A Study of Pre-processing Fairness Intervention Methods for Ranking People

Clara Rus(✉), Andrew Yates, and Maarten de Rijke

University of Amsterdam, Amsterdam, The Netherlands
{c.a.rus,a.c.yates,m.derijke}@uva.nl

Abstract. Fairness interventions are hard to use in practice when ranking people due to legal constraints that limit access to sensitive information. Pre-processing fairness interventions, however, can be used in practice to create more fair training data that encourage the model to generate fair predictions without having access to sensitive information during inference. Little is known about the performance of pre-processing fairness interventions in a recruitment setting. To simulate a real scenario, we train a ranking model on pre-processed representations, while access to sensitive information is limited during inference. We evaluate pre-processing fairness intervention methods in terms of individual fairness and group fairness. On two real-world datasets, the pre-processing methods are found to improve the diversity of rankings with respect to gender, while individual fairness is not affected. Moreover, we discuss advantages and disadvantages of using pre-processing fairness interventions in practice for ranking people.

Keywords: Fairness · Ranking · Recruitment

1 Introduction

A ranking system's goal is to create an ordered list of items having relevant items at top positions given the search terms and the user's preferences. Applications of ranking systems include not only ranking of items such as digital artefacts like documents, books, or movies, or real-world entities like hotels, but also *ranking of people*. Increasingly, recruiters rely on automatic tools to process the large amount of applications received for a vacancy [10].

Fairness in Ranking. Ensuring fairness in a ranking system is especially important when the items to be ranked are real persons, whose lives could be impacted by the system's decisions. When a ranking system is designed for recruitment, its outcomes influence who receives more interviews, the job's quality, and even the wage [2]. Systems for ranking people in a recruitment setting encode stereotypes and biases that already exist in recruitment [7,11,20,25], leading to actions that discriminate against minority groups [3,13,24].

There are several fairness interventions that can be used to improve the fairness of a ranking [32,33]. However, due to legal constraints, their use in a practical application of ranking people can be limited. Briefly, most fairness interventions require

N. Goharian et al. (Eds.): ECIR 2024, LNCS 14611, pp. 336–350, 2024.
https://doi.org/10.1007/978-3-031-56066-8_26

access to the sensitive information of the candidates, but the European General Data Protection Regulation (GDPR) [1] states that the processing of a special category of sensitive information is prohibited, with exceptions for gender and age [4]. While many fairness intervention methods require information about these sensitive categories, *pre-processing intervention methods* do not suffer from this limitation. Pre-processing fairness intervention methods aim at debiasing the data representation of the candidates, encouraging the (downstream) ranking model to generate fair predictions without having access to sensitive information at inference time. Similarly, in-processing methods can be trained offline and used during inference time without access to sensitive information. However, they require a definition of fairness to optimize for, which in practice might be hard to define as the European law does not give exact guidelines about the minimum requirements as opposed to the US ("80% rule").

Comparing Pre-processing Fairness Intervention Methods. Pre-processing fairness intervention methods can be applied offline on training data containing sensitive information, which was acquired in compliance with the GDPR [4]. In practice, a model trained on fair data representation can be applied on the real candidate data without having access to the sensitive information of the candidates. However, there is little knowledge about the performance of such methods with respect to group fairness, individual fairness, utility, and how the methods compare to each other when ranking people in a recruitment setting. We address this knowledge gap. We consider a scenario where access to sensitive attributes during inference time is limited, and compare three pre-processing fairness intervention methods: (i) CIF-Rank [27] aims at achieving group fairness, (ii) LFR [34] aims at achieving both group and individual fairness, while (iii) iFair [19] aims at achieving individual fairness. Neither LFR nor iFair require access to sensitive information during inference time, however, CIF-Rank does.

Fairness intervention methods are typically evaluated w.r.t. both group fairness and individual fairness, while considering the utility of the ranking. Here, *group fairness* means that members of different protected groups should be treated the same. *Individual fairness* means that similar individuals should be treated similarly, based on a defined similarity metric. In recruitment, the primary focus is on group fairness, but due to limited access to sensitive information, it may be more feasible to focus on individual fairness. Indeed, striving for individual fairness could contribute to group fairness. This works under the assumption that the measure of similarity between candidates is computed free of bias, which in practice might not be the case, resulting in segregation by membership in the sensitive groups.

Aim and Summary. Our aim in this paper is to investigate (i) how well pre-processing interventions generalize on two datasets for ranking people given an occupation, (ii) what the trade-off is between group and individual fairness obtained by the pre-processing interventions, and (iii) which method is more suitable to be used in practice. Our main finding is that the pre-processing methods do obtain an increase in diversity in most occupations on both datasets, without affecting individual fairness. Finally, we discuss the trade-offs using pre-processing interventions for ranking people in practice. CIF-Rank offers transparency and in-group fairness with minimal changes to candidate data, but it requires access to sensitive information during inference time. In contrast, LFR and iFair treat fairness as an optimization problem, thus, there is no need for access to sensitive information and the changes in diversity are more noticeable, though they are less explainable and iFair does not guarantee group fairness.

2 Related Work

Fairness interventions can be applied to prevent ranking systems from exacerbating discriminatory behavior towards minority groups. Fairness interventions can be categorized as *pre-processing*, *in-processing*, and *post-processing* methods [32].

Pre-processing methods [19,27,34] aim to debias the data used to represent the candidates and then either re-rank the candidates based on the new representations or use the data to train a model. CIF-Rank [27] aims to achieve group fairness by creating counterfactual representations of the candidates. LFR [34] aims at achieving both group and individual fairness by creating representations that obfuscate information about the protected groups, while also ensuring a good encoding of useful information. iFair [19] aims at achieving individual fairness by creating representations that encourage similar outcomes for similar individuals regardless of the sensitive attributes.

In-processing methods [5,15,30] aim to optimize the ranking system for both fairness and utility. For example, Zehlike and Castillo [30] propose to optimize the model for both utility of the ranking and fairness of exposure by ensuring that the groups have the same probability of being placed at the top position of the ranking.

Post-processing methods [6,24,26,29,31] adjust the recommended ranking based on some minimum and maximum constraints regarding the desired proportion of each sensitive group in the top positions. For example, FA*IR [29] creates separate ranked lists for each protected group and merges them such that a minimum number of candidates from the disadvantaged group is present in the top of the list. The work of [6] focuses on achieving individual fairness over time by re-ordering the items such that the unfairness is minimized over time in an online setting. The unfairness is measured as the cumulative difference between the attention an item receives and the relevance.

Previous work [27,34] experimented with pre-processing methods on various datasets, including recruitment datasets in some cases [19,23]. In this paper, we go beyond prior work by performing a systematic comparison of the above mentioned pre-processing methods in a ranking setting, showing how they generalize on two recruitment datasets, while considering that the bias direction may vary between occupations, w.r.t. group fairness, individual fairness, utility of the ranking, and their use in practice. Moreover, we adapt the work of [34] for a ranking setting.

3 Fairness Interventions

The use of fairness interventions for ranking people is constrained by the GDPR [1], which prohibits processing sensitive information that many fairness interventions require. We argue that pre-processing interventions can be used in practice as they can be applied offline on the training data containing sensitive information that has been acquired in compliance with the GDPR (e.g., a synthetic anonymous dataset reflecting a real distribution) [4]. Then, a ranking model is trained on the fair data, encouraging the model to generate a fair ranking without having access to sensitive information of

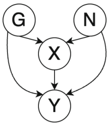

Fig. 1. Causal model describing the data with sensitive attributes gender (G) and nationality (N), non-sensitive attributes (X), and utility scores (Y).

the real candidates at inference time. As pre-processing methods do not guarantee the diversity of the ranking, we compare them with FA*IR, a post-processing method that does guarantee the desired proportion of disadvantaged candidates in the top positions. Below, we describe the fairness interventions we compare in this work.

CIF-Rank [27] estimates "How would this candidate data look like if they were part of another protected group?" As input it uses a reference protected group towards which we want to transform the candidates in a counterfactual world, and a causal model describing the data and the effects of the sensitive attributes on the data. A causal graph is a directed acyclic graph (DAG) where nodes represent variables, and directed edges between nodes represent causal relationships. A directed edge from node A to node B indicates that variable A causally influences variable B. Figure 1 shows a possible causal model that can be used to represent the data, having the following nodes: sensitive attributes, G (gender) and N (nationality), non-sensitive attributes of the candidates (X), and the utility score used to rank the candidates for a given occupation (Y), with edges from the features to the scores, and from the sensitive attributes to the features and the scores of the candidates. It estimates the total effect composed of the direct effect and the indirect effect. The indirect effect is the effect mediated by the non-sensitive attributes, which are called mediators.

The causal effects are estimated using the mma R package [28], which performs mediation analysis with multiple mediators. After computing the causal effects of the sensitive attributes on the data, which represent the bias encoded in the data, one can compute the counterfactual representations. These are computed by changing the observed representations by boosting them up or down according to the difference in the total effect of the actual group and the control group, the group to which we want to transform the data. This method satisfies transparency requirements as by using the causal estimates to generate the counterfactual representations we have a way to explain how changes in the representations were produced and why.

LFR [34] formulates fairness as an optimization problem of finding a good representation of the data such that the sensitive information encoded in the data is obfuscated, while also encoding useful information. This method can achieve both group and individual fairness. The authors formulate the new representation in terms of a probabilistic mapping to a set of prototypes (points in the input space). Thus, the model can be defined as a discriminative clustering model, where the prototypes act as the clusters as each representation is assigned to a prototype. In order to ensure that the information regarding the membership in a protected group is lost, the authors make use of statistical parity by ensuring that a random candidate from group A should map to a particular prototype with an equal probability as a random candidate from group B to the same prototype: $L_z = |\sum_k^K M_k^A - M_k^B|$, where M_k is the probability that X maps to prototype v_k. In order to keep the useful information for each candidate, the representations should be close to the original ones: $L_x = (X - X')^2$ with X being the original representation and $X' = \sum_k^K M_k v_k$ the new representations; while also ensuring that the representations are still predictive of the label (Y): $L_y = -Y \log(Y') - (1 - Y)\log(1 - Y')$, where Y is the label and $Y' = \sum_k^K M_k w_k$ is the prediction. As LFR was designed for a binary classification task, we adapt it to a ranking scenario by considering Y to be a binary label indicating the positive and negative samples used for training the ranking model. In the end, the model optimizes the compound loss $L = A_z L_z + A_x L_x + A_y L_y$ by learning two sets of parameters: the prototype locations v_k and the parameters w_k that govern the mapping from the prototypes to classification decisions Y.

iFair [19] aims at achieving individual fairness, meaning that candidates who are similar in all task-relevant attributes such as job qualification, and disregarding all potentially discriminating attributes such as gender, should have similar outcomes. iFair's main idea is constructed on top of LFR [34]. However, in contrast to LFR, iFair directly optimizes for individual fairness, implicitly obtaining group fairness by obfuscating the information regarding the membership of protected groups. To directly optimize for individual fairness the authors propose the following loss: $L_z = \sum_{i,j}^M (d(X_i', X_j') - d(X_i^*, X_j^*))^2$, where X' is the new representation, X^* the original representation excluding the sensitive attributes, and $d(X_i, X_j) = [\sum_n^N (\alpha_n(x_i, n - x_j, n)^p]^{1/p}$ is a distance function. M is the number of candidates in the data, N is the number of features that each candidate has and α_n is a learnable weight for different data attributes. Ideally, the learnable weights for the sensitive attributes should be near zero. Intuitively this ensures that similar individuals in the original input space, excluding the sensitive attributes, should have similar low-rank representations, regardless of the membership in the protected groups. Thus, the distance between individuals should be preserved in the transformed space. This directly optimizes for individual fairness, while also obfuscating the membership in protected groups. To ensure the utility of the representations, the representations should still be close to the original ones: $L_x = (X - X')^2$, with X containing both protected and non-protected features. The new representations are computed in the same manner as LFR proposes: $X' = \sum_k^K M_k v_k$. In the end, the model optimizes the compound loss $L = A_z L_z + A_x L_x$ by learning two sets of parameters: the prototype locations v_k and the parameters α_n that govern the mapping from the prototypes to classification decisions Y.

FA*IR [29] is a post-processing intervention that aims to ensure that the proportion of candidates from the disadvantaged group remains statistically above or indistinguishable from a given minimum, at every top-k. To ensure the utility of the ranking, candidates included in the top-k should be more qualified than every candidate not included, and for every pair of candidates in the top-k, the more qualified candidate should be ranked above. It pre-computes a table having fairly represent the protected group with minimal proportion p and significance α at various top-k. Next, the algorithm creates a queue for the privileged group ($P0$) and a queue for the disadvantaged group ($P1$). If the table demands a disadvantaged candidate at the current position, the algorithm extracts the best candidate from queue $P1$ and adds it to the final ranking, otherwise it extracts the best candidate from $P0 \cup P1$.

4 Experimental Setup

We aim to answer three research questions: **(RQ1)** How well do the pre-processing fairness interventions listed in Sect. 3 generalize on datasets for ranking people? **(RQ2)** What is the trade-off between group and individual fairness obtained by the pre-processing interventions? **(RQ3)** Which preprocessing intervention method is more suitable to be used in practice?

Datasets. The *BIOS dataset* [14] consists of real biographies collected from the web by filtering for lines starting with a name followed by the string "is a(n) (xxx) title," where "title" is an occupation from the BLS SOC system.[1] In total there are 28 occupations. Each candidate is represented by non-sensitive features extracted from the text biography (term frequency of the occupation in the biography, length and number of words of the biography) and sensitive features (gender, provided by the dataset). To simulate bias in the ranking, candidates are ranked for each occupation by the cosine similarity between word2vec [21] embeddings of the occupation title and the text biography, as word2vec embeddings are known to create stereotypical associations [8, 18].

The *XING dataset* was collected to study gender biases in the returned ranked search results given each user's profile details [29]. The dataset consists of anonymized real user profiles collected from XING[2] in response to 57 queries representing job titles. For each query the first 40 user profiles were selected. For each user profile the following non-sensitive information was gathered: duration of job experiences, duration of education and profile popularity, and sensitive information, gender (provided by the dataset, inferred from name and profile picture when available). The score to rank the candidates is computed as provided by the original implementation,[3] based on educational features, job experience and hits on profile.

Parameters and Settings. We model a scenario where experienced candidates apply for jobs in the same field. Train-test splits (30% BIOS and 40% XING for test) are stratified across sensitive groups, ensuring five consistent splits per occupation. For XING occupations fully over-represented by one gender are excluded, as no improvements in diversity can be measured, resulting in 44 occupations. Sensitive attributes are not used as training features.

As a ranker, we use the Ranklib [12] implementation of RankNet [9], a pairwise learning to rank algorithm as it can better capture the changes in ranks produced by the pre-processing methods [23]. Relevance judgments are based on candidate scores, with negative/positive samples determined using a threshold (0.4 BIOS and 0 XING).

We apply pre-processing methods separately for each occupation to account for varying bias directions [23]. We experimented with $A_x, A_y, A_z \in \{0.1, 0.01, 1, 5, 10\}$ with $K = 10$, selecting $A_x = 0.1$, for LFR $A_y = 1$, $A_z = 1$, while for iFair $A_z = 5$, as it obtained the best trade-off between group/individual fairness and utility.

The rankings are evaluated in terms of group fairness, individual fairness, and utility. *Group fairness* is evaluated as the percentage of each sensitive group among the top 10.

[1] https://www.bls.gov/soc/.

[2] https://www.xing.com/.

[3] https://github.com/MilkaLichtblau/xing_dataset/.

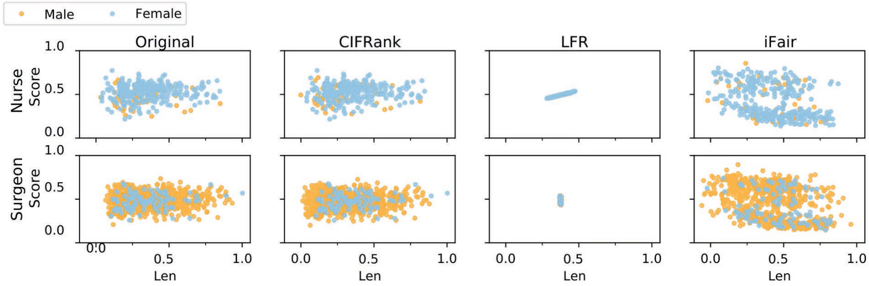

Fig. 2. Example of data points for the BIOS dataset.

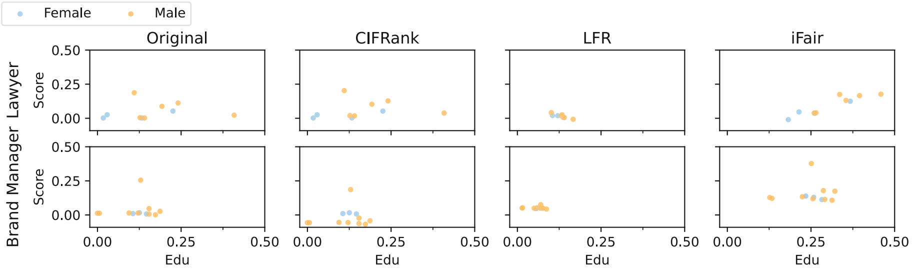

Fig. 3. Example of data points for the XING dataset.

Our aim is to create a diverse ranked list of the candidates with respect to the sensitive groups by increasing the proportion of the underrepresented groups, without producing a swap between the underrepresented and over-represented group. *Individual fairness* is evaluated by doing a pairwise comparison between candidates distance in the features space and their achieved exposure [16]: $d(E_i, E_j) \leq d(X_i, X_j)$. The difference in the candidates exposure (E) should be as close as possible to the distance between the candidates in the feature space. The closer the value to 1 the better. *Utility* of the ranking is measured using normalized discounted cumulative gain across the top 10 (NDCG@10). We focus on the top 10 since it is unlikely that recruiters will scroll down to view more candidates [17,22]. To support reproducibility of our study we provide the code to run the experiments on GitHub.[4]

5 Results and Discussion

In this section we describe our results obtained on the BIOS and XING dataset when training a ranking model on the fairer data generated by the pre-processing interventions. Reported results are an average over the five runs. The test set is pre-processed in experiments involving LFR and iFair as they do not require access to sensitive information. In contrast, for CIF-Rank, which requires access to sensitive information during

[4] https://github.com/ClaraRus/A-Study-of-Pre-processing-Fairness-Intervention-Methods-for-Ranking-People.

inference, the test set is not subjected to pre-processing. The pre-processed test set has the same candidates, but their representations differs.

Table 1. Evaluation of the ranking in the top 10 in terms of utility (NDCG@10), group fairness (%protected@10 gender), and individual fairness (yNN).

Dataset	Method	NDCG@10	%protected(G)@10	yNN
XING	Original	0.57	30	0.85
	CIF-Rank	0.79	32	0.86
	LFR	0.70	32	0.85
	iFair	0.97	30	0.85
	FA*IR ($p = 0.5$)	0.98	38	0.85
	FA*IR ($p = 0.9$)	0.93	38	0.85
BIOS	Original	0.86	20	0.72
	CIF-Rank	0.93	22	0.72
	LFR	0.77	26	0.72
	iFair	0.52	27	0.72
	FA*IR ($p = 0.5$)	0.98	33	0.72
	FA*IR ($p = 0.9$)	0.90	77	0.72

5.1 Examining the Effect of Interventions on Individual Data Points

Before addressing our research questions, we examine the effect of interventions on individual data points to better understand the methods we consider.

Figure 2 and 3 show how the data points representing the candidates change when applying the pre-processing methods on the BIOS and XING dataset for two example occupations. The data points are plotted in a space defined by the score of the candidate and one feature. Similar patterns can be observed between the two datasets. Looking at the data points generated by CIF-Rank, one can observe that the representations are slightly changed by the interventions. For Surgeon (Fig. 2) and Brand Manager (Fig. 3), which are male dominated occupations, there is a slight decrease for the males with respect to score, however, this does not create a positive change in the diversity of the generated ranking, as the change in rank is small, thus, the model is not strongly penalized. For Nurse (Fig. 2), a female-dominated occupation, the score of the males is slightly increased, and we also observe an increase in the diversity of the generated ranking. For Lawyer (Fig. 3), the score of the over-represented group (males) is increased, due to a difference in bias direction between the test and train set. For LFR and iFair, the changes in the representations of the candidates are more noticeable. Specifically, LFR tends to produce representations that cluster near similar values in the space. It seems that for the BIOS dataset the candidates are clustered near the threshold between positive and negative candidates. In contrast, iFair tends to scatter the candidates across the space, creating a clear distance between positive and negative candidates. For Brand Manager (Fig. 3) each female candidate is close in space to a

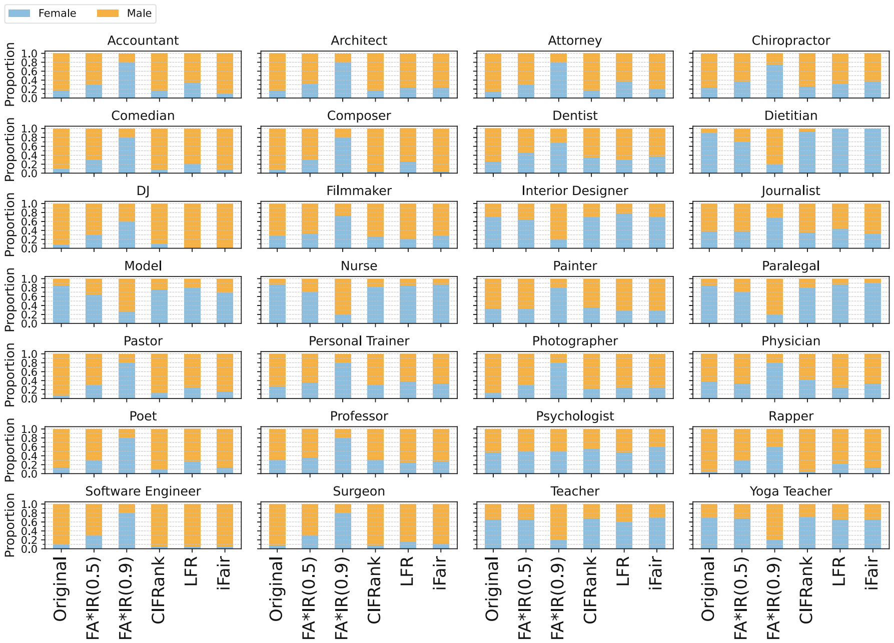

Fig. 4. Distribution of the gender groups in the top 10 for the BIOS dataset.

male candidate, resembling the idea that best candidates from the female group should be treated similarly with the best candidates from the male group.

5.2 Group Fairness

Next, we turn to group fairness and the first of our research questions.

BIOS Dataset. Fig. 4 and Table 1 (BIOS) show the proportion of the gender groups in the top 10 when applying pre-processing methods on the BIOS dataset. All methods increase the representation of underrepresented groups in some of the gender-biased occupations. These include occupations like attorney, dentist, pastor and photographer, which are male-dominated occupations, and female-dominated occupations such as model, nurse, and paralegal.

LFR obtains improvements for the underrepresented group in 20 (71%) occupations with an average increase in proportion of 6%, followed by iFair in 17 (60%) occupations with an average increase of 7%, and CIF-Rank in 13 (46%) occupations with an average increase in proportion of 2%. CIF-Rank negatively affects the diversity of the ranking in 8 (28%) occupations, while LFR and iFair do so in 7 (25%) occupations. Using FA*IR, the proportion of the underrepresented group is increased in all occupations. However, when using ($p = 0.9$), it can be observed that the previously underrepresented group is now over-represented.

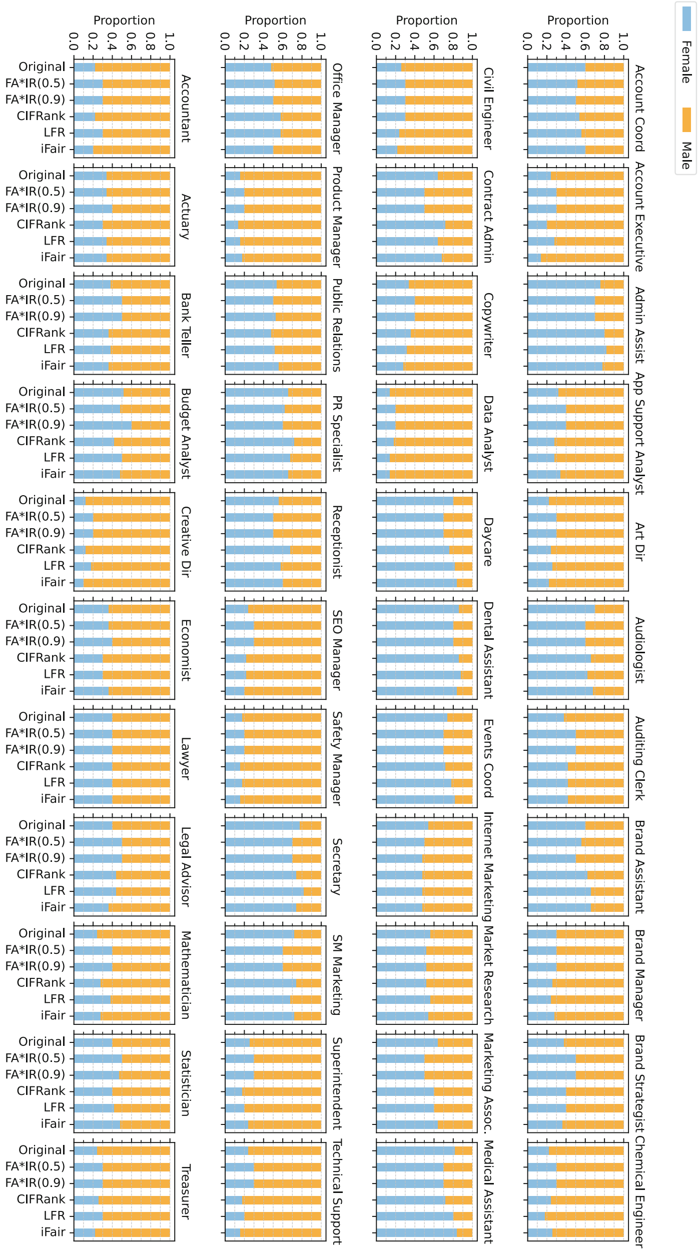

Fig. 5. Distribution of the gender groups in the top 10 for the XING dataset.

XING Dataset. Fig. 5 shows the proportion of the gender groups in the top 10, when applying the pre-processing interventions on the XING dataset. All three methods improved the proportion of the underrepresented group for male-dominated occupations such as auditing clerk, brand manager, office manager and mathematician, and for female-dominated occupations such as internet marketing coordinator and audiologist.

For some occupations the increase in proportion is observed for the group that is already over-represented. CIF-Rank and LFR increase the proportion of the over-represented groups in 17 (38%) occupations, while iFair in 20 (45%) occupations. Overall, CIF-Rank and LFR obtain the most improvements for the underrepresented groups in 22 (50%), respectively 19 (43%) occupations with an average increase in proportion of 2%, while iFair in 14 (31%) occupations with no increase as it produces more negative changes than positive changes.

RQ1: How Well do the Pre-processing Interventions Generalize on Two Datasets for Ranking People? It seems that for both datasets all methods achieve improvements in the diversity of the ranking in most occupations. However, in some occupations the already over-represented group is increased in proportion. We expected CIF-Rank and LFR to create a more diverse ranking, as their main goal is to create representations that aim to achieve group fairness. However, it is surprising that iFair achieves improvements, especially on the BIOS dataset, where the increase in proportion for the under-represented group is higher than of CIF-Rank. The goal of iFair is to create representations such that similar candidates receive similar outcomes regardless of the sensitive attributes. This can encourage group fairness as it could encourage treating candidates from the female group the same as similar candidates from the male group. This works under the assumption that the measure of similarity between candidates is computed free of bias. However, this might not be the case, resulting in representations that are still segregated by the membership of the candidates in the sensitive groups. Intuitively, this means that males should be treated similarly, but different from females. This can be observed in the occupations where the already over-represented group is increased in proportion. However, as CIF-Rank and LFR aim to achieve group fairness it is surprising that, although they improved the diversity in some occupations, they still harmed the diversity in other occupations. For CIF-Rank, especially on the XING dataset, we observe that the bias direction estimated on the train set is different from the test set in some of the occupations where a negative change was present, indicating that CIF-Rank is not robust towards changes in the bias direction. For all methods, given some occupations, we observe that when re-ranking the candidates given the new score, there is an improvement in diversity; however, when training and testing on the fairer data, there is no improvement in diversity. This could be because the change in rank was small, thus, the model is not penalized if it misses this change, but the ranking's diversity is affected.

In answer to RQ1, the pre-processing methods generalize well on the two datasets. We observed similar patterns on both datasets, where the diversity of the ranking increased in most occupations, with some exceptions where there was an increase in the over-represented group.

5.3 Individual Fairness and Utility

RQ2: What is the Trade-off between Group and Individual Fairness? Individual fairness and group fairness might not involve making trade-offs if the measure of similarity between individuals is computed independent of the sensitive attributes. However, this is hard to achieve in practice resulting in potential trade-offs between the two fairness measures. Table 1 shows how the methods affect the individual fairness of the ranking. Surprisingly, in this setting, neither LFR and iFair show improvements in individual fairness, however, the individual fairness is already reasonably high. CIF-Rank is expected to obtain similar individual fairness to the Original setting, as the representations produced by CIF-Rank are slightly changed w.r.t. the original ones. To measure individual fairness, one needs to define a way to measure the similarity between candidates; this implies defining a similarity metric and the features to use in the similarity metric. Following [19,34], we use the original features of the candidates in the distance function without considering the sensitive information. However, the original features might be a proxy to sensitive information, implying that to satisfy individual fairness it might be that females and males should not be treated the same.

Transforming the representation of the candidates to make them more fair carries the risk of harming the utility of the ranking (compared to the original utility). Considering the NDCG scores in Table 1, we see that, on the BIOS dataset, when pre-processing the data using LFR and iFair, the utility of the ranking decreases, while for CIF-Rank the utility of the ranking is slightly increased. As RankNet is trained and tested on pre-processed data for LFR and iFair, the generated rankings are noticeably different from the original one. For CIF-Rank the changes in the ranking are small, and the ranking model is tested on the original representations, which produces fewer changes. On the XING dataset, the ranking's utility is increased when training on pre-processed data.

In answer to RQ2, when using pre-processing fairness interventions, the individual fairness of the ranking is not affected, thus, we do not observe a trade-off between group and individual fairness.

5.4 Fairness Interventions in Practice

RQ3: Which method is more suitable to be used in practice? CIF-Rank satisfies transparency requirements, as explained in Sect. 3. Additionally, it guarantees in-group fairness by applying uniform changes to all candidates within a group. In contrast, LFR and iFair treat fairness as an optimization problem, making it less straightforward to explain any changes in the candidate representation. As opposed to CIF-Rank, LFR produces more noticeable changes to candidate data, leading to a more significant impact on the diversity of the rankings. In contrast with CIF-Rank, both LFR and iFair offer the advantage of being applicable at inference time without requiring access to sensitive attributes. Even though iFair can increase the diversity of the ranking, it may also exacerbate group segregation as we have seen in Sect. 5.2.

In practice, it is important to evaluate and model fairness with respect to intersectional groups, as candidates belonging to multiple disadvantaged groups are more likely to be discriminated [26]. CIF-Rank offers an intersectional framework by modelling the causal effects of the intersectional sensitive attributes, which was shown to increase the proportion of intersectional groups [23,27]. iFair supports multiple non-binary groups, however, there is no work analyzing its impact on intersectional groups. LFR has been designed for binary groups, thus needing an adaptation for intersectional settings.

In answer to RQ3, pre-processing fairness intervention methods can be used in the practice of ranking people, e.g., for recruitment. Depending on transparency requirements and on fairness goals, preference may be given to CIF-Rank or LFR, as they both aim at achieving group fairness with the former creating less noticeable changes in the representations, but making it easier to explain the changes.

6 Conclusion

In the context of ranking people, e.g., for recruitment, most fairness intervention methods require access to the sensitive information of people, information that should not be processed according to the GDPR, making it hard to use such interventions in practice. In this work we argue that *pre-processing* fairness interventions can be used in practice without having access to sensitive information during inference time. We have compared the performance of three pre-processing methods for the task of ranking people in a recruitment scenario on two real-world datasets with respect to group fairness, individual fairness, and utility of the ranking. Finally, we discuss the advantages and disadvantages of using the pre-processing methods in practice.

We find that for most occupations the methods achieved an improvement in the diversity of the ranking with respect to the protected groups, while the individual fairness was not affected. However, for some occupations the already over-represented group is increased in proportion. Consequently, the methods should be used carefully in order to avoid making a positive change in some occupations while other occupations are negatively affected. Advantages of using CIF-Rank include transparency, guaranteed in-group fairness and an intersectional framework, however it requires access to sensitive information during inference time and the changes to candidate data are minimal. In contrast, LFR and iFair treat fairness as an optimization problem, thus, there is no need for access to sensitive information and the changes in diversity are more noticeable, however they are less explainable and iFair does not guarantee group fairness.

While the BIOS and the XING dataset are real-world datasets, the features used to rank the candidates might not fully reflect a real-world scenario, as recruiters might take different attributes into account. Especially, the BIOS dataset, with extracted textual features and including artistic occupations, might not be mainstream in a hiring process. In addition, our study focused only on gender as a binary variable, as this was the sensitive information available in the datasets. Future work should compare the performance of fairness intervention methods w.r.t. other sensitive groups with non-binary values and towards intersectional groups.

Acknowledgments. We thank our reviewers for valuable feedback. This research was supported by the FINDHR (Fairness and Intersectional Non-Discrimination in Human Recommendation) project that received funding from the European Union's Horizon Europe research and innovation program under grant agreement No 101070212, the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, and project LESSEN with project number NWA.1389.20.183 of the research program NWA ORC 2020/21, which is (partly) financed by the Dutch Research Council (NWO).

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

1. General data protection regulation (GDPR). https://eur-lex.europa.eu/eli/reg/2016/679/oj (2023), (Accessed 20 Oct 2023)
2. Albert, E.T.: AI in talent acquisition: a review of AI-applications used in recruitment and selection. Strateg. HR Rev. **18**(5), 215–221 (2019)
3. Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., Rieke, A.: Discrimination through optimization: how Facebook's Ad delivery can lead to biased outcomes. Proc. ACM Hum. Comput. Interact. **3**(CSCW), 1–30 (2019)
4. van Bekkum, M., Borgesius Zuiderveen, F.: Using sensitive data to prevent discrimination by artificial intelligence: does the GDPR need a new exception? Comput. Law Sec. Rev. **48**, 105770 (2023)
5. Beutel, A., et al.: Fairness in recommendation ranking through pairwise comparisons. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2212–2220 (2019)
6. Biega, A.J., Gummadi, K.P., Weikum, G.: Equity of attention: amortizing individual fairness in rankings. In: The 41st international ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 405–414 (2018)
7. Bisschop, P., ter Weel, B., Zwetsloot, J.: Ethnic employment gaps of graduates in the Netherlands. De Economist **168**(4), 577–598 (2020)
8. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Advances in Neural Information Processing Systems 29 (2016)
9. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to rank using gradient descent. In: Proceedings of the 22nd International Conference on Machine Learning, pp. 89–96 (2005)
10. Chapman, D.S., Webster, J.: The use of technologies in the recruiting, screening, and selection processes for job candidates. Int. J. Sel. Assess. **11**(2–3), 113–120 (2003)
11. Ciminelli, G., Schwellnus, C., Stadler, B.: Sticky floors or glass ceilings? The role of human capital, working time flexibility and discrimination in the gender wage gap. Tech. Rep. 1668, OECD Publishing (2021)
12. Dang, V., Ascent, C.: The Lemur project (2023). http://lemurproject.org
13. Dastin, J.: Amazon scraps secret AI recruiting tool that showed bias against women. In: Ethics of Data and Analytics, pp. 296–299. Auerbach Publications (2022)
14. De-Arteaga, M., et al.: Bias in bios: a case study of semantic representation bias in a high-stakes setting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 120–128 (2019)
15. Diaz, F., Mitra, B., Ekstrand, M.D., Biega, A.J., Carterette, B.: Evaluating stochastic rankings with expected exposure. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 275–284 (2020)
16. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226 (2012)
17. Granka, L.A., Joachims, T., Gay, G.: Eye-tracking analysis of user behavior in WWW search. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 478–479 (2004)
18. Kurita, K., Vyas, N., Pareek, A., Black, A.W., Tsvetkov, Y.: Measuring bias in contextualized word representations. arXiv preprint arXiv:1906.07337 (2019)

19. Lahoti, P., Gummadi, K.P., Weikum, G.: iFair: learning individually fair data representations for algorithmic decision making. In: 2019 IEEE 35th International Conference on Data Engineering (ICDE), pp. 1334–1345, IEEE (2019)
20. Matteazzi, E., Pailhé, A., Solaz, A.: Part-time employment, the gender wage gap and the role of wage-setting institutions: evidence from 11 European countries. Eur. J. Ind. Relat. **24**(3), 221–241 (2018)
21. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of Workshop at ICLR 2013 (Jan 2013). https://doi.org/10.48550/arXiv.1301.3781
22. O'Brien, M., Keane, M.T.: Modeling result-list searching in the world wide web: the role of relevance topologies and trust bias. In: Proceedings of the 28th Annual Conference of the Cognitive Science Society, vol. 28, pp. 1881–1886 (2006)
23. Rus, C., de Rijke, M., Yates, A.: Counterfactual representations for intersectional fair ranking in recruitment. In: RecSys in HR 2023: The 3rd Workshop on Recommender Systems for Human Resources, in conjunction with the 17th ACM Conference on Recommender Systems (2023)
24. Singh, A., Joachims, T.: Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2219–2228 (2018)
25. Thijssen, L., Lancee, B., Veit, S., Yemane, R.: Discrimination against Turkish minorities in Germany and the Netherlands: field experimental evidence on the effect of diagnostic information on labour market outcomes. J. Ethn. Migr. Stud. **47**(6), 1222–1239 (2021)
26. Yang, K., Gkatzelis, V., Stoyanovich, J.: Balanced ranking with diversity constraints. arXiv preprint arXiv:1906.01747 (2019)
27. Yang, K., Loftus, J.R., Stoyanovich, J.: Causal intersectionality for fair ranking. arXiv preprint arXiv:2006.08688 (2020)
28. Yu, Q., Li, B.: mma: An R package for mediation analysis with multiple mediators. J. Open Res. Softw. **5**(1) (2017)
29. Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., Baeza-Yates, R.: FA*IR: A fair top-k ranking algorithm. In: Proceedings of the 2017 ACM Conference on Information and Knowledge Management, pp. 1569–1578. ACM (2017)
30. Zehlike, M., Castillo, C.: Reducing disparate exposure in ranking: a learning to rank approach. In: Proceedings of the Web Conference 2020, pp. 2849–2855 (2020)
31. Zehlike, M., Sühr, T., Baeza-Yates, R., Bonchi, F., Castillo, C., Hajian, S.: Fair top-k ranking with multiple protected groups. Inform. Process. Manag. **59**(1), 102707 (2022)
32. Zehlike, M., Yang, K., Stoyanovich, J.: Fairness in ranking, Part I: score-based ranking. ACM Comput. Surv. **55**(6), 1–36 (2022)
33. Zehlike, M., Yang, K., Stoyanovich, J.: Fairness in ranking, Part II: learning-to-rank and recommender systems. ACM Comput. Surv. **55**(6), 1–41 (2022)
34. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning, pp. 325–333. PMLR (2013)