Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists

Piotr Sapiezynski
Northeastern University
p.sapiezynski@northeastern.edu

Wesley Zeng
Northeastern University
zeng.wes@husky.neu.edu

Ronald E. Robertson
Northeastern University
robertson.ron@husky.neu.edu

Alan Mislove
Northeastern University
amislove@ccs.neu.edu

Christo Wilson
Northeastern University
cbw@ccs.neu.edu

ABSTRACT

In this work, we introduce a novel metric for auditing group fairness in ranked lists. Our approach offers two benefits compared to the state of the art. First, we offer a blueprint for modeling of user attention. Rather than assuming a logarithmic loss in importance as a function of the rank, we can account for varying user behaviors through parametrization. For example, we expect a user to see more items during a viewing of a social media feed than when they inspect the results list of a single web search query. Second, we allow non-binary protected attributes to enable investigating inherently continuous attributes (e.g., political alignment on the liberal to conservative spectrum) as well as to facilitate measurements across aggregated sets of search results, rather than separately for each result list. By combining these two elements into our metric, we are able to better address the human factors inherent in this problem. We measure the whole sociotechnical system, consisting of a ranking algorithm and individuals using it, instead of exclusively focusing on the ranking algorithm. Finally, we use our metric to perform three simulated fairness audits. We show that determining fairness of a ranked output necessitates knowledge (or a model) of the end-users of the particular service. Depending on their attention distribution function, a fixed ranking of results can appear biased both in favor and against a protected group.

CCS CONCEPTS

• **Information systems** → **Page and site ranking**; **Content ranking**; • **Human-centered computing** → *User interface design.*

KEYWORDS

information retrieval; group fairness; ranked lists

1 INTRODUCTION

The exponential growth of information available online has necessitated the development of Information Retrieval (IR) algorithms that decide what content is *relevant* to users. For example, upwards of 5.5 billion searches are conducted on Google every day [18], and in response to each, Google filters and sorts a list of ∼15 results [46]. Similarly, over 1.4 billion people visit Facebook daily [25] and scroll through a list of content from friends and advertisers deemed relevant by the News Feed algorithm. Finally, tens of millions of worker profiles are available on LinkedIn, filtered and sorted when recruiters search for prospective employees.

Recently, a concern has been growing that even seemingly benign IR systems may negatively impact people. It has been shown that algorithms can reflect societal biases [2], and ranking mechanisms are no different. Kay et al. found that Google Image Search returned images portraying men and women in stereotypical roles in response to occupation-related queries, and that these results reinforced stereotypical gender roles [34]. Others have examined partisan slant in search results [12, 38, 46] in light of user studies demonstrating that partisan search results can significantly influence voting behavior [15, 16]. Lastly, two studies have examined the relationships between gender, race, and ranking of job seekers on employment websites [9, 23], where systematic biases that push members of protected classes into lower ranks could result in the loss of employment opportunities and earnings [36].

Only recently have researchers started addressing the problem of fairness in ranked outputs. From the IR side, this includes novel ranking algorithms that aim to achieve *representational parity* (a.k.a. *group fairness*): the ranker is required to assign a certain fraction of top ranks to people in the protected or minority class) [3, 8, 50, 54]. From the *algorithm auditing* side (i.e., investigators who look for fairness problems in black-box systems) [48], Yang and Stoyanovich introduced metrics for quantifying whether the outputs from a given search engine are group-fair [51]. This enables auditors to examine real-world search engines and hopefully hold them accountable for producing unfair outputs.

However, there is a reoccurring challenge in the extant literature on fair ranking: how to model user attention? Eye-tracking studies and click-stream data show that users do not distribute their attention evenly over ranked lists of information [13, 19, 21, 42]. This unequal distribution of attention must be taken into account when designing fair IR systems and evaluating whether a given IR system is fair. The trouble is that the distribution of attention for a given search engine may be unknown, since it varies based on the user interface of the service (e.g., pagination boundaries) and context (e.g., searching for a specific movie trailer versus searching for a new employee).

Most of the previous work on fairness in ranked lists has assumed logarithmic discounting of attention [50, 51]. However, because of its flattening shape for low ranks, logarithmic drop-off is impractical for modeling attention. For example, modeling attention this way would mean assuming that on a list of 100 results, the sum of attention given to last eight results is bigger than the attention paid to the first. Biega et al. use a geometric distribution instead, but do not investigate the consequences of varying its steepness [3].

In this work, we extend the literature on fair ranking by introducing a novel metric for measuring group fairness in ranked outputs. Our metric, the *Viable-Λ Test*, is designed for auditors and answers two questions: (1) does there exist a distribution of user attention $P(\Lambda)$ such that the output of a search engine is group fair, and (2) if so, what is the parameterization $\Lambda = (\lambda_1, ..., \lambda_m)$ of this distribution? In contrast to prior work that attempts to "score" the fairness of a ranking algorithm [51], our metric fundamentally re-frames the question of fairness to involve the consumer of the ranking and their attention. If the fitted model of attention $P(\Lambda)$ does not match empirical observations of user attention in the given search engine, then the system does not achieve representational parity.

Overall, our paper makes three key contributions:

(1) We introduce a novel metric, the Viable-Λ Test that binds the usage patterns of a list to the measurement of fairness.
(2) We enable fairness measurements in situations with class assignment uncertainty, results aggregation, multiple protected classes, and continuous protected variables.
(3) We evaluate the Viable-Λ Test on data from three real-world services: a resume search engine, a dating service, and a web search engine. Our results demonstrate that the choice and parameterization of the attention function can lead to dramatically different conclusions about whether (and how) the rankings are biased.

Note that it is not possible to determine with certainty whether a given set of search results are biased without knowing the true attention distribution function for users of the corresponding service. Therefore, our work should *not* be seen as an audit study, but rather a showcasing of a metric that can be used by the operators or internal auditors of these services to ensure fair delivery of results.

The remainder of the paper is organized as follows. In Section 2 we introduce related work on fairness in ranked lists and auditing of ranking algorithms. In Section 3 we provide a detailed description of the mechanics of the proposed metric. In Section 4we explain how the parameters of our metric should be set and interpreted based on context of its use. In Section 5 we analyze three case studies: a hiring service, a dating service, and Google search. In Section 6 we explain the limitations of our approach. Section 7 suggests the directions for further research, and Section 8 concludes the article.

2 BACKGROUND

In Pursuit of Fairness. As use of large, observational datasets has proliferated, so have concerns that systems leveraging this data may have a negative impact on people. The machine learning community has mapped the legal concepts of *disparate treatment* and *disparate impact* to *direct* and *indirect discrimination* by algorithms, respectively [5, 22, 43]. Zafar et al. introduced the concept of *disparate mistreatment* to refer to situations where false positives and negatives are not equally distributed across subpopulations [52, 53]. Such situations have been shown to occur e.g., in pre-trial assessments [1] and academic performance predictors [49].

While direct discrimination can be corrected by removing the protected attributes from the data, indirect discrimination is more challenging to address. Dwork et al. proposed two potential objectives for mitigating indirect discrimination: under *individual fairness*, similar people should be treated similarly by the algorithm, while under *group fairness* subpopulations should be treated equivalently to the whole population [14]. There is a large and growing literature on how to achieve these objectives in machine learning-based classifiers [5–7, 14, 17, 22, 28–33, 39, 55].

Fairness in Ranking. Achieving fairness in IR systems has received less attention in the academic community. One challenge is that research on fair classification does not necessarily generalize to the ranking context. A second challenge is accounting for *order effects* [41], i.e., the well-established tendency of human beings to pay more attention to items at the top[1] of a ranked list.

A few methods for generating group-fair search results have been proposed. Zehlike et al. leverage randomization by positing that a given ranked output is fair if it could have been generated by a random Bernoulli process [54]. Celis et al. propose a more general approach allowing the user to specify the fairness constraints [8]. Unfortunately, neither of these take user attention into account: their methods do not distinguish between different orderings of a set as long as a minimum fraction of items from the minority class are presented at each rank. In contrast, Singh and Joachims argue that even if the ranking itself is unbiased, small differences in placement may lead to large discrepancies of attention [50]. Biega et al. point out that any single ranking of similarly relevant items is individually unfair because of the uneven distribution of attention [3]. Therefore, they propose achieving individual equity (attention corresponding to relevance) within a certain number of realizations by systematic reshuffling of the list.

Algorithm-in-the loop Approach. Most of the work we discussed so far focus on measuring or correcting the algorithm without explicitly involving its users. In contrast, Green and Chen emphasize the need for considering the whole sociotechnical system [20]; they show that rather than focusing solely on the bias in an automated risk-assessment system, one needs to include contextual information on the system is actually used by judges and how it affects their decisions.

Auditing Search Engines. There is a growing body of work from the algorithm auditing [48] community that aims to measure whether real-world search engines are fair and unbiased. Kay et al. found that Google Image Search presented results that were stereotypically gendered [34], while Hannak et al. and Chen et al. showed that search engines on employment websites were not group fair with respect to race and gender [9, 23]. Audit studies have also examined the political partisanship of search results from Twitter and Google [12, 38, 46].

An open challenge for the auditing community is selecting appropriate metrics for assessing whether search engine results are group fair. For example, Kay et al. only looked at simple metrics like average representation that fail to take order effects into account [34]. Other audits have used group representation in top K ranks [24], logarithmic discounting [9, 23, 51] and linear normalization by rank [38, 46] to model the decay of attention. In this work, we argue that these ad hoc methods do not accurately model users'

[1] We use "top" and "high" to refer to the numerically lowest ranks in lists, e.g., rank one, in keeping with the norms of the IR literature [10, 27].

attention, and may lead to incorrect conclusions about (un)fairness of IR systems.

3 METHODS

In this section, we introduce a novel metric for measuring group fairness in search results. This metric, the Viable-Λ test, combines existing research from algorithmic auditing, IR, and human-computer interaction to address the human factors inherent in this problem.

3.1 Overview

Suppose we are given an ordered list of search results $R = [r_1, \ldots, r_n]^T$. Our goal is to measure the representation of some target property p that is shared by each $r_i \in R$. The metric proposed in this paper requires the auditor to specify the following five elements: (1) an *alignment vector* L_R, (2) an *attentional weight vector* W_R, (3) a *population estimator* $\hat{p}$, (4) a *distance metric* d, and (5) a *maximum allowable distance* $\delta_{\max}$.

Below we briefly introduce these elements and explain their role in the Viable-Λ Test. In Section 4 we explain the design choices behind each element and how to allocate them appropriately in practice.

The **alignment vector** L_R is a vector of probability distributions $[l_1, \ldots, l_n]^T$ that describe the group membership (or *alignment*) of r_i with respect to the target property p. The subscript $_R$ indicates that L_R has the same length as R, and that each $l_i \in L_R$ corresponds to the alignment of each respective $r_i \in R$. l_i can be either discrete or continuous.

The **attentional weight vector** W_R is a probability vector $[\omega(r_1), \ldots, \omega(r_n)]^T$ that models the relative user attention allocated to each $r_i \in R$. While it is difficult to determine the exact distribution of W_R, we can make assumptions about its *shape*. Formally, $W_R \sim P(\Lambda)$ where P is a family of n-truncated discrete probability distributions with an unknown set of true parameters Λ.

We calculate the expected cumulative exposure E_R of group representation in R by taking the dot product of L_R and W_R:

$$E_R = L_R^T \cdot W_R \tag{1}$$

Note that E_R is a probability distribution with the same domain as each $l_i \in L_R$ (i.e., a distribution describing the target property p).

The **population estimator** $\hat{p}$ is a probability distribution that estimates the true demographics of p. For reasons described in § 4.3, the following formula is often a well-motivated choice:

$$\hat{p} \stackrel{\text{def}}{=} \bar{L}_R = \sum_{r_i \in R} \frac{l(r_i)}{n} \tag{2}$$

where $\hat{p}$ has the same domain as l_i.

The **distance metric** d quantifies the statistical difference between the probability distributions E_R and $\hat{p}$.

The **maximum acceptable distance** $\delta_{\max}$ is the threshold that separates group fair from unfair search results. It is chosen in a context-dependent manner in conjunction with d. As we describe in § 4.4, d is essentially a statistical significance test, and δ_{min} is the test statistic threshold. Both components draw inspiration from the principles of traditional sampling statistics.

3.2 The Viable-Λ Test for Representational Parity

Given $L_R, \hat{p}, d$, and $\delta_{\max}$, we define group exposure as $E_R = L_R^T \cdot W_R$. Assume that $W_R \sim P(\lambda_1, \ldots, \lambda_m)$ where P is an n-binned discrete probability distribution with an unknown set of true parameters $\Lambda = (\lambda_1, ..., \lambda_m)$ in some domain space D with known bounds. Then R is unfair if:

(1) $\nexists \lambda \in D$ such that $d(e_r, \hat{p}) < \delta_{\max}$, i.e., there is no way to parameterize the attention distribution such that representational parity is attained; and
(2) For Λ satisfying the above condition, $W(\Lambda)$ matches reasonable expectations and data about true user behavior.

4 DESIGN CHOICES

This section describes the key components of Viable-Λ and the motivating decisions behind their design.

4.1 The Alignment Vector L_R

Alignment as a Probability Distribution. Given an ordered list of result items $R = [r_1, \ldots, r_n]^T$, we use an *alignment function* ι to map each result r_i to a probability distribution describing its *alignment* in terms of the target property p (e.g., race, gender, political alignment, etc.). As a motivating example, consider a resume search engine [9] that returns a list of job candidates R in response to a query (e.g., "software engineer"). Suppose we want to model the gender alignment p of these results; namely, p is a discrete probability distribution across three classes: 'Male,' 'Female,' and 'Unknown.' If the search results *explicitly* display the gender of each candidate r_i, then we can define alignment as follows:

$$l_i \stackrel{\text{def}}{=} \begin{cases} \{\text{Female: 1.0, Male: 0.0, Unknown: 0.0}\} & \text{if } r_i \text{ is female} \\ \{\text{Female: 0.0, Male: 1.0, Unknown: 0.0}\} & \text{if } r_i \text{ is male} \\ \{\text{Female: 0.0, Male: 0.0, Unknown: 1.0}\} & \text{otherwise} \end{cases} \tag{3}$$

In realistic scenarios, defining l_i may not be so trivial. Continuing our example, resume search engines typically do not explicitly state the gender or race of job candidates. However, recruiters (and auditors) may still be able to infer them using other information, like a user's profile picture, given name, etc. Most existing approaches to measuring fairness assume that all l_i are explicitly known and binary (canonically between protected and non-protected classes S and S^C) [14, 51]. However, this assumption has not generalized to empirical studies. By using a probability distribution rather than a binary indicator, we cover the cases where the class assignment is ambiguous, where there are multiple classes, or there are multiple realizations of the ranking.

4.2 Attentional Weights W_R

W_R encapsulates the well-documented fact that search engine users do not treat all search results equally [10, 19, 21, 45]. For example, the first result on Google Search is estimated to receive approximately 30% of all clicks, and the results on the first page account for approximately 90% of clicks [35]. This observation remains true to a large extent even if the order of the search results is inverted [35].

Modelling Attentional Weights ω. W_R is a probability vector that models the amount of user attention that each $r_i \in R$ receives. $W_R = [w_1, \ldots, w_n]^T$, where $w_i = \omega(r_i) \in [0, 1]$ is the output of a weighting function ω applied to r_i. ω is influenced by (1) the user interface design of the search engine (for example pagination or highlighted results) and (2) the use context of the search engine. For an intuition on the latter, consider that a user may only view several web search results before finding an acceptable answer to their query, whereas that same user might view dozens of resumes from a resume search engine if they are tasked with hiring a new employee.

Fitting W_R from Empirical Data. A tempting way to approximate W_R is to use empirical data such as organic Click Volume (CV) and Click-Through Rate (CTR), widely used in digital marketing and by search engine proprietors. Unfortunately, there are three serious impediments to using an empirically derived W_R. First, click data is often proprietary, and thus unavailable to an external auditor. Second, click data is only an approximation for what users see in search results. Truly measuring user attention might require expensive eye tracking studies [10, 19, 21, 45]. Third, as we noted above, the way users distribute their attention over search results depends on website design and use context. Thus, although eye-tracking studies are available for sites like Google Search [10, 19, 21, 45], the data may not be applicable to other search engines.

Potential Choices for W_R. In algorithmic auditing, we need a way to measure user attention without (1) having access to the vendor's analytic data, and (2) having to conduct multiple eye tracking studies. We partially address this issue by making assumptions about the shape of W_R. Formally, we assume that $W_R \sim P(\Lambda)$ for some discrete truncated probability distribution P. W_R should meet two criteria:

(1) W_R is an n-truncated discrete probability distribution.
(2) Higher-ranked results receive substantially more attention than lower-ranked ones; i.e., for reasonably large n, $\omega(r_1) \gg \omega(r_i)$ as $i \rightarrow n$.

Some families of distributions that fit the above criteria are presented in Figure 1 and include Truncated Geometric Distribution, Truncated Log-series Distribution, and Truncated Discrete Pareto distribution [37]. For the remainder of the paper we use the truncated geometric distribution but in an actual measurement scenario, another choice may be more appropriate.

The Case of Small n. The above distributions are applicable when n is reasonably large, i.e., well beyond the human attention span. When n is small, we can draw inspiration from psychology: Miller's Law famously states that the average human's working memory can hold roughly 7 objects at a time [40]. When n is sufficiently small, we expect the user to read all of the results.[2] Formally, when n is small, $W_R \dot{\sim} unif\{0, n\}$. Note that when choosing

[2]This may occur in practice when a user queries for obscure information or the vendor lacks data relevant to the search query

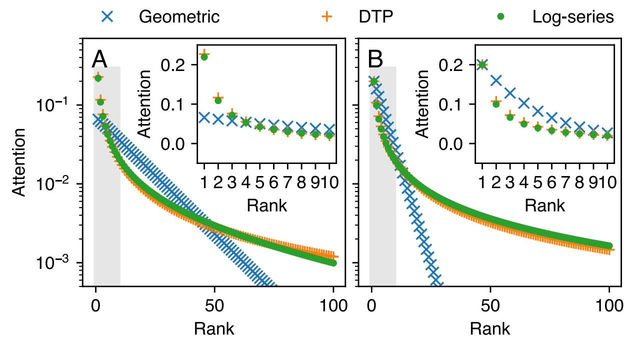

Figure 1: Comparison of Geometric, Log-series, and Discrete Truncated Pareto (TDP) on a list with length $n = 100$. In (A), the parameters are set so that $E[W_R]$ (i.e., the mean number of results seen) is 15. In (B), the parameters are chosen so that $\omega(r_1) = 0.2$.

$W_R \stackrel{\text{def}}{=} unif\{0, n\}$ and $\hat{p} \stackrel{\text{def}}{=} \bar{L}_R$, E_R is precisely equal to $\hat{p}$; therefore all short lists exhibit approximate representational parity.

The Problems of Inverse Log Scaling. We notably left out inverse log:

$$W_R = normalized\left(\left[\frac{1}{\ln(2)}, \ldots, \frac{1}{\ln(n+1)}\right]^T\right). \quad (4)$$

Inverse logarithmic scaling is commonly used in IR relevancy metrics such as nDCG, and also appears in some metrics of fairness in ranking [51]. However, this choice of W_R has two major flaws: *first*, it decays at a very slow rate and does not meet the relative convergence requirements described above in criteria (2). Even when $n = 1000$, $\frac{\omega(r_1)}{\omega(r_{1000})}$ is approximately 9.964, which implies that the last 10 search results are, in aggregate, as influential as the first result. *Second*, since there are no parameters for this choice of W_R, it incorrectly assumes that user behavior is the static across all platforms and for all search queries.[3]

4.3 Estimating Population Demographics $\hat{p}$

Our metric requires the specification of an estimator $\hat{p}$. This section describes how to choose $\hat{p}$ so that it serves as a reasonable prior for statistical parity.

Implicit Estimators for p. A tempting proposal would be to choose an *implicit estimator* for p based on intuition or observational data. This is plausible when the search query Q is relatively simple. For example, consider the case when the user is querying a resume search engine for a list of certified nurses: recent data shows that the US national gender ratio for this profession is approximately 9.5:1 female to male [44], thus we expect fair rankings for this query to reflect this. For more complex queries, however, demographic data may unavailable. If the user instead queries for Android Developers in Greenville, AL with at least three years of experience, we lack an external data reference. Thus, we are unable to choose nor justify an implicit estimator $\hat{p}$.

[3]The base of the logarithm is irrelevant; after normalization, they all evaluate to the same vector of values.

Furthermore, Q itself cannot be directly examined in many IR systems. Online services rank their feeds using proprietary algorithms that rely on personalization. From an auditor's standpoint, a generalized fairness metric must have the ability to estimate p regardless of whether or not we have access to Q.

Choosing $\hat{p} \stackrel{\text{def}}{=} \bar{L_R}$. An IR-motivated alternative to circumvent these issues is to determine $\hat{p}$ based on the vendor's data R. Suppose that within an IR system, the vendor evaluates Q then filters their corpus to yield a subset of results R in which all $r_i \in R$ meet some relevancy threshold. Then we can calculate $\hat{p}$ as defined in Equation 2, i.e., an equally-weighted sum $\forall l(r_i) \in L_R$. Since this calculation relies on the vendor's data R, the validity of $\hat{p}$ is dependent on the integrity of the vendor's data. Thus, it is imperative to first audit the vendor's *data curation* for sampling bias and *result scoring* for direct discrimination. A hypothetical example of unfair data curation would be a job site that refuses to add female software engineering candidates to their database. Unfair query evaluation could occur when a vendor has female candidates in their database but fails to show them when a recruiter queries for "software engineers" (i.e., being a woman directly impacts the relevancy score). Fortunately, this style of audit is often straightforward: Chen et al. tested for direct discrimination in scoring by posting two identical resumes that varied only by gender, and showing that they appeared at directly adjacent positions in search results [9]. If a preliminary audit finds that any of these assumptions are jeopardized, we can immediately deem R as unfair without needing to calculate $\hat{p}$ or other components of Viable-Λ.

Consequences. In choosing $\hat{p} \stackrel{\text{def}}{=} \bar{L_R}$, our prior for statistical parity is fitted to the vendor's data. Thus, if the vendor's knowledge of p is lacking or biased (for example, unknowingly exhibits sampling bias), our estimate of p will be as well. In this situation, Viable-Λ becomes a metric of how well the vendor's ranking *represents their own knowledge of* p.

4.4 Distance Metric d and the $\delta_{\max}$ Threshold

We use d and $\delta_{\max}$ to delineate an *acceptance region* around $\hat{p}$.

Distance Metric d. d is a statistical distance metric that quantifies the difference between the two probability distributions E_R and $\hat{p}$. The choice of d follows naturally based on the domain of E_R and $\hat{p}$. In Section § 5, we demonstrate the use of Z-approximation for the binomial test statistic when E_R and $\hat{p}$ are both binomial distributions.

Maximum Acceptable Distance $\delta_{\max}$. $\delta_{\max}$ establishes an acceptable range of values around $\hat{p}$ in which we can safely assume that group fairness is preserved. In general, $\delta_{\max}$ is the *test statistic* threshold to the statistical significance test d. In conjunction, d and δ_{min} constitute a statistical significance test with h_0: $E_R \sim \hat{p}$, i.e., E_R has the same sampling distribution as $\hat{p}$. A well-constructed $\delta_{\max}$ takes the following factors into consideration:

(1) *Statistically significant difference between E_R and $\hat{p}$ (size of R).* Suppose that we perform a search to yield results R with length $n = 100$, for some binomial alignment property. One interpretation of this scenario is that our result R is a sampling distribution of size $n = 100$ from a universe U of *all* relevant candidates. We can formalize this by saying that R is a simple random sample of U, where $n = 100 \ll |U|$.[4] U has a true binomial parameter p. Thus, we can employ statistical inference to construct a confidence interval for p. For a binomial distribution, the maximum likelihood estimate (MLE) of p yields $\hat{p} = \bar{L_R}$ with standard error $s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{100}}$. By specifying a confidence level (e.g., 95% corresponding to $z = 1.96$), we can build a confidence interval for p. Thus, we cannot reject H_0 if E_R falls inside this range of values. For categorical p, we can check for statistically significant difference between $\hat{p}$ and E_R using the appropriate test statistics (e.g., Pearson's χ^2 Test for Independence).

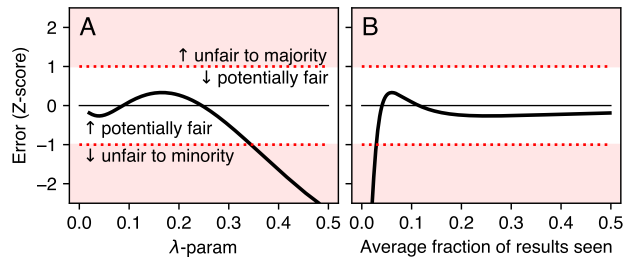

Figure 2: We find the range of the λ-parameter (and, thus, the corresponding average fraction of seen results (B)) for which the result list is potentially fair. If the attention given to a group is more than $1Z$ away from its representation in the population, we deem the results unfair.

(2) *Number of Search Queries Made.* This factor is applicable only in the case where the the vendor randomizes their search results. We can model $l(r_i)$ as the sampling distribution of S across k search realizations. We calculate the standard error in the same manner as above using k in lieu of n. An important distinction is that the previous standard error calculated with n represents uncertainty in p, while this standard error calculated with k represents uncertainty at each rank in $l(r_i)$. In the case where the vendor does not randomize their rankings or when the auditor can make an unlimited number of queries, this factor is irrelevant.[5]

Plotting $d(E_R, \hat{p})$. In the case where W_R has one parameter, we can plot $d(E_R, \hat{p})$ as a function of λ. An example of this is illustrated in Figure 2.

4.5 Restricting the Domain of the Parameter Space D

In our model, we accept the null hypothesis that a ranking R is fair if there exists a set of parameters Λ within the parameter space D which brings us within an acceptable range of parity. However,

[4] Frequentist statistics relies on the assumption that $n \ll |U|$, which reflects the fact that it is often impractical or expensive to collect a census over the entirety of U.

[5] With an unlimited number of queries, we can arbitrarily reduce the standard error to any $\epsilon > 0$. This is not possible with the confidence interval described in the previous section since it is the vendor—not the auditor—who determines the size of R. In the case where the vendor is able to increase n (i.e., by acquiring more data), then both the auditor and the vendor will be able to make more accurate estimates of p.

we are susceptible to type-II errors since not all $\Lambda \in D$ match realistic expectations about user behavior. Thus, by truncating our parameter space D, we can increase the *power* (the probability of *not* making a type-II error) of our hypothesis test.

Context-based Assumptions. To illustrate this, consider a ranked list with length $n = 100$. Suppose that we have deduced that $W_R \sim Geom(\lambda, 100)$. The mathematically-permitted domain of the success parameter is $\lambda \in (0, 1)$. However, if we know for certain that $E[W_R]$ is between 2 and 50 (i.e., the average user views between 2 and 50 profiles), thus we can restrict our domain space to $\lambda \in (0.02, 0.5)$.[6]

Empirically-informed Truncations. Furthermore, we can also set up a small-scale experiment to estimate $E[W_R]$. Suppose that we have a small group of users with $N = 16$. Assume that the average number of results viewed is approximately normal with $\bar{x} = 27$ and standard error $s = 12$. Using the maximum likelihood estimator (MLE) we can construct a 95% confidence interval for $\bar{x}$ between approximately 21 and 33. Since $E[W_R] \approx 1/\lambda$, we can say that the parameter λ lies between 1/21 and 1/33, corresponding to the interval $(0.03, 0.47)$[7].

As we can see in the above example, it is possible to leverage data from small-scale experiments to estimate the likelihood of λ. This offers us an avenue to improve the statistical power of the Viable-Λ test.[8]

4.6 Generating Fair Ranked Lists

Other researchers have focused on creating fair ranked lists and their cost in terms of individual fairness in depth [3, 8, 50, 54]. Here, we provide a few examples of ranked lists that are fair for fixed λ. Figure 3 shows four best-attempt fair lists with varying class imbalance (1:10 in top row, 5:10 in bottom row; minority class A with light blue, majority class B with dark blue) and attention distributions ($\lambda = 0.1, 0.5$ from left column to right). In the top row, we see that as the distribution becomes steeper, the one minority sample is placed higher to receive proportional attention. In the bottom row, the flattest distribution requires a list of results were both classes are quite mixed. However, in the case of the steepest distribution, all elements of class B are placed in a block from rank two on to match the attention already given to class A by its representative at first rank.

5 CASE STUDIES

In this section, we apply the Viable-Λ Test to three different search engines: (1) gender fairness on a hiring site used by recruiters to search for candidates, (2) racial fairness on a dating site, and (3) political fairness on a search engine. As a disclaimer, we are only using these results to demonstrate a use of our metric on real-world data. We do not *not* claim that any of these are biased or unfair.

[6]The expected value of $Geom(\lambda)$ is $1/\lambda$. Since W_R is truncated, $E[W_R]$ asymptotically approaches this value for large n.

[7]We can use the confidence interval of μ to restrict the parameter space of other families of probability distributions by replacing $1/\lambda$ with the corresponding equation for $E(W_R)$.

[8]This approach is different from the eye-tracking studies mentioned in § 4.2. Instead of using eye-tracking heatmaps to construct the entirety of W_R, use a singular dimension of the data (namely $E[W_R]$) to inform a reasonable range of values for Λ.

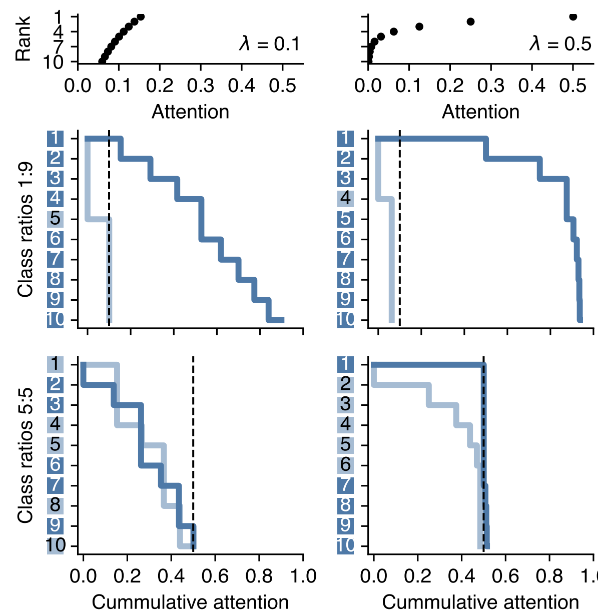

Figure 3: We generate rankings that satisfy our definition of attention fairness. Note that depending on the attention distribution function, fair rankings with the same proportions of classes differ.

Ethics of Data Collection. While conducting our measurements we were considerate both of the services we collected the data from and of people who this data represents. All collected data is available to any person with an account on the corresponding service. We did not interact with any users of these services as part of the collection. Additionally, we minimized any impact on the operations of the services by using a low query intensity (at most one query every 30 seconds). Finally, we adhered to the usage quotas of Face++.[9]

Assumptions In our case studies, we assume that $W_R \sim Geom(\lambda)$. As discussed in Section § 4.2, $Geom(\lambda)$ meets the desired properties of an attention distribution, and thus serves as a reasonable starting approximate for the true W_R.

We used brute-force optimization over our one-dimensional parameter space D to check if $\exists \lambda \in D$ such that $d(E_R, \hat{p}) < \delta_{\max}$.

5.1 Gender in Hiring

Our first cast study examines the ranking of job candidates on a resume search engine. Data was collected and made available to us by Chen et al. [9]. Using the recruiter's interface to the service, the authors queried 35 different search terms (such as "bartender", "electrical engineer", "laborer", "pharmacist", and "software engineer") in 20 US cities, resulting in 692 non-empty result lists. Of these results, we look at the 412 with length $n \geq 100$.

Modeling Gender Alignment L_R Chen et al. determined the gender of each candidate based on their given name (see [9] for

[9]https://www.faceplusplus.com

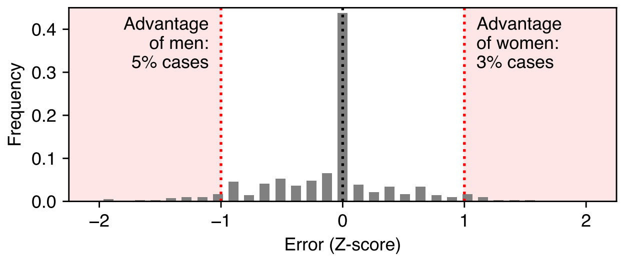

Figure 4: Best case scenario of error distribution among location-job pairs in the resume search engine. For each location-job query we attempt to find a Λ for which the ranking could be considered fair. We fail to do so in 8% of cases.

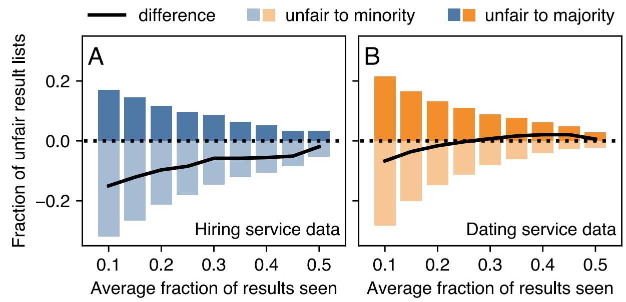

Figure 5: The more results a user sees, the higher the chance that the ranking can be considered fair (in extreme cases when all results are seen, the ranking is always fair). A) Regardless of λ, women are underrepresented more often than men in the job-location searches. B) Depending on the true value of λ, Black profiles may be underrepresented more frequently than non-Black profiles.

details). The gender estimation is a real number describing the probability of a candidate being male between 0 (female) and 1 (male). The original authors dichotomized the data, assuming 'male' for probabilities >0.8, 'female' for probabilities <0.2, and omitted other profiles. These constituted 8% of all candidates. In this study, we use the gender alignment probabilities directly. We do neither omit ambiguous candidates, nor project alignments to their most-likely class.

Determining $\hat{p}$ In this study, the true population demographics for each query are unknown. As per Section § 4.3, we cannot choose an *implicit estimator* and instead leverage the vendor's data by choosing $\hat{p} \overset{\text{def}}{=} \bar{L}_R$ in each query.

Evaluating d and δ_{max}. $\hat{p}$ and E_R are both binomial distributions representing gender alignment. To check for statistically significant difference between them, we use the Z-test approximation for the binomial test. This $d(E_R, \hat{p}$ is our test statistic. We deem R_i unfair if $\nexists \lambda$ such that this test statistic is less than 1Z. This represents a 68% confidence that fair representation is impossible.

Viable-Λ in Hiring. In Figure 4, we run Viable-Λ for each list and plot the minimum attainable $d(E_R, \hat{p}$ for each R_i. About 92% of rankings can be considered fair at the 1Z threshold. Still, 6% of these lists appear biased against women and 3% appear unfair to men regardless of the distribution function. All rankings are deemed fair at the 2Z threshold

In Figure 5A, we demonstrate the effect of λ on $d(E_R, \hat{p})$. As per Section § 4.5, we assume within reason that $E[W_R] \in [.1n, .5n]$ in the context of this search engine. The horizontal axis depicts nine λ-parameter choices corresponding to a user viewing between 10% and 50% of all results on average. For all sampled values of λ, rankings under-represent women more often than men.

While the gender ratio is balanced in the dataset, the gender proportions vary widely between queries. We calculated the $\hat{p} = \bar{L}$ estimate for p for each query separately. Although the vast majority of rankings passed the Viable-Λ test, rankings tended to under-represent women more frequently for most values of λ.

5.2 Race in Dating

We used an author's personal account to query an online dating service's API. The script ran once every 15 minutes over the course of one week, collecting 672 lists each with length $n = 100$. We determined that each of these lists sampled from a pool 4,407 unique profiles. Even though we were running the same query repeatedly, we observed significant *shuffling* and *churn* in the results, with some profiles being replaced more often than others.

Modelling Race Alignment. We used Face++ to infer the race of each person from their profile picture. Because the data is only used for illustration purposes, the precision of the race detection is not crucial.[10] In 813 (18.4%) profiles, Face++ detected > 1 face; among the identified profiles 2,411 (54.7%) were classified as white, 573 (13.0%) as Black, 534 (12.1%) as Asian, and 76 (1.7%) as Indian. For each profile photo, Face++ returns only the *most likely* race, not its certainty. One potential option is to model alignment as a categorical probability distribution among the five aforementioned classes. In doing so, we would use the χ^2 Independence Test to compute the statistical difference between E_R and $\hat{p}$. This would allow us measure the representation of all classes simultaneously. To simplify our illustrations, we project our alignment into a binomial distribution - i.e., Black vs. non-Black profiles.

Determining $\hat{p}$, d, and δ_{max}. In this example, we ran the same query across each of our searches, yielding 4,407 unique results that meet the vendor's relevancy threshold. In our first approach, we chose $\hat{p} \overset{\text{def}}{=} \bar{L}$ on the set of all unique results. In doing so, we can build a 95% confidence interval for p; namely $\hat{p}$ is a *binomial sampling distribution* with $n = 4407$. Since we have a large sample size, our test statistic has a very small error threshold. We use the Z-test approximate for binomial test as our distance metric d again. Using

[10]Face++ inferred a different gender than self-reported for 13% profiles identified as white, 11% profiles identified as Asian, and 8% of profiles identified as Black. Previous studies have shown high gender misclassification rates for Black women in particular [4]).

this definition of $\hat{p}$, Viable-Λ showed that rankings unequivocally over-represented Black profiles. While Black profiles constituted only 13% of all unique profiles, the average ranking displayed 16% Black profiles.

Potential for Correcting Societal Bias. Black users have been shown to be disadvantaged in online dating [47]. Thus, it is possible that these profiles are over-sampled by this dating service to compensate for their lower click-through rates. Suppose that we want to measure how well rankings match the vendors' *manipulated* population demographics. Then in this case, p is the true percentage of Black profiles displayed by the vendor's ranking algorithm. At each rank k, we have a binomial sampling distribution of Black profiles with size $n = 672$ (i.e. the total number of searches). Thus, we can continue using the Z-test approximation for binomials as our distance metric.

Evaluating R_i Individually. We begin by evaluating each ranked list R_i individually. First, we find that for all values of the λ-parameter, the majority of lists can be considered fair, see Figure 5B. For the steepest distributions (smallest value of the λ-parameter) more result lists are unfair towards the Black users than towards non-Black users; when we assume an attention distribution function such that the users see on average 30% to 45% of results, there are more realizations in which Black users get more attention than proportional to their representation, and the situation equalizes for the least steep distributions. This effect is caused by the fact that white profiles appear on the first position in 54% of realizations, even though they only constitute 48% of the observed population on average, but for the rest of the high ranks, Black profiles are presented more often than the population estimator would indicate.

Evaluating R_i in Aggregate. Next, we evaluate the fairness of several realizations sampled in aggregate. A similar notion was proposed by Biega et al. [3]. As shown in Figure 6, even if each single ranking realization is unfair, the aggregate of multiple unfair rankings can be considered fair. Our metric can capture this because the alignment vector is not binary. We note that the more realizations included in the aggregate, the higher the fraction of fair aggregates, regardless of the assumed attention distribution function. Still, the steeper the function, the more runs are necessary to achieve a fair aggregate. In a hypothetical case of multiple rankings generated by a random (and, thus, unbiased [51, 54]) ranker, each rank will converge to contain a proportional representation of classes, and the ranking will be fair regardless of the assumed attention distribution function.

Summary. This case study highlights several important findings. The perception of the existence and even the direction of bias depends on the attention distribution. Furthermore, the bias can be corrected over time by reshuffling the results. Finally, our metric can accommodate population estimates that are different from the underlying populations for example to correct for societal biases.

5.3 Political Bias in Web Search

Our third case study differs from the previous two in that there is no protected group; instead, the *alignment* l_i is a proxy for political leaning of each item in the ranked list. The dataset we use was collected and made available to us by Robertson et al. [46]. There are two elements to the dataset: (1) search results and (2) partisan audience bias estimates. The first part comprises the first pages of results to 1,443 different web search queries. The second part maps 19,022 domains that appear in the search results to the bias scores on a liberal/conservative axis. Based on the tweets from registered voters, it assigns a number from -1 (only Democratic voters share content from that domain) through 0 (Democratic and Republican voters are equally likely to share content from this domain) to 1 (shared only by Republican voters). For example, `blacklivesmatter.com` scores -0.94, `en.wikipedia.com` scores -0.22, `dhs.gov` scores -0.01, `youtube.com` scores 0.13, and `catholics4trump.com` scores 0.98. Note that the score is assigned to a domain, not a specific webpage.

In this case study, we measure whether the aggregated partisan bias of search results is cancelled, given the attention distribution. Among the 1,443 searches in the dataset, we select three examples that best highlight the importance of considering attention distribution in the audit. In the interest of brevity, we omit the distance metric steps of Viable-Λ. We do, however, report the difference between the source biases weighted by the attention per rank and 0. Positive values indicate that a result list leans conservative; negative values indicate liberal lean.

Figure 7 presents the search results to three queries: "financial regulation", "obamacare continue", and "medicare reform", along with perceived partisan bias of each of these lists for different attention distribution functions. Note that in the first panel, small values of the λ-param correspond to a flatter distribution; as the λ-param grows, more attention is given to top results. The search engine might return multiple items from the same domain, therefore some domains appear multiple times in the list (for example `healthcare.gov` in Figure 7b).

The results in Figure 7 for "financial regulation" are neutral regardless of how attention is distributed since most of the results are apolitical. The results for "obamacare continue", on the other hand, lean republican: they feature three items from highly conservative sources. Still, because the first result is liberal-leaning—and among the first four results, three lean liberal—the result list will appear

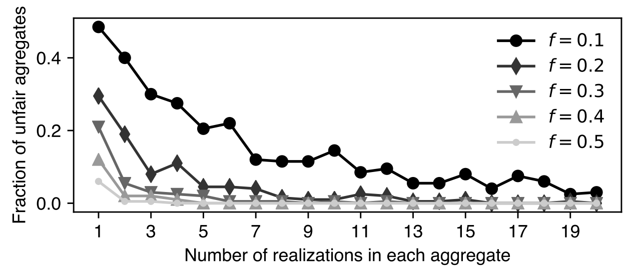

Figure 6: The investigated dating service returns a different set of results to the same query issued by the same person. As a consequence, while each ranking might be unfair, the results in aggregate are fair. The number of runs necessary to expect a fair ranking depends on the chosen distribution parameters.

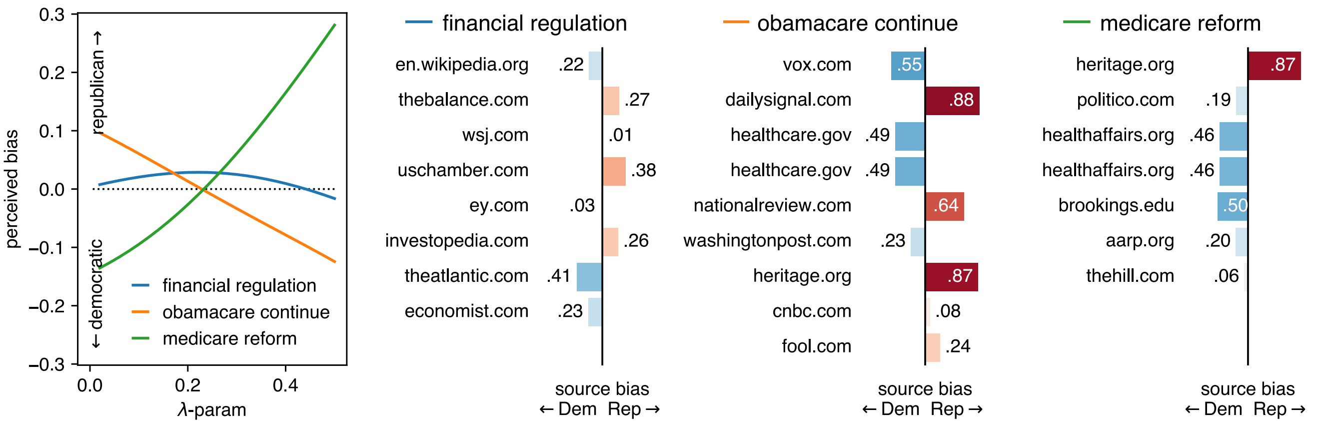

Figure 7: The perception of bias may depend on how the users of a service distribute their attention to the presented results. The neutral results for "financial regulation" form a list that appears neutral regardless of usage patterns. However, "obamacare continue" and "medicare reform" might appear partisan in either direction, depending on the attention distribution function.

to be liberal-leaning overall if the attention distribution function is steeper. Finally, the results for "medicare reform" are almost exclusively from democratic-leaning sources. However, because the top result comes from a strongly conservative outlet, the list exhibits strong conservative-bias if the attention distribution function is steep. Thus, these examples illustrate how the shape of the attention distribution function can dramatically alter conclusions about the fairness/bias of ranked outputs.

6 LIMITATIONS

The Shape of W_R. While our framework allows for arbitrary families of W_R , we only considered the truncated geometric distribution in our examples. The results of the Viable-Λ test rely on an accurate model of human attention; thus further research into human perception and the quantification of the SEME would improve the basis of this metric.

Additionally, singlely-parametered W_R may not be sufficient in modelling expected attention. One important parameter is pagination; researchers have found that the CTR of the last result on a page getting more clicks than the pre-to-last [42], and each page introducing a disproportional drop-off of attention [26]. Furthermore, modern search engines often add variation into their search results. For example, Google search might display the actual content the user is seeking directly in the result page, or group results by type (i.e., "Sponsored", "Video", etc.).

Population Estimators. In this work, we derived the population estimators $\hat{p}$ directly from R using the $\bar{L}$ estimator. Hence, we assumed that the items the vendor chooses to show in the top k results are a proportional representation of all N potential results. It is likely, however, especially with large N, that they are not. For example, a real-world candidate ranking system employed by Amazon was shown to systematically rank women lower than men [11]. If we audited it and only had access to the top 100 results out of 1000, we would be likely to assume a $\hat{p}$ that underestimates the fraction of female candidates.

7 DISCUSSION

Studies have shown that swapping search results can cause significant changes in users' eye-gaze and click patterns. For example, after reshuffling Google Search results, unwitting users still tend to click on the top results, but some do shift more attention to lower ranks [35]. While we can use our metric to check if Λ exists such that representational parity is achieved (or—from the operator's point of view—verify that the results are fair given the known Λ), it is not guaranteed that altering results based on the metric's measurements will create fair rankings, since user behavior may change in uncertain ways. One possible way to create fair rankings is by means of a continuous, iterative process: the operator reshuffles the results to achieve parity under a measured Λ, users potentially change their behavior as a response, the operator updates the Λ estimate, and so on.

8 CONCLUSION

In this work we introduced a novel metric of group fairness in ranked lists, tying the measurement to the consumers' attention distribution. We showed how our approach could be used by auditors on three real world examples. Our results highlight the need for modelling attention specifically for the audited service: depending on the attention distribution function, the same list of results can appear biased both in favor and against the protected group. All code will be made publicly available upon publication.

ACKNOWLEDGMENTS

This research was supported in part by NSF grant IIS-1553088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016).

[2] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Cal. L. Rev.* 104 (2016), 671.

[3] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *Proc. of SIGIR*. ACM.

[4] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proc. of FAT**. 77–91.

[5] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building Classifiers with Independency Constraints. In *Proc. of ICDM Workshops*. IEEE.

[6] Toon Calders, Asad Karim, Faisal Kamiran, Wesam Ali, and Xiangliang Zhang. 2013. Controlling Attribute Effect In Linear Regression. In *Proc. of ICDM*. IEEE.

[7] Toon Calders and Sicco Verwer. 2010. Three Naive Bayes Approaches For Discrimination-free Classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.

[8] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2017. Ranking with Fairness Constraints. *CoRR* abs/1704.06840 (2017). http://arxiv.org/abs/1704.06840

[9] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the Impact of Gender on Rank in Resume Search Engines. In *Proc. of CHI*. ACM.

[10] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An Experimental Comparison of Click Position-bias Models. In *Proc. of WSDM*. ACM.

[11] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[12] Nicholas Diakopoulos, Daniel Trielli, Jennifer Stark, and Sean Mussenden. 2018. I Vote For—How Search Informs Our Choice of Candidate. In *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, M. Moore and D. Tambini (Eds.). 22.

[13] Dimitar Dimitrov, Philipp Singer, Florian Lemmerich, and Markus Strohmaier. 2017. What Makes a Link Successful on Wikipedia?. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 917–926. https://doi.org/10.1145/3038912.3052613

[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proc. of ITCS*. ACM, 214–226.

[15] Robert Epstein and Ronald E Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* 112, 33 (2015), E4512–E4521.

[16] Robert Epstein, Ronald E. Robertson, David Lazer, and Christo Wilson. 2017. Suppressing the Search Engine Manipulation Effect (SEME). *Proceedings of the ACM: Human-Computer Interaction* 1, 2 (November 2017).

[17] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proc. of KDD*. ACM.

[18] Google. 2016. Ads and analytics innovations for a mobile-first world. Google Blog. Accessed: 2018-10-29, https://www.blog.google/products/ads/ads-and-analytics-innovations-for-a-mobile-first-world/.

[19] Laura A Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. In *Proc. of SIGIR*. ACM.

[20] Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *ACM Conference on Fairness, Accountability, and Transparency*.

[21] Zhiwei Guan and Edward Cutrell. 2007. An eye tracking study of the effect of target rank on web search. In *Proc. of CHI*. ACM.

[22] Sara Hajian and Josep Domingo-Ferrer. 2013. A Methodology For Direct And Indirect Discrimination Prevention In Data Mining. *IEEE Transactions on Knowledge and Data Engineering* 25, 7 (2013), 1445–1459.

[23] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr.. In *Proc. of CSCW*. ACM.

[24] Eduardo Hargreaves, Claudio Agosti, Daniel Menasché, Giovanni Neglia, Alexandre Reiffers-Masson, and Eitan Altman. 2019. Fairness in online social network timelines: Measurements, models and mechanism design. *Performance Evaluation* 129 (2019), 15–39.

[25] Facebook Inc. 2018. https://investor.fb.com/investor-news/press-release-details/2018/Facebook-Reports-Second-Quarter-2018-Results/default.aspx.

[26] Chitika Insights. 2013. The value of Google result positioning. *Online Advertising Network* (2013).

[27] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased Learning-to-Rank with Biased Feedback. In *Proc. of WSDM*. ACM.

[28] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *Proc. of Conference on Computer, Control and Communication*.

[29] Faisal Kamiran and Toon Calders. 2010. Classification with No Discrimination by Preferential Sampling. In *Proc. of Machine Learning Conference of Belgium and The Netherlands*.

[30] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination Aware Decision Tree Learning. In *Proc. of ICDM*. IEEE.

[31] Faisal Kamiran, Asad Karim, and Xiangliang Zhang. 2012. Decision Theory for Discrimination-Aware Classification. In *Proc. of ICDM*. IEEE.

[32] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Proc. of ECML PKDD*.

[33] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware Learning through Regularization Approach. In *Proc. of ICDM Workshops*. IEEE.

[34] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proc. of CHI*. ACM.

[35] Mark T Keane, Maeve O'Brien, and Barry Smyth. 2008. Are people biased in their use of search engines? *Commun. ACM* 51, 2 (2008), 49–52.

[36] Pauline T. Kim. 2017. Data-Driven Discrimination at Work. *William & Mary Law Review* 58 (2017).

[37] Tomasz J Kozubowski, Anna K Panorska, and Matthew L Forister. 2015. A discrete truncated Pareto distribution. *Statistical Methodology* 26 (2015), 135–150.

[38] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna Gummadi, and Karrie Karahalios. 2017. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. In *Proc. of CSCW*. ACM.

[39] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. k-NN As an Implementation of Situation Testing for Discrimination Discovery and Prevention. In *Proc. of KDD*. ACM.

[40] George A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63, 2 (1956), 81––97.

[41] Bennet B Murdock. 1962. The serial position effect of free recall. *Journal of Experimental Psychology* 64, 5 (1962), 482–488.

[42] Jamie Murphy, Charles Hofacker, and Richard Mizerski. 2006. Primacy and Recency Effects on Clicking Behavior. *Journal of Computer-Mediated Communication* 11, 2 (2006), 522–535. https://doi.org/10.1111/j.1083-6101.2006.00025.x

[43] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware Data Mining. In *Proc. of KDD*. ACM.

[44] Emily Rappleye. 2015. Gender ratio of nurses across 50 states. https://www.beckershospitalreview.com/human-capital-and-risk/gender-ratio-of-nurses-across-50-states.html.

[45] Matthew Richardson. 2007. Predicting clicks: Estimating the click-through rate for new ads. In *Proc. of WWW*. IW3C2.

[46] Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM: Human-Computer Interaction* 2 (2018).

[47] Christian Rudder. 2014. *Dataclysm: Love, Sex, Race, and Identity–What Our Online Lives Tell Us about Our Offline Selves*. Crown.

[48] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. In *Proceedings of "Data and Discrimination: Converting Critical Concerns into Productive Inquiry", a preconference at the 64th Annual Meeting of the International Communication Association.*

[49] Piotr Sapiezynski, Valentin Kassarnig, Christo Wilson, Sune Lehmann, and Alan Mislove. 2017. Academic performance prediction in a gender-imbalanced environment. In *FATREC Workshop on Responsible Recommendation Proceedings.*

[50] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proc. of KDD*. ACM.

[51] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management.*

[52] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment: Disparate Impact: Learning Classification Without Disparate Mistreatment. In *Proc. of WWW*. IW3C2.

[53] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2015. Fairness Constraints: A Mechanism For Fair Classification. In *Proc. of FATML*.

[54] 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 CSCW*. ACM.

[55] Indre Zliobaite, Faisal Kamiran, and Toon Calders. 2011. Handling Conditional Discrimination. In *Proc. of ICDM*. IEEE.