Evaluating Stochastic Rankings with Expected Exposure

Fernando Diaz
Microsoft
Montréal, QC
diazf@acm.org

Bhaskar Mitra
Microsoft
Montréal, QC
bmitra@microsoft.com

Michael D. Ekstrand
People & Information Research Team
Boise State Computer Science
Boise, ID
michaelekstrand@boisestate.edu

Asia J. Biega
Microsoft
Montréal, QC
asia.biega@microsoft.com

Ben Carterette
Spotify
New York, NY
carteret@acm.org

ABSTRACT

We introduce the concept of *expected exposure* as the average attention ranked items receive from users over repeated samples of the same query. Furthermore, we advocate for the adoption of the principle of equal expected exposure: given a fixed information need, no item should receive more or less expected exposure than any other item of the same relevance grade. We argue that this principle is desirable for many retrieval objectives and scenarios, including topical diversity and fair ranking. Leveraging user models from existing retrieval metrics, we propose a general evaluation methodology based on expected exposure and draw connections to related metrics in information retrieval evaluation. Importantly, this methodology relaxes classic information retrieval assumptions, allowing a system, in response to a query, to produce a *distribution over rankings* instead of a single fixed ranking. We study the behavior of the expected exposure metric and stochastic rankers across a variety of information access conditions, including *ad hoc* retrieval and recommendation. We believe that measuring and optimizing expected exposure metrics using randomization opens a new area for retrieval algorithm development and progress.

CCS CONCEPTS

• **Information systems** → **Evaluation of retrieval results**; *Learning to rank*.

KEYWORDS

evaluation, fairness, diversity

ACM Reference Format:
Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340531.3411962

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3411962

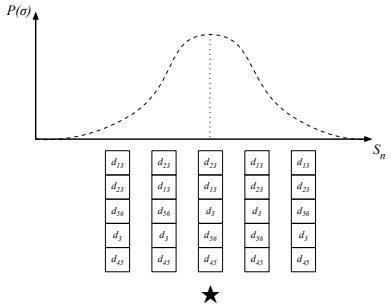

Figure 1: Distribution over rankings. Traditional evaluation methodologies consider only a single ranking (indicated by the ⋆) while stochastic rankers consider a distribution over rankings.

1 INTRODUCTION

Information access systems such as retrieval and recommendation systems often respond to an information need with a ranking of items. Even with more sophisticated information display modalities, the ranked list is a central feature of most interfaces. Since users often inspect a ranked list in a nonrandom–usually linear–order, some items are exposed to the user before others. Even if a system can perfectly model relevance and rank items accordingly, it still must put items in a particular order, breaking relevance ties in some way and reifying small differences in relative relevance into distinct rank positions.

Nonuniform exposure of relevant items resulting from ranking has multiple effects. It strongly affects the allocation of user attention (and therefore content exposure, visibility, and consumption-related revenue) to results and their producers, giving rise to fairness concerns for content producers [3, 6, 39]; if there are qualitative differences in comparably-relevant results, systematically favoring results preferred by one group of users affects other groups' quality of service [24] and may affect user retention [17]; similarly, in recall-oriented search or in scenarios where a searcher is interested in broad subtopical exposure, systematically promoting some relevant documents over others may risk overlooking important content;

it can homogenize users' information experiences and promote rich-get-richer effects [8]; and, although not often analysed in the design of algorithms, nonuniform exposure to relevant content may affect users' perception of the makeup of relevant information and its production community. There may also be difference of degree: if there is a small difference in the relative relevance of two documents, but a large difference in the attention users tend to pay to the positions in which they are ranked, the ranking may amplify small difference in content value into a large difference in the producers' return for providing that value.

Unfortunately, providing a static ranking for a query (in retrieval) or context (in recommendation) limits the ability for an algorithm to distribute exposure amongst relevant items. We propose to evaluate information access systems using *distributions over rankings* in response to a query or context. Figure 1 depicts this approach. More precisely, for a fixed query, we assume that a ranker, π, samples a permutation σ from a distribution over the set of all permutations S_n of n documents. This allows it to provide equal exposure to relevant items *in expectation*. And while current evaluation metrics and methods measure the relevance or utility of a single ranking per query, with a distribution of rankings, we can compute the expected value of the metric.

This paper provides the foundation for an exposure-based approach to evaluating rankings and advocates for exploring the family of stochastic ranking policies within that framework. To that end, we (i) define the concept of *expected exposure* and ways to operationalize it; (ii) discuss its relationship to existing retrieval metrics, including diversity, novelty, and fairness metrics; (iii) apply it to measure item exposure under stochastic versions of existing retrieval and recommendation algorithms. We argue that exposure provides a means of looking at several different concerns in the evaluation and impact of information access systems, and believe generalizing evaluation from deterministic rankers to stochastic rankers provides a broad area of study with implications for classic and contemporary problems in information access.

We begin by discussing the connection of previous work with our exposure-based evaluation and stochastic ranking (§2). We will then present the framework for evaluating with expected exposure and stochastic ranking together (§3). These definitions of expected exposure have deep connections to existing metrics, which we describe in §4. We then describe our experimental apparatus for analyzing these metrics in §5. We also propose a procedure to optimize towards these metrics in §6. We conclude with a discussion of our findings (§7).

2 RELATED WORK

Our work is inspired by and draws together two areas of work: (i) metrics recently developed in the context of algorithmic fairness, and (ii) randomized ranking algorithms developed in the context of online learning and optimization.

2.1 Fairness

Exposure optimization has been proposed as a means of achieving fairness in ranking: fairness for *individuals* means that exposure should be proportional to relevance for every subject in a system [3], while fairness for *groups* means that exposure should be equally distributed between members of groups defined by sensitive attributes such as gender or race [39]. From an optimization point of view, Singh and Joachims [40] and Yadav et al. [45] consider a similar notion of exposure fairness over multiple rankings as our work. Our work situates exposure-based measures in the context of information retrieval evaluation, allowing us to (i) extend them with user models from existing retrieval metrics, (ii) relate them with the objectives and formalisms of other retrieval metrics, and (iii) introduce a new experimentation protocols based on stochastic ranking.

Gao and Shah [15] recently proposed a randomized policy for diversifying search results very similar to our work, albeit in the context of group fairness. While studying connection between fairness and diversity empirically, we attempt to more formally elucidate the relationship and study broader connections beyond group fairness.

Beyond the definitions explicitly focusing on exposure, other fairness definitions in practice lead to enhanced equality of exposure, for instance, by requiring equal proportions of individuals from different groups in ranking prefixes [7, 47]. Similarly, Yang and Stoyanovich [46] measure fairness by computing sum of position-discounted set-wise parity at different rank thresholds. Beutel et al. [2] approach fair ranking by conducting pairwise analysis of user engagement with the protected groups in a ranking. Zehlike and Castillo [48] propose a supervised learning to rank method to optimize for fair exposure but focus only on the top position in ranking. It is not obvious how their proposed approach can be extended beyond the first rank position. In constrast to this literature, we study metrics that have clear user behavior model amenable to extension.

The notion of *meritocratic fairness*, originally introduced as a fairness objective for online bandit learning [20] and then applied to the problem of selecting a group of individuals from incomparable populations [21], intuitively requires that less qualified candidates do not have a higher chance of getting selected than more qualified candidates. In our setting, this translates to ensuring that less-relevant documents are not likely to be ranked above more-relevant documents. Our construct of target exposure connects this work to meritocratic fairness, in that a system satisfying equity of expected exposure will satisfy the goals of meritocratic fairness by allocating more exposure to relevant documents than to non-relevant documents, it also imposes a stronger constraint by requiring documents with comparable relevance to receive comparable exposure, preventing runaway popularity feedback loops that meritocratic fairness allows.

2.2 Stochastic Ranking

Randomization (either explicit or implicit) is ubiquitous in many information access systems and has been shown to be useful for eliciting user feedback and lead to desirable system properties. Pandey et al. [26] first proposed randomized ranking motivated by click exploration. Further strategies [18, 31, 32, 43] have been developed following this approach for collecting unbiased feedback for learning to rank. Instead of using randomization to collect unbiased training data, Joachims et al. [19] use it to estimate the parameters of a click propensity model that allows ranking models to be trained

on biased feedback. Using randomness in ranking may also be a means of improving diversity [33].

Recently, Bruch et al. [4] demonstrate that learning to rank models can be optimized towards expected values of relevance metrics computed over multiple rankings sampled based on estimated relevance. While not developed in the context of deploying a stochastic ranker, we adopt some of the methodologies therein in our experiments.

3 EXPECTED EXPOSURE

Given a query, we are interested in measuring the expected exposure of an item to a searcher with respect to items of similar relevance. Specifically, we would like to define a metric that quantifies a system's deviation from an ideal expected exposure of items of the same relevance. To this end, we adopt the following principle of *equal expected exposure*,[1]

> *Given a fixed information need, no item should be exposed (in expectation) more or less than any other item of the same relevance.*

This principle complements the existing core principle of ranked retrieval that more relevant documents should appear before less relevant documents. In this section, we will introduce an evaluation methodology based on the principle of equal expected exposure.

We note that existing relevance metrics do not measure the extent to which systems satisfy this principle, as they typically ignore differences in exposure amongst items of the same relevance. As a result, existing relevance metrics will not be able to distinguish a system that satisfies this principle from one that does not.

We will start by remaining agnostic about how items are exposed to searchers, only that there is some way in which searchers interact with a ranking of items that is related to the exposure. More formally, let a ranking be defined as a permutation of the n documents in the corpus. The set of all permutations of size n is referred to as the symmetric group or S_n in abstract algebra. Given a query $q \in Q$ with m relevant documents, an optimal permutation would place some ordering of the m relevant items at the top positions, followed by some ordering of the $(n - m)$ nonrelevant documents. Per existing models, exposure monotonically—often-exponentially—decreases with position in a ranking. Therefore, for a static ranking, we can see that (i) some relevant documents receive more exposure than other relevant documents, and (ii) some nonrelevant documents receive more exposure than other nonrelevant documents. A static ranking will therefore always violate equal expected exposure. Unfortunately, classic retrieval systems only provide and are evaluated according to static rankings.

However, we know that there are $m!(n - m)!$ optimal rankings. If an oracle provided us with an optimal ranking at random, any relevant item would be ranked in position $0 \leq i < m$ with the same probability.[2] As a result, all relevant items would receive the same exposure in expectation; similarly all nonrelevant items would receive the same exposure in expectation. Such a oracle would satisfy equal expected exposure. We will refer to the expected exposure of all items under the oracle policy as the *target exposure*, represented as a $n \times 1$ vector ϵ^*.

Just as we can satisfy ideal expected exposure by using a stochastic oracle, a retrieval system can improve the distribution of exposure by using a *stochastic policy*, a protocol where, in response to a query, a distribution over rankings is provided. Formally, given a query q, a ranking policy π provides a distribution over all permutations, $\sum_{\sigma \in S_n} \pi(\sigma|q) = 1$. Classic ranking algorithms are a special case which only assign probability to a single, static permutation. We will refer to such an algorithm as a *deterministic policy*. We note that most classic evaluation metrics (e.g. mean average precision) only evaluate a single, static permutation from a deterministic policy.

Given a policy π and a model of how the searcher might interact with a ranking, we can compute the expected exposure of all of the items in the corpus. We will represent the expected exposure of all items under π as a $n \times 1$ vector ϵ.

In order to measure the deviation from equal expected exposure, we compare the target exposure ϵ^* and sytem exposure ϵ. One simple way to do this is to compute the squared error between ϵ^* and ϵ,

$$\ell(\epsilon, \epsilon^*) = \|\epsilon - \epsilon^*\|_2^2 \tag{1}$$

$$= \underbrace{\|\epsilon\|_2^2}_{\text{EE-D}} \underbrace{- 2\epsilon^\mathsf{T}\epsilon^*}_{\text{EE-R}} + \|\epsilon^*\|_2^2 \tag{2}$$

where EE-D or *expected exposure disparity* measures inequity in the distribution of exposure; EE-R or *expected exposure relevance* measures how much of the exposure is on relevant documents; the remaining term is constant for a fixed information need.

This derivation allows us to clearly decompose expected exposure into a relevance and disparity components. A system that achieves optimal EE-R may maximize disparity (e.g. a static ranking with all relevant items at the top). Similarity a system that minimizes EE-D will have very bad expected exposure relevance (e.g. a random shuffling of the corpus every time a query is submitted).

We empirically observed (in §5.3) a tradeoff between the disparity (EE-D) and relevance (EE-R). This tradeoff is often controllable by a parameter in a stochastic policy that affects the degree of randomization. So, at one extreme, the parameter results in a deterministic policy that can achieve high relevance but also incurs high disparity. At the other extreme, the parameter results in a policy that randomly samples from amongst all permutations, achieving the lowest disparity but the lowest relevance. Given that such a parameter can often be swept between a minimum and maximum disparity, we can plot a disparity-relevance curve reflecting the nature of this tradeoff. We use the area under this curve, EE-AUC, as a summary statistic of this curve.

While we expect EE-R to behave similar to traditional relevance-based metrics–especially those sharing similar assumptions about how searchers interact with a ranking, reasoning about relevance and disparity within a single formalism allows us to compose aggregate metrics like EE-AUC, which traditional metrics do not capture (§5.3).

[1] This principle is related to *equity of attention* [3], which also ties exposure to relevance. However, *equity of attention* was originally amortized across information needs. While this paradigm accounts for changing relevance, the system might increase exposure of items for inappropriate information needs. Thus, in this paper we propose to measure exposure per information need. In this sense, the distinction between equal expected exposure and equity of attention is similar to the difference between macroaveraging and microaveraging of relevance metrics.

[2] Note that we use base-0 ranks throughout this manuscript.

3.1 Computing Exposure with User Browsing Models

So far, we have remained agnostic about how items are exposed to users. In this section, we will describe how we can compute the exposure vector ϵ for an arbitrary ranker, including the oracle ranker. Unlike previous fair ranking metrics, we approach exposure by adopting user models from existing information retrieval metrics. We focus on models from two metrics, rank-biased precision and expected reciprocal rank, although this analysis can be extended to more elaborate browsing models [12].

Rank-biased precision (RBP) is a metric that assumes that a user's probability of visiting a position decreases exponentially with rank [25],

$$\text{RBP}(\sigma) = (1-\gamma) \sum_{i \in [0,k)} \mathbf{y}^*_{\sigma_i} \gamma^i \tag{3}$$

where $\mathbf{y}^*$ is the $n \times 1$ binary relevance vector; γ is referred to as the *patience parameter* and controls how deep in the ranking the user is likely browse; and k is the maximum browsing depth. The multiplicative factor $1-\gamma$ ensures that the measure lies in the unit range.

We consider that the expected exposure of a document d is computed, in expectation, as,

$$\epsilon_d = \sum_{\sigma \in S_n} \pi(\sigma|q) \gamma^{\overline{\sigma}_d} \tag{4}$$

where $\overline{\sigma}$ is a map from document indexes to ranks. This allows us to compute ϵ for an arbitrary policy π.

Recall that the oracle policy selects randomly amongst all rankings with all of the relevant documents at the top. Since each document occurs at each of the top m positions equally, the target expected exposure for a relevant document is,

$$\begin{aligned} \epsilon^*_d &= \frac{1}{m} \sum_{i \in [0,m)} \gamma^i \\ &= \frac{1-\gamma^m}{m(1-\gamma)} \end{aligned}$$

Since the set of nonrelevant documents is usually very large, all nonrelevant documents will have equal expected exposure close to zero.

Expected reciprocal rank (ERR) is a metric that assumes that a user's probability of visiting a position is dependent on how many relevant documents appear a earlier positions [9]. The intuition is that earlier relevant documents may satisfy the user and prompt them to stop scanning the ranking. We adopt generalized expected reciprocal rank, a model which incorporates a patience parameter similar to that used in RBP [9, §7.2].

$$\text{ERR}(\sigma) = \sum_{i \in [0,k)} \phi(\mathbf{y}^*_{\sigma_i}) \prod_{j \in [0,i)} \gamma(1 - \phi(\mathbf{y}^*_{\sigma_j})) \tag{5}$$

where ϕ converts relevance to a probability of stopping the browsing. Normally this is zero for nonrelevant documents and some value between 0 and 1 for relevant documents. As with RBP, the expected exposure of document d can be computed as,

$$\epsilon_d = \sum_{\sigma \in S_n} \pi(\sigma|q) \gamma^{\overline{\sigma}_d} \prod_{j \in [0, \overline{\sigma}_d)} (1 - \phi(\mathbf{y}^*_{\sigma_j}))$$

Similarly, the target expected exposure of a relevant document is,

$$\begin{aligned} \epsilon^*_d &= \frac{1}{m} \sum_{i \in [0,m)} \gamma^i (1 - \phi(\mathbf{y}^*_{d^*}))^i \\ &= \frac{1 - \gamma^m (1 - \phi(\mathbf{y}^*_{d^*}))^m}{m(1 - \gamma(1 - \phi(\mathbf{y}^*_{d^*})))} \end{aligned}$$

and close to zero for nonrelevant documents.

3.2 Extension to Graded Judgments

So far, we have focused on binary relevance. For graded judgments, the ideal ranker always orders documents correctly by grade. We take all permutations satisfying this requirement and assume the ideal ranker has nonzero support only for these values. We then compute the expected exposure for documents by grade. Let m_g be the number of documents with relevance grade g and $m_{>g}$ the number of documents with relevant grade strictly larger than g. Without loss of generality, assume that grades take integer values. Given an RBP browsing model, the optimal exposure for a document d with grade g is,

$$\begin{aligned} \epsilon^*_d &= \frac{1}{m_g} \sum_{i \in [m_{>g}, m_{>g-1})} \gamma^i \\ &= \frac{\gamma^{m_{>g}} - \gamma^{m_{>g-1}}}{m_g(1-\gamma)} \end{aligned}$$

The derivation for the ERR user model is similar.

We note that this extension assumes that the a searcher will always prefer to see items with higher grade. In situations where, for example, the grade of an item is inversely correlated with some important property of a document (e.g. a subtopic, authors from underrepresented groups), then these groups will be under-exposed. In such cases, an alternative definition of ϵ^* may be more appropriate (see §4.2).

4 RELATIONSHIP TO OTHER METRICS

Expected exposure, both in motivation and in definition, has connections to existing retrieval metrics. In this section, we will discuss those relationships, highlighting the unique properties that expected exposure measures.

4.1 Retrieval Metrics

Measures such as RBP and ERR could be considered *precision metrics*, as they reward rankers for retrieving relevant material higher in the ranking. While based on the same user model, it is *not* the case that optimizing RBP will also minimize Equation 1, even if exposure is based on an RBP browsing model. To see why, consider a deterministic policy that outputs a static optimal ranking. Although EE-R will be optimal, EE-D will be very large since exposure is concentrated at the top ranks. Indeed, the value of EE-D for a static optimal ranking will be as bad as a static ranking that places all of the relevant document at the bottom since disparity is based only on the exposure and not on relevance. The converse, that minimizing Equation 1 also optimizes RBP, *is* true. If expected exposure is based on the RBP user model, a system that optimizes expected exposure will essentially be shuffling relevant documents at the top of the

ranking and nonrelevant items in the bottom, just as with the oracle in §3.

Optimizing recall means focusing on ensuring that all of the relevant items in the corpus occur at high ranks. Several of our motivating examples might be considered addressable by a retrieval system optimized for high recall (e.g. e-discovery, academic search, systematic review). However, if we assume, as many user models do, that a user may terminate their scan of a ranking early, then there is a chance that even a high-recall system, especially in situation where there are numerous relevant documents, a user will not be exposed to all relevant items. As a result, we would argue that expected exposure reduces the risk of overlooking a relevant document.

4.2 Fairness

Algorithmic fairness, in the context of information retrieval and recommendation, deals with the treatment of individuals associated with retrievable items [6]. These might be document authors in text retrieval, job candidates in recruiting, or musicians in song recommendation.

Individual Fairness. Expected exposure is closely related to various notions of individual fairness that quantify the extent to which models are fair to all individuals. Dwork et al. defined individual fairness in the context of classification models seen as mappings from individuals to probability distributions over outcomes [13]. In this setting, individual fairness is defined using the Lipschitz condition: the distributions of classification outcomes P_C of two individuals u_1, u_2 who are sufficiently similar according to a chosen similarity metric d should be close according to a distribution similarity metric D. Formally, if $d(u_1, u_2) < \delta$, then $D(P_C(u_1), P_C(u2)) < \Delta$. When will a retrieval policy be individually fair according to this definition? Assume we define δ and d such that two documents of equal relevance grade satisfy the above inequality, and two documents of different relevance grades do not. Assume furthermore that outcomes are measured as the expected exposure of individual documents. A stochastic ranker that distributes exposure (almost) equally among the documents of equal relevance grades (in particular if it achieves optimal expected exposure according to Eq. 1) is individually fair according to the above definition. However, the reverse does not hold: It is possible that an individually fair and an unfair stochastic rankers lead to similar values of the expected exposure measure (the total loss value in Eq. 1 can be aggregated equitably from documents of the same relevance level or from only few documents within a relevance grade).

Group Fairness. We can use exposure to define a group notion of provider fairness by measuring whether deviation from expected exposure differs between different groups of documents (or their authors). Let $\mathcal{A}$ be the set of k attributes that a document might be associated with. Attributes may be related to, for example, demographic or other group information about the provider. Let $\mathbf{A}$ be a $n \times k$ binary matrix of the group identity associated with each document in the corpus. We can then compute the total exposure for all documents with an attribute by $\xi_{\mathbf{A}} = \mathbf{A}^\mathsf{T}\epsilon$. If we are interested in equal exposure across groups, we can define the target group exposure as $\xi^*_{\mathbf{e}} = c\mathbf{A}^\mathsf{T}\mathbf{e}$ where $\mathbf{e}$ is a $k \times 1$ vector of ones and c is a normalizing constant based on the total exposure given a browsing model. We can then use Equation 1 as a measure of *demographic parity* [39, §4.1]. If desired, we can replace $\mathbf{e}$ with some other distributions, such as population level proportions [37]. Target exposures like $\xi^*_{\mathbf{e}}$ only balance group representation, but some groups may produce more relevant content than others. If we are interested in exposure proportional to relevance, we can define the target exposure as $\xi^* = \mathbf{A}^\mathsf{T}\mathbf{y}^*$, referred to as *disparate treatment* [39, §4.2]. Finally, if we are interested in ensuring the exposed items are relevant, we can define a new matrix $\tilde{\mathbf{A}} = \text{diag}(\mathbf{y}^*)\mathbf{A}$ and exposure vector $\xi_{\tilde{\mathbf{A}}} = \tilde{\mathbf{A}}^\mathsf{T}\epsilon$. If we let $\xi^*_{\tilde{\mathbf{A}}} = c\tilde{\mathbf{A}}^\mathsf{T}\mathbf{y}^*$, then we recover *disparate impact* [39, §4.3].

4.3 Topical Diversity

Exposure metrics are closely related to topical diversity metrics [36]. One common way to measure topical diversity is to consider so-called 'intent-aware' metrics defined as,

$$\text{IA-}\mu(\sigma) = \sum_{a \in \mathcal{A}} p(a|q)\mu(\sigma|a)$$

where $\mu(\sigma|a)$ computes a standard metric considering only those documents with aspect a as relevant. The intent-aware RBP metric is defined as

$$\text{IA-RBP}(\sigma) = \sum_{a \in \mathcal{A}} p(a|q)(1-\gamma) \sum_{i \in [0,k)} \mathbf{y}^{a,*}_{\sigma_i}\gamma^i$$

If we assume that $p(\mathcal{A}|q)$ is proportional to the frequency of a in the set of relevant documents, then $\text{IA-RBP}(\sigma) \propto \xi^\mathsf{T}_{\tilde{\mathbf{A}}}\xi^*_{\tilde{\mathbf{A}}}$. In other words, topic diversity reduces to a scaled relevance term in the disparate impact metric (§4.2). In the event that we are interested in uniform $p(a|q)$, then we can redefine the target exposure accordingly and recover the relevance term in the demographic parity metric. Both of these formulations ignore EE-D and it is worth observing that intent-aware metrics often include a 'subtopic recall' factor to [35] to ensure that all subtopics are retrieved. We believe that the disparity term captures precisely this behavior.

5 METRIC ANALYSIS

We are interested in empirically studying the EE-D and EE-R. Specifically, we will answering the following questions in our experiments: (i) can the metric distinguish between different randomization strategies? (ii) does an exposure-based relevance metric measure something different from a static ranking metric based on the same user model?

5.1 Randomizing a Deterministic Policy

The focus of this paper is on evaluation. However, we were interested in studying our metrics for stochastic rankers, which are not readily available outside of specialized online learning environments. As such, we developed several stochastic rankers for our experiments based on post-processing a precomputed set of retrieval scores.

Plackett-Luce (PL). Our first randomization strategy uses Plackett-Luce sampling to sample a permutation [22, 28]. To do this, we create a multinomial distribution $p(d|q)$ over the corpus using the

ℓ_1 normalization of the retrieval scores. The Plackett-Luce model samples a permutation by first sampling the document at position 0 using $p(d|q)$. We then set the probability of the selected document to 0, renormalize, and sample the document at position 1 from this modified distribution. We continue this process until we exhaust the scored documents. In order to control the randomness of the process, we use a modified sampling distribution,

$$p(d|q) = \frac{\mathrm{y}_d^\alpha}{\sum_{d'} \mathrm{y}_{d'}^\alpha}$$

where $\alpha \geq 0$. When $\alpha = 0$, all permutations are equally likely and EE-D is minimized; as α increases π concentrates around original static ranking and disparity degrades. We refer to this as the Plackett-Luce (PL) policy.

Rank Transpositions (RT). Our second randomization strategy ignores the retrieval scores and samples permutations by shuffling the original ranked list. We shuffle by repeatedly sampling pairs of positions and swapping the documents. Such a process takes $\frac{1}{2} n \log n + cn$ iterations to converge to sampling a random permutation [11]. This is precisely a random walk on S_n where permutations are connected by pairwise transpositions. As such, we can introduce a 'restart probability' to teleport the random walker back to the original ranked list. If this probability is θ, then the number of steps of the random walk follows a geometric distribution with support $[0, \infty)$. Our randomization strategy then first samples the number of steps k from the geometric distribution and then conducts k random transpositions. We refer to this as the rank transposition (RT) policy.

These two methods are intentionally constructed to perform differently. The PL policy takes a deterministic policy's scores into consideration and will, therefore, be more conservative in removing high-scoring items from the top of the ranked list. The RT policy, on the other hand, randomly swaps pairs, regardless of score or position. As a result, we suspect that the PL policy should outperform the RT policy, given a fixed base deterministic policy.

5.2 Method

We analyze the behavior of expected exposure metrics using the postprocessing of deterministic policies in two domains. The first is based on archival TREC submissions focus in information retrieval conditions. The Robust2004 dataset consists of 440 runs submitted to the TREC 2004 Robust track which evaluated systems on a set of 249 queries and binary relevance labels. We adopt this dataset because it has been well-studied in the context of evaluation metrics.

Our second dataset, MovieLens25M, is a movie recommendation dataset consisting of 25 million ratings of 59 thousand movies by 163 thousand users [16]. We used LensKit 0.8.4 [14] to generate runs representing binary implicit-feedback matrix factorization (IMF) [27] and Bayesian personalized ranking (BPR) [34].[3] We adopt implicit feedback instead of ratings in order to study the behavior of expected exposure under binary relevance.

We use a $\gamma = 0.50$ for all of our experiments, as consistent with standard TREC evaluation protocol. RBP and ERR are evaluated at depth 20. For stochastic rankers, we sample 50 rankings during evaluation to estimate expected exposure metrics. We found that this was sufficient to converge to appropriate expected metric values. Experiments randomizing deterministic policies rerank the top 100 documents from the original static ranking.

[3] BPR is implemented by the `implicit` package (https://github.com/benfred/implicit).

5.3 Results

Before analyzing our metrics in aggregate, we present our metrics on an example run from Robust2004. In Figure 2, we show the behavior of our randomization model for EE-R and EE-D, under both the ERR and RBP user models. We compare these metrics to RBP and ERR, two classic static ranking metrics. We also measure the generalized entropy of exposure on the relevant set of documents [41]; this allows us to assess the disparity amongst relevant items.

Comparing classic metrics and EE-R in the first and second rows, we observe correlated behavior as randomization changes. Across a sample of runs, we found that the expected RBP and EE-R were strongly correlated ($r = 0.99$, $p < 0.01$); a perfect correlation was observed between expected ERR and EE-R with an ERR model. This is unsurprising given that the relevance factor in the expected exposure metric is precisely the expectation of the static ranking metric. The imperfect correlation for RBP is due to normalization term in the classic RBP model.

Comparing generalized entropy and EE-D, we also observe correlated behavior across both RBP ($r = 0.47$, $p < 0.01$) and ERR user models ($r = 0.65$, $p < 0.01$). Because the generalized entropy is computed over only relevant documents, this suggests that EE-D is sensitive to changes in expected exposure to relevant documents, not the dominant, nonrelevant set.

Comparing the behavior of EE-D and EE-R in Figure 2, we notice the disparity-relevance tradeoff mentioned in §3. In order to visualize this tradeoff more clearly, we present example disparity-relevance curves for randomization of an arbitrary Robust2004 run and our two recommender systems on MovieLens25M in Figure 3. Disparity-relevance curves, when randomizing the same base policy, will have the same value of EE-R for EE-D = 1 because this recovers the original static ranking. Similarly, all disparity-relevance curves begin with EE-R = 0 at EE-D = 0 because a completely random ranker will achieve minimal relevance by dint of the number of nonrelevant documents in the corpus (i.e. a random shuffle will mean that, in expectation, every document receives a tiny amount of attention). Turning to the randomization policies being studied, across both domains and multiple runs, PL randomization policies dominate RT policies across all disparity points, confirming our intuition that incorporating score information improves post-processing performance. This provides us with the ability to test the ability of exposure to distinguish between stochastic policies.

Given two stochastic rankers, we are interested in understanding whether our exposure metrics can more accurately identify the superior algorithm compared to a metric based on a static ranking. To that end, we randomly assigned the runs for the Robust2004 dataset to either PL or RT randomization. This provided us with EE-AUC for each run as well as an RBP value for its base deterministic policy. We ordered runs by the RBP and then inspected the EE-AUC for the associated run. In Figure 4, we can see that, while RBP, a metric based on a static ranking, can approximately order runs for a fixed randomization policy ($\tau = 0.89$, $p < 0.05$), it cannot distinguish between the PL and RT policies.

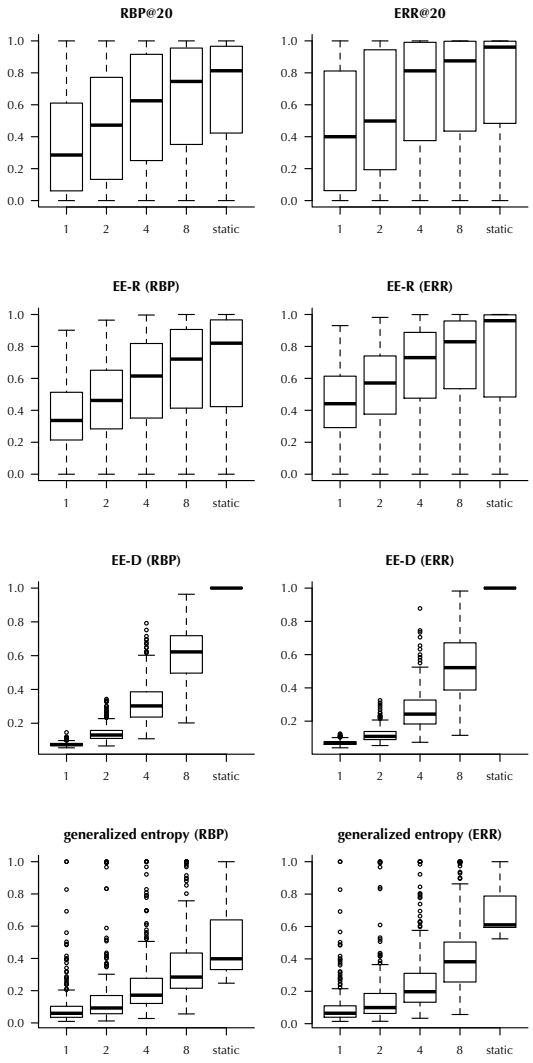

Figure 2: Behavior of expected exposure metrics for a deterministic run from the Robust2004 **dataset randomized using the Plackett-Luce model. Each horizontal axis indicates the value of α, where lower values indicate more randomization. Each vertical axis reflects the performance of policies on static ranking relevance metrics (top row), expected exposure relevance metrics (second row), expected exposure disparity metrics (third row), and generalized entropy on the relevant set of documents (fourth row) using two browsing models (left: RBP; right: ERR).**

6 OPTIMIZING FOR EXPECTED EXPOSURE

In the previous section, we introduced post-processing techniques to build stochastic rankers. Given a model that is perfectly able to predict relevance, Plackett-Luce randomization should perform optimally, especially for binary relevance. As such, a classic pointwise learning to rank model [10] with Plackett-Luce randomization may

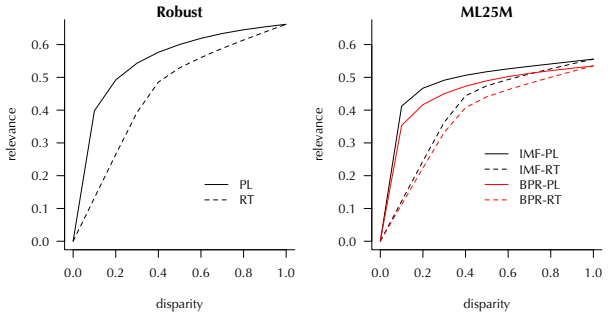

Figure 3: Disparity-relevance tradeoff curve for a random Robust2004 **run and our two recommendation runs on** MovieLens25M **with Placket-Luce randomization and rank transposition randomization.**

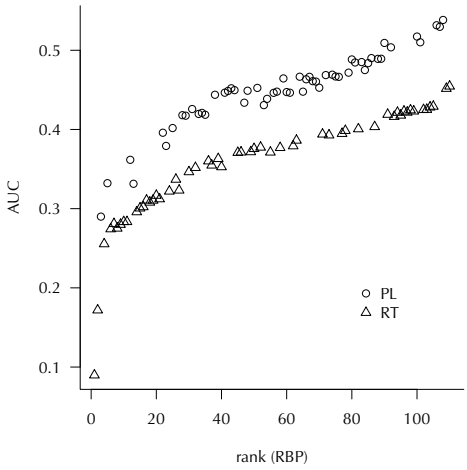

Figure 4: Sorting PL and RT runs by RBP. Half of the runs submitted to Robust2004 **were subjected to PL randomization and half to RT randomization. Runs were ranked by the RBP of the original, static ranking. EE-AUC for the randomized runs, according to its treatment, is plotted on the vertical axis.**

be an effective approach for expected exposure. Moreover, calibration of relevance does not happen with pairwise learning to rank models [5] and so we would expect these models, even if perfect, to perform worse than pointwise models, even with Plackett-Luce randomization. However, learning to rank models are not perfect estimators of relevance. Therefore, we believe there should be some advantage to optimizing directly for expected exposure.

In this section, we will examine the relationship between the performance of these approaches in the context of graded relevance as well as demographic parity (§4.2). We focused on a shared model architecture with varying loss functions in order to measure differences due to the objective alone, instead of artifacts resulting from the functional form of the models. We begin by describing how we optimize for expected exposure before proceeding to our empirical results.

6.1 Algorithm

Although optimizing for pointwise or pairwise loss has been well-studied in the information retrieval community, directly optimizing for a metric based on a distribution over rankings has received less attention.

We begin by defining an appropriate loss function for our model. Turning to Equation 1, we can drop the constant term and add a hyperparameter to balance between disparity and relevance,

$$\ell_\lambda(\epsilon, \epsilon^*) = \lambda\|\epsilon\|_2^2 - (1-\lambda)\epsilon^\mathsf{T}\epsilon^* \tag{6}$$

where ϵ^* is based on graded relevance (§3.2).

Let $f_\theta : \mathcal{D} \to \mathfrak{R}$ be an item scoring function parameterized by θ. Given a query, $\mathbf{y}$ is a $n \times 1$ vector of item scores for the entire collection such that, $\mathbf{y}_d = f_\theta(d)$. Using a Plackett-Luce model, we can translate the raw scores into sampling probabilities,

$$p(d) = \frac{\exp(\mathbf{y}_d)}{\sum_{d' \in \mathcal{D}} \exp(\mathbf{y}_{d'})}$$

This allows us to construct a ranking σ by sampling items sequentially. Unfortunately, this sampling process is non-differentiable and, therefore, prohibitive to a large class of models, including those that learn by gradient descent. We address this by adopting the method proposed by Bruch et al. [4]. To construct a sampled ranking σ, we reparameterize the probability distribution by adding independently drawn noise samples G from the Gumbel distribution [23] to $\mathbf{y}$ and sorting items by the "noisy" probability distribution $\tilde{p}$,

$$\tilde{p}(d_i) = \frac{\exp\left(\mathbf{y}_{d_i} + G_i\right)}{\sum_{d_j \in \mathcal{D}} \exp\left(\mathbf{y}_{d_j} + G_j\right)} \tag{7}$$

Given the perturbed probability distribution $\tilde{p}$, we compute each document's smooth rank [30, 44] as,

$$\overline{\sigma}_d = \sum_{d' \in \mathcal{D}/d} \left(1 + \exp\left(\frac{\tilde{p}(d) - \tilde{p}(d')}{\tau}\right)\right)^{-1} \tag{8}$$

The smooth rank is sensitive to the temperature τ. At high temperatures the smooth rank is a poor approximation of the true rank and at low temperatures may result in vanishing gradients. To rectify this issue, we employ the straight-through estimator [1] to compute the true ranks in forward pass but differentiating the gradients with respect to the smooth ranks during backpropagation.

Using the estimated ranks and a specified user model we compute the exposure for each document. For example, assuming RBP as the user model the exposure of document d from a single ranking σ is given by $\epsilon_d = \gamma^{\overline{\sigma}_d}$. We compute expected exposure by averaging over n_{train} different rankings—each generated by independently sampling different Gumbel noise in Equation 7.

We use this expected exposure vector ϵ in Equation 6 to compute the loss that we minimize through gradient descent. The relevance grades are not used for training beyond computing target exposure. We set τ in Equation 8 to 0.1.

We can adapt this model to optimize group-level exposure metrics like demographic parity (§4.2). To do so, we replace $\|\epsilon\|_2^2$ with $\|\xi\|_2^2$ in Equation 6 to define an optimization objective that trades-off relevance and demographic parity.

$$\ell_{\text{group},\lambda} = \lambda\|\xi\|_2^2 - (1-\lambda)\epsilon^\mathsf{T}\epsilon^* \tag{9}$$

This loss function assumes that the ideal policy distributes exposure equally across all demographics.

6.2 Experiment

Models. We restrict our choice of baselines to neural networks so that the exposure-based optimization can be compared to baseline ranking loss functions with respect to the same model. Our base model consists of a fully-connected neural network with two hidden layers of size 256 nodes per layer and rectified linear unit for activation function. We choose a learning rate of 0.001 and a dropout rate of 0.1 and perform early-stopping for all models based on validation sets. Stochastic rankings are then derived by employing Plackett-Luce sampling over these deterministic policies (i.e. pointwise and pairwise models), with varying softmax temperatures to obtain different trade-off points between disparity and relevance. We set n_{train} to 20 for our model and n_{test} to 50 for all models.

Objectives. We consider three models in our experiments. The pointwise model minimizes the squared error between the model prediction and true relevance. The pairwise model minimizes misclassified preferences using a cross-entropy loss. The expected exposure model minimizes the loss in Equation 6 and, in our demographic parity experiments, Equation 9.

Data. Our experiments use the MSLR-WEB10k dataset [29], a learning-to-rank dataset containing ten thousand queries. We perform five-fold cross validation (60/20/20 split between training, validation, and testing sets). Each query-document pair is represented by a 136-dimensional feature vector and graded according to relevance on a five point scale. For the demographic parity experiments, we discretize the PageRank feature in the ranges <1000, 1000–10000, and ≥10000 and treat it as a demographic attribute. We confirm that this discretization scheme is reasonable as roughly 70% of the queries have at least one document corresponding to each demography with a relevance grade greater than one.

6.3 Results

We present the results of our experiments in Table 1.

In terms of expected exposure, we did not observe a difference in performance between pointwise and pairwise models. However, directly optimizing for expected exposure resulted in a 3.9% improvement in EE-AUC over the pointwise and pairwise models. We confirm that the difference in EE-AUC follows a normal distribution and accordingly perform a paired student's t-test to check their statistical significance. The EE-AUC differences between our proposed method and the baselines are statistically significant ($p < 0.01$).

In terms of demographic parity, we observe a difference in performance between pointwise and pairwise models. Moreover, directly optimizing for expected exposure results in improved performance while directly optimizing for demographic parity further boosts performance. The gap in EE-AUC between all pairs of models are statistically significant ($p < 0.01$) in this case.

Table 1: Results for optimizing towards expected exposure and demographic parity using different ranking objectives. We report average EE-AUC for both tasks and highlight the best performance for each in bold. Optimizing directly for expected exposure and demographic parity using our proposed method achieves best performance in both cases.

Loss function	AUC	
	Expected exposure	Demographic parity
Pointwise loss	0.229	0.112
Pairwise loss	0.229	0.108
Our methods		
Expected exposure	**0.238**	0.141
Demographic parity		**0.178**

7 DISCUSSION

Our theoretical results draw clear connections to several areas of information retrieval research. We believe, moreover, that our empirical results suggest that expected exposure metrics capture important aspects of a retrieval system that are not currently measured in information retrieval evaluation. Our experiments furthermore demonstrated that these metrics are not only effective for distinguishing systems with varying degrees of expected exposure but also that they can be optimized toward.

Although previously studied in the context of algorithmic fairness, we have demonstrated that there are deep connections to existing core areas of information retrieval research. These results warrant revisiting algorithms and results in classic tasks such as *ad hoc* retrieval, legal search, and diversity-sensitive retrieval.

Beyond relevance, fairness, and diversity, we believe this approach to evaluation opens avenues for studying probabilistic search systems in probabilistic way. Many search systems are defined as probabilistic models, capable of handling uncertainty about document relevance [49], sometimes using online learning to refine scoring and ranking models and adapt to changing information needs. These models produce rankings in accordance with a probabilistic policy, so they naturally result in a distribution over rankings associated with each query. Expected exposure, along with computing expected values of other information retrieval metrics, provides a way to evaluate these models and study the effects of uncertainty. Moreover, modern search engines also randomize their rankings to reduce bias in feedback data [18]. Although these systems are often evaluated log data and off-policy evaluation techniques, in the case of pre-launch batch evaluation, we can explicitly model the impact of randomization by evaluating the distribution over rankings.

Randomization and improving equal expected exposure may also help with user retention. In search systems, we often want to make sure that we do not overemphasize dominant intents, which can often homogenize populations [17, 24]. As such, randomization can allow us to balance exposure across heterogeneous intents. Exposure balancing may also prevent churn caused by starvation of producers in two-sided economy systems such as ride-sharing platforms [42].

Our exposure model is flexible enough to incorporate more elaborate browsing models. Several exist others beyond RBP and ERR exist in the literature for rankings which deserve exploration. Furthermore, as searchers begin to interact with interfaces that are not based on rankings (e.g. two-dimensional grids, three-dimensional environments), alternative user models will need to be developed and incorporated.

We would also like to note possible limitations of this approach. First, the impact of randomization on user satisfaction is still an active area of research and we believe cumulative effects of randomization may be a novel extension to explore in the future work [38]. Second, from an evaluation perspective, stochastic policies introduce logistical constraints on distribution representation and permutation sampling. Centralized evaluations like TREC would need to support a method for either interrogating a stochastic policy or requiring a large pool of samples, incurring data storage costs. Third, although we have focused on randomization in order to increase exposure, we believe that drawing a connection to sequential decision-making scenarios like amortized evaluation are exciting areas of future work.

Notwithstanding these limitations, evaluation through expected exposure, when coupled with stochastic policies, opens a new perspective for the study, understanding, and design of information retrieval systems.

8 ACKNOWLEDGEMENTS

Michael Ekstrand's contribution to this work was supported by the National Science Foundation under Grant No. IIS 17-51278.

REFERENCES

[1] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432* (2013).
[2] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in Recommendation Ranking Through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. 2212–2220.
[3] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. 405–414.
[4] Sebastian Bruch, Shuguang Han, Mike Bendersky, and Marc Najork. 2020. A Stochastic Treatment of Learning to Rank Scoring Functions. In *Proceedings of the 13th ACM International Conference on Web Search and Data Mining (WSDM 2020)*.
[5] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*. 89–96.
[6] Robin Burke. 2017. Multisided Fairness for Recommendation. (July 2017). arXiv:cs.CY/1707.00093
[7] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2018. Ranking with Fairness Constraints. In *45th International Colloquium on Automata, Languages, and Programming, ICALP 2018, July 9-13, 2018, Prague, Czech Republic*. 28:1–28:15.
[8] Allison J B Chaney, Brandon M Stewart, and Barbara E Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18)*. 224–232.
[9] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09)*. 621–630.
[10] David Cossock and Tong Zhang. 2006. Subset ranking using regression. In *International Conference on Computational Learning Theory*. Springer, 605–619.
[11] Persi Diaconis and Mehrdad Shahshahani. 1981. Generating a random permutation with random transpositions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 57, 2 (Jun 1981), 159–179.

[12] Fernando Diaz, Ryen W. White, Georg Buscher, and Dan Liebling. 2013. Robust Models of Mouse Movement on Dynamic Web Search Results Pages. In *Proceedings of the 22nd ACM conference on Information and knowledge management (CIKM 2013)*. 1451–1460.

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. 214–226.

[14] Michael D. Ekstrand. 2020. LensKit for Python: Next-Generation Software for Recommender System Experiments. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. ACM.

[15] Ruoyuan Gao and Chirag Shah. 2020. Toward creating a fairer ranking in search engine results. *Information Processing & Management* 57, 1 (2020), 102138.

[16] F Maxwell Harper and Joseph A Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems* 5, 4 (Dec. 2015), 19:1–19:19.

[17] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness Without Demographics in Repeated Loss Minimization. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. 1929–1938.

[18] K. Hofmann. 2013. *Fast and Reliable Online Learning to Rank for Information Retrieval*. phd. University of Amsterdam.

[19] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 781–789.

[20] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in Learning: Classic and Contextual Bandits. In *Advances in Neural Information Processing Systems 29*, D D Lee, M Sugiyama, U V Luxburg, I Guyon, and R Garnett (Eds.). 325–333.

[21] Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. 2017. Meritocratic Fairness for Cross-Population Selection. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. 1828–1836.

[22] R. Duncan Luce. 1959. *Individual Choice Behavior*.

[23] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables.

[24] Rishabh Mehrotra, Amit Sharma, Ashton Anderson, Fernando Diaz, Hanna Wallach, and Emine Yilmaz. 2017. Auditing Search Engines for Differential Satisfaction Across Demographics. In *Proceedings of the 26th International Conference on World Wide Web*.

[25] Alistair Moffat and Justin Zobel. 2008. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* 27, 1 (Dec. 2008), 2:1–2:27.

[26] Sandeep Pandey, Sourashis Roy, Christopher Olston, Junghoo Cho, and Soumen Chakrabarti. 2005. Shuffling a Stacked Deck: The Case for Partially Randomized Ranking of Search Engine Results. In *VLDB*. 781–792.

[27] István Pilászy, Dávid Zibriczky, and Domonkos Tikk. 2010. Fast Als-based Matrix Factorization for Explicit and Implicit Feedback Datasets. In *Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys '10)*. 71–78.

[28] R. L. Plackett. 1975. The Analysis of Permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 24, 2 (1975), 193–202.

[29] Tao Qin and Tie-Yan Liu. 2013. Introducing LETOR 4.0 Datasets. *CoRR* abs/1306.2597 (2013).

[30] Tao Qin, Tie-Yan Liu, and Hang Li. 2010. A general approximation framework for direct optimization of information retrieval measures. *Information retrieval* 13, 4 (2010), 375–397.

[31] Filip Radlinski and Thorsten Joachims. 2006. Minimally Invasive Randomization for Collecting Unbiased Preferences from Clickthrough Logs. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2 (AAAI'06)*. 1406–1412.

[32] Filip Radlinski and Thorsten Joachims. 2007. Active exploration for learning rankings from clickthrough data. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 570–579.

[33] F. Radlinski, R. Kleinberg, and T. Joachims. 2008. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, Andrew McCallum and Sam Roweis (Eds.). 784–791.

[34] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI '09)*. 452–461.

[35] Tetsuya Sakai and Ruihua Song. 2011. Evaluating Diversified Search Results Using Per-Intent Graded Relevance. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. 1043–1052.

[36] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2015. *Search Result Diversification*. Foundations and Trends in Information Retrieval, Vol. 9. Now Publishers Inc.

[37] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attentionon Fair Group Representation in Ranked Lists. In *Companion Proceedings of The 2019 World Wide Web Conference on - WWW '19*. 553–562.

[38] Tobias Schnabel, Paul N. Bennett, Susan T. Dumais, and Thorsten Joachims. 2018. Short-Term Satisfaction and Long-Term Coverage: Understanding How Users Tolerate Algorithmic Exploration. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. 513–521.

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

[40] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for Fairness in Ranking. In *Advances in Neural Information Processing Systems*.

[41] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. 2239–2248.

[42] Tom Sühr, Asia J Biega, Meike Zehlike, Krishna P Gummadi, and Abhijnan Chakraborty. 2019. Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 3082–3092.

[43] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. 2016. Learning to rank with selection bias in personal search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 115–124.

[44] Mingrui Wu, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. 2009. Smoothing DCG for learning to rank: A novel approach using smoothed hinge functions. In *Proc. CIKM*. ACM, 1923–1926.

[45] Himank Yadav, Zhengxiao Du, and Thorsten Joachims. 2019. Fair Learning-to-Rank from Implicit Feedback. *arXiv preprint arXiv:1911.08054* (2019).

[46] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management (SSDBM '17)*.

[47] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. 1569–1578.

[48] Meike Zehlike and Carlos Castillo. 2020. Reducing Disparate Exposure in Ranking: A Learning To Rank Approach. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (Eds.). 2849–2855.

[49] Jianhan Zhu, Jun Wang, Ingemar J. Cox, and Michael J. Taylor. 2009. Risky business: modeling and exploiting uncertainty in information retrieval. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 99–106.