Estimation of Fair Ranking Metrics with Incomplete Judgments

Ömer Kırnap
University College London
London, UK
omer.kirnap.18@ucl.ac.uk

Fernando Diaz*
Microsoft Research
Montreal, Quebec, Canada
diazf@acm.org

Asia Biega†
Max Planck Institute
for Security and Privacy
Saarbrücken, Saarland, Germany
asia.biega@acm.com

Michael Ekstrand
People & Information Research Team
Boise State University
Boise, Idaho, USA
michaelekstrand@boisestate.edu

Ben Carterette
Spotify Research
New York, USA
benjaminc@spotify.com

Emine Yılmaz
University College London
London, UK
emine.yilmaz@ucl.ac.uk

ABSTRACT

There is increasing attention to evaluating the *fairness* of search system ranking decisions. These metrics often consider the membership of items to particular groups, often identified using *protected attributes* such as gender or ethnicity. To date, these metrics typically assume the availability and completeness of protected attribute labels of items. However, the protected attributes of individuals are rarely present, limiting the application of fair ranking metrics in large scale systems. In order to address this problem, we propose a sampling strategy and estimation technique for four fair ranking metrics. We formulate a robust and unbiased estimator which can operate even with very limited number of labeled items. We evaluate our approach using both simulated and real world data. Our experimental results demonstrate that our method can estimate this family of fair ranking metrics and provides a robust, reliable alternative to exhaustive or random data annotation.

KEYWORDS

information retrieval, evaluation, fairness, fair ranking

ACM Reference Format:
Ömer Kırnap, Fernando Diaz, Asia Biega, Michael Ekstrand, Ben Carterette, and Emine Yılmaz. 2021. Estimation of Fair Ranking Metrics with Incomplete Judgments. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3442381.3450080

1 INTRODUCTION

Information retrieval (IR) evaluation often focuses on the effectiveness of a system, but there is also a significant history of measuring additional properties of a system's output or behavior, such as novelty and diversity [13]. In the last several years, there has been increased interest in the *fairness* of information retrieval systems [31], with a number of metrics being proposed [2, 5, 15, 34, 41, 44]. While fairness metrics and constructs come in different flavors and aim at different goals, they all attempt to measure the *societal* impacts of the system decisions, and in particular to ensure that those impacts are equitably distributed.

In this paper, we study metrics for *provider group fairness*. This means that the fairness goal is to ensure that the *providers* of items or documents are treated fairly (as opposed to consumers or other stakeholders in multi-sided information access [8]). One way to evaluate this is by measuring whether the exposure different providers receive from the system is equitably distributed among providers of documents with comparable relevance [5, 15]. In this spirit, one class of metrics seeks to ensure that socially-salient *groups* of providers receive comparable exposure; that is, to measure whether providers of, for example, a particular gender, ethnicity, professional seniority, or other group potentially subject to discrimination are systematically disadvantaged in the system's results. This can be measured by aggregating exposure over provider groups, or by measuring the representation of provider groups in result lists [41].

These metrics require the availability of group membership annotations: in order to determine if system results are unfairly discriminating against particular groups, we need to know which groups the various entities in its corpus are associated with. These annotations are often difficult to acquire; reasons for this difficulty include general unavailability, legal and/or ethical restrictions on its collection or use (particularly since group membership is often sensitive personal information), or the cost of expert annotation to e.g. identify content creators' group identities from publicly-available data. These challenges are reflected in analyses of the needs of practitioners building fair systems. In a recent survey of machine learning practitioners by Holstein et al. [20], practical approaches to auditing fairness with limited data were mentioned as one of the most pressing issues. To address the needs, recent work has looked at mechanisms for auditing [23] and optimizing [19, 25] systems in the absence of such labeled data, with some success but also notable limitations, and none of this work has yet been applied to ranking, retrieval, or recommendation systems.

In this work, we consider the case where group membership annotations are available, but at a cost, so it is not practical to obtain

*Now at Google.
†Work in part done while at Microsoft Research.

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

complete labels for the underlying data but a strategically-selected sample can be labeled. This is the case, for instance, when annotations must be provided by human annotators, and researchers or system maintainers wish to make effective use of a limited budget for hiring annotators.

Our goal is to develop *statistical estimation techniques* that enable accurate estimation of a provider group fairness metric, applied to an IR system's ranked outputs, by acquiring group membership annotations for a sample of documents in its corpus. Inspired by the work by Pavlu *et al.* [29] on estimating information retrieval effectiveness metrics using incomplete judgments, along with other work in this line [1, 7, 9, 32, 33, 42, 43], we show how unbiased estimates of various fairness metrics can be computed using estimators based on the Horvitz-Thompson estimator [37]. To our knowledge, this is the first application of information retrieval metric estimation from incomplete data to the problem of assessing system fairness. In particular, we show how unbiased estimates of a few proportion-based metrics [41] and an exposure based metric[15] can be approximated with a significantly smaller number of group membership annotations. While we focus on these particular metrics in this paper, the techniques can easily be extended to estimate other fairness metrics when group membership annotations are incomplete.

2 RELATED WORK

In this section, we present the connection between information retrieval (IR) evaluation techniques and the fairness of IR systems.

2.1 Evaluating Information Retrieval

IR systems find information believed to match a user's information need from a *corpus* of *documents* (or other items). In their most common configuration, which we study here, they return a ranked list of *documents* in response to a *query* (for a search task) or a user's context and interest profile (for a recommendation task).

The performance of these systems is often evaluated with a variant of the "Cranfield protocol" [39], where the system's rankings are compared with a set of ground-truth *relevance judgements* and its accuracy measured using a metric such normalized discounted cumulative gain (nDCG) [22] or expected reciprocal rank [11]. nDCG and related metrics assess the system's ability to place the most relevant documents at the top of its ranked result lists. They also prioritize accuracy at the top of the list, because users tend to pay more attention to the first few results.

The relevance judgments come from variety of sources, depending on the experimental setting. In some cases, they are provided by expert assessors; in others, they are derived from user signals such as clicks, purchases, and ratings. Most metrics, in their naive forms, assume that relevance data is complete: that the evaluation not only know the relevance of every ranked document, but also the relevance of every document in the corpus so it can penalize a system for failing to retrieve relevant documents.

Sampling techniques estimate IR metrics with with incomplete judgments [1, 42, 43]. These methods assume that relevance *can* be assessed for any document with respect to a particular query, but at a cost; they approach the problem by strategically selecting documents to assess in order to accurately estimate the metric based on annotations for a subset of the corpus. We extend these ideas to assess the fairness of a system's rankings instead of its performance; this presents both new opportunities (since a document's protected group affiliation is independent of any particular query) and challenges (methods exploiting the structure of a performance metrics to improve sampling efficiency do not directly translate to fairness metrics).

2.2 Fairness in Algorithmic Systems

Algorithmic fairness is a broad field with many interlocking concepts and sometimes competing; Mitchell et al. [26] provide a useful overview. Most of this work, however, has focused on classification and regression models.

One key distinction in the algorithmic fairness literature is the line between *individual* and *group* fairness [16]. Individual fairness is concerned with ensuring that similar individuals receive similar decisions; in an IR system, this could mean that documents with similar relevance to a query should have equivalent probabilities of being retrieved in a valuable ranking position [5]. Group fairness, as we discussed in the introduction, ensures that data subjects don't experience disparate treatment, decision outcomes, or error on the basis of their group identity or membership. This is often operationalized through the concepts of *protected groups* and *sensitive attributes*, inspired by U.S. anti-discrimination law, resulting in fairness objectives such as "equality of opportunity", the constraint that system decisions should be conditionally independent of group membership given true outcomes [18].

An additional distinction that is particularly relevant to information retrieval systems is the difference between ensuring fairness for information providers, consumers, and other stakeholders in multi-sided systems [8].

2.3 Fairness of Ranked Outputs

Ranking systems, including search, recommendation, and matchmaking, have dynamics that are different than the classification and regression models typical of algorithmic fairness research. This flows from two interconnected problems: first, rank positions are exclusive, in that only one document can be in the first position of a given set of search results; second, such systems often exhibit *position bias*, where users are more likely to inspect results higher in the ranking [12, 14]. Documents that are ranked on top receive higher click rates even if they are not actually relevant to a query [21]. Broadly, there are two families of methods used for measuring the fairness of ranking systems:

Exposure Based Methods. Exposure can be defined as user's discovery of different documents in a ranked list. In other words, it is kind of the distribution of user's attention to documents in ranked list. Exposure-based metrics can quantify the level of attention discrepancy spent in some documents on top but not the lower ranks. In the context of amortized evaluation, Biega *et al.* [5] studies equity of attention among positions in rankings that are relevant to a query. Morik *et al.* [28] introduces a dynamic ranking scheme that optimizes the exposure metric introduced in [5]. Diaz *et al.* [15] extend [5] to the context of stochastic ranking, including both individual and group fairness definitions. Pairwise rank fairness [2, 24] does not directly measure exposure, but is related in that it

measures the system's propensity to correctly or incorrectly rank relevant documents above irrelevant ones based on the relevant document's protected attribute: a system that is systematically more likely to correctly surface documents from the majority group than the protected group is deemed unfair.

Proportion Based Methods. One other framework in algorithmic fairness dictates that all groups in data should be treated similarly [30]. This criterion has been met via statistical parity in fairness settings. Yang and Stoyanovic [41] propose a family of fraction based measures by comparing the distributions of different groups to adapt statistical parity into ranked outputs. Zehlike et al. [44] introduces the notion of following similar distribution of corpus with every position in ranking. They systematically check whether the distribution is preserved or not in each rank.

3 FAIR RANKING METRICS

In this section, we will summarize a broad family of fair ranking metrics. Given a query (for IR) or context (for recommendation), assume a system ranks items from an underlying corpus $\mathcal{D}$ producing ranking π. A document ranked at position i is denoted as π_i; the set of top k documents is denoted as $\pi_{\leq k}$. Let $\mathcal{G}$ be the set of group labels and $\mathcal{D}_g \subseteq \mathcal{D}$ be the subset of documents with group label $g \in \mathcal{G}$. We consider the situation where there are two groups which partition the corpus (i.e. $\mathcal{G} = \{A, B\}$ and $\{\mathcal{D}_A, \mathcal{D}_B\}$ is a partition of $\mathcal{D}$).

Given a ranking π, a group-based fair ranking metric is composed of three computations: (a) measuring the group representation in π, (b) defining a target group representation for that query, and (c) comparing the group representation in π with the target group representation.

3.1 Measuring Group Representation

Group representation can be computed as either the proportion of the groups in the top of the ranking or the probability of examination of groups in the ranking.

Proportion-based representation [10, 41, 44] measures the proportion of items belonging to different groups present in the top k positions of π. Proportion of group g in the ranking π can be computed as:

$$\tilde{\mathbf{P}}_g = \frac{|\pi_{\leq k} \cap \mathcal{D}_g|}{k} \tag{1}$$

Throughout this paper, we use $k = 30$ for all metrics that depend on the top k portion of the ranking.

Exposure-based representation measures the allocation of attention of searchers to items belonging to different groups [5, 15, 34]. Exposure is generally assumed to exponentially decrease with rank, albeit the exact formulations have differed in prior work [15]. There are some scenarios in IR tasks where the fraction of a particular group is the same for all systems. Thus, we need to pick an exposure based metric including individual ranks of documents to measure the fairness. In this paper, we adopt a discounted exposure metric inspired by Rank Biased Precision (RBP) [27] and Expected Exposure in [15]. For the protected group g, the following equation denotes the exposure of g in the top-k ranking results:

$$\tilde{\mathbf{E}}_g = (1-\gamma) \sum_{i=1}^{k} \gamma^{(i-1)} \mathbb{I}\left(\pi_i \in \mathcal{D}_g\right) \tag{2}$$

Here the discount factor, γ, is a decay parameter controlling the importance of higher ranks. We term this metric Exposure, and use this to measure the exposure of protected group.

3.2 Defining a Representation Target

The *representation target* refers to the ideal representation (i.e. proportion or distribution of exposure). The choice of representation target depends on the application domain. In this paper, we consider three targets often used in the literature,

- **Parity:** where the resource allocation should be equal for each group.
- **Proportionality to the corpus presence:** where the resource allocation should be proportional to the number of items in the corpus that belong to a given group.
- **Proportionality to the relevance:** where the resource allocation should be proportional to the number of items belonging to a given group that are relevant to the ranking query.

We use the notation $\mathbf{P}_g$ to refer to target proportion. We do not use a target for the Exposure metric, i.e. $\mathbf{E}_g$; instead we report Equation 2 as protected group's exposure and use this as Exposure metric, a divergence from [15] that keeps with the normative origin of our proportion-based metrics, focusing on the representation of protected group members in the ranking while leaving document relevance as a separate concern.

3.3 Divergence-Based Fairness Metrics

Fairness measures compare the system's proportion or exposure of a protected group with the ideal proportion or exposure suggested by the selected representation target. In this paper, we consider four divergence measures. For proportion-based representations, these are defined as,

$$\Delta_{\text{diff}} = \sum_{g \in \mathcal{G}} (\mathbf{P}_g - \tilde{\mathbf{P}}_g) \qquad \text{Difference} \tag{3}$$

$$\Delta_{\text{abs}} = \sum_{g \in \mathcal{G}} |\mathbf{P}_g - \tilde{\mathbf{P}}_g| \qquad \text{Absolute Difference} \tag{4}$$

$$\Delta_{\text{sq}} = \sum_{g \in \mathcal{G}} \left(\mathbf{P}_g - \tilde{\mathbf{P}}_g\right)^2 \qquad \text{Squared Difference} \tag{5}$$

$$\Delta_{\text{KL}} = \sum_{g \in \mathcal{G}} \mathbf{P}_g \log\left(\frac{\mathbf{P}_g}{\tilde{\mathbf{P}}_g}\right) \qquad \text{KL Divergence} \tag{6}$$

Definitions for exposure-based representations follow analogously. As with the representation target, the choice of divergence measure depends on the application context of the search system.

4 PROBLEM DEFINITION

Given a ranking π and a fixed annotation budget for obtaining group labels, our goal is to develop a sampling based method that can be used to produce an unbiased estimate of the actual value of a fairness metric Δ. That is, we would like to efficiently select only

a subset of items in the corpus to be annotated for membership, and use those to estimate the metric of interest.

5 ESTIMATION METHODOLOGY

There are various ways in which the need for annotations could be reduced such as techniques based on active learning [9]. However, for most of these techniques, there are no guarantees that the values of metrics computed using these techniques will be unbiased estimates for the actual metric value.

Our proposed approach for computing unbiased estimates of fairness metrics with incomplete judgments is based on the statistical estimation framework developed by Pavlu et. al. [29], which was originally proposed for reducing annotation effort in context of IR evaluation.

In this section, we first describe the sampling strategy we use, and show how unbiased estimates of fairness metrics can be computed using the sampled documents.

5.1 Sampling

The first step in our statistical estimation approach based on Pavlu et al. [29] is to select which items should be annotated, which will be done using sampling. One of the advantages of the statistical estimation technique we use in this paper (described in the next section) is that it can be applied to compute unbiased estimates of metrics regardless of what sampling distribution is used in the sampling stage. However, the particular sampling distribution used could have an impact on the variance of the estimator, which would affect the confidence of the final estimator.

There are many different potential sampling distributions that can be used in the sampling process. Which sampling distribution to use could depend on the quantity that needs to be estimated as the estimation could be made more efficient by adopting a sampling distribution that is ideal for the fairness metric in which we are interested. For instance, if the goal is to estimate an exposure based metric, which gives more weights to the items towards the top end of the ranking compared to the bottom, it would be better to use a sampling distribution that gives more weights to items towards the top. If uniform sampling, i.e., sampling documents uniformly at random and label them for group membership, were to be used instead, it is likely that we will be spending our annotation budget on items that have little to no impact on the value of the chosen metric.

In this paper, we adopt a sampling strategy proposed by Pavlu and Aslam [29], which was shown to achieve good performance in estimation of top heavy IR metrics such as average precision. The approach is based on using a prior distribution that associates each rank position with a weight. Let R be the length of a given ranked list of items, whose quality we are trying to estimate. Then, the sampling weight associated with an item that is retrieved at rank r can be computed as:

$$W(r) = \frac{1}{2R}\left(1 + \frac{1}{r} + \frac{1}{r+1} + \ldots + \frac{1}{R}\right) \approx \frac{1}{2R}\log\left(\frac{R}{r}\right) \tag{7}$$

Typically, we would have many ranked lists (systems), the quality of which needs to be estimated at the same time, using the same sampled dataset. Hence, in order to obtain the final sampling distribution that can work reasonably well across all these systems, we first generate these weights for each items retrieved by each system and then average the weights of each item across all systems, resulting in a single weight for each item. Finally, these weights are converted to a probability distribution by normalizing with the sum of weights over all the items.

For the actual sampling process, the stratified sampling strategy by Stevens [6, 35] that has also been used by Pavlu et al. [29] has been shown to have the advantage of achieving reduced variance. Hence, we also adopt this sampling strategy in this paper.

Let m be the amount of annotation budget we have available. The stratified sampling process works as follows [29]:

(1) Order the items in decreasing order based on their sampling probability (Eq. 7), and partition them into buckets of size m.
(2) For each bucket, assign a sampling probability for that particular bucket by taking the mean of items' probabilities that fall under each bucket, and normalizing across all buckets.
(3) Sample buckets with replacement m times.
(4) Uniformly sample items without replacement from each sampled bucket. If a bucket is sampled n times, sample uniformly n items from that bucket without replacement.

Note that this particular sampling strategy works by first sampling buckets, and then sampling items from each bucket, as opposed to directly sampling items. We call this strategy as *weighted sampling* throughout the paper.

We would like to further emphasize that while we decided to use the sampling distribution described in Equation 7 in this work, the estimation technique we use in this paper can work with *any* distribution and produce unbiased estimates. Thus, our approach has the flexibility towards incorporating prior knowledge of group distribution to stakeholders in fairness, i.e. law makers, practitioners such that the sampling distribution can be changed to meet the prior knowledge.

5.2 Statistical Estimation of Fairness Metrics

After obtaining samples drawn from a sampling distribution, we need to derive an estimator to compute the estimated value of a fairness metric. Our approach is based on extending the statistical estimation method from Pavlu and Aslam [29], which uses the Horvitz-Thompson estimator [38] for estimating values of IR metrics, to estimating values of fairness metrics.

5.2.1 Horvitz-Thompson Estimator for Estimation of the Mean. Suppose we are given a sample set S of size m sampled from a population D. According to the Horvitz-Thompson estimator [38], an unbiased estimate of the population mean X can be computed as:

$$\widehat{X} = \frac{\sum_{i \in S} \frac{f(\pi_i)}{\theta_i}}{|D|} \tag{8}$$

where $f(\pi_i)$ is the value associated with item i and θ_i is the *inclusion probability* for item i, which represents the probability that item i would be included in *any* sample of size m.

One should note that when samples are drawn using the stratified sampling strategy described in the previous section, the inclusion

probability of item i is different than the sampling probability associated with this item. Yet, inclusion probabilities can be derived from the sampling probabilities as follows: Let b be the sampling probability for a bucket of size m (where b is the sum of the sampling probabilities of all items that fall under this bucket), and T be a random variable that indicates the number of times this bucket has been selected in the stratified sampling process. Then, θ_i for an item i within this bucket can be computed as [29]:

$$\theta_i = \sum_{k=1}^{m} P(T=k)\frac{k}{m} = \frac{1}{m}E[T] = \frac{1}{m}mb = b \tag{9}$$

5.2.2 Horvitz-Thompson Estimator for Estimation of Fairness Metrics. Recall that $\tilde{\mathbf{P}}_g$ value described in Equation 1, indicates the proportion of group g within the top k ranking results. Then, using Horvitz-Thompson estimator, $\widehat{\tilde{\mathbf{P}}_g}$, an unbiased estimator for $\tilde{\mathbf{P}}_g$, can be computed as:

$$\widehat{\tilde{\mathbf{P}}_g} = \frac{1}{k}\sum_{i\in S, rank(i)\leq k} \frac{\mathbb{I}\left(\pi_i \in \mathcal{D}_g\right)}{\theta_i} \tag{10}$$

Hence, the estimators for the four divergence based fairness metrics defined in Equation 3 can be computed as:

$$\widehat{\Delta_{\text{diff}}} = \sum_{g\in\mathcal{G}} (\mathbf{P}_g - \widehat{\tilde{\mathbf{P}}_g}) \tag{11}$$

$$\widehat{\Delta_{\text{abs}}} = \sum_{g\in\mathcal{G}} |\mathbf{P}_g - \widehat{\tilde{\mathbf{P}}_g}| \tag{12}$$

$$\widehat{\Delta_{\text{sq}}} = \sum_{g\in\mathcal{G}} \left(\mathbf{P}_g - \widehat{\tilde{\mathbf{P}}_g}\right)^2 \tag{13}$$

$$\widehat{\Delta_{\text{KL}}} = \sum_{g\in\mathcal{G}} \mathbf{P}_g \log\left(\frac{\mathbf{P}_g}{\widehat{\tilde{\mathbf{P}}_g}}\right) \tag{14}$$

Similarly, the estimator $\widehat{\tilde{\mathbf{E}}_g}$ for group exposure $\tilde{\mathbf{E}}_g$ described in Equation 2 can be computed by substituting the value of each item in the sample with $\gamma^{(rank(i)-1)}\mathbb{I}\left(\pi_i \in \mathcal{D}_g\right)$, which leads to the formula below:

$$\widehat{\tilde{\mathbf{E}}_g} = (1-\gamma)\sum_{i\in S, rank(i)\leq k} \gamma^{(i-1)}\frac{\mathbb{I}\left(\pi_i \in \mathcal{D}_g\right)}{\theta_i} \tag{15}$$

6 EXPERIMENTAL SETUP

We evaluated the quality of our proposed estimation method across three different experimental conditions. In this section, we will describe our experimental setup, including datasets, metrics, and baselines.

6.1 Data

6.1.1 Synthetic Data. We develop a simulator to generate various simulated rankings (systems) in order to analyse the performance of our estimation methods in depth over multiple fair ranking metrics, and we list the variables used in our simulator as follows:

- Q: number of queries.
- D: number of documents in corpus.
- M: number of systems submitted.
- N: number of documents retrieved for each query, i.e. $N \leq D$.
- h_q: parameter modeling the "easiness" of query q, that is, how difficult the query is for a baseline retrieval system.
- r_{dq}: simulated relevance of document d for query q.
- α_m: parameter modeling the average effectiveness of system m, independent of query.
- γ_m: parameter modeling the bias of a system towards or against a sensitive group.
- β: parameter modeling the distribution of the protected group, $\mathbb{I}\left(d \in \mathcal{D}_A\right)$ defined in Section 3, in corpus.
- π_{qm}: system's retrieval result for query q.
- $\Delta(\pi_{qm})$: fairness performance of a system, in query q.

The simulator begins by assuming the existence of a test collection with D documents, Q queries, and relevance judgments between every ⟨ query, document ⟩ pair. These relevance judgments are simulated per query as follows: we first sample, for each of the Q queries, a "query easiness" parameter h_q from a prior beta distribution with parameters. This models the proportion of expected relevant documents for the query q. This parameter is then used in a Bernoulli distribution to sample the binary relevance r_{dq} of each document d to the query q. Via this procedure, we obtain a simulated set of queries with varying numbers of relevant documents, some much more than others, which is typical of an information retrieval test collection.

We next simulate protected class. We make the assumption that protected class is independent of relevance, and independent of query; we assume it is a global property of a document. Parameter β models the expected proportion of protected group members. Similarly to relevance, protected group status is sampled from a Bernoulli distribution with parameter β.

Retrieval systems are then simulated by sampling document scores S_{dq} and ranking them in decreasing order. A score S_{dq} is a function of the query easiness parameter h_q described above, a "system goodness" parameter α_m, the relevance r_{dq}, the protected group status, and a global "group bias" parameter γ. Specifically, the score is sampled as follows, with different cases for different combinations of relevance r_{dq} and protected group membership $I(d \in \mathcal{D}_A)$:

$$S_{dq} \sim \begin{cases} \mathcal{N}(0,\sigma) & \text{if } r_{dq}=0 \text{ and } I(d\in\mathcal{D}_A)=0 \\ \mathcal{N}(\alpha_m + h_q, \sigma) & \text{if } r_{dq}=1 \text{ and } I(d\in\mathcal{D}_A)=0 \\ \mathcal{N}(\gamma,\sigma) & \text{if } r_{dq}=0 \text{ and } I(d\in\mathcal{D}_A)=1 \\ \mathcal{N}(\alpha_m + h_q + \gamma, \sigma) & \text{if } r_{dq}=1 \text{ and } I(d\in\mathcal{D}_A)=1 \end{cases}$$

In effect, this process results in relevant documents having higher scores correlated to both system goodness and query easiness, and protected group members having higher scores correlated to global bias. Once scores have been generated for all documents for a query q, they are ranked in decreasing order by score (π_{qm}), and then all standard IR and fairness measures can be computed over the ranking.

We can simulate a wide variety of different experimental conditions by manipulating the variables D, Q, M and parameters for prior beta distributions and group membership proportion. By setting the β parameter for protected group membership to $(1, 1)$, we

made sure that both groups have equal presence in the generated corpus. Hence, a fairness target of 0.5 is being used for divergence based metrics described in Section 3.3.

In our experiments, we generated $M = 800$ systems retrieving $N = 100$ documents from a corpus containing $D = 1000$ documents for $Q = 50$ queries.

Number of Systems (M)	800
Number of Queries (Q)	50
Number of Documents (D)	1000
Number of Retrieved Documents (N)	100
Average documents in protected group per query	≈50

Table 1: Summary of Synthetic Data.

6.1.2 Real-World Data. Beyond the synthetic data, we test the performance of our method using two real-world datasets: (1) TREC Fair Ranking dataset consisting of submissions to the TREC Fair Ranking Track [3], and (2) book recommendation data from Ekstrand and Kluver [17] consisting of ranked lists of books recommendations. Below we describe each dataset in more detail:

TREC Fair Ranking Dataset. In the TREC Fair Ranking Track, each participant was asked to re-rank a given initial list of documents for each query. This dataset contains the binary protected attributes of group membership for each of document, as well as relevance judgments for each query-document pair. Since each participant of the Fair Ranking Track was given the same initial list of documents to re-rank, the proportion of each group is identical for all submitted systems [4]. Hence, divergence based metrics that depend on group proportions are not directly applicable to this dataset. Instead, Exposure based fairness metrics (explained in Equation 2) can be used to compare fairness among different systems. We use exposure of group called *advanced* in the annotations to compute our Exposure metric. Details about this dataset can be seen in the table below.

Number of Systems (M)	10
Number of Queries (Q)	635
Number of Documents (D)	2671
Average number of documents per query	7
Mode of documents per query	6
Average member from protected group per query	≈4

Table 2: Summary of TREC Data.

Book Recommendation. We also test our estimation method on the book data and recommendation models assembled by Ekstrand and Kluver [17]. This dataset combines user-book interaction data from GoodReads [40] with book metadata from the Library of Congress and OpenLibrary. For each book, the dataset also contains a binary atrribute dictating the gender of the author[1]. From this data, we generated 100-item recommendation lists for 5000 users via matrix factorization method using implicit feedback [36]. The dataset contains a ranked list of book recommendations for each user. Hence, the fairness metrics can be computed for each user, treating women as the protected group [17]. The gender distribution in the pooled corpus of *all* recommended books is used as our fairness target for divergence based metrics. Table 3 shows more details about this dataset.

Number of Users (M)	5000
Number of Books (D)	≈500K
Average number of books per user	100
Average books from protected group per user	≈24

Table 3: Summary of Book Recommendation Data.

6.2 Sampling Setup

We simulate the setup of not having complete annotations available by sampling from the set of complete judgments using different sampling rates $p \in \{0.1, 0.2, 0.3, 0.4, ..., 0.9\}$, where sampling using a sampling rate of $p = 0.1$ corresponds to generating a dataset that contains 10% of the complete judgements.

The simulated dataset and the TREC Fair dataset contains two types of annotations: the relevance judgements associated with each query document pair, as well as the protected attribute associated with each document. In our experiments, we assume that complete relevance judgments are available, and that only annotations related to the protected attributes are incomplete. Hence, when we generate our sampled datasets, we only sample from the protected attribute annotations.

For the Book Recommendation dataset, we assume that all recommended items are relevant and sample from the gender attribute.

6.3 Comparison of Estimated vs. Actual Values

Given a sampled dataset, we compute the estimated values of the various fairness metrics using our proposed estimators for each system. We then evaluate the quality of our estimations by comparing them with actual metric values (computed using all the available judgments, as opposed to just the sampled ones) using the following statistics:

(i) *Kendall's τ*: This statistic is used to compute the correlation between two system rankings. Its value can range from -1 (perfectly negative rank correlation) to 1 (perfectly correlated).

(ii) *Root Mean Squared Error (RMSE)*: This statistic is used to compare the actual and estimated metric values across all the systems in the experiment.

7 BASELINES

We compare are method against two baseline approaches, a non-sampling technique and uniform sampling.

7.1 Induced Method Baseline

We first compare against *induced metrics* proposed by Yilmaz and Aslam [42], which are shown to achieve reasonable performance for

[1]Binary gender is a limitation of the underlying author metadata; see Ekstrand and Kluver [17] for a detailed discussion of data limitations.

estimating IR metrics when judgments are incomplete. The induced version of a metric is computed by filtering the ranking to only include the labelled items instead of measuring the whole ranking based on the sample. This is done by removing the unlabelled items from the list, as a result of which lower-ranked labelled items move up in the ranking. The metric is then computed over this induced ranking that only contains labelled items.

7.2 Uniform Sampling Baseline

In a uniform sampling setup, instead of using our top heavy sampling distribution, sampling is performed uniformly at random, all items having equal likelihood of being included in the sample.

When uniform sampling is used in the sampling phase, estimation becomes quite straightforward as the actual mean can simply be estimated by computing the simple sample mean. Hence, in such a setup the estimator for $\tilde{\mathbf{P}}_g$ can be computed as:

$$\widehat{\tilde{\mathbf{P}}_g} = \frac{1}{k} \sum_{i \in S, rank(i) \leq k} \mathbb{I}\left(\pi_i \in \mathcal{D}_g\right) \tag{16}$$

Substituting $\widehat{\tilde{\mathbf{P}}_g}$ in the set of equations from Equation 11 to Equation 14 corresponds to the estimations of various divergence based fairness metrics under uniform sampling.

Similarly, the estimator $\widehat{\tilde{\mathbf{E}}_g}$ for the Exposure metric can be computed as:

$$\widehat{\tilde{\mathbf{E}}_g} = (1-\gamma)\frac{1}{k} \sum_{i \in S, rank(i) \leq k} \gamma^{(i-1)} \mathbb{I}\left(\pi_i \in \mathcal{D}_g\right) \tag{17}$$

8 RESULTS

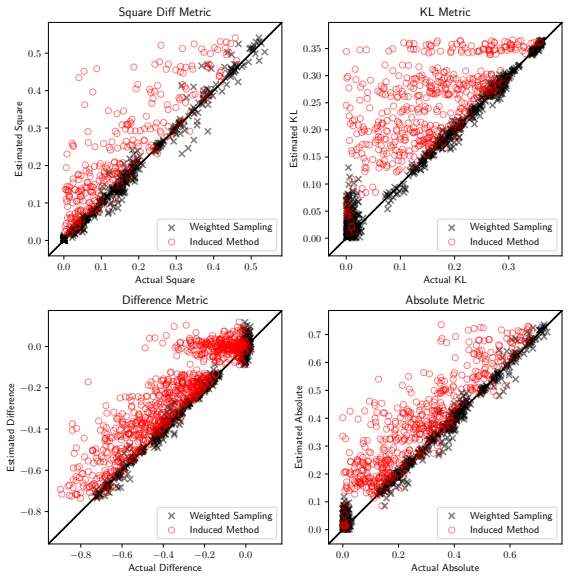
In this section, we examine our method's performance under different experimental conditions. We mainly focus on the scenario when 10% of the complete annotations are available (corresponding to sampling rate $p = 0.1$), although we also report aggregate results for different sampling rates.

In what follows, we first compare our fair ranking metric estimations against the induced method on both the synthetic and the real-world datasets. We then compare the quality of estimators obtained using weighted sampling with that of uniform sampling.

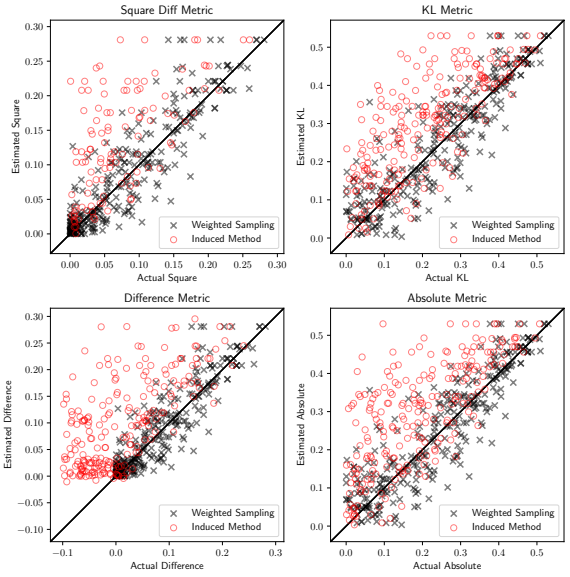
8.1 Estimated vs. Induced Metrics

We test how our estimation method performs against the *induced* method using both the synthetic and the real world-data.

Figures 1a and 1b depict our results for the four divergence based fair ranking metrics for the synthetic and the book recommendation datasets, respectively. The x axis in these figures show the actual metric value (if we had complete judgments available), while the y axis shows the estimated values computed using incomplete judgments. For comparison purposes, we also plot the line $y = x$ in the plots. As it can be seen, our estimates are generally well-distributed around the line $y = x$, validating that they are unbiased. On the other hand, the induced baseline tends to over-estimate the actual values consistently for both datasets. Since the proportions of different groups are identical in each system's output in the TREC Fair dataset, we do not report any proportion based metric results for this dataset.



(a) Synthetic Data



(b) Book Data

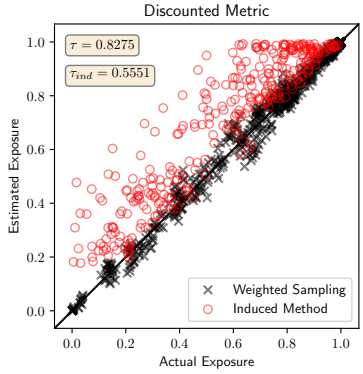
Figure 1: Estimation of divergence based fairness metrics using the proposed method versus the induced baseline. Each dot in the synthetic data represents the mean performance of each system over different queries; in book recommendation data, each dot represents the performance of recommendations per user.

Next, we examine the estimation of the discounted metrics on the synthetic and the book recommendation datasets, results of which

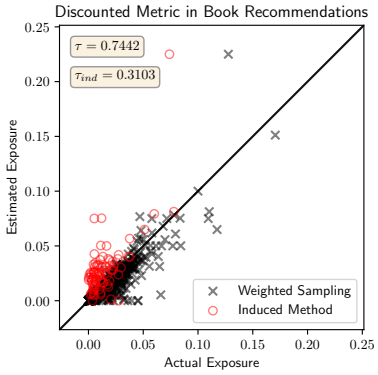
	Weighted		Induced	
	RMSE	τ	RMSE	τ
Δ_{abs}	0.0332	0.8112	0.1361	0.5792
Δ_{sq}	0.0303	0.8014	0.1103	0.5978
Δ_{KL}	0.0298	0.8413	0.1131	0.5725
Exposure	0.0341	0.8275	0.1412	0.5551

Table 4: RMSE and Kendall's τ values for the proposed vs. induced method in synthetic data.

can be seen in Figures 2a and 2b. The plots show the Kendall's τ correlations between the actual and estimated values.



(a) Synthetic Data



(b) Book Data

Figure 2: Estimation of the exposure metric. τ and τ_{ind} denote the Kendall's τ correlation coefficients for weighted and induced methods respectively. Each dot in synthetic data represents the mean performance of each system over different queries; in book recommendation data, each dot represents the performance of recommendations per user. The horizontal axis denotes the actual value for exposure metric and vertical axis is the estimated/induced values in both figures.

Note that all the previous reported results are over a *single* sampled dataset, which could affect the conclusions reached due to the randomness in the sampling process. In order to make sure our results are not affected by the particular sample chosen in the sampling phase, we further generate 10 different samples using a sampling rate of $p = 0.1$ and report the mean RMSE and mean Kendall's τ values across these 10 samples. Table 4 demonstrates the RMSE and Kendall's τ values when comparing actual and estimated metric values for the various divergence based metrics and the Exposure metric using the synthetic dataset. As it can be seen, the proposed method results in much lower RMSE values and higher Kendall's τ correlations when compared to the induced method.

While the results we have presented until now mainly focus on a sampling rate $p = 0.1$, our results seem consistent across different sampling rates. Figure 3a and Figure 3b show how the quality of our proposed method compares with that of induced method for various sampling rates $p \in \{0.1, 0.2, ...0.9\}$. In these plots, we focus on the Squared Difference metric and the Exposure metric[2] as the evaluation metrics. The x axis in the plots show the rate of unjudged documents $(1 - p)$ when sampled datasets are generated using various sampling rates p, and the y axis shows the Kendall's *tau* correlations between the actual vs. estimated values. In order to ensure that the results are not affected by a particular sample, the reported values for each sampling rate are the average Kendall's τ correlation values across 10 different randomly sampled datasets. It can be seen that the estimation method consistently outperforms the induced method over all the different sampling percentages, with the gap between the two methods increasing as judgments become more incomplete.

Overall, our results show that the proposed statistical estimation technique together with the weighted sampling strategy results in more accurate estimates of fairness metrics compared to the induced method when judgments are incomplete.

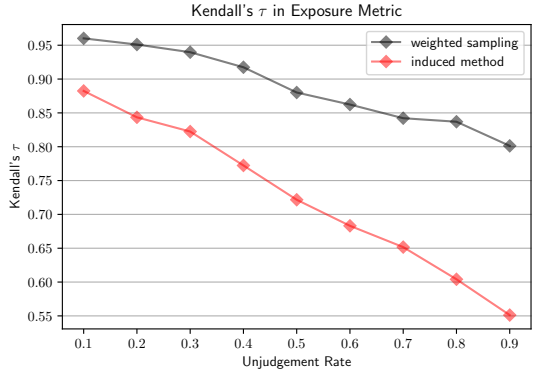
8.2 Weighted vs. Uniform Sampling

Having shown that our weighted sampling strategy outperforms the induced method, we now use the TREC Fair Ranking and the synthetic datasets (Section 6.1.2) to show how the estimators based on our proposed method using weighted sampling compare with estimators based on uniform sampling. In this section, we use the formulas based on a simple mean estimator (as described in Section 8.2) when uniform sampling is used.

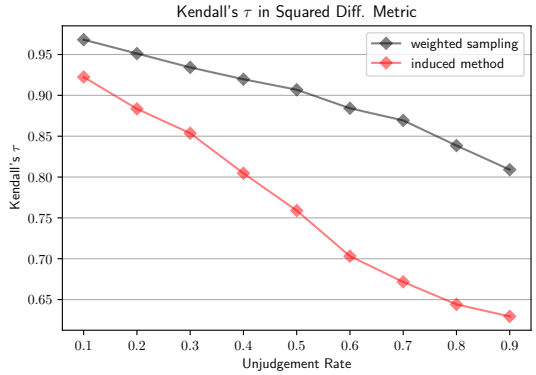
Figure 4a and Figure 4b show the quality of estimations with a sampling rate of $p = 0.1$ for the protected group exposure metric on the TREC dataset. The proposed estimator based on weighted sampling results in much better estimates compared to uniform sampling.

Table 5 demonstrates the RMSE and Kendall's τ values when comparing the actual and the estimated metric values for the various proportion based metrics and the Protected Group Exposure metric using the simulated dataset when weighted vs uniform sampling is used. Similar to the setup in Table 4, in order to make sure our results are not affected by the particular sample chosen, for this experiment, we generate 10 different samples using a sampling rate of $p = 0.1$ and report the mean RMSE and mean Kendall's τ

[2]We observed similar results for other divergence metrics.



(a) Protected Group Exposure Metric



(b) Squared Difference Proportion Metric

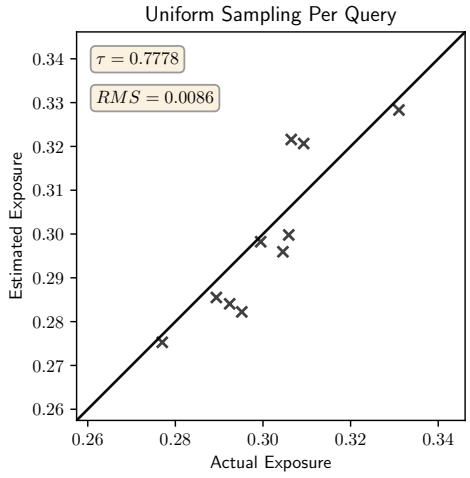
Figure 3: Comparison of weighted sampling and induced method for (top) Protected Group Exposure metric and (bottom) Squared Difference metric.

	Weighted		Uniform	
	RMSE	τ	RMSE	τ
Δ_{abs}	0.0332	0.8112	0.0253	0.8013
Δ_{sq}	0.0303	0.8014	0.0348	0.8045
Δ_{KL}	0.0298	0.8413	0.0289	0.7981
Exposure	0.0341	0.8275	0.0421	0.7147

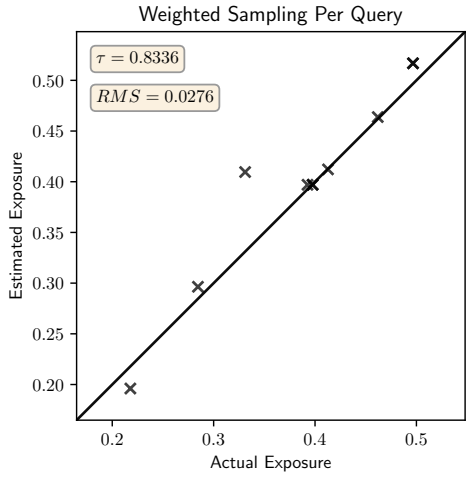
Table 5: RMSE and Kendall's τ values for weighted vs. uniform sampling in synthetic data.

values across these 10 samples. In order to further compare how the quality of the two estimators change across different sampling rates, Figure 5a and Figure 5b show how the weighted sampling estimates compare with that of using uniform sampling for the Protected Group Exposure metric and Squared Difference metric, for various sampling rates p.

Since Protected Group Exposure is a top-heavy metric, giving more weight to the items towards the top end of a ranking, the estimates obtained using our proposed method, which also uses



(a)



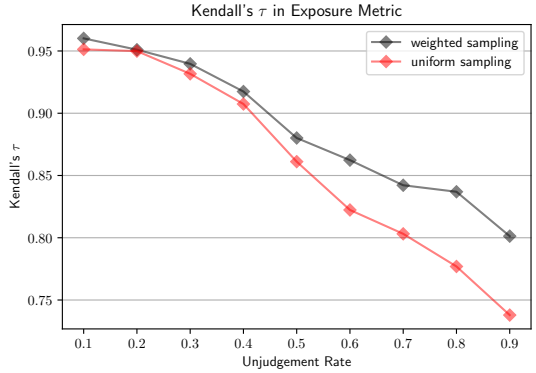
(b)

Figure 4: Estimations on TREC Fair Ranking data for the Protected Group Exposure metric.

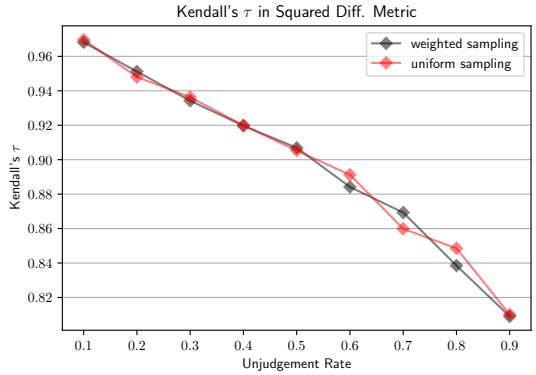
a top-heavy sampling distribution gives much better results compared to uniform sampling. On the other hand, for the estimation of divergence based metrics such as Squared Difference Metric, which gives equal weight to all items in the ranking, our proposed method performs very similar to uniform sampling. This result further validates that our proposed method produces unbiased estimates of metrics even for metrics that are not top-heavy, even though the sampling distribution used in the sampling process is a top-heavy one.

9 CONCLUSION AND FUTURE WORK

In this paper, we have adopted the statistical estimation framework developed by Pavlu et al. [29] to estimating values of various fair ranking metrics in a scenario where protected group annotations



(a) Protected Group Exposure Metric



(b) Squared Difference Proportion Metric

Figure 5: Comparison of different estimation techniques for(top) Protected Group Exposure metric and (bottom) Squared Difference metric.

are incomplete. In particular, we developed techniques based on the Horvitz-Thompson estimator [38] for estimating five different fairness metrics that fall under two classes of fairness notions—proportion-based and exposure-based.

Our results show that the proposed method, which uses a top heavy sampling strategy, results in unbiased estimates of fair ranking metrics and outperforms naive baselines such as the induced baseline by Yilmaz et al. [42], as well as uniform random sampling.

For future work, we aim to explore three directions. First, although focusing on binary protected attributes covers many common scenarios, expanding our estimation technique to more than two groups and multiple protected attributes is still an important aspect for intersectional fairness issues. Second, we plan to further extend our statistical estimation method towards the estimation of new fairness metrics. Last but not least, we plan to work on identifying sampling distributions that would minimize the number of annotations needed, while achieving high confidence estimates.

ACKNOWLEDGMENTS

This project was funded by the EPSRC Fellowship titled "Task Based Information Retrieval", grant reference number EP/P024289/1. Michael Ekstrand's contribution to this work was supported by the National Science Foundation under Grant No. IIS 17-51278.

REFERENCES

[1] Javed A Aslam, Virgil Pavlu, and Emine Yilmaz. 2006. A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 541–548.

[2] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2212–2220.

[3] 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*.

[4] 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).

[5] 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*. 405–414.

[6] Ken RW Brewer and Muhammad Hanif. 2013. *Sampling with unequal probabilities*. Vol. 15. Springer Science & Business Media.

[7] Chris Buckley and Ellen M. Voorhees. 2004. Retrieval Evaluation with Incomplete Information *(SIGIR '04)*. Association for Computing Machinery, New York, NY, USA, 25–32.

[8] Robin Burke. 2017. Multisided Fairness for Recommendation. (July 2017). arXiv:1707.00093 [cs.CY]

[9] Ben Carterette, James Allan, and Ramesh Sitaraman. 2006. Minimal Test Collections for Retrieval Evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Seattle, Washington, USA) *(SIGIR '06)*. Association for Computing Machinery, New York, NY, USA, 268–275.

[10] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2017. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840* (2017).

[11] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (Hong Kong, China) *(CIKM '09)*. ACM, New York, NY, USA, 621–630. https://doi.org/10.1145/1645953.1646033

[12] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. Click models for web search. *Synthesis lectures on information concepts, retrieval, and services* 7, 3 (2015), 1–115.

[13] Charles L A Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. ACM, New York, NY, USA, 659–666. https://doi.org/10.1145/1390334.1390446

[14] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*. 87–94.

[15] Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. *arXiv preprint arXiv:2004.13157* (2020).

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[17] Michael D. Ekstrand and Daniel Kluver. 2021. Exploring Author Gender in Book Rating and Recommendation. *User Modeling and User-Adapted Interaction* (2021). https://doi.org/10.1007/s11257-020-09284-2

[18] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[19] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness Without Demographics in Repeated Loss Minimization *(Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholmsmässan, Stockholm Sweden, 1929–1938. http://proceedings.mlr.press/v80/hashimoto18a.html

[20] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[21] Thorsten Joachims and Filip Radlinski. 2007. Search engines that learn from implicit feedback. *Computer* 40, 8 (2007), 34–40.
[22] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 4 (Oct. 2002), 422–446. https://doi.org/10.1145/582415.582418
[23] Nathan Kallus, Xiaojie Mao, and Angela Zhou. 2019. Assessing algorithmic fairness with unobserved protected class using data combination. *arXiv preprint arXiv:1906.00285* (2019).
[24] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. Fare: Diagnostics for fair ranking using pairwise error metrics. In *The World Wide Web Conference*. 2936–2942.
[25] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without Demographics through Adversarially Reweighted Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 728–740.
[26] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2018. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. (Nov. 2018). arXiv:1811.07867 [stat.AP]
[27] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), 1–27.
[28] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling Fairness and Bias in Dynamic Learning-to-Rank. *arXiv preprint arXiv:2005.14713* (2020).
[29] V Pavlu and J Aslam. 2007. A practical sampling strategy for efficient retrieval evaluation. *College of Computer and Information Science, Northeastern University* (2007).
[30] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2009. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the 2009 SIAM international conference on data mining*. SIAM, 581–592.
[31] Adam Roegiest, Aldo Lipani, Alex Beutel, Alexandra Olteanu, Ana Lucic, Ana-Andreea Stoica, Anubrata Das, Asia Biega, Bart Voorn, Claudia Hauff, Damiano Spina, David Lewis, Douglas W Oard, Emine Yilmaz, Faegheh Hasibi, Gabriella Kazai, Graham McDonald, Hinda Haned, Iadh Ounis, Ilse van der Linden, Jean Garcia-Gathright, Joris Baan, Kamuela N Lau, Krisztian Balog, Maarten de Rijke, Mahmoud Sayed, Maria Panteli, Mark Sanderson, Matthew Lease, Michael D Ekstrand, Preethi Lahoti, and Toshihiro Kamishima. 2019. FACTS-IR: Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval. *SIGIR Forum* 53, 2 (Dec. 2019), 20–43.
[32] Kevin Roitero, Andrea Brunello, Giuseppe Serra, and Stefano Mizzaro. 2020. Effectiveness evaluation without human relevance judgments: A systematic analysis of existing methods and of their combinations. *Information Processing & Management* 57, 2 (2020), 102149.
[33] Kevin Roitero, Marco Passon, Giuseppe Serra, and Stefano Mizzaro. 2018. Reproduce. generalize. extend. on information retrieval evaluation without relevance judgments. *Journal of Data and Information Quality (JDIQ)* 10, 3 (2018), 1–32.
[34] 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*. 2219–2228.
[35] WL Stevens. 1958. Sampling without replacement with probability proportional to size. *Journal of the Royal Statistical Society: Series B (Methodological)* 20, 2 (1958), 393–397.
[36] Gábor Takács, István Pilászy, and Domonkos Tikk. 2011. Applications of the conjugate gradient method for implicit feedback collaborative filtering. In *Proceedings of the fifth ACM conference on Recommender systems (RecSys '11)*. ACM, 297–300. https://doi.org/10.1145/2043932.2043987
[37] S. K. Thompson. 2002. *Sampling*. Wiley-Interscience.
[38] Steven K. Thompson. 2002. *Sampling*. Wile-Interscience, second edition, 2002.
[39] Ellen M Voorhees. 2001. The Philosophy of Information Retrieval Evaluation. In *Evaluation of Cross-Language Information Retrieval Systems*, Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck (Eds.). Springer Berlin Heidelberg, 355–370. http://link.springer.com/chapter/10.1007/3-540-45691-0_34
[40] Mengting Wan and Julian McAuley. 2018. Item Recommendation on Monotonic Behavior Chains. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 86–94. https://doi.org/10.1145/3240323.3240369
[41] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. 1–6.
[42] Emine Yilmaz and Javed A Aslam. 2006. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM international conference on Information and knowledge management*. 102–111.
[43] Emine Yilmaz, Evangelos Kanoulas, and Javed A Aslam. 2008. A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 603–610.
[44] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1569–1578.