The Impact of Group Membership Bias on the Quality and Fairness of Exposure in Ranking

Ali Vardasbi
Spotify
Amsterdam, The Netherlands
aliv@spotify.com

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

Fernando Diaz
Carnegie Mellon University
Pittsburgh, The USA
diazf@acm.org

Mostafa Dehghani
Google DeepMind
San Francisco, The USA
dehghani@google.com

ABSTRACT

When learning to rank from user interactions, search and recommender systems must address biases in user behavior to provide a high-quality ranking. One type of bias that has recently been studied in the ranking literature is when sensitive attributes, such as gender, have an impact on a user's judgment about an item's utility. For example, in a search for an expertise area, some users may be biased towards clicking on male candidates over female candidates. We call this type of bias *group membership bias*.

Increasingly, we seek rankings that are fair to individuals and sensitive groups. Merit-based fairness measures rely on the estimated utility of the items. With group membership bias, the utility of the sensitive groups is underestimated, hence, without correcting for this bias, a supposedly fair ranking is not truly fair. In this paper, first, we analyze the impact of group membership bias on ranking quality as well as merit-based fairness metrics and show that group membership bias can hurt both ranking and fairness. Then, we provide a correction method for group bias that is based on the assumption that the utility score of items in different groups comes from the same distribution. This assumption has two potential issues of sparsity and equality-instead-of-equity; we use an amortized approach to address these. We show that our correction method can consistently compensate for the negative impact of group membership bias on ranking quality and fairness metrics.

CCS CONCEPTS

• **Information systems** → **Learning to rank**.

KEYWORDS

Ranking fairness, Unbiased learning to rank

ACM Reference Format:
Ali Vardasbi, Maarten de Rijke, Fernando Diaz, and Mostafa Dehghani. 2024. The Impact of Group Membership Bias on the Quality and Fairness of Exposure in Ranking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3626772.3657752

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657752

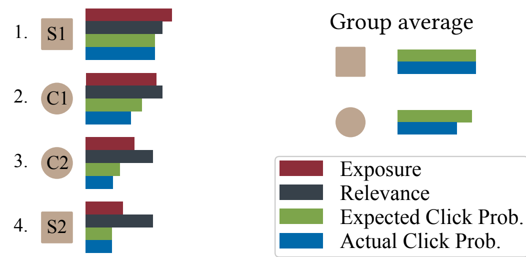

Figure 1: The effect of group bias on user clicks.

1 INTRODUCTION

Online search and recommender systems leverage user interaction data to enhance their ranking quality. When using human interactions, however, we need to account for human bias and the possibility of learning unfair ranking policies. In the context of learning to rank (LTR), the term *bias* usually refers to unequal treatment of items with equal utility by users [20]. Studies show that if the bias is ignored, this leads to a degradation in the ranking quality of a system trained on the user interactions [1, 21, 51, 54]. Correcting for bias is a necessary step for high-quality rankings, but it is not sufficient. A system should return rankings that strive, to a certain extent, for *fairness of exposure*. There are different definitions for fairness of exposure in ranking, leading to different metrics [12, 13, 19, 43, 47, 55]. However, the core idea is the same: items with similar levels of utility should receive similar exposure by the system. Without meeting fairness of exposure, bias towards the privileged groups or individuals is reinforced, both in what the system learns from the ongoing interactions [16, 18, 46], and in users' judgments about the utility of items [25, 49].

Group bias. A search or recommender system can only ensure that items with similar utility receive comparable exposure to users, by arranging them accordingly. However, this alone is insufficient. Users' judgments about the utility of items are affected by their perception of the item's group membership [25, 28, 34, 49]. This means that, even when the exposure of two high-utility items from two different groups is the same, users may judge them differently and one group may receive more clicks than the other. Fig. 1 provides a toy example for this phenomenon. Assume a job application system where there are four applicants coming from two groups, shown in the figure by squares ($S1$ and $S2$) and circles ($C1$ and $C2$). The average relevance over the groups is equal and both groups receive almost equal exposure on average, i.e., the ranking is fair w.r.t. disparate treatment ratio (DTR) definition [47]. As a result of equal average relevance and exposure, we expect both groups to receive an (almost) equal number of clicks from the employers. This is shown by the equal-length green bars in the "group average" part of the figure. However, employers' judgments are affected by group bias: they *think* that candidates coming from the square group are more suitable for this type of job, compared to the circle group candidates. Consequently, while being equally relevant and equally exposed, candidates from the circle group are selected less often, as shown by the blue bars in the "group average" part of the figure. We refer to this behavior as *group membership bias*. Our study focuses on the scenario of two groups, where the term "affected" refers to the group whose items are prone to underestimation and receive fewer clicks than they ideally should (the circle group in the example). Considering clicks as the primary measure of user interaction, we provide both theoretical and empirical analyses for the impact of group biased clicks on ranking quality as well as two merit-based fairness of exposure metrics.

Impact on ranking. Similar to other types of bias, group bias can potentially degrade the ranking quality of systems. For example, in Fig. 1, if the clicks are used to infer the relevance of the candidates, without correcting for group bias, inference would be biased towards members from the square group, negatively impacting the ranking quality of the system. In this work, we first theoretically quantify this degradation with an approximation formula for the normalized discounted cumulative gain (NDCG) metric. Then, we experimentally analyze the change in the ranking quality of an LTR model trained on clicks that suffer from group bias, compared to the full information case. Fig. 2 (left) shows an example of the impact of group bias on the ranking performance, measured by NDCG on the Yahoo! dataset with feature number 426 as the sensitive attribute (see Section 6.1). In this plot, the bars associated with the "(non-) affected group" label show the NDCG@10 when *only* the relevant items from the (non-) affected group are considered relevant. Note that a lower group underestimation factor means a higher group bias, and a factor equal to 1 (the leftmost bars) means no group bias. We observe that the affected group is hurt by group bias, while the other group has gained. Importantly, the overall ranking quality is degraded by increasing group bias.

Impact on fairness. Unlike other types of bias that may affect fairness indirectly, group bias has a direct impact on fairness: Clicks suffering from group bias can lead the system to undervalue the utility scores of a particular group (see, e.g., Fig. 2). Consequently, when the expected exposure is assigned to groups based on these biased estimates of the utility, the ranking may not be *truly* fair. For example, in Fig. 1, without correcting for group bias, the square members are inferred to be noticeably more relevant than the circle members and a fair system would swap $S2$ and $C2$ to give more exposure to the group that is more relevant. This means that, based on the *observed* clicks, the $[S1, C1, S2, C2]$ ranking is considered fairer than the original ranking in the figure. But based on the *latent* true relevance values, this ranking is far from being fair, giving the square group too much exposure. For our analyses, we consider two widely used metrics for fairness of exposure, namely *disparate treatment ratio* (DTR) [47] and *expected exposure loss* (EEL) [4, 13]. Each metric has a definition for the *ideal expected exposure* in terms of the utility, that leads to the fairest ranking. Distinguishing between the true (unbiased) utility and observed (biased) utility, we provide formulas for the change in the true fairness metrics, when

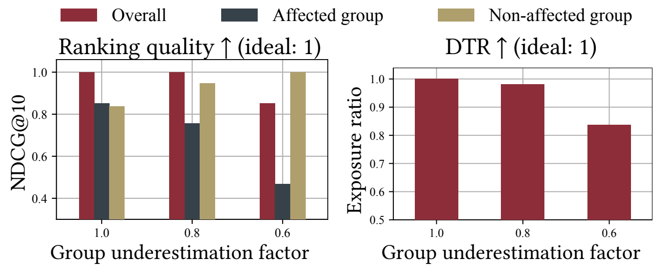

Figure 2: The impact of group bias on ranking performance (left) and DTR fairness metric (right) for the Yahoo! dataset.

the target expected exposure is obtained from the biased utility. Fig. 2 (right) shows an example of the impact of group bias on the DTR fairness metric. With DTR, a ratio of 1 means the fairest exposure, i.e., the leftmost bar with no group bias. Similar to the ranking quality, here we also observe that group bias leads to noticeable deviations from the full information case in DTR metric.

Correction. We follow previous work on implicit bias [26] and model group bias with a multiplicative factor. This allows us to use *inverse propensity scoring* (IPS) to correct for bias [21, 54]. Measuring group bias, however, is not as simple as measuring position bias. We argue that group bias measurement requires assumptions on the distribution of the true utility scores. Following [8, 15, 26], one can assume that the true utility scores of both groups come from the same distribution. However, since *equity* (i.e., merit-based fairness) is based on the premise that exposure should be distributed based on utility, assuming that the utility of different groups is equal for each query, means that different groups should receive equal exposure all the time, which means equality. To counter this equality-instead-of-equity issue, we propose to consider a set of queries (instead of one query) with their corresponding associated items and measure the group bias parameter over this aggregated set of scores. We show that our correction method based on the above amortized measurement of the bias parameter is effective for restoring both the ranking quality and fairness metrics.

Research questions. We answer the following questions:

(RQ1) What is the impact of group bias on the ranking quality and the true fairness metric of head and tail queries?
(RQ2) How can we effectively correct for group bias, without substituting equality for equity?

The rest of the paper is organized as follows. We first list several existing user studies showing that group bias exists in user interactions (Section 2). In Section 3 we give a formal definition of group bias. Answering the research questions starts from Section 4 where we derive mathematical formulas for the impact of group bias on the ranking quality and fairness metrics. To do so, we put some simplifying assumptions such as binary latent relevance and uniform distribution of the observed clicks (i.e., attractiveness) over the items. These assumptions may not be precisely met in real-world datasets, but we believe the derived formulas that are based on these assumptions are beneficial in giving insights into the impact of group bias on ranking. To close the gap between theory and practice, we perform extensive experiments on various real-world datasets and show the negative impact of group bias on the ranking quality and fairness metrics in Section 6.2. Regarding the second research question, in Section 5 we discuss the challenges and solutions for measuring and correcting for group bias from user interactions. Then we experimentally examine the effectiveness of our theoretical propositions for group bias correction in Section 6. Finally, Section 7 concludes the paper.

2 RELATED WORK

An increasing number of studies indicate the existence of group bias. Implicit bias, a special case of group bias, in which the preference of one group over the other is unintentional, has been widely studied in human behavior studies [e.g., 7, 17, 22]. More recently, implicit bias has been formalized in the set selection problem [26] and extended to the ranking scenario [8, 15].

Here, we list a small number of example studies indicating that group membership affects users' judgment. In [25], it is observed in a user study that in a career search, results that are consistent with stereotypes for a career are rated higher. Sühr et al. [49] pose the important question of whether "fair ranking improve[s] minority outcomes?" and arrive at the result that persistent gender preferences of employers can limit the effectiveness of fair ranking algorithms. Krieg et al. [28] in their user study on gender sensitive queries from [27] show that perceived gender bias affects judgment. In [53] it is shown that societal and algorithmic gender bias affect each other: the algorithmic outputs of search engines track pre-existing societal-level gender biases; and, at the same time, exposure of users to these biased results shape users' cognitive concepts and decisions. Liu et al. [34] study gender bias in the evaluation and selection of future leaders.

We study the impact of group bias on ranking and fairness measures and propose a method to correct for it. Closest to our paper are [8, 36], which show that implicit bias degrades ranking quality and that by ensuring equality of exposure, the ranking quality can be improved. What we add on top of this work is to provide a formalization of the change of ranking and merit-based fairness metrics as a result of group bias. We also provide experimental analyses of the impact of group bias on the output of an LTR model.

The idea of our amortized correction to counter sparsity and equality-instead-of-equity has similarities to the notion of amortized fairness of exposure [6], where the exposure and utility of individuals (or groups) are aggregated across multiple queries and the fairness metric is calculated according to the aggregated exposure and utility. This corresponds to fairness evaluation. In contrast, we aggregate the items associated with multiple queries to find the group bias parameter that minimizes the distance between the utility distribution of the affected and non-affected groups. This corresponds to group bias correction.

Recently, Balagopalan et al. [2] have provided a set of desired criteria that relevance scores should satisfy in order to meaningfully guide fairness interventions, and show that, e.g., learning from click data can violate some of the desiderata of relevance for fair ranking. Though their focus is not on group bias, their approach to showing that a violation of assumptions on the utility of items can misguide fairness interventions has similarities to this work.

Remark 1. Our terminology of *group bias* should not be confused with *in-group bias*, where a user favors members from their own group over out-of-group members [37, 39, 57], or *conformity bias*, where users tend to behave similarly to the others in a group [10, 23]. Hence, issues such as loyalty versus neutrality are out of scope.

3 GROUP MEMBERSHIP BIAS

As discussed in Section 2, prior work shows that the judgment of a user about the relevance of an item may be affected by the item's group. Either unconsciously (as in implicit bias [15, 26]) or due to stereotypical bias [25, 28, 46], users tend to rate one group higher than the other. In this paper, we do not aim to deal with the source of this biased behavior and only focus on its impact on algorithms and metrics. We call this behavior the *group membership bias*. Following the well-known examination hypothesis [11] that says that an item is clicked by a user if it is (i) examined and (ii) found attractive by that user, one can attribute group bias to the attractiveness part:

$$P(A \mid q, d, g) = f\big(P(R \mid q, d), g\big), \tag{1}$$

where A, R, q, and d stand for attractiveness, relevance, query, and document, respectively, and g is the group of which d is a member. Eq. (1) states that the attraction of an item to the user not only depends on the item's true relevance to the query, but is also a function of the item's group. Following the literature on implicit bias and gender bias [26, 46], we assume this dependency to have a multiplicative form as follows:

$$P(A \mid q, d, g) = \beta_g \cdot P(R \mid q, d). \tag{2}$$

We call β_g the group underestimation factor, or *group propensity*[1]. Our implicit assumption is that clicks for the affected group are missing completely at random (MCAR) with β_g being the missingness probability. This brings the bias-correction problem back to IPS correction, since the clicks missing due to group bias are analogous to clicks missing based on position. Notice that β_g is not necessarily fixed across all queries. For instance, in the Grep-BiasIR dataset [27], bias-sensitive queries have different expected gender stereotypes, and users are expected to be biased toward the respective gender stereotype of each query.

Remark 2. Group bias is only meaningful in settings where there is a global (objective) notion of relevance in contrast to personalized (subjective) relevance. For example, when ranking students for college entrance, or selecting among job applicants hiring, it is desired to base the decision on an *unbiased* and *unpersonalized* criterion. On the other hand, when searching for a roommate, the relevance is subjective and the group bias concern does not apply.

3.1 Ranking Regimes

We distinguish between two LTR regimes: (i) tabular search for head queries; and (ii) general LTR model for other queries. Note that the majority of previous studies focused only on the general LTR regime [e.g., 38, 48], or the tabular regime [e.g., 6, 47]. In contrast, we follow [52] and consider both LTR regimes.

Tabular search for head queries. In tabular search, users' historical interactions with head queries are directly used to estimate items' utility [24, 30–33, 42, 58]. In this regime, we assume that $P(A)$ can accurately be inferred from clicks: other types of bias such as position and trust bias are corrected for and only group

[1] We consider one sensitive attribute here. Extending our discussions to more attributes with intersectional groups is possible using the formulation in [8, 36].

bias remains in the signals. Our theoretical results on the impact of group bias on different metrics, lie in this regime.

General LTR model for tail queries. For new queries and ones that a tabular model is not confident about, an LTR model is used. We assume that this LTR model is trained over accurate estimations of $P(A)$ from the head queries. Writing $r_{q,d}$ for the relevance of item d to query q, ranking metrics per query can usually be expressed as: $\Delta_q = \sum_d \lambda_{q,d} \cdot r_{q,d}$, where λ is a metric-specific coefficient. Using attractiveness instead of the true relevance to train an LTR model, means that instead of Δ_q, the following metric is being optimized:

$$\hat{\Delta}_q = \sum_d \lambda_{q,d} \cdot P(A \mid q,d) \overset{(2)}{=} \sum_g \beta_g \sum_{d \in g} \lambda_{q,d} \cdot P(R \mid q,d). \quad (3)$$

Comparing Δ_q and Eq. (3), it is easy to see that $\hat{\Delta}_q$ is biased:

$$\mathbb{E}_R\left[\Delta_q\right] = \textstyle\sum_d \lambda_{q,d} \cdot P(R \mid q,d) \neq \hat{\Delta}_q, \quad (4)$$

unless β_g is equal across all groups, i.e., there is no group bias.

4 THEORETICAL RESULTS

We assume that there are two groups $G_{\mathcal{A}}$ (affected) and $G_{\mathcal{N}}$ (non-affected), with $\beta_{\mathcal{A}} < 1$ and $\beta_{\mathcal{N}} = 1$. We further assume binary latent relevance, and that within each group, relevant items are more attractive than non-relevant items:

$$\forall d, d' \in G_i,\ \text{if } r_d = 1 \text{ and } r_{d'} = 0, \text{ then } P(A \mid q,d) > P(A \mid q,d'). \quad (5)$$

For brevity, we write a_d for the attractiveness probability of item d, assuming that there is no confusion about the query. Let the number of candidate items for a query be n, out of which $n_{\mathcal{A}}$ and $n_{\mathcal{N}}$ items belong to groups $G_{\mathcal{A}}$ and $G_{\mathcal{N}}$, respectively. We indicate the number of relevant items with $n^+_{\mathcal{A}}$ and $n^+_{\mathcal{N}}$. To assess the impact of group bias on different metrics, we measure the change in the target metric when the observable attractiveness probabilities are considered as a proxy for the true relevance scores.

4.1 Ranking Quality

For ranking quality, we calculate the NDCG of the list obtained from sorting items based on their attractiveness probability and measure its deviation from the ideal NDCG, i.e., 1. By definition, group bias affects the attractiveness probabilities for $G_{\mathcal{A}}$ only. Let a^* be the minimum attractiveness value for the relevant items of $G_{\mathcal{N}}$:

$$a^* = \min_{d \in G_{\mathcal{N}}} \{a_d \mid r_d = 1\}. \quad (6)$$

Items in $G_{\mathcal{A}}$ with higher attractiveness values than a^* are ranked correctly with probability 1: Group bias has dampened their attractiveness probabilities, but still none of the non-relevant items is ranked higher than them. We define an auxiliary random variable v to be the fraction of relevant items from the affected group $G_{\mathcal{A}}$ that are ranked correctly with probability 1:

$$v = \frac{|\{d \mid d \in G_{\mathcal{A}} \wedge r_d = 1 \wedge a_d > a^*\}|}{|\{d \mid d \in G_{\mathcal{A}} \wedge r_d = 1\}|}. \quad (7)$$

For uniformly distributed scores in the interval of $[0,1]$, we have $\mathbb{E}[v] = \max(2 - \beta_{\mathcal{A}}^{-1}, 0)$.

Theorem 4.1. *In the presence of group bias, for uniformly distributed attractiveness scores, the change in the NDCG of the list, sorted based on items' attractiveness, can be approximated by a linear function of $\mathbb{E}[v]$, i.e., the fraction of affected relevant items that are still as attractive as the non-affected relevant items.*

Proof. Our monotonicity assumption of the within-group attractiveness (Eq. (5)) ensures that no relevant item is ranked lower than non-relevant items of $G_{\mathcal{A}}$. This means that the $1-v$ fraction of the affected relevant items lies somewhere between $n^+_{\mathcal{N}} + vn^+_{\mathcal{A}}$ and $n_{\mathcal{N}} + n^+_{\mathcal{A}}$ positions. The expected discounted cumulative gain (DCG) of the list would be as follows:

$$\mathbb{E}[DCG] = \sum_{i=1}^{n^+_{\mathcal{N}} + vn^+_{\mathcal{A}}} \frac{1}{\log(1+i)} + \sum_{i=n^+_{\mathcal{N}} + vn^+_{\mathcal{A}}+1}^{n_{\mathcal{N}} + n^+_{\mathcal{A}}} \frac{\xi_i}{\log(1+i)}, \quad (8)$$

where ξ_i depends on the distribution of the attractiveness scores. For a uniform distribution, we have:

$$\xi_i = \frac{(1-v)n^+_{\mathcal{A}}}{n_{\mathcal{N}} - n^+_{\mathcal{N}} + (1-v)n^+_{\mathcal{A}}}. \quad (9)$$

Finally, using numerical analysis to approximate the average DCG in Eq. (8) by a linear function of v, leads to a small approximation error, e.g., a relative error of at most 5% in a top-20 setup. □

4.2 Merit-Based Fairness Metrics

Next, to see the impact of group bias on fairness metrics, we analyze two well-known merit-based fairness of exposure metrics, viz. EEL [4, 13] and DTR [47]. For both, we calculate the target exposure in two cases: (i) the full information case where the true relevance scores are used to compute target exposure, and (ii) the group biased case where attractiveness probabilities are used as proxies for relevance to compute target exposure. By *change in true target exposure* we mean the difference between these two cases.

4.2.1 EEL. In the next theorem, we calculate the change in the target exposure of $G_{\mathcal{N}}$ as a result of group bias.

Theorem 4.2. *In the presence of group bias, assuming the Position-Based Model (PBM) as the user browsing model with logarithmic decay of exposure as in DCG,*[2] *the change in the target exposure of EEL can be approximated as follows:*

$$\Delta(EEL) = c \cdot \log\left(\frac{\text{\# True relevant items}}{\text{\# Perceived relevant items}}\right), \quad (10)$$

where c is a constant depending on $n^+_{\mathcal{N}}$, $n_{\mathcal{N}}$, and $n^+_{\mathcal{N}}$.

Proof. As we are working with two groups, and the sum of the group exposures is fixed, to measure the change in the target exposure vector, it is sufficient to measure the change in the target exposure of one group and multiply it by 2.

To compute the expected exposure for EEL, the utility values should be discrete. With a slight abuse of notation, we assume that a^* (instead of Eq. (6)) is the threshold used for discretization of the attractiveness probabilities,[3] and we use $\bar{a}_d$ for the discretized value of a_d. Since $\beta_{\mathcal{N}} = 1$, we assume that $\bar{a}_d = r_d$ for $d \in G_{\mathcal{N}}$. However, for the affected items, because $\beta_{\mathcal{A}} < 1$, not all the scores are necessarily correct. We re-use v from Eq. (7) to show the fraction of affected relevant items that are still recognized as relevant.

For the average exposure of the relevant and non-relevant items we use the following two approximations:

$$\frac{1}{m}\sum_{i=1}^{m} \frac{1}{\log(1+i)} \approx \alpha \log(m) + c \quad (11)$$

[2] Here we follow [14, 52]. Similar analyses can be performed for other exposure models.
[3] Usually, $a^* = 0.5$ is the least controversial threshold.

$$\frac{1}{n-m}\sum_{i=m}^{n}\frac{1}{\log(1+i)} \approx \alpha' \log(m) + c', \tag{12}$$

where α and α' are constants, obtained by numerical analysis. For example, for $n = 20$, $\alpha = -0.146$ and $\alpha' = -0.022$ lead to relative approximation errors of at most 5%. In the full information case, there are a total of $n^+_{\mathcal{N}} + n^+_{\mathcal{A}}$ relevant items, i.e., $m = n^+_{\mathcal{N}} + n^+_{\mathcal{A}}$ in Eq. (11) and (12). But with group bias, only $m = n^+_{\mathcal{N}} + \nu n^+_{\mathcal{A}}$ of the items are recognized as relevant. Consequently, the change in the target exposure as a result of group bias can be approximated as follows:

$$\Delta(\mathrm{EEL}) = 2\big(\alpha n^+_{\mathcal{N}} + \alpha'(n_{\mathcal{N}} - n^+_{\mathcal{N}})\big)\log\left(\frac{n^+_{\mathcal{N}}+n^+_{\mathcal{A}}}{n^+_{\mathcal{N}}+\nu n^+_{\mathcal{A}}}\right). \qquad \square$$

4.2.2 DTR. DTR is a multiplicative metric. To have a meaningful measure for the change in DTR in the presence of group bias, one has to compute the ratio of the target expected exposure in the full information (E) and group-biased ($\tilde{E}$) settings.

Theorem 4.3. *In the presence of group bias, the change in the target exposure of DTR, equals the fraction of affected relevant items that are still as attractive as the non-affected relevant items.*

Proof. Using the same notation as in previous sections, and noting that because of the binary relevance assumption the utility of each group is equal to the number of its relevant items, this ratio is computed as follows:

$$\rho(\mathrm{DTR}) = \frac{\tilde{E}_{\mathcal{A}}}{\tilde{E}_{\mathcal{N}}} \cdot \frac{E_{\mathcal{N}}}{E_{\mathcal{A}}} = \frac{\nu n^+_{\mathcal{N}}+\nu n^+_{\mathcal{A}}}{n^+_{\mathcal{N}}+n^+_{\mathcal{A}}} = \nu. \qquad \square$$

Upshot. In this section we derived formulisations of the negative effect of group bias on ranking. In proving our theorems, we relied on some assumptions on the relevance and distribution of attractiveness. The objective of this section is to provide theoretical insights on the impact of group bias on ranking, i.e., RQ1. Since our assumptions here may not be precisely met in real-world scenarios, in Section 6.2 we get back to RQ1 in a semi-synthesized experimental setup with real-world utility scores and show that, though not necessarily in complete alignment with the formulisations of this section, the message is the same: group bias *does* hurt the quality and fairness of ranking in real-world scenarios, and the degradation gets stronger with more severe group bias.

5 GROUP BIAS CORRECTION

Our multiplicative formulation of group bias in Eq. (2) allows us to use IPS to correct for group bias, once we know the value of the propensity β. The unbiasedness proof of IPS for this case is exactly the same as that of position bias in [21, 54]. However, similar to position bias, the unbiasedness proof depends entirely on an accurate estimation of the bias parameter [50].

Unlike position bias, group bias cannot be measured by intervening in the ranked list of items. The reason is that the bias attribute in position bias can be changed without modifying the content of the items: Each item can be shown in different positions, hence, detaching propensity from relevance. In contrast, for group bias, the bias attribute, i.e., group membership, is a characteristic of the item that *cannot* be changed. As such, users' interactions with items cannot be measured for different values of the bias attribute. Instead, to measure group bias, previous work on implicit bias (with the same problem formulation as Eq. (2)), assumes that the utility scores of different groups come from the same distribution [8, 15, 26]. We use the same assumption, but extend it to an amortized criterion.

5.1 Measurement

Let $\mathbf{A}_{\mathcal{A}}$ and $\mathbf{A}_{\mathcal{N}}$ be the set of (observed) attractiveness scores, and $\mathbf{R}_{\mathcal{A}}$ and $\mathbf{R}_{\mathcal{N}}$ the set of (latent) relevance scores of $G_{\mathcal{A}}$ and $G_{\mathcal{N}}$, respectively. Let $\Delta_{\mathbb{D}}$ be a non-parametric test for the equality of one-dimensional probability distributions such as the Kolmogorov-Smirnov (KS) [35] test. The assumption that the utility scores of the two groups come from the same distribution means that:

$$\lim_{|\mathbf{R}_{\mathcal{A}}|,|\mathbf{R}_{\mathcal{N}}|\to\infty} \Delta_{\mathbb{D}}\left(\mathbf{R}_{\mathcal{A}}, \mathbf{R}_{\mathcal{N}}\right) = 0. \tag{13}$$

Assuming Eq. (2) to be the relation between $\mathbf{A}_{\mathcal{A}}$ and $\mathbf{R}_{\mathcal{A}}$, the best estimation of $\beta_{\mathcal{A}}$ is given by the following optimization problem:

$$\hat{\beta}_{\mathcal{A}} = \operatorname*{argmin}_{\beta_{\mathcal{A}}} \Delta_{\mathbb{D}}\left(\frac{\mathbf{A}_{\mathcal{A}}}{\beta_{\mathcal{A}}}, \mathbf{A}_{\mathcal{N}}\right), \tag{14}$$

where $\mathbf{A}_{\mathcal{A}}/\beta_{\mathcal{A}}$ is the set obtained by dividing all the scores in $\mathbf{A}_{\mathcal{A}}$ by $\beta_{\mathcal{A}}$. In our experiments, we choose the KS test for $\Delta_{\mathbb{D}}$ and use grid search to solve the one-dimensional optimization of Eq. (14).

It only remains to define how the sets $\mathbf{A}_{\mathcal{A}}$ and $\mathbf{A}_{\mathcal{N}}$ should be constructed. Naively constructing these sets per query has two issues: (i) *Sparsity*: Usually, we do not have a large number of items with non-zero exposure, associated with one query in real-world search engines. On the other hand, statistical tests measuring the distance between probability distributions work best with large numbers of data points. (ii) *Equality-instead-of-equity*: Assuming the same distribution for the utility of different groups can make the notion of *equity* meaningless, as the implicit assumption in merit-based fairness metrics is that different groups may have different utilities.

Remark 3. Our assumption that the utility scores come from the same distribution derives from the principle of maximum entropy: unless there are explicit and justified reasons indicating that different groups have different utility score distributions, it is only reasonable to assume the same distribution. Prior work on implicit bias [8, 15, 26] is based on this same assumption.

5.2 Amortized Correction

Instead of using Eq. (14) per-query, in the amortized correction method we consider a set of queries with the same group propensity and aggregate the utility scores of their associated items. The sets $\mathbf{A}_{\mathcal{A}}$ and $\mathbf{A}_{\mathcal{N}}$ contain the attractiveness scores of these aggregated items. This aggregation addresses both issues mentioned in Section 5.1: (i) With multiple queries, the size of the sets $\mathbf{A}_{\mathcal{A}}$ and $\mathbf{A}_{\mathcal{N}}$ grows, reducing the variance. (ii) Amortized equality does not force per-query equality. Fig. 3 shows an example of the difference between per-query and amortized correction. In this example, there are three queries, each with two results from the squares group and two from the circles group. In all of the queries, the circles group is the affected group. In per-query correction for group bias, the bias parameter is over-estimated for query 1 (the blue bar is shorter than the relevance for the circles group members), but under-estimated for query 2 (the blue bar is longer). For query 3, since the square members have a lower average number of clicks compared to the circle members, the per-query correction wrongly detected the squares group as the affected group and the discrimination between the

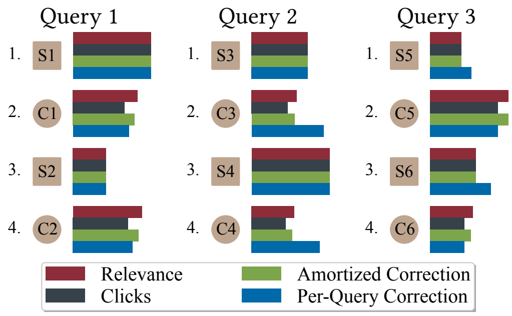

Figure 3: Contrasting per-query and amortized correction.

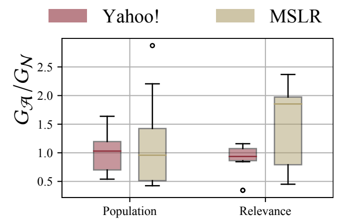

Figure 4: Ratio of affected to non-affected group members in terms of population and average utility score (relevance) for different sensitive attributes in Yahoo! and MSLR datasets.

groups has been boosted after the correction. On the other hand, by aggregating the clicks on all six square members and six circle members and measuring a single group bias parameter for all the queries, the amortized correction has a noticeably better performance at recovering the true relevance (the green bars are close to the red bars).

The amortized correction, however, introduces a new challenge: How to detect queries with the same group propensity, before measuring their group propensity? One way to break this cyclic dependency is by using extra knowledge. Notice that in order to detect queries with almost the same group propensity, it is only required to have a clustering of queries. Previous work shows that such a clustering exists for a number of group attributes such as gender [27]. In this paper, we first assume that such a clustering of queries is given. Then, in an ablation experiment, we further show that even loosely clustering the queries, when an accurate and more specific clustering is not available, improves the ranking quality and fairness metrics over the naive case of not correcting for group bias.

Upshot. We relied on prior studies for the existence of group bias in user interactions and provided theoretical results about its impact on the ranking and merit-based fairness metrics. Then, we proposed an amortized correction method for group bias that addresses the equality-instead-of-equity issue of the per-query correction. Next, we test our theoretical findings and arguments experimentally.

6 EXPERIMENTAL RESULTS

In our experiments we investigate the following questions regarding group bias: (i) Is the impact of group bias on degrading the ranking quality and fairness metrics consistent for different sensitive attributes and in different datasets? (ii) Can our correction method effectively correct for group bias? (iii) How does the amortized approach compare to the per-query approach for correction? (iv) How robust is our correction method to the accuracy of clustering the queries based on their group propensity?

6.1 Setup

Dataset. We use four datasets with provided sensitive attributes and two with synthesized sensitive attributes. (i) **IIT-JEE**: The dataset comprises the scores of candidates who took the Indian Institutes of Technology Joint Entrance Exam (IIT-JEE) in 2009. This information was made public in June 2009, following a Right to Information request [29]. It contains the scores of about $385k$ students, the student's gender ($98k$ women and $287k$ men), their birth category (see [3]), and zip code. This dataset was used in prior work on implicit bias [e.g., 8]. We normalize the scores to the $[0, 1]$ interval using min-max normalization. Furthermore, we simulate queries by grouping the students based on their birth category and zip code. This gives $48.6k$ queries, among which we only keep the ones with both genders and at least one normalized score above 0.5 and one below 0.5. The filtering gives us $2.9k$ queries with a total of $205k$ scores. (ii–iii) **TREC 2019 and 2020**: The academic search dataset provided by the TREC Fair Ranking track 2019 and 2020 [5]. These datasets come with 632 and 200 training queries, respectively, with an average of 6.7 and 23.5 documents per query. Following [45, 52], we divide the items (i.e., papers) into two groups based on their authors' h-index. (iv) **MovieLens** $1M$: The classic movie recommendation dataset comprising $1M$ movie ratings that were provided by $6k$ users for $3.9k$ different movies. We scraped IMDB to obtain the country of origin and box office cumulative worldwide gross values for each item (i.e., movie). For the sensitive attributes, we consider two groupings as follows. In MovieLens$_{[Co.]}$, we divide the movies based on their first listed country of origin into United States (US) and non-US groups with $2.7k$ and $1.2k$ movies and $807k$ and $193k$ ratings, respectively. In MovieLens$_{[BO]}$, we divide the movies based on their box office with a threshold of $100M\$$ into high and low-grossing groups with 388 and $3.5k$ movies and $324k$ and $676k$ ratings, respectively.

LTR dataset. To analyze the impact of group bias in the general LTR regime, following prior work on unbiased LTR [21, 50, 51], we the Yahoo! Webscope [9] and MSLR-WEB30k [41] datasets that are represented by query-document feature vectors of lengths 501 and 131, respectively, and both have graded relevance labels from 0 to 4. For our experiments on the tabular regime, we use the training set of the Yahoo! and MSLR datasets, with $20k$ and $19k$ queries together with $473k$ and $2.2M$ documents, respectively. The test set of Yahoo! and MSLR dataset contains $6.7k$ and $6k$ queries together with $163k$ and $749k$ documents, respectively. Test queries are used for our experiments on the general LTR regime.

Sensitive attribute for LTR datasets. We extend prior work [13, 52, 56] and use a data-driven approach for selecting features as sensitive attributes and dividing items into two groups based on some threshold on that feature. We notice that synthetic sensitive attributes are not a substitute for the real sensitive attributes in real-world datasets. As mentioned above, we perform experiments on four real-world datasets with provided *meaningful* sensitive attributes. Our objective for using LTR datasets with synthetic sensitive attributes is to extend and complement the results on real-world sensitive attributes with 32 new setups with various group utility and population dynamics as will be discussed shortly. Our criterion for selecting a feature as a sensitive attribute is as

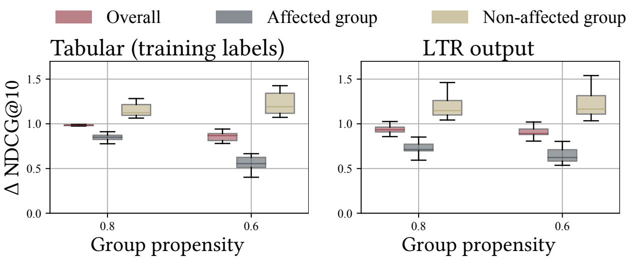

Figure 5: The impact of group bias on ranking quality for the Yahoo! and MSLR datasets with different sensitive attributes.

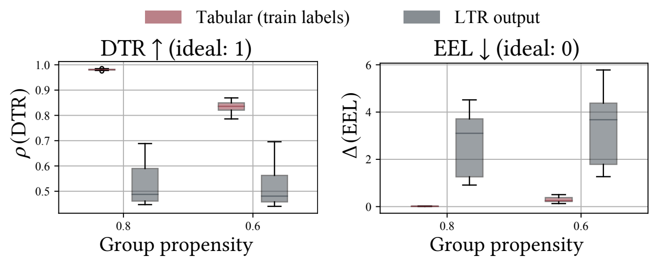

Figure 6: The impact of group bias on fairness metrics for the Yahoo! and MSLR datasets with different sensitive attributes.

follows: For each feature we divide the items in two groups based on a threshold equal to the mean minus one standard deviation of that feature. If more than 95% of queries have at least one item from both groups, we select the feature as a candidate for sensitive attribute. Based on this criterion, we have selected features $\{5, 88, 100, 141, 155, 264, 393, 426\}$ and $\{11, 14, 15, 126, 127, 130, 131, 132\}$ from the Yahoo! and MSLR datasets, respectively. Fig. 4 gives an overview of the ratio of affected to non-affected group members in terms of population and average utility score. In what follows we use, e.g., Yahoo!$_{[426]}$ for the Yahoo! dataset with feature number 426 as the sensitive feature. For each feature, we assume two groupings based on two thresholds: (i) mean value; and (ii) mean minus one standard deviation. This yields a total of 32 setups.

Bias simulation. To simulate group bias, we use Eq. (2), but to make the simulation more realistic, we add a normal noise to the $\beta_{\mathcal{A}}$ value for each query. We experiment with two propensities $\beta_{\mathcal{A}} \in \{0.6, 0.8\}$ and use $\sigma_\beta = 0.1$ for the standard deviation of the normal noise. In Sec. 6.4.2, to add to the uncertainty of the setup, we also experiment with higher noise variances of $\sigma_\beta \in \{0.2, 0.3\}$. For our correction method, we found that adding a small amount of noise to the scores for breaking the ties, without swapping the order of the grades, helps to have a smoother curve for β in Eq. (14).

LTR model. For the general LTR model (for tail queries) we use a neural network with attention and LambdaRank Loss as in [40].

6.2 Impact of Group Bias

First, we show that group bias, on both tabular and LTR regimes, consistently has a negative impact on the ranking quality and fairness metrics. To do so, we run experiments on two datasets, namely Yahoo! and MSLR, each with 8 different features as the sensitive attribute, and two thresholds for separating the groups (see the "Sensitive attribute" paragraph in Section 6.1). This gives us a total of 32 different setups. For each setup, we simulate the attractiveness probabilities with $\beta_{\mathcal{A}} \in \{0.8, 0.6\}$ as *mild* and *severe* cases of group bias, and compare the NDCG@10, DTR, and EEL metrics against the full information case. Our experiments with other values of bias parameters led to consistent results.

Fig. 5 shows a summary of the impact of group bias on the ranking performance of both tabular and LTR regimes. These results confirm that group bias degrades the ranking quality of the affected group in the tabular regime and this damage is also reflected in the LTR output, trained over the biased training labels. As a result of pushing down the relevant members of the affected group, the non-affected group gains ranking quality, i.e., the NDCG of the non-affected group with group bias is higher than the full information case. However, the overall ranking is worsened with group bias. Comparing the tabular (left) and LTR output (right) plots in Fig. 5, we observe that increasing the severity of group bias from 0.8 to 0.6, affects the tabular regime more. This may be because the impact of group bias on the LTR outputs is indirect.

Fig. 6 shows a summary of the change in fairness metrics of both tabular and LTR regimes as a result of group bias. For example, a value of $\rho(\text{DTR}) = 0.5$ in the left plot means that on average, the target (i.e., ideal) exposure computed by the biased attractiveness scores differs from the true target exposure computed in the full information case by a factor of 0.5. Similarly, a value of $\Delta(\text{EEL}) = 3$ in the right plot means that on average, the target exposure of the biased case has an ℓ_2 distance of 3 to the full information target exposure. These results confirm that group bias changes the target exposure in the tabular regime and this change is reflected in the LTR output, trained over the biased training labels. Consequently, when the system distributes the exposure according to the target exposure to make a ranking fair, if the scores are suffering from group bias, the result is not truly fair.

6.3 Amortized Correction

In the next set of experiments, we show the effectiveness of our proposed correction method in Section 5 for compensating for the negative effect of group bias. Table 1 compares the ranking quality, in terms of NDCG@10, as well as two merit-based fairness metrics, DTR and EEL, between the biased and corrected cases in the tabular regime. Note here that in the tabular regime, it is assumed that accurate relevance estimations, up to the group bias, are available, meaning that the unbiased metrics get their ideal values, i.e., unbiased NDCG@10 equals 1. When the corrected NDCG@10 reaches 1 (e.g., MovieLens with both bias parameter values), it means a full recovery from group bias. In all setups, our correction method improves the ranking quality and fairness metrics over the biased case. With some exceptions for the ranking quality with mild group bias ($\beta = 0.8$), the improvements are significant. For each bias parameter value, we have also included the estimated value obtained from Eq. (14). We observe that our estimated bias values, i.e., $\hat{\beta}_{\mathcal{A}}$, are close to their corresponding true values β.

We further analyze the effectiveness of our correction method in the general LTR regime. Table 2 contains the comparison of ranking quality and fairness metrics between the biased and corrected cases in our tested LTR datasets. We also report the full information case in the table. Here, we only report the results for one sensitive attribute for each dataset, noting that the results for other sensitive attributes lead to similar observations. Similar to the tabular regime, here we also observe performance improvements as a result of our correction method, compared to the biased case. Except for the

Table 1: The impact of our amortized group bias correction on ranking and fairness metrics in the tabular regime. $\hat{\beta}_{\mathcal{A}}$ shows the estimated value of the bias parameter as in Eq. (14). For each metric, the columns "*B*" and "C" show the "*Biased*" and "Corrected" performances, respectively. Superscripts * indicate a significant improvement over the biased case with $p < 0.001$.

	$\beta = 0.8$							$\beta = 0.6$						
	$\hat{\beta}_{\mathcal{A}}$	NDCG@10 ↑		ρ(DTR) ↑		Δ(EEL) ↓		$\hat{\beta}_{\mathcal{A}}$	NDCG@10 ↑		ρ(DTR) ↑		Δ(EEL) ↓	
		B	C	*B*	C	*B*	C		*B*	C	*B*	C	*B*	C
Yahoo!$_{[426]}$	0.825	*0.987*	0.996*	*0.820*	0.955*	*0.447*	0.120*	0.626	*0.885*	0.988*	*0.641*	0.941*	*1.809*	0.172*
MSLR$_{[127]}$	0.843	*0.975*	0.991*	*0.813*	0.948*	*1.687*	0.308*	0.648	*0.780*	0.966*	*0.627*	0.926*	*7.146*	0.471*
IIT-JEE	0.727	*0.989*	0.991	*0.799*	0.906*	*0.504*	0.341*	0.547	*0.970*	0.985*	*0.600*	0.901*	*1.401*	0.410*
MovieLens$_{[Co.]}$	0.822	*1.000*	1.000	*0.800*	0.962*	*1.101*	0.513*	0.612	*0.998*	1.000*	*0.602*	0.958*	*7.113*	1.908*
MovieLens$_{[BO]}$	0.781	*1.000*	1.000	*0.799*	0.974*	*2.330*	0.895*	0.579	*0.994*	1.000*	*0.600*	0.961*	*24.800*	2.831*
TREC 2019	0.838	*0.997*	1.000	*0.888*	0.954*	*0.041*	0.020*	0.634	*0.986*	0.999*	*0.771*	0.937*	*0.129*	0.028*
TREC 2020	0.821	*0.995*	0.999	*0.815*	0.954*	*0.356*	0.114*	0.614	*0.945*	0.995*	*0.627*	0.941*	*1.152*	0.137*

Table 2: The impact of our amortized group bias correction on ranking and fairness metrics in general LTR regime. For each metric, the columns "*B*" and "C" show the "*Biased*" and "Corrected" performances, respectively.

	Full info.			$\beta = 0.8$						$\beta = 0.6$					
	NDCG	ρ(DTR)	Δ(EEL)	NDCG@10 ↑		ρ(DTR) ↑		Δ(EEL) ↓		NDCG@10 ↑		ρ(DTR) ↑		Δ(EEL) ↓	
				B	C	*B*	C	*B*	C	*B*	C	*B*	C	*B*	C
Yahoo!$_{[426]}$	0.649	0.477	0.673	*0.615*	0.645*	*0.319*	0.428*	*1.970*	0.741*	*0.580*	0.645*	*0.313*	0.456*	*2.711*	0.728*
MSLR$_{[127]}$	0.320	0.681	1.779	*0.283*	0.313*	*0.675*	0.684	*4.380*	2.016*	*0.265*	0.309*	*0.671*	0.683	*5.803*	2.147*

DTR metric in MSLR, all the improvements are significant with $p < 0.001$. Compared with the full information case, we observe that in the Yahoo! dataset, our correction method leads to full recovery of NDCG@10, while in the MSLR dataset, there remains a slight gap toward the full information quality. One reason for this difference could be the distribution of relevant items in the affected and non-affected groups: In Yahoo!$_{[426]}$ the ratio between the mean relevance of items in $G_{\mathcal{A}}$ to $G_{\mathcal{N}}$ is 1.05, whereas the same quantity in MSLR$_{[127]}$ is 2.21. Therefore, the assumption of similar utility score distributions for both groups is closer to reality in Yahoo!$_{[426]}$ than in MSLR$_{[127]}$. Similarly to NDCG, we observe that DTR and EEL are almost fully recovered from group bias in the Yahoo! dataset, but not in the MSLR dataset.

Finally, Fig. 7 shows a summary of the ranking quality of biased (left) and corrected (right) utility scores in the tabular regime on all 32 setups of the LTR datasets mentioned in Sec. 6.2. In all but two cases, we observe that our correction method effectively improves the ranking quality over the biased case and achieves NDCG@10 close to 1. The two outlier cases correspond to Yahoo!$_{[5]}$ (for each feature, two different thresholds for separating the groups are used), where the ratio between the average utility of the affected group and the non-affected group is as low as 0.3. This is the same outlier as in Fig. 4. It is worth mentioning that in a slightly less severe violation of the same distribution assumption, i.e., MSLR$_{[130]}$ with a utility ratio of 0.45, our correction method is able to improve the ranking quality over the biased case. One interesting future direction would be to find out if this phenomenon, i.e., having the true average utility of the underrepresented group considerably lower than the other group, happens in real-world settings and how to correct for the bias in such cases.

Remark 4. This paper introduces a novel bias paradigm. Our study emphasizes that even the supposedly ideal fair rankings are not actually fair when group bias is unaddressed (RQ1); and that employing amortized correction yields more robust results (RQ2). Baseline comparisons are only pertinent to RQ2. We compare our amortized correction method to the per-query method which is the *only* existing method for group bias measurement and correction.

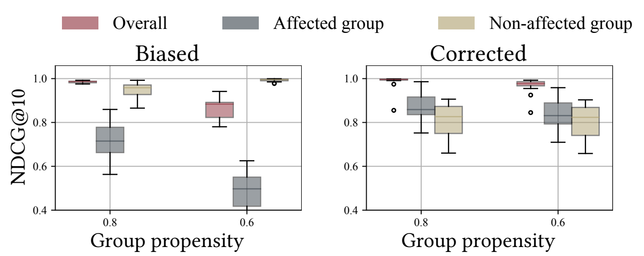

Figure 7: The ranking performance of biased (left) and corrected (right) scores for the Yahoo! and MSLR datasets with different sensitive attributes.

6.4 Ablation Study

6.4.1 Impact of cluster size on correction. In Section 5.2 we argued against measuring group propensity for each query. Here, we analyze the impact of cluster size on the correction method. Fig. 8 shows the ranking quality of the corrected scores as a function of the cluster size. The overall ranking quality (red line) improves as the cluster size grows and it converges to its final value at around a cluster size of 10. For severe group bias ($\beta_{\mathcal{A}} = 0.6$), which we omit due to space restrictions, the same pattern is observed, but with a convergence point of 100. In both cases, using a cluster size below the convergence point leads to corrected rankings that are even worse than the biased ranking. Comparing the ranking quality of

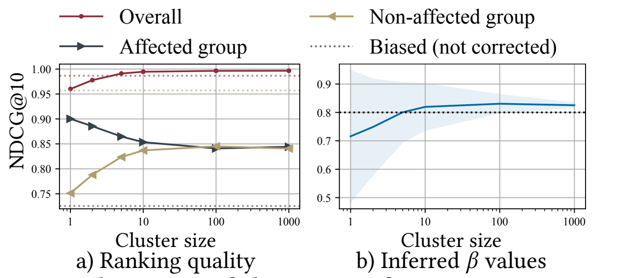

Figure 8: The impact of cluster size of group propensity on the amortized correction for group bias ($\beta_{\mathcal{A}} = 0.8$) for the Yahoo!$_{[426]}$ dataset. Ranking quality of corrected scores (a), and accuracy of the inferred group propensity (b).

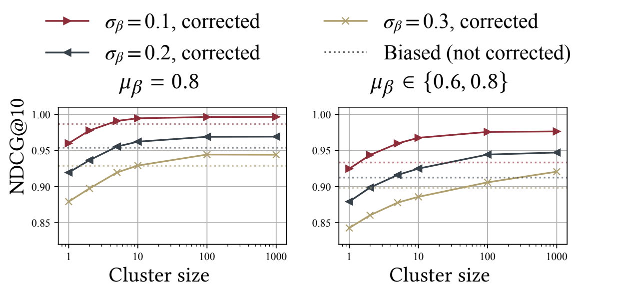

Figure 9: Effectiveness of amortized correction method when accurate clustering is not available. Yahoo!$_{[426]}$ dataset.

the affected group (black line) with the non-affected group (golden line), we observe that smaller clusters result in over-compensation of group bias. The reason is revealed in Fig. 8(b): for smaller cluster sizes, the inferred propensity is under-estimated, leading to larger corrected scores for the affected group members. One other interesting observation in Fig. 8(b) is the high variance of the inferred propensity for small clusters (issue (i) in Section 5.2).

6.4.2 Impact of clustering accuracy. Finally, we address the challenge of inaccurate clustering of queries based on their group propensity that we raised at the end of Section 5.2. The main goal of the following sets of experiments is to show that our correction method, even when accurate clustering of queries is not available, is still effective in improving the ranking quality over the biased case. To confirm this, we add to the uncertainty of our simulation setup in two different ways: (i) *Higher variance*: We increase the variance of the group propensity when simulating attractiveness. We consider $\sigma_\beta \in \{0.2, 0.3\}$. (ii) *Two modes*: Instead of using a unimodal normal distribution to simulate group propensity, we use a mixture model with two modes $\{0.6, 0.8\}$. This means that for half of the queries, the group propensity follows a normal distribution with a mean of 0.6 while for the other half, the mean is 0.8 and during inference, we are not given the information about which query belongs to which mode. Fig. 9 shows the ranking quality of the corrected scores w.r.t. different cluster sizes. In both plots, we observe that increasing the variance of the simulated group propensity both increases the negative impact of group bias on ranking (dotted lines) and makes it harder to correct for the bias (solid lines). The important result of these experiments, however, is that even though the uncertainty about group propensity is high, our amortized correction method almost always improves the ranking quality over the biased case. Note that in all setups, per-query correction as well as clusters with a small size lead to worse ranking qualities than the biased scores. Interestingly, when there are two modes of group propensity (right plot), our correction method, oblivious to the mode membership and assuming a fixed propensity, is able to correct the scores and achieve a ranking performance higher than the biased case.

7 CONCLUSION AND FUTURE WORK

We have addressed group membership bias, which is based on the observation that a user's perception of an item's group membership may affect their judgment about the utility of an item. We have provided extensive theoretical and empirical analyses of the impact of group bias on the ranking quality and two fairness of exposure metrics, DTR and EEL. By utilizing an auxiliary variable v as the fraction of affected relevant items that are still as attractive as the non-affected relevant items, we have shown that, in the presence of group bias, NDCG and DTR change linearly with v, while the change in EEL has a more complex form in terms of v.

Correcting for group bias, which is a type of *content-based* bias, is not as easy as *context-based* types of bias such as position and trust bias. To measure group bias, assumptions based on fairness constraints should be made about the utility distribution of different groups. However, such assumptions can potentially make the equity-based notion of fairness meaningless. Amortized correction for the group bias is our solution to this issue, as global equality does not contradict local equity. We have experimentally confirmed that our correction method, when its assumptions are met, is able to fully recover the scores suffering from group bias.

There are several future directions to this study. One way to extend our results is to propose measurement and correction methods that perform better with increased uncertainty of group propensity. Moreover, as group bias is based on users' perception of group membership, it can change over time. Analyzing group bias in a dynamic setting is therefore another direction. This study deals with the impact of group bias on fair exposure and, hence, we only consider the so-called treatment-based fairness metrics. In contrast, some studies focus on impact-based fairness [44], where the objective is to make sure that items receive a fair amount of impact, e.g., clicks. While our work suggests a way to correct for group bias in the historical clicks in order to make the exposure in future rankings fair, a next direction would be to account for group bias when optimizing for impact-based fairness.

CODE AND DATA

To ensure the reproducibility of the reported results, this work only made use of publicly available data and our experimental implementation can be accessed publicly at https://github.com/AliVard/groupbias.

ACKNOWLEDGMENTS

This research was supported by Elsevier, by the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.001.551, 024.004.022, NWA.1389.20.183, and KICH3.LTP.20.006, and by the European Union's Horizon Europe program under grant agreement No. 101070212. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Qingyao Ai, Keping Bi, Cheng Luo, Jiafeng Guo, and W Bruce Croft. 2018. Unbiased Learning to Rank with Unbiased Propensity Estimation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 385–394.

[2] Aparna Balagopalan, Abigail Z. Jacobs, and Asia J. Biega. 2023. The Role of Relevance in Fair Ranking. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2650–2660. https://doi.org/10.1145/3539618.3591933

[3] Surender Baswana, Partha Pratim Chakrabarti, Sharat Chandran, Yashodhan Kanoria, and Utkarsh Patange. 2019. Centralized Admissions for Engineering Colleges in India. In *Proceedings of the 2019 ACM Conference on Economics and Computation*. 323–324.

[4] Asia J. Biega, Fernando Diaz, Michael D. Ekstrand, Sergey Feldman, and Sebastian Kohlmeier. 2021. Overview of the TREC 2020 Fair Ranking Track. In *The Twenty-Ninth Text REtrieval Conference Proceedings (TREC 2020)*.

[5] Asia J. Biega, Fernando Diaz, Michael D. Ekstrand, and Sebastian Kohlmeier. 2019. Overview of the TREC 2019 Fair Ranking Track. In *The Twenty-Eighth Text REtrieval Conference (TREC 2019) Proceedings*.

[6] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 405–414. https://doi.org/10.1145/3209978.3210063

[7] Michael Brownstein. 2017. Implicit Bias. In *Stanford Encyclopedia of Philosophy*, Edward Zalta (Ed.).

[8] L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. 2020. Interventions for Ranking in the Presence of Implicit Bias. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 369–380. https://doi.org/10.1145/3351095.3372858

[9] Olivier Chapelle and Yi Chang. 2011. Yahoo! Learning to Rank Challenge Overview. In *Proceedings of the Learning to Rank Challenge (Proceedings of Machine Learning Research)*, Vol. 14. PMLR, 1–24. https://proceedings.mlr.press/v14/chapelle11a.html

[10] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and Debias in Recommender System: A Survey and Future Directions. *ACM Trans. Inf. Syst.* 41, 3, Article 67 (feb 2023), 39 pages. https://doi.org/10.1145/3564284

[11] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. *Click Models for Web Search*. Morgan & Claypool Publishers.

[12] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2023. Fairness in Recommender Systems: Research Landscape and Future Directions. *User Modeling and User-Adapted Interaction* (2023), 1–50.

[13] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/3340531.3411962

[14] Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. 2022. Overview of the TREC 2021 Fair Ranking Track. In *The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings*.

[15] Vitalii Emelianov, Nicolas Gast, Krishna P. Gummadi, and Patrick Loiseau. 2020. On Fair Selection in the Presence of Implicit Variance. In *Proceedings of the 21st ACM Conference on Economics and Computation (EC '20)*. Association for Computing Machinery, New York, NY, USA, 649–675. https://doi.org/10.1145/3391403.3399482

[16] Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. 2020. Gender Stereotype Reinforcement: Measuring the Gender Bias Conveyed by Ranking Algorithms. *Information Processing & Management* 57, 6 (2020), 102377.

[17] Chloë FitzGerald and Samia Hurst. 2017. Implicit Bias in Healthcare Professionals: A Systematic Review. *BMC Medical Ethics* 18, 1 (2017), 1–18. https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-017-0179-8

[18] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 2125–2126. https://doi.org/10.1145/2939672.2945386

[19] Di Jin, Luzhi Wang, He Zhang, Yizhen Zheng, Weiping Ding, Feng Xia, and Shirui Pan. 2023. A Survey on Fairness-aware Recommender Systems. *Information Fusion* 100 (2023), 101906. https://doi.org/10.1016/j.inffus.2023.101906

[20] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately Interpreting Clickthrough Data as Implicit Feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 154–161.

[21] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased Learning-to-Rank with Biased Feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 781–789.

[22] Christine Jolls and Cass R. Sunstein. 2006. The Law of Implicit Bias. *Calif. L. Rev.* 94 (2006), 969.

[23] Kilicheva Karomat. 2023. Unconscious Bias in the Workplace. *International Journal on Integrated Education* 6, 4 (2023), 266–268.

[24] Sumeet Katariya, Branislav Kveton, Csaba Szepesvari, and Zheng Wen. 2016. DCM Bandits: Learning to Rank with Multiple Clicks. In *International Conference on Machine Learning*. 1215–1224.

[25] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3819–3828. https://doi.org/10.1145/2702123.2702520

[26] Jon Kleinberg and Manish Raghavan. 2018. Selection Problems in the Presence of Implicit Bias. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[27] Klara Krieg, Emilia Parada-Cabaleiro, Gertraud Medicus, Oleg Lesota, Markus Schedl, and Navid Rekabsaz. 2022. Grep-BiasIR: A Dataset for Investigating Gender Representation-Bias in Information Retrieval Results. *arXiv preprint arXiv:2201.07754* (2022).

[28] Klara Krieg, Emilia Parada-Cabaleiro, Markus Schedl, and Navid Rekabsaz. 2022. Do Perceived Gender Biases in Retrieval Results Affect Relevance Judgements? *arXiv preprint arXiv:2203.01731* (2022).

[29] Rajeev Kumar. 2009. RTI Complaint. Decision No. CIC/SG/C/2009/001088/5392, Complaint No. CIC/SG/C/2009/001088.

[30] Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. 2015. Cascading Bandits: Learning to Rank in the Cascade Model. In *International Conference on Machine Learning*. 767–776.

[31] Paul Lagrée, Claire Vernade, and Olivier Cappé. 2016. Multiple-play Bandits in the Position-based Model. In *Advances in Neural Information Processing Systems*. 1597–1605.

[32] Tor Lattimore and Csaba Szepesvári. 2020. *Bandit Algorithms*. Cambridge University Press.

[33] Chang Li, Branislav Kveton, Tor Lattimore, Ilya Markov, Maarten de Rijke, Csaba Szepesvári, and Masrour Zoghi. 2020. BubbleRank: Safe Online Learning to Rerank via Implicit Click Feedback. In *Uncertainty in Artificial Intelligence*. PMLR, 196–206.

[34] Zhi Liu, Aneeta Rattan, and Krishna Savani. 2023. Reducing Gender Bias in the Evaluation and Selection of Future Leaders: The Role of Decision-makers' Mindsets about the Universality of Leadership Potential. *Journal of Applied Psychology* (2023). https://doi.org/10.1037/apl0001112

[35] Frank J Massey Jr. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Amer. Statist. Assoc.* 46, 253 (1951), 68–78.

[36] Anay Mehrotra, Bary S. R. Pradelski, and Nisheeth K. Vishnoi. 2022. Selection in the Presence of Implicit Bias: The Advantage of Intersectional Constraints. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 599–609. https://doi.org/10.1145/3531146.3533124

[37] Pascal Molenberghs. 2013. The Neuroscience of In-group Bias. *Neuroscience & Biobehavioral Reviews* 37, 8 (2013), 1530–1536. https://doi.org/10.1016/j.neubiorev.2013.06.002

[38] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling Fairness and Bias in Dynamic Learning-to-Rank. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 429–438. https://doi.org/10.1145/3397271.3401100

[39] Lucy O'Sullivan, Whitney Kagabo, Niyathi Prasad, Dawn Laporte, and Amiethab Aiyer. 2023. Racial and Ethnic Bias in Medical School Clinical Grading: A Review. *Journal of surgical education* (2023).

[40] Przemyslaw Pobrotyn, Tomasz Bartczak, Mikolaj Synowiec, Radoslaw Bialobrzeski, and Jaroslaw Bojar. 2020. Context-Aware Learning to Rank with Self-Attention. *arXiv preprint arXiv:2005.10084* (2020).

[41] Tao Qin and Tie-Yan Liu. 2013. Introducing LETOR 4.0 Datasets. *arXiv preprint arXiv:1306.2597* (2013).

[42] Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. 2008. Learning Diverse Rankings with Multi-armed Bandits. In *Proceedings of the 25th International Conference on Machine Learning*. 784–791.

[43] Amifa Raj and Michael D. Ekstrand. 2022. Measuring Fairness in Ranked Results: An Analytical and Empirical Comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 726–736. https://doi.org/10.1145/3477495.3532018

[44] Yuta Saito and Thorsten Joachims. 2022. Fair Ranking as Fair Division: Impact-Based Individual Fairness in Ranking. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*. Association for Computing Machinery, New York, NY, USA, 1514–1524. https://doi.org/10.1145/3534678.3539353

[45] Fatemeh Sarvi, Maria Heuss, Mohammad Aliannejadi, Sebastian Schelter, and

Maarten de Rijke. 2022. Understanding and Mitigating the Effect of Outliers in Fair Ranking. In *WSDM 2022: The Fifteenth International Conference on Web Search and Data Mining*. ACM, 861–869.

[46] Shrikant Saxena and Shweta Jain. 2021. Exploring and Mitigating Gender Bias in Recommender Systems with Explicit Feedback. *arXiv preprint arXiv:2112.02530* (2021).

[47] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2219–2228. https://doi.org/10.1145/3219819.3220088

[48] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for Fairness in Ranking. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/9e82757e9a1c12cb710ad680db11f6f1-Paper.pdf

[49] Tom Sühr, Sophie Hilgard, and Himabindu Lakkaraju. 2021. Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 989–999. https://doi.org/10.1145/3461702.3462602

[50] Ali Vardasbi, Maarten de Rijke, and Ilya Markov. 2021. Mixture-Based Correction for Position and Trust Bias in Counterfactual Learning to Rank. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 1869–1878. https://doi.org/10.1145/3459637.3482275

[51] Ali Vardasbi, Harrie Oosterhuis, and Maarten de Rijke. 2020. When Inverse Propensity Scoring Does Not Work: Affine Corrections for Unbiased Learning to Rank. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 1475–1484. https://doi.org/10.1145/3340531.3412031

[52] Ali Vardasbi, Fatemeh Sarvi, and Maarten de Rijke. 2022. Probabilistic Permutation Graph Search: Black-Box Optimization for Fairness in Ranking. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 715–725. https://doi.org/10.1145/3477495.3532045

[53] Madalina Vlasceanu and David M. Amodio. 2022. Propagation of Societal Gender Inequality by Internet Search Algorithms. *Proceedings of the National Academy of Sciences* 119, 29 (2022), e2204529119. https://doi.org/10.1073/pnas.2204529119 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2204529119

[54] Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. 2018. Position Bias Estimation for Unbiased Learning to Rank in Personal Search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 610–618.

[55] Haolun Wu, Bhaskar Mitra, Chen Ma, Fernando Diaz, and Xue Liu. 2022. Joint Multisided Exposure Fairness for Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 703–714. https://doi.org/10.1145/3477495.3532007

[56] Himank Yadav, Zhengxiao Du, and Thorsten Joachims. 2021. Policy-Gradient Training of Fair and Unbiased Ranking Functions. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 1044–1053. https://doi.org/10.1145/3404835.3462953

[57] Bozena Zdaniuk and John M. Levine. 2001. Group Loyalty: Impact of Members' Identification and Contributions. *Journal of Experimental Social Psychology* 37, 6 (2001), 502–509. https://doi.org/10.1006/jesp.2000.1474

[58] Masrour Zoghi, Tomáš Tunys, Lihong Li, Damien Jose, Junyan Chen, Chun Ming Chin, and Maarten de Rijke. 2016. Click-based Hot Fixes for Underperforming Torso Queries. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 195–204.