Fairness in Recommendation Ranking through Pairwise Comparisons

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao,
Lichan Hong, Ed H. Chi, Cristos Goodrow
alexbeutel,jilinc,tulsee,hqian,liwei,wuyish,heldt,zhezhao,lichan,edchi,cristos@google.com
Google

ABSTRACT

Recommender systems are one of the most pervasive applications of machine learning in industry, with many services using them to match users to products or information. As such it is important to ask: what are the possible fairness risks, how can we quantify them, and how should we address them?

In this paper we offer a set of novel metrics for evaluating algorithmic fairness concerns in recommender systems. In particular we show how measuring fairness based on pairwise comparisons from randomized experiments provides a tractable means to reason about fairness in rankings from recommender systems. Building on this metric, we offer a new regularizer to encourage improving this metric during model training and thus improve fairness in the resulting rankings. We apply this pairwise regularization to a large-scale, production recommender system and show that we are able to significantly improve the system's pairwise fairness.

1 INTRODUCTION

What should we expect of recommender systems? Recommenders are pivotal in connecting users to relevant content, items or information throughout the web, but with both users and content producers, sellers or information providers relying on these systems, it is important that we understand who is being supported and who is not. In this paper we focus on the risk of a recommender system under-ranking groups of items [9, 21, 39]. For example, if a social network under-ranked posts by a given demographic group, that could limit the group's visibility on the service.

While there has been an explosion in fairness metrics for classification [13, 17, 19, 23] with researchers fleshing out when each metric is appropriate [31], there has been far less coalescence of thinking for recommender systems. Part of the challenge in studying fairness in recommender systems is that they are complex. They
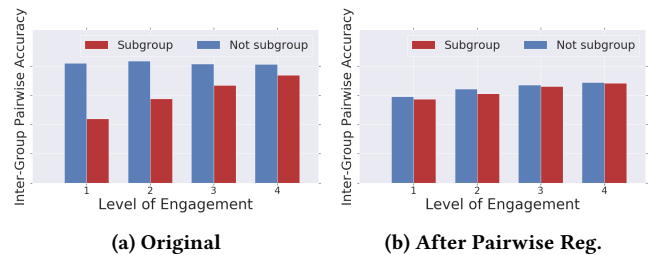
(a) Original   (b) After Pairwise Reg.

Figure 1: We find significant differences in *inter-group pairwise accuracy*, but using pairwise regularization we significantly close that gap.

often consist of multiple models [25, 47], must balance multiple goals [36, 50], and are difficult to evaluate due to extreme and skewed sparsity [9] and numerous dynamics [3, 33]. All of these issues are hardly resolved in the recommender system community, and present additional challenges in improving recommender fairness.

One challenging split in recommender systems is between treating recommendation as a pointwise prediction problem and applying those predictions for ranked list construction. Pointwise recommenders make a prediction about user interest for each item and then a ranking of recommendations is determined based on those predictions. This setup is pervasive in practice [16, 25, 34, 36, 38], but significant research goes into bridging the gap between pointwise predictions and ranking construction [1, 6]. Fairness falls into a similar dilemma. Recent research developed fairness metrics centered around pointwise accuracy [9, 49], but this does not indicate much about the resulting ranking that the user actually sees. In contrast, [11, 44, 45, 52] explore what is a fair ranking, but focus on unpersonalized rankings with complete information (where relevancy is known for all items) and in most cases require a post-processing algorithm with item group memberships, which is often not possible in practice [7].

Further, evaluation of recommender systems is notoriously difficult due to the ever changing dynamics in the system. What a user was interested in yesterday they may not be interested in tomorrow, and we only know a user's preferences if we recommend them an item. As a result, metrics are often biased (in the statistical sense) by the previous recommender system [3], and while a large body of research works to do unbiased offline evaluation [42, 43], this is very difficult due to the large item space, extreme sparsity of feedback, and evolving users and items. These issues only become more salient when trying to measure recommender system fairness, and even more so when trying to evaluate complete rankings.

We address all of these challenges through a *pairwise recommendation fairness metric*. Using easy-to-run, randomized experiments we are able to get unbiased estimates of user preferences. Based on these observed pairwise preferences, we are able to measure the fairness of even a pointwise recommender system, and we show that these metrics directly correspond to ranking performance. Further, we offer a novel regularization term that we show improves the ultimate ranking fairness of a pointwise recommender, as seen in Figure 1. We test this on a large-scale recommender system in production and show the practical benefits and trade-offs both theoretically and empirically. In summary, we make the following contributions:

- **Pairwise Fairness:** We propose a set of novel metrics for measuring the fairness of a recommender system based on pairwise comparisons. We show that this pairwise fairness metric directly corresponds to ranking performance and analyze its relation with pointwise fairness metrics.
- **Pairwise Regularization:** We offer a regularization approach to improve model performance for the given fairness metric that even works with pointwise models.
- **Real-world Experiments:** We apply our approach in a large-scale production recommender system and demonstrate that it produces significant improvements in pairwise fairness.

2 RELATED WORK

This research lies at the intersection of and builds on a myriad of research from the recommender systems community and the machine learning fairness community.

Recommender Systems. Historically, much of the recommender systems research built on collaborative filtering approaches, with a focus on ratings prediction spurred by the Netflix Prize [34]; another line of work has fleshed out pairwise models of user preferences [14, 28], particularly for information retrieval in search engines. As a core component in many industrial applications of machine learning, a significant amount of work has been published on production recommenders, such as ad click prediction for ranking ads [25, 38]. These systems often follow cascading patterns with sequences of models being applied [25, 47]. More recently, there has been a strong growth in using neural networks to improve recommender accuracy [16, 24, 36].

Building real-world recommenders face a variety of challenges. Two that relate to the challenges in fairness are the temporal dynamics [10, 26, 33, 48] and biased training data [3, 15, 29]. These issues do not just make training difficult but also evaluation of recommender performance [42].

Machine Learning Fairness. The machine learning fairness community has primarily focused on fairness in classification, with a myriad of definitions being proposed [13, 17, 19, 23]. Group fairness based definitions, where a model's treatment of two groups of examples is compared, has become the most prevalent structure, but even there researchers have shown tension between different definitions [31, 40]. We primarily follow the equality of opportunity intuition from Hardt et al. [23], where we are concerned with differences in accuracy across groups. Our metric most closely builds on the AUC-based fairness metrics for classification and regression

proposed by Dixon et al. [18] and expanded in [12] to be framed as different Mann-Whitney U-tests.

Recommender System Fairness. There has been a small amount of work on fairness in ranking and recommendation, but with each piece of working taking significantly different perspectives. Zehlike et al. [52] laid out goals for fair ranking, but do not touch upon recommender systems, where data is far sparser. Similarly, Singh and Joachims [44] take a full-ranking view of fairness but are able to apply this to recommender systems through a post-processing algorithm for model predictions; follow-up work [45] moves this into model training. All of this work [11, 44, 45, 52] focuses on an unpersonalized information retrieval setting where relevance labels are known for each item; we focus on personalized recommendations where data sparsity and biases must be handled. In contrast, [9, 49] focus on collaborative filtering pointwise accuracy differences across groups but do not connect these metrics to resulting rankings.

More distant is research on statistical parity in recommenders, which argues that in some applications items should be shown at the same rate across groups [55]. Diversity [32, 39, 46, 52], filter bubbles [4], and feedback loops [27], while related to machine learning fairness, are not the focus of this paper.

Fairness Optimization. Many approaches have been proposed to address fairness issues. Post-processing can provide elegant solutions [23, 44], but often requires knowing group memberships for all examples, which is rarely known for demographic data. Rather, numerous approaches have been developed for optimizing fairness metrics during classifier training, such as constraint-based optimization [2, 22], adversarial learning [8, 20, 35, 37, 53, 54], and regularization over model predictions [5, 7, 30, 51]. We build on these regularization approaches for improving the fairness properties of our recommender system.

3 PAIRWISE FAIRNESS FOR RECOMMENDATION

We begin now with a description of our recommender system, the fairness concerns, and our metrics for them.

3.1 Recommendation Environment

We consider a production recommender system that is recommending a personalized list of K items to users. We consider a cascading recommender [16, 25, 47], with a set of retrieval systems [15] followed by a ranking system [16, 36]. We assume that the retrieval systems return a set $\mathcal{R}$ of M' relevant items from the total corpus $\mathcal{J}$ of M items, where $M \gg M' \geq K$. The ranking model then scores and ranks M' items in $\mathcal{R}$ to get a final ranking of K items. Herein, we focus primarily on the role of the ranker.

Whenever a recommendation is made, the system observes user features $\mathbf{u}_i$ for user i and a set of context features $\mathbf{c}$, such as timing or device information; together we will refer to this as the query $\mathbf{q} = (\mathbf{u}, \mathbf{c})$. In addition, for each item $j \in \mathcal{J}$ we observe feature vector $\mathbf{v}_j$; this can include sparse representation or learned embeddings for the item as well as any other properties tied to the item. The ranker performs its ranking based on estimates of user feedback, which could be clicks, ratings [34], dwell-time on articles [50], later purchase of items, etc. For our system, we will estimate if the user

Symbol	Definition
$\mathbf{q}$	Query consisting of user and context features
$\mathcal{R}_{\mathbf{q}}$	The set of relevant items for $\mathbf{q}$
$y, \hat{y}$	User click feedback and prediction of it
$z, \hat{z}$	Post-click engagement and prediction of it
$f_\theta(\mathbf{q}, \mathbf{v})$	Predictions $(\hat{y}, \hat{z})$ for $\mathbf{q}$ on item $\mathbf{v}$
$g(\hat{y}, \hat{z})$	Monotonic ranking function from predictions
$c_{\mathbf{q}}(j, j')$	Comparison between items j, j' based on g and f
$\ell_{\mathbf{q}}(j)$	Position of j in ranked list of $\mathcal{R}_{\mathbf{q}}$
s_j	Binary sensitive attribute for item j
$\mathcal{D}$	Total dataset of tuples $\{(\mathbf{q}, \mathbf{v}, y, z)\}$
$\mathcal{P}$	Dataset of comparisons $\{((\mathbf{q}, \mathbf{v}, y, z), (\mathbf{q}, \mathbf{v}', y', z'))\}$

Table 1: Notation used throughout the paper.

clicked on the item $y \in \{0, 1\}$ as well as user engagement after a click on the item $z \in \mathbb{R}$, such as dwell-time, purchases, or ratings. As such, our dataset consists of historical examples $\mathcal{D} = \{(\mathbf{q}, \mathbf{v}, y, z)\}$. (Note, because z is user engagement after a click, if no click occurs, $z = 0$.) $\mathcal{D}$ only contains examples that have been recommended to the user previously.

The ranker is a model f_θ parameterized by θ; the model is trained to predict the user engagement $f_\theta(\mathbf{q}, \mathbf{v}) \rightarrow (\hat{y}, \hat{z}) \approx (y, z)$. Finally, a ranking of the items is produced by a monotonic scoring function $g(\hat{y}, \hat{z})$ and the user is shown the top K items from the relevant items $\mathcal{R}$ ordered by g.

3.2 Motivating Fairness Concerns

As discussed previously, a wide variety of fairness concerns have been highlighted in the literature. In this work, we primarily focus on the risk to groups of items from being under-recommended [9, 21, 39, 45]. For example, if a social network under-ranked posts by a given demographic group, that could limit the group's visibility and thus engagement on the service. If a comment section of a website was personalized, and if a demographic group of users' comments were under-ranked, then that demographic would have less of a voice on the website. In a more abstract sense, we assume that each item j has sensitive attribute $s_j \in \{0, 1\}$. We will work to measure if items from one group are systematically under-ranked.

Although not our primary focus, these issues could align with user group concerns if a group of items is preferred by a particular user group. This framework could be explicitly extended to incorporate user groups as well. If each user has a sensitive attribute, we can compute *all* of the following metrics over each user group and compare performance across the groups. For example, if we are concerned that a social network is under-ranking items about a particular topic to a particular demographic, we could compare the degree of under-ranking of that topic's content across demographic groups.

3.3 Pairwise Fairness Metric

While the above fairness goals may seem worthwhile, we must make precise what it means for an item to be "under-ranked." Here, we draw on the intuition of Hardt et al. [23] for equality of odds, where the fairness of a classifier is quantified by comparing either its false positive rate and/or false negative rate. Stated differently, given an item's label is positive, what is the probability the classifier will predict it to be positive. This works well in classification because the model's prediction can be compared to a predefined threshold.

In recommender systems, it is less clear what a positive prediction is, even if we restrict our analysis to clicks (y) and ignore engagement (z). For example, if an item is clicked, $y = 1$, and the predicted probability of click is $\hat{y} = 0.6$, is this a positive prediction? It can be perceived as an under-prediction of 0.4, but it may still be the top-ranked item if all other items have a predicted $\hat{y} < 0.6$. As such, understanding errors in the pointwise predictions requires comparing the predictions of items for the same query.

We begin with defining a pairwise accuracy: what is the probability that a clicked item is ranked above another relevant unclicked item, for the same query:

PairwiseAccuracy =

$$P(g(f_\theta(\mathbf{q}, \mathbf{v}_j)) > g(f_\theta(\mathbf{q}, \mathbf{v}_{j'})) | y_{\mathbf{q}, j} > y_{\mathbf{q}, j'}, j, j' \in \mathcal{R}_{\mathbf{q}}) \quad (1)$$

With this definition, we have a sense of how often the ranking system ranks the clicked item well. For succinctness, we will use $c_{\mathbf{q}}(j, j') \triangleq \mathbb{1}[g(f_\theta(\mathbf{q}, \mathbf{v}_j)) > g(f_\theta(\mathbf{q}, \mathbf{v}_{j'}))]$ to represent the comparison between the predictions for item j and j' on query $\mathbf{q}$; we will hide the term $j, j' \in \mathcal{R}_{\mathbf{q}}$, but we only consider comparisons among relevant items for all following definitions.

As with much of the rest of fairness research, we are concerned with the relative performance across groups, not the absolute performance. As such, we can compare:

$$P(c_{\mathbf{q}}(j, j') | y_{\mathbf{q}, j} > y_{\mathbf{q}, j'}, s_j = 0) = P(c_{\mathbf{q}}(j, j') | y_{\mathbf{q}, j} > y_{\mathbf{q}, j'}, s_j = 1)$$

That is, is the PairwiseAccuracy for items from one group $S = 0$ higher or lower than the PairwiseAccuracy for items from the other group $S = 1$ [1].

While this is an intuitive metric, it is problematic in that it ignores entirely user engagement z and thus possibly runs the risk of promoting clickbait that users don't ultimately value. As such, we can follow the approach of conditioning on other dependent signals as in [7, 41].

Definition 1 (Pairwise Fairness). A model f_θ with ranking formula g is considered to obey pairwise fairness if the likelihood of a clicked item being ranked above another relevant unclicked item is the same across both groups, conditioned on the items have been engaged with the same amount:

$$P(c_{\mathbf{q}}(j, j') | y_{\mathbf{q}, j} > y_{\mathbf{q}, j'}, s_j = 0, z_{\mathbf{q}, j} = \tilde{z})$$
$$= P(c_{\mathbf{q}}(j, j') | y_{\mathbf{q}, j} > y_{\mathbf{q}, j'}, s_j = 1, z_{\mathbf{q}, j} = \tilde{z}), \quad \forall \tilde{z}. \quad (2)$$

This definition gives us an aggregate notion of ranker accuracy for items from each group.

While this is valuable, it does not distinguish between types of mis-orderings. This can be problematic in systematically under-exposing items from one group [44]. For illustration, consider the following two examples where in both cases there are three items from each group $\{A_j\}_{j=1}^3 \cup \{B_j\}_{j=1}^3$ and in the first case assume that A_1 is clicked and in the second case assume that B_1 is clicked. If in the first case the system gives a ranking $[A_2, A_3, B_1, A_1, B_2, B_3]$ and in the second case the systems gives $[A_1, A_2, A_3, B_1, B_2, B_3]$, we see

---

[1] We focus on s_j rather than $s_{j'}$ as item j is the clicked item and we are concerned with ranking the clicked item well. We will incorporate $s_{j'}$ in later metrics.

that the overall pairwise accuracy is the same in both cases, $\frac{2}{5}$, but in the second case, even when an item from group B was of interest (clicked), all group B items ranked below group A items. Both are problematic in ranking the clicked item low, but the second is more problematic in systematically preferring one group to the other, independent of user preferences.

To deal with this, we can split the above pairwise fairness definition into two separate criteria: pairwise accuracy between items in the same group and pairwise accuracy between items from different groups; we will refer to these metrics as *intra-group pairwise accuracy* and *inter-group pairwise accuracy*, respectively:

$$\text{Intra-Group Acc.} \triangleq P(c_\mathbf{q}(j,j')|y_{\mathbf{q},j} > y_{\mathbf{q},j'}, s_j = s_{j'}, z_{\mathbf{q},j} = \tilde{z})$$

$$\text{Inter-Group Acc.} \triangleq P(c_\mathbf{q}(j,j')|y_{\mathbf{q},j} > y_{\mathbf{q},j'}, s_j \neq s_{j'}, z_{\mathbf{q},j} = \tilde{z})$$

From these we can define *Intra-Group Pairswise Fairness* and *Inter-Group Pairwise Fairness* criteria.

Definition 2 (Intra-Group Pairwise Fairness). A model f_θ with ranking formula g is considered to obey intra-group pairwise fairness if the likelihood of a clicked item being ranked above another relevant unclicked item from the same group is the same independent of group, conditioned on the items have been engaged with the same amount:

$$P(c_\mathbf{q}(j,j')|y_{\mathbf{q},j} > y_{\mathbf{q},j'}, s_j = s_{j'} = 0, z_{\mathbf{q},j} = \tilde{z})$$
$$= P(c_\mathbf{q}(j,j')|y_{\mathbf{q},j} > y_{\mathbf{q},j'}, s_j = s_{j'} = 1, z_{\mathbf{q},j} = \tilde{z}), \forall \tilde{z} \quad (3)$$

Definition 3 (Inter-Group Pairwise Fairness). A model f_θ with ranking formula g is considered to obey inter-group pairwise fairness if the likelihood of a clicked item being ranked above another relevant unclicked item from the opposite group is the same independent of group, conditioned on the items have been engaged with the same amount:

$$P(c_\mathbf{q}(j,j')|y_{\mathbf{q},j} > y_{\mathbf{q},j'}, s_j = 0, s_{j'} = 1, z_{\mathbf{q},j} = \tilde{z})$$
$$= P(c_\mathbf{q}(j,j')|y_{\mathbf{q},j} > y_{\mathbf{q},j'}, s_j = 1, s_{j'} = 0, z_{\mathbf{q},j} = \tilde{z}), \forall \tilde{z} \quad (4)$$

The Intra-Group Pairwise Fairness to some degree acts similarly to the overall Pairwise Fairness notion as it indicates the ability of the recommender system to rank well the item of interest to the user. The Inter-Group Pairwise Fairness gives us further insight into whether mistakes in ranking are at the cost of the group as a whole.

We can see this more clearly by decomposing the overall pairwise accuracy as follows:

$$P(c_\mathbf{q}(j,j')|y_{\mathbf{q},j} > y_{\mathbf{q},j'}, s_j = \tilde{s}, z_{\mathbf{q},j} = \tilde{z}, j, j' \in \mathcal{R}) \quad (5)$$
$$= \sum_{\tilde{s}' = \{\tilde{s}, 1-\tilde{s}\}} P(s_{j'} = \tilde{s}'|y_{\mathbf{q},j} > y_{\mathbf{q},j'}, s_j = \tilde{s}, z_{\mathbf{q},j} = \tilde{z}, j, j' \in \mathcal{R})$$
$$\times P(c_\mathbf{q}(j,j')|y_{\mathbf{q},j} > y_{\mathbf{q},j'}, s_j = \tilde{s}, s_{j'} = \tilde{s}' z_{\mathbf{q},j} = \tilde{z}, j, j' \in \mathcal{R})$$

That is, we find we can break up the pairwise comparisons into two sets, intra-group and inter-group comparisons, and that the overall pairwise accuracy is a weighted sum of the inter-group accuracy and intra-group accuracy, where the weights are determined by the probability of seeing a pair of that form (inter-group or intra-group) with the corresponding click and engagement. Together, these metrics give us a better sense of the fairness of the recommender system.

3.4 Measurement

While the above definitions offer a goal of how we would like a recommender system to perform, measuring the degree to which a recommender system meets these goals presents unique challenges. As discussed in the introduction, users and items in recommender systems are highly dynamic, and we typically only observe user feedback on previously recommended items, which makes metrics vulnerable to bias in the previous recommender system.

However, for all three fairness definitions given above, we would like to have unbiased estimates of user preferences between pairs of items. In order to do this we run *randomized experiments* over a small percentage of queries to the recommender system. The experimental description below is all assumed to operate over the subset of queries in the experimental slice.

For the experimental queries we will show the user a pair of items in positions two and three of the recommended slate; this prevents any position bias [3] where items that the recommender system ranks low are less likely to be clicked than items ranked high, irrespective of the particular items. Because the definitions above are all over arbitrary pairs of items from the set of relevant items for the given query, for each query two items are chosen at random from $\mathcal{R}_\mathbf{q}$ and their ordering in positions two and three is also randomized.

Among all queries in the experimental slice only a small fraction will have clicks on one of the items in the randomized item pair. Whenever an item in the randomized item pair is clicked, we record the query, pair, which item was clicked, and the subsequent engagement z. With this, we can compute all of the probabilities in the fairness definitions above. In practice we discretize z into buckets for easier comparison.

Note, as can be seen through this experiment, we cannot know the engagement we would have observed if the unclicked item had been clicked. This motivates our current metric design of conditioning on z rather than estimating the accuracy of $\hat{z}$, since we can only know $\hat{z}$ for one item in the pair.

Discussion. These metrics connect the performance of the ranking model to the end fairness properties of the resulting ranking. An underlying assumption is that the retrieval system that determines the set of relevant items, $\mathcal{R}_\mathbf{q}$, is "fair." We believe further research is needed to understand what does it mean for a retrieval system to be "fair" and how any bias in the retrieval system propagates through the ranker to effect the end ranking experience.

4 THEORETICAL ANALYSIS

While hopefully the above definitions are well-motivated, we find that upon further inspection they obey a number of fascinating properties.

4.1 Ranking Interpretation

While the metrics have thus far primarily been described similar to pairwise accuracy, they can be interpreted through the lens of ranking. That is, the recommender system sorts $\mathcal{R}_\mathbf{q}$ according to g and f_θ. We will use $\ell(j)_\mathbf{q}$ to denote the position of item $j \in \mathcal{R}_\mathbf{q}$ in

the sorted list of items:

$$\ell(j)_{\mathbf{q}} = |\mathcal{R}_{\mathbf{q}}| - \sum_{j' \in \mathcal{R}_{\mathbf{q}} \setminus j} \mathbb{1}[g(f_{\theta}(\mathbf{q}, \mathbf{v}_j)) > g(f_{\theta}(\mathbf{q}, \mathbf{v}_{j'}))] \qquad (6)$$

From this perspective, we find that we can connect pairwise fairness to fairness with respect to the ranked position:

Theorem 1. *If a recommender system achieves pairwise fairness then the expected position of a clicked item with z engagement is the same across groups.*

PROOF. This falls out of the definition of pairwise accuracy and pairwise fairness:

$$\mathbb{E}_{\mathbf{q}|y_{\mathbf{q},j}=1, s_j=0, z_{\mathbf{q},j}=\tilde{z}}[\ell(j)_{\mathbf{q}}]$$
$$= \mathbb{E}_{\mathbf{q}}[|\mathcal{R}_{\mathbf{q}}|(1 - \text{PairwiseAccuracy}_{\mathbf{q}}(j))]$$
$$= \mathbb{E}_{\mathbf{q}}[|\mathcal{R}_{\mathbf{q}}|(1 - P(c_{\mathbf{q}}(j, j')|y_{\mathbf{q},j} > y_{\mathbf{q},j'}, s_j = 0, z_{\mathbf{q},j} = \tilde{z}))]$$
$$= \mathbb{E}_{\mathbf{q}}[|\mathcal{R}_{\mathbf{q}}|(1 - P(c_{\mathbf{q}}(j, j')|y_{\mathbf{q},j} > y_{\mathbf{q},j'}, s_j = 1, z_{\mathbf{q},j} = \tilde{z}))]$$
$$= \mathbb{E}_{\mathbf{q}|y_{\mathbf{q},j}=1, s_j=1, z_{\mathbf{q},j}=\tilde{z}}[\ell(j)_{\mathbf{q}}]$$
□

As such, we see that we can interpret pairwise recommender fairness as equivalent to the notion that the position of a clicked and engaged with item should not depend on the group membership on average, aligning with position bias highlighted in [11, 45]. (This analysis is similar to probabilistic interpretations in traditional pairwise IR [14], but now in the context of recommender system fairness.)

The inter-group and intra-group pairwise accuracies also connect to the rank position of the clicked item. That is, we can decompose:

$$\ell(j)_{\mathbf{q}} = |\mathcal{R}_{\mathbf{q}}| - \sum_{j' \in \mathcal{R}_{\mathbf{q}} \setminus j} \mathbb{1}[g(f_{\theta}(\mathbf{q}, \mathbf{v}_j)) > g(f_{\theta}(\mathbf{q}, \mathbf{v}_{j'}))]$$

$$= |\mathcal{R}_{\mathbf{q}}| - \left(\sum_{j' \in \mathcal{R}_{\mathbf{q}} \setminus j | s_{j'} = s_j} c_{\mathbf{q}}(j, j') + \sum_{j' \in \mathcal{R}_{\mathbf{q}} \setminus j | s_{j'} \neq s_j} c_{\mathbf{q}}(j, j') \right)$$

$$= |\mathcal{R}_{\mathbf{q}}|(1 - P(s_{j'} = s_j | j' \in \mathcal{R}_{\mathbf{q}}) P(c_{\mathbf{q}}(j, j')|s_{j'} = s_j, j' \in \mathcal{R}_{\mathbf{q}})$$
$$\quad - P(s_{j'} \neq s_j | j' \in \mathcal{R}_{\mathbf{q}}) P(c_{\mathbf{q}}(j, j')|s_{j'} \neq s_j, j' \in \mathcal{R}_{\mathbf{q}}))$$

$$= |\mathcal{R}_{\mathbf{q}}| - |\{j' | j' \in \mathcal{R}_{\mathbf{q}}, s_j = s_{j'}\}| P(c_{\mathbf{q}}(j, j')|s_{j'} = s_j, j' \in \mathcal{R}_{\mathbf{q}})$$
$$\quad - |\{j' | j' \in \mathcal{R}_{\mathbf{q}}, s_j \neq s_{j'}\}| P(c_{\mathbf{q}}(j, j')|s_{j'} \neq s_j, j' \in \mathcal{R}_{\mathbf{q}})$$

Here too, we see that the overall ranked position can be decomposed into the position within the ranked list from the same group and position among the ranked list from the other group. However, because of the possibly varying distributions of the number of comparisons of each type, we believe it makes sense to focus on each of these terms as probabilities.

4.2 Relation to Pointwise Metrics

While our pairwise fairness metric aligns with previously stated goals of fair ranking, we find that it lies in tension with traditional pointwise metrics. For example, recommender systems are often evaluated in terms of calibration or RMSE [34], and these metrics have been espoused as important fairness metrics in classification [17] and in recommendation [9, 49]. We show here that these pointwise metrics are insufficient for guaranteeing pairwise fairness.

For the following proofs we consider a simplified case where $z = 0$ and $g(y, z) \triangleq y$. This can be thought of as ranking items by predicted click through rate (pCTR). For each group s we denote its average label y by $\bar{y}_s \triangleq \mathbb{E}_{\mathbf{q},j}[y_{\mathbf{q},j}|s_j = s]$.

Calibration. We begin with examining the relationship between calibration and pairwise fairness. A pCTR model $f(x)$ for labels y is considered calibrated if and only if:

$$E[y|f(x) = \tilde{y}] = \tilde{y} \qquad (7)$$

That is, among examples receiving a particular prediction, the average label for those examples needs to be equal to the predicted value. In the context of fairness, this would be evaluated over examples from one group.

Lemma 1. *A calibrated model is insufficient for guaranteeing pairwise ranking fairness.*

PROOF. In order to prove this, we offer an example of a calibrated model that does not obey pairwise ranking fairness. Let's assume that we learn a model that predicts $f(\mathbf{q}, \mathbf{v}_j) \triangleq \bar{y}_{s_j}$ for all examples with an item from group s. This model is by definition calibrated per group.

If we have two groups, $\tilde{s}$ and $\tilde{s}'$, where $\bar{y}_{\tilde{s}} > \bar{y}_{\tilde{s}'}$ then $P(c_{\mathbf{q}}(j, j')) = 1$ for all items j with $s_j = \tilde{s}$ and all items j' with $s_{j'} = \tilde{s}'$. As such, $P(c_{\mathbf{q}}(j, j')|y_j > y_{j'}, s_j = \tilde{s}, s_{j'} = \tilde{s}') = 1$ and $P(c_{\mathbf{q}}(j, j')|y_{j'} > y_j, s_j = \tilde{s}, s_{j'} = \tilde{s}') = 0$. From this it is clear that inter-group fairness does not hold. We assume ties are split randomly, giving us $P(c_{\mathbf{q}}(j, j')|y_j > y_{j'}, s_j = s_{j'}) = 0.5$. Based on Eq. (5), we find that as long as $P(s_j \neq s_{j'}|j, j' \in \mathcal{R}_{\mathbf{q}}, y_j > y_{j'}) > 0$ then overall pairwise fairness does not hold. □

This problem is highly similar to the issue pointed out by [44] with respect to ranking and exposure, and we see here holds true even among pairwise comparisons.

Squared Error. Another pervasive metric in recommender systems is mean squared error (MSE) [34]. This metric, and modifications of it, have been proposed for evaluating the fairness of collaborative filtering systems [9, 49]. While this may be worthwhile to encourage accuracy across groups, we find that it too is insufficient for guaranteeing pairwise fairness.

Lemma 2. *Equal MSE across groups is insufficient for guaranteeing pairwise ranking fairness.*

PROOF. As above we will demonstrate an example of a model that achieves equal MSE across groups but does not obey pairwise fairness. Again, let us assume that we learn a model that predicts $f(\mathbf{q}, \mathbf{v}_j) \triangleq \bar{y}_{s_j}$. We assume we have two groups, $\tilde{s}$ and $\tilde{s}'$, where $\bar{y}_{\tilde{s}} = 1 - \bar{y}_{\tilde{s}'}$ and $\bar{y}_{\tilde{s}} > \bar{y}_{\tilde{s}'}$. We see that:

$$\text{MSE}_s = \bar{y}_s(1 - \bar{y}_s)^2 + (1 - \bar{y}_s)(0 - \bar{y}_s)^2 = \bar{y}_s - \bar{y}_s^2$$

Through simply substituting in $\bar{y}_{\tilde{s}} = 1 - \bar{y}_{\tilde{s}'}$ in the definition above we can see that $\text{MSE}_{\tilde{s}} = \text{MSE}_{\tilde{s}'}$.

Just as in the proof above, because $\bar{y}_{\tilde{s}} > \bar{y}_{\tilde{s}'}$, we find $P(c_{\mathbf{q}}(j, j')) = 1$ for all items j with $s_j = \tilde{s}$ and all items j' with $s_{j'} = \tilde{s}'$. As such, $P(c_{\mathbf{q}}(j, j')|y_j > y_{j'}, s_j = \tilde{s}, s_{j'} = \tilde{s}') = 1$ and $P(c_{\mathbf{q}}(j, j')|y_{j'} > y_j, s_j = \tilde{s}, s_{j'} = \tilde{s}') = 0$. Again, it is clear that inter-group fairness does not hold. If we split ties randomly, such that $P(c_{\mathbf{q}}(j, j')|y_j > y_{j'}, s_j =$

$s_{j'}) = 0.5$, and as long as $P(s_j \neq s_{j'}|j, j' \in \mathcal{R}_\mathbf{q}, y_j > y_{j'}) > 0$, then we again find that overall pairwise fairness does not hold either. □

As a result, while matching MSE across groups is intuitively valuable for fairness, it is insufficient for making any claims about the end ranking. Looking at the example in the proof, it is clear that this is in part due to the fact that MSE does not distinguish between over- and under-prediction. However, even taking that into account, MSE ignores the relative ranking and thus it is hard to determine what an improvement of ϵ in MSE means for ranking accuracy.

5 PAIRWISE REGULARIZATION TO IMPROVE FAIRNESS

With an understanding of our fairness goals, we now ask: how can we learn a recommender system that achieves these fairness properties? As discussed previously, most production recommenders are pointwise recommenders trained to predict y and z, so we would like a modeling approach that does not require throwing out existing techniques.

To encourage fairness during training, we build on the regularization approach first proposed by Zafar et al. [51] and expanded upon by Beutel et al. [7]. In particular, Beutel et al. [7] optimized for equality of opportunity in classification [23] by minimizing the correlation between the group membership and the model's predictions among examples with the same label. In our setting, we are concerned with the relative ordering, so we must modify this objective.

We assume that our model is trained with a loss $L(f_\theta(\mathbf{q}, \mathbf{v}), (y, z))$; for example, if squared error were used then $L(f_\theta(\mathbf{q}, \mathbf{v}), (y, z)) \triangleq (\hat{y} - y)^2 + (\hat{z} - z)^2$. Further, we assume that we know $g(\hat{y}, \hat{z})$ and that it is differentiable. Given this, we train our model f_θ with the following objective:

$$\min_\theta \left(\sum_{(\mathbf{q}, j, y, z) \in \mathcal{D}} L(f_\theta(\mathbf{q}, \mathbf{v}_j), (y, z)) \right) + |\text{Corr}_\mathcal{P}(A, B)| \qquad (8)$$

Here, $\mathcal{D}$ is our original training data and $\mathcal{P}$ is the experimental data from Section 3.4 consisting of pairs of tuples $((\mathbf{q}, j, y, z), (\mathbf{q}, j', y', z'))$. The second term, the absolute correlation is computed as the correlation between two terms, A and B, both random variables over pairs from $\mathcal{P}$:

$$A = (g(f_\theta(\mathbf{q}, \mathbf{v}_j)) - g(f_\theta(\mathbf{q}, \mathbf{v}_{j'})))(y - y') \qquad (9)$$
$$B = (s_j - s_{j'})(y - y') \qquad (10)$$

That is, the pairwise regularizer calculates the correlation between the residual between the clicked and unclicked item and the group membership of the clicked item. As a result, the model is penalized if the its ability to predict which item was clicked is better for one group than the other.

To make sure there is sufficient data for a meaningful calculation, we rebalance $\mathcal{P}$ to have approximately half of the data with the clicked item belonging to group $s = 0$ and the other half with the clicked item belonging to group $s = 1$. Further data restrictions can be applied for alternative goals. If we are concerned with intra-group pairwise fairness, we can restrict $\mathcal{P}$ to the set of pairs where $s_j \neq s_{j'}$, and if we are concerned with a large differences in

engagement z, we can create buckets of $\mathcal{P}_z$ where all pairs in the set resulted in engagement z. This approach is general enough to be used with pointwise recommenders as well as pairwise recommenders.

As in [7], this approach does not provably achieve pairwise fairness but we follow it due to its strong empirical performance and ease of use, crucial for production applications.

6 EXPERIMENTS

To understand our pairwise fairness metric and our proposed modeling improvements, we study the performance of a large-scale, production recommender system. We offer analysis of the state-of-the-art production model's performance as well as how our modeling changes effect the system.

6.1 Experimental Setup

As described in Section 3.1, we study a cascading recommender system where multiple retrieval systems return the set of relevant items for a given query, followed by a ranking model. Here we evaluate the ranking model's performance. The ranking model is a multi-layer neural network trained in a pointwise-fashion with multiple heads to predict the probability of a click, y, and a set of user engagement signals after the click, which we refer to in the aggregate by z; this is a similar setup to [24, 36]. This model is continuously trained on a dataset of interactions with previous recommendations.

We study the performance of the ranker with respect to a sensitive subgroup of items, comparing the performance of this subgroup to the rest of the data, denoted by "not subgroup." The subgroup represents approximately 0.2% of all items, but it is a subgroup that we feel is important to analyze for recommendation fairness. As mentioned previously, we only know the group membership for a small percent of items; this prevents using serving-time approaches to improve the pairwise fairness metrics. Following the description in Section 3.4, we gather a dataset $\mathcal{P}$ of random pairs of relevant items shown to the user and recorded when the user clicks on one of the items. We use a random half of this dataset for the pairwise regularization and the other half for evaluating the model.

We compare two versions of the model: (1) the production model trained without any fairness objective, (2) a test model, trained with the same architecture but with the pairwise regularization to optimize for inter-group pairwise fairness. (As this is a live system with both data and training dynamics, we present a model chosen at random from a set of test models.) As we will see below, we focus on inter-group pairwise fairness as this is the area we find needing more improvement.

Due to the sensitive nature, we cannot report absolute accuracy measures. Rather, we report the relative performance between the subgroup and the rest of the data. That is, we aggregate the pairwise accuracy measures across engagement levels through a simple average; and we report the relative ratio of the average accuracy for the "not subgroup" divided by the average accuracy for the subgroup. All plots group engagement z into four levels and maintain the same y-axis scaling so that relative comparisons can be made across them.
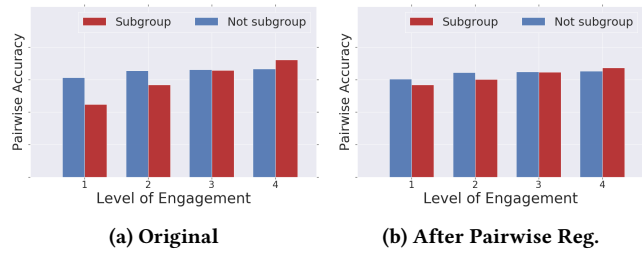
Figure 2: We find some gaps in *overall pairwise accuracy* that are improved through the pairwise regularization.

Figure 3: We observe slight differences in *intra-group pairwise accuracy.*

6.2 Baseline Performance

We begin with an analysis of the production system's performance. As discussed in Section 3.3, we analyze the system's performance in terms of: (1) pairwise fairness, (2) intra-group pairwise fairness, and (3) inter-group pairwise fairness.

As the overall pairwise fairness in Figure 2a shows, the production system under-ranks items from the subgroup when the subsequent level of engagement is low, but interestingly slightly over-ranks items from the subgroup when the subsequent level of engagement is high. In total, we find that the non-subgroup items have an 8.3% advantage overall. (That is, the pairwise accuracy of "not subgroup" divided by the pairwise accuracy of "subgroup" is 1.083.)

Second, we examine the performance within each group – the intra-group pairwise accuracy. As can be seen in Figure 3a, across all levels of engagement the model has more difficulty selecting the clicked item when comparing subgroup items than when comparing non-subgroup items. In total, this puts the non-subgroup items at a 14.9% advantage in intra-group pairwise fairness. We have found that this is in part due to the subgroup being small while there is far more diversity among the non-subgroup items, making comparisons easier. When further filtering the subgroup comparisons to remove highly-similar item comparisons, we find no meaningful difference in performance between the subgroup and the non-subgroup.

While both of the above results suggest some deficiencies, we find that the story is significantly more dramatic when looking at the inter-group pairwise accuracy. As seen in Figure 1a, across all levels of engagement we find that the subgroup items are significantly under-ranked relative to the non-subgroup items. Overall, we find that the non-subgroup items have a 35.6% advantage. Further, we see that the pairwise accuracy for non-subgroup items in inter-group pairs is notably higher than in intra-group pairs, suggesting that the model is taking advantage of this difference in items. This suggests that subgroup items, even when of interest to the user, are ranked under non-subgroup items. Because of the implication on subgroup experience and the more dramatic nature of the results, we focus herein on improving the inter-group pairwise fairness.

6.3 Fairness Improvements

As described above, we apply the pairwise regularization from Section 5 over inter-group pairs of examples so as to optimize for inter-group pairwise fairness.
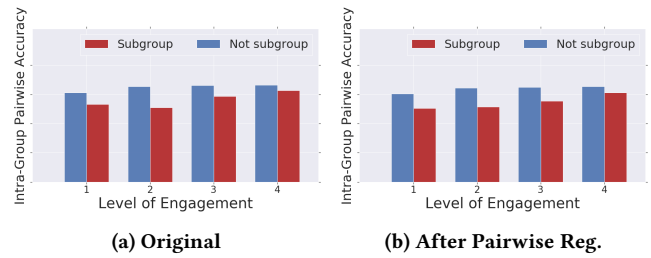
We see in Figure 1 the effect of pairwise regularization on the inter-group pairwise fairness, the metric that it is most aligned with. While the regularization decreases the pairwise accuracy of the non-subgroup items, it effectively closes the gap in the inter-group pairwise fairness metric, resulting in only a 2.6% advantage for non-subgroup items in inter-group pairwise fairness, down from 35.6%. Further, while the decrease in pairwise accuracy for the non-subgroup may appear discouraging, the pairwise accuracy for the non-subgroup items in the test model is approximately on-par with the pairwise accuracy metrics we see in intra-group comparisons for non-subgroup items, suggesting the model is no longer taking advantage of the difference in items.

While not our immediate goal, we also examine how improving the inter-group fairness effects the overall pairwise fairness. As we see in Figure 2, there is a notable improvement in the pairwise accuracy for subgroup items and the gap between the groups is largely closed. Quantitatively, we observe that the relative benefit to the non-subgroup items decreases to 2.5%, down from 8.3%. Intra-group accuracy is not optimized by our pairwise regularization configuration, and as expected we see little change in the intra-group accuracy (Figure 3, with a 16.7% advantage for non-group items).

Interestingly, in most of our live experiments using models trained with pairwise regularization we found overall engagement metrics were neutral relative to the production system. Given the subgroup is a tiny fraction of the overall system, it is reassuring to see that the above fairness benefits do not come at a cost to overall performance.

Together, this shows that pairwise regularization is effective in improving the fairness properties of the ranker.

6.4 How are improvements achieved?

While the results are compelling, we do further analysis to understand how the regularization is able to close fairness gaps. To do this, we examine the exposure of items from each group compared to the user preferences, similar in principle to a coarse pairwise calibration analysis.

To understand the user preferences, we measure the percentage of inter-group pairs for which users prefer (click on) the subgroup item versus the non-subgroup item. This presents a base rate click-through-rate (CTR) for each group, similar to the analysis in Section 4.2. As we see in Figure 4a, across nearly all levels of engagement, the subgroup items are less likely to be clicked when juxtaposed
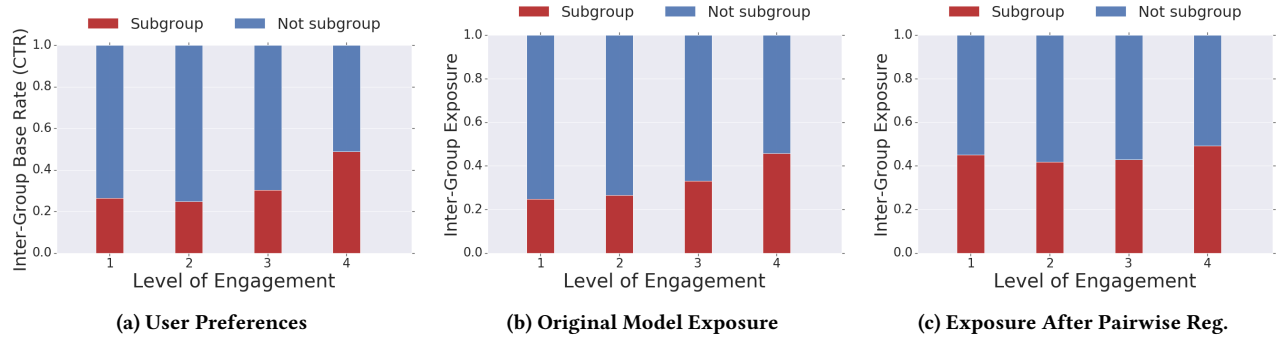
(a) User Preferences
(b) Original Model Exposure
(c) Exposure After Pairwise Reg.

Figure 4: We find that the original model's exposure closely matched the observed user preferences in the data. In correcting for pairwise fairness we observe that we comparably show more items from the subgroup.

with a non-subgroup item; interestingly, high-engagement interactions show a nearly even balance of likelihood of a click across the groups.

To understand how the model performs compared to this base CTR, we measure *exposure*: the probability of the model ranking one group's item above that of the other group, irrespective of the user preference[2]. To be precise:

$$\text{Exposure}_{\tilde{s}, \tilde{z}} = P(c_{\mathbf{q}}(j, j') | s_j = \tilde{s}, s_{j'} = 1 - \tilde{s}, z_{\mathbf{q}, j'} = \tilde{z})$$

As we see in Figure 4b, the production model exposes each group at approximately the same rate as the group's base CTR.

As we can see in Figure 4c, the exposure of items from each group changes when the model is trained with the pairwise regularization. Even with lower levels of engagement, items from the subgroup are ranked higher at a higher rate than the base CTR. This suggests that the regularizer has the effect of showing subgroup items at a higher rate than is "natural" to make sure users interested in subgroup items are recommended them. This aligns with Lemma 1 suggesting a general tension between calibration and pairwise fairness. We believe further research on this relationship and more generally how to improve model accuracy can help alleviate this tension, but for the time-being find this to be a reasonable trade-off.

7 CONCLUSION

In this work we have provided a tractable way to get unbiased measurements of recommender system ranking fairness. We are able to do this through pairwise experiments to observe user preferences. Based on this experimental data, we can evaluate and decompose recommender system fairness to see if a model systematically mis-ranks or under-ranks items from a particular group. We show that this measure aligns with ranking fairness definitions but is not covered by pointwise fairness measures. We ultimately offer a novel pairwise regularization approach to improve recommender system fairness during training, and show that it significantly improves fairness metrics in a large-scale production system.

Acknowledgements: We would like to thank Ben Packer, Xuezhi Wang, and Andrew Cotter for their helpful comments during the paper's preparation.

---

[2]This is a slight modification of exposure as defined by Singh and Joachims [44], using a probabilistic form over inter-group comparisons.

REFERENCES

[1] Ryan Prescott Adams and Richard S Zemel. 2011. Ranking via Sinkhorn Propagation. *arXiv preprint arXiv:1106.1925* (2011).
[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453* (2018).
[3] Aman Agarwal, Ivan Zaitsev, Xuanhui Wang, Cheng Li, Marc Najork, and Thorsten Joachims. 2018. Estimating Position Bias without Intrusive Interventions. *arXiv preprint arXiv:1812.05161* (2018).
[4] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.
[5] Y Bechavod and K Ligett. 2017. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044* (2017).
[6] Irwan Bello, Sayali Kulkarni, Sagar Jain, Craig Boutilier, Ed Chi, Elad Eban, Xiyang Luo, Alan Mackey, and Ofer Meshi. 2018. Seq2Slate: Re-ranking and Slate Optimization with RNNs. *arXiv preprint arXiv:1810.02019* (2018).
[7] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements. In *AIES*.
[8] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075* (2017).
[9] Alex Beutel, Ed H Chi, Zhiyuan Cheng, Hubert Pham, and John Anderson. 2017. Beyond globally optimal: Focused learning for improved recommendations. In *WWW*. 203–212.
[10] Alex Beutel, Paul Covington, Sagar Jain, Can Xu, Jia Li, Vince Gatto, and Ed H Chi. 2018. Latent Cross: Making Use of Context in Recurrent Recommender Systems. In *WSDM*. 46–54.
[11] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *SIGIR*. 405–414.
[12] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. (2019).
[13] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.
[14] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *ICML*.
[15] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H. Chi. 2019. Top-K Off-Policy Correction for a REINFORCE Recommender System. In *WSDM*. 456–464.
[16] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *RecSys*. 191–198.
[17] Cynthia S Crowson, Elizabeth J Atkinson, and Terry M Therneau. 2016. Assessing calibration of prognostic risk scores. *Statistical methods in medical research* 25, 4 (2016), 1692–1706.
[18] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. (2018).
[19] 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*. ACM, 214–226.
[20] Harrison Edwards and Amos Storkey. 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897* (2015).

[21] Michael D Ekstrand, Mucun Tian, Mohammed R Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring author gender in book rating and recommendation. In *RecSys*. 242–250.

[22] Gabriel Goh, Andrew Cotter, Maya R. Gupta, and Michael P. Friedlander. 2016. Satisfying Real-world Goals with Dataset Constraints. In *Advances in Neural Information Processing Systems*. 2415–2423.

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

[24] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*. 173–182.

[25] Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, et al. 2014. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*. ACM, 1–9.

[26] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).

[27] Ray Jiang, Silvia Chiappa, Tor Lattimore, Andras Agyorgy, and Pushmeet Kohli. 2019. Degenerate Feedback Loops in Recommender Systems. (2019).

[28] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *KDD*. 133–142.

[29] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *WSDM*. 781–789.

[30] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 643–650.

[31] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[32] Jon Kleinberg and Manish Raghavan. 2018. Selection Problems in the Presence of Implicit Bias. *arXiv preprint arXiv:1801.03533* (2018).

[33] Yehuda Koren. 2009. Collaborative filtering with temporal dynamics. In *KDD*. 447–456.

[34] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 8 (2009), 30–37.

[35] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. 2016. The Variational Fair Autoencoder. In *ICRL*.

[36] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *KDD*. 1930–1939.

[37] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *ICML*.

[38] H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. 2013. Ad click prediction: a view from the trenches. In *KDD*.

[39] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. In *CIKM*. 2243–2251.

[40] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.

[41] Ya'acov Ritov, Yuekai Sun, and Ruofei Zhao. 2017. On conditional parity as a notion of non-discrimination in machine learning. *arXiv preprint arXiv:1706.08519* (2017).

[42] Tobias Schnabel, Adith Swaminathan, Peter I Frazier, and Thorsten Joachims. 2016. Unbiased comparative evaluation of ranking functions. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*. ACM, 109–118.

[43] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. *arXiv preprint arXiv:1602.05352* (2016).

[44] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *KDD*. 2219–2228.

[45] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for Fairness in Ranking. *arXiv preprint arXiv:1902.04056* (2019).

[46] Julia Stoyanovich, Ke Yang, and HV Jagadish. 2018. Online set selection with fairness and diversity constraints. In *EDBT*.

[47] Lidan Wang, Jimmy Lin, and Donald Metzler. 2011. A cascade ranking model for efficient ranked retrieval. In *SIGIR*. 105–114.

[48] Chao-Yuan Wu, Amr Ahmed, Alex Beutel, Alexander J Smola, and How Jing. 2017. Recurrent recommender networks. In *WSDM*. 495–503.

[49] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems*.

[50] Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. 2014. Beyond clicks: dwell time for personalization. In *RecSys*. 113–120.

[51] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2015. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259* (2015).

[52] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *CIKM*. ACM, 1569–1578.

[53] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *ICML*. 325–333.

[54] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. *CoRR* abs/1801.07593 (2018). arXiv:1801.07593

[55] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-Aware Tensor-Based Recommendation. In *CIKM*. 1153–1162.