Data Mining and Knowledge Discovery (2024) 38:2186–2222
https://doi.org/10.1007/s10618-024-01029-8

Intersectional fair ranking via subgroup divergence

Eliana Pastor[1] · **Francesco Bonchi**[2,3]

Received: 5 December 2023 / Accepted: 28 April 2024 / Published online: 21 May 2024

Abstract

Societal biases encoded in real-world data can contaminate algorithmic decisions, perpetuating preexisting inequalities in domains such as employment and education. In the fair ranking literature, following the doctrine of affirmative action, fairness is enforced by means of a *group-fairness constraint* requiring "enough" individuals from protected groups in the top-k positions, for a ranking to be considered valid. However, which are the groups that need to be protected? And how much representation is "enough"? As the biases affecting the process may not always be directly observable nor measurable, these questions might be hard to answer in a principled way, especially when many different *potentially discriminated* subgroups exist. This paper addresses this issue by automatically identifying the disadvantaged groups in the data and mitigating their disparate representation in the final ranking. Our proposal leverages the notion of *divergence* to automatically identify which subgroups, defined as combination of sensitive attributes, show a statistically significant deviation, in terms of ranking utility, compared to the overall population. Subgroups with negative divergence experience a disadvantage. We formulate the problem of re-ranking instances to maximize the minimum subgroup divergence, while maintaining the new ranking as close as possible to the original one. We develop a method which is based on identifying the divergent subgroups and applying a re-ranking procedure which is monotonic w.r.t. the goal of maximizing the minimum divergence. Our experimental results show that our method effectively eliminates the existence of disadvantaged subgroups while producing rankings which are very close to the original ones.

Responsible editor: Rita P. Ribeiro.

✉ Eliana Pastor
eliana.pastor@polito.it

Francesco Bonchi
bonchi@centai.eu

[1] Politecnico di Torino, Turin, Italy

[2] CENTAI, Turin, Italy

[3] Eurecat, Barcelona, Spain

Keywords Fair ranking · Intersectional fairness · Divergence

1 Introduction

Ranking is a fundamental primitive in many algorithmic decision-making contexts, such as health (e.g., solid organ transplantation priority list), education (e.g., university admission), or employment (e.g., selection for a job). Typically, different information about an individual might be collected and processed by some machine learning model to produce a final score of "fitness" of a candidate, which then forms the basis of the final ranking. However, bias might be hiding in the underlying data, potentially interfering with the definition of the fitness score and ultimately leading to unfair ranking, which might substantially impact people's lives. Of particular concern are historically disadvantaged groups, whose information in the underlying data might correlate with lower fitness due to historical reasons and preexisting societal inequalities. The growing awareness of the risks associated with algorithmic decision-making has been attracting an increasing research effort toward devising fair ranking systems (Zehlike et al. 2017; Yang and Stoyanovich 2017; Singh and Joachims 2018; Celis et al. 2018; Yang et al. 2019; Celis et al. 2020; García-Soriano and Bonchi 2021; Zehlike et al. 2022; Ekstrand et al. 2023). The bulk of this literature deals with fair ranking as a constrained optimization problem, where the fairness constraint requires that a valid ranking must exhibit in the top-k positions, for any k, a certain fraction of individuals from some protected groups, defined on the basis of sensitive attributes such as ethnicity, gender, or age. A main limitation of this approach is that it needs someone to define the group-fairness constraint. This requires (i) to identify the potentially disadvantaged groups and (ii) to decide which is the minimum representation for each of these groups in the top-k positions. As the potential biases hiding in the underlying data may not always be directly observable nor measurable, these questions might be hard to answer in a principled way, especially in the intersectional case, when many different potentially discriminated subgroups exist.

The notion of *intersectionality* (Crenshaw 1990) refers to individuals belonging to multiple protected groups who may experience a unique disadvantage. Intersectionality is complex because guaranteeing a fair representation or treatment for every single attribute does not guarantee the fair representation of their intersection. As shown by Celis et al. (2018), when each of the elements to be ranked belongs to one and only one group, the constrained optimization problem can be solved exactly in polynomial time. Instead, when each element can belong to more than one group, the problem becomes hard.

To show such complexity, consider the law school admission `LSAT` dataset (Wightman 1998), containing information on law students, and suppose that we rank the students purely based on the LSAT scores. Table 1 reports, for various subgroups, their support (fraction of the population), *divergence* Pastor et al. (2021) (denoted Δ) of the group from the overall population w.r.t. the ranking: a

Table 1 Subgroups from the LSAT dataset with support (fraction of the population), the divergence Δ from the overall average score, and its statistical significance (Welch's t-test)

Subgroup description	Sup	Δ	W-t
ethn = Afr-Am, gender = Female	0.04	− 7.701	37.8
ethn = Afr-Am, gender = Male	0.02	− 6.773	24.2
gender = Female	0.44	− 0.487	7.22
ethn = Caucasian, gender = Female	0.35	0.479	7.3

negative value indicates discrimination, while a positive one indicates that the group is favored (formal definition will be provided later in Section 3). The group {*gender = female, ethnicity = African-American*} is so much discriminated ($\Delta = -7.701$), to make it look as if the whole group {*gender = female*} was discriminated when instead the subgroup {*gender = female, ethnicity = Caucasian*} is not discriminated ($\Delta > 0$). By creating an affirmative action to support the whole group {*gender = Female*}, one produces a positive impact on all women, favoring white women who are not disadvantaged, while damaging already disadvantaged groups, e.g., {*gender = male, ethnicity = African-American*} ($\Delta = -6.773$).

In this paper, we tackle the *intersectional fair ranking* problem in which the individuals to rank have multiple sensitive attributes and thus can belong to one or more protected groups. *Our proposal does not require predefining the group-fairness constraint*, thus avoiding the complication of deciding, a-priori, which are the disadvantaged groups and how much representation each disadvantaged group should have in the top-*k*. Instead, it leverages the notion of divergence to *automatically* identify which subgroups, defined as a combination of known protected attributes, show a statistically significant deviation in ranking utility compared to the overall population. We formulate our intersectional fair ranking problem as a bi-criteria optimization problem. The first criterion aims at mitigating subgroup disparities, thereby promoting equity: in particular, following Rawls' theory of justice (Rawls 1971) which advocates arranging social and financial inequalities to the benefit of the worst-off, requires maximizing the utility of the subgroup with the worst negative and statistically significant divergence.

The second criterion requires that the final ranking is as close as possible to the original one. While improving group fairness is a primary objective, we want to preserve its similarity with the original ranking to maintain the performance of the utility and the consistency in decision-making processes. To solve this fair ranking problem, we develop a method that is based on identifying the divergent subgroups and applying a re-ranking procedure, which is monotonic with the goal of maximizing the minimum divergence. Our approach directly processes the outcome of a ranking process. Our post-processing method enhances the fairness of the ranking without the need to access or modify the original ranker or its generating function. Our experiments show that in all real-world datasets we consider, *our approach can always eliminate the existence of all the disadvantaged groups while maintaining a ranking very similar to the original one*.

2 Related work

Fair ranking. Assessing and ensuring fairness in rankings has recently attracted growing research attention [see recent surveys (Zehlike et al. 2022a, b; Patro et al. 2022; Pitoura et al. 2022)]. Research on fair machine learning can first be divided between two targets of fairness: individual and group fairness Mitchell et al. (2021). Individual fairness requires that similar individuals should be treated similarly by an algorithm process Dwork et al. (2012). Group fairness ensures equal treatment across groups of individuals. With a few exceptions that look at individual and group fairness jointly (García-Soriano and Bonchi 2021; Zehlike et al. 2020), the bulk of the literature on fair ranking focuses on group-level fairness (Asudeh et al. 2019; Celis et al. 2018; Singh and Joachims 2018; Yang et al. 2019; Yang and Stoyanovich 2017; Feldman et al. 2015). Group fairness is assessed and addressed for groups that are protected from discrimination. The group definition is hence typically based on the knowledge of a set of sensitive (or protected) attributes and corresponding protected values Zehlike et al. (2022a). The protected attributes can denote the membership of instances in demographic groups (e.g., female gender or African American ethnicity) that represent a minority or are historically disadvantaged.

Several works (Zehlike et al. 2017; Yang and Stoyanovich 2017; Feldman et al. 2015) consider algorithm fairness for a single sensitive attribute. The work Feldman et al. (2015) aligns the probability distribution of the candidates of the protected group with the non-protected ones. Other researchers have approached fair ranking as a constrained optimization problem, in which the group fairness constraint requires that a minimum fraction of individuals from the protected group to be included among the top-k positions (Zehlike et al. 2017; Celis et al. 2018; Singh and Joachims 2018). We focus on ensuring fairness for multiple protected groups over the entire ranking, and we do not impose explicit group fairness constraints.

Intersectional fair ranking. Intersectionality (Crenshaw 1990) refers to the discrimination that affects individuals who belong to multiple protected groups simultaneously. The fair-ranking survey (Zehlike et al. 2022a, b) notes that, in the presence of multiple protected attributes, we can distinguish approaches that handle the attributes independently, i.e., detecting and mitigating the bias separately for each attribute (Celis et al. 2020, 2018; Yang et al. 2019), and approaches that truly tackle the intersectional problem of dealing with multiple attributes together (Zehlike et al. 2020; Yang et al. 2021; Zehlike et al. 2022).

The Continuous Fairness Algorithm CFAθ Zehlike et al. (2020) aligning the score distributions with the Wasserstein barycenter of all group distributions. The setting handles multiple protected groups and protected attributes. Using the barycenter notion avoids imposing a privileged or majority group as opposed to protected or minority groups. As our approach, the method does not distinguish between (predefined) protected and not-protected groups but only between protected (or sensitive) and non-protected attributes. The target is an adequate treatment for all groups defined by protected criteria. For CFAθ, the fulfillment of group fairness across all groups over multiple protected attributes entails considering a number of subgroups that grows exponentially with the number of attributes and values. We instead control the re-ranking mitigation process automatically by (*i*) considering groups with a size above a user-chosen frequency threshold and (*ii*) directly focusing on problematic groups only.

Yang et al. (2021) study intersectional fairness in ranking by modeling the causal effects of sensitive attributes on other variables and removing these effects to induce fairer rankers. They use counterfactuals to model how the score would change if the sensitive attributes of a given instance were different. The study models each instance as belonging to one specific intersectional subgroup. In our work, each individual can belong to multiple (and overlapping) subgroups (e.g., women and African American women); we focus on identifying whether the subgroups face disadvantage and then implement measures to mitigate their specific disadvantages.

Multi-FAIR Zehlike et al. (2022) extends the top-k algorithm of FA*IR Zehlike et al. (2017) to handle multiple protected groups. The approach ensures that the proportion of protected candidates at any point of the top-k ranking is statistically above a minimum percentage for each protected group, defined via fairness constraints. The statistical test is based on a multinomial distribution. This approach differs substantially from our proposal, as we do not require protected groups to be pre-defined, instead, we focus on detecting subgroups that are disadvantaged by analyzing disparities across the entire ranking, and not only in the top-k positions. Nevertheless, as we show in the experiments in Appendix 1.5, we can adapt our method to deal with top-k settings by suitably tailoring the utility ranking function (i.e., giving a null utility to positions beyond k). Our experiments show that our approach, although not specifically designed for the top-k problem, it is able to address fairness concerns within the top-k subset by enhancing group representation in these top positions.

Anomalous subgroup identification. Fairness assessment and mitigation algorithms typically assume the knowledge of the protected group or set of protected groups (Zehlike et al. 2022; Feldman et al. 2015) to address or generalize to all subgroups over protected attributes. Recently, several works have been proposed to automatically identify the data subgroups associated with a biased behavior in the context of classification (Pastor et al. 2021, 2023; Sagadeeva et al. 2021; Chung et al. 2019) and rankings (Pastor et al. 2021; Li et al. 2023). Both Pastor et al. (2021), Li et al. (2023) focus on identifying groups with a biased representation in the rankings and propose efficient exploration strategies to avoid the exponential enumeration of all subgroups. The approach proposed by Li et al. (2023) automatically detects groups with biased representation in the top-k positions of the ranking using search algorithms based on fairness measures, bounding the global or proportional representation of groups in the ranking by imposing fairness constraints. The work Pastor et al. (2021) is an adaptation of Pastor et al. (2021) for rankings. The approach explores all subgroups that occur sufficiently frequently in the dataset based on a frequency threshold leveraging frequent pattern mining techniques. It identifies overlapping subgroups whose rankings differ, both in terms of advantaged or disadvantaged representation in the ranking. Both approaches Pastor et al. (2021), Li et al. (2023) are limited to detecting groups with a biased representation. Instead, we focus on mitigating the biased representation via method (Pastor et al. 2021) and propose a re-ranking algorithm to mitigate the disadvantage rather than just identify it. Existing intersectional fair-ranking approaches do not inherently support the mitigation of these identified disadvantaged subgroups as they do not directly accommodate overlapping subgroups. As we explain in Sect. 4.4, we adapt the method in Pastor et al. (2021) to our context for automatically identifying the subgroups that have an adequate representation, are disadvantaged in the ranking, and thus require a mitigation process. Our solution mitigates such disparities without the need to impose fairness constraints; instead, it automatically quantifies the necessary degree of mitigation for each disadvantaged group.

3 Preliminaries

We next provide the basic definitions, introduce the notion of utility divergence, and present the problem statement.

Candidates and attributes. We are given a set of candidates $C = \{c_1, \dots, c_n\}$ and for each candidate we have a set of attributes $X = \{X_1, \dots, X_m\}$. We assume that every attribute $X_i \in X$ can be mapped to a discrete,[1] finite set of values $\mathcal{V}_{X_i}$, and we denote with $c(X_i)$ the value of the attribute X_i for the candidate c. Without loss of generality, X are *protected attributes* (e.g., gender, ethnicity, age). If other non-sensitive attributes exist, they are simply disregarded since they are not relevant to our problem. We are also given a relevance function $S : C \rightarrow \mathbb{R}$ that assigns a relevance score to each candidate: this could be, e.g., a fitness score computed by a machine learning algorithm or the result of an aptitude test.

Groups description. Any *pair of attribute-value*, $[X_i = v]$, where $v \in \mathcal{V}_{X_i}$ (e.g., $[gender = female]$), uniquely identifies a group $g(C)$ of candidates: i.e., if $g := [X_i = v]$ then $g(C) = \{c \in C | c(X_i) = v\}$. Similarly, any conjunction of attribute-value pairs uniquely identifies the subgroup of candidates having all the required features. More formally, let G denote the set of all possible attribute-value pairs, and let $\mathbb{G} = 2^G$ denote the set of all possible subsets of G. An element $\{g_1, \dots g_k\} \in \mathbb{G}$ uniquely identifies the group of candidates that have all the attribute-value pairs in $g_1, \dots g_k$, i.e., $\{g_1, \dots g_k\}(C) = \bigcap_{i=1}^{k} g_i(C)$. For instance, consider $g_1 := [gender = female]$ and $g_2 := [ethnicity = Black]$, then:

$$\{g_1, g_2\}(C) = \{c \in C | c(gender) = female \wedge c(ethnicity) = Black\}.$$

In the rest of this paper we will call *subgroup description* any element of G or $\mathbb{G}$, and denote it with g_i; while we call *subgroup* $g_i(C) \subseteq C$ the set of candidates that satisfy the description.

[1] Using numerical values for the subgroup identification may result in subgroups that are too fine-grained. Statistical analyses over small subgroups are prone to fluctuations and may lack statistical significance. To avoid such issues, we map to a discrete set of values for the subgroup identification and the subsequent group-aware re-ranking. Nevertheless, the ranker model or the general function generating rankings operates on the original values.

Adequately represented subgroups. Borrowing from the frequent-pattern terminology, given a subgroup description $g_i \in \mathbb{G}$ we call support the fraction of the overall population belonging to the group: $sup(g_i) = |g_i(C)|/n$. We can also use a minimum support threshold $s \in [0, 1]$ to disregard subgroups that are not adequately represented in C. Given such a threshold s we denote the set of adequately represented subgroups as $\mathbb{G}^s = \{g_i \in \mathbb{G} | sup(g_i) \geq s\}$.

Ranking and utility. A *ranking* r is a permutation of C. In a ranking r each candidate $c \in C$ has a rank $r(c) \in [n]$, where 1 is the top rank.[2] Given a ranking r, the utility function $\gamma : C \rightarrow \mathbb{R}$ represents the utility of ranking candidate c at position $r(c)$. In general, different utility functions can be adopted. For instance, in the fair-ranking literature, to take into account *position bias*, the individual utility is typically combined with a decreasing function of the position in the ranking as in the *discounted cumulative gain* $\gamma(c) = S(c)/\log_2(r(c) + 1)$. Another possibility is to use the ranking position itself, with $\gamma(c) = r(c)$, or of the position in the top-k ranking, i.e., $\gamma(c) = r(c)$ if $r(c) \leq k$ or 0 otherwise. In our setting, as we do not have a fairness constraint, we will always consider only the ranking by decreasing relevance S as natural ranking, and our framework will focus on "adjusting" such relevance function. Thus, for the sake of simplicity of presentation, in defining utility we can drop the ranking and simply define $\gamma(c) = S(c)$.

The utility function $\gamma(g_i(C))$ of a subgroup of candidates $g_i(C)$ is the average utility for the group described by g_i, i.e., $\gamma(g_i(C)) = \frac{1}{|g_i(C)|} \sum_{c \in g_i(C)} \gamma(c)$. When C is obvious from the context, we drop it and use the simpler notation $\gamma(g_i)$.

Divergent subgroups. As we do not have any additional information besides each candidate relevance and attributes, and we know that the relevance score might be biased, the assumption at the basis of our approach is that relevance is distributed uniformly among the population, i.e., no subgroup over protected attributes is deemed a-priori to be more or less skilled. Therefore, we would expect that the relevance score for each subgroup does not deviate substantially from that of the overall population. Still, due to pre-existing societal inequalities and historical biases, we might observe a disparate ranking relevance in subgroups over protected attributes.

We define the divergence of a subgroup described by g_i as a measure of how it deviates from the behavior of the entire population C with respect to the utility function γ, following (Pastor et al. 2021):

$$\Delta_\gamma(g_i) = \gamma(g_i) - \gamma(C) \tag{1}$$

The divergence $\Delta_\gamma(g_i)$ represents how the average value is higher or lower compared with the general population: if it is positive, we say that the subgroup described by g_i is advantaged; if it is negative we say it is disadvantaged. We then measure the statistical significance of the divergence of a subgroup by means of Welch's t-test (Welch 1947).[3] We test the hypothesis that the subgroup and overall population have equal means, computed as follows.

[2] Our approach accounts for ties in the ranking as we consider the utility for the mitigation process.

[3] While there may be some level of dependency between $\gamma(g_i)$ and $\gamma(C)$, we assume it is not substantial, given that g_i is a smaller subset of C. Hence, we expect the results to remain robust to this moderate dependency violation.

$$W\text{-}t(g_i) = \frac{\gamma(g_i) - \gamma(C)}{\sqrt{\sigma^2_{\gamma(g_i(C))}/|g_i(C)| + \sigma^2_{\gamma(C)}/|C|}} \tag{2}$$

where σ^2 is the variance. The W-t statistic is then compared to a critical value t to determine if the hypothesis should be accepted. When the null hypothesis is rejected, we say the subgroup has a statistically significant divergence.

We denote with $\mathbb{D}^{s,t}_{\gamma} \subseteq \mathbb{G}^s$ the set of descriptions of statistically significant disadvantaged subgroups, with $\Delta(g_i) < 0$ and W-$t(g_i) > t$ for each $g_i \in \mathbb{D}^{s,t}_{\gamma}$. Conversely, we denote with $\mathbb{A}^{s,t}_{\gamma} \subseteq \mathbb{G}^s$ the set of descriptions of advantaged subgroups, with positive and statistically significant $\Delta_{\gamma}(g_i)$. Our analysis ignores the subgroups with divergences that are not statistically significant. Adopting a common rule of thumb (Siegel 2012), if Welch's t-statistic for a subgroup described by g_i is larger in absolute value than 2, we reject the null hypothesis, and we identify the divergence of g_i as statistically significant. The values s and t are primary and fixed thresholds of our problem. Without loss of generality, we use $\mathbb{G}$, $\mathbb{D}_{\gamma}$ and $\mathbb{A}_{\gamma}$ when thresholds s and t are clear from the context.

Problem statement. We are now ready to define the objective of our work. We are given as input the set of candidates $C = \{c_1, \dots, c_n\}$, with their attributes $X = \{X_1, \dots, X_m\}$, the relevance function $S : C \to \mathbb{R}$, the utility function $\gamma : C \to \mathbb{R}$, the two thresholds for minimum support s and statistical significance t, as discussed above.

Let r denote the optimal ranking w.r.t. γ: when we assume $\gamma(c) = S(c)$ as we do in the rest of the paper, r is simply the ranking by decreasing relevance, without any group fair representation consideration.

We want to identify a utility score γ^* such that it optimizes two criteria:

1. $\max \min_{g_i \in \mathbb{D}_{\gamma^*}} \Delta_{\gamma^*}(g_i)$;
2. $\min\ dist(r^*, r)$.

where r^* is the ranking induced by γ^*. The first criterion, following Rawls' theory of justice (Rawls 1971) (which advocates arranging social and financial inequalities to the benefit of the worst-off), requires maximizing the utility of the subgroup with the worst negative and statistically significant divergence. The second criterion, requires that the final ranking is as close as possible to the original one. A proper optimization problem can be obtained by optimizing one of the two criteria, while using the other as a constraint. In the method introduced in the next section, we just focus on the first criteria, aiming at having as few disadvantaged groups as possible while using $dist(r^*, r)$ as a measure of quality. In fact, in our experiments in Sect. 5, we show that, in all real-world datasets we consider, we can always eliminate all the disadvantaged groups (i.e., producing $\mathbb{D}_{\gamma^*} = \emptyset$) while maintaining a ranking very similar to the original one (high Kendall's τ similarity).

Table 2 Disadvantaged (top) and advantaged (bottom) subgroups for LSAT dataset

Subgroup description	sup	Original		1 iteration		Mitigation	
		Δ	W-t	Δ	W-t	Δ	W-t
ethn = Afr-Am, gender = Female	0.04	– **7.70**	37.8	0.00	0.0	0.15	0.7
ethn = Afr-Am	0.06	– **7.35**	43.9	– **2.66**	13.8	0.02	0.1
ethn = Afr-Am, gender = Male	0.02	– **6.77**	24.2	– **7.07**	25.2	– 0.19	0.7
ethn = Mexican	0.02	– **3.84**	13.6	– **4.13**	14.6	– 0.16	0.6
ethn = Mexican, gender = Male	0.01	– **3.68**	9.5	– **3.98**	10.2	– 0.13	0.3
ethn = Hisp, gender = Male	0.01	– **3.13**	8.7	– **3.43**	9.5	– 0.24	0.7
ethn = Hisp	0.02	– **2.97**	11.0	– **3.27**	12.1	– 0.09	0.3
ethn = Hisp, gender = Female	0.01	– **2.80**	6.9	– **3.09**	7.6	0.07	0.2
ethn = Other	0.01	– **1.11**	3.1	– **1.40**	4.0	– 0.06	0.2
gender = Female	0.44	– **0.49**	7.2	– 0.11	1.7	– 0.09	1.5
ethn = Asian	0.04	– **0.41**	2.1	– **0.70**	3.6	– 0.01	0.1
ethn = Cauc, gender = Male	0.49	0.96	15.8	0.67	11.2	0.15	2.5
ethn = Cauc	0.84	0.76	14.7	0.46	9.2	0.06	1.2
ethn = Cauc, gender = Female	0.35	0.48	7.3	0.19	2.8	-0.06	1.0
gender = Male	0.56	0.38	6.2	0.09	1.4	0.07	1.2

We report the divergence Δ and the statistical significance of divergence *W-t* for (a) the original utility scores, (b) after one iteration step of Div-Rank and (c) after the full mitigation process

4 Ranking divergence mitigation

We next introduce the notion of divergence mitigation for disadvantaged subgroups. Recall that our goal is to modify $\gamma : C \rightarrow \mathbb{R}$ into γ^* to maximize the minimum divergence across all disadvantaged groups. Intuitively, we want to reduce, in absolute terms, the (negative) divergence of the disadvantaged groups to the extent that their divergence is no longer statistically significant. In the following, we first specify the desired properties of the mitigation step; then, we provide an intuitive transformation of γ satisfying them. Finally, we present an iterative approach for divergence mitigation for fair ranking.

Example 1 (Running example—part 1) To illustrate the divergence mitigation process of disadvantaged groups, we use a running example for the LSAT dataset. The dataset contains information on 21,791 law students. We consider their ethnicity and gender as protected attributes and the LSAT score as the target utility γ for the ranking. The ranking defined via γ induces 11 disadvantaged subgroups $\mathbb{D}_{\gamma}$ according to our definition in Sect. 3 (i.e., groups with a negative and statistically significant divergence), and 4 advantaged subgroups $\mathbb{A}_{\gamma}$. In Table 2 we report the disadvantaged and advantaged groups with their divergence scores Δ_{γ} and the statistical significance of the divergence. Disadvantaged subgroups have an average utility that is statistically significantly different from the one of the entire ranking. This divergence in the average utility indicates that candidates of these subgroups tend to occupy lower positions in the ranking. For instance, the subgroup characterized by {ethnicity =

African-American, gender = Female} has a divergence Δ_{γ} equal to -7.7, i.e., their utility score is by -7.7 lower than the average. In the rest of this section, we will refer again to this running example while describing the process to mitigate the disadvantage of groups $\mathbb{D}_{\mathbb{\gamma}}$.

4.1 Desired properties of the mitigation process

We refer to the term *mitigation* of the divergence of a disadvantaged subgroup as the process that reduces, in absolute terms, its divergence, bringing it close to 0. Consider a disadvantaged subgroup described by $g_i \in \mathbb{D}_{\mathbb{\gamma}}$. We seek a mitigation process of $g_i(C)$ that turns $\gamma(c)$ to $\gamma'(c)$ for each $c \in C$, with $\gamma' : C \rightarrow \mathbb{R}$ be the utility scores after the mitigation step of subgroup $g_i(C)$ is applied. We require the mitigation process of a disadvantaged subgroup $g_i(C)$ to satisfy some properties.

The first, obvious, property states that the mitigation action should reduce the negative divergence of the disadvantaged subgroup described by $g_i \in \mathbb{D}_{\mathbb{\gamma}}$.

Property 4.1 (Mitigate the divergence of a subgroup) *We say that we mitigate subgroup described by* $g_i \in \mathbb{D}_{\gamma}$ *if* $\Delta_{\gamma'}(g_i) > \Delta_{\gamma}(g_i)$[4], *with* $\Delta_{\gamma'}(g_i)$ *being the divergence of* g_i *for scores* γ'. *We say that we fully mitigate a subgroup* g_i *if* $g_i \in \mathbb{D}_{\mathbb{\gamma}}$ *and* $g_i \notin \mathbb{D}_{\mathbb{\gamma}'}$, *i.e., the divergence of* g_i *becomes either non-negative or not statistically significant.*

The second property requires that the mitigation action should increase the minimum divergence among all subgroups. This is to avoid that, e.g., by mitigating the divergence of a subgroup, we worsen the condition of another disadvantaged subgroup. In other terms, we want the mitigation process to be monotonic in improving the minimum divergence.

Property 4.2 (Monotonicity of the mitigation process) *We mitigate the divergence monotonically if, for all subgroup descriptions* $g \in \mathbb{G}^s$, *the mitigated* γ' *has a greater minimum divergence* $\underset{g \in \mathbb{G}}{min}\ \Delta_{\gamma'}(g) > \underset{g \in \mathbb{G}}{min}\ \Delta_{\gamma}(g)$.

The third property requires that the overall ranking utility is maintained in the population. Given that divergence is assessed with respect to $\gamma(C)$, our goal is to prevent changes in its value.

Property 4.3 (Constant average overall behavior) *The average behavior of the overall population C after mitigation is preserved, i.e.,* $\gamma(C) = \gamma'(C)$.

[4] Since $g_i \in \mathbb{D}_{\gamma}$ is a disadvantaged groups, we have $\Delta_{\gamma}(g_i) < 0$. Therefore, in the mitigation process, the goal is to increase its divergence, intending to counteract its disadvantage. We opted for this formulation as it makes the definition of Property 4.2 more general for all subgroups $\mathbb{G}$ and it is not restricted to $\mathbb{D}$. We note that, in absolute terms, the mitigation process consists of reducing the divergence of the disadvantaged subgroup, bringing it closer to 0. Hence, another way for assessing the mitigation of a subgroup described by g_i is if $|\Delta_{\gamma'}(g_i)| < |\Delta_{\gamma}(g_i)|$.

4.2 Mitigation step

Intuitively, we can mitigate the divergence of a disadvantaged subgroup described by $g_i \in \mathbb{D}_\Upsilon$ by increasing the utility scores of the candidates in $g_i(C)$. In this way, however, we would vary the properties of the dataset: the average ranking utility of the overall population changes, i.e., $\gamma(C) \neq \gamma'(C)$. Hence, the property of constant overall behavior would not be satisfied. We can address this issue by correspondingly decreasing the utility scores of other candidates. To avoid introducing a disparate treatment, we decrease the score of *all* candidates in $C \backslash g_i(C)$. The following mitigation function satisfies both this intuition and the desired properties. Let $\tau \in \mathbb{R}_{>0}$ and let $g_i \in \mathbb{D}$ describe a subgroup whose divergence we want to mitigate. We derive $\gamma'(c)$ that mitigates g_i by tweaking the scores $\gamma(c)$ as follows.

$$\gamma'(c) = \begin{cases} \gamma(c) + \tau & c \in g_i \\ \gamma(c) - \frac{\tau \cdot |g_i(C)|}{|C| - |g_i(C)|} & c \notin g_i \end{cases} \tag{3}$$

The term $\frac{\tau \cdot |g_i(C)|}{|C| - |g_i(C)|}$ distributes the reduction in score equally among all candidates in $C \backslash g_i(C)$ to counterbalance the τ given to the candidates in $g_i(C)$. Equation 3 thus satisfies by definition Property 4.3. It is also straightforward that, if $\tau > 0$, we have $\Delta_{\gamma'}(g_i) > \Delta_\gamma(g_i)$ (thus satisfying Property 4.1). Finally, if $\tau = -\Delta_\gamma(g_i)$, after the mitigation step we have that $\Delta_{\gamma'}(g_i) = 0$ (fully mitigated).

Example 2 (Running example—part 2) Consider our example for the LSAT dataset. The subgroup with the highest disadvantage (i.e., highest negative divergence) is described by $g_i = \{\text{ethnicity} = \text{African-American}, \text{gender} = \text{Female}\}$ with $\Delta(g_i) = -7.7$ (first row in Table 2). Fully mitigating the disadvantage of this subgroup entails increasing the score γ of each member of g_i by $\tau = 7.7$. Then, to satisfy Property 4.3 and preserve the average behavior of the overall population, we decrease the score of all candidates not satisfying g_i by $\frac{\tau \cdot |g_i(C)|}{|C| - |g_i(C)|}$. Decreasing by an equal amount ensures that the mitigation process avoids unfairly disadvantaging any specific group.

Unfortunately, the score tweaking of Eq. 3 is not enough to guarantee the monotonicity property, as the corresponding mitigation of the divergence of g_i impacts the divergence $\Delta_{\gamma'}(g_j)$ of other $g_j \in \mathbb{G}$ as follows.

$$\begin{aligned} \Delta_{\gamma'}(g_j) =& \Delta_\gamma(g_j) + \tau \cdot \frac{|\{g_i, g_j\}(C)|}{|g_j(C)|} \\ & - \frac{\tau \cdot |g_i(C)|}{|C| - |g_i(C)|} \cdot \frac{|g_j(C)| - |\{g_i, g_j\}(C)|}{|g_j(C)|} \end{aligned} \tag{4}$$

We can rewrite Eq. 4 dividing it in three cases as in Eq. 5. The first case never decreases the divergence of g_j since $\tau \in \mathbb{R}_{>0}$.

$$\Delta_{\gamma'}(g_j) = \begin{cases} \Delta_\gamma(g_j) + \tau & \text{if } g_j(C) \supseteq g_i(C) \\ \Delta_\gamma(g_j) + \tau \cdot \frac{|g_i(C)|}{|g_j(C)|} - \frac{\tau \cdot |g_i(C)|}{|C| - |g_i(C)|} \cdot \frac{|g_j(C)| - |g_i(C)|}{|g_j(C)|} & \text{if } g_j(C) \subset g_i(C) \\ \Delta_\gamma(g_j) + \tau \cdot \frac{|\{g_i, g_j\}(C)|}{|g_j(C)|} - \frac{\tau \cdot |g_i(C)|}{|C| - |g_i(C)|} \cdot \frac{|g_j(C)| - |\{g_i, g_j\}(C)|}{|g_j(C)|} & \text{otherwise} \end{cases} \tag{5}$$

Also the second case cannot decrease the divergence: with a minor rewriting, we can check that the divergence of g_j would decrease only when $|C| < |g_j(C)|$, which cannot happen in this case as $|C| \geq |g_i(C)| > |g_j(C)|$. However, the divergence of g_j can decrease for the last case when:

$$\Delta_\gamma(g_j) > \\ \left(\Delta_\gamma(g_j) + \tau \cdot \frac{|\{g_i, g_j\}(C)|}{|g_j(C)|} - \frac{\tau \cdot |g_i(C)|}{|C| - |g_i(C)|} \cdot \frac{|g_j(C)| - |\{g_i, g_j\}(C)|}{|g_j(C)|} \right)$$

which can be rewritten as:

$$|g_i(C)| \cdot |g_j(C)| > |\{g_i, g_j\}(C)| \cdot |C|$$

We next discuss how we avoid this case of non-monotonicity.

4.3 Ensuring monotonicity

To ensure the monotonicity of the mitigation process, we want to avoid that, by mitigating the divergence of a subgroup g_i, we decrease the divergence of another subgroup $g_j \neq g_i$ such that $\Delta_{\gamma'}(g_j) < \min_{g \in \mathbb{G}} \Delta_\gamma(g)$. We presented above the case in which the mitigation step of Eq. 3 can decrease the divergence of some other subgroups. We now study the impact of subgroup mitigation on the minimum divergence. Specifically, we define the maximum τ we can apply to mitigate a subgroup g_i such that the mitigation satisfies the monotonicity property.

Let $g_i \in \mathbb{D}_\gamma$ be a subgroup whose divergence we want to mitigate via Eq. 3. Let $a_\gamma = \underset{g_j \in \mathbb{G}}{min} \Delta_\gamma(g_j)$ be the minimum divergence. The maximum $\tau \in \mathbb{R}_{>0}$ for a subgroup $g_j \in \mathbb{G}$ ensuring $\Delta_{\gamma'}(g_j) > a_\gamma$ when mitigating g_i is computed as follows.

$$\tau_{cap}(g_i, g_j, a_\gamma) = \frac{(a_\gamma - \Delta_\gamma(g_j))}{\frac{|\{g_i, g_j\}(C)|}{|g_i(C)|} - \frac{|g_i(C)|}{|C| - |g_i(C)|} \frac{|g_j(C)| - |\{g_i, g_j\}(C)|}{|g_j(C)|}} \tag{6}$$

For $\tau < \tau_{cap}$, we ensure that the divergence of g_j for γ' is greater than a_γ.

Recall that we could break the monotonicity constraint only for the last case of Eq. 5, the only case for which we could decrease divergence. The definition of the maximum threshold τ_{cap} directly derives from imposing $\Delta_{\gamma'}(g_j)$ as equal to a_γ.

Equation 6 defines the mitigation threshold for g_i with respect to a single subgroup g_j. We now define the maximum mitigation we can adopt to ensure the monotonicity of the process across *all* subgroups $\mathbb{G}$. Let $\mathbb{G}^l \subset \mathbb{G}$ be the set of $g_j \in \mathbb{G}$ such that $g_j(C) \not\supseteq g_i(C)$ and $g_j(C) \not\subset g_i(C)$ (last case of Eq. 5). The maximum $\tau \in \mathbb{R}_{>0}$ across all subgroups when mitigating g_i is defined as follows.

$$\tau_{cap}(g_i, a_\gamma) = \underset{g\in\mathbb{G}^l}{min}\, \tau_{cap}(g_i, g_j, a_\gamma) \tag{7}$$

For $\tau < \tau_{cap}(g_i, a_\gamma)$, we ensure the monotonicity of the minimum divergence of the mitigation process.

Example 3 (Running example—part 3) In our running example, when we mitigate g_i = {ethnicity = African-American, gender = Female}, we have that $\tau = \Delta(g_i) = 7.7 < \tau_{cap}(g_i, a_\gamma) = 24.1$. Hence, we can fully mitigate the disadvantage of g_i while satisfying the monotonicity property.

4.4 An iterative mitigation approach

We next introduce Div-Rank, an iterative approach to mitigate the divergence of disadvantaged subgroups. Div-Rank mitigation process involves iteratively applying the mitigation step discussed above. The algorithm (whose pseudocode is outlined in Algorithm 1) iteratively selects the subgroup with the highest disadvantage and mitigates its divergence. Div-Rank ensures that the monotonicity property is satisfied by applying a mitigation lower than the maximum admitted one, defined by Eq. 7.

Algorithm 1 Div-Rank mitigation approach

```
Require: $C$, $X$, $\gamma$, $s$, $t$
Ensure: $\gamma'$
$\gamma' = \gamma$    initialize
$\mathbb{G}^s$, $\Delta_{\gamma'}$ = ExtractSubgroups&Divergence($C, X, \gamma', s$)
    via DivExplorer [12]
$\mathbb{D}^{s,t}_{\gamma'}$ =GetDisadvantagedSubgroups($\mathbb{G}^s$, $\Delta_{\gamma'}$, $t$)
$a_{\gamma'} = \underset{g\in\mathbb{G}^s}{min}\, \Delta_{\gamma'}(g)$
for $g_i \in$ sorted($\mathbb{D}^{s,t}_{\gamma'}$) do
    $\tau_{cap}$ = DefineCap($g_i, \mathbb{G}^s, a_{\gamma'}, \Delta_{\gamma'}$ )    via Eq. 7
    $\tau = min(-\Delta_{\gamma'}(g_i), \tau_{cap})$
    if $\tau > 0$ then
        $\gamma'$ = MitigateSubgroup($C, \gamma', \mathbb{G}^s, g_i, \tau$)    via Eq. 3
        $\Delta_{\gamma'}$ = UpdateDivergence($\gamma', \mathbb{G}^s$)
        $\mathbb{D}^{s,t}_{\gamma'}$=GetDisadvantagedSubgroups($\mathbb{G}^s$, $\Delta_{\gamma'}$, $t$)
        $a_{\gamma'} = \underset{g\in\mathbb{G}^s}{min}\, \Delta_{\gamma'}(g)$
    end if
end for
return $\gamma'$
```

The algorithm requires as input the set of candidates C, the protected attributes X, the utility score γ, the minimum support s, and critical value t for the divergence significance. After initializing the output utility score γ' (Line 1), the first step is the extraction of the adequately represented subgroups and their divergence.

To perform this step, we adopt the subgroup identification algorithm DivExplorer (Pastor et al. 2021) (Line 2). We opt for DivExplorer for two main reasons. First, it leverages frequent pattern mining techniques for the exploration. The approach extracts all subgroups with support greater than the frequency s (we use the FP Growth (Han et al. 2000) frequent pattern mining algorithm in the experiments). This ensures that the subgroups we consider are well-represented in the data and that their average utility scores are statistically significant. Second, it directly defines and integrates the notion of divergence, which is fundamental to our notion of disadvantaged groups. Specifically, it efficiently computes the subgroup divergence and its statistical significance during the subgroup extraction process. Other subgroup discovery approaches (Herrera et al. 2011) could also be considered. We will explore alternative methodologies in future work.

Next, we compute the disadvantaged subgroups as the set of subgroups with negative and statistically significant divergence with respect to the critical value t (Line 3). The minimum divergence across all subgroups is then computed (Line 4). Div-Rank iteratively selects the subgroup $g_i \in \mathbb{D}^{s,t}_{\gamma'}$ with the highest negative divergence from the set of disadvantaged groups (Line 5). In the case of ties, it selects the one with the highest statistical significance.

A disadvantaged subgroup g_i diverges from the overall behavior by $-\Delta_{\gamma'}(g_i)$, i.e., $\gamma'(g_i) - \gamma'(C) = E\{\gamma'(c) \mid c \in g_i\} - E\{\gamma'(c) \mid c \in C\} = -\Delta_{\gamma'}(g_i)$. Setting $\tau = -\Delta_{\gamma'}(g_i)$ ensure the full mitigation of g_i divergence. On the other hand (Line 6), the maximum mitigation a ranking can handle while satisfying the monotonicity constraint is τ_{cap}, defined in Eq. 7. To allow the maximum mitigation while ensuring monotonicity, we set τ as the minimum value among $-\Delta(g_i)$ and τ_{cap} (Line 7). In the case of a positive τ, we can proceed with the mitigation by applying Eq. 3 (Line 9). We then update the divergence scores (Line 10) and the set of disadvantaged subgroups given the updated scores γ' (Line 11). In Line 12, we update the minimum divergence value. The process stops when there are no disadvantaged subgroups with statistically significant divergence ($\mathbb{D} = \emptyset$) or no disadvantaged subgroup could be mitigated without breaking the monotonicity constraint (i.e., $\tau \leq 0$ for all currently disadvantaged subgroups).

Example 4 (Running example—part 4) Considering our running example, we start by mitigating the highest disadvantage, which corresponds to the subgroup described by $g_i = \{\text{ethnicity} = \text{African-American}, \text{gender} = \text{Female}\}$. As detailed in Sect. 4.3, we can fully mitigate its disadvantage by applying Eq. 3 with $\tau = \Delta(g_i)$. Subsequently, we update the divergence scores. We can observe the impact of the first iteration of Div-Rank in Table 2. The divergence of g_i $\{\text{ethnicity} = \text{African-American}, \text{gender} = \text{Female}\}$ is 0 as we fully mitigate its disadvantage. The mitigation step on g_i also reduces the disadvantages of its components, i.e., $\{\text{ethnicity} = \text{African-American}\}$ and $\{\text{gender} = \text{Female}\}$. Specifically, the disadvantage for the group of African-American candidates reduces from −7.35 to −2.66

while the disadvantage for the group of female candidates is mitigated, as the t-value is lower than 2, with Δ from -0.49 to -0.11.

We then proceed by mitigating the new highest disadvantage subgroup. The iterative process continues until $\mathbb{D} = \emptyset$, hence it successfully mitigates all disadvantages.

5 Experiments

We assess Div-Rank w.r.t. its capability of mitigating divergence of disadvantaged groups while producing a ranking as close as possible to the original one. We do this on real-world and synthetic datasets, comparing against baselines in the literature.

The source code of Div-Rank and all the conducted experiments are available at https://github.com/elianap/divrank.

5.1 Experimental setup

Datasets. We use five real-world datasets commonly adopted in the fairness literature: COMPAS (Angwin et al. 2016), LSAT (Wightman 1998), German credit (Lichman 2013), IIT-JEE (Technology IIO 2009), and folktables (Ding et al. 2021). We also leverage a synthetic dataset to further benchmark our approach. Table 3 provides, for each dataset, the protected attributes and their values and the target score used for the ranking. We report a detailed description of the five real-world datasets and the target scores in the "Appendix".

Table 3 Dataset score, number of instances (D), protected attributes and their values

Dataset	Score	D	Protected attributes	Values
LSAT	LSAT score	21,791	`gender`	Female, male
			`ethnicity`	Afric-Am., Amerind., Asian, Cauc., Hisp., Mex., Other, Puert
COMPAS	Inverse recidivism score	6172	`age`	<25, 25–45, ≥45
			`ethnicity`	Cauc., non Cauc
			`gender`	Female, male
German credit	Credit score	999	`age`	young, adult, elder
			`gender`	Female, male
IIT-JEE	Test scores	384,970	`gender`	female, male
			`birth category`	GE, ON, OC, SC, ST, OB
folktables	Income	195,665	`gender`	Female, male
			`ethnicity`	Afric-Am., Alask., Amerind., Amerind./Alask., Asian, Cauc., Two+, Other
Synthetic	Custom	10,000	`a, b, c, d, e`	0, 1

Full details on their preprocessing are available in our repository

Synthetic is a dataset created to have more protected attributes than what can be found in the adopted real-world ones. It has 10,000 instances with 5 attributes with domain {0, 1} and relevance score in [0, 100]. We create the score by randomly setting it and then injecting a controlled bias. Specifically, we decrease the score of all instances with $a = b = 1$ or $c = 1$ with an original score higher than 70. We use the perturbed score as the target score for the ranking. As a result, instances with $a = b = 1$ or $c = 1$ will be associated with lower positions in the ranking. The dataset and the code to generate it are available in our repository.

Evaluation measures. For the evaluation, we adopt divergence-based subgroup measures and ranking performance measures. The former assesses the divergence in the ranking. We consider the minimum and maximum divergence across all subgroups, defined as $\min\Delta_\gamma = \underset{g \in \mathbb{G}^s}{min}\ \Delta_\gamma(g)$ and $\max\Delta_\gamma = \underset{g \in \mathbb{G}^s}{max}\ \Delta_\gamma(g)$ respectively. We also compute the number of disadvantaged subgroups $|\mathbb{D}^{s,t}|$ and advantaged ones $|\mathbb{A}^{s,t}|$. In addition, we consider the Gini index as a measure of inequality (Gini 1921). It quantifies the extent of score distribution among the members of a population. The index ranges from 0 to 1, where 0 represents perfect equality, and 1 represents perfect inequality. The closer the Gini index is to 0, the more equal the distribution of the candidates' scores is.

The ranking performance indexes measure the quality of the derived ranking compared to the original, which maximizes utility and imposes no fairness constraints. In the following, we refer to the ranking based on the original scores as *original ranking*. We consider Kendall's τ and the Normalized Discounted Cumulative Gain loss indexes. Kendall's τ measures the similarity between the mitigated ranking and the original one. The closer the value to 1, the stronger the similarity between the two rankings is. The Normalized Discounted Cumulative Gain (*ndcg*) is a standard measure of ranking quality. It is the weighted summation of candidates' scores in the ranking using a logarithmic discount in the ranking position as weights, normalized to obtain a score between 0 and 1. We compute the loss of *ndcg* (*ndcgLoss*) between the top-$K = 300$ of the original and mitigated ranking. The lower the *ndcgLoss*, the lower the loss when considering the mitigated ranking.

Baselines. Our method automatically detects the disadvantaged groups, with a statistically significant deviation, before applying the needed mitigation. To the best of our knowledge, there is no other method in the literature that approaches intersectional fair ranking without receiving in input the subgroups that need attention. Nevertheless, for the sake of comparison, we adopt three methods from the literature: *it is worth stressing again that these methods need in input the protected groups, while our method automatically identifies the disadvantaged groups to address.*

The technique in Feldman et al. (2015) was not originally proposed for intersectional fair ranking: we use it by taking one protected subgroup for each experiment and the rest of the dataset as non-protected. The method aligns the probability distribution of candidates belonging to a protected group with that of the non-protected group. The approach substitutes the score of a candidate belonging to a protected group with one of a non-protected candidate whose score is in the same quantile, considering protected and non-protected distributions separately. We consider as protected subgroups the ones analyzed in Zehlike et al. (2022). We note that these also generally align with the ones derived by our automatic identification. We include the detailed list in the "Appendix".

CFAθ (Zehlike et al. 2020) aligns the score distribution with the Wasserstein barycenter of all group distributions, with the parameter θ controlling the alignment. We set $\theta = 1$ in all the experiments, which corresponds to enforcing group fairness and imposing equal distribution of scores among the groups. This setting is closer to ours, in which we aim to mitigate disparities in subgroups compared to the overall population. The approach can handle multiple protected groups. We evaluate two configurations. In the former, we consider the same groups specified for the previous baseline, but in this case, the multiple groups are approached *simultaneously*. The latter considers all groups for a given set of protected attributes (e.g., all ethnicities). Note that it entails enumerating the cartesian product of the values of all protected attributes.

Multi-FAIR (Zehlike et al. 2022) is a top-K fair-ranking approach able to handle multiple protected groups. The approach ensures the proportion of protected candidates of the top-k ranking to be statistically above a minimum percentage for each protected group, defined via fairness constraints. We consider a fairness constraint equal to the minimum proportion for each protected group since it is close to our setting. It requires the protected groups to address, and we specify the same as before. Unlike Div-Rank which focuses on disparities in the entire ranking, *Multi-FAIR* targets the top-K. Moreover, *Multi-FAIR* re-ranks the candidates, while Div-Rank adjusts the utility scores that define the ranking. To compare the results, we set K equal to the number of instances. With this configuration, *Multi-FAIR* struggles to terminate within a reasonable time. For this reason, we cannot report the results for this configuration. To still enable a comparison, we consider a setting with a lower K and explore how we can define the utility function γ for Div-Rank to support a top-k fair ranking scenario. We analyze this setting and compare the results in Appendix 1.5.

Parameters. We consider $s = 0.01$ to identify and mitigate disadvantaged groups represented at least 1% in the dataset. We set the critical value t for the statistical significance of divergence to 2 (Siegel 2012).

5.2 Divergence mitigation

This section provides anecdotal and qualitative analysis to illustrate the behaviour of Div-Rank in mitigating the divergence of disadvantaged subgroups. We focus the analysis on the LSAT dataset, where the utility function γ is the LSAT score of each individual. Table 4 (top) reports the top-10 candidates with the highest utility.

As noted in Sect. 4, the original ranking defined via γ induces 11 disadvantaged subgroups (Table 2, top block). We observe that African-American students are associated with lower positions in the ranking. This is accentuated for women. We can quantify their relative contribution to the subgroup disadvantage using the notion of Shapley value (Shapley 1952) from game theory, as adopted in Pastor et al. (2021) for analyzing divergence. Figure 1 (left) shows the Shapley value for the subgroup with the highest disadvantaged. Indeed the highest contribution is of

Table 4 Top-10 positions of the original (top) and mitigated (bottom) rankings

Ethnicity	Gender	Original rank	Mitigated rank
Original ranking			
Cauc	Male	1	209
Cauc	Male	2	210
Cauc	Male	3	211
Asian	Male	4	70
Cauc	Female	5	105
Cauc	Male	6	212
Cauc	Male	7	213
Cauc	Male	8	214
Cauc	Female	9	106
Cauc	Male	10	215
Mitigated ranking			
Afr-Am	Female	777	1
Afr-Am	Male	402	2
Afr-Am	Male	501	3
Afr-Am	Female	1470	4
Afr-Am	Male	798	5
Afr-Am	Male	808	6
Afr-Am	Female	1759	7
Afr-Am	Female	1848	8
Afr-Am	Female	2189	9
Mexican	Male	94	10

LSAT dataset. We report the original and mitigated rank positions

the African-American ethnicity, followed by the female gender. Candidates of the 11 disadvantaged subgroups would experience a disadvantage if the ranking is adopted, for example, to decide who can access an internship program. While this example is purely illustrative, it raises concerns about the use of ranking without fairness constr aints.

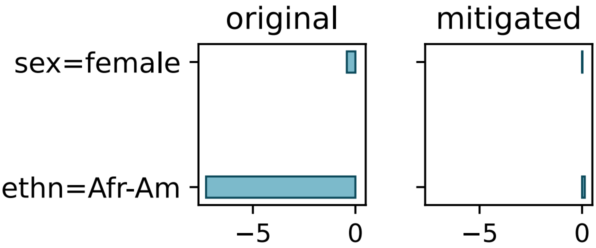

Fig. 1 Shapley value of the highest disadvantaged before (left) and after (right) mitigation. LSAT dataset

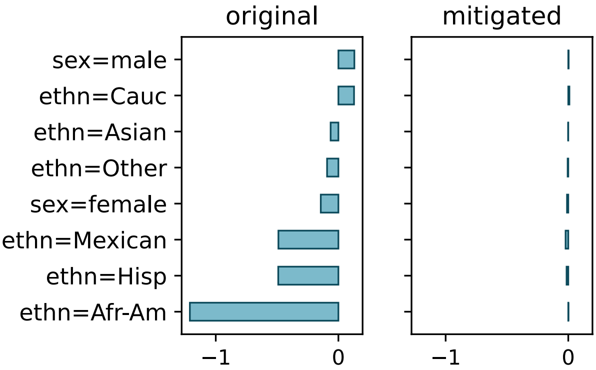

Fig. 2 Global Shapley value before (left) and after (right) mitigation. LSAT dataset

Figure 2 (left) reports the disparate impact of groups on the ranking of individuals in terms of Global Shapley Value (GSV) (Pastor et al. 2021), which measures how much each attribute value contributes to the divergence across *all explored subgroups*. The lower the value, the more the attribute and its value are associated with lower scores and lower positions in the ranking. Women and all ethnicities other than Caucasian are associated with a lower-than-average score. The highest discrepancy is observed for African-American students.

We apply the Div-Rank algorithm to reduce the disadvantage of the ranking. The results are reported in the last block of Table 6. Div-Rank successfully allows mitigating the bias. After mitigation, we have *no disadvantaged* subgroups ($|\mathbb{D}| = 0$). We note that also the number of advantaged subgroups ($|\mathbb{A}|$) reduces from 4 to 1. The minimum subgroup divergence increases from −7.7 to −0.24. The mitigation process, by mitigating the divergence of disadvantaged groups, as a by-product also reduces the advantage of some groups. Indeed, the divergence of the subgroup with the highest advantage decreases substantially: from 0.96 to 0.15. These mitigation results are obtained by producing a re-ranking that is still quite close to the original one (Kendall's $\tau = 0.88$). Table 4 (bottom) reports the top-10 candidates after the mitigation. The first candidate in the mitigated ranking is an African-American female candidate, originally at position 777. This position marked the first occurrence of a candidate of African-American ethnicity and female gender in the original ranking. Similar observations apply to other candidates. For example, the second and third positions are two candidates of African-American ethnicity and male gender, whose original first occurrence was in positions 402 and 501. The top-10 includes in the tenth position a candidate of Mexican ethnicity and male gender, originally in position 94. Therefore, Div-Rank raised the position in the ranking of candidates who belong to disadvantaged groups in the original ranking.

We further analyze the impact of the mitigation process on ranking positions in Table 5. The table details each disadvantaged and advantaged group, including the minimum, 25th percentile, 50th percentile, and maximum ranking positions, before and after mitigation. In the original ranking, candidates from disadvantaged groups typically held much lower positions than those from advantaged groups. For instance, in the original ranking, the first female African-American candidate occupied position 777, the first male African-American candidate held position 402, and the first Mexican candidate was in position 94. In contrast, the top position was held by a Caucasian male, with the first Caucasian female candidate ranking at position 5. This discrepancy is particularly evident at the 25th and 50th percentiles. Specifically, for African American candidates, the 25th percentile rankings ranged from approximately 16,000 positions for males to 17,000 for females, with 50th percentile rankings around 19,500 and 20,000, respectively. In contrast, Caucasian candidates had rankings ranging from 4500 to 5000 for the 25th percentile and approximately 10,000 for the 50th percentile, indicating significant disparities. Following mitigation, the 25th and 50th percentiles for both disadvantaged and advantaged groups fall within the range of 5000 and 10,000–11,000 positions, indicating a more equitable distribution across the ranking.

Table 5 also includes each group's representation in the top-K positions [with K = 300, in line with the value adopted in Zehlike et al. (2022)]. Before the mitigation, most of the candidates belonged to a single ethnicity (94% Caucasians and 60.67% male Caucasians), and none were African Americans. After the mitigation, we have a representation of all ethnicities in the top 300.

We further analyze qualitatively the impact of the mitigation for specific subgroups and overall across subgroups. Consider the subgroup with the highest disadvantage for the original ranking. The divergence goes from −7.7 (first in Table 2) to zero (0.1); the contribution of the two terms after mitigation is reported in Fig. 1 (right). Figure 2 (right) shows the impact on the Global Shapley value of Div-Rank mitigation. The GSV reduces for *all* terms to negligible values. Hence, no term is highly associated with divergence.

Further insights in the mitigation process. The mitigation algorithm stops when either no statistically significant disadvantaged subgroups remain in the ranking or further mitigation would violate the monotonicity constraint. In all experiments on the four datasets, we met the first condition. Hence, we obtain a ranking in which no subgroups face a statistically significant disadvantage. In this section, we analyze the iterative mitigation process, still for the LSAT dataset, but similar considerations apply to the others.

Figure 3 (left) shows the minimum subgroup disadvantage and Kendall's τ during the iteration of the mitigation process. Figure 3 (right) shows the number of disadvantaged and advantaged subgroups. The minimum divergence monotonically increases (Fig. 3 (left)). At iteration 12, the minimum divergence is −0.24, and no subgroup is statistically significantly disadvantaged. As expected, Kendall's τ decreases while we proceed with the mitigation. The more we adjust the utility of the candidates to mitigate the divergence of subgroup disparities, the more the mitigated ranking deviates from the original one. Still, users can easily control and bind the dissimilarity to the original ranking while reducing the disadvantage in the iterative process by imposing a threshold on ranking quality indexes as a further stopping condition.

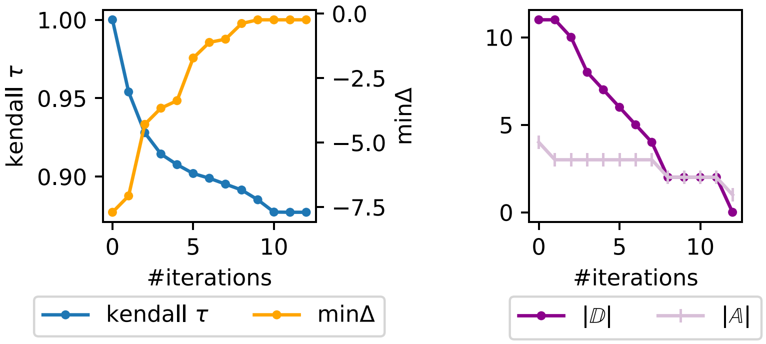

Fig. 3 Kendall's τ and minimum divergence (left) and number of advantaged and disadvantaged subgroups (right) over the steps of the mitigation process. LSAT dataset

5.3 Comparison with baselines

We evaluate Div-Rank on the four datasets. We compare ranking results with (*i*) the original (score-based) ranking, (*ii*) ranking produced for Feldman et al.'s technique (Feldman et al. 2015), and (*iii*) *CFAθ* (Zehlike et al. 2020) and we experimentally evaluate (iv) Multi-FAIR (Zehlike et al. 2022). As explained before, a direct and fair comparison is not possible. The method by Feldman et al. (2015) considers a disadvantaged group at a time, while *CFAθ* (Zehlike et al. 2020) can handle multiple subgroups at once; Multi-FAIR (Zehlike et al. 2022) can handle multiple groups as well, but it is designed explicitly for fair top-K ranking. For all these methods, however, we need to specify the protected groups as part of the input. Instead, Div-Rank automatically identifies the disadvantaged groups (i.e., groups with a negative and statistically significant divergence) to mitigate.

Tables 6, 7, 8, 9, 10, and 11 in the appendix compare the results for the LSAT, Synthetic, COMPAS, German credit, IIT-JEE and folktables datasets. We first remark that, for all experiments, Div-Rank is able to mitigate the disadvantage of *all* disadvantaged subgroups.

As mentioned, Multi-FAIR (Zehlike et al. 2022) addresses top-K fair ranking. To enable the comparison, we set K equal to the number of instances of the dataset. With this configuration, however, the algorithm did not terminate in a practical time compared to our and the other approaches (we interrupted its execution after two days of computation). The unfeasibility of the computation can be traced back to the time complexity which increases exponentially with the number K. Instead, our approach mitigates the disadvantage of all groups in at most 20 s (further details in the following performance analysis). We explore a setting with $K \ll |C|$ in Appendix 1.5.

Consider Table 6 and the LSAT dataset. As also revealed by our analysis in Fig. 2 (left), women are associated with lower positions in the ranking. Hence, we first compare *feldman* and *CFAθ*'s ability to mitigate the disparate impact on the female gender with Div-Rank. In the latter case, we just provide gender as a sensitive

Table 5 Disadvantaged (top) and advantaged (bottom) subgroups for LSAT dataset

Subgroup description	Original					Mitigated				
	Min	25%	50%	Max	%-K	Min	25%	50%	Max	%-K
ethn = Afr-Am, sex = female	777	17,027	20,163	21,789	0.00	1	4819	10,868	21,737	11.00
ethn = Afr-Am	402	16,581	19,968	21,791	0.00	1	4883	10,970	21,785	18.67
ethn = Afr-Am, sex = male	402	15,982	19,598	21,791	0.00	2	4896	10,980	21,785	7.67
ethn = Mexican	94	11,319	17,029	21,778	0.67	10	4908	11,000	21,766	4.00
ethn = Mexican, sex = male	94	11,294	17,014	21,754	0.67	10	4908	10,997	21,731	3.33
ethn = Hisp, sex = male	290	9598	15,354	21,727	0.33	12	4780	10,805	21,709	2.33
ethn = Hisp	125	9074	15,366	21,790	1.00	12	4779	10,810	21,791	4.67
ethn = Hisp, sex = female	125	8295	15,377	21,790	0.67	13	3471	10,819	21,791	2.33
ethn = Other	150	6199	12,181	21,783	1.00	34	4039	10,783	21,787	2.67
sex = female	5	6110	11,602	21,790	36.00	1	5319	11,142	21,791	51.00
ethn = Asian	4	5394	11,576	21,735	3.00	49	4058	10,027	21,768	6.67
sex = male	1	4973	10,344	21,791	63.67	2	5520	10,612	21,789	48.67
ethn = Cauc, sex = female	5	5474	10,532	21,729	33.33	105	5347	11,153	21,778	32.67
ethn = Cauc	1	4963	10,028	21,773	94.00	105	5522	10,752	21,789	63.00
ethn = Cauc, sex = male	1	4596	9690	21,773	60.67	209	5552	10,554	21,789	30.33

We report the min, 25th, and 50th percentiles, and max ranking positions for the original rank and after the mitigation of Div-Rank. %-K reports the percentage of representation of the group in the top-K positions, with K = 300

Table 6 Original and re-ranking results for Div-Rank, *CFAθ*, and *feldman*

Method	Prot. attr	Prot. groups	$\|\mathbb{A}\|$	$\|\mathbb{D}\|$	min Δ	max Δ	gini	Kendall's τ	ndcg Loss
Original	gender		1	1	– 0.49	0.38	0.12	1.00	0.00
Div-Rank		–	0	**0**	– 0.00	0.00	0.12	0.98	0.00
CFAθ		{female}, {male}	0	**0**	– 0.02	0.02	0.09	0.98	0.00
feldman		{female}	0	**0**	– 0.05	0.06	0.12	0.97	0.00
original	ethn		1	5	– 7.35	0.76	0.12	1.00	0.00
Div-Rank		–	0	**0**	– 0.22	0.07	0.13	0.89	0.03
CFAθ		{Afr-Am}, {*other*}	1	5	– 4.75	0.35	0.09	0.93	0.00
CFAθ		{e}, ∀ e ∈ ethn	0	1	– 0.35	0.04	0.09	0.88	0.00
feldman		{Afr-Am}	1	4	– 4.30	0.30	0.11	0.92	0.00
feldman		{≠Cauc}	3	1	– 2.52	2.74	0.13	0.89	0.00
Original	ethn,gender		4	11	– 7.70	0.96	0.12	1.00	0.00
Div-Rank		–	1	**0**	– 0.24	0.15	0.13	0.88	0.03
CFAθ		{Afr-Am, female}, {*other*}	3	12	– 7.68	0.74	0.09	0.95	0.00
CFAθ		{e, Female} ∀ e ∈ ethn, {*other*}	2	8	– 7.79	0.61	0.09	0.93	0.00
CFAθ		{e, g} ∀ e ∈ ethn∀ g ∈ gender	0	1	– 0.37	0.05	0.09	0.87	0.00
feldman		{Afr-Am, female}	3	11	– 7.07	0.67	0.11	0.95	0.00
feldman		{≠Cauc, female}	5	8	– 7.19	3.21	0.11	0.93	0.00

LSAT dataset. Each block specifies the considered set of protected attributes (prot. attr.). The subgroups column indicates the considered subgroups (prot. groups). The subgroup {*other*} indicates the *remainder* group. We specify with '–' for Div-Rank since no subgroup specification is required

Table 7 Original and re-ranking results

Method	Prot. attr	Prot. groups	$\|\mathbb{A}\|$	$\|\mathbb{D}\|$	min Δ	max Δ	gini	Kendall's τ	ndcg Loss
original	a,b,c,d,e		66	97	– 5.58	8.40	0.31	1.00	0.00
Div-Rank		-	11	**0**	– 1.80	4.76	0.28	0.93	0.00
CFAθ		$\{v_a, v_b, v_c, v_d, v_e\}$, $\forall v_a \in [0,1]$, $\forall v_b \in [0,1]$, $\forall v_c \in [0,1]$, $\forall v_d \in [0,1]$, $\forall v_e \in [0,1]$	0	**0**	– 0.46	0.44	0.3	0.91	0.07
feldman		{1,1,1,*,*}	53	77	– 5.36	7.91	0.31	0.98	0.02
feldman		{1,1,1,1,1}	61	93	– 5.05	8.22	0.31	0.99	0.01

Synthetic dataset. For *feldman*, we test two configurations: (i) a = 1, b = 1, and c = 1 as protected subgroup, regardless of d and e values (i.e., {1,1,1,*,*}) and (ii) all values equal to 1 as protected (i.e., {1,1,1,1,1}) attribute (i.e., without directly indicating 'female' as protected). They all mitigate the divergence of the subgroups, ($|\mathbb{D}| = 0$) with Div-Rank achieving slightly better results in terms of Kendall's τ compared to *feldman*, while *feldman* achieves the lowest Gini index.

The second block reports the results when considering only the ethnicity attribute. For *feldman*, we consider two configurations. The first considers only 'African-American' candidates as protected. *feldman* mitigates this subgroup, while other ethnicities still experience a disadvantage ($|\mathbb{D}|$ from 5 to 4). The second configuration groups all ethnicities except 'Caucasian' into a single protected group. *feldman* mitigates 4 out of 5 disadvantaged subgroups. Yet, African-American candidates still are associated with lower positions. Moreover, the number of advantaged subgroups rises from 1 to 3. We then apply *CFAθ* considering all ethnicities as input groups. The number of disadvantaged subgroups drops to 1 and the advantaged ones drops to 0. Div-Rank eliminates the presence of both disadvantaged and advantaged subgroups related to the ethnicity attribute with a high Kendall's τ.

We then apply the mitigation for the intersection of both *ethnicity* and *gender*. For *feldman*, we consider as protected subgroups African-American women (which we previously observed to be the most divergent one) and non-Caucasian women. For the latter group, the number of disadvantaged subgroups drops from 11 to 8. For *CFAθ*, we consider three configurations. The first considers only the subgroup of African-American women as protected. The approach is not effective in this case. The second considers the intersection of all ethnicities in the dataset (8 in total) and the female gender. The approach reduces the number of disadvantaged subgroups $\mathbb{D}$ from 11 to 8. The third considers all subgroups at the intersection of ethnicity and gender values (16 in total). The number of disadvantaged groups drops to 1. Div-Rank mitigates all disadvantaged subgroups. The mitigation results has still high Kendall's τ, but a slightly higher Gini index.

Similar results and considerations apply also for COMPAS (Table 8), German credit (Lichman 2013) (Table 9), IIT-JEE (Technology IIO 2009) (Table 10), and folktables (Ding et al. 2021) (Table 11) datasets reported in the appendix. In all cases, Div-Rank reduces to 0 the number of disadvantaged subgroups. For COMPAS and German credit datasets, also *CFAθ* mitigates disadvantage when all subgroups at the intersection of multiple attributes are considered. Still, Div-Rank results have higher Kendall's τ (0.68 vs. 0.66 for COMPAS, 0.96 vs. 0.92 for German credit) and generally lower execution time (as we discuss in the following). The experimental results show the ability of Div-Rank to mitigate divergence and remove disadvantaged subgroups without specifying the protected subgroups of interest.

Div-Rank identifies the subgroups to mitigate *automatically*. This ability is particularly relevant when the number of protected attributes increases. Applying the mitigation process of *CFAθ* for all possible subgroups becomes computationally expensive. Conversely, Div-Rank considers only adequately represented subgroups and iteratively mitigates the disadvantaged ones. We demonstrate this by considering the Synthetic dataset and all its five attributes as protected. Table 7 shows the mitigation results. Given that the dataset is synthetic, we know that instances with $a = b = 1$ or $c = 1$ are located in lower positions in the ranking. We leverage this knowledge for selecting the protected values for *feldman* mitigation. In a real-case scenario, we should conduct preliminary analyses of the ranking. Yet, *feldman* mitigation still yields numerous disadvantaged subgroups. This again highlights the need to mitigate multiple subgroups simultaneously. Div-Rank and *CFAθ* reduce to 0 the number of disadvantaged subgroups. However, for *CFAθ*, we need to enumerate and consider all possible subgroups of the protected attributes. Addressing the complete enumeration makes the mitigation process computationally expensive, requiring 7.5 h. Div-Rank instead automatically identifies the subgroups to mitigate and efficiently mitigates all disadvantaged subgroups in 20 s. Moreover, it achieves higher Kendall's τ and lower Gini index. While we have a higher alignment with the original ranking, CFAθ demonstrates greater improvement in reducing the subgroup divergence (lower minimum and maximum divergence in absolute terms). We attribute this to CFAθ considering all subgroups across protected attributes in this experiment, even those without statistically significant disadvantages. In contrast, our method focuses only on mitigating statistically significant disadvantages, thus reducing the alterations to the original ranking and maintaining a more targeted intervention.

To further illustrate the effectiveness of Div-Rank in handling an increased number of attributes, we conducted experiments considering all attributes of the three evaluated datasets, thus disregarding the distinction between protected and non-protected ones. Note that this setting is for demonstration purposes, as our goal is fair ranking and reducing disparities among subgroups over protected attributes. We report the results and a detailed discussion in Sect. 1.4 of the Appendix. The findings reveal that both *feldman* and *CFAθ* fail to mitigate disparities for all disadvantaged groups. In contrast, Div-Rank demonstrates its effectiveness by reducing the number of disadvantaged groups to zero.

Computational performance analysis. We performed the experiments on a Ubuntu server with Intel Xeon CPU 12 cores, 32GB memory. Div-Rank required 7 s, 11 s, 0.5s, 4.9 min, 3.05 min and 20 s for LSAT, COMPAS, German credit, IIT-JEE, folktables and Synthetic respectively, when considering all sensitive attributes. Execution time is also low for the *feldman* method: the maximum running time is 4.5s, 1.5s, 0.1s, 8.5 min, 4.4 min, and 1.5s, respectively. However, the results show that it is not effective in reducing the disadvantaged subgroups to zero. *CFAθ* requires more time, especially when the number of protected groups increases. The maximum running time for LSAT, COMPAS, German credit, IIT-JEE, folktables and Synthetic is respectively 13.5 min, 2.2s, 13.7s, 5 s, 13.4 min and 7.5 h.

Sensitivity analysis. Div-Rank identifies and mitigates the disadvantaged subgroups represented at least s% in the dataset. We vary the minimum support s, considering values 0.001, 0.005, 0.01 (as in the previous experiments), 0.1, 0.2, and 0.3, and we study the impact on the mitigation process. The lower the value, the higher the number of frequent subgroups and disadvantaged ones we expect. For all values of s and all datasets, Div-Rank terminates while satisfying the monotonicity constraint and reduces the number of disadvantaged subgroups to 0.

6 Conclusions

We propose a framework aimed at reducing inequalities in a ranking task for automatically identified disadvantaged subgroups. The approach leverages the notion of divergence to automatically identify data subgroups that experience a disadvantage in terms of ranking utility compared to the overall population. We first outline the desired properties of the mitigation process of disadvantaged subgroups. We propose a mitigation step to mitigate the divergence of disadvantaged subgroups, analyzing its properties and its impact on subgroup divergence. We then propose the re-ranking algorithm Div-Rank which iteratively applies the mitigation step while satisfying the desired properties. The experimental results show the effectiveness of the proposed approach in removing the presence of all disadvantaged subgroups.

Our approach prioritizes group fairness over individual fairness when ranking candidates. Hence, we mitigate disparities in demographic groups, even if it results in individuals from originally advantaged groups being displaced by those from disadvantaged groups. This decision acknowledges systemic biases in decision-making processes, aiming to promote equity. However, it entails trade-offs. While it addresses historical inequalities, it may disadvantage individuals who would have performed well under the original system. Practitioners must weigh these trade-offs based on their context. While prioritizing group fairness may rectify systemic biases, a balanced approach may be necessary in contexts where individual performance is critical.

Transparent communication is crucial in explaining and justifying this process to impacted individuals, adhering to trustworthy decision-making and AI principles. It is critical to inform that the goal is to address systemic biases and promote equity, rather than unfairly disadvantage individuals. Providing examples and illustrating the positive impact on historically disadvantaged groups can help individuals understand the rationale behind the adjustments.

In future work, we plan to take into consideration the *individual-level fairness* aspects in the mitigation process and other subgroup discovery approaches for subgroup identification.

Appendix 1: Additional experiments

This appendix presents additional experimental assessments. First, we provide a comprehensive description of the adopted datasets (Appendix 1.1), followed by an outline of the protected groups under consideration for the comparative baselines (Appendix 1.2). We then present the experimental comparison against the baselines on COMPAS, German credit, IIT-JEE and folktables datasets (Appendix 1.3). Then, we present an extended computational performance analysis, showcasing the efficacy of our approach by considering all attributes available in the datasets (Appendix 1.4). Lastly, we show how we can define the utility function γ of Div-Rank to support a top-k fair ranking scenario, and we experimentally compare our approach with Multi-FAIR (Appendix 1.5). Bibliographic references correspond to the bibliography in the main body of the paper.

Appendix 1.1: Dataset description

We use five real-world datasets commonly adopted in the fairness literature.

COMPAS (Angwin et al. 2016) dataset contains demographic information and the criminal history of 6,172 defendants from the Broward County Sheriff's Office in Florida in 2013 and 2014 collected by ProPublica. In the experiment, we consider the inverse of recidivism scores as target score, derived as in Zehlike et al. (2022). The higher the score, the less likely a defendant is considered as likely to recidivate. We consider *age range*, *ethnicity*, and *gender* as protected attributes.

LSAT (Wightman 1998) dataset derived from a survey across 163 law schools in the United States in 1998 conducted by the Law School Admission Council. The dataset contains information on 21,791 law students, such as their entrance exam scores (LSAT), and their ethnicity and gender, which we considered as protected. We use the LSAT score as the target score for the ranking.

German credit (Lichman 2013) contains the financial information of 1,000 individuals. We consider as sensitive attributes gender and age, where age is categorized into young, adult, or elder as in Zehlike et al. (2022). The credit score used to rank individuals is a weighted sum of account status, credit duration, credit amount, and employment length (Zehlike et al. 2022).

ITT-JEE (Technology IIO 2009) (Indian Institutes of Technology - Joint Entrance Examination) consist of the engineering entrance assessment conducted for admission to engineering colleges in India for the year 2009. The dataset includes the gender and birth category [according to the traditional Indian socio-demographic groups, see (Baswana et al. 2019)] of the students, we consider these two attributes as protected. We use the test scores as the target score for the ranking.

folktables (Ding et al. 2021) dataset contains US Census information. We use the Census data from the California state and the year 2018. We consider the ethnicity and the gender as protected attributes. We use the income of the individuals in the dataset as the target score for the ranking.

Appendix 1.2: Considered protected groups for the compared baselines

The baseline approaches require, as input, the protected groups to mitigate. We consider as protected subgroups the ones analyzed in Zehlike et al. (2022). These also generally align with the ones derived by our automatic identification.

For the LSAT dataset, we consider female and non-Caucasian candidates as protected groups for the gender and ethnicity attributes. We also consider the intersection of the protected attributes: African-American women and non-Caucasian women. For COMPAS, we consider female and non-Caucasian candidates as protected and, iteratively, all values of the age attribute as protected (i.e., {<25}, {25–45}, and {>45} as protected). We also consider the intersection of the three protected attributes: (i) non-Caucasian women, (ii) Caucasian women, and (iii) non-Caucasian men, all with ages lower than 25 (i.e., (i) {<25, non Cauc, female}, {<25, Cauc, female}, {<25, non Cauc, male}) for attributes age, ethnicity, and gender respectively). For the German credit dataset, we consider female, young, and elder as protected groups. At the intersection, we consider 5 groups: (i) young, (ii) adult and (iii) elder female candidates and (iv) adult and (v) elder males. For the IIT-JEEE dataset, we consider female candidates and candidates belonging to the SC and ST birth categories as protected [see Baswana et al. (2019)]. At the intersection, we consider 3 groups: (i) female candidates of the SC birth category, (ii) female candidates of the ST birth category, and (iii) female candidates of the SC or ST birth categories. For the folktables dataset, we consider 2 groups: (i) African-American female candidates, and (ii) non-Caucasian female candidates.

Appendix 1.3: Comparison with baselines: COMPAS, German Credit, IIT–JEE and folktables datasets

Tables 8, 9, 10 and 11 compare the original ranking and re-ranking results of Div-Rank, *feldman* and *CFAθ* for the COMPAS, German credit, IIT-JEE and folktables datasets.

Across all cases, Div-Rank successfully eliminates the count of disadvantaged subgroups. *feldman* ranking struggles when multiple subgroups at the intersections of multiple protected attributes experience a disadvantage. Consider COMPAS and all the three sensitive attributes, *feldman* mitigates at most one subgroup. For German credit and all the two sensitive attributes, *feldman* mitigates the disadvantage in only two out of the five cases. This again highlights the need for the automatic identification of the subgroups to mitigate and handle multiple subgroups. For IIT-JEE and folktables, *feldman* does not mitigate the disadvantage. *CFAθ*, similarly to Div-Rank, handle multiple groups. It achieves disadvantage mitigation when considering the exhaustive enumeration of all the subgroups at the intersection of multiple attributes for the COMPAS and German credit datasets. However, Div-Rank exhibits superior Kendall's τ values (0.68 vs. 0.66 for COMPAS and 0.96 vs. 0.92 for German credit), along with generally reduced execution time. For the IIT-JEE and folktables, *CFAθ* reduces the disadvantage, as the minimum divergence and the number of disadvantaged subgroups reduces, but it does not fully mitigate all disadvantages. After the mitigation, there are still some subgroups with a statistically significant divergence. Div-Rank instead successfully reduces to 0 the number of disadvantaged subgroups.

Table 8 Re-ranking results for DIV-RANK, *CFAθ* and *feldman* rankings. COMPAS dataset

Method	Prot. attr	Prot. groups	$\mid \mathbb{A} \mid$	$\mid \mathbb{D} \mid$	min Δ	max Δ	gini	Kendall's τ	ndcg Loss
original	`age`		1	1	– 0.16	0.15	0.19	1.00	0.00
DIV-RANK		–	1	**0**	– 0.01	0.02	0.14	0.71	0.05
CFAθ		{<25}, {25–45}, {≥ 45}	0	**0**	– 0.01	0.01	0.17	0.71	0.03
feldman		{<25}	1	1	– 0.04	0.10	0.17	0.80	0.03
feldman		{25–45}	1	1	– 0.15	0.16	0.21	0.95	0.00
feldman		{≥45}	1	1	– 0.06	0.02	0.10	0.81	0.00
original	`ethn`		1	1	– 0.04	0.07	0.19	1.00	0.00
DIV-RANK		–	0	**0**	– 0.00	0.00	0.19	0.87	0.00
CFAθ		{non Cauc}, {Cauc}	0	**0**	– 0.00	0.00	0.18	0.87	0.00
feldman		{non Cauc}	0	**0**	– 0.00	0.00	0.16	0.87	0.00
original	`gender`		1	1	– 0.01	0.05	0.19	1.00	0.00
DIV-RANK		–	0	**0**	– 0.00	0.00	0.19	0.94	0.00
CFAθ		{female}, {male}	0	**0**	– 0.00	0.00	0.19	0.94	0.00
feldman		{female}	0	**0**	– 0.00	0.00	0.20	0.95	0.00
original	`age,ethn,gender`		19	14	– 0.18	0.21	0.19	1.00	0.00
DIV-RANK		–	8	**0**	– 0.01	0.08	0.14	0.68	0.03
CFAθ		{<25, non Cauc, female}, {<25, Cauc, female}, {<25, non Cauc, male}, {*other*}	19	11	– 0.15	0.17	0.17	0.84	0.02
CFAθ		{a, e, v}, $\forall a \in$ `age`, $\forall e \in$ `ethn`, $\forall v \in$ `gender`	0	**0**	– 0.01	0.01	0.16	0.66	0.03
feldman		{<25, non Cauc, female}	19	14	– 0.19	0.20	0.19	0.98	0.01
feldman		{<25, Cauc, female}	19	13	– 0.19	0.21	0.19	0.99	0.00
feldman		{<25, non Cauc, Male}	18	14	– 0.16	0.18	0.17	0.87	0.02

Table 9 Re-ranking results for Div-Rank, *CFAθ* and *feldman* rankings

Method	Prot. attr	Prot. groups	| $\mathbb{A}$ |	| $\mathbb{D}$ |	min Δ	max Δ	gini	Kendall's τ	ndcg Loss
original	age, gender		0	2	– 0.26	0.15	0.46	1.00	0.00
Div-Rank		–	0	**0**	– 0.13	0.13	0.45	0.96	0.00
CFAθ		$\{a, v\}$, $\forall a \in$ age, $\forall v \in$ gender	0	**0**	– 0.02	0.01	0.45	0.92	0.02
feldman		{young, female}	0	2	– 0.29	0.12	0.46	0.97	0.00
feldman		{adult, female}	0	**0**	– 0.13	0.13	0.46	0.98	0.01
feldman		{elder, female}	0	1	– 0.26	0.14	0.46	0.99	0.00
feldman		{young, male}	0	**0**	– 0.16	0.24	0.48	0.93	0.01
feldman		{elder, male}	0	2	– 0.25	0.07	0.45	0.99	0.00

German credit dataset

Table 10 Re-ranking results for Div-Rank, *CFAθ* and *feldman* rankings

Method	Prot. attr	Prot. groups	| $\mathbb{A}$ |	| $\mathbb{D}$ |	min Δ	max Δ	gini	Kendall's τ	ndcg Loss
Original	gender, birth cat.		3	13	– 22.04	10.33	0.21	1.00	0.0
Div-Rank		–	1	**0**	– 0.31	0.65	0.19	0.80	0.0
CFAθ		$\{a, v\}$, $\forall a \in$ age, $\forall v \in$ birth cat.	2	10	– 2.92	0.97	0.26	0.88	0.05
feldman		{Female, SC}	3	12	– 22.15	9.87	0.21	0.99	0.01
feldman		{Female, SC}	3	13	– 22.21	10.16	0.21	1.00	0.01
feldman		{Female, SC or ST}	4	12	– 21.20	9.69	0.21	0.98	0.01

IIT-JEE dataset

Appendix 1.4: Computation performance analysis: mitigation for all the attributes

We assess the effectiveness of our approach in handling a growing number of attributes. To do so, we evaluate Div-Rank by considering *all* attributes, thus disregarding the distinction between protected and non-protected ones. For this set of experiments, we consider the LSAT, COMPAS and the German credit datasets.

Table 12 summarizes the characteristics of the adopted real-world datasets. Specifically, for LSAT, we also include the *UGPA* and normalized score *ZFYA* attributes. For COMPAS, we include the attributes *c_charge_degree* and *length_of_stay*. Finally, for German credit, we include the following set: credit_history, purpose, savings_status, installment_commitment, other_parties, residence_since,

Table 11 Re-ranking results for Div-Rank, *CFAθ* and *feldman* rankings

Method	Prot. attr	Prot. groups	$\mid \mathbb{A} \mid$	$\mid \mathbb{D} \mid$	min Δ	max Δ	gini	Kendall's τ	ndcg Loss
Original	`gender, ethnicity`		5	12	– 35,544	20,364	0.53	1.00	0.0
Div-Rank		–	2	**0**	– 761	2110	0.42	0.73	0.0
CFAθ		$\{g, v\}, \forall g \in$ `gender`, $\forall v \in$ `ethnicity`	5	5	– 5996	2499	0.44	0.87	0.06
feldman		{Afr-Am, Female}	5	11	– 35,945	19,963	0.53	1.00	0.01
feldman		{non Cauc, Female}	6	10	– 28,672	19,940	0.53	0.95	0.01

folktables dataset

Table 12 Dataset characteristics

Dataset C	$\mid C \mid$	$\mid X \mid$	$\mid \overline{X} \mid$	$\mid A \mid$
LSAT	21,791	2	2	4
COMPAS	6,172	3	2	5
German credit	1000	2	15	17

$\mid C \mid$ indicates the number of candidates. X and $\overline{X}$ are the set of protected attributes, unprotected ones, while A is the entire set

Table 13 Re-ranking results of Div-Rank, *CFAθ* and *feldman* considering all attributes as input, LSAT dataset

method	prot. attr	prot. groups	$\mid \mathbb{A} \mid$	$\mid \mathbb{D} \mid$	min Δ	max Δ	gini	Kendall's τ	ndcg Loss
Original	$X_1 \dots X_m$		54	88	– 8.22	3.02	0.12	1.00	0.00
Div-Rank		–	15	**0**	– 0.25	1.24	0.13	0.79	0.05
CFAθ		$X_1 \times X_2 \times \dots \times X_m$	0	5	– 0.46	0.19	0.08	0.77	0.01
feldman		top-10 disadvantaged groups	53	92	– 5.83	3.54	0.11	0.93	0.0
feldman		Disadvantaged groups composed of one attribute	62	38	– 1.43	1.54	0.13	0.79	0.0

Minimum subgroup size equal to 100

Table 14 Re-ranking results of Div-Rank, *CFAθ* and *feldman* considering all attributes as input, COMPAS dataset

Method	Prot. attr	Prot. groups	$\mid \mathbb{A} \mid$	$\mid \mathbb{D} \mid$	min Δ	max Δ	gini	Kendall's τ	ndcg Loss
Original	$X_1 \dots X_m$		126	109	– 0.24	0.23	0.19	1.0	0.00
Div-Rank		–	56	**0**	– 0.01	0.10	0.14	0.62	0.06
CFAθ		$X_1 \times X_2 \times \dots \times X_m$	0	5	– 0.02	0.02	0.16	0.6	0.02
feldman		top-10 disadvantaged groups	126	109	– 0.24	0.23	0.19	1.0	0.0
feldman		Disadvantaged groups composed of one attribute	122	89	– 0.20	0.17	0.15	0.81	0.0

Minimum subgroup size equal to 100

property_magnitude, other_payment_plans, housing, existing_credits, job, num_dependents, own_telephone, foreign_worker, civil_status.

In the case of continuous attributes, we discretize them into three ranges using frequent-based discretization to enable subgroup identification. We evaluate two settings. We consider all groups represented by at least 100 instances in the dataset to ensure the mitigation of adequately represented groups. This corresponds to s equal to 0.05 for LSAT, 0.016 for Compas, and 0.1 for German credit. Using the support count to set the support threshold exemplifies how it is easy for practitioners to set the input threshold: the minimum size of the subgroups we want to mitigate.

Tables 13, 14 and 15 show the mitigation results for the LSAT, COMPAS and German Credit datasets. For *CFAθ*, we consider all groups at the intersection of all attribute values, i.e., we consider the Cartesian product of attribute values as the set of all possible combinations of values from different attributes. Being $X_1, X_2, \dots, X_m$ the set of attributes, we refer to all groups over the m attributes with $X_1 \times X_2 \times \dots \times X_m$. For *feldman* setting, we note that we do not know the groups to protect over the entire set of attributes. So, we proceed by first identifying the disadvantaged groups as described in Sect. 3. We then consider two settings: (i) we iteratively mitigate the 10 groups with the highest disadvantage, and (ii) we iteratively mitigate disadvantaged subgroups composed of only one attribute (e.g., '*sex* = *Female*'). Note that these experiments serve also to test whether the identification of disadvantaged groups could be directly used by other mitigation techniques or novel solutions as our Div-Rank.

CFAθ and *Feldman* do no mitigate all disadvantaged subgroups. For Feldman, we also note that mitigating the 10 most disadvantaged groups could even increase the number of disadvantaged ones. Hence, these results show the need for ad-hoc solutions as ours. Div-Rank successfully terminates and mitigates disparities among subgroups for the three datasets, albeit with some (expected) impact on ranking quality measures. We again remark that this experimental set is for demonstration purposes of Div-Rank effectiveness, as our goal is fair ranking and reducing disparities among subgroups over protected attributes.

Table 15 Re-ranking results of Div-Rank, *CFAθ* and *feldman* considering all attributes as input, German credit dataset

Method	Prot. attr	Prot. groups	$\lvert \mathbb{A} \rvert$	$\lvert \mathbb{D} \rvert$	min Δ	max Δ	gini	Kendall's τ	ndcg Loss
Original	$X_1 \dots X_m$		990	6188	– 0.59	0.83	0.46	1.00	0.0
Div-Rank		-	223	**0**	– 0.09	0.38	0.30	0.52	0.1
CFAθ		$X_1 \times X_2 \times \dots \times X_m$	1006	5913	– 0.24	0.32	0.21	0.98	0.00
feldman		top-10 disadvantaged groups	990	6188	– 0.59	0.83	0.46	1.0	– 30.22
feldman		Disadvantaged groups composed of one attribute	305	2821	– 0.75	0.87	0.43	0.67	– 23.540

Minimum subgroup size equal to 100

Table 16 Execution time of Div-Rank, *CFAθ* and *feldman* considering all attributes as input, LSAT, COMPAS and German Credit datasets

Dataset	Div-Rank	*CFAθ*	*feldman* - top-10 disadvantaged groups	*feldman* - disadvantaged groups composed of one attribute
LSAT	1 min 21 s	1 h 45 min —	4.89 s	18.28 s
COMPAS	28 s	23 s	0.41 s	1.49 s
German credit	17 min —	16 min —	0.1 s	0.5 s

Minimum subgroup size equal to 100

Table 16 shows the execution time Div-Rank, *CFAθ* and *feldman* when considering all attributes as input on the three datasets LSAT, COMPAS and German Credit. We note that Div-Rank execution time is lower or comparable to the other methods while it is the only one that mitigates all the disadvantages.

Appendix 1.5: Top-k ranking and comparison with Multi-FAIR

The computational limitations of *Multi-FAIR* make re-ranking the entire list (i.e., with $K = |C|$) infeasible. In order to compare with *Multi-FAIR* we adapt our approach to deal with a top-K fair ranking scenario, with $K \ll |C|$, as follows.

We use as a utility function for each candidate a function of its position in the top-K ranking. Specifically, we set the utility score as the candidate score itself for the candidates in the top-K positions, while we set it to 0 for the candidates not in the top-K. Our approach still considers the average utility across all candidates as the baseline for the divergence computation and adjusts the score for all candidates.

As a workaround for *Multi-FAIR*'s computational bottleneck, we retrieved precomputed re-ranking results from *Multi-FAIR*'s repository. The pre-computed reranking results include the LSAT, COMPAS, and German credit datasets. For LSAT, the analysis involves four groups, defined by the protected attributes of gender and ethnicity (with values *Caucasian* and *protected*), and K set to 300. For COMPAS, the setting considers 4 groups: three pre-identified disadvantaged groups and the remainder group; $K = 300$. For German credit, the setting considers 6 groups (5 as protected), defined by the protected attributes of gender and age (with values *young*, *adult* and *elder*), and K set to 50.

Recall that *Multi-FAIR* re-ranks only the top K, while ours readjusts the score of the entire ranking to generate the mitigated ranking. To make the results comparable, we assigned a utility score based on the position on the mitigated ranking, from K to 1 for the candidates in the top-K position of the ranking and 0 for the others. We report the results in Tables 17 and 18 for LSAT and COMPAS respectively. We do not report the results for German since, with the used setting and $K = 50$, we identified no disadvantaged group. For LSAT, our method reduces the number of disadvantaged groups from 3 to 0, while *Multi-FAIR*

Table 17 Re-ranking results for Div-Rank and *Multi-FAIR* rankings

Method	Prot. attr	Prot. groups	| $\mathbb{A}$ |	| $\mathbb{D}$ |	min Δ	max Δ	gini	Kendall's τ	ndcg Loss
original	`gender, ethn`		1	3	– 1.39	0.57	0.99	1.00	0.00
Div-Rank		-	0	**0**	– 0.78	0.41	0.99	1.0	0.05
Multi-FAIR		$\{g, v\}, \forall g \in \{\text{female}, \text{male}\}, \forall v \in \{\text{Cauc}, \text{prot.}\}$	3	2	– 0.91	1.18	0.99	0.74	0.16

LSAT dataset. $k = 300$. For Div-Rank, we define the utility score as the score a candidate receives if they are in the top-k positions of the ranking; 0 otherwise

Table 18 Re-ranking results for Div-Rank and *Multi-FAIR* rankings

Method	Prot. attr	Prot. groups	| $\mathbb{A}$ |	| $\mathbb{D}$ |	min Δ	max Δ	gini	Kendall's τ	ndcg Loss
original	`age, ethn, gender`		12	15	– 7.32	27.99	0.97	1.00	0.00
DivRank		–	6	**9**	– 7.32	7.56	0.97	0.98	0.07
Multi-FAIR		{<25, prot., male}, {<25, prot., female}, {<25, Cauc., female}, {*other*}	11	13	– 7.32	24.51	0.97	0.87	0.02

COMPAS dataset. $k = 300$. For Div-Rank, we define the utility score as the score a candidate receives if they are in the top-k positions of the ranking; 0 otherwise

reduces them to 2. For the COMPAS dataset, our approach reduces the number of disadvantaged groups from 15 to 9, while *Multi-FAIR* reduces them to 13. The results from *Multi-FAIR* are based on a subset of subgroups (3) identified as protected, which explains the outcome when considering a broader range of groups. Div-Rank adjusts the utility scores considering the entire ranking. Hence, it is not specifically optimized for the top-K ranking scenario.

These experiments not only demonstrate the effectiveness of our approach compared to the *Multi-FAIR* but also demonstrate how we can set the utility function of Div-Rank to support a top-K re-ranking scenario.

Acknowledgements This work is partially supported by the spoke "FutureHPC & BigData" of the ICSC - Centro Nazionale di Ricerca in High-Performance Computing, Big Data and Quantum Computing, funded by the European Union - NextGenerationEU.

Author contributions All authors were involved in conceptualizing the study and designing the methodology. E.P. performed the experiments and developed the software used in the study. E.P. contributed to writing the original draft of the manuscript. F.B. contributed to editing the manuscript. All authors reviewed the final manuscript.

Funding Open access funding provided by Politecnico di Torino within the CRUI-CARE Agreement.

Availability of data and materials No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate The datasets and the target scores used as the utility function in our experiments are commonly adopted in the fairness literature. Still, they were not specifically designed for ranking candidates in real applications. Nevertheless, these real-world datasets serve as experimental demonstrations of the effectiveness of the proposed approach. We do not necessarily endorse using such scores to rank people.

Competing interests The authors declare no Conflict of interest.

References

Zehlike M, Bonchi F, Castillo C, Hajian S, Megahed M, Baeza-Yates R (2017) FA*IR: a fair top-k ranking algorithm. In: Proceedings of the 2017 ACM on conference on information and knowledge management, CIKM 2017, Singapore, November 06–10, 2017, pp 1569–1578. https://doi.org/10.1145/3132847.3132938

Yang K, Stoyanovich J (2017) Measuring fairness in ranked outputs. In: Proceedings of the 29th international conference on scientific and statistical database management. SSDBM'17. https://doi.org/10.1145/3085504.3085526

Singh A, Joachims T (2018) Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining. KDD'18, pp 2219–2228. https://doi.org/10.1145/3219819.3220088

Celis LE, Straszak D, Vishnoi NK (2018) Ranking with fairness constraints. In: 45th International colloquium on automata, languages, and programming (ICALP 2018), vol 107, p 28. https://doi.org/10.4230/LIPIcs.ICALP.2018.28

Yang K, Gkatzelis V, Stoyanovich J (2019) Balanced ranking with diversity constraints. In: Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI. https://doi.org/10.24963/ijcai.2019/836

Celis LE, Mehrotra A, Vishnoi NK (2020) Interventions for ranking in the presence of implicit bias. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. FAT*'20, pp 369–380. https://doi.org/10.1145/3351095.3372858

García-Soriano D, Bonchi F (2021) Maxmin-fair ranking: individual fairness under group-fairness constraints. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery and data mining. KDD'21. ACM, pp 436–446. https://doi.org/10.1145/3447548.3467349

Zehlike M, Sühr T, Baeza-Yates R, Bonchi F, Castillo C, Hajian S (2022) Fair top-k ranking with multiple protected groups. Inf Process Manag. https://doi.org/10.1016/j.ipm.2021.102707

Ekstrand MD, McDonald G, Raj A, Johnson I (2023) Overview of the TREC 2022 fair ranking track. arXiv preprint arXiv:2302.05558

Crenshaw K (1990) Mapping the margins: intersectionality, identity politics, and violence against women of color. Stanf Law Rev 43:1241

Wightman LF (1998) LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. ERIC. https://eric.ed.gov/?id=ED469370

Pastor E, Alfaro L, Baralis E (2021) Looking for trouble: analyzing classifier behavior via pattern divergence. In: Proceedings of the 2021 international conference on management of data. SIGMOD'21. ACM, New York, pp 1400–1412. https://doi.org/10.1145/3448016.3457284

Rawls J (1971) A theory of justice: original edition, 1st edn. Harvard University Press. http://www.jstor.org/stable/j.ctvjf9z6v

Zehlike M, Yang K, Stoyanovich J (2022a) Fairness in ranking, part I: score-based ranking. ACM Comput Surv. https://doi.org/10.1145/3533379

Zehlike M, Yang K, Stoyanovich J (2022b) Fairness in ranking, part II: learning-to-rank and recommender systems. ACM Comput Surv. https://doi.org/10.1145/3533380

Patro GK, Porcaro L, Mitchell L, Zhang Q, Zehlike M, Garg N (2022) Fair ranking: a critical review, challenges, and future directions. In: 2022 ACM conference on fairness, accountability, and transparency. FAccT'22. ACM, New York, pp 1929–1942. https://doi.org/10.1145/3531146.3533238

Pitoura E, Stefanidis K, Koutrika G (2022) Fairness in rankings and recommendations: an overview. VLDB J. https://doi.org/10.1007/s00778-021-00697-y

Mitchell S, Potash E, Barocas S, D'Amour A, Lum K (2021) Algorithmic fairness: choices, assumptions, and definitions. Ann Rev Stat Appl 8(1):141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. ITCS'12. ACM, New York, pp 214–226. https://doi.org/10.1145/2090236.2090255

Zehlike M, Hacker P, Wiedemann E (2020) Matching code and law: achieving algorithmic fairness with optimal transport. Data Min Knowl Discov 34(1):163–200. https://doi.org/10.1007/s10618-019-00658-8

Asudeh A, Jagadish H, Stoyanovich J, Das G (2019) Designing fair ranking schemes. In: Proceedings of the 2019 international conference on management of data, pp 1259–1276. https://doi.org/10.1145/3299869.3300079

Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. KDD'15. ACM, pp. 259–268. https://doi.org/10.1145/2783258.2783311

Yang K, Loftus JR, Stoyanovich J (2021) Causal intersectionality and fair ranking. In: 2nd Symposium on foundations of responsible computing (FORC). https://doi.org/10.4230/LIPIcs.FORC.2021.7

Pastor E, Baralis E, Alfaro L (2023) A hierarchical approach to anomalous subgroup discovery. In: 2023 IEEE 39th international conference on data engineering (ICDE), pp 2647–2659. https://doi.org/10.1109/ICDE55515.2023.00203

Sagadeeva S, Boehm M (2021) Sliceline: fast, linear-algebra-based slice finding for ML model debugging. In: Proceedings of the 2021 international conference on management of data, pp 2290–2299. https://doi.org/10.1145/3448016.3457323

Chung Y, Kraska T, Polyzotis N, Tae KH, Whang SE (2019) Slice finder: automated data slicing for model validation. In: 2019 IEEE 35th international conference on data engineering (ICDE). IEEE, pp 1550–1553. https://doi.org/10.1109/ICDE.2019.00139

Pastor E, Alfaro L, Baralis E (2021) Identifying biased subgroups in ranking and classification. Presented at the Responsible AI @ KDD 2021 Workshop (non archival). arXiv preprint arXiv:2108.07450

Li J, Moskovitch Y, Jagadish H (2023) Detection of groups with biased representation in ranking. In: 2023 IEEE 39th international conference on data engineering (ICDE). IEEE, pp 2167–2179. https://doi.org/10.1109/ICDE55515.2023.00168

Welch BL (1947) The generalization of Student's' problem when several different population variances are involved. Biometrika 34(1–2):28–35. https://doi.org/10.2307/2332510

Siegel AF (2012) Chapter 10—Hypothesis testing: deciding between reality and coincidence. In: Siegel AF (ed) Practical business statistics, 6th edn. Springer, Berlin, pp 249–287. https://doi.org/10.1016/B978-0-12-385208-3.00010-9

Han J, Pei J, Yin Y (2000) Mining frequent patterns without candidate generation. ACM Sigmod Rec 29(2):1–12. https://doi.org/10.1145/335191.335372

Herrera F, Carmona CJ, González P, Del Jesus MJ (2011) An overview on subgroup discovery: foundations and applications. Knowl Inf Syst 29:495–525

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine Bias. ProPublica, New York

Lichman M (2013) UCI machine learning repository. https://archive.ics.uci.edu

Technology IIO (2009) IIT–JEE. https://indiankanoon.org/doc/1955304/

Ding F, Hardt M, Miller J, Schmidt L (2021) Retiring adult: new datasets for fair machine learning. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW (eds.), Advances in Neural Information Processing Systems, vol. 34, pp. 6478–6490. Curran Associates, Inc., Virtual. https://proceedings.neurips.cc/paper/2021/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html

Gini C (1921) Measurement of inequality of incomes. Econ J 31(121):124–125. https://doi.org/10.2307/2223319

Shapley LS (1952) A value for n-person games. https://doi.org/10.1515/9781400881970-018

Baswana S, Chakrabarti PP, Kanoria Y, Patange U, Chandran S (2019) Joint seat allocation 2018: an algorithmic perspective. arXiv preprint arXiv:1904.06698