Abstract

Search and recommendation systems, such as search engines, recruiting tools, online marketplaces, news, and social media, output ranked lists of content, products, and sometimes, people. Credit ratings, standardized tests, risk assessments output only a score, but are also used implicitly for ranking. Bias in such 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.

In this paper, we study a trade-off between individual fairness and group fairness in ranking. We define individual fairness based on how close the predicted rank of each item is to its true rank, and prove a lower bound on the trade-off achievable for simultaneous individual and group fairness in ranking. We give a fair ranking algorithm that takes any given ranking and outputs another ranking with simultaneous individual and group fairness guarantees comparable to the lower bound we prove. Our algorithm can be used to both pre-process training data as well as post-process the output of existing ranking algorithms. Our experimental results show that our algorithm performs better than the state-of-the-art fair learning to rank and fair post-processing baselines.

1 Introduction

Search, and recommendation systems have revolutionized the way we consume an overwhelming amount of data and find relevant information quickly [BP98, AT05]. They help us find relevant documents, news, media, people, places, products and rank them based on our interests and intent [KLH16, PZZ+19]. Examples of these include rankings and recommendations in online marketplaces, recruitments, college admissions, news, and social media. Rankings can also be implicit when systems only screen or rate people, products, places as well as the social and economic exchange of goods, money, information, while a downstream application uses their scores or ratings for ranking. Credit ratings, standardized test scores, health risk assessment scores are some common examples of the above.

Information presented through ranked lists influences our worldview [Par11, Tav20]. 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 [Nob18]. 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’N16].
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 [BS16]. Protected groups could vary for specific statutes and include race, gender, age, religion, national origin, etc. Group-fairness in machine learning literature has focused on outcome-based or proportion-based definitions of fairness (e.g., demographic parity, equality of opportunity) in classification, ranking, and selection problems [HPS16, BHN19].

Another notion of fairness studied in fairness literature is individual fairness. Achieving individual fairness in classification often means similar predictions for two similar individuals or two similar data points in terms of their features or risks [DHP+12, CDPF+17]. Group-fairness is a desirable goal but arbitrary corrections to achieve group-fairness can cause further harm if they are perceived as individually unfair [Cro04]. In classification as well as ranking, if we consider individual fairness on average, then it is closely tied to the overall accuracy or relevance. However, if we remove the aggregation or averaging to focus on the parts where individual-fairness really matters, it is not always the same as accuracy or relevance. Recent work has pointed out these subtleties between group-fairness and individual-fairness in both classification and ranking [BGW18, Bin20, KRW17].

**Fairness in ranking.** Fairness in ranking has three broad requirements: sufficient presence of items belonging to different groups, consistent treatment of similar individuals or individual fairness, and proper representation to avoid representational harm to members of protected groups [Cas19a]. The first and the third requirements are about group-fairness, whereas the second requirement is about individual-fairness. For example, diversity alone in top ranks satisfies sufficient presence for the user who consumes the ranking but need not provide consistent treatment and proper representation in the way items are ranked. Fair ranking algorithms can be divided into two categories. First, re-ranking algorithms that modify a given ranking of high utility to incorporate fairness constraints while trying to preserve the utility. Second, learning-to-rank algorithms that incorporate fairness and utility objectives into learning a ranker from training data. Re-ranking can be used to post-process the prediction of any given ranker as well as pre-process the training data of any given ranker. We survey previous work on fair ranking, with the above distinction in mind.

Fair ranking can be framed as an integer optimization problem [CSV18]. Given a set of \( n \) items along with their group memberships (where a single item can belong to multiple groups), and a matrix \( W \) whose entries \( W_{ij} \) indicate the utility of assigning rank \( j \) to item \( i \), the objective is to maximize the total utility of rank assignments while satisfying the given group-fairness constraints for the top \( k \) positions. They consider group-fairness constraints as lower and upper bounds on the group-wise utilities in the top \( k \) positions and allow such constraints for all values of \( k \). For \( W \) matrices corresponding to most practical utility metrics, e.g., Discounted Cumulative Gain (DCG), a greedy assignment of the highest valued item available at each rank maximizes the total utility. If \( \Delta \) is the maximum number of groups an item belongs to, then a fair and greedy re-ranking gives \((\Delta + 2)\)-approximation to the group-fair ranking of maximum utility [CSV18].

The fair top-\( k \) selection problem gives another formulation of fair re-ranking [ZBC+17]. For a given \( k \ll n \), and list of \( n \) items with a numerical quality value assigned to each item, the objective of fair top-\( k \) selection problem is to select \( k \) items to maximize utility while ensuring a minimum proportion from a protected group in the top-\( l \) ranks, for all \( l \leq k \). The authors divide utility into two objectives, selection and ordering. Selection utility quantifies if every candidate in the top-\( k \) is more qualified than the rest, and ordering utility quantifies if every pair in the top-\( k \) is ranked according to their numerical quality values. They give an efficient algorithm called FA*IR to solve the fair top-\( k \) selection problem by re-ranking a given true or color-blind ranking, which orders all the items by their numerical quality values.

Fair ranking problem can also be defined in the learning-to-rank (LTR) setting, where a model is trained to maxi-
mize utility subject to fairness constraints. In LTR setting, the ranking is probabilistic, and the fairness guarantees are often on average. Given a query-document pair, the probability of each document being ranked at top-1 is called its *exposure*. ListNet is a neural network model trained to rank a list of documents by minimizing a loss function based on their true and predicted exposure [CQL+07]. Building upon this, DELTR [ZC20] 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 [SJ19] learns a ranking that satisfies fairness of exposure. Experimental results on the above fair LTR algorithms give better fairness and utility both when compared to post-processing algorithms on real-world datasets, e.g., Yahoo! LTR and GermanCredit data.

There is related work on defining and maximizing various group-fairness metrics overall top-$k$ prefixes of the top-$k$ ranks [YS17], for a given $k$, using an optimization algorithm to learn fair representations [ZWS+13]. There are also other measures of group-fairness in ranking based on pairwise comparisons [NCGW20, BCD+19]. 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 [GAK19]. Intersectional fairness, where the items belong to more than one group, and counterfactually fair ranking, measured for their group fairness (demographic parity at top-$k$, equal opportunity at top-$k$) and ranking utility (utility loss at top-$k$, average precision at top-$k$) are studied in [YLS20].

To the best of our knowledge, all existing fair ranking algorithms guarantee group fairness but can provide only an aggregate guarantee for individual fairness. They study trade-offs between group fairness and utility of fair rankings but do not provide guarantees for the worst-case individual fairness. In this work, we address this gap in the fair ranking literature. Our group-fairness definition ensures sufficient presence of all groups, similar to previous work, but we give a new, natural definition of individual-fairness. Our main contributions can be summarized as follows.

- We define individual-fairness based on the worst-case deviation of re-ranking from the true merit-based (or color-blind) ranking for the top-$k$ items, for any $k$. This directly captures the loss of visibility suffered by items of high merit that may get ranked lower in order to achieve high group-fairness. We prove a lower bound on the trade-off achievable between individual-fairness and group-fairness simultaneously.

- We propose a Fair Individual and Group-fair Ranking (FIGR) algorithm that takes a given merit-based (or color-blind) ranking and outputs another ranking with simultaneous individual and group-fairness guarantees, comparable to the lower bound mentioned above. Our algorithm can be used to both pre-process the training data as well as post-process the output of the existing ranking algorithms.

- We do extensive experiments to show that our algorithm performs better than the state-of-the-art fair LTR and fair post-processing baselines on standard real-world datasets such as COMPAS recidivism, German credit risk, and ChileSAT used in fair ranking literature.

## 2 Individual and Group Fair Rankings

In the rest of this paper, we say that rank $i$ is “lower” than rank $j$ if $i < j$, and we say that rank $i$ is “higher” than rank $j$ if $i > j$. 
We now formally define the notion of *group fairness*; this definition is similar to the notions studied in the literature [Cas19a, CSV18].

**Definition 2.1 (Group Fairness).** A ranking is said to satisfy \((\alpha, k)\) group fairness if any \(k\) consecutive ranks have at most \(\alpha k\) items from any group.

This notion of group fairness has the desirable property that even if a few low ranked items are removed from the ranking, the remaining ranking still satisfies the group fairness conditions. In case the given “true ranking” of the items doesn’t already satisfy group fairness conditions, the re-ranking algorithms rearrange the items in the true ranking such that the group fairness conditions are satisfied. Using the notion of *individual fairness*, 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 (Individual Fairness).** A ranking is said to satisfy \(\alpha\) individual fairness if the rank of each item is at most \(1/\alpha\) 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 strictly less than 1 individual fairness. This manifests the trade-off between the group fairness and the individual fairness in ranking. We also 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.

Closely related to individual fairness is the well studied notion of PRECISION@\(K\) of ranking [JK00, MRS08, ZC20]. For a given ranking, PRECISION@\(K\) is defined as the number of items in the top \(K\) ranks of the true ranking which also appear in the top \(K\) ranks of the given ranking. We get the following relation between individual fairness and PRECISION@\(K\).

**Corollary 2.3.** A ranking satisfying \(\alpha\) individual fairness also has PRECISION@\(K\) at least \(\lfloor \alpha K \rfloor\), \(\forall K \in \mathbb{Z}^+\).

**Proof.** Fix a ranking having \(\alpha\) individual fairness. By definition, the top \(\lfloor \alpha K \rfloor\) items in the true ranking get displaced at most to the rank \(\lfloor \alpha K \rfloor / \alpha \leq K\). Hence, at least the top \(\lfloor \alpha K \rfloor\) items in the true ranking are also in the top \(K\) ranks in an \(\alpha\) individually fair ranking. Therefore, PRECISION@\(K\) is at least \(\lfloor \alpha K \rfloor\).

Our first main result is a lower bound on the trade-off achievable for simultaneous individual and group fairness in ranking.

**Theorem 2.4.** Fix \(\alpha \in [1/2, 1] \cap \mathbb{Q}\) and \(k \in \mathbb{Z}^+\). For every \(n_0 \in \mathbb{Z}^+\), there exists an \(n\) such that \(n \geq n_0\), and there exists a true ranking of \(2n\) items grouped into two groups of \(n\) items each, such that the following holds. Any ranking satisfying \(\alpha\) individual fairness (w.r.t. the true ranking) and \((\beta, k)\) group fairness in the first \(n/\alpha\) ranks must have \(\beta \geq \alpha\).

**Proof.** Let \(\alpha = a/b\) where \(a, b \in \mathbb{Z}^+\). Set \(n\) to be any integer multiple \(ak\) such that \(n \geq n_0\). Consider a true ranking where all the \(n\) items from group 1 are placed in first \(n\) ranks followed by the \(n\) items from group 2. Now, consider any ranking of these items satisfying \(\alpha\) individual fairness and \((\beta, k)\) group fairness in the first \(n/\alpha\) ranks. Observe that by our choice of parameters, \(n/\alpha\) is an integer. By the definition of individual fairness, we get that the first \(n/\alpha\) ranks must contain all the \(n\) items from group 1. Since the ranking satisfies \((\beta, k)\) group fairness, any \(k\) consecutive ranks have at most \(\beta k\) items from group 1. This implies that for any \(c \in \mathbb{Z}^+\), any consecutive \(ck\) ranks have at most \(\beta ck\) items from group 1. By our choice of parameters, \(n/(\alpha k)\) is an integer; let
Algorithm 1: FIGR Algorithm

Input: A ranking of the $N$ items and parameters $\alpha, k$ satisfying the conditions in Theorem 2.5.

1. Set $\epsilon := \frac{2}{k} \left(1 + \frac{1}{\alpha - \frac{1}{\ell}}\right)$;
2. for $i = \lceil \frac{N}{\lfloor \alpha \lfloor ek/2 \rfloor \rfloor} \rceil$ down to 1 do
   3. for $j = 1$ to $\min\{\lfloor \alpha \lfloor ek/2 \rfloor \rfloor, N - (i - 1) \lfloor \alpha \lfloor ek/2 \rfloor \rfloor\}$ do
      4. Move item at rank $(i - 1) \lfloor \alpha \lfloor ek/2 \rfloor \rfloor + j$ to rank $(i - 1) \lfloor ek/2 \rfloor + j$.
3. end
4. end
5. for $i = 1$ to $\lceil \frac{N}{\lfloor \alpha \lfloor ek/2 \rfloor \rfloor} \rceil$ do
   6. for $j = (i - 1) \lfloor ek/2 \rfloor + 1$ to $i \lfloor ek/2 \rfloor$ do
      7. if rank $j$ is unoccupied then
         8. move the first item at a rank higher than $j$ that can be moved to rank $j$ without violating the
            $(\alpha, \lfloor ek/2 \rfloor)$ group fairness constraints among the items in ranks $(i - 1) \lfloor ek/2 \rfloor + 1$ to $i \lfloor ek/2 \rfloor$, if any such element is available.
      9. end
6. end
7. end
8. for $i = 1$ to $N$ do
9. if rank $i$ is unoccupied then
10. move to rank $i$, the first item at rank higher than $i$.
11. end
12. end
13. end
14. for $i = 1$ to $N$ do
15. Output final ranking:
16. end
17. end
18. end
19. end

$c = n/(\alpha k)$. Therefore, the first $ck = n/\alpha$ ranks contain at most $\beta ck = \beta n/\alpha$ items from group 1. If $\beta < \alpha$, then first $n/\alpha$ ranks contain strictly less than $n$ elements from group 1, which is a contradiction. Therefore, we must have $\beta \geq \alpha$.

Our next main result is a fair ranking algorithm that takes any given ranking and outputs another ranking with individual and group fairness guarantees comparable to that of Theorem 2.4.

**Theorem 2.5.** Given a true ranking of $N$ items grouped into $\ell$ disjoint groups, with each group having at least $n$ items, and fairness parameters $\alpha \in (1/\ell, 1]$ and $k \in \mathbb{Z}^+$, there exists a polynomial time algorithm to compute a ranking satisfying

1. $\left(\alpha - \frac{1}{\left[1 + \frac{1}{\alpha - \frac{1}{\ell}}\right]}\right)$ individual fairness.
2. $\left(\alpha \left(1 + \frac{2}{k}\left(1 + \frac{1}{\alpha - \frac{1}{\ell}}\right)\right), k\right)$ group fairness in the first $\lfloor n/\alpha \rfloor - \left[1 + \frac{1}{\alpha - \frac{1}{\ell}}\right]$ ranks.

For the rest of the section, let $\epsilon := \frac{2}{k} \left(1 + \frac{1}{\alpha - \frac{1}{\ell}}\right)$. Let the $i$th “block” of ranks refer to the ranks $(i - 1) \lfloor ek/2 \rfloor + 1$ to $i \lfloor ek/2 \rfloor$. 

Lemma 2.6. The ranking output by Algorithm 1 satisfies \( \left( \alpha - \frac{1}{\lfloor \epsilon k/2 \rfloor} \right) \) individual fairness.

Proof. Fix an item having true rank \( j \in [N] \). At the end of step 4 its rank is

\[
\left\lfloor \frac{j}{\alpha \lfloor \epsilon k/2 \rfloor} \right\rfloor \lfloor \epsilon k/2 \rfloor + \left( j - \left\lfloor \frac{j}{\alpha \lfloor \epsilon k/2 \rfloor} \right\rfloor \lfloor \epsilon k/2 \rfloor \right) \left( \lfloor \epsilon k/2 \rfloor - \left\lfloor \alpha \lfloor \epsilon k/2 \rfloor \right\rfloor \right) + j
\]

\[
\leq \frac{j}{\alpha \lfloor \epsilon k/2 \rfloor} \left( \lfloor \epsilon k/2 \rfloor - \left\lfloor \alpha \lfloor \epsilon k/2 \rfloor \right\rfloor \right) + j \frac{\lfloor \epsilon k/2 \rfloor}{\alpha \lfloor \epsilon k/2 \rfloor} \leq \frac{j}{\alpha \lfloor \epsilon k/2 \rfloor - 1} = \alpha - \frac{1}{\lfloor \epsilon k/2 \rfloor}.
\]

Subsequent steps do not increase the ranking of any item. \( \square \)

Lemma 2.7. At the end of step 13, no positions in the first \( \lfloor n/\alpha \rfloor - \lfloor \epsilon k/2 \rfloor \) ranks will be empty.

Proof. Consider step 10 of the algorithm. A rank \( j \) will be left unoccupied if either (i) each group already has \( \lfloor \alpha \lfloor \epsilon k/2 \rfloor \rfloor \) elements in the block containing rank \( j \), or (ii) the groups which have less than \( \lfloor \alpha \lfloor \epsilon k/2 \rfloor \rfloor \) elements in the block containing rank \( j \) do not have any elements in ranks higher than \( j \). By our choice of parameters, we have

\[
\lfloor \epsilon k/2 \rfloor = \left\lfloor 1 + \frac{1}{\alpha - \frac{1}{\ell}} \right\rfloor \geq 1
\]

and

\[
\epsilon = \frac{2}{k} \left( 1 + \frac{1}{\alpha - \frac{1}{\ell}} \right) \implies (\epsilon k/2 - 1) = \frac{1}{\alpha - \frac{1}{\ell}} \implies \lfloor \epsilon k/2 \rfloor > \frac{1}{\alpha - \frac{1}{\ell}}
\]

\[
\implies \ell (\alpha \lfloor \epsilon k/2 \rfloor - 1) > \lfloor \epsilon k/2 \rfloor \implies \ell \lfloor \alpha \lfloor \epsilon k/2 \rfloor \rfloor > \lfloor \epsilon k/2 \rfloor.
\]

Therefore, if each group has \( \lfloor \alpha \lfloor \epsilon k/2 \rfloor \rfloor \) elements in the block containing rank \( j \), then every rank in this block has to be occupied. Therefore, case (i) can not happen.

For any block \( i \in \mathbb{Z}^+ \), the first \( i \lfloor \epsilon k/2 \rfloor \) elements in any intermediate ranking (including the final ranking) contain at most \( i \lfloor \alpha \lfloor \epsilon k/2 \rfloor \rfloor \) items from each group. Since there are at least \( n \) elements from group, we have that as long as \( i \) satisfies \( i \lfloor \alpha \lfloor \epsilon k/2 \rfloor \rfloor \leq n \), there will be at least one element available from each group to move into an empty spot in the first \( i \) blocks without violating \((\alpha \lfloor \epsilon k/2 \rfloor)\) group fairness constraints for any of the first \( i \) blocks. Thus, the first \( i \) blocks will be filled at the end of step 13. Therefore, the number of ranks filled is at least

\[
i \lfloor \epsilon k/2 \rfloor = \left\lfloor \frac{n}{\lfloor \alpha \lfloor \epsilon k/2 \rfloor \rfloor} \right\rfloor \lfloor \epsilon k/2 \rfloor > \left( \frac{n}{\lfloor \alpha \lfloor \epsilon k/2 \rfloor \rfloor} - 1 \right) \lfloor \epsilon k/2 \rfloor
\]

\[
\geq \left( \frac{n}{\alpha \lfloor \epsilon k/2 \rfloor - 1} \right) \lfloor \epsilon k/2 \rfloor = n/\alpha - \lfloor \epsilon k/2 \rfloor \geq \lfloor n/\alpha \rfloor - \lfloor \epsilon k/2 \rfloor.
\]

Therefore, case (ii) will not happen for the first \( \lfloor n/\alpha \rfloor - \lfloor \epsilon k/2 \rfloor \) ranks. Thus, at the end of step 13 no positions in the first \( \lfloor n/\alpha \rfloor - \lfloor \epsilon k/2 \rfloor \) ranks will be empty. \( \square \)

Lemma 2.8. At the end of step 13 each block has at most \( \lfloor \alpha \lfloor \epsilon k/2 \rfloor \rfloor \) items from any particular group.
Proof. For any block $i$, we observe that at the end of step 4, block $i$ of size $\lfloor \frac{\epsilon k}{2} \rfloor$ has at most $\lfloor \alpha \lfloor \frac{\epsilon k}{2} \rfloor \rfloor$ non-empty positions and therefore has at most $\lfloor \alpha \lfloor \frac{\epsilon k}{2} \rfloor \rfloor$ items from any particular group. Step 10 ensures that when the algorithm terminates, each block has at most $\lfloor \alpha \lfloor \frac{\epsilon k}{2} \rfloor \rfloor$ items from any particular group.

Lemma 2.9. The ranking output by Algorithm 1 satisfies $(\alpha (1 + \epsilon), k)$ group fairness in the first $\lfloor \frac{n}{\alpha} \rfloor - \lfloor \frac{\epsilon k}{2} \rfloor$ ranks.

Proof. Lemma 2.7 shows that none of the first $\lfloor \frac{n}{\alpha} \rfloor - \lfloor \frac{\epsilon k}{2} \rfloor$ ranks will be empty at the end of step 13; therefore, these ranks will remain unchanged in the steps after step 13.

Consider any $k$ consecutive ranks $j, \ldots, j + k - 1$. Let $i_1 \overset{\text{def}}{=} \lfloor j / \lfloor \frac{\epsilon k}{2} \rfloor \rfloor$ and $i_2 \overset{\text{def}}{=} \lfloor (j + k - 1) / \lfloor \frac{\epsilon k}{2} \rfloor \rfloor$. By construction, the blocks $i_1 + 1, \ldots, i_2 - 1$ are fully contained in the ranks $\{j, j + 1, \ldots, j + k - 1\}$. For any $l \in [\ell]$, the number of items from group $l$ in ranks $j$ to $j + k - 1$ is at most the number of items from group $l$ in blocks $i_1$ to $i_2$. Using Lemma 2.8 we get that this is at most $\lfloor \alpha k \rfloor + 2 \lfloor \alpha \lfloor \frac{\epsilon k}{2} \rfloor \rfloor \leq \alpha (1 + \epsilon) k$.

We note that this bound also holds for cases when $i_2 = i_1 + 1$ or $i_2 = i_1$.

Proof of Theorem 2.5. Follows from the choice of $\epsilon$ and from Lemma 2.6, Lemma 2.7 and Lemma 2.9.

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.

Theorem 2.10. Given a true ranking of $N$ items grouped into $\ell$ disjoint groups, with each group having at least $n$ items, and fairness parameters $\alpha \in (1/\ell, 1]$ and $k \in \mathbb{Z}^+$ such that $k \geq \frac{1}{\alpha - \frac{1}{\ell}}$, there exists a polynomial time algorithm to compute a ranking satisfying

1. $(\alpha - \frac{1}{\ell})$ individual fairness.

2. $(\alpha, k)$ group fairness in each of the first $\lfloor \frac{n}{\lfloor \alpha k \rfloor} \rfloor$ blocks of size $k$.

Proof. We use Algorithm 1 with $\epsilon := 2$. Now, the $i$th “block” is of size $\lfloor \frac{\epsilon k}{2} \rfloor = k$.

Fix an item $j \in [N]$ in the true ranking. The proof of Lemma 2.6 shows that its final rank will be at most

$$\frac{j}{\alpha - \lfloor \frac{\epsilon k}{2} \rfloor} = \frac{j}{\alpha - \frac{1}{\ell}}.$$  

Here, the equality follows from our choice of $\epsilon = 2$. Hence, the ranking output by Algorithm 1 with $\epsilon = 2$ satisfies $(\alpha - \frac{1}{\ell})$ individual fairness.

Lemma 2.8 shows that at the end of step 13 each block has at most $\lfloor \alpha k \rfloor$ items from any particular group. For $k \geq \frac{1}{\alpha - \frac{1}{\ell}}$, we have,

$$k \geq \frac{1}{\alpha - \frac{1}{\ell}} \implies k (\alpha \ell - 1) \geq \ell \implies \ell (\alpha k - 1) \geq k \implies \ell \lfloor \alpha k \rfloor > k.$$
Therefore, if a block contains \( \lceil \alpha k \rceil \) items from each of the \( \ell \) groups, it can not have any empty ranks. Consequently, as long as \( i \lceil \alpha k \rceil \) items from each group are available, blocks 1 to \( i \) will not contain empty ranks. Since there are at least \( n \) items from each group, no rank in the first \( i \) blocks will be empty, where \( i \) satisfies \( i \lceil \alpha k \rceil \leq n \). Hence, for \( i \in \mathbb{Z}^+ \) such that \( i \leq \frac{n}{\lceil \alpha k \rceil} \), blocks 1 to \( i \) each of size \( k \) contain at most \( \alpha k \) items from each group. That is, the first \( \frac{n}{\lceil \alpha k \rceil} \) blocks each of size \( k \) satisfy \((\alpha, k)\) group fairness.

3 Experimental Validation

In this section, we study the trade-off between individual and group fairness achieved by FIGR on various real-world datasets. We also compare our results with the fair post-processing and fair LTR baselines.

3.1 Datasets

We consider two types of datasets in our experiments. The first type has a global ranking on the entire dataset calculated using a subset of attributes. Bias in one or more of these attributes might introduce bias in the ranking output. Hence, a fair post-processing algorithm can be used to correct these biases in the global ranking. However, a listwise LTR model, such as ListNet [CQL+07], requires that each training sample consists of a list of items on which a ranking is available. Hence, the experiments based on LTR model are performed on the second type of datasets. These datasets have a query-document format and support training of a listwise LTR model.

**German credit risk dataset.** This is a dataset of credit risk scoring of adult German residents [DG17]. It consists of a set of 1000 candidates from various demographics applying for a loan. Features of the candidate include 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 [ZBC+17] and [YS17]. [Cas19b] 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) \( \text{age} < 25 \) as protected group, and (ii) \( \text{age} < 35 \) as protected group similar to [ZBC+17].

**COMPAS recidivism dataset.** This dataset consists of violent recidivism assessment of nearly 7000 criminal defendants by the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) tool based on answers to a questionnaire consisting of 137 questions. This dataset is curated by [ALMK16] to analyze the biases in the tool. Their study points out racial as well as gender biases in the predictions of COMPAS. In our experiments, we consider ranking based on the recidivism score (individual with highest ~recidivism is ranked at top-1). The protected groups are (i) \( \text{gender} (=\text{female}) \) and (ii) \( \text{race} (=\text{African American}) \) similar to [ZBC+17].

**ChileSAT dataset.** This dataset consists of a Chilean University’s admission test scores of the students admitted, their highschool grades, and a score based on their academic performance after one year at the university. Given these test scores and their highschool grades, an LTR model has to predict in advance, a ranking over these students about their performance at the end of the first academic year. This dataset has a query-document format since each academic year can be considered as a query and students in that academic year are the list of candidates to be ranked. There are around 450 to 600 student records in each of the 5 academic years in the dataset. In the experiments, we perform a 5-fold cross validation with four academic years for training and one for validating the

---

1Non-binary genders were not annotated in any of the datasets used in this paper.
LTR models. We consider two protected groups in the dataset, gender (=female) and high-school type (=public) to study the behaviour of FIGR when two different types of bias exist. These biases are also studied by [ZC20].

In our experiments, we use the processed subsets of ChileSAT (Engineering Students\textsuperscript{2}, German credit risk and COMPAS recidivism datasets\textsuperscript{3}).

3.2 Experiments

We compare our results with baselines, such as (i) ListNet [CQL+07], a listwise LTR model that ranks a list of documents based on their relevance scores with respect to the query, (ii) DELTR [ZC20], an in-processing LTR model that is trained with ListNet’s objective along with the group fairness constraint – fairness of exposure, and (iii) FA*IR [ZBC+17], a post-processing algorithm that re-ranks a given ranking to maintain significant proportions of the protected group in every prefix of the ranking. For brevity, we omit detailing the intricacies of these baseline algorithms. Following is a brief explanation of our experimental setup\textsuperscript{4}.

1. **color-blind LTR.** ListNet trained in a ‘colorblind’ fashion – no access to protected (or sensitive) attributes.
2. **color-aware LTR.** ListNet trained on all the attributes.
3. **DELTR.** Trained with $\gamma = 100K$. (refer [ZC20] for more details).
4. **FA*IR pre $p$.** FA*IR with parameter $p$ used to pre-process the training data. ListNet is trained on this data.
5. **FA*IR post $p$.** ListNet’s predictions on test data post-processed using FA*IR with parameter $p$.
6. **FA*IR $p$.** FA*IR with parameter $p$ used to re-rank the global ranking in the COMPAS recidivism and German credit risk datasets.
7. **FIGR pre $p$.** FIGR with parameter $\alpha = \max\{1 - p, \frac{1}{\ell} + \delta\}$ used to pre-process the training data. We add a small number $\delta$ to $1/\ell$ since FIGR requires that $\alpha$ is strictly greater than $1/\ell$. ListNet is trained on this data.
8. **FIGR post $p$.** ListNet’s predictions on test data post-processed using FIGR with parameter $\alpha = \max\{1 - p, \frac{1}{\ell} + \delta\}$.
9. **FIGR $p$.** FIGR with parameter $\alpha = \max\{1 - p, \frac{1}{\ell} + \delta\}$ used to re-rank the global ranking in the COMPAS recidivism and German credit risk datasets.

**Training details.** For FA*IR, the parameter $p$ is chosen from $\{p^+, p^-, p^*\}$, where $p^*$ is the proportion of the protected group, $p^+ = p^* + 0.1$ and $p^- = p^* - 0.1$. This parameter setting is adopted from [ZC20]. In all the experiments, the number of groups are 2 (protected and non-protected). Hence, we use $\ell = 2$ and $\delta = 0.01$ in FIGR. For the values of $p$, $\ell$ and $\delta$, the parameter $\alpha$ in FIGR will be $\alpha = \max\{1 - p, 0.51\}$. For all the datasets in the experiments, we set the value of $k$ to 100. This is a reasonable choice since the number of ranked items is at least 500 in all the datasets. For the LTR models – ListNet and DELTR – we use the same parameter and hyper-parameter settings as [ZC20] and report average results across 5-folds of the ChileSAT dataset. Further, to account for variability in learning the parameters for LTR, we run all the LTR based experiments 5 times and report the final average\textsuperscript{5}.

\textsuperscript{2}https://github.com/MilkaLichtblau/DELTR-Experiments/tree/master/data/EngineeringStudents
\textsuperscript{3}https://github.com/DataResponsibly/FairRank/tree/master/datasets
\textsuperscript{4}Code for FIGR and baselines is available at [https://github.com/sruthigorantla/FIGR](https://github.com/sruthigorantla/FIGR)
\textsuperscript{5}Our runs showed minor variations in the learned LTR parameters with almost same ranking predictions.
For every combination of a dataset and a protected group, we show a pair of plots to understand better the trade-off between group and individual fairness in the real-world datasets. For example, Figure 1(a) and Figure 1(b) show group and individual fairness respectively on the German credit risk dataset with age < 25 as protected group. In Figure 1(a), X-axis represents the rank $k$ and Y-axis shows the proportion of protected group items in the top-$k$ ranks. In this plot, the line $y = p^*$ shows the proportion of the protected group in the entire (training) data, whereas the line ‘True’ represents the proportion of the protected group in the top-$k$ ranks of the true ranking. These two lines serve as guidelines to understand the behavior of various algorithms and pick the best performing algorithm. In Figure 1(b), X-axis represents blocks of size 100 – $i$th block consists of items $100(i−1) + 1$ to $100i$ of the true ranking. For any item $j$ in the block, $\alpha_j = \frac{j}{\text{pred}(j)}$, where $\text{pred}(j)$ is the rank of item $j$ in the output ranking. Y-axis shows the minimum $\alpha_j$ among the items in the block. We call this $\min \alpha$. Higher the value of $\min \alpha$, higher the individual fairness, according to the Definition 2.2. Since we measure individual fairness with respect to the true ranking, the line ‘True’ has 1-individual fairness. We say that there is a trade-off between group and individual fairness when the algorithms placing higher proportions of protected groups in the top-$k$ ranks consistently suffer from lesser individual fairness in these ranks. In the following section we study the trade-offs achieved by FIGR and the baselines on the real-world datasets.

Figure 1: Trade-offs between individual and group fairness in the German credit risk dataset. In all the experiments with FIGR $p$, the parameter settings are $\ell = 2$, $\delta = 0.01$ and $k = 100$. Then, $\alpha = \max \{1 − p, 0.51\}$. 

Reading the plots.
3.3 FIGR vs. Fair Post-Processing Methods on German Credit Risk and COMPAS Datasets

Figures 1(a)-1(d) show the experimental results on the German credit risk dataset. The protected group age < 25 (younger adults) is significantly underrepresented in the dataset with \( p^* = 0.15 \) (see Figure 1(a)). Moreover, their ‘True’ proportions given by the Schufa score based ranking in the top-400 ranks is even less. This indicates bias against younger adults in the ranking. All three variants of FIGR allocate almost 1.5-2 times the true proportion and increase the representation of younger adults in the top-400 ranks. FIGR also achieves approximately 0.70-individual fairness, which is a reasonable trade-off for group fairness in this case. FA*IR on the other hand fails to correct biases with \( p^+ \) and \( p^- \). Even with \( p^+ \), it fails to achieve significant improvement in the representation of the younger adults.

In case of the protected group age < 35 (young adults), although their representation in the entire dataset is almost
half \((p^* = 0.55)\), they are all ranked in the tail of the true ranking. This shows a strong bias against young adults (see Figure [1c]). FA*IR \(p^-\) stays too close to the ‘True’ proportions. On the other hand, with slightly larger values of \(p\), FA*IR \(p^+\) and \(p^-\) clearly overcompensate for the lack of representation of the protected group at the cost of very low individual fairness (0.50-0.65). This also reduces the proportion of the non-protected group to much lower than it’s true proportion, leading to inversion of bias. FIGR, by design, avoids this problem. For values of \(p\) higher than \(1/\ell\), \(\alpha\) is set to \(\frac{1}{2} + \delta\), where \(\delta\) is a small value (\(\delta = 0.01\) in our experiments). Hence, the representation of both protected and non-protected groups is bounded (see Theorem 2.5). This is also evident in our results. With \(p^+\) and \(p^-\), FIGR achieves same results because \(\alpha\) in both the cases is set to 0.51. FIGR with each of these parameter settings finds the best trade-off between group fairness (45% in the top 100 positions) and individual fairness (0.75 in the top 100 positions).

Figures [2(e)-2(h)] show experimental results on the COMPAS recidivism dataset. Females in the dataset \((p^* = 0.19)\) face bias in the top-200 ranks due to the biases introduced by the COMPAS tool (see Figure [2e]). Once again, FA*IR with \(p^+\) and \(p^-\) stays close to the ‘True’ proportions. With \(p^+\), it improves the female representation in the top-400 ranks. FIGR \(p^+\) also shows same trends as FA*IR \(p^+\). Both FIGR \(p^+\) and \(p^-\) trade off individual fairness for group fairness. Nevertheless, this doesn’t cause an inversion of bias in the output ranking.

The protected group \(race = African\ American\) is substantially underrepresented in the dataset in the top-500 positions even though their representation in the entire data is high, \(p^* = 0.51\) (see Figure [2g]). This shows bias against the protected group in the true ranking based on the recidivism score. Since this case of bias is similar to the bias towards young adults in the German credit risk dataset, we see similar results. FIGR improves their representation in the top-500 ranks while achieving reasonable individual fairness. FA*IR \(p^-\) is a close competitor to FIGR \(p^-\) for the best choice of algorithm. However, FA*IR \(p^+\) and \(p^-\) once again overcompensate for the lack of representation of African-Americans in the top-1000 positions since they actually are the majority group in the data.
Figure 4: Trade-offs between individual and group fairness in the ChileSAT dataset with public school as the protected group. Comparison of post-processing, in-processing and standard LTR methods. In all the experiments with FIGR $p$, the parameter settings are $\ell = 2$, $\delta = 0.01$ and $k = 100$. Then, $\alpha = \max \{1 - p, 0.51\}$. See Table 1 in Appendix A for tabulated results.

### 3.4 FIGR vs. Fair LTR Methods on the ChileSAT Dataset

Figure 3 and Figure 5 show the group and individual fairness trade-offs achieved by the LTR models, DELTR and pre-processing training data with FIGR and FA*IR. Whereas Figure 4 and Figure 6 show these results for same LTR and DELTR models compared with post-processing methods applied on the ListNet predictions.

Representation of the students from public school in the top-300 ranks is higher than their total representation ($p^* = 0.34$) (see Figure 3(a)). This means that the students from public schools having same test scores as private school students indeed have caliber to perform better at the university. Hence, it is evident that the test scores are biased against the students from the public school. Such biases are naturally adjusted by a color-aware LTR, whereas a color-blind LTR re-inforces these biases (see Figure 5(a)). DELTR does not further adjust the ranks. Interestingly, pre-processing with all three variants of FIGR and FA*IR with $p^+$ rank higher proportions of the public school students in the top-200 ranks and at the same time achieve high individual fairness compared to the LTR models (see Figure 3(b)). Although color-aware LTR corrects biases, its impact is limited. DELTR, on the other hand, limits itself from achieving higher group fairness as well as individual fairness since fairness of exposure is already satisfied. In the case of post-processing the ListNet predictions, only FIGR with $p^*$ achieves similar results (see Figure 4(a)), and still has as much individual fairness as others.

Female students are substantially under-represented in the dataset ($p^* = 0.20$). The color-blind LTR ranks almost the same proportion of the females in all top-$k$ ranks as that of ‘True’ proportions (see Figure 5(a)). This means that the academic performance of females both before and after the first year at the university is bad. In this case, a color-aware LTR will further learn to discriminate against the females because in this case, placing females in the top few ranks would hurt the utility. As expected, the color-aware LTR stays below the color-blind LTR. DELTR acts on the difference in the exposure of the groups. Hence, it places more number of females in the top-200 ranks purely to achieve fairness of exposure. But, DELTR is oblivious to its impact on the individual fairness. Although
Figure 5: Trade-offs between individual and group fairness in the ChileSAT dataset with female as the protected group. Comparison of pre-processing, in-processing and standard LTR methods. In all the experiments with FIGR, the parameter settings are $\ell = 2, \delta = 0.01$ and $k = 100$. Then, $\alpha = \max \{1 - p, 0.51\}$. See Table 2 in Appendix A for tabulated results.

pre-processing with FIGR achieves less individual fairness than DELTR, post-processing with FIGR places the highest number of females in the top-200 while achieving as much individual fairness as DELTR (see Figure 6).

In both the cases, FIGR with most of the parameter settings achieves the best group fairness as well as individual fairness. Even in cases where it loses some individual fairness compared to the baselines, the trade-off is minimal. In contrast, the fair LTR and fair post-processing baselines do not achieve these trade-offs.

4 Conclusion

Fair ranking is crucial to search and recommendations, and has been a matter of global concern in the quest towards responsible AI. We studied group and individual fairness notions in ranking. We defined individual fairness based on how close the predicted rank of each item is to its true rank, and proved a lower bound on the trade-off achievable for simultaneous individual and group fairness in ranking. While other works (e.g., [CSV18], etc.) have studied “aggregate forms” of individual fairness, to the best of our knowledge, our work is the first to give provable guarantees on the worst-case displacement of an item in the output ranking with respect to its true rank. We presented the first (to the best of our knowledge) algorithm that takes any given ranking and outputs another ranking with simultaneous individual and group fairness guarantees comparable to the lower bound we proved. Our algorithm performed better than the state-of-the-art fair learning to rank and fair post-processing baselines.

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.
Figure 6: Trade-offs between individual and group fairness in the ChileSAT dataset with female as the protected group. Comparison of post-processing, in-processing and standard LTR methods. In all the experiments with FIGR \( p \), the parameter settings are \( \ell = 2, \delta = 0.01 \) and \( k = 100 \). Then, \( \alpha = \max \{1 - p, 0.51\} \). See Table 2 in Appendix A for tabulated results.

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

[ALMK16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias, 2016.

[AT05] G. Adomavicius and A. Tuzhilin. 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(6):734–749, 2005.

[BCD+19] Alex Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Zhe Zhao, L. Hong, Ed Huaihsein Chi, and Cristos Goodrow. Fairness in recommendation ranking through pairwise comparisons. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019.

[BGW18] Asia J. Bienga, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR ’18, page 405414, New York, NY, USA, 2018. Association for Computing Machinery.

[BHN19] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and Machine Learning. fairmlbook.org, 2019. http://www.fairmlbook.org.
[Bin20] Reuben Binns. On the apparent conflict between individual and group fairness. In FAT* ‘20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020, pages 514–524. ACM, 2020.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In Proceedings of the Seventh International Conference on World Wide Web 7, WWW7, page 107117, NLD, 1998. Elsevier Science Publishers B. V.

[BS16] Solon Barocas and Andrew D. Selbst. Big data’s disparate impact. California Law Review, 104(3):671–732, 2016.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In Proceedings of the Seventh International Conference on World Wide Web 7, WWW7, page 107117, NLD, 1998. Elsevier Science Publishers B. V.

[BS16] Solon Barocas and Andrew D. Selbst. Big data’s disparate impact. California Law Review, 104(3):671–732, 2016.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In Proceedings of the Seventh International Conference on World Wide Web 7, WWW7, page 107117, NLD, 1998. Elsevier Science Publishers B. V.

[Cas19a] Carlos Castillo. Fairness and transparency in ranking. SIGIR Forum, 52(2):6471, January 2019.

[Cas19b] Carlos Castillo. Fairness and transparency in ranking. volume 52, page 6471, New York, NY, USA, January 2019. Association for Computing Machinery.

[CDPF+17] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’17, page 797806, New York, NY, USA, 2017. Association for Computing Machinery.

[CQL+07] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In Proceedings of the 24th International Conference on Machine Learning, ICML 07, page 129136, New York, NY, USA, 2007. Association for Computing Machinery.

[Cro04] F.J. Crosby. Affirmative Action is Dead: Long Live Affirmative Action. Current perspectives in psychology. Yale University Press, 2004.

[CSV18] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. Ranking with fairness constraints. In Ioannis Chatzigiannakis, Christos Kaklamanis, Dániel Marx, and Donald Sannella, editors, 45th International Colloquium on Automata, Languages, and Programming, ICALP 2018, July 9-13, 2018, Prague, Czech Republic, volume 107 of LIPIcs, pages 28:1–28:15. Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik, 2018.

[DG17] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[DHP+12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS '12, page 214226, New York, NY, USA, 2012. Association for Computing Machinery.

[GAK19] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD ’19, page 22212231, New York, NY, USA, 2019. Association for Computing Machinery.

[HPS16] M. Hardt, E. Price, and Nathan Srebro. Equality of opportunity in supervised learning. In NIPS, 2016.

[JK00] Kalervo Järvelin and Jaana Kekäläinen. 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, page 4148, New York, NY, USA, 2000. Association for Computing Machinery.
A Appendix for Experimental Results
Table 1: Individual and group fairness results for the ChileSAT dataset with **public school** as the protected group. The plots for these results are as shown in Figure 3 and Figure 4. The column *Grp.* shows the proportion of the protected group items in the top-$k$ ranks and the column *Ind.* shows min $\alpha$ in the ranks $k - 99$ to $k$. Note that $k - 99$ to $k$ is the $\frac{k}{100}$th block of size 100 shown in the plots.

| Method \(\downarrow\) | Grp. Ind. \(k = 100\) | Grp. Ind. \(k = 200\) | Grp. Ind. \(k = 300\) | Grp. Ind. \(k = 400\) | Grp. Ind. \(k = 500\) |
|----------------------|------------------|------------------|------------------|------------------|------------------|
| True                 | 0.42 1           | 0.38 1           | 0.36 1           | 0.35 1           | 0.31 1           |
| color-blind LTR      | 0.32 0.17        | 0.32 0.46        | 0.32 0.63        | 0.33 **0.88**    | 0.3 1            |
| color-aware LTR      | 0.38 0.18        | 0.38 0.4         | 0.37 0.72        | 0.36 0.87        | 0.31 1           |
| DELTR                | 0.41 0.19        | 0.41 0.42        | 0.39 **0.76**    | 0.36 0.77        | 0.31 1           |
| FA*IR pre \(p^+\)   | 0.39 0.18        | 0.39 0.4         | 0.38 0.64        | 0.36 0.87        | 0.31 1           |
| FA*IR pre \(p^-\)   | 0.45 0.2         | **0.54**         | 0.43 0.65        | 0.38 0.82        | 0.32 1           |
| FIGR pre \(p^+\)    | 0.39 0.18        | 0.39 0.4         | 0.38 0.61        | 0.36 0.87        | 0.31 1           |
| FIGR pre \(p^-\)    | 0.45 0.2         | **0.54**         | 0.43 0.63        | 0.38 0.82        | 0.32 1           |
| FA*IR post \(p^+\)  | 0.43 0.19        | 0.44 **0.54**    | 0.41 0.62        | 0.37 0.82        | 0.32 1           |
| FA*IR post \(p^-\)  | 0.45 0.29        | **0.52**         | 0.43 0.65        | 0.39 0.82        | 0.32 1           |
| FIGR post \(p^+\)   | 0.43 0.19        | 0.44 **0.54**    | 0.41 0.62        | 0.37 0.82        | 0.32 1           |
| FIGR post \(p^-\)   | 0.45 0.29        | **0.52**         | 0.43 0.65        | 0.39 0.82        | 0.32 1           |
| Method (↓) | Grp. Ind. | Grp. Ind. | Grp. Ind. | Grp. Ind. | Grp. Ind. | Grp. Ind. |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| True      | 0.1 1     | 0.14 1    | 0.17 1    | 0.19 1    | 0.17 1    | 0.17 1    |
| color-blind LTR | 0.13 0.17 | 0.14 0.46 | 0.17 0.63 | 0.18 0.88 | 0.17 1    | 0.17 1    |
| color-aware LTR | 0.1 0.17  | 0.12 0.46 | 0.15 0.59 | 0.16 0.86 | 0.16 1    | 0.16 1    |
| DELTR     | 0.18 0.16 | 0.22 0.49 | 0.22 0.6  | 0.21 0.85 | 0.18 1    | 0.18 1    |
| FA*IR pre $p^*$ | 0.12 0.16 | 0.14 0.47 | 0.17 0.72 | 0.18 0.9  | 0.16 1    | 0.16 1    |
| FA*IR pre $p^+$ | 0.23 0.14 | 0.27 0.39 | 0.26 0.71 | 0.23 0.96 | 0.19 1    | 0.19 1    |
| FA*IR pre $p^-$ | 0.11 0.17 | 0.12 0.45 | 0.15 0.62 | 0.17 0.88 | 0.16 1    | 0.16 1    |
| FIGR pre $p^*$ | 0.27 0.14 | 0.3 0.38  | 0.27 0.6  | 0.23 0.95 | 0.19 1    | 0.19 1    |
| FIGR pre $p^+$ | 0.35 0.13 | 0.35 0.35 | 0.3 0.63  | 0.24 0.92 | 0.19 1    | 0.19 1    |
| FIGR pre $p^-$ | 0.27 0.14 | 0.31 0.36 | 0.29 0.65 | 0.23 0.91 | 0.19 1    | 0.19 1    |
| FA*IR post $p^*$ | 0.15 0.16 | 0.17 0.42 | 0.17 0.75 | 0.18 0.92 | 0.16 1    | 0.16 1    |
| FA*IR post $p^+$ | 0.25 0.15 | 0.26 0.45 | 0.27 0.6  | 0.24 0.88 | 0.2 1     | 0.2 1     |
| FA*IR post $p^-$ | 0.11 0.17 | 0.12 0.46 | 0.15 0.6  | 0.16 0.86 | 0.16 1    | 0.16 1    |
| FIGR post $p^*$ | 0.3 0.13   | 0.3 0.46  | 0.27 0.6  | 0.24 0.87 | 0.2 1     | 0.2 1     |
| FIGR post $p^+$ | 0.35 0.13  | 0.35 0.44 | 0.3 0.6   | 0.24 0.89 | 0.2 1     | 0.2 1     |
| FIGR post $p^-$ | 0.31 0.13  | 0.31 0.52 | 0.28 0.6  | 0.22 0.85 | 0.18 1    | 0.18 1    |

Table 2: Individual and group fairness results for the ChileSAT dataset with female as the protected group. The plots for these results are as shown in Figure 5 and Figure 6. The column $Grp.$ shows the proportion of the protected group items in the top-$k$ ranks and the column $Ind.$ shows $\min \alpha$ in the ranks $k - 99$ to $k$. Note that $k - 99$ to $k$ is the $\frac{k}{100}$th block of size 100 shown in the plots.