Fairness in Ranking: Robustness through Randomization without the Protected Attribute

Andrii Kliachkin*, Eleni Psaroudaki†‡, Jakub Mareček§, and Dimitris Fotakis†‡
* Department of Mathematics, Università degli Studi di Padova, Padua, Italy.
† School of Electrical & Computer Engineering, National Technical University of Athens, Greece.
‡ Athena Research & Innovation Center in Information Communication & Knowledge Technologies, Athens, Greece.
§The Department of Computer Science, Czech Technical University in Prague, the Czech Republic.

***Abstract*—There has been great interest in fairness in machine learning, especially in relation to classification problems. In ranking-related problems, such as in online advertising, recommender systems, and HR automation, much work on fairness remains to be done. Two complications arise: first, the protected attribute may not be available in many applications. Second, there are multiple measures of fairness of rankings, and optimization-based methods utilizing a single measure of fairness of rankings may produce rankings that are unfair with respect to other measures. In this work, we propose a randomized method for post-processing rankings, which do not require the availability of the protected attribute. In an extensive numerical study, we show the robustness of our methods with respect to P-Fairness and effectiveness with respect to Normalized Discounted Cumulative Gain (NDCG) from the baseline ranking, improving on previously proposed methods.**

I. Introduction

There has been a great interest in fairness in machine learning [1, e.g.] over the past decade With the coming regulation of AI both in the European Union and the US, the interest is expected to grow substantially. Much of the work on fairness in machine learning has focused on fairness in classification, so far.

Beyond classification, there are first papers on fairness in forecasting [2, e.g.], decision making [3, e.g.], and ranking [4], [5, cf.]. In some ways, one may argue that ranking and decision-making have a much larger impact on humans than most classification problems. Consider, for example, ranking within HR automation: a recruiter may obtain hundreds of applications for a job, and needs to shortlist 10 best candidates for the hiring manager to interview. This task is increasingly being automated (cf. workable.com, linkedin.com), and the ranking algorithms utilized in the online platforms need to be fair. (For the legal perspective, see the AI Act, Annex III, fifth paragraph.) Similarly, recommender algorithms in social media increasingly drive political preferences, and their performance is of paramount importance for the continued existence of political systems of many countries. Many other applications arise in information retrieval [6].

Two complications arise. First, there are multiple measures of fairness of rankings. Proportionate fairness (P-Fairness) may be the best known, but many others are possible. Optimization-based methods for post-processing or aggregating rankings often utilize only a single measure of the fairness of rankings, which may produce rankings that are not unfair with respect to other measures. We call this the challenge of robustness. Moreover, achieving fairness may compromise efficiency, which is usually quantified by normalized discounted cumulative gain (NDCG) or distance from a baseline ranking.

Second, the protected attribute may not be available in many applications. Consider, for example, the HR-automation use case, as above. There, neither the hard-copy resumes, nor the profiles on online platforms such Workable or LinkedIn, contain protected attributes such as gender, religion, or ethnicity. In many legal systems (e.g., France), it is illegal to collect such information. At the same time, both the employer and the vendor of the AI system can be fined, if they are found guilty of direct or indirect discrimination. Following pioneering work on uncertainty in the group membership [7], [8], fairness without demographics has attracted much recent attention in classification [9], [10], [11], [12, e.g.], but has not been studied within fairness in ranking, yet.

Here, we propose a randomized method for postprocessing rankings to improve their fairness. Inspired by approaches of differential privacy [13], where noise is admixed to data to protect privacy, the proposed method admixes Mallow's noise to improve fairness. Thus, our method does not require the availability of the protected attribute to improve fairness of rankings in a way oblivious to the specific protected attributes. At the same time, our method addresses the challenge of robustness. In an extensive numerical study, we show the robustness of our method with respect to P-Fairness and competitive efficiency with respect to Normalized Discounted Cumulative Gain (NDCG) and KT distance from the baseline ranking, improving upon previously proposed methods ApproxMultiValuedIPF [14] and DetConstSort [15], where the latter has been developed by a team at LinkedIn.

Overall, our contributions are:

- a randomized method for post-processing of rankings improving fairness
- a computational study of five algorithms evaluated with respect to P-Fairness and multiple protected attributes and with respect to two different measures of efficiency.

II. Related Work

Geyik et al. [15] introduced the problem of computing efficient rankings under proportionate fairness constraints [16],

[17]. Proportionate fairness (or P-fairness, in short) requires that each protected group is represented by at least a given proportion of individuals in each prefix of the final ranking. On the other hand, efficiency requires that the final ranking is close to an initial quality-optimal ranking, where the individuals appear in non-increasing order of their quality-based scores, without any fairness or other considerations. In [15], efficiency is quantified by NDCG. Geyik et al. [15] presented and experimentally evaluated a few natural heuristics (including DETCONSTSORT, which we compare against in this work) that aim to balance between violating P-fairness and maximizing NDCG.

Subsequently, Wei et al. [14] and Chakraborty et al. [18] presented efficient algorithms for computing P-fair rankings, under a stricter definition of proportionate fairness using both upper and lower bounds on the representation of each protected group, where efficiency is measured with respect to the distance of the final ranking to a given initial ranking. They proved that the best P-fair ranking can be computed in polynomial time for the Kendall tau distance with two protected groups (Algorithm GRBINARYIPF, inspired by mergesort, in [14]) and with any number of protected groups (Theorem 3.4 in [18]), for the Spearman's footrule distance with any number of protected groups (Algorithm APPROXMULTIVALUEDIPF, based on the computation of a minimum weight bipartite matching, in [14]) and for the Ulam distance with a constant number of protected (Theorem 3.10 in [18]).

Wei et al. [14] and Chakraborty et al. [18] also considered the more general problem of aggregating a collection of rankings to a P-fair ranking that minimizes the total distance to the given ones. In a nutshell, their approach is to aggregate the given rankings into a near-optimal ranking with respect to the objective of minimizing the total distance, which is a well studied problem [19], [20] from the viewpoint of polynomial-time approximation algorithms, and then to transform that ranking into a P-fair one using the algorithms above.

An orthogonal direction, extensively studied in computational social choice (see e.g., [21] and the references therein), is to aggregate a collection of rankings, submitted by a set of voters, into a final ranking under proportional representation constraints for any group of sufficiently large and sufficiently cohesive group of voters. For an excellent discussion of a substantial volume of work in this direction under different notions of proportional representation and different objectives, we refer the reader to [21, Section 6].

In this work, we propose and experimentally evaluate the use of Mallows noise [22] in order to produce approximately efficient rankings, starting from a given one and considering the objectives of maximizing NDCG and minimizing the Kendall tau distance, which are also approximately P-fair with respect not only to some given protected subgroups, but also to other (possibly unknown) protected subgroups that are sufficiently large. The Mallows model has received considerable attention from the viewpoints of both modelling and quantifying noise in computational learning tasks that involve rankings, e.g., [23], [24], [25], and of computationally and statistically efficient learning of Mallows distributions [26], [27]. To the best of our knowledge, our work is the first that investigates the use of Mallows noise towards achieving P-fairness with respect to groups possibly defined by unknown protected attributes.

III. TECHNICAL BACKGROUND

A. Preliminaries

Let a set C of d of candidates or items. We denote the symmetric group over d elements with $\mathbb{S}_d$ and, for incomplete rankings, $\mathbb{S}_{\leq d}$. Let $\sigma \in \mathbb{S}_d$ be a **ranking** (or permutation) of the k elements. For $i \in [d]$, we denote $\sigma(i)$ as the position of the i-th element. Each candidate $i \in C$ has a **protected attribute** $\mathcal{A}(i)$, that can take any of g values, leading to g **protected groups** of candidates. If $g = 2$ then we have a binary protected attribute, leading to 2 groups G_1 and G_2, while if $g > 2$ we have a multi-valued protected attribute leading to more groups $G_1, \ldots, G_g$.

B. Fairness Metrics for Rankings

We want to satisfy some fairness metrics in the suggested ranking. A class of fairness measures, usually referred to as P-fairness, corresponds to the proportional representation of each protected group in the top-k or in every prefix of the top-k of the ranking. The following definitions formalize two different variants of P-fair rankings, namely $(\vec{\alpha}, \vec{\beta})$-$k$ fair rankings and $(\vec{\alpha}, \vec{\beta})$-weak k-fair rankings.

Definition 1 ($(\vec{\alpha}, \vec{\beta})$-$k$ fair ranking, Def. 2.4 of [18]): Consider a set C of d candidates partitioned into g groups $G_1, \ldots, G_g$ and $\vec{\alpha} = (\alpha_1, \ldots, \alpha_g) \in [0,1]^g$, $\vec{\beta} = (\beta_1, \ldots, \beta_g) \in [0,1]^g$, $k \in [d]$. A ranking $\pi \in \mathbb{S}_d$ is said to be $(\vec{\alpha}, \vec{\beta})$-$k$ fair if for any prefix P of π of length at least k and each group $i \in [g]$, there are at least $\lfloor \alpha_i \cdot |P| \rfloor$ and at most $\lceil \beta_i \cdot |P| \rceil$ elements from the group G_i in P, i.e.,

$$\forall_{prefix}\, P : |P| \geq k, \forall i \in [g], \lfloor \alpha_i \cdot |P| \rfloor \geq |P \cap G_i| \geq \lceil \beta_i \cdot |P| \rceil$$

Definition 2 ($(\vec{\alpha}, \vec{\beta})$-weak k-fair ranking, Def. 2.5 of [18]): Consider a set C of d candidates partitioned into g groups $G_1, \ldots, G_g$ and $\vec{\alpha} = (\alpha_1, \ldots, \alpha_g) \in [0,1]^g$, $\vec{\beta} = (\beta_1, \ldots, \beta_g) \in [0,1]^g$, $k \in [d]$. A ranking $\pi \in \mathbb{S}_d$ is said to be $(\vec{\alpha}, \vec{\beta})$-weak k-fair if for the k-length prefix P of π and each group $i \in [g]$, there are at least $\lfloor \alpha_i \cdot k \rfloor$ and at most $\lceil \beta_i \cdot k \rceil$ elements from the group G_i in P, i.e.,

$$\forall i \in [g], \lfloor \alpha_i \cdot k \rfloor \geq |P \cap G_i| \geq \lceil \beta_i \cdot k \rceil$$

A quantitative notion of fairness in rankings, introduced in [14] and referred to as Infeasible Index, accounts for percentage of positions satisfying P-fairness.

Definition 3 (Two-Sided Infeasible Index):

$$\text{LowerViol}(\pi) = \sum_{k=1}^{|\pi|} \mathbb{1}(\exists G_i \in G,\ s.t.\ \text{count}_k(G_i, \pi) < \lfloor \beta_i \cdot k \rfloor$$

$$\text{UpperViol}(\pi) = \sum_{k=1}^{|\pi|} \mathbb{1}(\exists G_i \in G,\ s.t.\ \text{count}_k(G_i, \pi) > \lceil \alpha_i \cdot k \rceil$$

where $\text{count}_k(G_i, \pi)$ is the number of elements of group G_i in the top k positions of ranking π. Then,

$$\text{TwoSidedInfInd}(\pi) = \text{LowerViol}(\pi) + \text{UpperViol}(\pi)$$

Definition 4 (Percentage of P-Fair Positions):

$$\text{PPfair}(\pi, i) = 100 * (1 - \text{TwoSidedInfInd}(\pi, i)/|\pi|)$$

C. Distance Metrics in Rankings

There are several well-known metrics for measuring the similarity between two rankings $\sigma, \pi \in \mathbb{S}_k$, including Spearman distance, Kendall Tau, Kendall's tau coefficient, etc. We define them all for completeness but focus our reporting on the Kendall Tau distance.

Spearman distance measures the total element-wise displacement of π from σ:

$$d_2(\pi, \sigma) = \sum_{i \in [k]} (\pi(i) - \sigma(i))^2$$

Kendall Tau (KT) distance measures the total number of discordant pairs between the two permutations:

$$d_{KT}(\pi, \sigma) = \sum_{i<j} \mathbb{1}\{(\pi(i) - \pi(j))(\sigma(i) - \sigma(j)) < 0\}$$

Kendall's tau coefficient is the normalization of d_{KT} to the interval $[-1, 1]$:

$$k_\tau = 1 - \frac{4 d_{KT}(\pi, \sigma)}{k(k-1)}$$

Kendall's tau coefficient measures the proportion of the concordant pairs between two rankings. The k_τ is 1 when perfect agreement between the two rankings occurs (i.e., the two rankings are the same), and negative when the agreement is less than expected by chance, with -1 when the two rankings have a perfect disagreement.

D. Ranking Quality Measures

In recommendation engines, a model predicts the rank of a list of items based on the search queries. There are relevance scores for each item, and the quality of a ranking is determined by comparing the relevance of the items to the relevance of the items in the optimal, in terms of quality, ranking. Hence, they aim to maximize a quality measure, such as Cumulative Gain (CG), Discounted Cumulative Gain (DCG), Ideal Discounted Cumulative Gain (IDCG), and *Normalized Discounted Cumulative Gain* (NDCG). In this paper, following the work of [15] we utilize the NDCG metric to assess the quality of the suggested rankings, which is defined as:

$$\text{NDCG}(\pi) = \frac{\text{DCG}(\pi)}{\text{IDCG}(\pi)}$$

Specifically, DCG is the sum of gains discounted by rank, computed as :

$$\text{DCG}(\pi) = \sum_{i=1}^{k} \frac{s(\pi(i))}{\log(1+i)},$$

while IDCG calculates the DCG of the ideal order based on the gains, i.e. provides the quality of the optimal ranking π^*

$$\text{IDCG}(\pi) = \sum_{i=1}^{k} \frac{s(\pi^*(i))}{\log(1+i)}.$$

The optimal ranking π^* is to list the items in decreasing order of their scores. For simplicity, we can say that since the DCG of the optimal ranking (i.e., the IDCG) is known, we focus on the optimization of the DCG.

E. Mallows Model

We utilize a distance-based probability model introduced by Mallows [22]. The standard Mallows model $\mathcal{M}(\pi_0, \theta)$ is a two-parameter model with the probability mass function for a permutation π to be equal to:

$$\mathbb{P}_{\pi \sim \mathcal{M}(\pi_0, \theta)}[\pi | \pi_0, \theta] = \frac{e^{-\theta d(\pi, \pi_0)}}{Z_k(\theta, \pi_0)}$$

π_0 is the *central ranking* (also known as location parameter) while θ is the *dispersion parameter* (also known as spread). The probability of any other permutation exponentially decreases as the distance from the central permutation increases, while the dispersion parameter controls how fast this occurs. $Z_k(\theta, \pi_0) = \sum_{\pi \in \mathbb{S}} e^{-\theta d(\pi, \pi_0)}$ is the *normalization constant*, which when $d = d_{KT}$, only depends on the dispersion θ and k and not on the central ranking, hence $Z_k(\theta, \pi_0) = Z_k(\theta)$.

F. Problem Definition

Given a ranking $\pi^* \in \mathbb{S}_d$ (which is supposed to have been optimized based on some quality-based scores that may be unknown to our algorithms) and g protected groups $G_1, \ldots, G_g$ with proportionate fairness requirements α and $\vec{\beta}$ as in definitions 1 and 2, we aim to compute a (possibly approximately and weakly) P-fair ranking π which is as close to π^* as possible. If the quality-based scores behind π^* are unknown, we use the Kendal tau distance to quantify the efficiency of π, as in [14], [18]. Otherwise, we use focus on maximizing DCG (and NDCG) subject to proportionate fairness constraints (as formalized by the Integer Linear Program in Section IV-B.

In Section V, we also evaluate the final ranking with respect to proportionate fairness against unknown protected groups, showing that the use of Mallows noise results in robustness of fairness against unknown protected attributes at the expense of reasonable deterioration of the efficiency objective.

IV. Our Algorithms

A. Post-hoc fairness through Mallows noise

In this work, we utilize to the Mallows model (see Section III-E) as a randomization mechanism that ensures fairness in a way oblivious to the groups. Specifically, given a central permutation and a dispersion parameter, we sample rankings from the corresponding Mallows distribution, and keep the best according to a specific metric.

The central ranking could be either the result of a rank aggregation problem or any ranking in general. For the problem

Algorithm 1 Fair ranking algorithm through Mallows noise

Input: set of items to rank S, number of samples m, criterion c

Output: a ranking π

```
noisy_ranking(m):
    π_0 ← find_central_permutation(S)
    for i ∈ {1, ..., m} do
        samples[i] ← M(θ, π_0)
    end for
    best_ranking ← choose_ranking(c, samples)
    return best_ranking
```

of fair rankings, we pick a weakly fair ranking of candidates ordered by their descending score as the center of our Mallows distribution.

B. Integer Linear Programm

The optimal $\vec{\alpha}, \vec{\beta}$)-$k$ fair ranking with respect to the metrics of DCG and NDCG can be computed via the following Integer Linear Program (ILP).

$$\begin{aligned}
&\max \sum_{i\in[k]}\sum_{j\in[k]} s(i)c(j)x_{ij} \\
&\text{s.t.} \sum_{i\in[k]} x_{ij} = 1 && \forall j \in [k] \\
&\sum_{j\in[k]} x_{ij} \leq 1 && \forall i \in [k] \\
&\lfloor \beta_p \ell \rfloor \leq \sum_{i\in G_p}\sum_{j=1}^{\ell} x_{ij} \leq \lceil \alpha_p \ell \rceil && \forall \ell \in [k], \forall p \in [g] \\
&x_{ij} \in \{0,1\}
\end{aligned}$$

where for the DCG and NDCG metrics, we use $c(j) = 1/\log(1+j)$.

V. Experimental Evaluation

We utilize Mallow's model as a randomization mechanism that ensures fairness in a way oblivious to groups, without a big loss in terms of optimality. Specifically, we perform three types of experiments, two with synthetic data (Sections V-A and V-B) and one with the German Credit dataset [28] (Section V-C). The code for the reproduction of our results can be found in https://github.com/andrewklayk/fairness_with_mallows_distribution.

A. Mallow's model and Infeasible Index

The first experiment aims to evaluate the impact of Mallow's randomization on the Infeasible Index of a ranking. The experimental setup and results follow.

1) Experimental Setup: We analyze a scenario of ten individuals, who belong to two equal-sized groups and create multiple rankings, denoted as σ_{II}, by adjusting the placement of candidates from each group to produce diverse values of the Infeasible Index (II). We then sample rankings from a Mallows distribution using the σ_{II} ranking as the central permutation and varying values of θ, observing the resulting Infeasible Index in the samples.

2) Experimental Results: In Fig. 1 we observe that as the dispersion parameter increases, the Infeasible Index of samples drawn from Mallow's model converges to the Infeasible Index of the central permutation. When the centrals' permutation Infeasible Index is small, as $\theta \to 0$, we notice that the Infeasible Index of the samples drawn from Mallows' model is higher, but without a significant difference. However, when the centrals' permutation Infeasible Index is large, we notice a significant drop in the Infeasible Index of the samples drawn from Mallows' model as $\theta \to 0$.

B. Mallow's model and NDCG

The aim of the second experiment is to evaluate the impact of Mallow's randomization on both fairness and ranking utility in a simple ranking setup.

1) Experimental Setup: We consider two equal-sized groups of five individuals each, where the candidates in the first group are assigned scores $S_1 \sim \mathcal{U}(0,1)$, and in the second group - $S_2 \sim \mathcal{U}(0+\delta, 1+\delta)$, where $\delta = \{0.0, 0.1, \ldots, 0.9, 1.0\}$. We sort the rankings according to the descending candidate scores and sample the Mallows distribution centered on the sorted rankings with difference values of θ. We evaluate the Infeasible Index and NDCG of the samples.

2) Experimental Results: Figs. 3 and 4 depict the experimental results in terms of evaluating how the Infeasible Index and the NDCG change as we start from the corresponding Infeasible Indexes as depicted in a rankings Fig. 2. As for the Infeasible Index, we notice similar behavior as in the Section V-A. We can see that as the dispersion parameter increases the NDCG converges to 1. We note that the NDCG of the central ranking is 1. Therefore we can conclude that there is a trade-off between the NDCG and the Infeasible Index while we sample from Mallow's distribution with different dispersion parameters.

Figs. 3 and 4 depict the experimental results in terms of evaluating how the Infeasible Index and the NDCG change as we start from the corresponding Infeasible Indexes of the central rankings as depicted in Fig. 2. As for the Infeasible index, we notice similar behavior as in Section V-A. We can see that as the dispersion parameter increases, the Infeasible Index converges to the Infeasible Index of the Central Ranking and the NDCG converges to 1, which is the NDCG of the Central Ranking. Therefore, we can conclude that there is a trade-off between the NDCG and the Infeasible Index.

C. Experiments with the German Credit Dataset

In the third experiment, we utilized the German Credit dataset [28] to evaluate how well the Mallows randomization method performs in a practical scenario, where we have partial knowledge regarding some of the protected attributes and aim to evaluate the result in terms of an unknown protected attribute. We conduct a thorough experimental analysis with several state-of-the-art postprocessing algorithms that are designed to ensure fairness in rankings regarding specific attributes. To emulate real-world conditions, we introduce noise

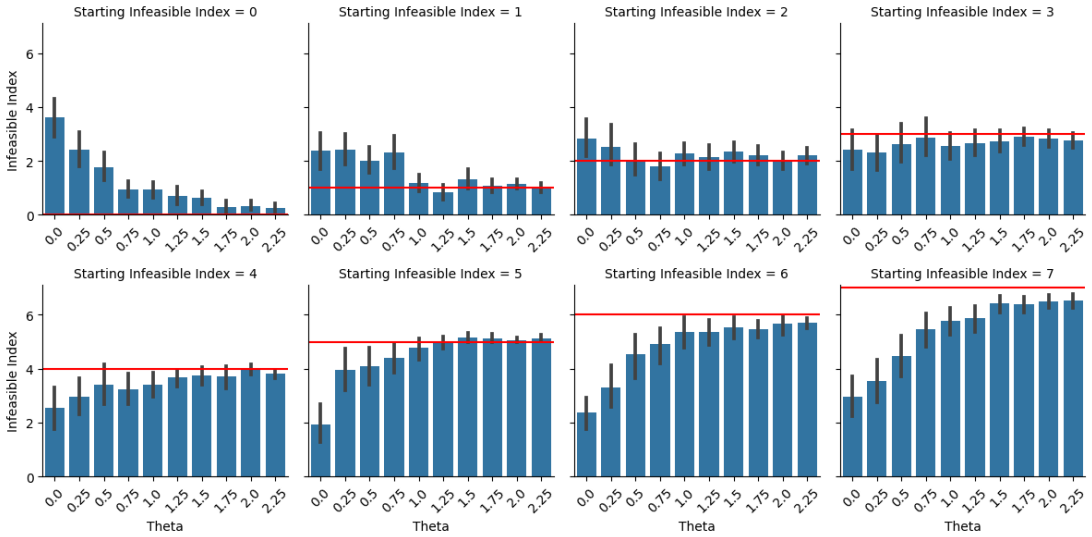


Fig. 1: Mallow's distribution and Infeasible Index. Each subplot corresponds to a different value of the central's ranking Infeasible Index. The Infeasible Index of the central ranking is shown as a red line. The bar plots depict the mean value of the Infeasible Index of the samples from the Mallows distribution centered on the initial ranking with two groups. Confidence intervals were obtained via bootstrapping ($n = 1000$).

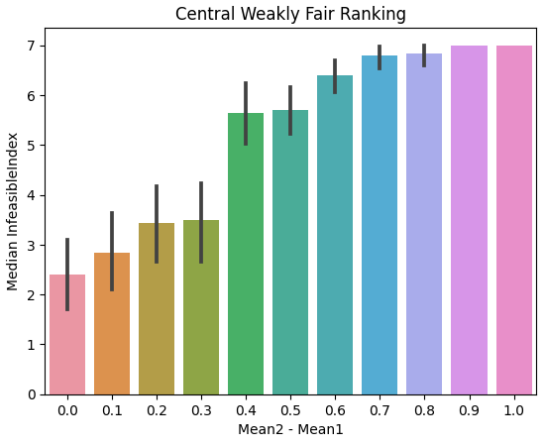


Fig. 2: The Infeasible Index of the Central Ranking, as constructed by sampling from score distributions for each of the two groups (Section V-B). Specifically, the x-axis depicts the difference in means between the score distributions of the two groups. Confidence intervals were obtained via bootstrapping ($n = 1000$).

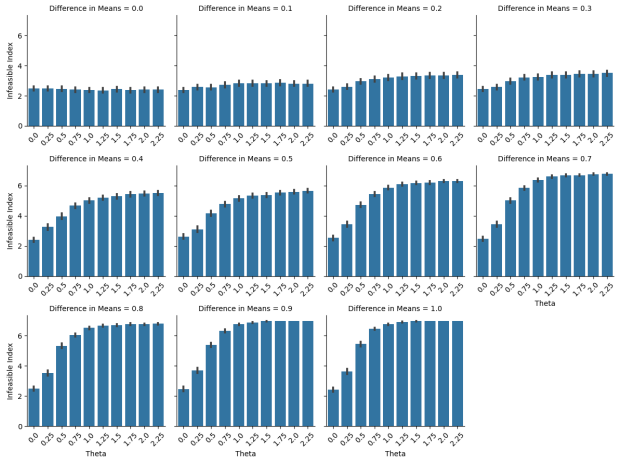


Fig. 3: Mallow's distribution and Infeasible Index. Each subplot corresponds to a difference in means between the score distributions of the two groups. We sample five individuals for each group, where the candidates in the first group are assigned scores $S_1 \sim \mathcal{U}(0,1)$, and in the second group - $S_2 \sim \mathcal{U}(0+\delta, 1+\delta)$, where δ is the difference in means. The subplots depict the mean value of the Infeasible Index of the samples from the Mallows distribution centered on the initial ranking. Confidence intervals were obtained via bootstrapping ($n = 1000$).

into their fairness constraints to simulate imperfect knowledge about group membership.

1) Dataset Description: We utilize the German Credit dataset [28], following the work of [29], [14]. For the ranking of the candidates, we use the Credit Amount attribute. We aggregate the binary attributes Sex and Age into the $Sex{-}Age$ protected attribute with four values and consider the information of this attribute as known, with little or no noise. We evaluate the fairness of the algorithms in terms of a third attribute, named $Housing$, with three values. We regarded the $Housing$ attribute as unknown; therefore, it could not be used as information for any algorithm. The distribution of the groups defined by these attributes is shown in Table I.

2) Experimental Setup: We executed the state-of-the-art ApproxMultiValuedIPF [14], DetConstSort [15] as well as the

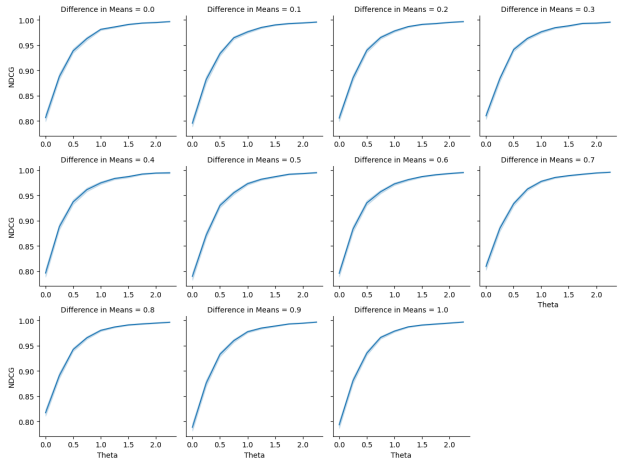


Fig. 4: Mallow's distribution and NDCG. Each subplot corresponds to a difference in means between the score distributions of the two groups. We sample five individuals for each group, where the candidates in the first group are assigned scores $S_1 \sim \mathcal{U}(0,1)$, and in the second group - $S_2 \sim \mathcal{U}(0+\delta, 1+\delta)$, where δ is the difference in means. The subplots depict the mean value of NDCG of the samples from the Mallows distribution centered on the initial ranking. Confidence intervals were obtained via bootstrapping ($n = 1000$).

TABLE I: Distribution of groups defined by Age, Sex, and Housing in the German Credit dataset.

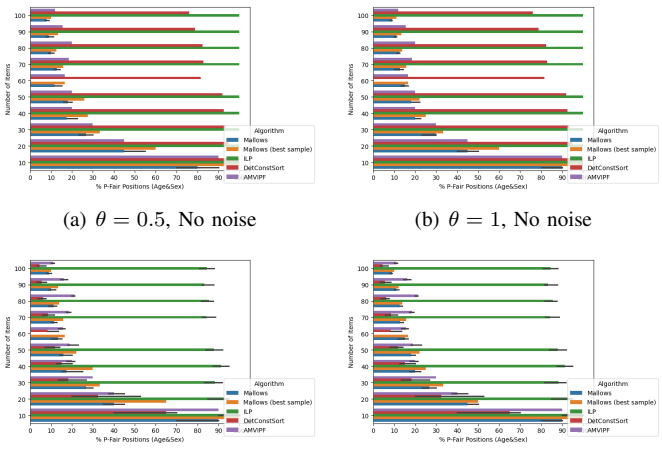
Age-Sex	Housing			Total
	free	own	rent	
<35 - female	2	131	80	213
<35 - male	23	261	51	335
≥ 35 - female	17	65	15	97
≥ 35 - male	66	256	33	355
Total	108	713	179	1000

ILP algorithm described in subsection IV-B, using as input a weakly-p-fair ranking with respect to the combined $Sex-Age$ protected attribute. The algorithms were run in their vanilla version and with some noisy representation constraints on the combined $Sex - Age$ protected attribute. Specifically, we introduced noise into the calculations of the constraints by each of the aforementioned algorithms in the following ways:

- ApproxMultiValuedIPF: we added an independent sample from $N(0,\sigma)$ to each of the weights at the weight calculation step (Algorithm 2, line 2 of [14])
- DetConstSort: we added an independent sample from $N(0,\sigma)$ to each of tempMinCounts (Algorithm 3, line 7 of [15])
- ILP: given $X_{ij}, Y_{ij} \sim |N(0,\sigma)|$, we modified the calculation of constraints for each group G_p such that:

$$\lfloor \beta_p \ell \rfloor - X_{ij} \leq \sum_{i \in G_p} \sum_{j=1}^{\ell} x_{ij} \leq \lceil \alpha_p \ell \rceil + Y_{ij} \quad \forall \ell \in [k]$$

This was done to lessen the probability of making the problem infeasible, while still retaining noise.



(a) $\theta = 0.5$, No noise

(b) $\theta = 1$, No noise

(c) $\theta = 0.5, \sigma = 1$

(d) $\theta = 1, \sigma = 1$

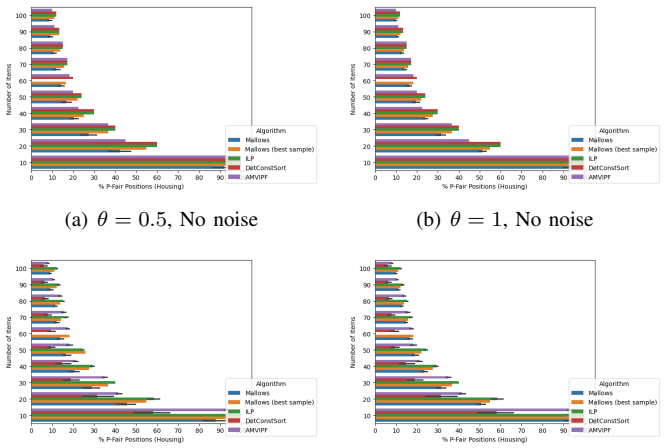
Fig. 5: Rankings constructed with noisy representation constraints on the combined $Age - Sex$ protected attribute from an initial weakly-p-fair ranking with respect to the combined $Age-Sex$ protected attribute. The plots show the median percentage of positions satisfying P-fairness w.r.t. the $Age-Sex$ protected attribute. Confidence intervals were obtained via bootstrapping ($n = 1000$). In Subfigure (a) the θ parameter of the Mallows distribution is set to 0.5, and no noise is added to the constraints. In Subfigure (b) $\theta = 1$ and no noise is added to the constraints. In Subfigure (c) $\theta = 0.5$ and Gaussian noise $\xi \sim \mathcal{N}(0,1)$ is added to the constraints. In Subfigure (d) $\theta = 1$ and Gaussian noise $\xi \sim \mathcal{N}(0,1)$ is added to the constraints.

We repeated this step 15 times to reliably measure the effect of the noise.

We also run the Mallows algorithm on the same weakly-p-fair ranking, using the weakly-p-fair ranking as a central ranking and dispersion parameters of 0.5 and 1, taking 1 or the best of 15 samples.

We evaluate the fairness of the output rankings using the Infeasible Index with respect to the $Housing$ attribute and the utility of the output rankings using NDCG. The experiment is repeated for rankings of size 10, 20, 30, 40, 50, 60, 70, 80, 90, 100.

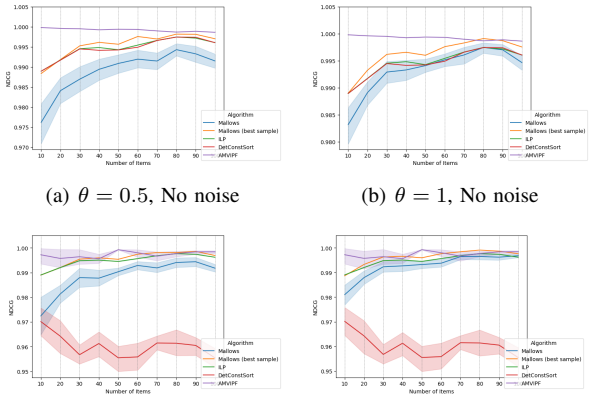
3) Experimental Results: The experimental results are presented in Figs. 6 and 7. Fig. 6 shows the median percentage of positions satisfying P-fairness with respect to the $Housing$ protected attribute. Since DetConstSort, ApproxMultiValuedIPF and the ILP use the $Age - Sex$ protected attribute to create the fair ranking, the resulting ranking fairness has to do with another attribute's distribution, therefore we cannot have any guarantees. For comparison only, we include Fig. 5, so that it is clear how the same ranking can have different fainress scores according to different attribute. We argue that the addition of noise can improve the results regarding different protected attributes, leading to a more balanced output, regardless of the target protected attribute. This approach seems like a compromise among all the different



(a) $\theta = 0.5$, No noise (b) $\theta = 1$, No noise

(c) $\theta = 0.5, \sigma = 1$ (d) $\theta = 1, \sigma = 1$

Fig. 6: Rankings constructed with noisy representation constraints on the combined $Age - Sex$ protected attribute from an initial weakly-p-fair ranking with respect to the combined $Age-Sex$ protected attribute. The plots show the median percentage of positions satisfying P-fairness w.r.t. the $Housing$ protected attribute. Confidence intervals were obtained via bootstrapping ($n = 1000$). In Subfigure (a) the θ parameter of the Mallows distribution is set to 0.5, and no noise is added to the constraints. In Subfigure (b) $\theta = 1$ and no noise is added to the constraints. In Subfigure (c) $\theta = 0.5$ and Gaussian noise $\xi \sim \mathcal{N}(0, 1)$ is added to the constraints. In Subfigure (d) $\theta = 1$ and Gaussian noise $\xi \sim \mathcal{N}(0, 1)$ is added to the constraints.



(a) $\theta = 0.5$, No noise (b) $\theta = 1$, No noise

(c) $\theta = 0.5, \sigma = 1$ (d) $\theta = 1, \sigma = 1$

Fig. 7: Mean NDCG (solid line) and ± 1 standard deviations away from the mean (shaded region) of the output rankings. Confidence intervals were obtained via bootstrapping ($n = 1000$). In Subfigure (a) the θ parameter of the Mallows distribution is set to 0.5, and no noise is added to the constraints. In Subfigure (b) $\theta = 1$ and no noise is added to the constraints. In Subfigure (c) $\theta = 0.5$ and Gaussian noise $\xi \sim \mathcal{N}(0, 1)$ is added to the constraints. In Subfigure (d) $\theta = 1$ and Gaussian noise $\xi \sim \mathcal{N}(0, 1)$ is added to the constraints.

protected attributes that may be present and about which we have no knowledge.

As shown in Fig. 5, under noisy conditions, the Mallows algorithm performs well compared to state-of-the-art algorithms (DetConstSort and ApproxMultiValuedIPF). Fig. 7 shows the mean NDCG (as a solid line) and ± 1 standard deviations (shaded region). We can notice that as the number of items increase, the NDCG score increases for Algorithm 1, and the performance of the best sample from Algorithm 1 approaches the NDCG curve for the ILP.

VI. Conclusions

We have introduced a randomized algorithm for post-processing rankings, demonstrating its efficacy in balancing fairness across multiple attributes through the strategic use of noise. Our findings highlight the potential of noise injection to address fairness concerns, particularly in scenarios where individual attribute information is limited or unavailable.

As future work, we propose developing a systematic methodology for incorporating noise into rankings. This could involve exploring various "noise distributions" or tuning parameters within the noise distribution, such as dispersion in the case of Mallow's model. This approach aims to provide fairness guarantees by leveraging knowledge about the distribution of multiple attributes while respecting individual attribute privacy. Such advancements will contribute significantly to the field of fairness-aware ranking algorithms and their practical applications in real-world settings.

Acknowledgment

This research has been supported by European Union's Horizon Europe research and innovation programme under grant agreement No. GA 101070568 (Human-compatible AI with guarantees).

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

[2] Quan Zhou, Jakub Mareček, and Robert Shorten. Fairness in forecasting of observations of linear dynamical systems. *Journal of Artificial Intelligence Research*, 76:1247–1280, 2023.

[3] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *International conference on machine learning*, pages 1617–1626. PMLR, 2017.

[4] Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, part i: Score-based ranking. *ACM Comput. Surv.*, 55(6), dec 2022.

[5] Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, part ii: Learning-to-rank and recommender systems. *ACM Comput. Surv.*, 55(6), dec 2022.

[6] Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

[7] Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. Equalized odds postprocessing under imperfect group information. In *International conference on artificial intelligence and statistics*, pages 1770–1780. PMLR, 2020.

[8] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. Robust optimization for fairness with noisy protected groups. *Advances in neural information processing systems*, 33:5190–5203, 2020.

[9] Junyi Chai, Taeuk Jang, and Xiaoqian Wang. Fairness without demographics through knowledge distillation. *Advances in Neural Information Processing Systems*, 35:19152–19164, 2022.

[10] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1433–1442, 2022.

[11] Xiaotian Han, Zhimeng Jiang, Hongye Jin, Zirui Liu, Na Zou, Qifan Wang, and Xia Hu. Retiring $\delta\{DP\}$: New distribution-level metrics for demographic parity. *Transactions on Machine Learning Research*, 2023.

[12] Quan Zhou and Jakub Marecek. Group-blind optimal transport to group parity and its constrained variants. *arXiv preprint arXiv:2310.11407*, 2023.

[13] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[14] Dong Wei, Md Mouinul Islam, Baruch Schieber, and Senjuti Basu Roy. Rank aggregation with proportionate fairness. In *Proceedings of the 2022 International Conference on Management of Data*, SIGMOD '22, page 262–275, 2022.

[15] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19. ACM, July 2019.

[16] Sanjoy K. Baruah, N. K. Cohen, C. Greg Plaxton, and Donald A. Varvel. Proportionate progress: A notion of fairness in resource allocation. *Algorithmica*, 15(6):600–625, 1996.

[17] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. Ranking with fairness constraints. In Ioannis Chatzigiannakis, Christos Kaklamanis, Dániel Marx, and Donald Sannella, editors, *Proc. of the 45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*, volume 107 of *LIPIcs*, pages 28:1–28:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018.

[18] Diptarka Chakraborty, Syamantak Das, Arindam Khan, and Aditya Subramanian. Fair rank aggregation. In *Proc. of 35th Conference on Advances in Neural Information Processing Systems (NeurIPS 2022)*, 2022.

[19] Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: Ranking and clustering. *J. ACM*, 55(5):23:1–23:27, 2008.

[20] Claire Kenyon-Mathieu and Warren Schudy. How to rank with few errors. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, California, USA, June 11-13, 2007*, pages 95–103. ACM, 2007.

[21] Yusuf Hakan Kalayci, David Kempe, and Vikram Kher. Proportional representation in metric spaces and low-distortion committee selection. *CoRR*, abs/2312.10369, 2023.

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

[23] Dimitris Fotakis, Alkis Kalavasis, and Konstantinos Stavropoulos. Aggregating incomplete and noisy rankings. In *International Conference on Artificial Intelligence and Statistics*, pages 2278–2286. PMLR, 2021.

[24] Dimitris Fotakis, Alkis Kalavasis, Vasilis Kontonis, and Christos Tzamos. Linear label ranking with bounded noise. *Advances in Neural Information Processing Systems*, 35:15642–15656, 2022.

[25] Dimitris Fotakis, Alkis Kalavasis, and Eleni Psaroudaki. Label ranking through nonparametric regression. In *Proc. of the 39th International Conference on Machine Learning (ICML 2022)*, volume 162 of *Proceedings of Machine Learning Research*, pages 6622–6659. PMLR, 2022.

[26] Róbert Busa-Fekete, Dimitris Fotakis, Balázs Szörényi, and Manolis Zampetakis. Optimal learning of mallows block model. In *Conference on Learning Theory*, pages 529–532. PMLR, 2019.

[27] Xujun Liu and Olgica Milenkovic. Finding the second-best candidate under the mallows model. *Theoretical Computer Science*, 929:39–68, 2022.

[28] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[29] Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs, 2016.