Since the distances on $\mathfrak{S}_n$ are bounded, in general, the full breakdown function needs to be considered and one cannot focus only on a particular level such as $\delta = 0^+$ or $\delta = +\infty$. From here and throughout, the distance d and the attack amplitude δ are normalized to lie between 0 and 1.

The robustness of the median statistic when an adversary is allowed to attack with any strategy a pairwise model has also been studied (Datar et al., 2022). They characterize the robustness of two statistics in terms of the L2 distance on distributions. We propose in Definition 2.1 a more general and natural measure for robustness as a function of the distance between the true and a corrupted statistic.

Bucket Rankings as a robustness candidate. In rankings, adversarial attacks often target pairs of items that are "close" in some sense (Agarwal et al., 2020): consecutive ranks, a pairwise marginal probability close to $\frac{1}{2}$, ... Thus, a simple and efficient way to robustify a ranking median is to accept *ties*, rather than being restricted to a strict order.

2.3. Challenges and Contributions

There is a wide number of median statistic studies motivated by the lack of analytical expression and the computational and statistical challenges that arise in the estimation process. However, robustness results for ranking statistics are rare and not rigorous enough for comparing different estimators.

Contribution 1. *Using Definition 2.1 with the Kendall tau distance provides a straightforward measure of robustness for ranking medians. In Section 3.1 we provide a lower-bound on the breakdown function for a ranking median (Theorem 3.2) and a tight upper-bound for the Kemeny consensus (Theorem 3.2).*

Moreover, slight perturbations in the pairwise relations of items that are similar to each other can imply breaking a median estimator, showing a lack of robustness. It is natural to propose more robust estimators by allowing pairs of items to be "equally ranked", i.e., by considering bucket ranking statistics. However, generalizations of the breakdown function for bucket rankings require the use of Kendall tau for buckets, which is computationally impractical.

Contribution 2. *In Section 3.2 we propose an extension of the breakdown function for bucket rankings which is built upon a Hausdorff generalization of the Kendall tau distance. We also develop an optimization algorithm to approximate this breakdown function that overcomes the computational issue of having a piece-wise constant objective function.*

We illustrate and show empirically that bucket rankings are more robust median estimators than rankings. However, finding the optimal bucket order statistic requires exhaustively searching the space of bucket rankings Π_n, which is even larger than the space of permutations, of factorial cardinality, and therefore, it is totally infeasible.

Contribution 3. *In Section 4 we propose a general method for robustifying medians: given a ranking median, our algorithm successively merges "similar" items together into the same bucket. We evaluate this statistic in Section 5, showing an improvement of robustness w.r.t. Kemeny's median without sacrificing its precision.*

3. Robustness - Breakdown Function for Ranking and Bucket Rankings

This section first details how to apply the notion of *breakdown function* $\varepsilon^\star_{d,p,T}$. This allows providing insights into the robustness of classical location statistics such as the Kemeny consensus. These results advocate for the introduction of a more robust type of statistics based on bucket orders that are also developed in this section.

3.1. Breakdown Function for the Kemeny Consensus

We explore the robustness of ranking medians $\sigma_d^{\text{med}}(p)$ as defined in Equation (1) for different metrics d over $\mathfrak{S}_n$ as defined by the breakdown function $\varepsilon^\star_{d_\tau,p,T}$. In particular, it is possible to tightly sandwich the breakdown function for the Kemeny median.

Theorem 3.1. *For $p \in \mathcal{M}_+^1(\mathfrak{S}_n)$, $\sigma_p^\star = \sigma_{d_\tau}^{\text{med}}(p)$ (Kemeny median) and $\delta \geq 0$, if $\varepsilon^+(\delta) \leq 2p(\sigma_p^*)$ then $\varepsilon^\star_{d_\tau,p,\sigma_p^\star}(\delta) \leq \varepsilon^+(\delta)$ with*

$$\varepsilon^+(\delta) = \min_{\substack{\sigma\in\mathfrak{S}_n \\ d_\tau(\sigma,\sigma_p^\star)\geq\delta}} \max_{\substack{\nu\in\mathfrak{S}_n \\ d_\tau(\nu,\sigma_p^\star)<\delta}} \frac{\mathbb{E}_{\Sigma\sim p}\left[d_\tau(\Sigma,\sigma) - d_\tau(\Sigma,\nu)\right]}{d_\tau(\sigma_p^\star,\sigma) - d_\tau(\sigma_p^\star,\nu)}.$$

Proof Sketch. Detailed Proof can be found in Appendix C.1 The proof relies on showing that, for $\varepsilon > 0$, the *attack* distribution $\bar{q}_\varepsilon = p - \frac{\varepsilon}{2}\mathbb{1}_{[\cdot=\sigma_p^*]} + \frac{\varepsilon}{2}\mathbb{1}_{[\cdot=\sigma_p^{\star,\text{rev}}]}$, where $\sigma_p^{\star,\text{rev}}$ is the reverse of $\sigma_p^\star$, is in the feasible set of the optimization problem $\sup_{q:\text{TV}(p,q)\leq\varepsilon} d_\tau(\sigma_p^*,\sigma_q^*)$ (see Definition 2.1).

Using $\bar{q}_\varepsilon$ provides a way to link ε and δ. The condition $\varepsilon^+(\delta) \leq 2p(\sigma_p^\star)$ ensures $\bar{q}_\varepsilon$ is well-defined. □

It is also possible to provide a lower bound on the breakdown function for any generic ranking median.

Theorem 3.2. *For $p \in \mathcal{M}_+^1(\mathfrak{S}_n)$, m and d being two metrics on $\mathfrak{S}_n$, $\sigma_p^\star = \sigma_d^{\text{med}}(p)$ and $\delta \geq 0$, we have $\varepsilon^\star_{m,p,\sigma_p^\star}(\delta) \geq \varepsilon^-(\delta)$ with*

$$\varepsilon^-(\delta) = \min_{\substack{\sigma\in\mathfrak{S}_n \\ m(\sigma,\sigma_p^\star)\geq\delta}} \max_{\substack{\nu\in\mathfrak{S}_n \\ \nu\neq\sigma}} \frac{\mathbb{E}_{\Sigma\sim p}\left[d(\Sigma,\sigma) - d(\Sigma,\nu)\right]}{\max_{\sigma'\in\mathfrak{S}_n} d(\sigma',\sigma) - d(\sigma',\nu)}$$

Proof. Detailed proof can be found in Appendix C.2. □

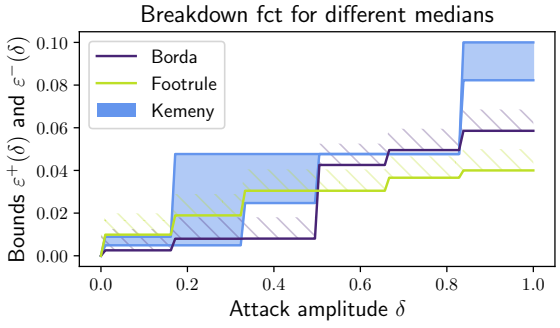

Figure 1. An illustration of $\varepsilon^+(\delta)$ and $\varepsilon^-(\delta)$ (from Theorem 3.1 and Theorem 3.2) for a distribution on permutations of 4 items. For Borda and the median associated with Spearman footrule, only the lower bound is displayed.

Figure 1 shows that no choice of d makes the median uniformly more robust than another. Then, unfortunately, it also illustrates the fragility of median statistics against corruption of the distribution. In this example, impacting the distribution p by less than 5% allows changing the Kemeny median by flipping more than half item pairs ($\delta \geq 0.5$).

Sensitivity to similar items. To further illustrate the fragility of Kemeny's median, Figure 2 shows its breakdown function on specific distributions. As could be expected, if all items are almost indifferent (uniform distribution - purple curve), then a ranking median is very fragile: a small nudge on p is enough to change the Kemeny median from one ranking to its reverse. On the contrary, when p is a point mass at a given ranking (blue curve), it requires a large attack on p to impact the median.

The green curve shows a weakness in the median: despite p being concentrated on two neighbouring rankings (identical up to a pair of adjacent items), the robustness is very low for $\delta \leq 0.2$. This highlights a mechanism underlying adversarial attacks in real-world recommender systems (ex: fake reviews...): at a small cost, it is possible to be systematically ranked on top of close alternatives. This calls for using the natural alternative to (strict) rankings, which incorporates indifference between items: *bucket rankings*.

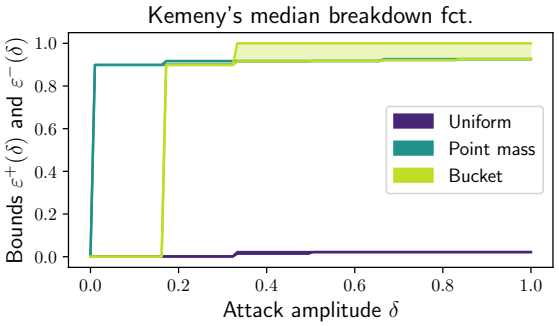

Figure 2. Breakdown function for Kemeny's median for different distributions p. "Uniform" denotes an almost uniform distribution; "Point mass" an almost point mass distribution, and "Bucket" an almost point mass distribution on two neighboring rankings.

3.2. Bucket Ranking - Extended Ranking Consensus

Intuitively, bucket rankings are rankings with ties allowed. Formally, they can equivalently be defined as a total preorder – *i.e.* a homogeneous binary relation that satisfies transitivity and reflexivity (preorder) in which any two elements are comparable (total) – or as a strict weak ordering – *i.e.* a strict total order over equivalence classes of items (buckets).

Definition 3.3. (BUCKET RANKING) A bucket order π is a strict weak order defined by an ordered partition of $[n]$, *i.e.* a sequence $(\pi^{(1)},\ldots,\pi^{(k)})$ of $k \geq 1$ pairwise disjoint non empty subsets (buckets) of $[n]$ such that:

(i) $i \prec_\pi j \;\Leftrightarrow\; \exists l < l' \in [k], (i,j) \in \pi^{(l)} \times \pi^{(l')}$,

(ii) $i \sim_\pi j \;\Leftrightarrow\; \exists l \in [k], (i,j) \in \pi^{(l)} \times \pi^{(l)}$,

We denote Π_n the set of bucket rankings, which is of size $\sum_{k=1}^{n} k!S(n,k)$[1] (vs $n!$ for $\mathfrak{S}_n$).

The indifference between items that bucket rankings can incorporate is an interesting feature to gain robustness, because the statistic can output alternatives between several strict orders, making it harder to attack.

As sets of permutations. A bucket ranking $\pi \in \Pi_n$ can be equivalently mapped to a subset of permutations, generated through the different ways to break ties. We say that a permutation $\sigma \in \mathfrak{S}_n$ is *compatible* with a bucket ranking $\pi \in \Pi_n$ – denoted $\sigma \in \pi$ – if for any $i, j \in [n]$, $\sigma(i) < \sigma(j) \;\Leftrightarrow\; i \prec_\pi j$ or $i \sim_\pi j$. For two bucket orders π_1, π_2, we say that π_1 is *stricter* that π_2, denoted $\pi_1 \subseteq \pi_2$, iff for any $\sigma \in \mathfrak{S}_n$, $\sigma \in \pi_1 \Rightarrow \sigma \in \pi_2$.

As a distribution. Being a set of permutations, a bucket order $\pi \in \Pi_n$ can also be seen as a uniform distribution with restricted support. This point of view is particularly intuitive from a robustness perspective: a randomized output is generally harder to attack for an adversary.

Distances between bucket rankings. A key to applying the breakdown function from Definition 2.1 to bucket orders statistics is to have a metric on Π_n that extends those defined on $\mathfrak{S}_n$. To this end, we use the previous remark that weak orders are sets of rankings as well as a classical Hausdorff extension of metrics to sets. More precisely, we define:

Definition 3.4. (NON-SYMMETRIC HAUSDORFF) Let d be a metric on $\mathfrak{S}_n$. The non-symmetric Hausdorff pseudoquasi-metric between two bucket rankings $\pi_1, \pi_2 \in \Pi_n$ is

$$H_d^{\text{NS}}(\pi_1, \pi_2) = \max_{\sigma_2 \in \pi_2} \min_{\sigma_1 \in \pi_1} d(\sigma_1, \sigma_2)\,.$$

Even though it is not a metric, H_d^{NS} is well-suited to ranking with ties. Intuitively, its lack of symmetry allows differentiating adversarial attacks whose effect is on the strict part of the bucket order (e.g. swapping two items that are strictly ordered) from those whose effect is "only" to disambiguate a tie. More precisely, if $\pi_2 \subseteq \pi_1$, then $H_d^{\text{NS}}(\pi_1, \pi_2) = 0$. Depending on the application, one may want to focus on the first type of attacks, in which case H_d^{NS} is a suitable choice to define the breakdown function as $\varepsilon^\star_{H_d^{\text{NS}},p,T}$. Otherwise, it is possible (and usual) to symmetrize the Hausdorff metric.

Definition 3.5. ($1/2$-SYMMETRIC HAUSDORFF) Let d be a metric on $\mathfrak{S}_n$. The $1/2$-symmetric Hausdorff metric between two bucket rankings $\pi_1, \pi_2 \in \Pi_n$ is defined by

$$H_d^{(1/2)}(\pi_1, \pi_2) = \frac{1}{2}\Big(H_d^{\text{NS}}(\pi_1, \pi_2) + H_d^{\text{NS}}(\pi_2, \pi_1)\Big)\,.$$

Usual symmetrization of the Hausdorff metric uses a maximum rather than an average (Fagin et al., 2006). However, under the Kendall-tau distance, the average version is computationally simpler (see Appendix D for more details).

3.3. The Breakdown Function in Ranking Data Analysis - Definition and Estimation

Definition. Putting all the pieces together, from now on, the statistic $T : \mathcal{M}_+^1(\mathfrak{S}_n) \to \Pi_n$ summarizes a distribution over $\mathfrak{S}_n$ by a bucket ranking in Π_n. Then, we use either $H_{d_\tau}^{(NS)}(\pi_1, \pi_2)$ (see Definition 3.4) or $H_{d_\tau}^{(1/2)}(\pi_1, \pi_2)$ on Π_n where d_τ is the Kendall tau (see Equation (2)). Finally, the breakdown function $\varepsilon^\star_{H_{d_\tau}^{(NS)},p,T}$ is the result of the following optimization problem

$$\inf\left\{\varepsilon > 0 \,\middle|\, \sup_{q:\text{TV}(p,q)\leq\varepsilon} H_{d_\tau}^{(NS)}(T(p), T(q)) \geq \delta\right\} \quad (3)$$

The Empirical Breakdown Function. Computing a closed-form expression for the breakdown point for any statistic T and distribution p is challenging in general. However, it can be estimated empirically: the extended expression of the breakdown function in Equation (3) can be simplified so that it is the solution to the following Lagrangian-relaxed optimization problem.

$$\inf_{q \in \Delta^{\mathfrak{S}_n}} \sup_{\lambda \geq 0} 1/2\|p - q\|_1 + \lambda(\delta - H_{d_\tau}^{(NS)}(T(p), T(q))) \quad (4)$$

Smoothing. As $H_{d_\tau}^{(NS)}(T(p), T(q)))$ is piece-wise constant as a function of q (with a combinatorial number of pieces), Problem (4) cannot directly be solve using standard optimization techniques. To solve this issue, we used a smoothing procedure by convolving this function with a smoothing kernel k_γ with scale γ. Thus, after the relaxation, the optimization problem (4) becomes:

$$\inf_{q \in \Delta^{\mathfrak{S}_n}} \sup_{\lambda \geq 0} 1/2\|p - q\|_1 + \lambda(\delta - \rho_T(p, q)), \quad (5)$$

with

$$\begin{aligned} \rho_T(p, q) &= H_{d_\tau}^{(NS)}(T(p), T(q)) \star k_\gamma(q) \\ &= \int_u H_{d_\tau}^{(NS)}(T(p), T(u)) \times k_\gamma(q - u)\mathrm{d}u, \end{aligned} \quad (6)$$

On a practical note, a simple way to build a convolution kernel k_γ on a simplex like $\mathcal{M}_+^1(\mathfrak{S}_n)$, is to use a convolution kernel κ_γ on the whole euclidean space –

[1] $S(n,k)$ are Stirling numbers of the second kind.

for instance an independent Gaussian density $\kappa_\gamma(x) = \frac{1}{\sqrt{(2\pi\gamma)^{n!}}} \exp\left\{(-\frac{x^\mathrm{T}x}{2\gamma^2})\right\}$ – and set k_γ to be the density of the push-forward through a *softmax* function. We denote $\varepsilon^\gamma_{p,T}(\delta)$ the limiting value of $\|p-q\|_1/2$ at the solution of (5). Note the bias induced by such definition of k_γ fades away when γ goes to 0 in the same way as the bias induced by the convolution. This smoothing ensures ρ_T is a continuous, differentiable function with respect to q. Moreover, it can easily be estimated using a Monte-Carlo sampling, using the following remark: $\rho_T(p,q) = \mathbb{E}_{u\sim k_{(p,\gamma)}}(H^{(NS)}_{d_\tau}(T(u),T(q))$.

Optimization. When using Monte-Carlo estimation for ρ_T, Equation (5) is a stochastic saddle-point problem. To solve such problems, gradient/ascent has a rate of convergence of $\mathcal{O}(t^{1/2})$ for its ergodic average (t being the number of steps) (Nemirovski & Rubinstein, 2002). Our empirical optimization algorithm for computing the breakdown functions relies on stochastic gradient descent and is able to provide good approximations, as illustrated in Figure 4. We denote $\hat{\varepsilon}^\gamma_{p,T}(\delta) = \|p-\bar{q}_t\|_1$, where $\bar{q}_t$ is the ergodic average of the iterates $(q_s)_{s\leq t}$ obtained during the optimization.

Let's make a couple of remarks on the empirical breakdown function $\hat{\varepsilon}^\gamma_{p,T}$. First, it is a noisy estimate of $\varepsilon^\gamma_{p,T}$ as ρ_T and its gradients are estimated via Monte-Carlo. Thus, the choice of γ and t should trade-off the variance of $\hat{\varepsilon}^\gamma_{p,T}$ and the bias $|\varepsilon^\gamma_{p,T} - \varepsilon^\star_{d_\tau,p,T}|$. Second, as the term $\|p-q\|_1$ is minimized in (5), it is expected $\hat{\varepsilon}^\gamma_{p,T}$ over-estimates $\varepsilon^\gamma_{p,T}$.

4. Robust Consensus Ranking Statistics

As proved by Theorem 3.1, the classical median statistics as defined by (1) can be easily broken, which motivates defining more robust statistics, based on bucket rankings. As illustrated by Figure 2, the weakness of median statistics comes from being "forced" to rank all items, even those which are (almost) indistinguishable. Bucket rankings seem to be a natural solution to this problem, but *what is a good way to build a bucket order statistic?*

As $H^{(NS)}_{d_\tau}$ defines a (pseudoquasi-) distance on Π_n, we could adapt the idea of a median as in (1) for bucket rankings. However, contrarily Borda medians which can be computed in a scalable way (Caragiannis et al., 2013), Hausdorff-based medians would require to optimize over Π_n. As its cardinality is larger than $\mathfrak{S}_n$ this problem can be more computationally challenging than Kemeny's median.

A more scalable approach is to start from a ranking median such as the Kemeny or Borda consensus and to robustify it using a plug-in method based on merging items that are close into buckets. Figure 3 illustrates this idea. The left graph describes pairwise marginal probabilities for which the Kemeny consensus is $A \prec B \prec C \prec D$. Intuitively, merging either C and D (as $\mathbb{P}(C \prec D) = 0.51$) or B and C (as $\mathbb{P}(B \prec C) = 0.52$) leads to bucket rankings (i) and (ii), which will be harder to attack. However, this example also highlights that there is no unique way of merging items. For instance, if the constraint is to only merge items whose pairwise preference probability is in $[0.4, 0.6]$, it is possible to merge B, C or C, D, but not B, C, D as $\mathbb{P}(B \prec D) = 0.7$: *pairwise indistinguishability is not transitive*.

4.1. Naïve Merge Statistic

In order to formalize the latter intuition and to derive a first (naïve) plug-in rule, we define the pairwise preference probability between two items, which provides a relevant notion of closeness between items.

Definition 4.1. (PAIRWISE PROBABILITIES). For $p \in \mathcal{M}^1_+(\mathfrak{S}_n)$, the pairwise preference probability between items i and j, denoted $P_{i,j}$, is defined for $i \neq j$ by: $P_{i,j} = \mathbb{P}_{\Sigma\sim p}(\Sigma(i) < \Sigma(j))$. By convention, $P_{ii} = 0.5$. We define the pairwise matrix of p as $P := [P_{i,j}]_{1\leq i,j\leq n}$.

Then, given a bucket ranking $\pi \in \Pi_n$, we formalize the notion that two buckets can be merged, with the constraint of not changing the strict order between buckets. To this end, we define $\bar{P}_i(\pi)$, the *strongest deviation from indifference* between any two items within the i^{th} bucket $\pi^{(i)}$.

$$\bar{P}_i(\pi) = \max\left\{|P_{l,l'} - 0.5| : (l,l') \in \pi^{(i)}\right\} \tag{7}$$

Then, one needs to quantify the value of $\bar{P}_i(\pi)$ that would result from merging bucket i to bucket j,

$$\bar{P}_{ij}(\pi) = \max\left\{\left|P_{l,l'} - \frac{1}{2}\right| : (l,l') \in \bigcup_{\substack{l\in[n]\\ i\leq l\leq j}} \pi^{(l)}\right\} \tag{8}$$

Finally, given a threshold $\theta \in [0, 0.5]$ on the acceptable deviation from indifference, we define the set of pairs of buckets that can be merged while keeping $\bar{P}$ below θ,

$$\mathcal{G}(\pi,\theta) = \left\{(i,j) \in [n]^2 : \bar{P}_{ij}(\pi) \leq \theta\right\} \tag{9}$$

The first intuition is to merge buckets iteratively, starting with the most indifferent ones, as described in Algorithm 1.

Termination of Algorithm 1 is guaranteed by the fact that the number of buckets in π strictly decreases at each iteration. Then, by definition of $\mathcal{G}(\pi,\theta)$, the resulting bucket ranking π is such that any of its bucket i satisfies $\bar{P}_i(\pi) \leq \theta$ – *i.e.* no two items with higher deviation than θ have been merged.

Despite being very natural, this algorithm suffers from an important limitation: when changing the threshold θ, its output only spans a limited subset of valid bucket rankings. In the example provided by Figure 3, the naïve merge method

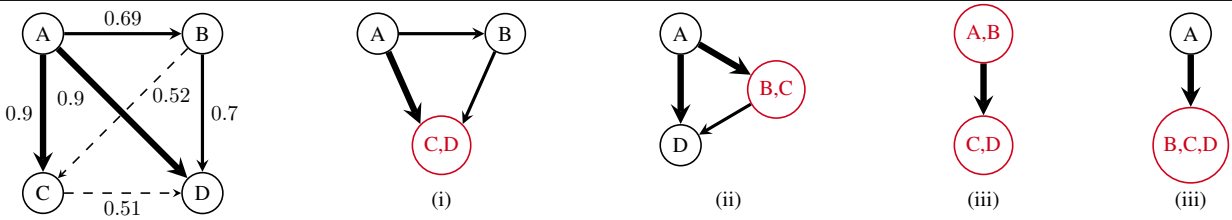

Figure 3. Left: Directed Graph that summarizes a pairwise marginal probability matrix. (i-iv) Graph representations of bucket orders that are compatible with merging items which pairwise preference probability is below 0.52 (i, ii) and below 0.7 (iii,iv).

Algorithm 1 Naïve Merge

Input: Pairwise matrix P, Ranking median σ, threshold $\theta \in [0, 0.5]$.
$\pi \leftarrow \sigma$ `// σ as a bucket ranking`
repeat
$(i^*, j^*) = \text{argmin}_{(i,j)\in\mathcal{G}(\pi,\theta)} \bar{P}_{ij}(\pi)$
update π by merging all buckets between i^* and j^*

$$\begin{cases} \pi^{(i)} & \leftarrow \pi^{(i)} \quad \text{for } i < i^* \\ \pi^{(i^*)} & \leftarrow \bigcup_{l\in[n], i^*\leq l\leq j^*} \pi^{(l)} \\ \pi^{(i-j^*+i^*)} & \leftarrow \pi^{(i)} \quad \text{for } i > j^* \end{cases}$$

until $\mathcal{G}(\pi, \theta) = \emptyset$
Output: π

Algorithm 2 Downward Merge

Input: Pairwise matrix P, Ranking median σ, threshold $\theta \in [0, 0.5]$.
$\pi \leftarrow \sigma$ `// σ as a bucket ranking`
repeat
$(i^*, j^*) = \text{argmax}_{(i,j)\in\mathcal{G}(\pi,\theta)} \bar{P}_{ij}(\pi)$
update π by merging all buckets between i^* and j^*

$$\begin{cases} \pi^{(i)} & \leftarrow \pi^{(i)} \quad \text{for } i < i^* \\ \pi^{(i^*)} & \leftarrow \bigcup_{l\in[n], i^*\leq l\leq j^*} \pi^{(l)} \\ \pi^{(i-j^*+i^*)} & \leftarrow \pi^{(i)} \quad \text{for } i > j^* \end{cases}$$

until $\mathcal{G}(\pi, \theta) = \emptyset$
Output: π

plugged-in on the Kemeny consensus can only output (i) and (iii). Whatever the value of θ, it can never output (ii) or (iv). This limitation is induced by its outputs being a monotonic (w.r.t. to inclusion) function of θ – *i.e.* for $\theta_1 \leq \theta_2$, the resulting bucket rankings satisfy $\pi_{\theta_1} \subseteq \pi_{\theta_2}$.

4.2. Downward Merge Statistic

Overcoming this limitation only requires a small change in the algorithm which results in our main plug-in method named *Downward Merge*, shown in Algorithm 2. Downward Merge algorithm selects the two buckets (i^*, j^*) whose deviation from indifference $\bar{P}_{ij}(\pi)$ is maximal among those $\bar{P}_{ij}(\pi) \leq \theta$. [2] Then, all the buckets l such that $i^* \leq l \leq j^*$ are merged. This process is repeated while there exist pairs of buckets whose deviation from indifference $\bar{P}_{ij}(\pi) \leq \theta$ and thus termination is guaranteed.

The Downward Merge method is thus able to span a larger set of bucket orders when varying θ. In the example from Figure 3, the Downward Merge method plugged-in on the Kemeny consensus can generate all four bucket rankings (i-iv) for $\theta \in \{0.51, 0.52, 0.69, 0.7)\}$.

The computation of the Downward Merge plugin is quite efficient: in the general setting, its complexity scales in $\mathcal{O}(n^3)$, which can be made more efficient in more specific cases. For example, if the distribution is strongly SST (meaning that it satisfies: $\forall(i, j, k)$, $p_{i,j} \geq \max(p_{i,k}, p_{k,j})$), the complexity reduces to $\mathcal{O}(n^2 \log(n))$; in the top-k / soft top-k ranking setting, the complexity is $\mathcal{O}(k^3)$ where $k = o(n)$; and finally, in the "small threshold" case, meaning $\#(i, j)|p_{i,j} \leq \theta \leq o(n)$, where θ is the threshold, then it is $\mathcal{O}(n \log(n))$. Note that this latter case is reasonable (especially for large values of n) since the purpose of our plugin is to robustify ranking statistics while not losing too much on precision: this means we prefer creating a small number of buckets rather than a large number.

The next experimental section illustrates the robustness improvement brought by this plug-in method over a ranking median.

5. Numerical Experiments

In this section, we illustrate the relevance of the statistic outputted by our Downward Merge plug-in on Kemeny's median (called our *Downward Merge statistic* for short) by running several illustrative experiments for various settings and comparing with the baseline provided by the usual Kemeny's median. The code is available here.

[2] Instead of taking the most similar buckets, as in the previous statistic, we take the most different pair among those that are "similar enough".

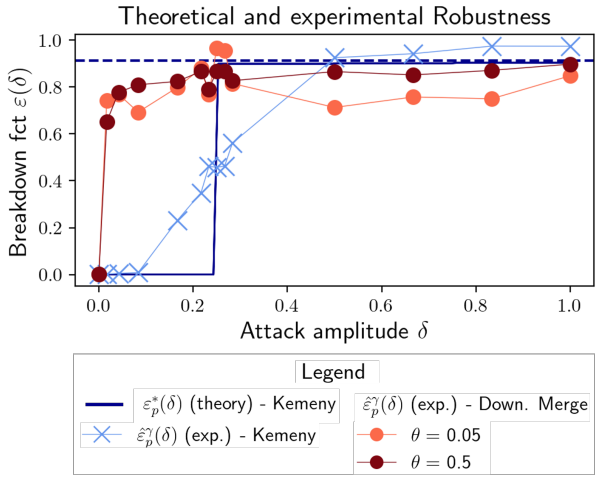

Figure 4. Breakdown function $\hat{\varepsilon}_{p,T}^{\gamma}(\delta)$ as a function of attack amplitude δ for a bucket distribution p (almost a point mass on two neighboring rankings) with $n = 4$. The plain blue line denotes the theoretical value for Kemeny's median $\varepsilon_p^*(\delta)$, blue crosses (resp. red dots) the empirical approximation $\hat{\varepsilon}_{p,T}^{\gamma}$ for Kemeny's median (resp. Down. Merge statistic for different thresholds θ).

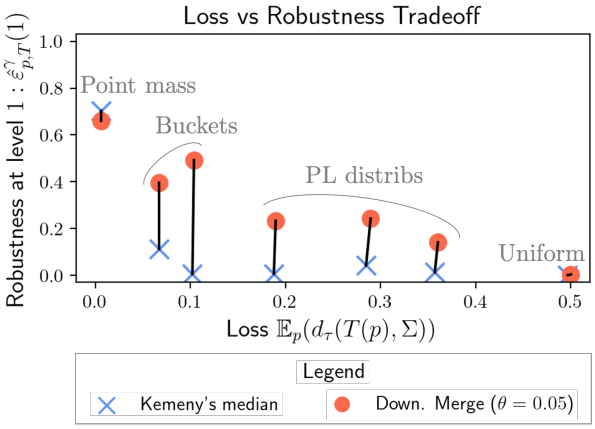

Figure 5. Loss/Robustness tradeoffs for different p with $\delta = 1$. Pairs of points linked by a black line denote results for Kemeny's median and Down. Merge statistics on the same distribution p with $n = 4$. "Buckets" are hand-picked distributions generated to be almost a point mass on a bucket order, "Uniform" (resp. "Point mass") "is an almost uniform (resp. point mass) hand-picked distribution, and "PL distribs." are random Plackett-Luce distributions.

5.1. Empirical Robustness

Our Downward Merge plug-in aims at providing a robustified statistic. To illustrate its usefulness, we ran experiments computing the approximate breakdown functions $\hat{\varepsilon}_{p,T}^{\gamma}(\delta)$ for the Kemeny's median as a baseline and our statistic when varying δ. Figure 4 shows the robustness as a function of attack amplitude δ and for a hand-picked distribution p that is almost a point mass on a bucket ranking.

When the threshold is set to a sensible value (here $\theta = 0.05$), the Downward Merge algorithm outputs a bucket order as a statistic: thus, the robustness increases very strongly to reach nearly optimal values even for very small values of δ, which illustrates its efficiency. When $\theta = 0.5$, the statistic is the bucket order regrouping all items. In this case, the statistic cannot be broken, and provide optimal values for the breakdown function. However, such a statistic does not provide any information about the distribution under analysis: its accuracy of location is very poor. Formally, the accuracy of location of a statistic T is defined by its closeness (under the same metric d used in its definition) to the whole ranking distribution: $AL_{d,p}(T) := \|d\|_\infty - \mathbb{E}_p(d(T(p), \Sigma))$, which is the opposite of the *loss*, as simply defined by $Loss_{d,p}(T) = \mathbb{E}_p(d(T(p), \Sigma))$. By definition, under metric $d = d_\tau$, Kemeny's median has the highest accuracy of location, *i.e.* the smallest loss. On the other hand, the Downward Merge statistic when $\theta = 0.5$ has a very high loss, which makes it irrelevant in most cases. These observations justify the analysis of the loss/robustness tradeoff of our Downward Merge statistic compared to Kemeny's median.

5.2. Tradeoffs between Loss and Robustness

We ran experiments for various distributions p and computed the loss and the breakdown function of Kemeny's median and our Downward Merge algorithm to show the loss/robustness tradeoff for each statistic. Figure 5 shows the results for different choices of distribution p when the number of items $n = 4$, and for $\delta = 1/6$ (normalized value of δ that requires at least a switch between two items to break the statistic).

The point mass (resp. the uniform) distribution represents the extreme case for which Kemeny's median is very robust (resp. not robust at all) and for which we expect no improvement from using the Downward Merge statistic. This intuition is verified in both cases, and we can see that the Downward Merge statistic yields the same results (in loss and in robustness) as Kemeny's median.

The bucket distributions (for which the gap between the probabilities for two rankings in the bucket order is respectively 0.1 and 0.01) represent the settings to which our Downward Merge is best suited. As expected, the improvement in robustness when using our Downward Merge statistic is high, and the increase in loss is negligible.

Finally, the Plackett Luce distributions (for which the parameters were generated randomly) represent a random setting. The results are interestingly very similar to those for the bucket distributions: the gain in robustness is high and the increase in loss is negligible. This random setting illustrates the usefulness of our Downward Merge statistic in

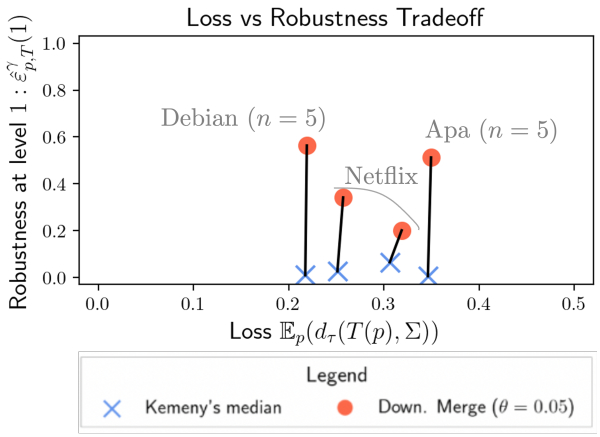

Figure 6. Loss/Robustness tradeoffs for different real-world datasets with $\delta = 1$. Pairs of points linked by a black line denote results for Kemeny's median and Down. Merge statistics on the same dataset.

general cases and shows that, overall, it yields a much better compromise than Kemeny's median.

To corroborate these findings, we also ran experiments using real-world datasets from the preflib library: two Netflix Prize datasets (resp. with $n = 3$ and $n = 4$ items), a Debian dataset (with $n = 5$ items) and an Apa dataset (with $n = 5$ items). The results are shown in Figure 6, and corroborate the synthetic results: our plugin always provides much better robustness, while the increase in the loss stays minimal.

6. Conclusion

In this paper, we developed a framework to study robustness in ranks: we defined breakdown functions for rankings, extended it to bucket rankings, and created an optimization algorithm to approximate its value in practice. We developed our Downward Merge statistic as a plug-in to the classical Kemeny's median to provide, as confirmed by our experiments, not only an improved robustness but also a better compromise between centrality and robustness. We ensured our Downward Merge algorithm is scalable to practical settings, but the evaluation of the breakdown function remains challenging because of the use of the Total-Variation distance as a metric for the budget constraint. The definition and study of further scalable approximations of the breakdown function are left for future work.

References

Agarwal, A., Agarwal, S., Khanna, S., and Patil, P. Rank aggregation from pairwise comparisons in the presence of adversarial corruptions. In *International Conference on Machine Learning*, pp. 85–95. PMLR, 2020.

Alvo, M. and Yu, P. L. H. *Statistical Methods for Ranking Data*. Springer, 2014.

Bartholdi III, J. J., Tovey, C. A., and Trick, M. A. The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6(3):227–241, 1989.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D. *Handbook of computational social choice*. 2016.

Calauzènes, C., Usunier, N., and Gallinari, P. Calibration and regret bounds for order-preserving surrogate losses in learning to rank. *Machine Learning*, 93(2):227–260, 2013.

Caragiannis, I., Procaccia, A. D., and Shah, N. When do noisy votes reveal the truth? pp. 143–160. ACM, 2013.

Coppersmith, D., Fleischer, L. K., and Rurda, A. Ordering by weighted number of wins gives a good ranking for weighted tournaments. *ACM Trans. Algorithms*, 6:1–13, 7 2010.

Critchlow, D. E. *Metric methods for analyzing partially ranked data*, volume 34. Springer Science & Business Media, 2012.

Datar, A., Rajkumar, A., and Augustine, J. Byzantine spectral ranking. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Davies, J., Katsirelos, G., Narodytska, N., and Walsh, T. Complexity of and algorithms for borda manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2011.

de Borda, J. C. Mémoire sur les élections au scrutin. 1781.

De Condorcet, N. et al. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Cambridge University Press, 1785.

Deza, M. and Deza, E. *Encyclopedia of Distances*. Springer, 2009.

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. Rank aggregation methods for the web. pp. 613–622. ACM, 2001.

Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., and Vee, E. Comparing partial rankings. *SIAM Journal on Discrete Mathematics*, 20(3):628–648, 2006.

Gibbard, A. et al. Manipulation of voting schemes: a general result. *Econometrica*, 41:587–601, 1973.

Goibert, M., Clémençon, S., Irurozki, E., and Mozharovskyi, P. Statistical Depth Functions for Ranking Distributions: Definitions, Statistical Learning and Applications. In *Proceedings of the Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Huber, P. J. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pp. 73–101, 1964.

Huber, P. J. and Ronchetti, E. M. *Robust Statistics*. 2nd edition, John Wiley & Sons, 2009.

Jiao, Y., Korba, A., and Sibony, E. Controlling the distance to a kemeny consensus without computing it. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

Jin, T., Xu, P., Gu, Q., and Farnoud, F. Rank aggregation via heterogeneous thurstone preference models. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Kemeny, J. G. Mathematics without numbers. *Daedalus*, 88:571–591, 1959.

Korba, A., Clémençon, S., and Sibony, E. A learning theory of ranking aggregation. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, (AISTATS)*, 2017.

Lerasle, M., Szabo, Z., Mathieu, T., and Lecué, G. Monk – outlier-robust mean embedding estimation by median-of-means. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

Luce, R. D. *Individual Choice Behavior: A Theoretical analysis*. Wiley, 1959.

Lugosi, G. and Mendelson, S. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 2019.

Nemirovski, A. and Rubinstein, R. Y. *An Efficient Stochastic Approximation Algorithm for Stochastic Saddle Point Problems*, pp. 156–184. New York, NY, 2002.

Procaccia, A. and Shah, N. Optimal aggregation of uncertain preferences. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 608–614, 2016.

Satterthwaite, M. A. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory*, 10:187–217, 1975.

Thurstone, L. L. A law of comparative judgement. *Psychological Review*, 34:278–286, 1927.

van der Vaart, A. *Asymptotic Statistics*. Cambridge University Press, 1998.

Zuckerman, M., Procaccia, A. D., and Rosenschein, J. S. Algorithms for the coalitional manipulation problem. *Artificial Intelligence*, 173(2):392–412, 2009.

Zuo, B. and Serfling, R. General notions of statistical depth function. *The Annals of Statistics*, 28(2):461–482, 2000.

A. Additional Metrics on $\mathfrak{S}_n$

The Kendall Tau is the metric used all along the main part of the paper, the proportion of misordered pairs,

$$d_\tau(\sigma,\nu) = \frac{2}{n(n-1)} \sum_{i<j} \mathbb{1}_{[\sigma(i)<\sigma(j)]} \mathbb{1}_{[\nu(i)>\nu(j)]} .$$

The Kemeny consensus is the median associated with the Kendall Tau metric.

The Spearman Rho is a normalized quadratic distance between the rank vectors,

$$d_\tau(\sigma,\nu) = \frac{6}{n(n^2-1)} \sum_i (\nu(i)-\sigma(i))^2 . \tag{10}$$

The Borda count is the median associated with the Spearman Rho (*e.g.* see Calauzènes et al. (2013)).

The Spearman footrule is a absolute value distance between the rank vectors,

$$d_\tau(\sigma,\nu) = \sum_i |\nu(i)-\sigma(i)| . \tag{11}$$

B. Notation for Appendix

For the sake of clarity of the proofs, we switch to matrix notation in the appendix. We fix an arbitrary indexation $\{\sigma^{(1)},\dots,\sigma^{(n!)}\}$ of $\mathfrak{S}_n$. Using this indexation, given a metric d on $\mathfrak{S}_n$, we can defined the (symmetric) metric matrix $D = (d(\sigma^{(i)},\sigma^{(j)}))_{i,j\in[n!]}$. Identifying a ranking σ with its corresponding basis vector $\mathbf{e}_i$ s.t. $\sigma = \sigma^{(i)}$, we write for two rankings $\sigma,\sigma',\nu \in \mathfrak{S}_n$,

$$\nu^\top D\sigma := d(\nu,\sigma) \qquad \text{or} \qquad \nu^\top D(\sigma-\sigma') := d(\nu,\sigma) - d(\nu,\sigma') \tag{12}$$

Further, a distribution $p \in \mathcal{M}_+^1(\mathfrak{S}_n)$ on permutation can now be seen as a $n!$-dimensional vector in $\mathbb{R}^{n!}$. This allows to write, for $p \in \mathcal{M}_+^1(\mathfrak{S}_n)$, $\sigma \in \mathfrak{S}_n$,

$$p^\top D\sigma := \mathbb{E}_{\Sigma\sim p}[d(\Sigma,\sigma)] \tag{13}$$

C. Proof: Bound on Breakdown Function for Ranking Medians

C.1. Upper-bound

We first remind Theorem 3.1.

Theorem 3.1. *For $p \in \mathcal{M}_+^1(\mathfrak{S}_n)$, $\sigma_p^\star = \sigma_{d_\tau}^{\mathrm{med}}(p)$ (Kemeny median) and $\delta \geq 0$, if $\varepsilon^+(\delta) \leq 2p(\sigma_p^*)$ then $\varepsilon^\star_{d_\tau,p,\sigma_p^\star}(\delta) \leq \varepsilon^+(\delta)$ with*

$$\varepsilon^+(\delta) = \min_{\substack{\sigma\in\mathfrak{S}_n \\ d_\tau(\sigma,\sigma_p^\star)\geq\delta}} \max_{\substack{\nu\in\mathfrak{S}_n \\ d_\tau(\nu,\sigma_p^\star)<\delta}} \frac{\mathbb{E}_{\Sigma\sim p}\left[d_\tau(\Sigma,\sigma) - d_\tau(\Sigma,\nu)\right]}{d_\tau(\sigma_p^\star,\sigma) - d_\tau(\sigma_p^\star,\nu)} .$$

We re-state the theorem with the matrix notation defined in Appendix B and used all along the appendix.

Theorem C.1. *For $p \in \mathcal{M}_+^1(\mathfrak{S}_n)$, $\sigma_p^\star = \sigma_{d_\tau}^{\mathrm{med}}(p)$ and $S_\delta = \{\sigma \in \mathfrak{S}_n | d_\tau(\sigma,\sigma_p^\star) \geq \delta\}$, if $\varepsilon^+(\delta) \leq 2p(\sigma_p^\star)$, then $\varepsilon^\star_{d_\tau,p,\sigma_p^\star} \leq \varepsilon^+(\delta)$.*

$$\varepsilon^+(\delta) = \min_{\sigma\in S_\delta} \max_{\nu\in N_\delta} \frac{p^\top D_\tau(\sigma-\nu)}{{\sigma_p^\star}^\top D_\tau(\sigma-\nu)} , \tag{14}$$

Proof.

$$\varepsilon^\star_{d_\tau,p,\sigma^\star_p} = \inf\left\{\varepsilon>0 \middle| \sup_{q:\mathrm{TV}(p,q)\leq\varepsilon} d_\tau(\sigma^\star_p,\sigma^\star_q)\geq\delta\right\} \tag{15}$$

$$= \inf\left\{\varepsilon>0 \middle| \exists q, s.t. \mathrm{TV}(p,q)\leq\varepsilon \text{ and } d_\tau(\sigma^\star_p,\sigma^\star_q)\geq\delta\right\} \tag{16}$$

$$= \inf\underbrace{\left\{\varepsilon>0 \middle| \exists q, s.t. \mathrm{TV}(p,q)\leq\varepsilon \text{ and } \operatorname*{argmin}_{\sigma\in\mathfrak{S}_n} q^\top D_\tau\sigma \subseteq S_\delta\right\}}_{=:E} \quad \text{with } S_\delta=\{\sigma\in\mathfrak{S}_n | d_\tau(\sigma,\sigma^\star_p)\geq\delta\} \tag{17}$$

Further, we define $N_\delta = \mathfrak{S}_n \setminus S_\delta$, $\sigma^{\star,\mathrm{rev}}_p$ the reverse of $\sigma^\star_p$, i.e., $\sigma^{\star,\mathrm{rev}}_p(i) = \sigma^\star_p(n-i-1)$ and the *attack* distribution $\bar{q}_\varepsilon = p - \frac{\varepsilon}{2}\mathbb{1}_{[\cdot=\sigma^\star_p]} + \frac{\varepsilon}{2}\mathbb{1}_{[\cdot=\sigma^{\star,\mathrm{rev}}_p]}$ that removes the probability mass from the median to put it on the farthest point. We also define $E = \left\{\varepsilon | \operatorname{argmin}_{\sigma\in\mathfrak{S}_n} \bar{q}_\varepsilon^\top D_\tau\sigma \subseteq S_\delta\right\}$ and $\tilde{E} = \left\{0<\varepsilon\leq 2p(\sigma^\star_p) \middle| \operatorname{argmin}_{\sigma\in\mathfrak{S}_n} \bar{q}_\varepsilon^\top D_\tau\sigma \subseteq S_\delta\right\} \subseteq E\cap(0,2p(\sigma^\star_p)]$. Let $\varepsilon>0$ be such that $\varepsilon\leq 2p(\sigma^\star_p)$. Then

$$\varepsilon\in\tilde{E} \Leftrightarrow \exists\sigma\in S_\delta, \forall\nu\in N_\delta, \bar{q}_\varepsilon^\top D_\tau\sigma \leq \bar{q}_\varepsilon^\top D_\tau\nu \tag{18}$$

$$\Leftrightarrow \exists\sigma\in S_\delta, \forall\nu\in N_\delta, p^\top D_\tau(\sigma-\nu) + \frac{\varepsilon}{2}\left(\sigma^\top D_\tau\sigma^{\star,\mathrm{rev}}_p - \sigma^\top D_\tau\sigma^\star_p + \nu^\top D_\tau\sigma^\star_p - \nu^\top D_\tau\sigma^{\star,\mathrm{rev}}_p\right)\leq 0 \tag{19}$$

$$\Leftrightarrow \exists\sigma\in S_\delta, \forall\nu\in N_\delta, p^\top D_\tau(\sigma-\nu) \leq \frac{\varepsilon}{2}\left((\sigma^\star_p-\sigma^{\star,\mathrm{rev}}_p)^\top D_\tau(\sigma-\nu)\right) \tag{20}$$

$$\Leftrightarrow \exists\sigma\in S_\delta, \forall\nu\in N_\delta, p^\top D_\tau(\sigma-\nu) \leq \varepsilon\left(\sigma^{\star\top}_p D_\tau(\sigma-\nu)\right) \qquad \text{as } \sigma^{\star,\mathrm{rev}\top}_p D_\tau\cdot = \|D_\tau\|_\infty - \sigma^{\star\top}_p D_\tau\cdot \tag{21}$$

$$\Leftrightarrow \exists\sigma\in S_\delta, \forall\nu\in N_\delta, \frac{p^\top D_\tau(\sigma-\nu)}{\sigma^{\star\top}_p D_\tau(\sigma-\nu)} \leq \varepsilon \tag{22}$$

$$\Leftrightarrow \min_{\sigma\in S_\delta}\max_{\nu\in N_\delta} \frac{p^\top D_\tau(\sigma-\nu)}{\sigma^{\star\top}_p D_\tau(\sigma-\nu)} \leq \varepsilon \tag{23}$$

Now, denoting $\varepsilon^+(\delta) = \min_{\sigma\in S_\delta}\max_{\nu\in N_\delta}\frac{p^\top D_\tau(\sigma-\nu)}{\sigma^{\star\top}_p D_\tau(\sigma-\nu)}$, by definition $\varepsilon^+(\delta)$ satisfies Equation (23), which means $\varepsilon^+(\delta)\in\tilde{E}$ iff $\varepsilon^+(\delta)\leq 2p(\sigma^\star_p)$. Thus, if $\varepsilon^+(\delta)\leq 2p(\sigma^\star_p)$, then

$$\varepsilon^+(\delta) = \inf\tilde{E} \geq \inf E = \varepsilon^\star_{d_\tau,p,\sigma^\star_p}. \tag{24}$$

□

C.2. Lower-bound

We first remind Theorem 3.2.

Theorem 3.2. *For $p\in\mathcal{M}^1_+(\mathfrak{S}_n)$, m and d being two metrics on $\mathfrak{S}_n$, $\sigma^\star_p = \sigma^{\mathrm{med}}_d(p)$ and $\delta\geq 0$, we have $\varepsilon^\star_{m,p,\sigma^\star_p}(\delta)\geq\varepsilon^-(\delta)$ with*

$$\varepsilon^-(\delta) = \min_{\substack{\sigma\in\mathfrak{S}_n\\ m(\sigma,\sigma^\star_p)\geq\delta}}\max_{\substack{\nu\in\mathfrak{S}_n\\ \nu\neq\sigma}} \frac{\mathbb{E}_{\Sigma\sim p}\left[d(\Sigma,\sigma)-d(\Sigma,\nu)\right]}{\max_{\sigma'\in\mathfrak{S}_n} d(\sigma',\sigma)-d(\sigma',\nu)}$$

We re-state the theorem with the matrix notation defined in Appendix B.

Theorem C.2. *For $p\in\mathcal{M}^1_+(\mathfrak{S}_n)$, d and m two metrics on $\mathfrak{S}_n$ and $\sigma^\star_p = \sigma^{\mathrm{med}}_d(p)$, we have*

$$\varepsilon^\star_{m,p,\sigma^\star_p} \geq \min_{\sigma\in S_\delta}\max_{\nu\in\mathfrak{S}_n:\nu\neq\sigma} \frac{p^\top D(\sigma-\nu)}{\|D(\sigma-\nu)\|_\infty}, \tag{25}$$

where $S_\delta = \{\sigma\in\mathfrak{S}_n | d_\tau(\sigma,\sigma^\star_p)\geq\delta\}$.

Proof. Let $S_\delta, N_\delta, E, \tilde{E}$ are defined as above.

$$\varepsilon^\star_{m,p,\sigma^\star_p} = \inf\left\{\varepsilon > 0 \middle| \sup_{q:\mathrm{TV}(p,q)\leq\varepsilon} m(\sigma^\star_p, \sigma^\star_q) \geq \delta\right\} \tag{26}$$

$$= \inf\left\{\varepsilon > 0 \middle| \exists q, s.t. \mathrm{TV}(p,q) \leq \varepsilon \text{ and } m(\sigma^\star_p, \sigma^\star_q) \geq \delta\right\} \tag{27}$$

$$= \inf \underbrace{\left\{\varepsilon > 0 \middle| \exists q, s.t. \mathrm{TV}(p,q) \leq \varepsilon \text{ and } \operatorname*{argmin}_{\sigma\in\mathfrak{S}_n} q^\top D\sigma \subseteq S_\delta\right\}}_{=:E} \quad \text{with } S_\delta = \{\sigma \in \mathfrak{S}_n | m(\sigma, \sigma^\star_p) \geq \delta\}. \tag{28}$$

Now,

$$\varepsilon \in E \Leftrightarrow \exists q, s.t. \mathrm{TV}(p,q) \leq \varepsilon \text{ and } \operatorname*{argmin}_{\sigma\in\mathfrak{S}_n} q^\top D\sigma \subseteq S_\delta \tag{29}$$

$$\Leftrightarrow \exists q \in \Delta^{\mathfrak{S}_n}, \mathrm{TV}(p,q) \leq \varepsilon \text{ and } \exists \sigma \in S_\delta, \forall \nu \in \mathfrak{S}_n, q^\top D\sigma \leq q^\top D\nu \tag{30}$$

$$\Leftrightarrow \exists q \in \Delta^{\mathfrak{S}_n}, \mathrm{TV}(p,q) \leq \varepsilon \text{ and } \exists \sigma \in S_\delta, \forall \nu \in \mathfrak{S}_n, p^\top D(\sigma - \nu) \leq (q_- - q_+)^\top D(\sigma - \nu) \tag{31}$$

$$\text{where } q_+ = (q-p)_+ \text{ and } q_- = (p-q)_+$$

$$\Rightarrow \exists q \in \Delta^{\mathfrak{S}_n}, \mathrm{TV}(p,q) \leq \varepsilon \text{ and } \exists \sigma \in S_\delta, \forall \nu \in \mathfrak{S}_n, p^\top D(\sigma - \nu) \leq \|q_+ - q_-\|_1 \|D(\sigma - \nu)\|_\infty \tag{32}$$

$$\Rightarrow \exists \sigma \in S_\delta, \forall \nu \in \mathfrak{S}_n, p^\top D(\sigma - \nu) \leq \varepsilon \|D(\sigma - \nu)\|_\infty \qquad \text{as } \|q_+ - q_-\|_1 \leq \varepsilon \tag{33}$$

$$\Rightarrow \exists \sigma \in S_\delta, \forall \nu \in \mathfrak{S}_n, s.t. \sigma \neq \nu, \frac{p^\top D(\sigma - \nu)}{\|D(\sigma - \nu)\|_\infty} \leq \varepsilon \tag{34}$$

$$\Rightarrow \min_{\sigma\in S_\delta} \max_{\nu\in\mathfrak{S}_n : \nu\neq\sigma} \frac{p^\top D(\sigma - \nu)}{\|D(\sigma - \nu)\|_\infty} \leq \varepsilon. \tag{35}$$

Finally,

$$\varepsilon^\star_{m,p,\sigma^\star_p} = \inf E \geq \min_{\sigma\in S_\delta} \max_{\nu\in\mathfrak{S}_n : \nu\neq\sigma} \frac{p^\top D(\sigma - \nu)}{\|D(\sigma - \nu)\|_\infty}. \tag{36}$$

□

D. Hausdorff Extensions of Kendall Tau

We remind first the Kendall-tau distance, defined by:

$$d_\tau : (\sigma_1, \sigma_2) \in \mathfrak{S}_n \times \mathfrak{S}_n \to \sum_{i<j} \mathbb{1}((\sigma_1(i) - \sigma_1(j))(\sigma_2(i) - \sigma_2(j)) < 0)$$

and the Definitions 3.4 and 3.5 of the Hausdorff extensions of the Kendall tau metric.

Definition 3.4. (NON-SYMMETRIC HAUSDORFF) Let d be a metric on $\mathfrak{S}_n$. The non-symmetric Hausdorff pseudoquasi-metric between two bucket rankings $\pi_1, \pi_2 \in \Pi_n$ is

$$H^{\mathrm{NS}}_d(\pi_1, \pi_2) = \max_{\sigma_2\in\pi_2} \min_{\sigma_1\in\pi_1} d(\sigma_1, \sigma_2).$$

Definition 3.5. ($1/2$-SYMMETRIC HAUSDORFF) Let d be a metric on $\mathfrak{S}_n$. The $1/2$-symmetric Hausdorff metric between two bucket rankings $\pi_1, \pi_2 \in \Pi_n$ is defined by

$$H^{(1/2)}_d(\pi_1, \pi_2) = \frac{1}{2}\left(H^{\mathrm{NS}}_d(\pi_1, \pi_2) + H^{\mathrm{NS}}_d(\pi_2, \pi_1)\right).$$

Proposition D.1. *For any $\pi_1, \pi_2 \in \Pi_n$, the computation cost of $H^{\mathrm{NS}}_{d_\tau}(\pi_1, \pi_2)$ and $H^{(1/2)}_{d_\tau}(\pi_1, \pi_2)$ is $\mathcal{O}(n^2)$.*

The average Hausdorff distance can be expressed with various expressions, necessitating the following notations (see (Fagin et al., 2006)):

1. $\forall\, i \in [\![1, n]\!] \quad \bar{\pi}(i) = \sum_{\sigma \in \pi} \sigma(i)$ is the rank of item i according to weak order π.
2. $S(\pi_1, \pi_2) = \{(i < j) \mid \bar{\pi}_1(i) \neq \bar{\pi}_1(j), [\bar{\pi}_1(i) - \bar{\pi}_1(j)][\bar{\pi}_2(i) - \bar{\pi}_2(j)] < 0\}$ is the set of item pairs $(i < j)$ that are in different buckets in both π_1 and π_2, and that are in different orders in π_1 and π_2.
3. $S(\pi_1 \setminus \pi_2) = \{(i < j) \mid \bar{\pi}_1(i) = \bar{\pi}_1(j)$ and $\bar{\pi}_2(i) \neq \bar{\pi}_2(j)\}$ is the set of item pairs $(i < j)$ such that both items are in the same bucket in π_1 but in different ones in π_2.
4. $\text{prof}(\pi) = (\text{prof}(\pi)_{i,j})_{i<j}$, where $\forall\ i < j, \text{prof}(\pi)_{i,j} = 1/2$ if $\bar{\pi}(i) < \bar{\pi}(j), = 0$ if $\bar{\pi}(i) = \bar{\pi}(j)$ and $= -1/2$ if $\bar{\pi}(i) > \bar{\pi}(j)$. $\text{prof}(\pi)$ is called the profile vector of π.

We have the following equivalent expressions for the average Hausdorff distance:

Proposition D.2 (Average Hausdorff distance)**.**

$$H_K^{(1/2)}(\pi_1, \pi_2) := \#S(\pi_1, \pi_2) + \frac{1}{2}\left(\#S(\pi_1 \setminus \pi_2) + \#S(\pi_2 \setminus \pi_1)\right) \tag{37}$$

$$= \sum_{i<j} \mathbb{1}\left([\bar{\pi}_1(i) - \bar{\pi}_1(j)][\bar{\pi}_2(i) - \bar{\pi}_2(j)] < 0\right) +$$
$$\frac{1}{2}\mathbb{1}\left([\bar{\pi}_1(i) = \bar{\pi}_1(j)]\right)\mathbb{1}\left([\bar{\pi}_2(i) \neq \bar{\pi}_2(j)]\right) +$$
$$\frac{1}{2}\mathbb{1}\left([\bar{\pi}_2(i) = \bar{\pi}_2(j)]\right)\mathbb{1}\left([\bar{\pi}_1(i) \neq \bar{\pi}_1(j)]\right) \tag{38}$$

$$= \|\text{prof}(\pi_1) - \text{prof}(\pi_2)\|_1 \tag{39}$$

Avergage Hausdorff distance - Proof. Let π_1, π_2 be two weak orders associated with buckets $(B_1^1, ...B_{t_1}^1)$ and $(B_1^2, ...B_{t_2}^2)$ respectively. Such buckets are sets of items i forming a partition of $[\![1, n]\!]$ such that $i \in B_k^1$ iif $\bar{\pi}_1(i) = \sum_{k'<k} \#B_{k'}^1 + \frac{\#B_k^1+1}{2}$ (see (Fagin et al., 2006) for a more formal definition). Let us define, as in (Critchlow, 2012; Fagin et al., 2006), $\forall\, i \leq t_1, \forall\, j \leq t_2, \quad n_{i,j} = \#(B_i \cap B_j)$.

Then we have (Critchlow, 2012)[Chapter IV]: $H_K^{(1/2)} = \frac{1}{2}\left(\sum_{i<i', j\geq j'} n_{i,j} n_{i',j'} + \sum_{i\leq i', j>j'} n_{i,j} n_{i',j'}\right)$.

By noting that $2\#S(\pi_1, \pi_2) = \sum_{i<i', j>j'} n_{i,j} n_{i',j'}$ and $2\#S(\pi_1 \setminus \pi_2) = \sum_{i=i', j>j'} n_{i,j} n_{i',j'}$, we derive our first equality. The second equality directly comes from re-expressing the first one. The third equality comes from (Fagin et al., 2006).

□