Reducing Disparate Exposure in Ranking: A Learning to Rank Approach

Meike Zehlike Humboldt Universität zu Berlin Max Planck Inst. for Software Systems meikezehlike@mpi-sws.org Gina-Theresa Diehn Technische Universität Berlin gina.diehn@googlemail.com Carlos Castillo Universitat Pompeu Fabra chato@acm.org

ABSTRACT

In this paper we introduce DELTR, a learning-to-rank framework that addresses potential issues of discrimination and unequal opportunity in rankings. Following long-standing empirical observations showing that users of information retrieval systems rarely look past the first few results, we measure these problems in terms of discrepancies in the *average group exposure*. Specifically, we define our notion of group exposure as the average probability of items from a *legally protected* social group to be ranked at the top position. With this we design a ranker that optimizes search results in terms of relevance, while at the same time reducing potential discrimination or inequality of opportunity.

We describe this objective formally, how to optimize it efficiently, and how to implement it. We perform an extensive experimental study showing that being "colorblind," i.e. ignoring protected attributes such as race or gender, can be among the best choices or the worst choices from the perspective of relevance and exposure, depending on how much and which kind of bias is present in the training set. As baselines for benchmarking our in-processing method we use pre-processing and post-processing methods based on FA*IR, a state-of-the-art algorithm to re-rank search results according to predefined fairness constraints. We show that our in-processing method performs better in terms of relevance and equality of exposure than pre-processing and post-processing across all tested scenarios.

Our proposed method neither makes assumptions about biases in the training data, nor does it ignore relevance scores of items and thus can reduce discrimination and inequality of opportunity without having to introduce large distortions in ranking relevance.

1 INTRODUCTION

Ranked search results, news feeds, and recommendations, have become the main mechanism by which we find content, products, places, and people online. These rankings are typically constructed to provide a maximum utility to searchers, for instance, by ordering items by decreasing probability of being relevant [26]. However, when the items to be ranked represent people, businesses or places, ranking algorithms have consequences that go beyond immediate utility for searchers. With hiring, selecting, purchasing, and dating being increasingly mediated by algorithms, rankings may determine career and business opportunities, educational placement, access to benefits, and even social and reproductive success.

Over the past decade, data mining researchers became increasingly concerned with various systematic biases [15] against any specific group (i.e., *group discrimination*), caused by historic and current discriminatory patterns in society making their way into data-driven models. A common element in this line of research is the presence of a historically and currently disadvantaged *protected group*, and the concern of *disparate impact*, i.e., loss of opportunity for the protected group independently of whether they are (intentionally) treated differently. In the case of rankings, a natural way of understanding disparate impact is by considering differences in exposure [29] or inequality of attention [3], which translate into systematic differences in access to economic or social opportunities.

Disparate exposure in rankings. There are a number of issues, sometimes appearing simultaneously, that call for reducing disparate exposure in information retrieval systems. First, there can be a situation in which minimal differences in relevance translate into large differences in exposure for different groups [3, 29], because of the large skew in the distribution of exposure brought by positional bias [21]. Second, there can be a legal requirement, policy, or voluntary commitment that requires that elements in the protected group are given sufficient visibility among the top positions in a ranking [9, 35]. Third, there can be systematic differences in the way in which documents are constructed, as in the case of different sections in online resumes, which are completed differently by men and women [2]; these differences may systematically affect ranking algorithms. Fourth, there can be systematic differences in the way ground truth rankings have been generated due to historical discrimination and/or annotator bias.

These situations point to two conceptually different goals: *reducing inequality of opportunities* and *reducing discrimination* (as defined by Roemer [27], chapter 12). Equality of opportunity seeks to correct a historical or present disadvantage for a group of society. Non-discrimination seeks to allocate resources in a way that does not consider irrelevant attributes, and is a matter of efficiency. DELTR can be applied in both cases, as our experiments show.

Post-processing methods for fair rankings. Fairness-aware datadriven methods can be classified into *pre-, in-* and *post-processing* approaches, in which pre-processing methods seek to mitigate bias in training data, in-processing methods learn a bias-free model, and post-processing methods re-rank output items [17].

For rankings, several post-processing methods have been presented in the scientific literature [3, 9, 29, 35], but these methods are problematic for various reasons. First, the post-processing idea suggests that there is *always* a trade-off between an optimally *fair* and an optimally *relevant* ranking, because a presumably "exact" model produces a "relevant" ranking that is then reordered to meet fairness constraints. We show in our experiments that this assumption is wrong, as reducing bias against a protected group can increase relevance (experiment 6.3). Second, it is likely that post-processing algorithms are legally open to criticism as the Ricci v. DeStefano ruling (2009) in the US Supreme Court illustrates. At the center of this case was a test administered by the firefighter department of New Haven, in which white firefighters scored significantly better than black ones. The result was then rejected by the department. The US Supreme Court ruled that the white firefighters were subjected to race discrimination, because the department could not demonstrate a "strong basis in evidence that using the results would cause them to lose a disparate-impact suit." However, "[e]mployers may consider potential racial impact during the test-design stage". [23] This decision states that it is as much unlawful to score people solely on the basis of race, as it is to reject scores produced from lawful exams and procedures solely on that very same basis, which is exactly what post-processing ranking algorithms do. One crucial advantage of in-processing approaches such as DELTR is that they satisfy the condition of "considering potential racial impact during the test-design stage". Pre-processing methods also satisfy this condition. However, they suggest that if we only had unbiased training data, we could use standard ranking tools without having to worry about biased models. We show that creating an unbiased training set is not trivial and may easily lead to reverse discrimination.

Our contribution. We address the problem of mitigating discrimination and inequality of opportunity in rankings by reducing disparate exposure under a learning-to-rank framework. We design DELTR as a list-wise learning-to-rank approach that provides an *in-processing* approach to fairness-aware rankings.

We perform extensive experiments in two different ranking tasks: expert search in a document retrieval setting, and ranking students by predicted performance. Our experiments comprise three realworld datasets, of which two are newly introduced (section 5.1 and 5.2). We further show that being "colorblind" on discriminatory training data, i.e., simply ignoring protected attributes, which is a naive attempt to overcome discriminatory models, can yield the best results in some cases, while in others it is among the worst results both in terms of performance and fairness. This makes it difficult to identify when or when not to include the protected attribute in the training process. We describe the reasons for this somewhat counterintuitive behavior in detail and show that DELTR performs well in terms of fairness and relevance in all tested scenarios.

As one baseline we demonstrate a pre-processing approach for fairness in rankings by applying a post-processing method, FA*IR [35] to our training data before the learning routine starts. These experiments show two interesting insights: (i) it is not easy to produce fair training data, because discrimination is baked into all attributes, and a truly bias-free dataset is just wishful thinking, as argued earlier in this section; and (ii) re-ordering items in a "fair" way can lead to significant performance decline and even to reverse discrimination. We show that DELTR does not suffer from the aforementioned problems of post- or pre-processing methods, as it makes no assumptions on whether or not bias against a protected group is present in the training data. It performs well in both cases.

Additionally, we show that the current understanding of a necessary trade-off between relevance and fairness can be sometimes misleading. Only if we seek to enhance equality of opportunity, we have to trade performance against fairness (experiments 6.2, 6.4 and 6.5). In a case of non-discrimination, optimizing for fairness as equal exposure will increase relevance (experiments 6.1 and 6.3).

2 RELATED WORK

Fairness in data-driven modeling. In recent years the data mining and machine learning communities have been increasingly concerned with *algorithmic bias*, particularly with the fact that sensitive attributes have been found to have an observable impact on machine learning outcomes [17]. Membership in socially salient groups defines a protected characteristic, while merely removing sensitive attributes from training data may have little or no effect on a data-driven model [18]. Algorithmic fairness as defined by Žliobaitė [30] seeks that: (1) people that are similar in terms of non-protected characteristics should receive similar predictions, and (2) differences in predictions across groups of people can only be as large as justified by non-protected characteristics. The study of algorithmic discrimination and fairness is connected to open debates in moral philosophy including, among other topics, distributive justice and egalitarianism (see, e.g., Binns [4]).

There are various approaches to algorithmic fairness. A basic method for algorithmic fairness seeks to reduce *disparate impact* by achieving *statistical parity* between the outcomes for protected and non-protected elements; however, this can be inadequate if the outcomes also depend on non-protected, legitimate attributes [14]. Other methods define algorithmic fairness in terms of predicted and actual outcomes, reducing differences in false positive or true positive rates that have been called *disparate mistreatment* [34] or unequal opportunity [19].

Fairness in rankings. Fairness in ranking is concerned with a sufficient presence, a consistent treatment, and a proper representation of different groups across all ranking positions [8]. At a high level, this research is motivated by producing rankings based on relevant characteristics of items, in which items belonging to the protected group are not under-represented or relegated systematically to lower ranking positions [33]. This requires new evaluation metrics that extend relevance-based ones [11].

Yang and Stoyanovich [32] introduce a generative model for fair rankings with two groups (protected and non-protected). They also introduce a series of ranking-aware measures of disparity, such as averaging differences in NDCG (Normalized Discounted Cumulative Gain) at different cut-off points across both groups.

Zehlike et al. [35] construct a statistical test for the generative model of Yang and Stoyanovich [32]. Given a parameter p and a statistical significance α , reject a ranking that has probably not been generated according to this process, based on an adjusted binomial test. They also provide a method for generating a ranking that passes this statistical test, given two separate rankings for the two groups. Celis et al. [9] consider a situation in which several protected groups exist and hence several vectors containing the exact number of protected elements (one per group) at each position are given as input.

Singh and Joachims [28] introduce the concept of *exposure* of a group based on empirical observations that show that the probability that a user examines an item ranked at a certain position, decreases rapidly with the position. Biega et al. [3], in work parallel to Singh and Joachims [28], introduce an integer linear programming formulation that receives a vector of relevance scores and produces a ranking that places high-scoring elements first, and at

Table 1: Summary of Notation

0	set of queries q with $ Q = m$
\widetilde{D}	set of documents
$d_i^{(q)}$	a document associated to query q
$s_i^{(q)}$	a general judgment on document $d_i^{(q)}$ for query q
$x_i^{(q)}$	feature vector for document $d_i^{(q)}$
$y^{(q)}$	list of training judgments
f	ranking model
$\hat{y}^{(q)}$	list of predicted judgments
Ľ	error between the training judgments and those predicted by
	model <i>f</i>
$P_s(i)$	probability for document <i>i</i> to be ranked at the top position
G_k	groups of documents identifiable by the presence or absence
	of sensitive attributes
v_i	position bias
Ů	disparate exposure metric
L_{DELTR}	loss function that incorporates L and U at the same time
Y	tuning parameter

the same time minimizes the accumulated attention received by elements in both groups.

Previous works in fair rankings [3, 9, 28, 32, 35] have been concerned with creating a fairness-aware ranking given a set of scores, and can be considered *post-processing* approaches—they are given a ranking and re-rank elements to achieve a desired objective. In contrast, our approach DELTR is *learning-based* and the first *inprocessing* approach to reduce discrimination and inequality of opportunity in rankings, because it learns a ranking function with an additional objective that reduces disparate exposure. In the experiments on this paper, we additionally describe how to use a postprocessing method on training data to implement a pre-processing approach.

Diversity in rankings. A classical definition of diversity in ranking is related to the marginal relevance of a document for a user, considering the documents ranked above it, that the user has already seen [7]. Another often-used definition is that diversity should be understood as a way of incorporating uncertainty over user intents, in the sense that all queries have some degree of ambiguity [1]. Both interpretations ("seeking variety" and "hedging bets") are present in contemporary accounts of diversity in data-driven methods [13].

In contrast with diversification approaches, we are not only concerned with the utility that search system users receive, but also with the exposure of the items being ranked, which can represent individual, organizations, or places. In other words, we also consider their utility. Another key difference is that diversification is usually symmetric so groups are interchangeable, while fairness-aware algorithms are usually asymmetric, as they focus on increasing the overall benefit received by a disadvantaged or protected group.

3 PRELIMINARIES: LISTNET

In this section we describe ListNet [6], a well-known list-wise learning to rank framework. Given that lists present a natural way to measure disparate exposure across groups for an entire ranking, we describe DELTR using ListNet as a base algorithm. Readers familiar with ListNet can skip this section. The notation we use is summarized on Table 1. **List-wise learning to rank.** We consider a set of queries Q with |Q| = m and a set of documents D with |D| = n. Each query q is associated with a list of candidate documents $d^{(q)} \subseteq D$. We denote by $s_i^{(q)} \in \mathbb{R}_0^+$ for $q \in Q$; $i = 1, 2, ..., |d^{(q)}|$ a judgment on document $d_i^{(q)}$ for query q, that indicates the extent to which document candidate $d_i^{(q)}$ is relevant for q.

For each query q the list of candidate documents is associated with a list of judgments: $d^{(q)} \rightarrow s^{(q)} = \left(s_1^{(q)}, s_2^{(q)}, \dots, s_{n^{(q)}}^{(q)}\right)$. For a clearer distinction between different judgment sets we call $y^{(q)}$ the judgments of the training data and $\hat{y}^{(q)}$ the judgments predicted by the model.

From each document $d_i^{(q)}$ we can derive a feature vector $x_i^{(q)}$. Each list of feature vectors $x^{(q)} = \left(x_1^{(q)}, x_2^{(q)}, \dots, x_{n^{(q)}}^{(q)}\right)$ and the corresponding list of judgments $y^{(q)} = \left(y_1^{(q)}, y_2^{(q)}, \dots, y_{n^{(q)}}^{(q)}\right)$ form an instance of the training set $\mathcal{T} = \left\{(x^{(q)}, y^{(q)})\right\}_{q \in Q}$. The standard learning-to-rank objective then is to learn a ranking function f that outputs a new judgment $\hat{y}_i^{(q)}$ for each feature vector $x_i^{(q)}$ which forms a second list of judgments $\hat{y}^{(q)} = \left(f(x_1^{(q)}), f(x_2^{(q)}), \dots, f(x_{n^{(q)}}^{(q)})\right)$.

Ideally, the function f should be such that the sum of the differences (or losses) L between the training judgments $y^{(q)}$ and the predicted judgments $\hat{y}^{(q)}$ is minimized:

$$\min\left(\sum_{q\in Q} L\left(y^{(q)}, \hat{y}^{(q)}\right)\right)$$

In list-wise learning to rank, training elements are processed as lists of elements (not as individual elements having scores, which corresponds to point-wise learning to rank, or as pairs of elements, which corresponds to pair-wise learning to rank).

Probability models. As rankings are combinatorial objects the naive approach to find an optimal solution for *L* leads to exponential execution time in the number of documents. Hence instead of considering an actual permutation of documents we will reuse Theorem 6 and Lemma 7 from Cao et al. [6], which focuses on the probability for a document $d_i^{(q)}$ to be ranked onto the top position:

$$P_{s^{(q)}}\left(d_{i}^{(q)}\right) = \frac{\phi\left(s_{i}^{(q)}\right)}{\sum_{j=1}^{n} \phi\left(s_{j}^{(q)}\right)} \tag{1}$$

with $\phi : \mathbb{R}_0^+ \longrightarrow \mathbb{R}^+$ being an increasing strictly positive function and in which $s_j^{(q)}$ denotes a relevance score/judgment for document *j*. The top-one-probabilities $P_{s^{(q)}}(d_i^{(q)})$ form a probability distribution $P_{s^{(q)}}$ over $d^{(q)}$.

In a general list-wise learning-to-rank setting the document judgments $s^{(q)}$ are given as lists of scores that represent the respective relevance degree of document $d_i^{(q)}$ to query q. Documents are sorted by decreasing top-one probabilities as predicted by the algorithm. **Loss function in list-wise learning to rank.** Setting $P_{s^{(q)}}$ to $P_{y^{(q)}}(x_i^{(q)})$ or $P_{\hat{y}^{(q)}}(x_i^{(q)})$ respectively, leads to a way of measuring the distance between the judgments provided in the training set $y^{(q)}$ and the judgments $\hat{y}^{(q)}$ produced by our function f. Following Cao et al. [6] we use the Cross Entropy metric for the loss function:

$$L\left(y^{(q)}, \hat{y}^{(q)}\right) = -\sum_{i=1}^{|d^{(q)}|} P_{y^{(q)}}(x_i^{(q)}) \log\left(P_{\hat{y}^{(q)}}(x_i^{(q)})\right) \tag{2}$$

4 OUR METHOD: DELTR

In this section we describe our method, DELTR (Disparate Exposure in Learning To Rank). We assume that the retrieved items represent people belonging to two distinct socially salient [22] groups (such as men and women, or majority and minority ethnicity). Using terminology from non-discriminatory data mining, we assume one of these groups is *protected* [25]. DELTR is a supervised learningto-rank algorithm that simultaneously seeks to minimize ranking errors with respect to training data, and to reduce disadvantages experienced by the protected group in terms of exposure.

At training time, we are given an annotated set consisting of queries and ordered lists of items for each query. The algorithm learns from training data by minimizing a loss function. At testing time, we provide a query and a document collection, and expect as output a set of top-k items from the collection that should be relevant for the query, and additionally should not exhibit disparate exposure.

4.1 Disparate Exposure

We assume that items in D belong to two different groups, which we denote by G_0 for the non-protected group, and G_1 for the protected groups. Items in the protected group have a certain protected attribute, such as belonging to an underprivileged group. As argued in Section 1, the protected group may, due to various causes including historic discrimination or erratic data collection procedures, have a significant disadvantage in the training dataset. This is likely to cause a model to predict rankings with a large discrepancy in exposure, and hence not only to incorporate but to reproduce discrimination and unequal opportunities for already disadvantaged groups. As a remedy we design a learning-to-rank objective to optimize the results not only for accuracy with respect to the training data, but also with respect to the unfairness of the predictions.

To define a measure of unfairness we borrow [29]'s definition of exposure of a document d in a ranked list generated by a probabilistic ranking P as:

Exposure
$$(d|P) = \sum_{j=1}^{n} P_{d,j} \cdot v_j$$
 (3)

where $P_{d,j}$ is the probability that document d will be ranked in position j, and v_j is the *position bias* of position j, indicating its relative importance for users of a ranking system. We use a logarithmic discount function $v_j = \frac{1}{\log(1+j)}$ which is commonly used [20]. In our current implementation of this framework we deal only with top-one probabilities, i.e., we adapt equation 3 such that the exposure of document $d_i^{(q)}$ represented by features $x_i^{(q)}$ is its probability of achieving the top position:

$$\operatorname{Exposure}\left(x_{i}^{(q)}|P_{\hat{y}^{(q)}}\right) = P_{\hat{y}^{(q)}}\left(x_{i}^{(q)}\right) \cdot \upsilon_{1}$$
(4)

Hence, the average exposure of documents in group G_p with $p \in \{0, 1\}$ is

$$\text{Exposure}(G_{p}|P_{\hat{y}^{(q)}}) = \frac{1}{|G_{p}|} \sum_{x_{i}^{(q)} \in G_{p}} \text{Exposure}(x_{i}^{(q)}|P_{\hat{y}^{(q)}}) \quad (5)$$

Finally, we adapt the first definition of equal exposure in Singh and Joachims [29], *demographic parity*, which compares the average exposure across items from all groups is equal. With this we can now introduce an unfairness criterion measured in terms of disparate exposure:

$$U(\hat{y}^{(q)}) = \max\left(0, \operatorname{Exposure}(G_0|P_{\hat{y}^{(q)}}) - \operatorname{Exposure}(G_1|P_{\hat{y}^{(q)}})\right)^2 \quad (6)$$

Note that using the squared hinge loss allows us to have a differentiable loss function that prefers rankings in which the exposure of the protected group is not less than the exposure of the nonprotected group *but not vice versa*. This means that our definition will optimize only relevance in cases where the protected group already receives as much exposure as the nonprotected group.

We note that other definitions of disparate exposure can be used as long as they can be optimized efficiently (e.g., differentiable), and that the definition in equation 6 can be trivially extended to multiple protected groups by considering average or maximum difference of exposure between a protected group and the nonprotected one.

4.2 Formal Problem Statement

Learning to rank obtains a ranking function f that is learned by solving a minimization problem with respect to a loss function L, as described in Section 3. In our case, we learn f by solving a minimization problem with respect to loss function L_{DELTR} , which incorporates a measure of accuracy with respect to the training data L, as well as a measure of unfairness U in terms of exposure with respect to the generated rankings. Specifically, we seek to minimize a weighted summation of the two elements, controlled by a parameter $\gamma \in \mathbb{R}_0^+$:

$$L_{DELTR}\left(y^{(q)}, \hat{y}^{(q)}\right) = L\left(y^{(q)}, \hat{y}^{(q)}\right) + \gamma U\left(\hat{y}^{(q)}\right)$$
(7)

with larger γ expressing preference for solutions that reduce disparate exposure for the protected group, and smaller γ expressing preference for solutions that reduce the differences between the training data and the output of the ranking algorithm. The parameter γ depends on desired trade-offs between ranking utility and disparate exposure that are application-dependent. In our experiments, we consider two settings: γ_{large} in which γ is comparable to the value of the standard loss *L*, and γ_{small} in which it is an order of magnitude smaller.

We remark that U only depends on \hat{y} and is hence not directly affected by biases in the training data, which is a great advantage compared to the naive "colorblind" approach. Our new objective can handle both cases, the one in which it is desirable to exclude the protected attribute during training as well as the case in which it is desirable to include it.

4.3 Optimization

For the ranking function to infer the document judgments we use a linear function $f_{\omega}(x_i^{(q)}) = \langle \omega \cdot x_i^{(q)} \rangle$, and Gradient Descent to find

an optimal solution for L_{DELTR} . We can now rewrite the top-oneprobability for a document (Equation 1) and set ϕ to an exponential function, which is strictly positive and increasing and convenient to derive:

$$P_{\hat{y}^{(q)}(f_{\omega})}(x_i^{(q)}) = \frac{\exp(f_{\omega}(x_i^{(q)}))}{\sum_{k=1}^n \exp(f_{\omega}(x_k^{(q)}))}$$
(8)

To use Gradient Descent we need the derivative of $L_{DELTR}(y^{(q)}, \hat{y}^{(q)})$ which in turn consists of the derivatives of the disparate exposure and accuracy metric respectively.

$$\frac{\partial L_{DELTR}\left(y^{(q)}, \hat{y}^{(q)}\right)}{\partial \omega} = \frac{\partial L(y^{(q)}, \hat{y}^{(q)})}{\partial \omega} + \gamma \cdot \frac{\partial U(\hat{y}^{(q)})}{\partial \omega} \tag{9}$$

The derivative of *L* can be found in [6]. For brevity we write $e_{i,q}$ instead of $\exp(f_{\omega}(x_i^{(q)}))$ in equation 8:

$$P_{\hat{y}^{(q)}(f_{\omega})}(x_{i}^{(q)}) = \frac{e_{i,q}}{\sum_{k=1}^{n} e_{k,q}}$$

Hence, the inner derivative of the top-one probability with respect to coordinate ω_j is:

$$\frac{\partial P_{\hat{y}^{(q)}(f_{\omega})}(x_{i}^{(q)})}{\partial \omega_{j}} = \frac{e_{i,q} x_{i,j}^{(q)} \cdot \sum_{k=1}^{n} e_{k,q} - e_{i,q} \sum_{k=1}^{n} e_{k,q} \cdot x_{k,j}^{(q)}}{\left(\sum_{k=1}^{n} e_{k,q}\right)^{2}}$$
(10)

We summarize these equations into a vector ω such that we can write a single equation:

$$\frac{\partial P_{\hat{y}^{(q)}(f_{\omega})}(x_i^{(q)})}{\partial \omega} = \frac{e_{i,q}x_i^{(q)} \cdot \sum_{k=1}^n e_{k,q} - e_{i,q}\sum_{k=1}^n e_{k,q} \cdot x_k^{(q)}}{\left(\sum_{k=1}^n e_{k,q}\right)^2}$$
(11)

Finally the gradient becomes,

$$\frac{\partial U(\hat{y}^{(q)})}{\partial \omega} = 2\left(\frac{1}{|G_0|} \sum_{x_i^{(q)} \in G_0} P_{\hat{y}^{(q)}(f_\omega)}(x_i^{(q)}) \cdot v_1 - \frac{1}{|G_1|} \sum_{x_i^{(q)} \in G_1} P_{\hat{y}^{(q)}(f_\omega)}(x_i^{(q)}) \cdot v_1\right) \\ \cdot \left(\frac{1}{|G_0|} \sum_{x_i^{(q)} \in G_0} \frac{\partial P_{\hat{y}^{(q)}(f_\omega)}(x_i^{(q)})}{\partial \omega} \cdot v_1 - \frac{1}{|G_1|} \sum_{x_i^{(q)} \in G_1} \frac{\partial P_{\hat{y}^{(q)}(f_\omega)}(x_i^{(q)})}{\partial \omega} \cdot v_1\right)\right)$$

5 EXPERIMENTS

In our experiments, we consider three real-world datasets that help us study *non-discrimination*, through experiments that seek to reduce exposure disparities due to biases that are unrelated to utility (experiments 6.1 and 6.3), and *equality of opportunity*, through experiments that seek to reduce exposure disparities due to utility differences that pre-exist (experiments 6.2, 6.4 and 6.5). As explained in the introduction, these are the two prototypical cases for applying this kind of method [27]. The datasets are presented in subsections 5.1, 5.2 and 5.3. *All of the data and code required, plus* instructions to reproduce all the experiments, will be available with the camera-ready version of this paper.

We apply DELTR with two different values of γ to each dataset, and compare the results against several baselines: (i) a "colorblind" learning-to-rank approach, which excludes protected attributes during training; (ii) a standard learning-to-rank method, which considers them during training; (iii) a post-processing approach that applies learning to rank and then re-ranks the output; and (iv) a pre-processing approach that modifies the training data. Baselines are described in subsection 5.4.

5.1 W3C experts (TREC Enterprise) Dataset

This dataset originates from the expert search task at the TREC 2005 Enterprise Track [10], where an algorithm has to retrieve a sorted list of experts for a given topic, given a corpus of e-mails written by possible candidates. It contains 198,395 mail messages in mailing lists of the World Wide Web Consortium (W3C). A series of 60 topics and a list of hand-picked experts for each topic are provided; each list contains between 7 and 20 experts, where the available expert relevance judgments are binary (expert or non-expert), hence all experts are considered equally expert. The number of candidates (people who authored at least one e-mail) is 1092.

We computed a series of text retrieval features for each querydocument pair, such as word count and normalized tf-idf scores by usage of the Learning to Rank ElasticSearch Plug-in [24]. Furthermore we created a set of query expansion terms for each topic manually to improve the quality of retrieval.

We consider a scenario in which women are the protected group. We manually attributed gender to each candidate on the basis of their given names. Women comprise 10.5% of e-mail authors and on average 13.93% of experts across queries. To create training data, for each query, we created a list of 200 people with all experts at the top, followed by random non-experts sampled using the same male/female proportion as for the entire set of candidates.

Given that, to the best of our knowledge, no real-world dataset proven to contain discriminatory patterns is readily available for document retrieval tasks, we injected a discriminatory pattern in this dataset. This was done by sorting the training list for each training query in the following order: 1. all male experts, 2. all female experts, 3. all male non-experts, and 4. all female non-experts. This simulates a scenario where expertise has been judged correctly, but training lists have been ordered with a bias against women, placing them systematically below men having the same level of expertise. For our training data we created a six-fold cross-validation dataset, (12) each containing 50 training queries and 10 testing queries. We will release the dataset with the camera-ready version of this paper.

5.2 Engineering Students Dataset

This dataset corresponds to the task of sorting a list of applicants to an engineering school by predicted academic performance, for instance, to give a scholarship to the k most promising candidates. Academic performance is measured by grades after the first year in university. The dataset contains anonymized historical information from first-year students at a large school in a university (name withheld for double-blind review). It covers 5 years and on average 675 students per year. Most of them (94.17%) are admitted

to the university based on a standardized, country-wide university admission test (details withheld for double-blind review) and their high-school grades. Other students are admitted through positive action programs aimed at outstanding students from public schools (4.06%), women right below the cut-off score (1.04%), and students who excel in sports (0.74%).

For each student, the following features are available: (i) their average high-school grades, computed using an official standardization formula; (ii) their scores in math, language, and science in a standardized test; (iii) the number of university credits taken, passed, and failed during the first year; and (iv) their average grades at the end of their first year. Additionally, we are given the gender for each student, and whether they come from a public high school or from a private one. We considered two scenarios, one in which women are the protected group (they comprise less than 21% of students), and one in which students from public high schools are the protected group (measures of educational quality have consistently shown public high schools lag behind private ones in this country [reference withheld for double-blind review]).

We created a five-fold cross-validation setup in which each fold contains four classes (years) of students for training and one class for testing. The ground truth is created by sorting students by decreasing grades upon finishing the first year, in which grades are weighted by the credits of each course they passed (and divided by the total number of credits they took). We will provide instructions for access to this dataset and a cleaning script with the camera-ready version of this paper.

5.3 Law Students Dataset

This dataset originates from a study by Wightman [31] that examined whether the LSAT (Law Students Admission Test in the US) is biased against ethnic minorities. It contains anonymized historical information from first-year students at different law schools and consists of 21,792 students in total. We use a uniform sample of 10% of this dataset, while maintaining the distribution of gender and ethnicity respectively. Our training data corresponds to 80% of all candidates in this sample, and the ground truth is created by sorting students by decreasing grades upon finishing the first year. For the different experiments we divided the candidates into protected and non-protected groups. When dividing by gender, we consider women as the protected group. They comprise 44% of the students in the dataset, compared to 21% in the engineering students dataset; there is also a much smaller difference in terms of academic performance between men and women than in the engineering students dataset. When dividing by ethnicity, we consider minority 'Black' (the term used in the US census for African American) as the protected group, while the non-protected group is the majority ('White'). 'Blacks' comprise less than 6% of the students. For each student, the following features are available: (i) average high-school grades; (ii) scores in the LSAT, a law-specific scholastic assessment test; and (iii) average grades at the end of the first year. Additionally, we are given the gender and ethnicity for each student.

The experimental setting is very similar to the engineering students and also simulates a scenario in which one seeks to anticipate the top-k students, e.g., to give a benefit such as a scholarship. We predict academic performance after the first year in university on the basis of high school grades and the LSAT score.

5.4 Baselines

We compare DELTR with a small and a large value for γ to the following pre-, in- and post-processing approaches: since we are still lacking another in-processing fair learning-to-rank approach, our in-processing baselines constitute (i) a "colorblind" LTR approach in which a standard learning to rank algorithm (Cao et al. [6], i.e., DELTR with $\gamma = 0$) is applied over all the non-sensitive attributes; and (ii) the same learning to rank approach in which all attributes are used (including the protected attribute). Note that in real life, a "true" colorblind result is likely to be unachievable, because of dependencies between protected and non-protected attributes (red lining effect [5]). These dependencies become particularly problematic in cases of different baselines across groups, in the sense that e.g. the same result in a test corresponds to different levels of performance or expertise, because the test's design happens to favor the non-protected group. We will investigate on this in section 6.3. Therefore a method is needed that neither takes any assumptions on correlations of features to sensitive characteristics such as standard LTR, nor is blind to biased non-sensitive features such as colorblind LTR. DELTR constitutes such a method.

In the *pre*- and *post-processing* baselines we apply the algorithm FA*IR [35] (i) to the training data in order to reduce bias before model training; and respectively (ii) to the rankings predicted by a standard LTR method to increase exposure of the protected group if necessary. FA*IR is a top-k ranking algorithm that ensures a minimum proportion of a protected group at every position in a ranking based on a statistical significance test. With a given minimum target proportion p of protected candidates, FA*IR reorders a given predicted ranking and places protected items at higher positions, if its statistical test would fail otherwise. If a top-k ranking passes the test, it just orders items according to their predicted scores.

In our *pre-processing* baseline experiments we process a given training dataset with FA*IR to free the data from potential bias and create fair training data. We use three different values of p, $p^* =$ the ratio of protected candidates in the dataset, $p^+ = p^* + 0.1$ and $p^- = p^* - 0.1$, to show how crucial the right choice of p is, especially in a pre-processing setting. Afterwards we use standard LTR [6] to train a ranker over all features, both sensitive and non-sensitive. The *post-processing* baseline uses the same learning to rank algorithm [6] and trains a ranker over all available features, including the protected one. Afterwards FA*IR is applied to the predicted rankings, potentially resulting in a reordering of the items. We use the same parameters p^* , p^+ and p^- as in the pre-processing experiments.

6 EXPERIMENTAL RESULTS

In this section we present the results of each experimental setting. Figure 1 depicts the relations between exposure of the protected group and overall relevance. Figure 2 illustrates the distribution of men and women across ranking positions for experiment 6.1.



Figure 1: (Best seen in color.) Comparison of relevance and exposure achieved by each approach. Each plot (a)-(e) is one experimental setting consisting of a dataset and a protected attribute; (f) contains the legend. In all plots, relevance is in the x axis, while the exposure of the protected group relative to the nonprotected group is in the y axis. We see that a trade-off between exposure and relevance is not a law of nature. Instead its presence or absence depends on the concrete underlying bias in the training data (plot (b) vs plot (c)). In case we observe a trade-off between performance and exposure (plot (b), (d) and (e)), DELTR mostly outperforms the pre- and post-processing approaches. The plots focus on high-relevance results, hence settings that obtain substantially lower relevance are omitted, their approximate position can be inferred from the respective lines joining settings using the same approach with different parameters. We give the respective value of γ and p^* in each evaluation subsection.

6.1 W3C experts (gender)

Experimental results are shown on Figures 1a and 2, averaged over all folds, using $\gamma_{small} = 20K$, $\gamma_{large} = 200K$ and $p^* = 0.105$, which is the proportion of women in the dataset.

In this experiment we expect the "colorblind" approach to achieve the best results, because we injected a strong bias against women *that was completely unrelated to their expertise.* This setting corresponds to a *non-discrimination case*, where we want to exclude the protected feature from training for relevance reasons and we expect to see no trade-off between relevance and exposure by usage of DELTR. Instead we expect that a larger value for γ corresponds to better exposure *and* better relevance.

Figures 1a and 2 confirm our expectations. Note that for this experiment we measure utility in terms of precision in the top ten positions instead of Kendall's tau, because we want to know which algorithm finds most of the true experts and ranks them accordingly. Colorblind L2R performs best in terms of relevance and achieves almost equal exposure for men and women, by distributing women evenly across rankings (Figure 2a). Standard L2R (including the

biased protected feature) performs worse in terms of relevance and exposure than most of the other approaches. Indeed, it exaggerates the bias against women, placing all women at the bottom of the ranking, as shown in Figure 2b, even those that were considered experts in the ground truth.

Our goal for DELTR in this experiment is therefore to get closest to the colorblind results and away from the standard L2R results. We achieve this goal, as shown in Figure 1a, where DELTR with a large γ and post-processing FA*IR with a large p produce the best results. DELTR reduces the gap in exposure between men and women, and scores best in terms of relevance compared to all other fair algorithms. Post-processing with p^+ achieves better exposure, but leads to a slight over-representation of women at the top-positions (Figure 2h), which causes the drop in relevance (Figure 1a).

A few more things are worth noticing: First, we see that the choice of p is crucial for the success of FA*IR and that this choice is not trivial. Particularly as a pre-processing approach (orange " \overline{F} " symbols), using the intuitive p^* , which corresponds to the proportion of women in the dataset, does not help to de-bias the L2R



Figure 2: (Best seen in color.) Distribution of men (blue) and women (orange) along rankings with DELTR and different baselines in the W3C experts dataset, using gender as the protected attribute. The training data has been biased against women by breaking ties in expertise always in favor of men. At the top we see that the colorblind learning-to-rank approach distributes women evenly (a), while the inclusion of gender leads learning-to-rank to place all women in the bottom positions (b). Plots (c), (d), (e) show the results of pre-processing: depending on the parameter used, this can quickly go from no change at all in exposure for women to the other extreme, in which women are placed in top positions because of their protected feature. Plots (f), (g), (h) show the results of post-processing: re-ranking the output of standard learning-to-rank algorithm may increase the share of women in the top positions. However if the expected proportion parameter p is not set well, it can lead to an *under- and over-representation* of women. Plots (i) and (j) show the results of DELTR, which reduces the impact of the biased training set to the model. When γ is too small, DELTR may behave similar to standard learning to rank and cause an under-representation of women in the top-positions. Contrary to the pre- and post-processing methods, *over-representation of the protected group is not possible* due to the design of the algorithm, even if γ were set to a very large value.

model. Figure 2d shows that this setting produces exactly the same biased distribution as standard L2R. Also in all other cases of FA*IR, post- and pre-processing, a too small p does not show any effect on the exposure of women in the rankings (Figure 2c and Figure 2f). In contrast, DELTR by design always results in better exposure, even if *y* is set low. The change may be little but it is never zero, unless $\gamma = 0$. Second, a too large p can result in an over-representation of protected elements at the top positions (Figure 2e and Figure 2h). Especially in the pre-processing case (Figure 2e) this may result into inverting the bias, such that non-protected items are now ranked lower by a standard LTR method because of their group membership. In addition this leads into a profound decline of result relevance, as shown in Figure 1a, where pre-processing using p^+ not only increases disparate exposure to the detriment of the non-protected group, but also performs significantly worse than all other approaches in terms of relevance.

We note that, in contrast to FA*IR, DELTR by design excludes the risk of reverse discrimination.

6.2 Engineering students (gender)

Figure 1b summarizes the averaged results obtained by applying each ranker on all five cross-validation folds ($\gamma_{small} = 3K$, $\gamma_{large} = 50K$, $p^* = 0.202$, which is the proportion of women in this dataset). We know that women score worse than men in the university entrance test and also worse in terms of academic success. We therefore expect a trade-off between utility and exposure, if we optimize for more exposure than the protected group should receive based on their true performance. This is desirable if one wants to achieve equality of opportunity.

In terms of performance the best approaches are standard L2R together with DELTR with small γ and FA*IR with small values of p. Among these, DELTR achieves the highest exposure for the protected group. In contrast, pre- and post-processing FA*IR with small ps do not have any effect on the rankings, which means that they produce the same results as a standard L2R setting. This is confirmed by the fact that their markers are very close to the standard L2R one (blue cross).

Compared to standard L2R, colorblind scores slightly less in terms of relevance, but a lot better in terms of exposure. Because women tend to have lower scores in the standardized test and also lower grades after the first university year, we expect neither colorblind nor standard L2R to be close to the line of total equality of exposure. The standard learning-to-rank however, by usage of the protected feature emphasizes the disadvantage of women and puts them even lower then they deserve. This problem can be mitigated using both DELTR and post-processing FA*IR. If one seeks to increase the exposure of women beyond the colorblind result, due to, e.g., requirements on equal opportunities or affirmative action policies, Figure 1b shows that this comes with a penalty in terms of relevance. Post-processing with p^+ and DELTR with a large γ come closest to equal exposure across groups (a ratio of 1 means total equality), but the relevance decrease is smaller for DELTR. This means when using DELTR we trade less relevance for the same fairness achievement in a search result than when using FA*IR.

As in the previous experiment, pre-processing with p^+ inverts the disparate exposure problem by ranking female students to the top positions just because they are female, which also causes a decline in relevance. This shows again that trying to de-bias a dataset before with a "fair" algorithm, can lead to reverse discrimination.

6.3 Engineering students (high school type)

In this experiment, we consider students coming from public high schools as the protected group and those from private high schools as the non-protected. Results appear in Figure 1c ($\gamma_{small} = 100K$, $\gamma_{large} = 5M$ and $p^* = 0.348$, which is the proportion of students from public high schools).

We know that students from public schools perform worse on average in the entrance test, but tend to have higher grades in university than students from private high schools with the same scores in the standardized test. One explanation is that public schools tend to provide an education of inferior quality compared to private schools in the country under study. This means that for achieving the same test scores, students from public schools need to have better academic aptitudes and/or more grit than students from private ones (similarly to observations by [16] with respect to community colleges in the US). Under these circumstances, which also correspond to a case of non-discrimination, *including* the protected attribute will lead to better performance in terms of relevance *and* exposure. We therefore expect the colorblind to be among the worst approaches and standard L2R to be among the best.

Given that, our goal for this experiment is to achieve results close to standard L2R and away from the colorblind results. The results in Figure 1c confirm our expectations. We can see that the colorblind method performs significantly worse than most approaches both in terms of exposure and in terms of relevance. DELTR, given that students from the protected group already receive more exposure than the students in the non-protected group in learning to rank, does not further increase their exposure, preserving the quality of the ranking result, both for small and large values of γ (due to the asymmetry of the method from the hinge loss function that we use). The same is true for FA*IR in pre- and post-processing with small values of *p*. As in the previous experiments, they *always* behave like standard L2R and do not change the rankings. DELTR only behaves like standard L2R when the exposure of the protected group exceeds the one of the non-protected group. In the post-processing setting FA*IR with p^+ achieves equal exposure ratios as DELTR, but less relevance. In the pre-processing experiment, if p becomes too large (p^* and p^+), more candidates than necessary are pushed towards the top positions in the training data, which in turn leads the LTR algorithm to place too much weight on the protected feature, resulting in a decline of relevance.

6.4 Law students (gender)

Figure 1d summarizes the results obtained by the different methods with $\gamma_{small} = 3K$, $\gamma_{large} = 50K$ and $p^* = 0.437$ (which is the proportion of women in this dataset). We observe that the standard L2R approach performs marginally better than colorblind in terms of relevance, which suggests that women's LSAT results lag slightly behind men's in the ground truth. However, the standard LTR exacerbates this small difference and the loss of exposure for women is large. DELTR corrects this mistake. While the experiment with a small γ only shows a marginal decrease in performance but a comparatively large increase in mitigating disparate exposure, DELTR with large γ achieves basically the same relevance as the colorblind approach.

Interestingly the post-processing setting for FA*IR did not perform well in this experiment, which suggests that its success depends on the underlying true score distribution of the training dataset. With all values p^*, p^- and p^+ we observe a substantial decrease of relevance, for the same achievements in exposure for women in comparison to DELTR. We interpret this as "too many women that performed poorly being pushed to the higher positions" by FA*IR, as it does not take relevance into consideration, while DELTR will balance exposure and relevance of a candidate.

Again the pre-processing approach with p^* and p^- show the same results as standard LTR, while in case of p^+ the results exhibit a too high exposure and a huge decline in relevance (the point lies far outside of the plot boundaries). This means that using FA*IR to de-bias the training data did not produce any meaningful fair results in this experiments.

6.5 Law students (race)

In this experiment African American students, the largest minority group in this dataset (6.4%), are considered the protected group. Results appear in Figure 1e ($\gamma_{small} = 1M$, $\gamma_{large} = 50M$ and $p^* = 0.064$). We did not use p^- because it would have been a negative number. Colorblind learning to rank places black students a little lower than white students, which corresponds to their lower high school and admission test performance found during our data analysis. In contrast, the standard learning to rank approach weights the sensitive feature 'race' to such extent that all black students are relegated to the lowest positions (not shown in the figures), even those students who scored well in terms of grades and standardized tests. This experiment corresponds to a scenario in which both objectives, non-discrimination and equal opportunity, are relevant simultaneously. Non-discrimination is required because standard L2R places all black students to the bottom of the ranking, even those who performed well in LSAT and have good results in university. DELTR corrects this bias, with a penalty in terms of relevance. Non-discrimination explains also, why colorblind achieves better relevance than the other methods. If we want to additionally achieve equal opportunity for blacks and increase their overall exposure beyond the colorblind level, this comes with a penalty in terms of relevance. DELTR allows us to balance between exposure and performance of the students, depending on how we choose γ .

The pre-processing method with p^* achieves approximately the same exposure as DELTR with a small γ , but less relevance. The post-processing approach with p^* performs slightly better in terms of exposure than DELTR with small γ , however from these three, it shows the least relevance. In case of p^+ both pre- and post-processing results show poor relevance, because they overcompensate for the protected feature.

Note again that the design of DELTR prevents the model from overrating protected candidates at the top positions, which avoids a severe decline of precision. DELTR optimizes for exposure until equality is achieved and not further. Then, only non-sensitive features are considered.

7 CONCLUSIONS

Rankings obtained using learning to rank can reproduce and exaggerate differences in exposure between groups that can be present in training data. In this paper we have presented DELTR, which extends a list-wise learning to rank method with an objective that reduces the extent to which non-protected elements receive less exposure than protected elements.

Our experiments showed that optimizing for fairer results does not necessarily come with a trade-off in relevance. On the contrary, when the training data is strongly biased against the protected group, without any relationship with utility, aiming for fair search results will *increase* relevance. We also showed that this objective can be achieved by *explicitly excluding* or *explicitly including* the protected feature. As it is hard to understand a-priori, what kind of underlying bias is present in the training data, and whether to include or exclude the protected feature, DELTR provides a convenient approach, which can handle both situations. At the same time it maintains high relevance compared to other state-of-the-art fair ranking approaches, and critically, it cannot "overcompensate" due to its asymmetry.

Limitations and future work. The parameter γ provides great flexibility for combining in the objective relevance with respect to the training data and avoiding large differences in exposure. To set it, we looked at the scale of *L* and *U*, which depend on many factors including the number of items to be ranked, and then started with a value of γ that reflected the ratio of these scales. However, more work is required to provide a systematic way of setting this parameter, and to understand the implications of different values.

Our differential exposure criterion of Equation ?? can be easily extended to multiple protected groups, for instance, by considering the maximum difference in exposure between the non-protected and any protected group. However, this needs to be experimentally validated.

Reproducibility. All of the data and code required, as well as instructions for reproducing all the experiments we have presented, including code implementing DELTR, will be made publicly available with the camera-ready version of this paper.

REFERENCES

- Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In Proceedings of the second ACM international conference on web search and data mining. ACM, 5–14.
- [2] Kristen M Altenburger, Rajlakshmi De, Kaylyn Frazier, Nikolai Avteniev, and Jim Hamilton. 2017. Are There Gender Differences in Professional Self-Promotion? An Empirical Case Study of LinkedIn Profiles Among Recent MBA Graduates. In ICWSM. 460–463.
- [3] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. arXiv:1805.01788 (pre-print) (2018).
- [4] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research), Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 149–159. http: //proceedings.mlr.press/v81/binns18a.html
- [5] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery 21, 2 (2010), 277–292.
- [6] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In Proceedings of the 24th international conference on Machine learning. ACM, 129–136.
- [7] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 335–336.
- [8] Carlos Castillo. 2018. Fairness and Transparency in Ranking. SIGIR Forum 52 (12 2018). Issue 2.
- [9] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2017. Ranking with fairness constraints. arXiv:1704.06840 (pre-print) (2017).
- [10] Nick Craswell, Arjen P de Vries, and Ian Soboroff. 2005. Overview of the TREC 2005 Enterprise Track. In Trec, Vol. 5. 199-205.
- [11] J. Shane Culpepper, Fernando Diaz, , and Mark D. Smucker (editors). 2018. Report from the Third Strategic Workshop on Information Retrieval (SWIRL). (2018).
- [12] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (10 2018).
- [13] Marina Drosou, HV Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. 2017. Diversity in Big Data: A Review. *Big Data* 5, 2 (2017), 73–84.
 [14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard
- [14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In Proc. of Innovations in Theoretical Computer Science (ITCS). ACM, 214–226.
- [15] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. ACM Transactions on Information Systems 14, 3 (1996), 330–347.
- [16] Jennifer Glynn. 2019. Persistence: The Success of Students Who Transfer from Community Colleges to Selective Four-Year Institutions. Technical Report.
- [17] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2125–2126.
- [18] Moritz Hardt. 2014. How big data is unfair: Understanding sources of unfairness in data driven decision making. *Medium* (2014).
 [19] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in
- [19] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (NIPS). 3315–3323.
- [20] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS) 20, 4 (2002), 422-446.
- [21] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In ACM SIGIR Forum, Vol. 51. Acm, 4–11.
- [22] Kasper Lippert-Rasmussen. 2014. Born free and equal?: a philosophical inquiry into the nature of discrimination. Oxford University Press.
- [23] Adam Liptak. 2009. Supreme Court finds bias against white firefighters. The New York Times A 1 (2009), A13.
- [24] OpenSource Connections. 2017. We're Bringing Learning to Rank to Elasticsearch. https://opensourceconnections.com/blog/2017/02/14/elasticsearchlearning-to-rank/.
- [25] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 560–568.
- [26] Stephen E Robertson. 1977. The probability ranking principle in IR. Journal of documentation 33, 4 (1977), 294–304.
- [27] John E Roemer. 1998. Equality of opportunity. Harvard University Press.
- [28] Ashudeep Singh and Thorsten Joachims. 2017. Equality of Opportunity in Rankings. In Workshop on Prioritizing Online Content (WPOC) at NIPS.
- [29] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. arXiv:1802.07281 (pre-print) (2018).
- [30] Indrė Žliobaitė. 2015. A survey on measuring indirect discrimination in machine learning. arXiv:1511.00148 (pre-print) (2015).

- [31] Linda F Wightman. 1998. LSAC National Longitudinal Bar Passage Study. LSAC
- [32] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In Proc. of International Conference on Scientific and Statistical Database Management or Conference on Scientific and Statistical Database Management (SSDB). ACM, 22.
- [33] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, HV Jagadish, and Gerome Miklau. 2018. A Nutritional Label for Rankings. arXiv:1804.07890 (pre-print) (2018).
- [34] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learn-ing classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1171–1180.
- Conterences steering Committee, 11/1–1180.
 [35] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA'IR: A fair top-k ranking algorithm. In Proc. of the 2017 ACM on Conference on Information and Knowledge Management. ACM, 1569–1578.