On the Problem of Underranking in Group-Fair Ranking

Sruthi Gorantla [1] Amit Deshpande [2] Anand Louis [1]

Abstract

Bias in ranking systems, especially among the top ranks, can worsen social and economic inequalities, polarize opinions, and reinforce stereotypes. On the other hand, a bias correction for minority groups can cause more harm if perceived as favoring group-fair outcomes over meritocracy. Most group-fair ranking algorithms post-process a given ranking and output a group-fair ranking. In this paper, we formulate the problem of underranking in group-fair rankings based on how close the group-fair rank of each item is to its original rank, and prove a lower bound on the trade-off achievable for simultaneous underranking and group fairness in ranking. We give a fair ranking algorithm that takes any given ranking and outputs another ranking with simultaneous underranking and group fairness guarantees comparable to the lower bound we prove. Our experimental results confirm the theoretical trade-off between underranking and group fairness, and also show that our algorithm achieves the best of both when compared to the state-of-the-art baselines.

1. Introduction

Search and recommendation systems have revolutionized the way we consume an overwhelming amount of data and find relevant information quickly (Brin & Page, 1998; Adomavicius & Tuzhilin, 2005). They help us find relevant documents, news, media, people, places, products and rank them based on our interests and intent (Kofler et al., 2016; Pei et al., 2019). Information presented through ranked lists influences our worldview (Pariser, 2011; Tavani). Rankings not only influence the users who consume them but also act as vehicles of opportunities for the items being ranked. Biased ranking of news, people, products raises ethical concerns and can potentially cause long-term economic and societal harm to demographics and businesses (Noble, 2018). Many state-of-the-art rankings that maximize utility or relevance reflect existing societal biases and are often oblivious to the societal harm they may cause by amplifying such biases. When these systems amplify societal biases observed in their training data, they worsen social and economic inequalities, polarize opinions, and reinforce stereotypes (O'Neil, 2016). In addition to ethical concerns, there are also legal obligations to remove bias. Disparate impact laws prohibit even unintentional but biased outcomes in employment, housing, and many other areas if one group of people belonging to a protected group is adversely affected compared to another (Barocas & Selbst, 2016). Protected groups could vary for specific statutes and include race, gender, age, religion, national origin, etc.

Fairness in ranking. Fairness in ranking has three broad requirements: *sufficient presence* of items belonging to different groups, *consistent treatment* of similar individuals, and *proper representation* to avoid representational harm to members of protected groups (Castillo, 2019). The first and the third requirements are about fairness to groups, whereas the second requirement is about fairness to individuals. For example, diversity alone in top ranks satisfies sufficient presence but need not provide consistent treatment and proper representation in the way the items are ranked. Fair ranking algorithms can be divided into two categories. Re-ranking algorithms that modify a given ranking of high utility to incorporate fairness constraints while trying to preserve the original utility, and learning-to-rank algorithms that incorporate fairness and utility objectives together into learning a ranker from training data. Re-ranking can be used to post-process the prediction as well as to pre-process the training data of any given ranker.

Most of the fair ranking algroithms are designed to output group-fair ranking. Group fairness in machine learning literature has focused on outcome-based or proportion-based definitions of fairness (e.g., demographic parity, equality of opportunity) (Hardt et al., 2016; Barocas et al., 2019). Although group-fairness is a desirable goal, affirmative action to achieve group-fairness is often misinterpreted as non-meritocratic by individuals and requires a deeper understanding (Crosby, 2004). In this context, we argue that

[1] Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India. [2] Microsoft Research, Bangalore, India.. Correspondence to: Sruthi Gorantla <gorantlas@iisc.ac.in>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

it is important to measure the worst-case effect of group-fair ranking on the individuals, which is not addressed by previous work.

Our contributions. To the best of our knowledge, our paper is the first to study how group-fair re-ranking affects individual ranks in the worst case. Previous work has looked at re-ranking or learning-to-rank with group-fairness and aggregate individual-fairness (or consistency) constraints. Our group-fairness definition ensures sufficient presence of all groups, similar to previous work, but we give a new, natural definition called *underranking* to study how re-ranking affects the merit-based ranks of individual items in the worst case. Our main contributions can be summarized as follows.

- We define underranking based on the worst-case deviation from the true, merit-based (or color-blind) rank of any individual item (see Definition 2.2) during re-ranking for group fairness. This directly captures the loss of visibility suffered by individual items of high merit in the process to achieve high group-fairness. We prove a lower bound on the trade-off achievable between underranking and group-fairness simultaneously.
- We propose a re-ranking algorithm that takes a given merit-based (or color-blind) ranking and outputs another ranking with simultaneous underranking and group-fairness guarantees, comparable to the lower bound mentioned above. Our algorithm is fast, flexible, and can accommodate multiple groups, each with a different constraint on their group-wise representation.
- We do extensive experiments to show that our algorithm achieves the best of both underranking and group-fairness compared to the baselines on standard real-world datasets such as COMPAS recidivism and German credit risk. Moreover, our algorithm runs significantly faster than the baselines.

Related work. The two most important baselines related to our work are the group-fair re-ranking algorithms (Celis et al., 2018; Zehlike et al., 2017). Fair ranking to maximize ranking utility subject to upper and lower bounds on group-wise representation in the top k ranks, for all values of k, can be framed as an integer optimization problem (Celis et al., 2018). The authors propose an exact dynamic programming (DP)-based algorithm, and a greedy approximation algorithm to achieve fairness in ranking for intersectional subgroups. The *fair top-k ranking problem* gives another formulation for fair re-ranking of a given true or *color-blind* ranking based on numerical quality values and a given k, so that the top-k re-ranking maximizes certain selection and ordering utilities subject to group-wise representation constraints (Zehlike et al., 2017). The authors give an efficient algorithm called FA*IR to solve the fair top-k ranking problem.

Fair ranking has also been studied in the learning-to-rank (LTR) setting, where the output ranking is probabilistic, so the fairness and utility guarantees are often on average instead of the worst case. Given a query-document pair, the probability of each document being ranked at the top rank is called its *exposure*. While the traditional ListNet framework simply minimizes a loss function based on the items' true and predicted exposure (Cao et al., 2007), an extension of this, DELTR (Zehlike & Castillo, 2020), learns fair ranking via a multi-objective optimization that maximizes utility and minimizes disparate exposure for different groups of items for group-fairness or different items for individual-fairness. This general learning-to-rank framework facilitates optimizing multiple utility metrics while satisfying equal exposure, and Fair-PG-LTR (Singh & Joachims, 2019) learns a ranking that satisfies fairness of exposure. Aggregate or average-case guarantees in ranking are more suited to the applications in search and recommendation, whereas the worst-case guarantees are more suited to the applications in recruitment, school admissions, healthcare etc. where the worst-case fairness to individuals could be critical.

There is related work on defining and maximizing various group-fairness metrics over each prefix of the top k ranks (Yang & Stoyanovich, 2017), for a given k, using an optimization algorithm to learn fair representations (Zemel et al., 2013). There are also other measures of group-fairness in ranking based on pairwise comparisons (Narasimhan et al., 2020; Beutel et al., 2019). Recent work has also studied fairness-aware ranking in search and recommendations for real-world recruitment tools using fairness metrics based on skew in the top k and Normalized Discounted KL-divergence (NDKL) divergence (Geyik et al., 2019). Fairness and ranking utility trade-offs have also been studied via counterfactually fair rankings (Yang et al., 2020).

2. Underranking and Group Fairness

Preliminaries. Let $M, N \in \mathbb{Z}^+$ and $N \leqslant M$. Then, a ranking is an assignment of M ranks to N items such that each rank (denoted by a number in $[M]$) is assigned to at most one item (denoted by a number in $[N]$) and each item is assigned exactly one rank. Whenever a rank is not assigned to any item, we call it an *empty rank*. Note the whenever $N = M$, there are no empty ranks in the ranking. We say that rank i is *lower* than rank j if $i < j$, and rank i is *higher* than rank j if $i > j$. In a ranking the *top m ranks* refer to the ranks $(1, 2, \dots, m)$, *each prefix of the top m ranks* is the set $\{(1, 2, \dots, i) | i \in \{1, \dots, m\}\}$, *every k consecutive ranks in the top m ranks* is the set $\{(i + 1, i + 2, \dots, i + k) | i \in \{0, \dots, m - k\}\}$, *$i$th block of size k* is the ranks $((i - 1)k + 1, (i - 1)k + 2, \dots, (i -$

$1)k + k)$, and hence, *top d blocks of size k* is the set $\{((i-1)k+1, (i-1)k+2, \ldots, (i-1)k+k) | i \in \{1, ..., d\}\}$. A *true ranking* is the ranking of the items based on a measure of merit of the items. A re-ranking algorithm rearranges the items in the true ranking and outputs another ranking of the items with some desired properties. We note that a true ranking is not always available for the real-world datasets. In our experiments, we use some natural substitutes for the true ranking; see Section 3 for details. An item's *true rank* is its rank in the true ranking. The set of N items is partitioned into ℓ groups based on the sensitive attributes of the items. We denote each group by the subscript l whenever we refer to a group $l \in [\ell]$. In all that follows, α_l, $\beta_l \in [0, 1]$ such that $\alpha_l \geqslant \beta_l$, and $\gamma \geqslant 1$. The fairness constraints are based on the representation (number of items) from each group in the ranking, and are denoted by $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_\ell)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_\ell)$, where α_l, β_l represent the fairness constraints for group l. And for any $c \in \mathbb{R}$, $c\boldsymbol{\alpha} = (c\alpha_1, c\alpha_2, \ldots, c\alpha_\ell)$, and similary $c\boldsymbol{\beta} = (c\beta_1, c\beta_2, \ldots, c\beta_\ell)$.

We now formally define the notion of *group fairness* of a ranking of N items.

Definition 2.1 (Group Fairness)**.** A ranking is said to satisfy $(\boldsymbol{\alpha}, \boldsymbol{\beta}, k)$ group fairness if every k consecutive ranks have at most $\alpha_l k$ and at least $\beta_l k$ items from group l, for every group $l \in [\ell]$.

That is, each element, ranks $(i+1, ..., i+k)$, in the set of every k consecutive ranks in top N ranks is such that for each group $l \in [\ell]$, at most $\alpha_l k$ and at least $\beta_l k$ of these ranks are assigned to the group l. The set of top $\lfloor N/k \rfloor$ blocks of size k is a strict subset of the set of every k consecutive ranks in top N ranks. Therefore, any ranking that satisfies group fairness constraints for every k consecutive ranks in top N ranks also satisfies group fairness constraints for of the top $\lfloor N/k \rfloor$ blocks of size k.

Using the notion of *underranking*, we would like to capture how much an item has been displaced from its true rank during re-ranking for group fairness.

Definition 2.2 (Underranking)**.** A ranking satisfies γ underranking if the rank of each item is at most γ times its true rank.

We remark that unless the true ranking satisfies the group fairness conditions, some items with high merit must suffer a loss of visibility during the process of re-ranking for group fairness. That is, the output group fair ranking has underranking strictly greater than 1. This manifests the trade-off between the group fairness and the underranking in ranking.

Closely related to underranking is the well studied notion of PRECISION@K of ranking (Järvelin & Kekäläinen, 2000; Manning et al., 2008; Zehlike & Castillo, 2020). For a given ranking, PRECISION@K is defined as the number of items in the top K ranks of the true ranking which are still in the top K ranks after re-ranking. We get the following relation between underranking and PRECISION@K.

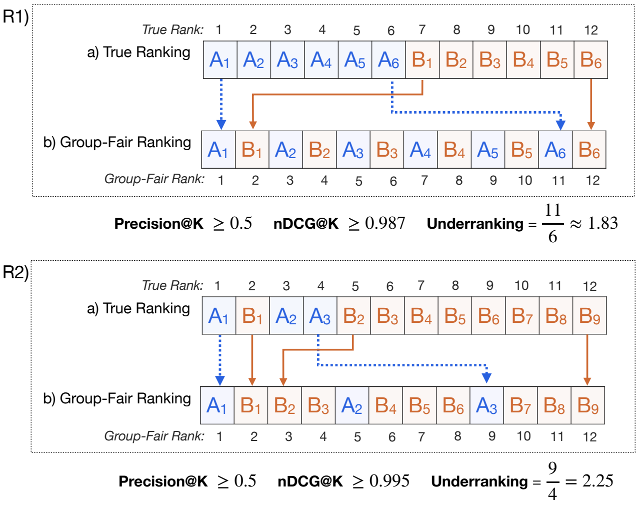

R1)
True Rank: 1 2 3 4 5 6 7 8 9 10 11 12
a) True Ranking
A1 A2 A3 A4 A5 A6 B1 B2 B3 B4 B5 B6
b) Group-Fair Ranking
A1 B1 A2 B2 A3 B3 A4 B4 A5 B5 A6 B6
Group-Fair Rank: 1 2 3 4 5 6 7 8 9 10 11 12
Precision@K ≥ 0.5 nDCG@K ≥ 0.987 Underranking = 11/6 ≈ 1.83
R2)
True Rank: 1 2 3 4 5 6 7 8 9 10 11 12
a) True Ranking
A1 B1 A2 A3 B2 B3 B4 B5 B6 B7 B8 B9
b) Group-Fair Ranking
A1 B1 B2 B3 A2 B4 B5 B6 A3 B7 B8 B9
Group-Fair Rank: 1 2 3 4 5 6 7 8 9 10 11 12
Precision@K ≥ 0.5 nDCG@K ≥ 0.995 Underranking = 9/4 = 2.25

Figure 1: High ranking utility does not imply better (lower) underranking in group-fair rankings.

Corollary 2.3. *A ranking satisfying γ underranking also has* PRECISION@K *at least* $\lfloor K/\gamma \rfloor$, $\forall K \in \mathbb{Z}^+$.

Proof. Fix a ranking having γ underranking. By definition, the top $\lfloor K/\gamma \rfloor$ items in the true ranking get displaced at most to the rank $\lfloor K/\gamma \rfloor \gamma \leqslant K$. Hence, at least the top $\lfloor K/\gamma \rfloor$ items in the true ranking are also in the top K ranks in ranking with γ underranking. Therefore, PRECISION@K is at least $\lfloor K/\gamma \rfloor$. □

We note that our definition of group fairness in ranking is slightly different from previous definitions in (Zehlike et al., 2017; Celis et al., 2018; Castillo, 2019). There, the group fairness constraints are at every prefix of the top k ranking, whereas, in Definition 2.1 group fairness constraints are for every k consecutive ranks. Our notion of group fairness has the desirable property that even if items from top few ranks are removed from the ranking, the remaining ranking still satisfies the group fairness constraints. Using this notion of group fairness, we propose Algorithm 1 that achieves simultaneous group fairness and underranking guarantees. Such theoretical guarantees are not available for the re-ranking algorithms with prefix group fairness constraints. We will also see in Section 3 that the algorithm proposed in this paper achieves better representation of the protected groups in the prefixes of the top k ranking as well.

We also note here that, even though low (better) underranking in the top K ranks implies high PRECISION@K, the converse need not be true. Consider two pairs of a true ranking and its corresponding group-fair ranking shown in

Figure 1, R1(a), R1(b) and R2(a), R2(b) with items from groups A and B. Both R1(b) and R2(b) satisfy proportional representation (50% of each group in R1 and 25% A's and 75% B's in R2) in every prefix of the ranking (ignoring the rounding errors), as well as in every k consecutive ranks, for a reasonable choice of k, which is 4 or 8. If we assume that in both R1(a) and R2(a), the merit of the item ranked at rank i is $1 - 0.01 * (i-1)$, then the nDCG (see Section 3 for the formula), which is also a ranking utility metric like PRECISION@K, is more for R2 than R1. We also note that, in both R1 and R2, the PRECISION@K in any prefix of the top 12 ranks is at least 0.5. From these two examples we observe that even though the utility of the group-fair ranking of R2 is greater than or equal to that of R1, the underranking in R2 is higher (worse) than that of R1. Hence, high ranking utility may not imply better underranking. This is also observed in our experiments on the real-world datasets.

2.1. Theoretical Results

Our first main result is a lower bound on the underranking when satisfying group fairness in blocks of size k.

Theorem 2.4 (Lower bound). *Fix $\ell \in \mathbb{Z}^+$. For each group $l \in [\ell]$, fix $\alpha_l, \beta_l \in (0,1] \cap \mathbb{Q}$ such that $\alpha_l \geqslant \beta_l$, $\sum_{l\in[\ell]} \alpha_l \geqslant 1$, and $\sum_{l\in[\ell]} \beta_l \leqslant 1$. Fix $k \in \mathbb{Z}^+$. For every $n_0 \in \mathbb{Z}^+$, there exists an n such that $n \geqslant n_0$, and there exists a true ranking of the $N = \ell n$ items grouped into ℓ groups of n items each, such that the following holds. Any ranking satisfying γ underranking (w.r.t. the true ranking) and $(\boldsymbol{\alpha}, \boldsymbol{\beta}, k)$ group fairness in the top $\frac{\gamma n}{k}$ blocks of size k must have $\gamma \geqslant \frac{1}{\min\{\alpha_{min}, 1-\sum_{l\neq l_*} \beta_l\}}$, where $\alpha_{min} = \min_l \alpha_l$ and $l_* = \operatorname{argmin}_l \beta_l$.*

Our next main result is a fair ranking algorithm that takes a true ranking and outputs another ranking with underranking and group fairness guarantees in any k consecutive ranks.

Theorem 2.5 (Trade-off 1). *Given a true ranking of N items grouped into ℓ disjoint groups, with each group having at least n items, and fairness parameters $k \in \mathbb{Z}^+$ and $\alpha_l, \beta_l, \forall l \in [\ell]$, where α_l, β_l define the upper and lower group fairness constraints for the group l respectively such that $0 \leqslant \beta_l \leqslant \alpha_l \leqslant 1$, $\sum_{l\in[\ell]} \alpha_l > 1$, $\sum_{l\in[\ell]} \beta_l < 1$. Let $\alpha_{min} := \min_l \alpha_l$, $\alpha_{max} := \max_l \alpha_l$, and $l_* = \operatorname{argmin}_l \beta_l$. Let $\epsilon := \frac{2}{k} \cdot \max\left\{ \left(1 + \frac{\ell}{\sum_{l\in[\ell]} \alpha_l - 1}\right), \left(1 + \frac{\ell}{1-\sum_{l\in[\ell]} \beta_l}\right), \max_{l\in[\ell]} \left(1 + \frac{2}{\alpha_l - \beta_l}\right) \right\}$.*
Then there exists a polynomial time algorithm to compute a ranking satisfying both of the following simultaneously,

1. $\frac{1}{\min\left\{\alpha_{min} - \frac{1}{\lfloor \epsilon k/2 \rfloor}, \left(1 - \sum_{l\neq l_*} \beta_l\right) - \frac{\ell - 1}{\lfloor \epsilon k/2 \rfloor}\right\}}$ *underranking,*
2. $((1+\epsilon)\boldsymbol{\alpha}, (1-\epsilon)\boldsymbol{\beta}, k)$ *group fairness in the top $\left\lfloor \frac{n}{\alpha_{max}} \right\rfloor - \lfloor \epsilon k/2 \rfloor$ ranks.*

Algorithm 1 ALG

```
Input: A true ranking of the $N$ items and parameters $\alpha_l, \beta_l, \forall l \in [\ell]$, and $k$ satisfying the conditions in Theorem 2.5.
Set $\epsilon, \alpha_{\min}, l_*$ as in Theorem 2.5, set $B := \left\lfloor \frac{\epsilon k}{2} \right\rfloor$
Set $b := \min\left\{ \lfloor \alpha_{\min} B \rfloor, B - \sum_{l \neq l_*} \lceil \beta_l B \rceil \right\}$
Set $M := \lceil N/b \rceil \cdot B$
for $i := \lceil N/b \rceil$ down to $1$ do
    for $j := 1$ to $\min\{b, N - (i-1)b\}$ do
        Move item at rank $(i-1)b + j$ to rank $(i-1)B + j$
    end
end
for each rank $j$ in $1$ to $M$ do
    if rank $j$ is empty then
        Set $i := \lceil j/B \rceil$
        for $j' := j + 1$ to $M$ such that rank $j'$ is not empty do
            Set $l :=$ group the item ranked at $j'$ belongs to
            if the lower bound for group $l$ in the block $i$ is not satisfied $\vee$ (lower bounds of all the groups are satisfied $\wedge$ upper bound for group $l$ would not be violated) then
                Move the item at rank $j'$ to rank $j$
                Break the loop
            end
        end
    end
end
for $j := 1$ to $N$ do
    if rank $j$ is empty then
        Move to rank $j$, the first item at rank higher than $j$
    end
end
Output final ranking from rank 1 to rank $N$
```

We note that ϵ need not be smaller than 1.

We also obtain slightly stronger guarantees if we only need group fairness in blocks of size k instead of group fairness guarantees for any k consecutive ranks. That is, in each of the top $\left\lfloor \frac{n}{\alpha_{\max} k} \right\rfloor$ blocks of size k, the output ranking has to be such that, for each group $l \in [\ell]$, at least $\beta_l k$ and at most $\alpha_l k$ ranks are assigned to items from group l.

Theorem 2.6 (Trade-off 2). *Given a true ranking of N items grouped into ℓ disjoint groups, with each group having at least n items, and fairness parameters $k \in \mathbb{Z}^+$, and $\boldsymbol{\alpha}, \boldsymbol{\beta}$, where α_l, β_l define the upper and lower group fairness constraints for the group l respectively such that $0 < \beta_l \leqslant \alpha_l \leqslant 1$, $\sum_{l\in[\ell]} \alpha_l > 1$ and $\sum_{l\in[\ell]} \beta_l < 1$, let $\alpha_{min} = \min_l \alpha_l$, $\alpha_{max} = \max_l \alpha_l$, and $l_* = \operatorname{argmin}_l \beta_l$. If the fairness parameters are also such that $\alpha_l k \in \mathbb{Z}^+$ and*

$\beta_l k \in \mathbb{Z}^+, \forall l \in [\ell]$, then there exists a polynomial time algorithm to compute a ranking satisfying both of the following simultaneously,

1. *$\frac{1}{\min\left\{\alpha_{min}, 1-\sum_{l \neq l_*} \beta_l\right\}}$ underranking*
2. *$(\boldsymbol{\alpha}, \boldsymbol{\beta}, k)$ group fairness in each of the top $\left\lfloor \frac{n}{\alpha_{max} k} \right\rfloor$ blocks of size k,*

2.2. Overview of the Algorithm 1 and Proof Outline.

We defer all the proofs to the supplementary material. Here we present an overview of Algorithm 1. We use Algorithm 1 to prove Theorem 2.5. We invoke Algorithm 1 with a different value of ϵ to prove Theorem 2.6.

Let ϵ, b, and B be as set in the algorithm. The ith *block* of size $\lfloor \epsilon k/2 \rfloor$ refers to the ranks $(i-1)\lfloor \epsilon k/2 \rfloor + 1$ to $i \lfloor \epsilon k/2 \rfloor$. We are given a true ranking of N items. Let $M = \lceil N/b \rceil \cdot B$. Then $M \geqslant N$. Hence we assume that the length of the true ranking is M such that the ranks $N+1$ to M are empty at the beginning of the algorithm. We first move the items to a rank higher than their true ranks in a fashion such that at the end of Step 8 the underranking of this intermediate ranking consisting of M ranks is bounded (see Lemma A.2 in supplementary). By our choice of parameters, this also guarantees that in each block the top $\min\left\{ \lfloor \alpha_{\text{min}} \lfloor \epsilon k/2 \rfloor \rfloor, \lfloor \epsilon k/2 \rfloor - \sum_{l \neq l_*} \lceil \beta_l \lfloor \epsilon k/2 \rfloor \rceil \right\}$ ranks are occupied and the rest of the ranks in the block are empty. Hence, the upper bound group fairness constraints in each block are not violated after Step 8. Next, we greedily fill up the empty ranks starting from the rank 1 while ensuring that the group fairness is not violated, until there are items available from each group. We use the fact that there are at least n items from each group to show that if there is any empty rank in the top $\lfloor n/\alpha_{\text{max}} \rfloor - \lfloor \epsilon k/2 \rfloor$ ranks, then there will be at least one higher ranked item available from each group which can be assigned to the empty rank without violating the condition in Step 14. Therefore, top $\lfloor n/\alpha_{\text{max}} \rfloor - \lfloor \epsilon k/2 \rfloor$ ranks will be unchanged after Step 20.

Then we fill the remaining empty ranks till N while ensuring that the underranking does not get worse. It is easy to show that after Step 25, each of the top $\left\lfloor \frac{n}{\lfloor \alpha_{\text{max}} \lfloor \epsilon k/2 \rfloor \rfloor} \right\rfloor$ blocks have at most $\lfloor \alpha_l \lfloor \epsilon k/2 \rfloor \rfloor$ items and at least $\lceil \beta_l \lfloor \epsilon k/2 \rfloor \rceil$ items from group l for each l. Finally we output the top N ranks. Observe that any k consecutive ranks must include some number of blocks completely, and will intersect at most two blocks partially. Therefore, in the worst case, the number of items from a group l in any k consecutive ranks of the top $\lfloor n/\alpha_{\text{max}} \rfloor - \lfloor \epsilon k/2 \rfloor$ ranks will be at most $\alpha_l k + 2\alpha_l \lfloor \epsilon k/2 \rfloor \leqslant \alpha_l(1+\epsilon)k$, and at least $\beta_l k - 2\beta_l \lfloor \epsilon k/2 \rfloor \geqslant \beta_l(1-\epsilon)k$. This gives us our group fairness guarantee in the top $\lfloor n/\alpha_{\text{max}} \rfloor - \lfloor \epsilon k/2 \rfloor$ ranks.

3. Experimental Validation

In this section, we give empirical observations about three broad questions – (i) Is there a trade-off between underranking and group fairness in the real-world datasets? (ii) How effective is underranking in choosing a group-fair ranking? (iii) Does ALG achieve best trade-off between group fairness and underranking?

Baselines. The baselines considered in this paper are described below,

1. **(Celis et al., 2018)'s DP algorithm:** In (Celis et al., 2018), W_{ij} represents the utility of the item i placed at rank j. Since we only have the scores (or relavance) of the item when placed at the top rank, we construct W_{ij} using positional discounting as described in the appendix of (Celis et al., 2018). We first sort the items in the decreasing order of their scores, which gives the true ranking. Wlog, let this ordering also represent the indices of the items, i.e., the item with highest score is indexed as item 1. Let y_i be the score of item i. Then, $W_{ij} = \frac{y_i}{\log_2(j+1)}$. Then W satisfies the Monge and monotonicity properties required by the DP algorithm in (Celis et al., 2018). Let there be ℓ groups the items can belong to. Let P_l contain the indices of the items that belong to group l, for each $l \in [\ell]$. Since we have N items, let $x \in \{0,1\}^{N \times N}$ be a ranking (or assignment) whose j-th column contains a one in the i-th position if item i is assigned to rank j. Note that each rank can be assigned to exactly one item and each item is assigned exactly one rank. Then, the fairness constraints at every prefix $k' \in [k]$ of the top k ranking are in the form of the following cardinality constraints,

$$L_{l,k'} \leqslant \sum_{1 \leqslant k' \leqslant k} \sum_{i \in P_l} x_{ik'} \leqslant U_{l,k'}.$$

The set of rankings that satisfy the above fairness constraints is represented with $\mathcal{B}$. Then the fair ranking problem posed as the following integer program,

$$\max_{x \in \mathcal{B}} \sum_{j \in [N]} \sum_{i \in [N]} x_{ij} W_{ij}.$$

The DP algorithm solves the above integer program exactly. In the expeirments with only one protected group represented by $l=1$ and one non-protected group represented by $l=2$, we use the lower bounds on the representation of the protected group, $L_{1,k'} = \lceil pk' \rceil, \forall k' \in [k]$ such that every prefix of the top k ranks has minimum p proportion of the items from the protected group. All other constraints are removed, i.e., $U_{1,k'} = k', U_{2,k'} = k', L_{2,k'} = 0$. In case of experiments with upper bound on the protected group (Figure 12 to Figure 17), we use

$U_{1,k'} = \lfloor pk' \rfloor, U_{2,k'} = k', L_{1,k'} = 0, L_{2,k'} = 0$, since we want a maximum proportion of p of the protected group in each prefix of the top k ranks. In our experiments, we show the trade-offs between representation and underranking by varying the parameter p. We use $p = p_l^* + \delta$ where p_l^* is the proportion of group l in the dataset and $\delta \in \mathbb{R}$.

2. (**Zehlike et al., 2017**): This greedy algorithm solves top-k ranking problem such that the proportion of the protected group stays significantly above the given minimum p, in every prefix $k' \in [k]$. Since, in our experiments with just one protected and one non-protected group, we want all the algorithms to achieve a minimum representation of $p^* + \delta$ of the protected group with true representation p^* and small number $\delta \in \mathbb{R}$, we run FA*IR by choosing the parameter $p = p^* + \delta$. Note that FA*IR can not handle the upper bound constraints on the representation of the groups. Hence in Figure 12 to Figure 17 we do not consider comparison with FA*IR.

In ALG and both the baselines, we choose $k = 100$.

Running Time and Space Complexity. Let $\phi = \min\{\alpha_{\text{min}} - \frac{1}{\lfloor \epsilon k/2 \rfloor}, 1 - \sum_{l \neq l_*} \beta_l - \frac{\ell - 1}{\lfloor \epsilon k/2 \rfloor}\}$. Then $\lceil N/\phi \rceil$ is the total length of the intermediate ranking. Hence, Steps 4 to 8 of ALG take time $O(N/\phi)$ altogether. In Steps 9 to 20, for each empty rank in the top 1 to top k, ALG searches for a suitable item in the rest of the intermediate ranking. This takes time $O(kN/\phi)$, and the space complexity is $O(N/\phi + \ell N/\min\{\lfloor \alpha_{\text{min}} \lfloor \epsilon k/2 \rfloor \rfloor, \lfloor \epsilon k/2 \rfloor - \sum_{l \neq l_*} \lceil \beta_l \lfloor \epsilon k/2 \rfloor \rceil\})$ to store the intermediate ranking and the counters for each block. Steps 21 to 25 again take time $O(N/\phi)$ altogether. The running time shown in Table 1 in the paper is for a naive implementation of ALG. We also note that the items within a group appear in the same order in both the true ranking as well as the ranking output by ALG. For each group we can maintain a list according to this ordering from the true ranking. When filling an empty rank, we can simply pick the best among each group's next item; whichever has the lowest rank and satsifies all the fairness constraints. The time complexity will then be $O(\ell k + N/\phi)$ with the same space complexity. Celis et al's DP algorithm runs in time $O(\ell^2 k^\ell + \ell N)$, and has space complexity $O(k^\ell)$. FA*IR runs in time $O(N + n \log n)$ time where n is the minimum number of items from each group, and has space complexity $O(N + \ell n)$.

Datasets. We experiment on two real-world datasets.

1. *German Credit Risk* dataset consists of credit risk scoring of 1000 adult German residents (Dua & Graff, 2017) along with their demographic information such as personal status, gender, age, etc. as well as financial status such as credit history, property, housing, job etc. Schufa scores of these individuals is used to get a global ranking on the dataset similar to Zehlike et al. (2017). Castillo (2019) observed that Schufa scoring is biased against young adults. Hence, we divide the dataset into protected and non-protected groups based on age. We consider two such cases (i) *age* < 25 as protected group, and (ii) *age* < 35 as protected group similar to Zehlike et al. (2017).

2. *COMPAS*[1] *recidivism* dataset consists of violent recidivism assessment of nearly 7000 criminal defendants based on a questionnaire. Angwin et al. (2016) have analysed this tool and pointed out the biases in the recidivism score against African Americans and females. We consider ranking based on the recidivism score (individual with the least negative recidivism score is ranked at the top) with (i) *gender* (=female)[2] and (ii) *race* (=African American) as protected groups similar to Zehlike et al. (2017). We use the processed subsets of German credit risk and COMPAS recidivism datasets[3]. The implementation of the algorithm proposed in this paper is also made public[4].

Fairness constraints. Representation of a group in a ranking is measured by its proportion in the ranking. FA*IR can only work with one protected and one non-protected group, and can only handle minimum representation requirements in each prefix of the top k ranks. Hence, in all the results shown in Figures 2 to 4 there is only one protected group and the algorithms are run only with lower bound constraints on representation of the protected group as follows. Let $l = 1$ and $l = 2$ correspond to protected and non-protected group respectively. Let p_l^* be the representation of the group l in the entire dataset. Sufficient representation of a group need not necessarily mean there has to be exactly p_l^* fraction of items in the top k ranks from group l. Hence, we run experiments by varying this sufficient representation requirement using a control parameter δ. FA*IR is run with $p = p_1^* + \delta$, the DP algorithm from (Celis et al., 2018) is run with the fairness constraints, $\forall k' \in [k], L_{1,k'} = \lceil (p_1^* + \delta)k' \rceil, L_{2,k'} = 0, U_{1,k'} = k', U_{2,k'} = k'$. ALG is also run with group fairness constraints $(\boldsymbol{\alpha} = (1, 1), \boldsymbol{\beta} = (p_1^* + \delta, 0), k = 100)$, and the parameter $\epsilon = 0.4$. In Figure 5, we run ALG with fairness constraints $(\boldsymbol{\alpha} = (p_1^* + \delta, p_2^* + \delta), \boldsymbol{\beta} = (p_1^* - \delta, p_2^* - \delta), k = 100)$.

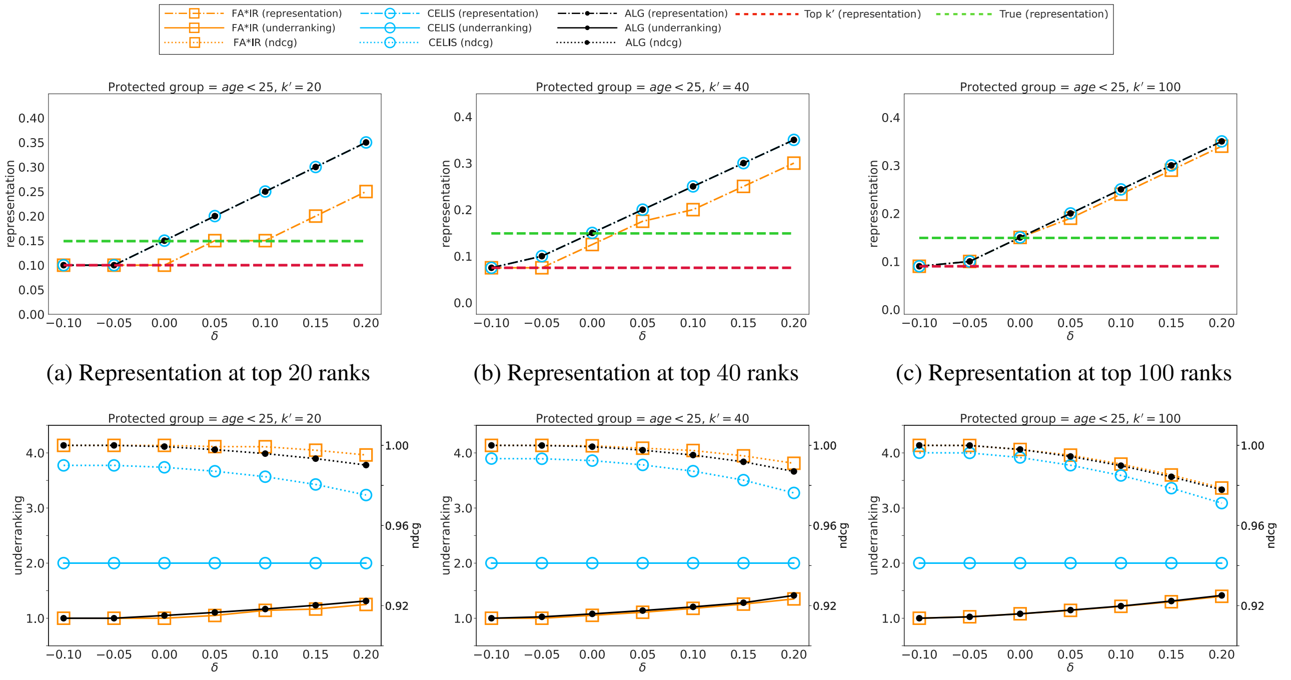
Evaluation metrics. In all the datasets, the true ranking is generated based on the decreasing order of the score (or

[1] Correctional Offender Management Profiling for Alternative Sanctions

[2] Non-binary genders were not annotated in any of the datasets used in this paper.

[3] https://github.com/DataResponsibly/FairRank/tree/master/datasets

[4] Implementation of ALG



(a) Representation at top 20 ranks

(b) Representation at top 40 ranks

(c) Representation at top 100 ranks

(d) Underranking, nDCG at top 20 ranks

(e) Underranking, nDCG at top 40 ranks

(f) Underranking, nDCG at top 100 ranks

Figure 2: Results on the German Credit Risk dataset with *age*$<$ 25 as the protected group.

relevance) of the item. For an algorithm run with $k = 100$, we evaluate its group fairness, underranking and ranking utility – normalised discounted cumulative gain (nDCG) – at top $k' = 20, 40, 100$ ranks since we are comparing with the baselines that have group fairness constraints in every prefix of the top k ranking.

1. Let G_1 represent the ranks assigned to items from the protected group. Then,

$$\text{repersentation in top } k' = \frac{|\{i \in G_1, i \leqslant k'\}|}{k'}.$$

2. For an item ranked at $j \leqslant k'$ in true ranking, let r_j be its rank in the group-fair ranking. Then,

$$\text{underranking for top } k' \text{ ranks} = \max_{j \in [k'], r_j} \frac{r_j}{j}.$$

3. Let y_i be the score of the item at rank i in true ranking and $\hat{y}_i$ be the score of the item at rank i in the group-fair ranking. Then,

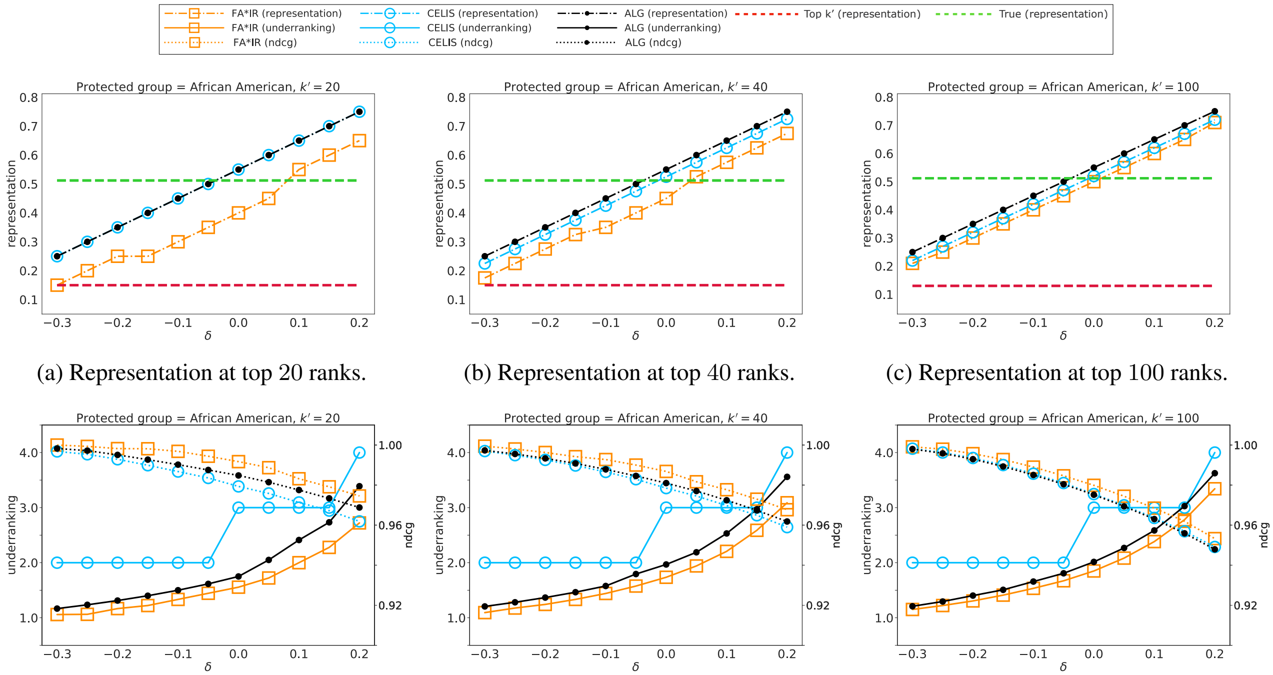
$$\text{nDCG}_{k'} = \frac{\sum_{i=1}^{k'} \frac{2^{\hat{y}_i}}{\log_2(i+1)}}{\sum_{i=1}^{k'} \frac{2^{y_i}}{\log_2(i+1)}}.$$

Despite being designed to satisfy group fairness constraints for every k consecutive ranks, ALG achieves representation in prefixes of the ranking similar to that of the baselines. In supplementary, we also show (1) evaluation of all the algorithms in consecutive ranks, (2) results for true ranking based on negative scores, for example, in the COMPAS dataset, if candidate with highest recidivism score is ranked at top 1 and so on, the protected groups are overrepresented in the top few ranks, and hence, the upper bound constraints could be used to achieve proportional representation, (3) results of experiments on the German Credit dataset with disjoint subgroups based on *age* and *gender*.

Reading the plots. For every combination of a dataset and a protected group, we show a *pair* of plots. Consider Figure 2a, 2d. Y-axis in Figure 2a shows the representation of the protected group *age* < 25 in the top 20 ranks for each run of the algorithm with fairness constraints controlled by δ on X-axis. Here, the dashed green line shows the proportion of *age* < 25 in the dataset, whereas the dashed red line shows their proportion in the top 20 ranks of the true ranking. These two lines serve as guidelines to understand the behavior of various algorithms. Figure 2d shows corresponding underranking and nDCG in the top 20 ranks.

3.1. Experimental Observations

Trade-off between underranking and group fairness. In the COMPAS dataset the female candidates are underrepresented in any of the top $k' \in [k]$ ranks (see dashed red lines in Figures 4a to 4c) compared to their true rep-



(a) Representation at top 20 ranks.

(b) Representation at top 40 ranks.

(c) Representation at top 100 ranks.

(d) Underranking, nDCG at top 20 ranks.

(e) Underranking, nDCG at top 40 ranks.

(f) Underranking, nDCG at top 100 ranks.

Figure 3: Results on the COMPAS Recidivism dataset with *African American* as the protected group.

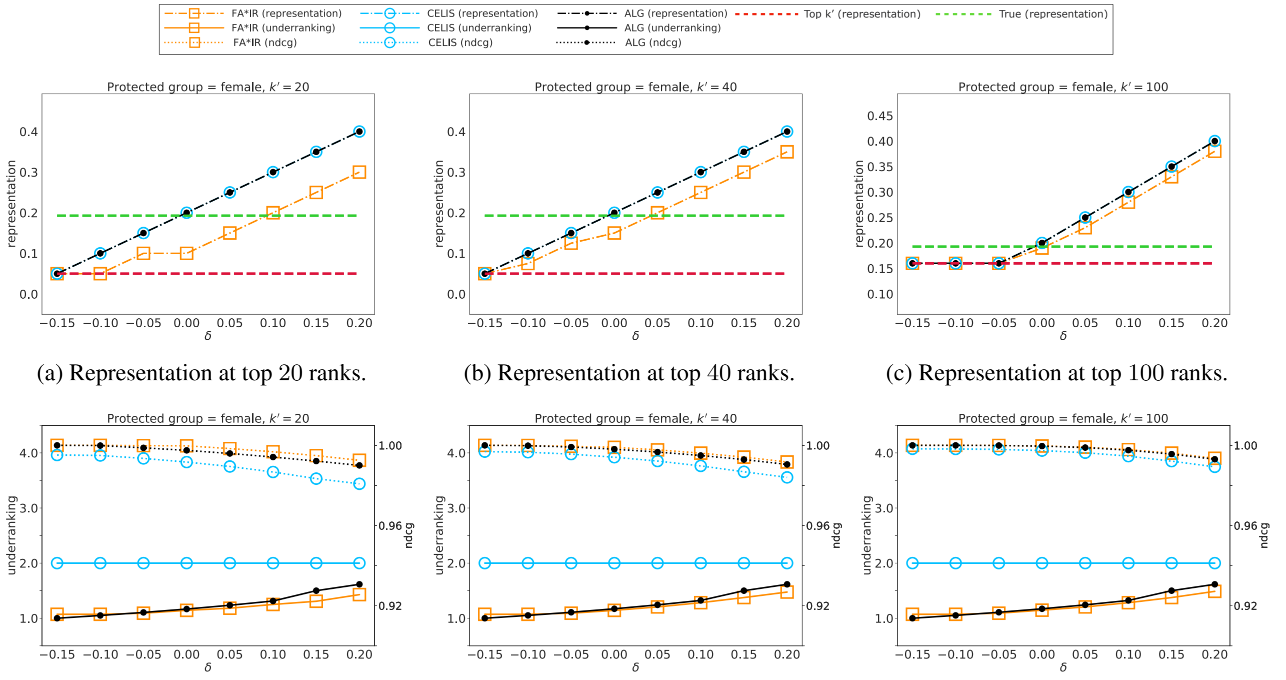
resentation in the dataset, $p^* = 0.19$ (dashed green lines). By varying δ (on X-axis), we run the experiments with the minimum female representation constraint, $p^* + \delta$. Now, comparing Figure 4a with Figure 4d for varying δ, we show the trade-off between group fairness and underranking. As the value of δ increases from -0.15 to 0.2, the underranking gets worse since more number of male candidates with lower true ranks have to be moved to higher ranks in order to accomodate for the required female repersentation in the top k ranks. Similarly, even though African Americans have very high representation, $p^* = 0.55$ (see green line in Figure 3a), in the true ranking, their representation in the top k' ranks is again significantly less (see red lines in Figures 3a to 3c). Even in this case, we observe a trade-off between group fairness and underranking in the top k' ranks. These trends are also observed at any of the top $k' = 20, 40, 100$ ranks in the German Credit dataset (see Figure 2).

We also partition the candidates in the German Credit dataset into three disjoint groups based on *age*, and enforce the constraints $\alpha_l = p_l^* + \delta$ and $\beta_l = p_l^* - \delta$ for each group with corresponding p_l^* (see Figure 5). Even in this case, the underranking gets worse with increase in the lower bound representation requirements because of the underrepresentation of the protected groups, hence confirming the trade-off even for more than two groups. These experimental results show evidence of the trade-off between group fairness and underranking in the real-world datasets.

Underranking for comparing different group-fair rankings. In previous work, only the trade-off between fairness and utility of the ranking such as nDCG has been studied. However, as established in Figure 1 and is evident from our experimental results, high nDCG does not imply anything for the underranking of a group-fair ranking. For example, in Figure 3, Celis et al.'s DP algorithm achieves almost same nDCG and group fairness as ALG but suffers badly in terms of underranking. Hence, underranking allows us to break ties when aggregate ranking utility and group fairness are same for any two group-fair rankings.

ALG achieves best trade-off between group fairness and underranking. In all the results in Figures 2 to 4, Celis et al.'s DP algorithm achieves worse underranking and same representation as ALG, and FA*IR achieves worse representation and same underranking as ALG. We note that even though in Figure 3d to 3f, FA*IR seems to achieve better underranking than ALG, it has significantly below the minimum protected group representation in the top k' ranks. Hence, we posit that ALG achieves the best trade-off between underranking and representation of the protected group in the real-world datasets.

ALG runs significantly faster than the baselines. Table 1 shows the average running time of each algorithm. The experiments were run on a Dual Intel Xeon 4110 processor consisting of 16 cores (32 threads), with a clock speed of



(a) Representation at top 20 ranks.

(b) Representation at top 40 ranks.

(c) Representation at top 100 ranks.

(d) Underranking, nDCG at top 20 ranks.

(e) Underranking, nDCG at top 40 ranks.

(f) Underranking, nDCG at top 100 ranks.

Figure 4: Results on the COMPAS Recidivism dataset with *Female* as the protected group.

Table 1: Average wall clock running time (in seconds) of five runs of the algorithms on the German Credit Risk dataset with *age* < 25 as the protected group ($n = 1000$, $p^* = 0.15$), $\ell = 2$. For these experiments we choose, $\delta = 0$. ALG is run for $((1,1),(0.15,0),k)$ group fairness, with $\epsilon = 0.4$. FA*IR is run with $p = 0.15$ and Celis et al. with $L_{age<25,k'} = \lceil 0.15 \cdot k' \rceil, \forall k' \in [k]$.

	$k = 100$	$k = 300$	$k = 500$	$k = 1000$
Celis et al.'s DP algorithm	11.0	100.0	186.0	301.0
FA*IR	3.0	3.3	3.3	3.5
ALG	**1.1**	**1.6**	**1.8**	**1.8**

2.1 GHz and DRAM of 128GB. ALG runs faster that the Celis et al.'s DP algorithm. Note that both the algorithms work for any number of groups. ALG also runs faster than the greedy algorithm, FA*IR.

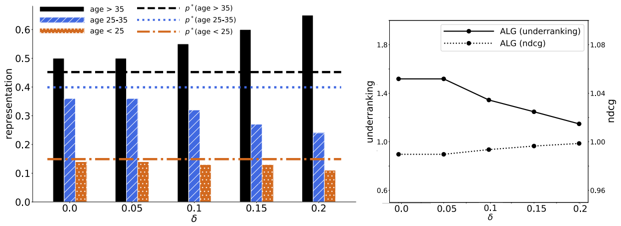


Figure 5: Results of ALG on the German Credit Risk dataset with three groups based on *age*. The bar plot (left) shows the representation achieved by ALG at top 100 ranks. The line p^* for each group shows its representation in the dataset. Corresponding underranking, nDCG shown on the right.

4. Conclusion

Previous works involving group-fair ranking are mainly focused on the trade-off between its utility and group fairness. We presented the first (to the best of our knowledge) algorithm that takes a true ranking and outputs another ranking with simultaneous group fairness and underranking guarantees. Our algorithm achieves the best of both underranking and group fairness compared to the state-of-the-art group-fair ranking algorithms. It also works in the case of more than two disjoint groups, and with different group fairness constraints for each of these groups. One limitation of our work (and other re-ranking algorithms) is that it requires the true ranking as input. All our theoretical guarantees are with respect to this true ranking; in practice, a true merit-based ranking may be debatable or unavailable due to incomplete data, unobserved features, legal and ethical considerations behind the downstream application of these rankings, etc.

Acknowledgements. AL was supported in part by SERB Award ECR/2017/003296 and a Pratiksha Trust Young Investigator Award. AL is also grateful to Microsoft Research for supporting this collaboration.

References

Adomavicius, G. and Tuzhilin, A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17:734–749, 2005.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias, 2016.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *California Law Review*, 104(3):671–732, 2016. ISSN 00081221.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2019.

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., hsin Chi, E. H., and Goodrow, C. Fairness in recommendation ranking through pairwise comparisons. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

Brin, S. and Page, L. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on World Wide Web 7*, pp. 107–117, NLD, 1998. Elsevier Science Publishers B. V.

Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., and Li, H. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pp. 129–136, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595937933. doi: 10.1145/1273496.1273513.

Castillo, C. Fairness and transparency in ranking. *Special Interest Group on Information Retrieval Forum*, 52(2): 64–71, January 2019. ISSN 0163-5840. doi: 10.1145/3308774.3308783.

Celis, L. E., Straszak, D., and Vishnoi, N. K. Ranking with fairness constraints. In *45th International Colloquium on Automata, Languages, and Programming, ICALP 2018, July 9-13, 2018, Prague, Czech Republic*, volume 107 of *LIPIcs*, pp. 28:1–28:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018. doi: 10.4230/LIPIcs.ICALP.2018.28.

Crosby, F. *Affirmative Action is Dead: Long Live Affirmative Action*. Current perspectives in psychology. Yale University Press, 2004. ISBN 9780300101294.

Dua, D. and Graff, C. UCI machine learning repository, 2017.

Geyik, S. C., Ambler, S., and Kenthapadi, K. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 2221–2231, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330691.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Järvelin, K. and Kekäläinen, J. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pp. 41–48, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132263. doi: 10.1145/345508.345545.

Kofler, C., Larson, M., and Hanjalic, A. User intent in multimedia search: A survey of the state of the art and future challenges. *ACM Comput. Surv.*, 49(2), 2016. ISSN 0360-0300. doi: 10.1145/2954930.

Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to Information Retrieval*. Cambridge University Press, USA, 2008. ISBN 0521865719.

Narasimhan, H., Cotter, A., Gupta, M., and Wang, S. Pairwise fairness for ranking and regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 5248–5255, Apr. 2020. doi: 10.1609/aaai.v34i04.5970.

Noble, S. U. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018. ISBN 9781479849949.

O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA, 2016. ISBN 0553418815.

Pariser, E. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group , The, 2011. ISBN 1594203008.

Pei, C., Zhang, Y., Zhang, Y., Sun, F., Lin, X., Sun, H., Wu, J., Jiang, P., Ge, J., Ou, W., and Pei, D. Personalized re-ranking for recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, pp. 3–11, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362436. doi: 10.1145/3298689.3347000.

Singh, A. and Joachims, T. Policy learning for fairness in ranking. In *Advances in Neural Information Processing Systems*, volume 32, pp. 5426–5436. Curran Associates, Inc., 2019.

Tavani, H. Search Engines and Ethics. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2020 edition.

Yang, K. and Stoyanovich, J. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, SSDBM '17, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450352826. doi: 10.1145/3085504.3085526.

Yang, K., Loftus, J. R., and Stoyanovich, J. Causal intersectionality for fair ranking. *ArXiv*, abs/2006.08688, 2020.

Zehlike, M. and Castillo, C. *Reducing Disparate Exposure in Ranking: A Learning To Rank Approach*, pp. 2849–2855. Association for Computing Machinery, New York, NY, USA, 2020. ISBN 9781450370233.

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., and Baeza-Yates, R. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pp. 1569–1578, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349185. doi: 10.1145/3132847.3132938.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 325–333, Atlanta, Georgia, USA, 2013. PMLR.