

Equity of Attention: Amortizing Individual Fairness in Rankings

Asia J. Biega
Max Planck Institute for Informatics
Saarland Informatics Campus
jbiega@mpi-inf.mpg.de

Krishna P. Gummadi
MPI-SWS
Saarland Informatics Campus
gummadi@mpi-sws.org

Gerhard Weikum
Max Planck Institute for Informatics
Saarland Informatics Campus
weikum@mpi-inf.mpg.de

ABSTRACT

Rankings of people and items are at the heart of selection-making, match-making, and recommender systems, ranging from employment sites to sharing economy platforms. As ranking positions influence the amount of attention the ranked subjects receive, biases in rankings can lead to unfair distribution of opportunities and resources such as jobs or income.

This paper proposes new measures and mechanisms to quantify and mitigate unfairness from a bias inherent to all rankings, namely, the *position bias* which leads to disproportionately less attention being paid to low-ranked subjects. Our approach differs from recent fair ranking approaches in two important ways. First, existing works measure unfairness at the level of subject *groups* while our measures capture unfairness at the level of *individual* subjects, and as such subsume group unfairness. Second, as no single ranking can achieve individual attention fairness, we propose a novel mechanism that achieves *amortized fairness*, where attention accumulated across a series of rankings is proportional to accumulated relevance.

We formulate the challenge of achieving amortized individual fairness subject to constraints on ranking quality as an online optimization problem and show that it can be solved as an integer linear program. Our experimental evaluation reveals that unfair attention distribution in rankings can be substantial, and demonstrates that our method can improve individual fairness while retaining high ranking quality.

KEYWORDS

Algorithmic Fairness, Fair Ranking, Position Bias, Exposure

1 INTRODUCTION

Motivation and Problem. Rankings of subjects like people, hotels, or songs are at the heart of selection, matchmaking and recommender systems. Such systems are in use on a variety of platforms that affect different aspects of life – from entertainment and dating all the way to employment and income. Notable examples of platforms with a tangible impact on people’s livelihood include two-sided sharing economy websites, such as Airbnb or Uber, or human-resource matchmaking platforms, such as LinkedIn or TaskRabbit.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-5657-2/18/07...\$15.00

<https://doi.org/10.1145/3209978.3210063>

The ongoing migration to online markets and the growing dependence of many users on these platforms in securing an income have spurred investigations into the issues of bias, discrimination and fairness in the platforms’ mechanisms [5, 26].

One aspect in particular has evaded scrutiny thus far – to be successful on these platforms, ranked subjects need to gain the *attention* of searchers. Since exposure on the platform is a prerequisite for attention, subjects have a strong desire to be highly ranked. However, when inspecting ranked results, searchers are susceptible to *position bias*, which makes them pay most of their attention to the top-ranked subjects. As a result, lower-ranked subjects often receive disproportionately less attention than they deserve according to the ranking relevance. Position bias has been studied in information retrieval in scenarios where subjects are documents such as web pages (e.g., [8, 10]). It has been shown that top-ranked documents receive most clicks often irrespective of their actual relevance [21].

Systemic correction for the bias becomes important when ranking positions potentially translate to financial gains or losses. This is the case when ranking people on platforms like LinkedIn or Uber, products on platforms like Amazon, or creative works on platforms like Spotify. For example, cumulating the exposure on a subset of drivers in ride-hailing platforms might lead to economic starvation of others, while low-ranked artists on music platforms might not get their deserved chance of earning royalties.

Observing that attention is influenced by a human perception bias, while relevance is not, uncovers a fundamental problem: there necessarily exists a discrepancy between the attention that subjects receive at their respective ranks and their relevance in a given search task. For example, attention could decrease geometrically, whereas relevance scores may decrease linearly as the rank decreases. If a ranking is displayed unchanged to many searchers over time, the lower-ranked subjects might be systematically and repeatedly disadvantaged in terms of the attention they receive.

Problem Statement. A vast body of ranking models literature has focused on aligning system relevance scores with the true relevance of ranked subjects, and in this paper we assume the two are proportional. What we focus on instead is the relation between relevance and attention. Since relevance can be thought of as a proxy for worthiness in the context of a given search task, the attention a subject receives from searchers should ideally be proportional to her relevance. In economics and psychology, a similar idea of proportionality exists under the name of equity [31] and is employed as a fairness principle in the context of distributive justice [17]. Thus, in this paper, we make a translational normative claim and argue for *equity of attention* in rankings.

Operationally, the problem we address in this paper is to devise measures and mechanism which ensure that, for all subjects in the system, the received attention approximately equals the deserved attention, while preserving ranking quality. For a single ranking this goal is infeasible, since attention is influenced by the position bias, while relevance is not. Therefore, our approach looks at a series of rankings and aims at measures of *amortized fairness*.

State of the Art and Limitations. Fairness has become a major concern for decision-making systems based on machine learning methods (see, e.g., [9, 29]). Various notions of *group fairness* have been investigated [14, 20, 23, 28, 35], with the goal of making sure that protected attributes such as gender or race do not influence algorithmic decisions. Fair classifiers are then trained to maximize accuracy subject to group fairness constraints. These approaches, however, do not distinguish between different subjects from within a group. The notion of *individual fairness* [12, 24, 37] aims at treating each individual fairly by requiring that subjects who are similar to each other receive similar decision outcomes. For instance, the concept of *meritocratic fairness* requires that less qualified candidates are almost never preferred over more qualified ones when selecting candidates from a set of diverse populations. Relevance-based rankings, where more relevant subjects are ranked higher than less relevant ones, also satisfy meritocratic fairness. A stronger fairness concept, however, is needed for rankings to be a means of distributive justice.

Prior work on *fair rankings* is scarce and includes approaches that perturb results to guarantee various types of group fairness. This goal is achieved by techniques similar to those for ranking result diversification [6, 34, 36], or by granting equal ranking exposure to groups [30]. Individual fairness is inherently beyond the scope of group-based perturbation.

Approach and Contribution. Our approach in this paper differs from the prior work in two major ways. First, the measures introduced here capture fairness at the level of *individual* subjects, and subsume group fairness as a special case. Second, as no single ranking can guarantee fair attention to every subject, we devise a novel mechanism that ensures *amortized fairness*, where attention is fairly distributed across a series of rankings.

For an intuitive example, consider a ranking where all the relevance scores are almost the same. Such tiny differences in relevance will push subjects apart in the display of the results, leading to a considerable difference in the attention received from searchers. To compensate for the position bias, we can reorder the subjects in consecutive rankings so that everyone who is highly relevant is displayed at the top every now and then.

Our goal is not just to balance attention, but to keep it proportional to relevance for all subjects while preserving ranking quality. To this end, we permute subjects in each ranking so as to improve fairness subject to constraints on quality loss. We cast this approach to an online optimization problem, formalizing it as an integer linear program (ILP). We moreover devise filters to prune the combinatorial space of the ILP, which ensures that it can be solved in an online system. Experiments with synthetic and real-life data demonstrate the viability of our method.

This paper makes the following novel contributions:

- To the best of our knowledge, we are the first to formalize the problem of individual equity-of-attention fairness in rankings, and define measures that capture the discrepancy between the deserved and received attention.
- We propose online mechanisms for fairly amortizing attention over time in consecutive rankings.
- We investigate the properties and behavior of the proposed mechanisms in experiments with synthetic and real-world data.

2 EQUITY-OF-ATTENTION FAIRNESS

We now formally define equity of attention accounting for *position bias*, which determines how attention is distributed over the ranking positions. We consider a sequence of rankings at different time points, by different criteria or on request of different users.

2.1 Notation

We use the following notation:

- u_1, \dots, u_n is a set of subjects ranked in a system,
- ρ^1, \dots, ρ^m is a sequence of rankings,
- r_i^j is the $[0..1]$ -normalized relevance score of subject u_i in ranking ρ^j ,
- a_i^j is the $[0..1]$ -normalized attention value received by subject u_i in ranking ρ^j ,
- A denotes the distribution of cumulated attention across subjects, that is, $A_i = \sum_{j=1}^m a_i^j$ for subject u_i ,
- R denotes the distribution of cumulated relevance across subjects, that is, $R_i = \sum_{j=1}^m r_i^j$ for subject u_i .

2.2 Defining Equity of Attention

Our fairness notion in this work is in the spirit of the *individual fairness* proposed by Dwork et. al. [12], which requires that “similar individuals are treated similarly”, where “similarity” between individuals is a metric capturing suitability for the task at hand. In the context of rankings, we consider *relevance* to be a measure of subject suitability. Further, in applications where rankings influence people’s economic livelihood, we can think of rankings not as an end, but as a means of achieving distributive justice, that is, fair sharing of certain real-world resources. In the context of rankings, we consider the *attention of searchers* to be a resource to be distributed fairly.

There exist different types of distributive norms, one of them being *equity*. Equity encodes the idea of proportionality of inputs and outputs [31], and might be employed to account for “differences in effort, in productivity, or in contribution” [33].

Building upon these ideas, we make a translational normative claim and propose a new notion of individual fairness for rankings called *equity of attention*, which requires that *ranked subjects receive attention that is proportional to their worthiness in a given search task*. As a proxy for worthiness, we turn to the currently best available ground truth, that is, the system-predicted relevance.

DEFINITION 1 (EQUITY OF ATTENTION). *A ranking offers equity of attention if each subject receives attention proportional to its relevance:*

$$\frac{a_{i1}}{r_{i1}} = \frac{a_{i2}}{r_{i2}}, \forall u_{i1}, u_{i2}.$$

Note that this definition is unlikely to be satisfied in any single ranking, since the relevance scores of subjects are determined by the data and the query, while the attention paid to the subjects (in terms of views or clicks) is strongly influenced by position bias. The effects of this mismatch will be aggravated if multiple subjects are similarly relevant, yet obviously cannot occupy the same ranking position and receive similar attention.

To operationalize our definition in practice, we propose an alternative fairness definition that requires *attention to be distributed proportionally to relevance, when amortized over a sequence of rankings.*

DEFINITION 2 (EQUITY OF AMORTIZED ATTENTION). *A sequence of rankings ρ^1, \dots, ρ^m offers equity of amortized attention if each subject receives cumulative attention proportional to her cumulative relevance, i.e.:*

$$\frac{\sum_{l=1}^m a_{i1}^l}{\sum_{l=1}^m r_{i1}^l} = \frac{\sum_{l=1}^m a_{i2}^l}{\sum_{l=1}^m r_{i2}^l}, \forall u_{i1}, u_{i2}.$$

Observe that this modified fairness definition allows us to permute individual rankings so as to satisfy fairness requirements over time. The deficiency in the attention received by a subject relative to her relevance in a given ranking instance can be compensated in a subsequent ranking, where the subject is positioned higher relative to her relevance.

2.3 Equality of attention

In certain scenarios, it may be desirable for subjects to receive the same amount of attention, irrespective of their relevance. Such is the case when we suspect the ranking is biased and cannot confidently correct for that bias, or when the subjects are not shown as an answer to any query but need to be visually displayed in a ranked order (e.g., a list of candidates on an informational website for an election). In such scenarios, the desired notion of fairness would be *equality of attention*. We observe that this egalitarian version of fairness is a special case of equity of attention, where the relevance distributions are uniform, i.e., $r_{i1} = r_{i2} \forall u_{i1}, u_{i2}$. As equity of attention subsumes equality of attention, we do not explicitly discuss it further in this paper.

2.4 Relation to group fairness in rankings

To our knowledge, all prior works on fairness in rankings have focused on notions of *group fairness*, which define fairness requirements over the collective treatment received by all members of a demographic group like women or men. Our motivation for tackling fairness at the individual level stems from the fact that position bias affects all individuals, independently of their group membership. It is easy to see, however, that when equity of attention is achieved for individuals, it will also be achieved at the group level: the cumulated attention received by all members of a group will be proportional to their cumulated relevance.

Prior works on fairness in rankings [6, 34, 36] has mostly focused on diversification of the results. These approaches are geared

for one-time rankings, and, as any static model, will steadily accumulate equity-of-attention unfairness over time. Since they were developed with a different goal in mind, they are not directly comparable to our dynamic approach.

Parallel with our work, Singh and Joachims have explored similar ideas of how position bias influences fairness of exposure [30]. Their probabilistic formulations are possibly a counterpart of our amortization ideas, and it will be interesting to see to what extent these formulations are interchangeable. In line with other prior works on fairness in rankings and different from our work, however, they focus on satisfying constraints on group rather than individual fairness, and on notions of equality rather than equity.

3 RANKINGS WITH EQUITY OF ATTENTION

3.1 Measuring (un)fairness

To be able to optimize ranking fairness, we need to measure to what extent a sequence of rankings ρ^1, \dots, ρ^m violates Definition 2. Since the proposed fairness criterion is equivalent to the requirement that the empirical distributions A and R be equal, we can measure unfairness as the distance between these two distributions. A variety of measures can be applied here, including KL-divergence, or L1-norm distance. In this paper, measure fairness using the latter:

$$\text{unfairness}(\rho^1, \dots, \rho^m) = \sum_{i=1}^n |A_i - R_i| = \sum_{i=1}^n \left| \sum_{j=1}^m a_i^j - \sum_{j=1}^m r_i^j \right|. \quad (1)$$

L1-norm is minimized with a value of 0 for distributions satisfying the fairness criterion from Definition 2, and is thus useful as an optimization objective. However, since the measure is cumulative and indifferent to the exact distribution of unfairness among individuals, other measures could be developed to *quantify* unfairness in the system at any given point.

3.2 Measuring ranking quality

Permuting a ranking to satisfy fairness criteria can lead to a quality loss when less relevant subjects get ranked higher than more relevant ones. We propose to quantify ranking quality using measures that draw from IR evaluation. Traditionally, ranking models are evaluated in comparison with ground-truth rankings based on human-given relevance labels. Here, we are interested in quantifying the divergence from the original ranking. Thus, we consider the *original ranking* ρ to be the ground-truth reference for evaluating the quality of a *reordered ranking* ρ^* . We assume that the ground truth scores are the relevance scores returned by the system, and that these scores reflect the best ordering of subjects. These considerations lead to the following definitions.

Discounted cumulative gain (DCG) quantifies the quality of a ranking by summing the relevance scores in consecutive positions with a logarithmic discount for the values at lower positions. The measure thus puts an emphasis on having higher relevance scores at top positions.

$$\text{DCG}@k(r) = \sum_{i=1}^k \frac{2^{r(i)} - 1}{\log_2(i + 1)}$$

This value can be further normalized by the DCG score of a perfect ranking ordered by the ground truth relevance scores. The normalized discounted cumulative gain (NDCG)-based quality measure can be thus expressed as:

$$NDCG\text{-quality}@k(\rho, \rho^*) = \frac{DCG@k(\rho^*)}{DCG@k(\rho)} \quad (2)$$

This measure is maximized with a value of 1 if the rankings do not differ or if swaps are only made within ties (i.e., subjects with equal relevance). Other measures, like Kendall's Tau or appropriately defined *MAP-quality*, could be applied as well.

3.3 Optimizing fairness-quality tradeoffs

As discussed in the previous section, there is “no free lunch”: to improve fairness, we need to perturb relevance-based rankings, which might lead to lower ranking quality. To address the tradeoff, we can formulate two types of constrained optimization problems: one where we minimize unfairness subject to constraints on quality (i.e., lower-bound the minimum acceptable quality), and another where we maximize quality subject to constraints on unfairness (i.e., upper-bound the maximum acceptable unfairness measure). In this paper, we focus on the former, since at the moment ranking quality measures are more interpretable, and so are the constraints on quality.

3.3.1 Offline optimization. Let ρ^1, \dots, ρ^m be a sequence of rankings where the subjects are ordered by the relevance scores. These rankings induce zero quality loss. We wish to reorder them into $\rho^{1*}, \dots, \rho^{m*}$ so as to minimize the distance between the distributions A and R with constraints on NDCG-quality loss in each ranking:

$$\begin{aligned} & \text{minimize} && \sum_i |A_i - R_i| \\ & \text{subject to} && NDCG\text{-quality}@k(\rho^j, \rho^{j*}) \geq \theta, \quad j = 1, \dots, m. \end{aligned}$$

where A_i and R_i denote the cumulated attention and relevance scores that subject u_i has gained across all the m rankings.

Instead of thresholding the loss in each individual ranking, an alternative would be to threshold the average loss over m rankings.

3.3.2 Online optimization. In practice, ranking amortization needs to be done in an online manner, one query at a time. Without the knowledge of future query loads, the goal is then to reorder the current ranking so as to minimize unfairness over the cumulative attention and relevance distributions in rankings seen so far, subject to a constraint on the quality of the current ranking. Thus, in the l -th ranking we want to :

$$\begin{aligned} & \text{minimize} && \sum_i |A_i^{l-1} + a_i^l - (R_i^{l-1} + r_i^l)| \\ & \text{subject to} && NDCG\text{-quality}@k(\rho^l, \rho^{l*}) \geq \theta \end{aligned}$$

where A_i^{l-1} and R_i^{l-1} denote the cumulated attention and relevance scores that subject u_i has gained up to and including ranking ρ^{l-1} .

3.4 An ILP-based fair ranking mechanism

3.4.1 ILP for online attention amortization. The optimization problem defined in Sec. 3.3.2 can be solved as an integer linear

program (ILP). Assume we are to rerank the l -th ranking in a series of rankings. We introduce n^2 decision variables $X_{i,j}$ which are set to 1 if subject u_i is assigned to the ranking position j , and set to 0 otherwise. At the time of reordering the l -th ranking, the following values are constants:

- relevance scores for each subject u_i in the current ranking: r_i^l ,
- attention values assigned to ranking positions: w_j ,
- relevance scores accumulated up to (and excluding) the current ranking for each subject: R_i^{l-1} ,
- attention values accumulated up to (and excluding) the current ranking for each subject: A_i^{l-1} ,
- IDCG@ k value computed over the current ranking ρ_l , which is used as a normalization score for NDCG-quality@ k .

For each subject u_i , the accumulated attention and relevance are initialized as $A_i^0 = 0$ and $R_i^0 = 0$ for all u_i .

The ILP is then defined as follows:

$$\begin{aligned} & \text{minimize} && \sum_{i=1}^n \sum_{j=1}^n |A_i^{l-1} + w_j - (R_i^{l-1} + r_i^l)| \cdot X_{i,j} \\ & \text{subject to} && \sum_{j=1}^k \sum_{i=1}^n \frac{2^{r_i^l} - 1}{\log_2(j+1)} X_{i,j} \geq \theta \cdot IDCG@k \\ & && X_{i,j} \in \{0, 1\}, \quad \forall_{i,j} \\ & && \sum_i X_{i,j} = 1, \quad \forall_j \\ & && \sum_j X_{i,j} = 1, \quad \forall_i \end{aligned} \quad (3)$$

The first constraint bounds the loss in ranking quality, in terms of the NDCG-quality measure, by the multiplicative threshold $0 \leq \theta \leq 1$. The other constraints ensure that the solution is a bijective mapping of subjects onto ranking positions. The terms $A_i^{l-1} + w_j$ and $R_i^{l-1} + r_i^l$ encode the updates of the cumulative attention and relevance, respectively, if and only if u_i is mapped to position j .

It is worth noting that:

- When $\theta = 1$, we do not allow any quality loss. This, however, does not mean that the ranking will remain unchanged. Subjects can be reordered within ties to minimize unfairness.
- When $\theta = 0$, any permutation of the ranking is allowed striving to minimize unfairness in the current iteration.

3.4.2 ILP with candidate pre-filtering. The ILP operates on a huge combinatorial space, with the number of binary variables being quadratic in the number of subjects. Real systems deal with millions of subjects, and the optimization needs to be carried out each time a new ranking is requested. Such a problem size is a bottleneck for ILP solvers, and in practice the optimization needs to use approximation algorithms, such as LP relaxations or greedy-style heuristics. This is one of the directions for further research.

To deal with the issue in this paper, instead of reranking all subjects in each iteration, we rerank only subjects from a prefiltered candidate set. Different strategies are possible for selecting the candidate sets. On the one hand, prefiltering the top-ranked subjects by relevance scores would let us satisfy the quality constraints, but may entail small fairness gains, especially for near-uniform

relevance distributions. On the other hand, prefiltering based on the objective function might lead to situations where the ILP cannot find any solution without violating the constraints.¹

Our strategy thus is as follows. Assume we want to select a subject candidate subset D of size t to be reranked, and we constrain the quality in Eq. 3 at rank k . Since the attention weights w_j are positive, the biggest contributors to the objective function are the subjects with the smallest values of $A_i - (R_i + r_i)$. These are the subjects with the highest deficit (negative value) of fair-share attention. We always select k subjects with the highest relevance scores in r^l , to make sure we satisfy the quality constraint, plus other $t - k$ subjects with the lowest $A_i - (R_i + r_i)$ values, who are most worthy of being promoted to high ranks. As a result, when no feasible solution can be found by reranking the most worthy subjects, the ILP will default to choosing the top- k candidates by relevance scores.

3.4.3 Extensions.

Granularity. The presented model assumes that attention and relevance are aggregated per ranked subject. It is straightforward to extend it to handle higher-level actors such as product brands or Internet domains, by summing the relevance and attention scores over the corresponding subjects. As a consequence of this modification, bigger organizations would obtain higher exposure. Deciding whether this effect is fair is a policy issue.

Handling dynamics. In a real-world system, the size of the population will vary over time, with new subjects joining and existing ones dropping out. Our model is capable of handling this kind of dynamics, since new users starting with no deserved attention will be positioned in between the users who got more than they deserved and those who got less. Moreover, ranking quality constraints will prevent such users from being positioned too low in rankings where they are highly relevant.

4 EXPERIMENTS

4.1 Data

The datasets we use are either synthetically generated or derived from other publicly available resources. They are freely available to other researchers.

4.1.1 Synthetic datasets. We create 3 synthetic datasets to analyze the performance of the model in a controlled setup under different relevance distributions. We assume the following distribution shapes: (i) **uniform**, where every user has the same relevance score, (ii) **linear**, where the scores decrease linearly with the rank position, and (iii) **exponential**, where the scores decrease exponentially with the rank position. Each dataset has 100 subjects.

4.1.2 Airbnb datasets. To analyze the model in a real-world scenario, we construct rankings based on Airbnb² apartment listings from 3 cities located in different parts of the world: Boston, Geneva, and Hong Kong. Airbnb is a two-sided sharing economy platform allowing people to offer their free rooms or apartments for short-term rental. It is a prime example of a platform where exposure and

¹Without prefiltering, the ILP always has at least one feasible solution (the original ranking).

²<https://www.airbnb.com/>

attention play a crucial role in the subjects' financial success. The data we use is freely available for research.³

Rankings are constructed using the attribute *id* as a subject identifier, and various review ratings as the ranking criteria, with the rating scores serving as relevance scores. Such crowd-sourced judgments serve as a good worthiness-of-attention proxy on this particular platform, although one has to have in mind that rating distributions tend to be skewed towards higher scores, which is confirmed by our experimental analysis.

For each of the 3 datasets, we run the amortization model on two types of ranking sequences:

- (1) **Single-query:** We examine the amortization effects when a single ranking is repeated multiple times. To construct the rankings, we use the values of the *review_scores_rating* attribute, which corresponds to the overall quality of the listing.
- (2) **Multi-query:** We examine the behavior of the model when a sequence of rankings, each with a different relevance distribution, is repeated multiple times. To this end, for each city, we construct 7 rankings based on different rating attributes: *review_scores_rating*, *review_scores_accuracy*, *review_scores_cleanliness*, *review_scores_checkin*, *review_scores_communication*, *review_scores_location*, and *review_scores_value*.

The datasets for Boston, Geneva, and Hong Kong contain 3944, 1728, and 4529 subjects, respectively.

Note that, for the purpose of model performance evaluation, the queries themselves become irrelevant once the relevance is computed. Since the values of the aforementioned attributes serve as relevance scores, the queries are abstracted out.

4.1.3 StackExchange dataset. We create another dataset from a querylog and a document collection synthesized from the StackExchange dump by Biega et al. [4], please refer to the original paper for details. We choose a random subset of users and order their queries by timestamps, creating a workload of around 20K queries. We use Indri⁴ to retrieve 500 most relevant answers for each query, and treat the author of the answer as the subject to be ranked. Using this dataset helps us gain an insight into the performance of the method in core IR tasks and with different sets of subjects ranked in each iteration.

4.2 Position bias

Our model requires that we assign a weight to each ranking position, denoting the fraction of the total attention the position gets. These weights will depend on the application and platform, and may be estimated from historical click data. In this paper we study the behavior of the equity-of-attention mechanism under generic models of attention distribution. We focus on the following distributions:

- (1) **Geometric:** The weights of the positions are distributed geometrically with the parameter p up to the position k , and are 0 for positions lower than k . Geometrically distributed weights are a special case of the cascade model [10], where

³Downloaded from <http://insideairbnb.com/>

⁴<https://www.lemurproject.org/indri/>

each subject has the same probability p of being clicked. Setting the weights of lower positions to 0 is based on an assumption that low-ranked subjects are not inspected.

$$w_j = \begin{cases} p(1-p)^{j-1} & j \leq k \\ 0 & j > k \end{cases} \quad (4)$$

- (2) **Singular:** The top-ranked subject receives all the attention. This is a special case of the geometric attention model with parameters $p = 1, k = 1$. Studying this attention model is motivated by systems such as Uber, which present only top-1 matches to the searchers by default.

$$w_j = \begin{cases} 1 & j = 1 \\ 0 & j > 1 \end{cases} \quad (5)$$

Before being passed on to the model, the weights are rescaled such that $\sum_j w_j = 1$. Studying the effects of position bias on individual fairness under more complex attention models is future work.

4.3 Implementation and parameters

We implement the ILP-based amortization defined in Section 3.4 using the Gurobi software.⁵ Constraints are set to be satisfied up to a feasibility threshold of $1e-7$. We prefilter 100 candidates for reranking in each iteration, as described in Section 3.4.2.

In the singular attention model, since all the attention is assumed to go to the first ranking position, the ILP constrains the NDCG-quality at rank $k = 1$. We construct the geometric attention model with $p = 0.5$ and $k = 5$, and in this case the ILP constrains the NDCG-quality at rank $k = 5$.

In the single-query mode, where a single ranking is repeated multiple times, we set the number of iterations to 20K. In the multi-query mode, with a repeated sequence of different rankings, we repeat the whole sequence 3K times, which leads to a total of 21K rankings.

Relevance scores in the framework need to be normalized to form a distribution. In this paper, we assume relevance is a direct proxy for worthiness and rescale the rating scores linearly. Note, however, that if additional knowledge is available to the platform regarding the correspondence between relevance and worthiness, other transformations can be applied as well.

4.4 Mechanisms under comparison

We compare the performance of the ILP-based online mechanism against two baseline heuristics.

- (1) **Relevance:** The first heuristic is to allow only relevance-based ranking, completely disregarding fairness.
- (2) **Objective:** The second heuristics is an objective-driven ranking strategy, which orders subjects by the increasing priority value: $A_i - R_i - r_i$ (see Sec. 3.4.2) for each ranking. Since all position weights w_j are positive, assigning highest weights to subjects with the lowest preference value is in line with the minimization goal. This ranking strategy aims at strong fairness amortization without any quality constraints, and is expected to perform similarly to the ILP with $\theta = 0$.

⁵<http://www.gurobi.com/>

4.5 Data characteristics: relevance vs. attention

Figure 1 shows the relevance score distributions in the single-query Airbnb datasets for Boston, Geneva, and Hong Kong. The seemingly flatter shape of the Boston and Hong Kong distributions is the result of a bigger size of these datasets when compared to the Geneva dataset, where each individual has, on average, a larger fraction of the total relevance. Overall, the distributions have a shape which complements the uniform, linear, and exponential shapes of distributions in the synthetic datasets.

Figure 2 presents an example strongly motivating our research. Namely, it compares the distribution of attention according to the geometric model with $p = 0.5$, where the weights closely follow the empirical observations made in previous position bias studies [21]. Observe that the relevance distribution plotted in green is the same as that in Figure 1. There is a huge discrepancy between these two distributions, while, as argued in this paper, they should ideally be equal to ensure individual fairness. Similar discrepancy exists in the two other Airbnb datasets.

4.6 Performance on synthetic data

Singular attention model. Figure 3 reveals a number of interesting properties of the mechanism for the Uniform relevance distribution. We plot the iteration number on the x-axis, and the value of the

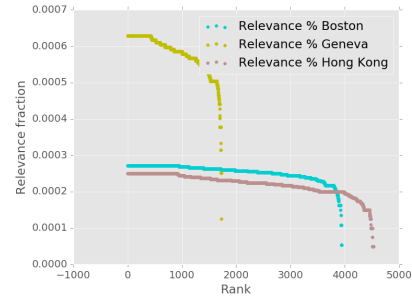


Figure 1: Relevance distributions in the Airbnb datasets.

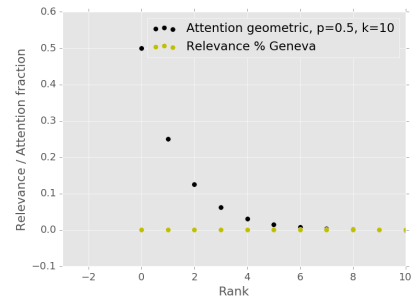


Figure 2: Comparison of the attention and relevance distributions for the top-10 ranking positions in the Geneva dataset. Note that the relevance distribution presented here is the same as in Fig. 1. To satisfy equity-of-attention fairness, the two distributions would have to be the same.

unfairness measure defined by Equation 1 on the y-axis. First, since reshuffling does not lead to any quality loss when all the relevance scores are equal, all the reshuffling methods perform equally well irrespective of θ . Their amortizing behavior should be contrasted with the black line denoting the relevance baseline. Unfairness for this method always increases linearly by a constant factor incurred by the single ranking. Second, amortization methods periodically bring unfairness to 0. The minimum occurs every n iterations, where n is the number of subjects in the dataset. Within the cycle, each subject is placed in the top position (receiving all the attention) exactly once.

Figure 4 with the results for the Linear dataset, confirms another anticipated behavior. With no ties in the relevance scores, it is not possible to improve fairness without incurring quality loss. Thus, all methods with $\theta > 0$ lead to higher unfairness when compared to the Objective baseline, although the unfairness is still lower in ILP with $\theta < 0.8$ than in the Relevance baseline.

When the relevance scores decrease exponentially (Figure 5), the ILP is not able to satisfy the quality constraint with any $\theta \geq 0.5$, and thus these rerankings become equivalent to those of the Relevance heuristic.

Geometric attention model. As shown in Figures 6, 7, and 8, the periodicity effect becomes less pronounced under the general geometric attention model. Figure 9 helps to understand this behavior by showing the unfairness values achieved by the Objective heuristic with different values of the attention cut-off k (see Equation 4). With $k = 1$, the model is equivalent to Singular. As we increase k , the distribution of the position weights becomes smoother, smoothing also the periodicity of the unfairness values.

The very good performance of the ILP-based rerankings with any $\theta < 1$ in Figure 8, stems from the fact that the relevance and attention distributions are almost the same (the only difference being that the scores in the relevance distribution are non-zero for more positions). Our results show that in this case the ILP performs a reordering only every now and then, when the subjects ranked lower than position 5 in the original ranking gather enough deserved attention. This causes the unfairness to go up and down periodically.

4.7 Performance on Airbnb data

4.7.1 Single-query, singular attention. We first analyze the model performance on the Airbnb datasets where a single ranking is repeated multiple times, and the attention model is set to singular. The results are shown in Figures 10, 11, 12 for Boston, Geneva, and Hong Kong, respectively. As in the analysis with the synthetic data, we plot the iteration number on the x-axis, and the value of the unfairness measure defined by Equation 1 on the y-axis. There are a number of observations:

- As noted before, the loss in the Relevance baseline (plotted in black) increases linearly by the constant unfairness factor incurred by the single ranking.
- Relaxing the quality constraint by decreasing θ allows us to achieve lower unfairness values in the corresponding ranking iterations.
- The Objective heuristic with no quality constraints and the ILP where $\theta = 0$ are able to amortize fairness over time well, with no significant growth of unfairness over time.
- The periodicity effect we observed on synthetic uniform data appears here as well. This is due to the relative closeness of the relevance distributions in the Airbnb data to the uniform distribution. Unfairness achieved by the amortizing methods is close to 0 every n iterations. The frequency of the minimum indeed corresponds to the size of the respective datasets.
- In some methods unfairness starts to grow linearly after a certain number of iterations (see, e.g., the blue curve in Figure 10). This is a side effect of the candidate prefiltering heuristic we chose. When the ILP receives a filtered candidate set where no subjects filtered based on the objective can be placed at the top of the ranking without violating the quality constraint, the ILP defaults to placing the most relevant subjects at the top, which causes the quality loss to be 0 and the unfairness growing linearly. This effect persists until some of the more relevant subjects gather enough deserved attention to be pre-selected - note the variability that occurs in the blue curve again starting around the 17K-th iteration.
- For a number of iterations at the beginning (equal to the number of ties at the top of the ranking), all the methods perform the same, irrespective of the quality constraints. This is due to the fact that unfairness is minimized by reshuffling the most deserving relevant subjects first, which does not incur any quality loss.

4.7.2 Multi-query, singular attention. Our methods amortize fairness better (achieving lower unfairness) on the Airbnb multi-query datasets (Figures 13, 14, and 15) when compared to the single-query datasets for two reasons. First, the variability in subject relevance and ordering in different iterations is a factor helpful in smoothing the deserved attention distributions over time. Second, distributions of the rating attributes in the Airbnb datasets used to construct the rankings are more uniform than the global rating score, and have more ties at the top of the ranking. These relevance distribution characteristics enable methods with conservative quality constraints (even the ILP with $\theta = 1$) to perform very well.

4.7.3 Single-query, geometric attention. The general geometric attention distribution is closer to the relevance distributions in the Airbnb datasets than the singular distribution is. As noted in the analysis with synthetic data, the closeness of the two distributions helps amortize fairness at a lower quality loss. We can observe a similar effect in Figure 16, with more ILP-based methods reaching the performance of the Objective heuristic. Note, however, that the improved performance here is also partly due to the fact that we constrain the quality at a higher rank when assuming the geometric attention, which is easier to satisfy.

4.7.4 Unfairness vs. quality loss. The results presented so far show the performance of the ILP-based fairness amortization under different quality thresholds. Since the thresholds bound the maximum quality loss over all iterations, the actual loss in most cases might be lower. To investigate these effects, we plot the actual NDCG-quality values of the rerankings done by different methods

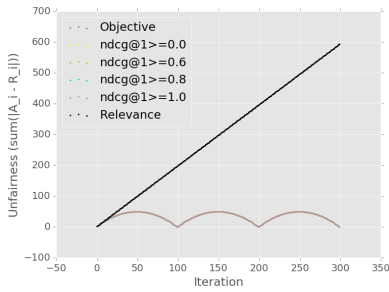


Figure 3: Model performance on the synthetic Uniform dataset. Attention singular.

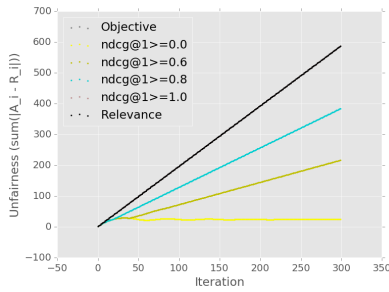


Figure 4: Model performance on the synthetic Linear dataset. Attention singular.

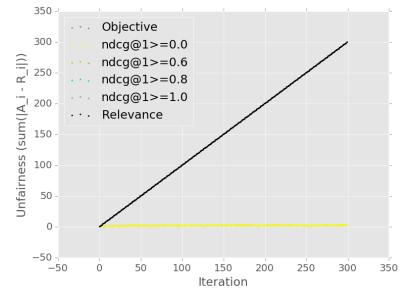


Figure 5: Model performance on the synthetic Exponential dataset. Attention singular.

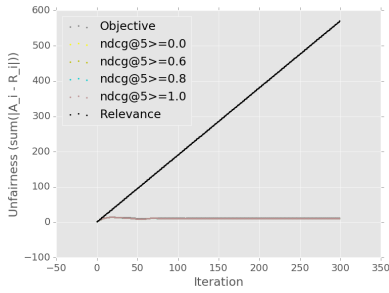


Figure 6: Model performance on the synthetic Uniform dataset. Attention geometric.

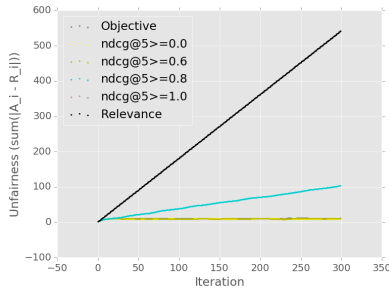


Figure 7: Model performance on the synthetic Linear dataset. Attention geometric.

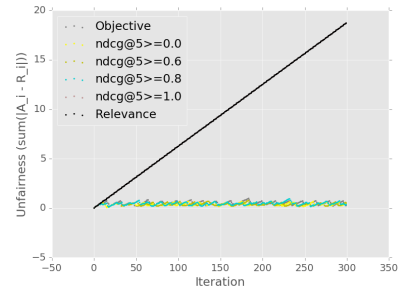


Figure 8: Model performance on the synthetic Exponential dataset. Attention geometric.

on the Boston dataset under the Singular attention model in Figure 17. The results confirm that the actual loss is often lower than the threshold enforced by the ILP. Observe that NDCG-quality is 1 for a number of initial iterations in all the methods. This is where reshuffling of the top ties happens. The quality starts decreasing as less relevant subjects gather enough deserved attention, and periodically goes back to 1, when the top-relevant subjects gain priority again. Similar conclusions regarding the absolute loss hold under the general geometric attention model.

4.8 Performance on StackExchange data

The relative trends in the performance of our method are the same here as in the results for other datasets. One of the characteristics

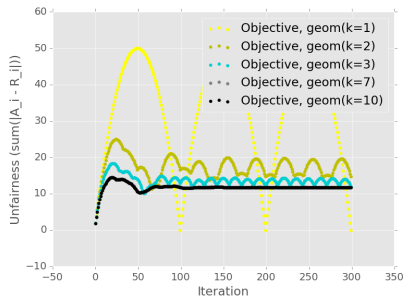


Figure 9: Performance of the Objective heuristic on the synthetic Uniform dataset under the geometric attention model with different attention cut-off points.

that distinguish the StackExchange dataset is that each individual subject occurs in relatively few rankings. An observation that follows is that longer amortization timeframe is necessary under such conditions - a subject obviously needs to appear in a number of rankings so that the model can reposition them to fairly distribute attention.

5 RELATED WORK

Fairness. The growing ubiquity of data-driven learning models in algorithmic decision-making has recently boosted concerns about the issues of fairness and bias (see, e.g., [9] and the pointers there). The problem of discrimination in data mining and machine learning has been studied for a number of years (e.g., [23, 28, 29]). The goal there is to analyze and counter data bias and unfair decisions that may lead to discrimination. Much prior work has centered around various notions of group fairness: preserving certain ratios of members of protected vs. unprotected groups in the decision making outcomes, with the groups derived from discrimination-prone attributes like gender, race, nationality, etc. [14, 20]. For example, the criterion of statistical parity requires that a classifier’s outcomes do not depend on the membership in the protected group. State-of-the-art mechanisms for dealing with such group fairness requirements are to solve constrained optimization, e.g. maximize prediction accuracy subject to certain bounds on group membership in the output labels. This has led to classification models with fairness-aware regularization (e.g., [35]). Beyond the fairness of outcomes, researchers have looked into the fairness of process in the decision-making systems [18].

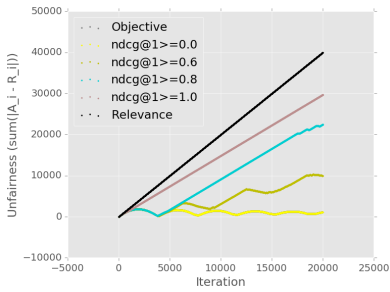


Figure 10: Model performance on the single-query Boston dataset. Attention singular.

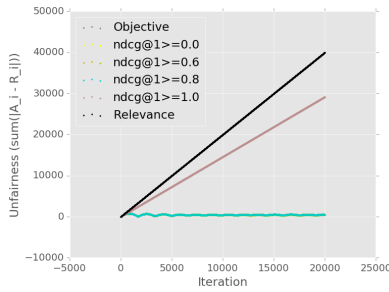


Figure 11: Model performance on the single-query Geneva dataset. Attention singular.

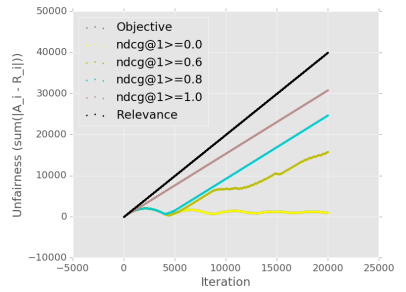


Figure 12: Model performance on the single-query Hong Kong dataset. Attention singular.

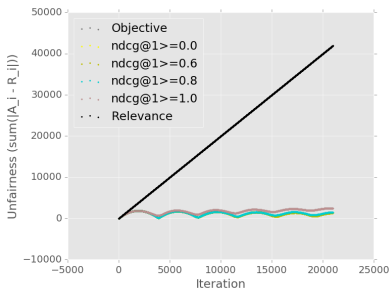


Figure 13: Model performance on the multi-query Boston dataset. Attention singular.

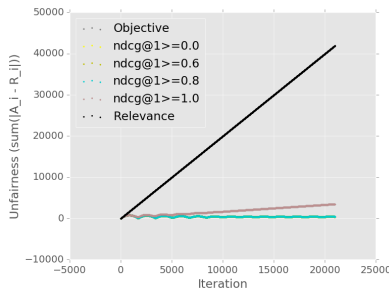


Figure 14: Model performance on the multi-query Geneva dataset. Attention singular.

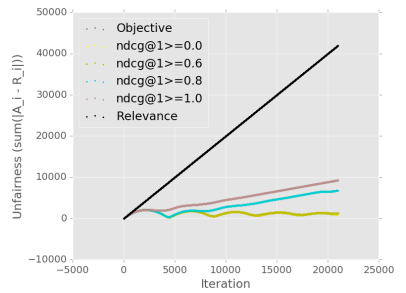


Figure 15: Model performance on the multi-query Hong Kong dataset. Attention singular.

Individual fairness [12] requires that individual subjects who have similar attributes should, with high probability, receive the same prediction outcomes. Literature to this end has so far focused on classification and selection problems [24, 37].

Other lines of work investigate mechanisms for fair division of resources [1], or how automated systems can assist humans in decision making [25].

Fairness in rankings. Prior work on fair rankings is scarce and recent. Some proposals show how to incorporate various notions of group fairness into ranking quality measures [34]. There have been approaches that diversify the ranking results in terms of presence of members of different groups in ranking prefixes, at the same time

keeping the ranking quality high [36]. This problem has also been studied from a theoretical perspective with the results provided for the computational complexity of the problem [6]. All of these approaches consider static rankings only, and all focus on group fairness. Parallel with our work, Singh and Joachims [30] have proposed a notion of group fairness based on equality of exposure for demographic groups. While technically complementary and similar in spirit to our approach, this method is also geared for a purpose different than individual fairness, and does not aim at binding attention to relevance.

Bias in IR. The existence of position bias in rankings of search results has been revealed by a number of eye-tracking and other

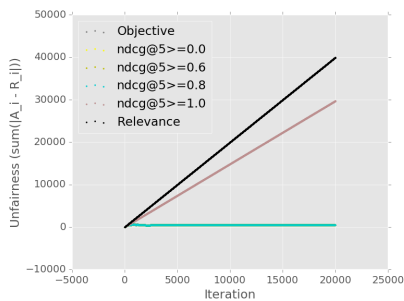


Figure 16: Model performance on the single-query Boston dataset. Attention geometric. Results are similar for the Geneva and Hong Kong datasets.

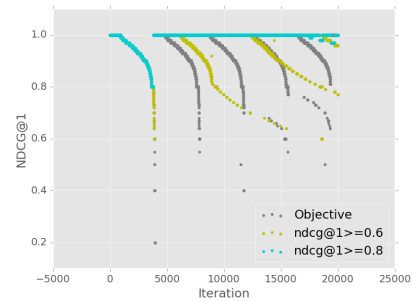


Figure 17: Actual values of ranking quality. Boston dataset, attention singular.

empirical studies [10, 11, 19]. Top-ranked answers have a much higher probability of being viewed and clicked than those at lower ranks. The effect persists even if the elements at different ranks are randomly permuted [21]. These observations have led to a variety of click models (see [8] for a survey), and several methods for bias-aware re-ranking (e.g., [22, 32]). However, position bias has been primarily studied in the context of document ranking and no prior work has investigated the influence of the bias on the fairness of ranked results. A large search engine has been investigated for presence of differential quality of results across demographic groups [27]. Similar studies have been carried out on other kinds of tasks such as credit worthiness or recidivism prediction [2].

Relation to other models. Fairness dimension has been considered for job dispatching at the OS level, for packet-level network flows [15], for production planning in factories [16], and even for two-sided matchmaking in call centers [3]. Fairness understood as envy-freeness is also investigated in computational advertising, including generalized second-price auctions [13]. In the context of rankings, a potential connection between fair rankings and fair queuing has recently been suggested [7].

6 CONCLUSION

This paper argues for equity of attention – a new notion of fairness in rankings, which requires that the attention ranked subjects receive from searchers is proportional to their relevance. As this definition cannot be satisfied in a single ranking because of the position bias, we propose to amortize fairness over time by reordering consecutive rankings, and formulate a constrained optimization problem which achieves this goal.

Our experimental study using real-world data shows that the discrepancy between the attention received from searchers and the deserved attention can be substantial, and that many subjects have equal relevance scores. These observations suggest that improving equity of attention is crucial and can often be done without sacrificing much quality in the rankings. Incorporating such fairness mechanisms is especially important on sharing economy or two-sided market platforms where rankings influence people's economic livelihood, and our work addresses this gap.

Equity of attention opens a number of interesting directions for future work, including calibration of ranker scores in economically-themed applications, all the way down the IR stack to properly training judges to provide relevance labels with fairness in mind.

Acknowledgements. This work was partly supported by the ERC Synergy Grant 610150 (imPACT). We thank Aniko Hannak and Abhijnan Chakraborty for inspiring discussions at the initial stage of this project.

REFERENCES

- [1] Rediet Abebe, Jon Kleinberg, and David C Parkes. 2017. Fair division via social comparison. In *AAMAS*.
- [2] Philip Adler, Casey Falk, Sorelle Friedler, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2016. Auditing Black-Box Models for Indirect Influence. In *ICDM*.
- [3] Mor Armony and Amy R. Ward. 2010. Fair Dynamic Routing in Large-Scale Heterogeneous-Server Systems. *Operations Research* (2010).
- [4] Asia J Biega, Rishiraj Saha Roy, and Gerhard Weikum. 2017. Privacy through Solidarity: A User-Utility-Preserving Framework to Counter Profiling. In *SIGIR*.
- [5] Ryan Calo and Alex Rosenblat. 2017. The taking economy: Uber, information, and power. *Columbia Law Review* (2017).
- [6] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2017. Ranking with Fairness Constraints. *arXiv preprint arXiv:1704.06840*.
- [7] Abhijnan Chakraborty, Asia J. Biega, Aniko Hannak, and Krishna P. Gummadi. 2017. Fair Sharing for Sharing Economy Platforms. In *FATREC@RecSys Workshop*.
- [8] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. Click Models for Web Search. In *Morgan & Claypool*.
- [9] FAT Conference. [n. d.]. Conference on Fairness, Accountability, and Transparency (FAT*). In <http://fatconference.org/resources.html>.
- [10] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *WSDM*.
- [11] Georges Dupret and Benjamin Piwowarski. 2008. A user browsing model to predict search engine click data from past observations. In *SIGIR*.
- [12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *ITCS*.
- [13] Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. 2007. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American economic review* (2007).
- [14] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *KDD*.
- [15] Ali Ghodsi, Vyas Sekar, Matei Zaharia, and Ion Stoica. 2012. Multi-resource fair queueing for packet processing. In *SIGCOMM*.
- [16] Ali Ghodsi, Matei Zaharia, Benjamin Hindman, Andy Konwinski, Scott Shenker, and Ion Stoica. 2011. Dominant resource fairness: Fair allocation of multiple resource types. In *NSDI*.
- [17] Jerald Greenberg. 1987. A taxonomy of organizational justice theories. *Academy of Management review* (1987).
- [18] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In *AAAI*.
- [19] Fan Guo, Chao Liu, and Yi Min Wang. 2009. Efficient multiple-click models in web search. In *WSDM*.
- [20] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NIPS*.
- [21] Thorsten Joachims and Filip Radlinski. 2007. Search Engines that Learn from Implicit Feedback. *IEEE Computer* (2007).
- [22] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased Learning-to-Rank with Biased Feedback. In *WSDM*.
- [23] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *ECML/PKDD*.
- [24] Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. 2017. Meritocratic Fairness for Cross-Population Selection. In *ICML*.
- [25] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human decisions and machine predictions. *The Quarterly Journal of Economics* (2017).
- [26] Karen Levy and Solon Barocas. 2018. Designing Against Discrimination in Online Markets. *Berkeley Technology Law Journal* (2018).
- [27] Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. 2017. Auditing Search Engines for Differential Satisfaction Across Demographics. In *WWW*.
- [28] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *KDD*.
- [29] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *Knowledge Eng. Review* (2014).
- [30] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. *arXiv preprint arXiv:1802.07281*.
- [31] Elaine Walster, Ellen Berscheid, and G William Walster. 1973. New directions in equity research. *Journal of personality and social psychology* (1973).
- [32] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. 2016. Learning to Rank with Selection Bias in Personal Search. In *SIGIR*.
- [33] Menahem E Yaari and Maya Bar-Hillel. 1984. On dividing justly. *Social choice and welfare* (1984).
- [34] Ke Yang and Julia Stoyanovich. 2007. Measuring fairness in ranked outputs. In *SSDBM*.
- [35] 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 *WWW*.
- [36] Meike Zehlke, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A fair top-k ranking algorithm. In *CIKM*.
- [37] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *ICML*.