Pairwise Fairness in Ranking as a Dissatisfaction Measure

Alessandro Fabris
Max Planck Institute for Security and Privacy
Bochum, Germany
University of Padova
Padova, Italy
fabrisal@dei.unipd.it

Gianmaria Silvello
University of Padova
Padova, Italy
silvello@dei.unipd.it

Gian Antonio Susto
University of Padova
Padova, Italy
sustogia@dei.unipd.it

Asia J. Biega
Max Planck Institute for Security and Privacy
Bochum, Germany
asia.biega@mpi-sp.org

ABSTRACT

Fairness and equity have become central to ranking problems in information access systems, such as search engines, recommender systems, or marketplaces. To date, several types of fair ranking measures have been proposed, including diversity, exposure, and pairwise fairness measures. Out of those, pairwise fairness is a family of metrics whose normative grounding has not been clearly explicated, leading to uncertainty with respect to the construct that is being measured and how it relates to stakeholders' desiderata.

In this paper, we develop a normative and behavioral grounding for pairwise fairness in ranking. Leveraging measurement theory and user browsing models, we derive an interpretation of pairwise fairness centered on the construct of producer dissatisfaction, tying pairwise fairness to perceptions of ranking quality. Highlighting the key limitations of prior pairwise measures, we introduce a set of reformulations that allow us to capture behavioral and practical aspects of ranking systems. These reformulations form the basis for a novel pairwise metric of producer dissatisfaction. Our analytical and empirical study demonstrates the relationship between dissatisfaction, pairwise, and exposure-based fairness metrics, enabling informed adoption of the measures.

CCS CONCEPTS

• **Information systems**;

KEYWORDS

algorithmic fairness, fair ranking, paiwise fairness

ACM Reference Format:
Alessandro Fabris, Gianmaria Silvello, Gian Antonio Susto, and Asia J. Biega. 2023. Pairwise Fairness in Ranking as a Dissatisfaction Measure. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23), February 27-March 3, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3539597.3570459

WSDM '23, February 27-March 3, 2023, Singapore, Singapore

ACM ISBN 978-1-4503-9407-9/23/02…$15.00
https://doi.org/10.1145/3539597.3570459

1 INTRODUCTION

Information Access Systems (IAS) facilitate user interactions with content by ranking and presenting items to their users according to their estimated merit or relevance [2, 29]. *Content producers* in IAS are increasingly recognized as stakeholders whose economic and societal needs must be taken into account, along with those of *consumers*, to foster a fruitful and equitable information ecosystem [19, 37, 45, 46]. Their needs can be considered individually [8, 10, 16] or based on group membership [6, 39, 40] determined by sensitive attributes such as gender or race. To this end, several measures of *fairness* in ranking have been proposed, capturing notions of equity of *exposure* [17, 40], *representation* [1, 42], or *pairwise accuracy* [5, 32].

When considering a measure, it is important to distinguish between its construct, that is, the theoretical property captured by the measure (e.g. fame), and its operationalization, that is, the mathematical formulation adopted to capture this property (e.g. number of followers) [26]. In this regard, exposure- and representation-based measures in fair ranking operationalize well-defined constructs, clearly connected to the desiderata of producers. They measure the presence of salient groups of providers in the most visible positions of a ranking, increasing their chance of being viewed by IAS users and consequently gain benefits, such as clicks, purchases, or downloads. In contrast, in the prior literature, pairwise fairness has not been clearly associated with a quantity of practical interest for producers [5, 32, 36]. In a nutshell, measures of pairwise fairness quantify how often the rank of two items from different groups reflects their merit and whether mismatched pairs are systematically in favor of one group. This notion of equity is less clearly connected with immediate producer benefits and thus deserves further scrutiny.

In this paper, we perform an in-depth study of pairwise fairness. First, we provide an interpretation of pairwise fairness grounded in browsing models [12], developing a rigorous distinction between the construct and its operationalization [26]. We show that pairwise fairness can capture perceived unfairness on part of item producers, and thus operationalize their *dissatisfaction* with the output of an IAS. Second, we highlight several limitations of existing pairwise fairness metrics, deriving a novel metric that overcomes the issues.

Our measure improves on previous proposals by modeling realistic browsing behaviors, individual user perspectives, and relevance ties. It captures key aspects of observed unfairness and dissatisfaction, connected with perceived quality of IAS by the producers, which is one of the central concerns for platform owners. Finally, we characterize the relationship between pairwise and exposure-based measures analytically and empirically. We show their key similarities, inherited from browsing models, and highlight their differences arising from the underlying normative constructs.

Overall, we make the following salient contributions:

- Interpretation of pairwise fairness, centered on producer dissatisfaction with IAS (§ 3).
- New measure of pairwise fairness (§ 5), overcoming the limitations inherent in the most popular measures (§ 4).
- An analytical (§ 5.2) and empirical (§ 6) study of the relationship between pairwise fairness and exposure-based fairness.

2 BACKGROUND AND RELATED WORK

Fair ranking is concerned with accurately ordering items without unjust discrimination. Fairness interventions are developed for bias mitigation [9, 13], equity [8, 40], and diversity [33, 41], and are technically challenging due to the existence of multiple protected groups [20, 44], outliers [39], and duplicate ranking items [16]. Recent surveys and comparative analyses of fair ranking omit measures of pairwise fairness [38, 44] or simply frame them as accuracy-based [19, 37]. A clear discussion of the construct underlying pairwise fairness is lacking in the literature [5, 32, 36], hindering an informed adoption of these measures and understanding of how they relate to item producers and equity towards them.

Notation. Let $\mathcal{I}$ denote a set of items to be ranked, and let i be an item from this set. Let r_i denote the relevance of item i in a given ranking. Moreover, let $g \in \mathcal{G} = \{A, B\}$ denote a (binary, for ease of exposition) sensitive attribute.[1] Let $i \in g$ denote the membership of i in group g. Let σ_* denote an "ideal" ranking, i.e., a permutation which orders items decreasingly by relevance: $\sigma_* = \text{argsort}(r_i)$. Finally, let σ denote a ranking returned by the IAS in response to a query, and $\sigma(k)$ indicate the item ranked by σ in position k.

Discordant pairs. Central to pairwise fairness is the definition of *discordant pair*. Two items $i, j \in \mathcal{I}$ represent a discordant pair if their relative ordering in σ and σ_* differs. More formally, let $\sigma^{-1}(i)$ denote the position of item i in ranking σ, i.e., $\sigma^{-1}(i) = k \iff \sigma(k) = i$. Given two rankings, σ and σ_*, the indicator function for a discordant pair is defined as

$$d(i,j) = \underbrace{\mathbb{1}(\sigma^{-1}(i) < \sigma^{-1}(j), \sigma_*^{-1}(i) > \sigma_*^{-1}(j))}_{d_F(i,j)} + \underbrace{\mathbb{1}(\sigma^{-1}(i) > \sigma^{-1}(j), \sigma_*^{-1}(i) < \sigma_*^{-1}(j))}_{d_U(i,j)}$$

In other words, i can be part of a discordant pair when ranking σ unfairly places it at an advantage (d_F) or a disadvantage (d_U) over another item j; subscripts F and U indicate that the first item is part of a Favorable Discordant Pair (FDP) or an Unfavorable Discordant Pair (UDP).

Pairwise Fairness. Inter-Group Inaccuracy (IGI) [5] and Rank Equality Error (REE) [32], the most popular measures of pairwise fairness, are defined as

$$M_{AB} = \frac{1}{C_{AB}} \cdot \sum_{i \in A} \sum_{j \in B} d_U(i,j). \tag{1}$$

The key difference between IGI and REE is the normalizing constant C_{AB}. We defer a detailed analysis of this aspect to Section 4.4. M_{AB} measures how often items $i \in A$ are in a UDP with items $j \in B$. Conversely, M_{BA} measures the frequency of cross-group UDPs where items from B are disadvantaged. Beutel et al. [5] then define fairness as

$$M_{AB} - M_{BA} = 0, \tag{2}$$

i.e., equality in the frequency of unfavorable discordant pairs between groups. An explicit discussion of the normative reasoning behind this measure and the construct it captures is lacking in the literature. To foster contextualized adoption of pairwise fairness, this paper develops an interpretation of the measured constructs and proposes a new generalized fairness metric.

3 WHAT DOES PAIRWISE FAIRNESS ACTUALLY MEASURE?

Following Jacobs and Wallach [26], we examine fairness measures distinguishing between the construct, i.e., the theoretical property that a measure intends to capture, and the operationalization, i.e., the particular mathematical formulation meant to model that property. Ideally, a fairness measure (operationalization) should be based on an a priori defined clear normative construct, explicitly enunciating what it means for an algorithm to be equitable and from whose perspective. However, fairness measures are often introduced as self-evident prerequisites for equity, resulting in downstream uncertainty as to what exactly is being measured or optimized.

These considerations are especially applicable to measures of pairwise fairness in ranking. For example, REE is based on the "postulate that there is value in considering error-based fairness criteria for rankings" [32]. Similarly, for IGI, Beutel et al. [5] "draw on the intuition of Hardt et al. [23] for equality of odds, where the fairness of a classifier is quantified by comparing its false positive rate and/or false negative rate." Although related to fairness in that they seek to equalize a certain property between groups, according to Equation (2), an explicit exposition of the construct behind these measures is lacking. In this section, we address this gap by analyzing pairwise fairness measures in depth and uncovering their underlying construct(s).

3.1 Implicit browsing models

In this section, we demonstrate and derive the implicit user browsing model in pairwise fairness metrics. Let us begin with an observation that REE and IGI are closely related to Kendall's Tau [30], a rank correlation measure defined as $\tau(\sigma, \sigma_*) = 1 - \frac{2}{C} \cdot \sum_i \sum_{j \neq i} d(i,j)$, with $C = n(n-1)/2$. In essence, computing Kendall's Tau requires enumerating every item pair and counting discordant ones. Following Equation (1), let us define *inaccuracy* as the frequency of

[1] We follow the literature on pairwise fairness and consider binary sensitive attributes. Extensions to settings with more than two groups can be defined in multiple ways starting from individual measures (§4.1) and are left to future work.

discordant pairs in σ and σ_*

$$M = \frac{1}{C} \cdot \sum_i \sum_{j \neq i} d(i, j), \tag{3}$$

from which Kendall's Tau is computed via the linear transformation $\tau = 1 - 2 \cdot M$. For the sake of simplicity, we will temporarily concentrate on Kendall's Tau and its interpretation(s), and subsequently reintroduce the complexity of sensitive attributes to specifically study REE and IGI.

Furthermore, note that item pairs can be enumerated by browsing the ranking σ according to a cascade model [14]. To enumerate every pair, we can browse σ from top ($k = 0$) to bottom ($k = n - 1$) and compare the current item $\sigma(k)$ (the item at rank k in σ) with items further up in the ranking, to determine whether they constitute a discordant pair.

$$M = \frac{1}{C} \cdot \sum_{k=1}^{n-1} \sum_{k'=0}^{k-1} d(\sigma(k), \sigma(k'))$$

With shorthand notation, we write the indicator function for a discordant pair of items ranked by σ at positions (k, k') as $d(k, k') = d(\sigma(k), \sigma(k'))$.

Moreover, let us define a trivial browsing model, according to which users visit the positions in a ranking with uniform (unit) probability across all ranks. More formally, $F(k) = 1 \ \ \forall k$, where $F(k)$ denotes the probability that users will visit the item $\sigma(k)$. With this notation, we can write the following alternative formulas for M:

$$\textbf{Item-centric} \quad M = \frac{1}{C} \cdot \sum_{k=0}^{n-1} \sum_{k'=0}^{k-1} F(k') d_U(k, k') \tag{4}$$

$$\textbf{User-centric:} \quad M = \frac{1}{C} \cdot \sum_{k=0}^{n-1} F(k) \sum_{k'=0}^{k-1} d_U(k, k') \tag{5}$$

In the next section, we show that these alternative formulations capture the perspectives and desiderata of item producers (item-centric) or item consumers (user-centric). They are equivalent under the trivial browsing model defined above, but generally yield different values for M. Both provide a way to count and weigh each pair of items by sequentially traversing a ranking according to a specified browsing model $F(k)$.

3.2 Interpretations

We provide two alternative interpretations of Kendall's Tau based on Equations (4) and (5), before generalizing those interpretations to pairwise fairness metrics.

- **Item-centric:** Producers of items at each rank k evaluate ranking σ by focusing on the most visible cases of unfair treatment against their item $\sigma(k)$. Their dissatisfaction with σ grows each time they encounter a UDP for $\sigma(k)$, that is, an item of lesser relevance ranked better than their own. The inner summation $\sum_{k'=0}^{k-1} F(k') d_U(k, k')$ is a weighted counter of UDPs, with a weight proportional to the visibility of the unjustly favored item. According to this interpretation, Kendall's Tau operationalizes aggregate producer dissatisfaction with σ for unjustly favoring other items.
- **User-centric:** Users browse the ranking σ sequentially, visiting items in rank k with probability $F(k)$. Each time they visit an item $\sigma(k)$, if an item of lower relevance was unduly positioned above it, users add 1 to a counter measuring wasted effort in arriving at the item in position k. According to this interpretation, Kendall's Tau operationalizes user dissatisfaction due to wasted browsing effort.

These interpretations are also applicable to group-based measures of pairwise fairness, such as IGI and REE (Equation 1), with the caveat of focusing on cross-group comparisons. To exemplify, let us focus on the item-centric formulation and consider

$$M_{AB} = \frac{1}{C_{AB}} \cdot \sum_{k=1}^{n-1} \sum_{k'=0}^{k-1} F(k') d_U(k, k') \cdot \mathbb{1}(\sigma(k) \in A, \sigma(k') \in B)$$

Item-centric interpretations for IGI and REE convey the dissatisfaction of items (and producers) from one group for being unjustly ranked worse than items of lesser relevance from a different group. More specifically, suppose that an item in position k' belongs to group B; the producers of group A evaluate whether this item is unjustly ranked above their items despite having lower merit. They contribute to an inter-group dissatisfaction counter, which is weighted according to the probability of a visit at rank k', i.e., to the visibility of the unjustly favored item. In other words, if an item j is unjustly ranked better than another item i, but in a position with low visibility (such as $\sigma^{-1}(j) = 900$ under a top-heavy browsing model), the producer of i is unlikely to notice, while they are more likely to observe the UDP and increase their dissatisfaction if j is very visible. According to this interpretation, M_{AB} represents the dissatisfaction of group A with the ranking σ, due to their items being unjustly ranked below the items of group B (in expectation over the browsing model $F(k)$ and after normalization). Pairwise fairness is thus connected to observed injustice, which can affect the perceived quality of platform service [18, 28], and, in turn, influence the loyalty of item producers [31]

User-centric interpretations, on the other hand, center on wasted effort due to user attention being diverted to items of lower interest from a different group. Users visit an item with probability $F(k)$, taking into account its group (say, $g = A$). They evaluate how much effort they wasted to reach this item because they examined items of inferior relevance from different groups. According to this interpretation, the counter measures wasted effort to reach items in group A that are unduly ranked below items in group B, and M_{AB} represents a normalized expectation of cross-group wasted effort over the browsing model.

4 TOWARD A DISSATISFACTION MEASURE

The fact that multiple interpretations are possible speaks to the flexibility of pairwise measures in operationalizing multiple constructs. Yet, both IGI and REE exhibit certain limitations when it comes to capturing phenomena that occur in ranking systems and user behavior in practice. In this section, we describe these limitations and propose new formulations of pairwise fairness metrics that address them.

4.1 Individual pairwise fairness

Limitation. Just like aggregate performance measures can obscure poor performance for groups of people [3, 22], group measures can obscure poor performance for individuals. IGI and REE focus on user groups, hiding the potential impact on individuals. This section presents an individual pairwise fairness metric.

New formulation. We define an individual version of pairwise fairness that captures the dissatisfaction of each ranked item (or implicitly, the item's producer):

$$M_i = \sum_{j=0}^{n-1} d_U(i,j) \tag{6}$$

Moreover, we may model the case where producers are especially alert about discordant ranking with items from a certain group (for example, marketplaces can selectively favor items based on brand ownership [18, 28], making this attribute particularly salient [15]):

$$M_{iA} = \sum_{j=0}^{n-1} d_U(i,j) \cdot \mathbb{1}(j \in A); \quad M_{iB} = \sum_{j=0}^{n-1} d_U(i,j) \cdot \mathbb{1}(j \in B).$$

Note that the group fairness metrics defined in Eq. (1) can be derived from these group-envy versions of individual fairness metrics as follows:

$$M_{AB} = \frac{1}{C_{AB}} \sum_{i \in A} M_{iB}; \quad M_{BA} = \frac{1}{C_{BA}} \sum_{i \in B} M_{iA}$$

This property provides an intuitive connection between individual and group perspectives, and guarantees that interventions at the individual level, making M_i smaller $\forall i \in \{0, \dots, n-1\}$, will also be beneficial at the group level for metrics M_{AB} and M_{BA}.

4.2 Top-heaviness

Limitation. Existing pairwise fairness metrics do not account for realistic browsing behaviors. As we have shown in Section 3.1, REE and IGI implicitly use a simple browsing model with a uniform visit probability for all ranks. Yet, UDPs at the top of a ranking in practice would be more visible (top ranking positions are more likely to be visited by searchers) and thus cause greater dissatisfaction.

New formulation. Pairwise fairness measures can be flexibly modified, both at the individual and group levels, to account for a suitable user browsing model $F(k)$:

$$M_i = \sum_{k=0}^{n-1} F(k) d_U(i, \sigma(k))$$

In other words, the dissatisfaction M_i of the item i is a weighted sum of UDPs, with weights proportional to the probability of visiting the item unfairly ranked better than i.

We can also incorporate group membership into the individual pairwise fairness measure defined in Sec. 4.1:

$$M_{iB} = \sum_{k=0}^{n-1} F(k) d_U(i, \sigma(k)) \cdot \mathbb{1}(\sigma(k) \in B)$$

and aggregate it to quantify cross-group dissatisfaction:

$$M_{AB} = \frac{1}{C_{AB}} \sum_{i \in A} \sum_{k=0}^{n-1} F(k) d_U(i, \sigma(k)) \cdot \mathbb{1}(\sigma(k) \in B). \tag{7}$$

Many top-heavy user models have been proposed and studied in the literature, including logarithmic ($F(k) \propto 1/\log(k)$ [27]) and exponential discount ($F(k) \propto \gamma^k$ [35]) models.

4.3 Tie handling

Limitation. Measures of pairwise fairness do not account for ties in relevance scores r_i, a common occurrence in practical applications. In recommender systems, for example, user ratings are often quantized [25], while, in information retrieval, relevance judgements are typically discrete (either binary or graded) [24]. IAS which prioritize a group by frequently breaking ties in its favour are not detected as problematic by either IGI or REE.

New formulation. Recall that $\sigma_* = \text{argsort}(r_i)$. We can rewrite the indicator function for UDPs as:

$$d_U(i,j) = \mathbb{1}(\sigma^{-1}(i) > \sigma^{-1}(j), r_i > r_j)$$

showing that relevance ties are unaccounted for. We propose to generalize the notion of UDP to handle ties as:

$$\begin{aligned} d_U(i,j) =& \mathbb{1}(\sigma^{-1}(i) > \sigma^{-1}(j), r_i > r_j) \\ &+ c_t \mathbb{1}(\sigma^{-1}(i) > \sigma^{-1}(j), r_i = r_j) \end{aligned} \tag{8}$$

where c_t models the dissatisfaction of an item ranked below another item of the same relevance. We call this case a *partial UDP*. Possible values for c_t range in $(0, 1)$, where $c_t = 1$ corresponds to equating partial UDPs to proper UDPs, while $c_t = 0$ indicates indifference to comparisons with items of the same relevance.

4.4 Normalization

Limitation. Recall that IGI and REE can be written as:

$$M_{AB}^{\text{IGI,REE}} = \frac{1}{C_{AB}^{\text{IGI,REE}}} \cdot \sum_{i \in A} \sum_{j \in B} d_U(i,j)$$

with different normalizing constants:

$$C_{AB}^{\text{IGI}} = \sum_{i \in A} \sum_{j \in B} \mathbb{1}(r_i > r_j); \quad C_{AB}^{\text{REE}} = N_A \cdot N_B, \tag{9}$$

where N_A and N_B denote the number of items in $\mathcal{I}$ that belong to group A and B, respectively. In other words, IGI is normalized with respect to a worst-case scenario which takes into account the ground truth relevance r_i and its distribution between groups, while REE is normalized with respect to the *a-priori* worst case which does not take r_i into account. As a result, the normalizing constant in REE is the same for M_{AB} and M_{BA} ($C_{AB}^{\text{REE}} = C_{BA}^{\text{REE}}$), while for IGI they typically differ ($C_{AB}^{\text{IGI}} \neq C_{BA}^{\text{IGI}}$).

The normalization scheme for IGI has a downside—it becomes unclear how to compare M_{AB}^{IGI} and M_{BA}^{IGI}. Let us visualize this issue with a toy example where $F(k) = 1, \forall k$, and the ideal ranking is $\sigma_* = [i_0^A, i_1^B, i_2^A, i_3^A]$; here, for ease of exposition, superscript g in i^g denotes membership of i in group g. In this situation, we have different constants for IGI ($C_{AB}^{\text{IGI}} = 1$, $C_{BA}^{\text{IGI}} = 2$) and equal constants for REE ($C_{AB}^{\text{REE}} = C_{BA}^{\text{REE}} = 3$). A ranking $\sigma = [i_2^A, i_1^B, i_0^A, i_3^A]$, obtained by exchanging i_0^A and i_2^A in σ_*, produces two UDPs, one (i_0^A, i_1^B) in favor of group B and another (i_1^B, i_2^A) in favor of group A. The resulting measures for IGI are $M_{AB}^{\text{IGI}} = 1 \gg M_{BA}^{\text{IGI}} = 0.5$. Taken at face value, this suggests that group B is largely favored over group A, and that the latter should be more dissatisfied with σ than

the former. We argue that this is not necessarily true since, from a groupwise perspective, σ and σ_* are equivalent. In fact, under IGI, comparing M_{AB}^{IGI} and M_{BA}^{IGI} is not straightforward. This is a very practical problem, since fairness, according to Equation (2), is defined precisely as the difference between these quantities.

New formulation. We propose an REE inspired normalization scheme, using the same constant for M_{AB} and M_{BA}, independently of the relevance scores. In Equation (7). we define:

$$C_{AB} = C_{BA} = \max\left(N_A \cdot \sum_{k=0}^{N_B-1} F(k), N_B \cdot \sum_{k=0}^{N_A-1} F(k)\right) \tag{10}$$

Inside the $\max(\cdot)$ function, the first term represents a worst-case scenario in which every item in group B is unduly ranked above every item in group A (hence the multiplying factor N_A) and occupies the most visible ranking positions (hence the summation). Analogously, the second term represents the case where every item in group A is unduly ranked above every item in group B.

This formulation has two desirable properties: (1) the difference $M_{AB} - M_{BA}$ (the unfairness measure) is bounded between $(-1, 1)$ and (2) the sign of the measure identifies the (dis)advantaged group, since positive (negative) values correspond to rankings σ with more visible UDPs against group A (B).

5 DISSATISFACTION INDUCED BY PAIRWISE SWAPS

Based on the proposed reformulations, we define a pairwise fairness measure termed Dissatisfaction Induced by Pairwise Swaps (DIPS):

$$M_{AB}^{\text{DIPS}} = \frac{1}{C_{AB}^{\text{DIPS}}} \sum_{i=0}^{n-1} \sum_{k=0}^{n-1} F(k) d_U(i, \sigma(k)) \cdot \mathbb{1}(i \in A, \sigma(k) \in B), \tag{11}$$

DIPS (i) handles ties through parameter c_t in the definition of $d_U(\cdot)$ in Equation (8), (ii) is normalized with a group-symmetric constant according to Equation (10), and (iii) inherits a top-heavy behavior from a suitable browsing model $F(k)$. Browsing models capture the fact that dissatisfaction is more likely to occur when unfair swaps happen at highly exposed ranking positions. The tunable parameters for DIPS are the browsing model $F(k)$ and the tie-handling constant c_t. For the latter, we recommend an intermediate value $c_t = 0.5$, while the former depends on the application and should be tuned to context-specific browsing behaviour.

Exposure-based measures are a popular family of fairness metrics typically also grounded in browsing models [7, 8, 40]. In the remainder of this section, we study the relationship between DIPS and exposure-based fairness.

5.1 Review of exposure-based fairness

Exposure-based measures, in their groupwise version, define an ideal target exposure (T_A, T_B) for each group and measure the distance between this target and actual exposure (E_A, E_B) in ranking σ. We define the normalized misallocation vector as:

$$\delta^\sigma = [\delta_A^\sigma, \delta_B^\sigma] = \left[\frac{T_A}{T_A + T_B} - \frac{E_A}{E_A + E_B}, \frac{T_B}{T_A + T_B} - \frac{E_B}{E_A + E_B}\right] \tag{12}$$

where, for a given group g, E_g is the sum of individual exposure values granted by σ to items in group g: $E_g = \sum_{i \in g} F(\sigma^{-1}(i))$. To compute the overall unfairness of ranking σ, we follow Biega et al. [8], and report the ℓ_1 norm of δ^σ. We consider three measures that differ in their normative reasoning for establishing the target exposure quotas.

Equity of Attention. According to Equity of Attention (EA) [8], the target exposure for a group g should be proportional to the sum of the relevance of the items in g:

$$T_g^{\text{EA}} = \sum_{i \in g} r_i \tag{13}$$

Under a different normative reasoning, we can define a version of EA inspired by demographic parity [4, 11], which requires that each group receives a share of attention that is proportional to the group's representation in the overall population:

$$T_g^{\text{EA-dp}} = N_g / N. \tag{14}$$

Expected Exposure. Expected Exposure (EE) [17] also relies on relevance scores to specify its target exposure; however, unlike EA, it assigns *ordinal* validity to relevance judgements: if item i is more relevant than (or as relevant as) j, it should get more (or as much) exposure. This property should be contrasted with EA, which assigns a *scale ratio* validity to relevance judgements: if item i is twice as relevant as j, it should get twice as much exposure. The amount of exposure in EE is not explicitly specified by the normative reasoning and is determined by the browsing model $F(k)$ in practice. Numerically, the target exposure in EE can be expressed as:

$$t_i = \text{mean}_{\{j | r_j = r_i\}}(F(\sigma_*^{-1}(j))) \tag{15}$$

$$T_g^{\text{EE}} = \sum_{i \in g} t_i \tag{16}$$

where t_i is the exposure target quota for item i. In a simple setting without relevance ties, t_i is equal to the exposure granted to i by the ideal ranking σ_* under $F(k)$. If ties are present, t_i is the average exposure granted by σ_* to items of the same relevance as i.

5.2 DIPS and exposure-based fairness

According to exposure-based measures, individual misallocation is the difference between the target exposure quota of an item and its actual exposure $F(\sigma^{-1}(i))$, i.e., its probability of a visit by a searcher given ranking σ. For example, EA defines the target quota of an item as its share of overall relevance $c_i = r_i / \sum_{i'} r_{i'}$. Under EA, individual misallocation M_i can be written as:

$$\begin{aligned} M_i^{\text{EA}} &= c_i \sum_{i'=0}^{n-1} F(\sigma^{-1}(i')) - F(\sigma^{-1}(i)) \\ &= \sum_{k=0}^{n-1} p_s(\sigma(k)) \left[c_i(k+1) - \Pr(\sigma^{-1}(i) \le k)\right], \end{aligned}$$

where $p_s(\sigma(k))$ denotes the probability of a user stopping browsing at position k, and $F(k)$ is the resulting probability of a visit.[2]

[2]Under cascade (sequential) browsing models, the probability of receiving a visit at rank k is equal to the sum of the probability of stopping at any rank greater than or equal to k [12]: $F(k) = \sum_{k'=k}^{n-1} p_s(\sigma(k'))$.

Moreover, recall that DIPS at the item level can be expressed as:

$$M_i^{\text{DIPS}} = \sum_{k=0}^{n-1} F(k) d_U(i, \sigma(k)) = \sum_{k=0}^{n-1} p_s(\sigma(k)) \sum_{k'=0}^{k} d_U(i, \sigma(k'))$$

These formulas show that EA and DIPS can both be expressed as a sum, weighted by stopping probabilities $p_s(\sigma(k))$, of two quantities that are directly related: DIPS counts the number of UDPs for the item i up to rank k, while EA computes the (negative) probability $\Pr(\sigma^{-1}(i) \leq k)$ that item i is among the top k. One can expect the probability of an item being in the top ranks to decrease with the number of its UDPs. For this reason, we expect DIPS and exposure-based measures to exhibit certain similarities in practice. At the same time, these measures operationalize different constructs; hence, we expect them to capture different properties of rankings. For example, a ranking can assign to an item i its ideal exposure quota ($M_i^{EA} = 0$), while granting the most visible positions to items of lesser relevance, thus causing substantial dissatisfaction of i due to highly visible UDPs ($M_i^{DIPS} \gg 0$).

6 EXPERIMENTS

DIPS is a measure of pairwise fairness, yet it is grounded in browsing models like exposure-based fairness. In this section, we empirically study the similarities and differences between DIPS, pairwise measures, and exposure-based measures on synthetic and real-world datasets.

6.1 Synthetic data

To compare fair ranking measures in a principled fashion, we build a synthetic dataset with full control on group representation, merit, and ranking policies. We consider a controlled setting with a binary sensitive attribute $g \in \{A, B\}$, where groups have equal representation over a total of $N = 1,000$ items, and with sizeable differences in relevance scores. More specifically, we set $N_A = N_B = 500$, and draw relevance scores from group-specific, uniform distributions $f_A(r_i) = \text{unif}(0.5, 1)$ and $f_B(r_i) = \text{unif}(0.2, 0.7)$. In other words, all items of high relevance ($0.7 < r_i \leq 1$) belong to group A, items of intermediate relevance ($0.5 \leq r_i \leq 0.7$) belong to both groups with the same probability, and low relevance items ($0 \leq r_i < 0.5$) are entirely from group B. The distribution of relevance scores between groups is depicted in panel (1) of Figure 1a. We choose the browsing model underlying rank biased precision [35], modeling a top-heavy probability of visit with exponential decay: $F(k) = \gamma^k, \quad \gamma = 0.9$.

6.1.1 *Experimental condition 1: systematic group advantage.*

Setup. To be able to compare metrics under controlled unfairness conditions, we create a rank promotion mechanism that allows us to control the amount of unfairness relative to a ranking purely based on relevance. In this experiment, the mechanism advances the 20 most relevant items from group B. We vary the top destination rank k for the promoted items, with k in $(0, 99)$. For example, setting $k = 0$, we promote the 20 most relevant items from $g = B$ to the ranks $\{0, 1, \ldots, 19\}$, while the relative positions of the remaining items remain unchanged, i.e., their rank increases according to $\sigma^{-1}(i) = \sigma_*^{-1}(i) + 20$.

Results. The results of this experiment are reported in Figure 1a. The values of $M_{AB} - M_{BA}$ (Equation 2) for REE and DIPS are shown in panel (2). No promotion takes place in favour of $g = A$, hence $M_{BA} = 0$. DIPS is very sensitive to the promotion rank of items in group B, showing an exponential decay, while REE is mostly flat. Furthermore, the value $M_{AB}^{\text{DIPS}} > 0.5$ for $k = 0$ captures a strong dissatisfaction, while $M_{AB}^{\text{REE}} \ll 0.1$ is much smaller in comparison.

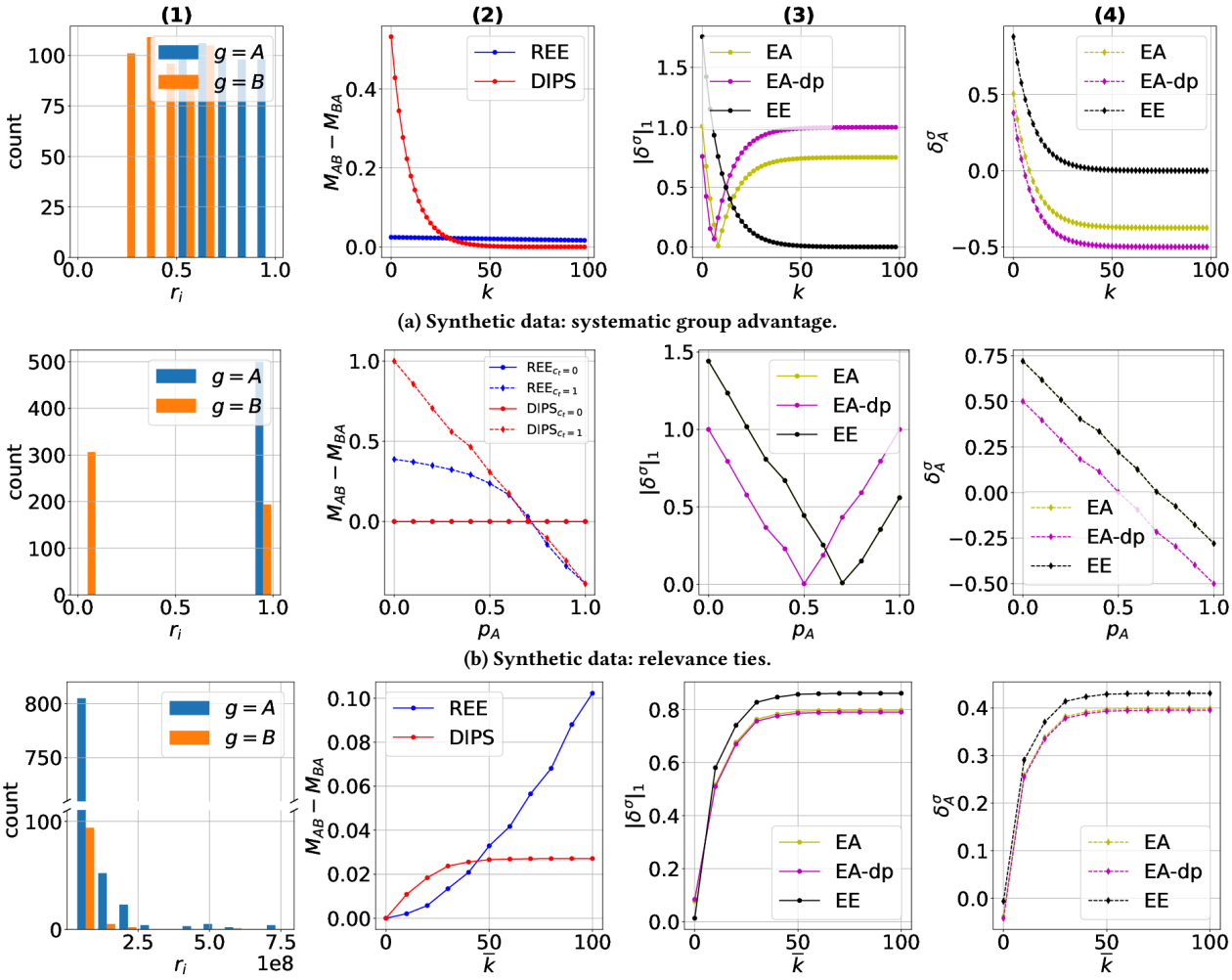
The remaining panels concentrate on three exposure-based measures (EE, EA, EA-dp). Panel (3) of Figure 1a reports the aggregate measure $|\delta^\sigma|_1$, i.e., the ℓ_1 norm of the misallocation vector in Equation (2), while Panel (4) reports the groupwise measure δ_A^σ for group A, i.e., the first component of Equation (2). The groupwise misallocation in panel (4) clearly shows a monotonic trend with exponential decay, as expected from the browsing model $F(k)$. It is worth recalling that positive values indicate underexposure for group A. Promoting items from group B to the most visible positions reduces the exposure E_A available for group A, and therefore δ_A^σ increases as items from group B are promoted to better positions, corresponding to lower values of k on the x axis. It should be noted that the aggregate measure ($|\delta^\sigma|_1$) in panel (3) derives directly from the groupwise measure δ_A^σ in panel (4). In the binary case considered in this example, it is equal to twice its absolute value, since $|\delta^\sigma|_1 = 2 \cdot \text{abs}(\delta_A^\sigma)$.

Interpretation. The large value $M_{AB}^{\text{DIPS}} > 0.5$ for $k = 0$ captures the strong dissatisfaction that is likely to arise in group A if many items in another group were unfairly promoted to the top ranks—unfairly in the sense that they do not reflect the merit reflected in r_i and σ_*. A large value for M_{AB}^{DIPS} adequately summarizes a situation where items in group A are highly dissatisfied, as the promoted items form visible UDPs with most items from group A. The same is not true for $M_{AB}^{\text{REE}} \ll 0.1$, suggesting that, under the (implicit) normative reasoning of REE, the dissatisfaction of group A would be very far from its theoretical maximum.

Turning to exposure-based measures, the disaggregated measures δ_A^σ, depicted in panel (4), are equal up to a constant, which depends on the differences in their normative reasoning presented in Section 5.1. Moreover, these measures have the same profile as DIPS in the left panel. As discussed in Section 5.2, UDPs (in the absence of FDPs) directly result in missed exposure and higher values of EA, EA-dp, and EE. Since the same top-heavy browsing model $F(k)$ is assumed across these measures, they end up having a similar profile with exponential decay. Hence, if item producers have a notion of merit r_i, any intervention that assigns exposure to a group beyond its merit, as encoded by r_i, may generate a proportional amount of dissatisfaction in the remaining groups.

6.1.2 *Experimental condition 2: relevance ties.*

Setup. Relevance ties are common in ranking problems and datasets [24, 25, 34, 47]. To study the behavior of DIPS and related measures in the presence of ties, we round relevance scores in the synthetic dataset to the nearest integer, leaving us with binary values $r_i^q = \text{round}(r_i)$, depicted in panel (1) of Figure 1b. We consider rankings of maximum utility $\sigma = \text{argsort}(r_i^q)$ where we vary the tie breaking policy. At each position of the ranking σ, a policy places the item of maximum relevance among those that have not already been placed



(a) Synthetic data: systematic group advantage.

(b) Synthetic data: relevance ties.

(c) Real-world data: fairness intervention to ensure minimum representation.

Figure 1: Distribution of relevance r_i (1) and comparison of pairwise fairness measures REE and DIPS (2) with exposure-based measures EE, EA, EA-dp: $|\delta^\sigma|_1$ (3) and δ^σ_A (4).

in better positions; if items of the same relevance are available from both groups, we draw the best available item from $g = A$ with probability $p_A \in \{0, 0.1, \ldots, 1\}$, or from $g = B$ with probability $p_B = 1 - p_A$. We consider a tie-aware and a tie-indifferent variant of REE and DIPS, obtained by setting $c_t = 1$ and $c_t = 0$, respectively, in Equation (8).

Results. Figure 1b shows the values for each measure, averaged over 100 repetitions. Panel (2) shows both versions of REE and DIPS. As expected, the tie-indifferent variant of both measures is flat at zero. Indeed, $\sigma = \text{argsort}(r_i^q)$ is a meritocratic ranking; therefore, there are no proper UDPs. For $c_t = 1$, both DIPS and REE span a wide range of values, capturing the large dissatisfaction between groups that is likely to arise in this setting with ranking policies that systematically favor one group over another in case of ties.

EE, EA, and EA-dp are represented in panels (3)-(4), with their aggregate ($|\delta_g|_1$) and groupwise component (δ^σ_A), respectively. The difference between EE and EA is negligible in this setting and they are therefore indistinguishable in the plots. Overall, EA and EE have the same profile as the tie-aware version of DIPS.

Interpretation. Measures of pairwise fairness can aptly model dissatisfaction in contexts where relevance ties are present, a situation that is fairly common in ranking problems. This is achieved by extending the concept of UDP to account for relevance ties. If, instead, we stick to the regular definition of UDP, any systematic advantage for one group will go unnoticed, as testified by the (constant and null) values of REE and DIPS instantiated with $c_t = 0$. Furthermore, this experiment confirms a close connection between DIPS and exposure-based measures.

6.2 Real-world data

In this section, we complement our discussion of similarities and differences between pairwise and exposure-based fairness measures by experimenting with a real-world dataset and a popular fair ranking intervention. We use the Entrepreneurs dataset [21], which

consists of a list of US startup founders who received Series A funding in the last 5 years, obtained from Crunchbase.[3] Entrepreneurs are ranked by inflation-adjusted funding, which is considered the merit parameter r_i, reported in panel (1) of Figure 1c. The sensitive attribute is binary gender, with a representation ratio of 9:1 in favor of men (group A). Notice that the y axis is broken, to highlight the prevalence of items of low relevance from the group A, while favoring readability at higher values of r_i.

6.2.1 *Enforcing minimum representation.*

Setup. We deploy the fairness intervention of Zehlike et al. [43], imposing a minimum representation for the protected group (women). More specifically, we require a minimum percentage $p_{\min} = 0.5$ of women in every prefix of the final ranking, up to a given ranking position $\bar{k}$. In this experiment, we vary $\bar{k} \in \{0, \dots, 100\}$.

As a motivating example for such an intervention, consider a trade magazine that compiles a chart of successful entrepreneurs with attention to gender representation. Relevance and gender representation goals can be achieved with a ranking σ that is aware of the raised funding while featuring a minimum percentage of women in every prefix up to a given rank $\bar{k}$. Low values of $\bar{k}$ correspond to mild gender parity requirements, enforced only at the top positions of the ranking (up to $\bar{k}$). On the other hand, high values of $\bar{k}$ correspond to more strict requirements, where the minimum representation must also be maintained further down the ranking.

Results. The results of this experiment are reported in Figure 1c. Panel (2) focuses on REE and DIPS. The latter increases sharply for small values ($\bar{k} < 20$), where an increased representation corresponds to highly visible UDPs under a top-heavy browsing model. Around rank $\bar{k} = 40$ DIPS becomes flat, as these ranks have low visibility. REE also increases with $\bar{k}$, but, unlike DIPS, the increase accelerates with $\bar{k}$. This is due to the fact that, to satisfy the minimum representation requirement, the number of UPDs increases superlinearly with $\bar{k}$.

EE, EA, and EA-dp are represented in panels (3)-(4). The groupwise measure displays a concave profile, similar to DIPS, since promotions after rank $k = 40$ have a negligible impact on exposure. As usual, EE is minimized by the null manipulation $\bar{k} = 0$; EA and EA-dp are very close to it as, in this particular setting, women entrepreneurs have a low overall representation ($T_B^{\text{EA-dp}} = N_B/N \simeq 0.1$) and, subsequently, a low share of the overall relevance ($T_B^{\text{EA}} = \sum_{i \in B} r_i / \sum_i r_i \simeq 0.1$). The sizeable values of EE, for $\bar{k} \geq 40$, suggest that group B (women) gains a significant exposure from this intervention, clearly at the expense of group A (men).[4]

DIPS, on the other hand, has low values $|M_{AB}^{\text{DIPS}} - M_{BA}^{\text{DIPS}}| \ll 0.1$. This is due to the fact that the women entrepreneurs occupying these highly visible positions in the final ranking σ have greater relevance (r_i) than most of the other entrepreneurs. In other words, despite a substantial visibility gain for female entrepreneurs, the most visible positions occupied by them do not represent a UDP for most male entrepreneurs. For example, when $\bar{k} \geq 20$, among the twenty most visible positions, accounting for more than 80% of overall exposure, we find ten female entrepreneurs who are in the top decile for raised funding overall. This follows from the fact that the fairness manipulation used is aware of relevance, so women entrepreneurs with higher r_i are promoted first. Different ranking policies, naïvely enforcing representation without paying attention to relevance, would yield high values of DIPS.

Interpretation. On the one hand, this experiment shows that, when DIPS and exposure-based measures are instantiated with the same top-heavy browsing model $F(k)$, they are similarly influenced by fairness interventions toward the top of a ranking, while ignoring swaps at less visible positions; they display similar profiles as a result. On the other hand, the absolute values of these measures can differ substantially. In essence, exposure-based measures are based on a comparison between *groupwise* merit and groupwise representation among the most visible items in the final ranking. Although DIPS is focused similarly on the most visible items, it takes into account their *individual* merits. For instance, an item whose relevance is in the highest decile can be promoted to the most visibile position, i.e, with a sizeable impact on exposure, without increasing the dissatisfaction counter of most items, i.e., with a small impact on the aggregate DIPS measure. While showing some clear similarities, DIPS and exposure-based measures operationalize different constructs and capture different properties. Overall, our analyses show that fairness-enhancing interventions in ranking may cause dissatisfaction for non-protected groups, but merit-based policies will mitigate this downside.

7 CONCLUSION

In this paper, we have provided a normative grounding for pairwise fairness measures (Inter-Group Inaccuracy (IGI) [5] and Rank Equality Error (REE) [32]), retrospectively mapping the measured construct to producer *dissatisfaction* induced by a non-meritocratic ranking, which is related to, yet different from, the construct of equitable exposure allocation.

We have highlighted the limitations of REE and IGI in capturing behavioral and practical aspects of rankings in information access systems, deriving a new measure called *Dissatisfaction Induced by Pairwise Swaps (DIPS)* to address them. DIPS operationalizes perceptions of injustice by ranked producers when they are positioned below less relevant items from other groups.

Finally, we have studied the relationship between DIPS, pairwise, and exposure-based fairness measures, including Equity of Attention and Expected Exposure. We have shown how to ground pairwise fairness in browsing models, highlighting the similarities between exposure-based measures and DIPS. At the same time, we have stressed the differences between the two families of measures which arise as they operationalize fundamentally different constructs.

Overall, this work grounds and generalizes measures of pairwise fairness, situates them more precisely in the practical context of information access systems, and contributes to the debate on the normative reasoning behind algorithmic fairness measures.

ACKNOWLEDGMENTS

The work of Gianmaria Silvello was supported by the ExaMode project, as part of the EU H2020 program under Grant Agreement no. 825292.

[3] https://crunchbase.com/

[4] Recall that δ_A^σ is a normalized quantity, i.e., $0 \leq \delta_A^\sigma \leq 1$

REFERENCES

[1] Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das. 2019. Designing Fair Ranking Schemes. In *Proc. of the 2019 International Conference on Management of Data* (Amsterdam, Netherlands) *(SIGMOD '19)*. ACM, 1259–1276. https://doi.org/10.1145/3299869.3300079

[2] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*. Vol. 463. ACM press New York.

[3] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W. Duncan Wadsworth, and Hanna Wallach. 2021. *Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs*. ACM, 368–378. https://doi.org/10.1145/3461702.3462610

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[5] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *Proc. of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. ACM, 2212–2220.

[6] Asia J Biega, Fernando Diaz, Michael D Ekstrand, Sergey Feldman, and Sebastian Kohlmeier. 2021. Overview of the TREC 2020 Fair Ranking Track. *arXiv preprint arXiv:2108.05135* (2021).

[7] Asia J Biega, Fernando Diaz, Michael D Ekstrand, and Sebastian Kohlmeier. 2020. Overview of the TREC 2019 fair ranking track. *arXiv preprint arXiv:2003.11650* (2020).

[8] 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* (Ann Arbor, MI, USA) *(SIGIR '18)*. ACM, 405–414.

[9] Ludovico Boratto, Gianni Fenu, and Mirko Marras. 2021. Connecting user and item perspectives in popularity debiasing for collaborative recommendation. *Information Processing & Management* 58, 1 (2021), 102387.

[10] Amanda Bower, Hamid Eftekhari, Mikhail Yurochkin, and Yuekai Sun. 2021. Individually Fair Ranking. *arXiv preprint arXiv:2103.11023* (2021).

[11] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data mining and knowledge discovery* 21, 2 (2010), 277–292.

[12] Ben Carterette. 2011. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In *Proc. of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11)*. ACM, 903–912. https://doi.org/10.1145/2009916.2010037

[13] L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. 2020. Interventions for Ranking in the Presence of Implicit Bias. In *Proc. of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. ACM.

[14] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An Experimental Comparison of Click Position-Bias Models. In *Proc. of the 2008 International Conference on Web Search and Data Mining* (Palo Alto, California, USA) *(WSDM '08)*. ACM, 87–94. https://doi.org/10.1145/1341531.1341545

[15] Abhisek Dash, Abhijnan Chakraborty, Saptarshi Ghosh, Animesh Mukherjee, and Krishna P. Gummadi. 2021. When the Umpire is Also a Player: Bias in Private Label Product Recommendations on E-Commerce Marketplaces. In *Proc. of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. ACM, 873–884. https://doi.org/10.1145/3442188.3445944

[16] Giorgio Maria Di Nunzio, Alessandro Fabris, Gianmaria Silvello, and Gian Antonio Susto. 2021. Incentives for Item Duplication Under Fair Ranking Policies. In *Advances in Bias and Fairness in Information Retrieval*, Ludovico Boratto, Stefano Faralli, Mirko Marras, and Giovanni Stilo (Eds.). Springer International Publishing, Cham, 64–77.

[17] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. In *Proc. of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. ACM, 275–284. https://doi.org/10.1145/3340531.3411962

[18] Renee Dudley. 2020. Amazon's New Competitive Advantage: Putting Its Own Products First. https://www.propublica.org/article/amazons-new-competitive-advantage-putting-its-own-products-first.

[19] Michael D Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2021. Fairness and discrimination in information access systems. *arXiv preprint arXiv:2105.05779* (2021).

[20] 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*.

[21] Avijit Ghosh, Ritam Dutt, and Christo Wilson. 2021. *When Fair Ranking Meets Uncertain Inference*. ACM, 1033–1043. https://doi.org/10.1145/3404835.3462850

[22] Sruthi Gorantla, Amit Deshpande, and Anand Louis. 2021. On the Problem of Underranking in Group-Fair Ranking. In *Proc. of the 38th International Conference on Machine Learning (Proc. of Machine Learning Research, Vol. 139)*. PMLR, 3777–3787.

[23] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proc. of the 29th Annual Conference on Neural Information Processing Systems (NIPS 2016)*. Barcelona, ES, 3323–3331.

[24] Donna Harman. 1992. The DARPA TIPSTER Project. *SIGIR Forum* 26, 2 (1992), 26–28.

[25] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (Dec. 2015), 19 pages. https://doi.org/10.1145/2827872

[26] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proc. of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. ACM, 375–385. https://doi.org/10.1145/3442188.3445901

[27] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (oct 2002), 422–446. https://doi.org/10.1145/582415.582418

[28] Adrianne Jeffries and Leon Yin. 2021. Amazon puts its own "brands" above better rated products. https://themarkup.org/amazons-advantage/2021/10/14/amazon-puts-its-own-brands-first-above-better-rated-products.

[29] Thorsten Joachims. 2002. Optimizing Search Engines Using Clickthrough Data. In *Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Edmonton, Alberta, Canada) *(KDD '02)*. ACM, 133–142. https://doi.org/10.1145/775047.775067

[30] Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1/2 (1938), 81–93.

[31] Junic Kim. 2021. Platform quality factors influencing content providers' loyalty. *Journal of Retailing and Consumer Services* 60 (2021), 102510. https://doi.org/10.1016/j.jretconser.2021.102510

[32] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. FARE: Diagnostics for Fair Ranking Using Pairwise Error Metrics. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. ACM, 2936–2942. https://doi.org/10.1145/3308558.3313443

[33] Graham McDonald, Craig Macdonald, and Iadh Ounis. 2022. Search results diversification for effective fair ranking in academic search. *Information Retrieval Journal* 25, 1 (2022), 1–26.

[34] Frank McSherry and Marc Najork. 2008. Computing Information Retrieval Performance Measures Efficiently in the Presence of Tied Scores. In *Advances in Information Retrieval*. Springer Berlin Heidelberg, Berlin, Heidelberg, 414–421.

[35] Alistair Moffat and Justin Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* 27, 1, Article 2 (dec 2008), 27 pages. https://doi.org/10.1145/1416950.1416952

[36] Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Serena Wang. 2020. *Proc. of the AAAI Conference on Artificial Intelligence* 34, 04 (Apr. 2020), 5248–5255. https://doi.org/10.1609/aaai.v34i04.5970

[37] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2021. Fairness in rankings and recommendations: an overview. *The VLDB Journal* (2021), 1–28.

[38] Amifa Raj and Michael D Ekstrand. 2022. Measuring Fairness in Ranked Results: An Analytical and Empirical Comparison. In *Proc. of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[39] Fatemeh Sarvi, Maria Heuss, Mohammad Aliannejadi, Sebastian Schelter, and Maarten de Rijke. 2022. Understanding and Mitigating the Effect of Outliers in Fair Ranking. In *Proc. of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM '22)*. ACM, 861–869.

[40] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. ACM, 2219–2228.

[41] Julia Stoyanovich, Ke Yang, and HV Jagadish. 2018. Online set selection with fairness and diversity constraints. In *Proc. of the EDBT Conference*.

[42] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proc. of the 29th International Conference on Scientific and Statistical Database Management* (Chicago, IL, USA) *(SSDBM '17)*. ACM, Article 22, 6 pages. https://doi.org/10.1145/3085504.3085526

[43] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proc. of the 2017 ACM on Conference on Information and Knowledge Management*. 1569–1578.

[44] Meike Zehlike, Tom Sühr, Ricardo Baeza-Yates, Francesco Bonchi, Carlos Castillo, and Sara Hajian. 2022. Fair Top-k Ranking with multiple protected groups. *Information Processing & Management* 59, 1 (2022), 102707.

[45] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in Ranking, Part I: Score-Based Ranking. *ACM Comput. Surv.* (apr 2022). https://doi.org/10.1145/3533379 Just Accepted.

[46] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in Ranking, Part II: Learning-to-Rank and Recommender Systems. *ACM Comput. Surv.* (apr 2022). https://doi.org/10.1145/3533380 Just Accepted.

[47] Ke Zhou, Gui-Rong Xue, Hongyuan Zha, and Yong Yu. 2008. Learning to Rank with Ties. In *Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Singapore, Singapore) *(SIGIR '08)*. ACM, 275–282. https://doi.org/10.1145/1390334.1390382