TWIX: Balancing Fairness and Utility in Item Exposure for Recommendation Systems

Maria de Lourdes M. Silva, Iago C. Chaves, André L. C. Mendonça, Eduardo R. D. Neto, Javam C. Machado

Laboratório de Sistemas e Banco de Dados (LSBD)
Departamento de Computação / UFC – Fortaleza – CE – Brazil

{malu.maia, iago.chaves, andre.luis,
eduardo.rodrigues, javam.machado}
@lsbd.ufc.br

Abstract. *Recommendation systems personalize user experiences but often introduce the "rich-get-richer" effect, where popular items dominate visibility. This bias disadvantages new items or those with few interactions, where historical data limits the exposure of new items. To address these challenges, fairness in exposure has been studied through balanced exposure, which ensures equal visibility but may reduce recommendation utility, and quality-weighted exposure, which prioritizes high-quality items. To mitigate the trade-off between fairness and utility, we propose TWIX, a method that applies quality-weighted and balanced exposures determined by a threshold. Our approach reduces popularity bias and cold start issues while maintaining recommendation utility.*

1. Introduction

Recommendation systems have received significant attention in recent academic research due to their ability to personalize information and assist users in finding potential items of interest. They generate suggestions by analyzing users' profiles and learning about their preferences. However, despite their benefits, recommendation systems can also introduce or propagate certain biases that affect item providers. A well-known issue in those systems is the "rich-get-richer" phenomenon, also known as the Matthew Effect [Merton 1968]. In a recommendation system, it expresses a popularity bias where popular items receive disproportionate exposure. This effect can limit the visibility of lesser-known or newer items or providers, making it difficult for independent or small-scale providers to compete.

As a result, smaller providers may abandon the platform due to inadequate exposure and engagement. New providers are affected by additional challenges, such as the cold start problem, which primarily arises in collaborative filtering algorithms that learn user preferences based on historical user-item interactions. It occurs when the platform receives new users, items, or providers for which there is insufficient historical data. In such cases, the recommendation algorithm struggles to estimate the quality or relevance of the new items, as it lacks interaction data to learn from. This problem reinforces the exposure gap between large-scale and independent providers. Unfair recommendations are primarily concentrated in the long-tail of the item exposure distribution, and several studies argue that popularity bias affects not only item providers but also users [Boratto et al. 2021, Abdollahpouri et al. 2019].

Popularity bias can be particularly problematic when low-quality content receives high exposure, such as clickbait, which attracts attention but offers little value to users. In such cases, items gain popularity not due to their inherent quality but because of manipulative or attention-grabbing elements, decreasing the utility of recommendations and potentially amplifying exposure inequality.

To handle these disparities, several studies have introduced the notion of fairness in exposure in their approaches. A naive approach involves incorporating constraints that balance exposure across all providers (or items), guaranteeing a balanced exposure. While this method helps mitigate issues such as popularity bias and the cold start problem, it can also significantly reduce the overall recommendation utility, highlighting the trade-off between fairness of exposure and utility.

Another fairness of exposure notion, quality-weighted exposure, leverages utility by defining that items with higher quality should receive higher exposure, i.e., an item's exposure should be proportional to its quality. This approach aims to align fairness with utility by prioritizing high-quality content. While promising, this approach faces limitations when item quality must be inferred from user interactions rather than metadata. In collaborative filtering settings, this makes quality estimation for new items especially difficult, as they lack historical interactions, a classic cold start scenario [Natarajan et al. 2020, Sang et al. 2024].

In the context of collaborative filtering recommenders and fairness of exposure, we identify two challenges in estimating item quality: (1) ensuring sufficient exposure for items with few interactions to enable accurate quality estimation, and (2) guaranteeing that high-quality items receive appropriately higher exposure.

Example 1: To illustrate the impact of different exposure fairness strategies and the cold start problem, consider a scenario with three items (A, B, and C) associated with providers P1, P2, and P3, respectively. In the example, each item has known quality and popularity scores. We compare the exposure allocations under the following strategies: i) *His.*: Exposure proportional to historical popularity; ii) *Bal.*: Equal exposure to all items; iii) *Qua.*: Exposure proportional to item quality.

Now, suppose provider P2 introduces a new item D. As a newly added item, D lacks an interaction history. Therefore, its quality and popularity are unknown. Table 1 summarizes the characteristics and exposure allocations under each strategy. This example illustrates how new items suffer from exposure lack under *His.* and *Qua.*, highlighting the cold start challenge.

Item	Quality	Popularity	*His.*	*Bal.*	*Qua.*
A	0.9	0.8	60%	25%	50%
B	0.6	0.3	30%	25%	33.3%
C	0.3	0.1	10%	25%	16.7%
D	—	—	0%	25%	—

Table 1. Exposure allocation under different fairness strategies. Item D represents a cold start case with no prior interactions, and item C potentially represents an unreliable quality estimation case due to the lack of interactions.

In this work, we address the fairness of item exposure in recommendation systems, which can naturally extend to provider-level fairness by grouping items according to their associated providers. To mitigate the challenges of poor quality estimation caused by limited interaction data, while still prioritizing the exposure of high-quality items, we propose a hybrid approach that combines two fairness notions: balanced exposure and quality-weighted exposure. Our method, called **T**runcated Quality-**w**eighted **I**tem E**x**posure (TWIX), introduces a minimum quality threshold: items above this threshold receive exposure proportional to their quality, while items below it are treated with balanced exposure. Although TWIX is an offline method, it effectively addresses issues of poor quality estimation for items with insufficient data. Hence, an online approach extension can alleviate the cold start problem, as new items inherently suffer from a lack of interactions.

We demonstrate that TWIX effectively prioritizes high-quality items while ensuring exposure for items with limited interactions, thereby promoting the fairness of exposure. Furthermore, we empirically show that TWIX achieves fairness while maintaining a high user recommendation utility.

Example 1: ***(continued)*** We now consider the running example under the TWIX strategy. In this setting, we assume that an item's quality is determined by the interactions it receives, and that its exposure is affected by this quality. To address the cold start problem, we define a minimum quality threshold of 0.4. Items with quality below this threshold are assigned a balanced exposure proportional to the threshold, ensuring a fair baseline visibility. In contrast, items with quality above the threshold are allocated quality-weighted exposure to maintain high utility. The highlighted cells indicate the updates resulting from the application of the threshold, and the final column shows the resulting exposure for each item under the TWIX strategy.

Item	**Quality**	**Popularity**	***His.***	***Bal.***	***Qua.***	***TWIX***
A	0.9	0.8	60%	25%	50%	39%
B	0.6	0.3	30%	25%	33.3%	27%
C	0.4	0.1	10%	25%	16.7%	17%
D	0.4	0.0	0%	25%	—	17%

Table 2. Exposure allocation of our hybrid strategy that combines balanced and quality-weighted exposure.

2. Preliminaries

In this section, we provide the necessary background to support the understanding of our proposed approach. We first formalize the notation and review essential concepts from collaborative filtering. Then, we introduce two key notions of fairness in exposure. Finally, we formalize the problem statement.

Formal Framework

We identify a provider by the set of items they contribute to the platform. Without loss of generality, we focus on fairness of exposure at the item level. However, this notion can

Symbol	Meaning
$\mathcal{I}, m$	Set of items and number of items
$\mathcal{U}, n$	Set of users and number of users
$\boldsymbol{u}, u_j$	Utility vector and for the item j
$\boldsymbol{q}, q_j$	Quality vector and for the item j
$\boldsymbol{v}, v_r$	Exposition vector and for the position $r \in [\![k]\!]$
$\boldsymbol{\sigma}, \sigma_i$	Ranking matrix and ranking vector for user i
P, P_{ijr}	Probability matrix and probability of item j being recommended to user i at rank r

Table 3. Notation used to formalize the framework and problem statement, along with their meaning.

be naturally extended to the provider-level by aggregating exposure across the items associated with each provider. Let $n, m \in \mathbb{N}$ be the number of users and items, respectively. We denote $[\![n]\!] = \{1, \ldots, n\}$ as a set of index. We denote the set of items as $\mathcal{I} = [\![m]\!]$ and the set of users as $\mathcal{U} = [\![n]\!]$. The user-item interaction dataset is represented by a matrix μ where each entry $\mu_{ij} \in \mathbb{Z}^+$ denotes the non-negative interaction between user $i \in \mathcal{U}$ and item $j \in \mathcal{I}$.

A recommendation system outputs a ranking list σ_i of the top-k items for each user $i \in \mathcal{U}$, where $k \leq m$. We assume that items in lower positions, i.e., ranked higher, receive more exposure. Therefore, we assign an exposure a position-based exposure weight vector $v_r \in \mathbb{R}^+$, i.e., $v_1 \geq v_2 \geq \cdots \geq v_k \geq 0$.

Each item's position in the ranking list is determined by a score that reflects its likelihood of being recommended. To model this, we define the matrix $P \in \mathbb{R}^{n \times m \times k}$, where P_{ijr} denotes the score of item j being recommended to user i at position r in the ranking. These scores must satisfy ranking constraints to ensure that the most relevant items are placed in lower positions. Figure 1 illustrates how the matrix P is computed. Given a 2-dimensional score matrix $\mu \in \mathbb{R}^{n \times m}$, where μ_{ij} represents the base score of item j for user i, we multiply each scalar μ_{ij} by an exposure weight vector $\boldsymbol{v} \in \mathbb{R}^k$, which encodes the positional bias across ranking positions. This element-wise operation, denoted by $\circledast$, performs a scalar-vector multiplication for each entry in μ, resulting in a vector of size k per (i, j) pair. The output of this operation is the 3-dimensional matrix $P \in \mathbb{R}^{n \times m \times k}$, where each slice along the third dimension corresponds to a ranking position. Thus, P captures the position-aware recommendation scores for every user–item pair across all positions in the ranking list.

We define the quality of each item as the sum of its interaction values across all users. Formally, let $\boldsymbol{q}$ be the quality vector, where each component is given by q_j. Let $\boldsymbol{u}$ denote the utility vector where each entry u_j corresponds to the cumulative utility that item j provides across all users and ranking positions. The utility is computed by aggregating the scores weighted by the exposure at each position in the ranked list. The Equation 1 summarizes the item quality and utility formally.

$$\textit{Item quality:} \quad q_j = \sum_{i \in \mathcal{U}} \mu_{ij} \qquad \textit{Item utility:} \quad u_j = \sum_{i \in \mathcal{U}} P_{ij}\boldsymbol{v} = \sum_{r \in [\![k]\!]} P_{ijr} v_r \qquad (1)$$

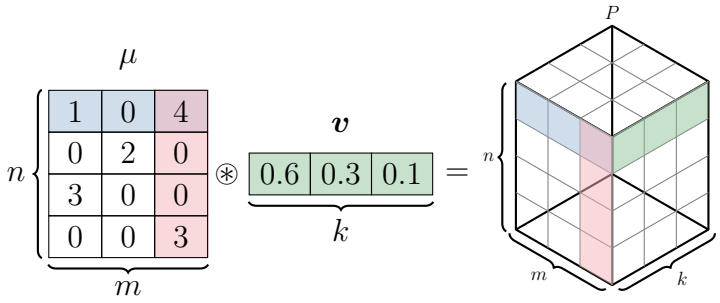

Figure 1. The computation of the matrix P relies on the scalar-vector multiplication of each entry in the 2D score matrix μ which is multiplied by the exposure weight vector v, producing a 3D matrix P. This matrix encodes the score of item j being recommended to user i at ranking position $1 \leq r \leq k$, incorporating position-aware exposure effects.

The exposure of an item is determined by the number of times it appears in the top-k ranking lists across all users, as defined by the set of rankings $\boldsymbol{\sigma}$.

Matrix Factorization

Matrix Factorization (MF) is a fundamental technique in collaborative filtering for modeling user-item interactions [Koren et al. 2009]. It represents users and items as vectors in a shared low-dimensional latent space, enabling the prediction of unobserved interactions based on latent similarities. Formally, given a partially observed interaction matrix $\mu \in \mathbb{R}^{n \times m}$, matrix factorization approximates μ as:

$$\mu \approx UV^{\top},$$

where $U \in \mathbb{R}^{n \times e}$ and $V \in \mathbb{R}^{m \times e}$ are latent factor matrices, and $e \ll \min(n, m)$ is the embedding dimension. Each user $i \in \mathcal{U}$ and item $j \in \mathcal{I}$ are thus associated with e-dimensional embeddings, U_i and V_j, respectively. The goal is to find U and V by minimizing the reconstruction loss over the observed entries, typically using the mean squared error.

Matrix factorization provides a compact representation of latent user and item characteristics, enabling the system to generalize recommendations to unseen interactions. However, it inherently depends on historical interaction data, making it vulnerable to cases involving items with limited interaction history.

Balanced Exposure

Balanced exposure or uniform exposure is a notion of exposure fairness that aims to distribute exposure uniformly across all items or groups of items [Wu et al. 2021, Singh and Joachims 2018, Khenissi and Nasraoui 2020]. Balancing the exposure of items in the recommendation lists $\boldsymbol{\sigma}$ promotes greater item diversity and ensures equal opportunity for items to be recommended. Applying this principle directly addresses the issue of popularity bias, as illustrated in Figure 2, where a small highlighted subset of items receives disproportionately high exposure compared to the majority of items in the long-tail [Abdollahpouri et al. 2017, Yin et al. 2012]. Achieving balanced exposure would ideally result in a uniform distribution of exposure across items or item groups. As a consequence, it also mitigates the cold start problem, since newly introduced items would receive equal exposure alongside existing items, regardless of their interaction history.

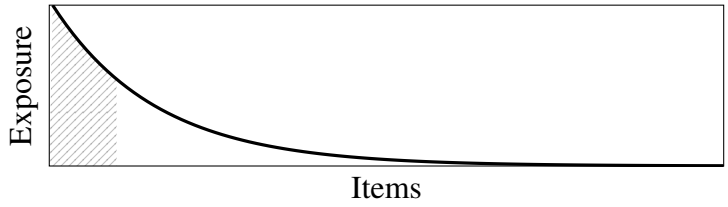

Figure 2. long-tail distribution of item exposure under popularity bias, where the short head represents frequently interacted (popular) items and the long-tail comprises less popular items with limited user interactions.

Quality-weighted Exposure

In contrast to balanced exposure, quality-weighted exposure is a fairness notion that assigns exposure proportionally to the intrinsic quality of items, ensuring that higher-quality items receive greater visibility [Biega et al. 2018, Wu et al. 2021, Morik et al. 2020]. The key idea is that exposure should reflect true quality, not historical bias. While this approach can enhance the overall utility of recommendations, it does not inherently solve the cold start problem, particularly when item quality is inferred from historical user-item interactions.

Problem Statement

Let a collaborative filtering recommender system $\mathcal{A} : \mathcal{U} \rightarrow \boldsymbol{\sigma}$ be given, where $\mathcal{A}$ produces a ranking list σ_i for each user $i \in \mathcal{U}$. Those rankings should prioritize items of high quality. However, a core limitation of collaborative filtering is its lack of adaptability to new items, a phenomenon referred to as the cold start problem. Newly introduced items, whose quality remains unknown, tend to receive low exposure regardless of their actual relevance or quality.

Naively increasing the exposure of such items can degrade overall recommendation utility, as the system risks promoting low-quality content. This highlights a fundamental trade-off between promoting fair exposure among items and preserving the utility of the recommendations.

To address these limitations, the recommender system $\mathcal{A}$ must balance two objectives: minimizing the quality loss associated with the items recommended to users in $\mathcal{U}$, and minimizing the cold start loss to ensure that newly introduced items also receive adequate exposure. We describe it as an optimization problem presented in Equation 2.

$$\begin{aligned} \min_{P} \quad & \underbrace{\mathcal{L}_{\text{qual}}(P)}_{\substack{\text{Quality-weighted} \\ \text{fairness loss}}} + \underbrace{\mathcal{L}_{\text{cold}}(P)}_{\text{cold start loss}} \\ \text{s.t.} \quad & P \subseteq \mathbb{R}^{n \times m \times k} \end{aligned} \tag{2}$$

3. TWIX: The Hybrid Approach

In this section, we present our proposed method, which introduces a hybrid notion of exposure fairness and a novel loss function designed to guide the fairness of exposure for items in user-item interaction learning. Our approach explicitly incorporates item quality into the training objective after updating the quality to a minimum threshold of items that lack interactions.

Hybrid Notion Exposure Our proposed method integrates both notions of exposure fairness: balanced exposure and quality-weighted exposure. Balanced exposure helps mitigate the effects of unreliable quality estimation caused by sparse user interactions, ensuring that items with limited feedback are not unfairly exposed. In contrast, quality-weighted exposure emphasizes the promotion of high-quality items by aligning their visibility with their estimated quality value. To incorporate these principles, we introduce **T**runcated Quality-**w**eighted **I**tem E**x**posure (TWIX), a model that learns user-item interactions while explicitly minimizing disparities between each item's actual exposure and its expected exposure. The expected exposure is computed using the two notions of exposure fairness. This encourages a fair allocation of exposure that supports both inclusiveness and merit-based.

TWIX is implemented within a matrix factorization framework, where the optimization objective is augmented with a fairness-aware loss term. This term penalizes discrepancies between the true and expected exposure-to-quality ratios, encouraging alignment between exposure levels and item quality. In this way, the model not only favors high-quality items but also ensures that items with few interactions are not entirely overlooked.

To address the unreliability of quality estimates for under-interacted items, we introduce a quality threshold, set at the 10th percentile of the estimated quality distribution. Items with quality scores below this threshold are assigned the threshold value, guaranteeing a minimum expected exposure level. This adjustment ensures that all items receive a baseline level of visibility, aligning with the platform's fairness objectives.

Truncated Quality-weighted Item Exposure

Our proposed loss function, referred to as **T**runcated Quality-**w**eighted **I**tem E**x**posure (TWIX), is designed to minimize the root mean squared error between each item's actual exposure and its expected exposure. The expected exposure is determined based on the item's estimated quality: for items with reliable quality estimates (above a defined threshold), it is proportional to their quality, and for items with low or uncertain quality (below the threshold), the expected exposure is computed using a fixed minimum quality value. This adjustment ensures fair treatment of underexposed items with limited interaction data while still prioritizing high-quality content.

TWIX is formally defined as follows:

Definition 3.1 (TWIX)**:** **T**runcated Quality-**w**eighted **I**tem E**x**posure loss is defined as:

$$\mathcal{L}(\boldsymbol{q}, \boldsymbol{u}, \boldsymbol{v}) = \frac{1}{\sqrt{m}} \left\| \boldsymbol{u} - \max\Bigg(\underbrace{\frac{E}{Q}\gamma \mathbb{1}}_{\substack{\text{Quality} \\ \text{threshold}}}, \underbrace{\frac{E}{Q}\boldsymbol{q}}_{\substack{\text{Expected} \\ \text{exposure}}} \Bigg) \right\|_2,$$

where $\gamma > 0$ is the quality threshold, and the max operation performs element-wise, the total exposure $E = n\, \mathbb{1}^T \boldsymbol{v}$, the total quality $Q = \mathbb{1}^T \boldsymbol{q}$, and $\mathbb{1}$ is a vector of ones.

The expected exposure for an item $j \in \mathcal{I}$ should be proportional to its quality, i.e., the proportion of j's quality to the total quality should match the proportion of its exposure to the total exposure. Mathematically, $\frac{q_j}{Q} = \frac{e_j}{E}, \forall j \in I$. Hence, the expected exposure is given by $e_j = \frac{E}{Q} q_j$ [Do et al. 2021, Usunier et al. 2022].

In cases where an item's quality falls below a predefined threshold γ, its quality is adjusted to γ. Consequently, the exposure assigned to the item should also be proportional to γ, ensuring a minimum guaranteed level of exposure for low-quality items.

Since optimization methods rely on gradients to determine the direction of the optimal solution, TWIX is defined as a differentiable function. Its gradient, which captures the underlying interactions, is presented below.

Lemma 3.2 (Gradient of TWIX loss)**:** Recall that TWIX loss, the derivative w.r.t. $\boldsymbol{u}$ is:

$$\frac{\partial \mathcal{L}(\boldsymbol{q}, \boldsymbol{u}, \boldsymbol{v})}{\partial \boldsymbol{u}} = \frac{1}{\sqrt{m}} \frac{\boldsymbol{u} - \max\big(\gamma \mathbb{1}, \phi\, \boldsymbol{q}\big)}{\big\|\boldsymbol{u} - \max\big(\gamma \mathbb{1}, \phi\, \boldsymbol{q}\big)\big\|_2},$$

where the max operation performs element-wise, the expected exposure-quality ratio $\phi = \frac{n}{\mathbb{1}^T \boldsymbol{q}} \mathbb{1}^T \boldsymbol{v}$, and $\mathbb{1}$ is a vector of ones.

4. Related Work

Before introducing our approach to generating fair rankings, we review key prior works that study fairness of exposure in rankings. This area broadly concerns ensuring that items, content, or providers are sufficiently represented in ranked outputs, regardless of their inherent popularity or group membership.

[Singh and Joachims 2018] propose a framework balancing item exposure fairness and user utility through constraints such as demographic parity, disparate treatment, and disparate impact. Their algorithms provide fairness guarantees but may reduce user utility when item relevance differs significantly across groups.

[Abdollahpouri et al. 2017] introduce a regularization-based learning-to-rank method to balance exposure between two item groups: short-head and medium-tail items. While improving fairness for these groups, the approach does not address distant-tail items, leaving cold start issues unresolved.

[Biega et al. 2018] propose *equity of attention*, aligning item exposure with relevance (similarly to quality-weighted) over time via amortized fairness. Their online optimization improves long-term fairness with minimal utility loss, but assumes accurate relevance estimates and is less suitable for one-time rankings.

[Do et al. 2021] extend fairness to both users and items using Lorenz efficiency, optimizing a novel loss function for social welfare and equitable exposure with the Frank-Wolfe algorithm [Frank et al. 1956]. However, quality estimates based on sparse data may lead to biased exposure allocations.

TWIX. We tackle fairness under unreliable quality estimates by introducing distinct exposure fairness criteria for short-head and long-tail items. This strategy ensures baseline visibility for all items, while still prioritizing high-quality items, thereby potentially enhancing overall utility [Boratto et al. 2021, Abdollahpouri et al. 2019]. Table 4 summarizes the limitations of prior work that our method seeks to overcome. In particular, we address the challenge of limited interactions for certain items, which can result in unreliable quality estimates, exacerbating the cold start problem and leading to unfair exposure, especially for independent or less popular items.

Paper	Fairness notion	Address quality unreliability	Favors high-quality
[Singh and Joachims 2018]	Group Fairness	✗	✗
[Abdollahpouri et al. 2017]	Balanced Exposure	✗	✗
[Biega et al. 2018]	Equity of attention	✗	✓
[Do et al. 2021]	Quality-Weighted OR Balanced Exposure	✗	✓
TWIX	Quality-Weighted AND Balanced Exposure	✓	✓

Table 4. Comparison between prior work and ours, highlighting the main limitations that we address.

5. Experiments

This section details the methodology employed in the experiments and the experimental evaluation, comparing our approach to the baseline.

5.1. Methodology

Our method is a collaborative filtering framework that uses matrix factorization to learn item rankings for users. The model is trained using the TWIX loss function (see Definition 3.1), which is specifically designed to incorporate fairness considerations into the recommendation process. For comparison, we adopt a baseline model that also employs matrix factorization but is trained using the standard mean squared error (MSE) loss. This baseline optimizes the predicted user-item scores by minimizing the squared difference between the predicted and observed scores, without incorporating any fairness-aware components for item exposure.

For evaluating recommendation utility, we use the widely adopted Normalized Discounted Cumulative Gain (NDCG) [Järvelin and Kekäläinen 2002]. NDCG considers the positions of relevant items in the ranked list, assigning higher scores, i.e., closer to one, to methods that place relevant items near the top, resulting in an effective ranking.

To assess fairness in item exposure, we introduce the Exposure Root Mean Squared Error (eRMSE), a metric that quantifies the discrepancy between the actual and ideal exposure of items. The actual exposure is computed based on item positions in the ranked lists produced by the model, typically using a position-based decay function. The ideal exposure reflects a fairness-driven target distribution, uniform or weighted by item quality, representing the intended visibility each item should receive. The eRMSE is calculated as the root mean squared difference between the actual and ideal exposure vectors across all items. A lower eRMSE indicates that the model distributes exposure more fairly, which is in line with the defined fairness objectives.

Importantly, eRMSE captures two complementary notions of fairness: quality-weighted exposure, where higher-quality items are expected to receive more exposure, and balanced exposure, which prevents overexposure of a few items at the expense of others. To address the unreliability of quality estimates, particularly for items with few or no interactions,

we apply a quality threshold: any item with an estimated quality below this threshold is assigned the threshold value. This adjustment helps avoid severely underestimating the ideal exposure of such items, thereby reducing unfair penalization due to sparse data and ensuring a more robust fairness evaluation.

5.2. Experimental evaluation

In this section, we present the experimental setup and results of our evaluation. We first describe the datasets used, followed by the model configuration and hyperparameter settings. Finally, we provide a comparative analysis of our approach against the baseline model, focusing on both utility and fairness metrics.

Datasets

This work utilizes three publicly available datasets: MovieLens, Amazon, and LastFM, each preprocessed as user-item interaction records, i.e., each entry represents the interaction/rating of a user i to an item j. Below, we provide a detailed description of each dataset.

The MovieLens 100K dataset [Harper and Konstan 2015] is a widely adopted benchmark in recommender systems research. It consists of 100,000 explicit ratings (on a 1–5 scale) provided by 943 users across 1,682 movies, with each user having rated at least 20 movies.

The Amazon Magazine Subscriptions dataset [Ni et al. 2019] contains product reviews from Amazon, including 89,689 reviews from 75,258 users on 2,789 magazines. Reviews are accompanied by explicit 1–5 star ratings. Due to hardware limitations, we subsampled the dataset to 40,000 reviews while preserving its statistical properties.

The LastFM 360K dataset [Celma 2010] comprises user listening histories, originally containing 92,831 users and 1,000,000 music tracks. Due to its large scale, we employed a subset of 10,000 listening histories to facilitate experimentation while maintaining representativeness for music recommendation tasks.

Model configuration

All experiments were conducted in Python using PyTorch with CUDA acceleration on an NVIDIA GPU. The datasets were partitioned into an 80-20 train-test split.

We evaluated our model across multiple hyperparameter configurations: recommendation size $k = [3, 5, 7, 15]$, latent factor dimensions $e = [10, 25, 50, 75]$, and a fixed training regimen of 200 epochs with a learning rate of 0.05 along with AdamW optimizer. Each configuration was run 5 times.

Comparative analysis and evaluation

In all configuration scenarios, TWIX outperformed the baseline model in terms of eRMSE, i.e., fairness of exposure metric. These results indicate that our approach effectively addresses the quality estimation limitations present in the baseline, leading to a more equitable exposure distribution.

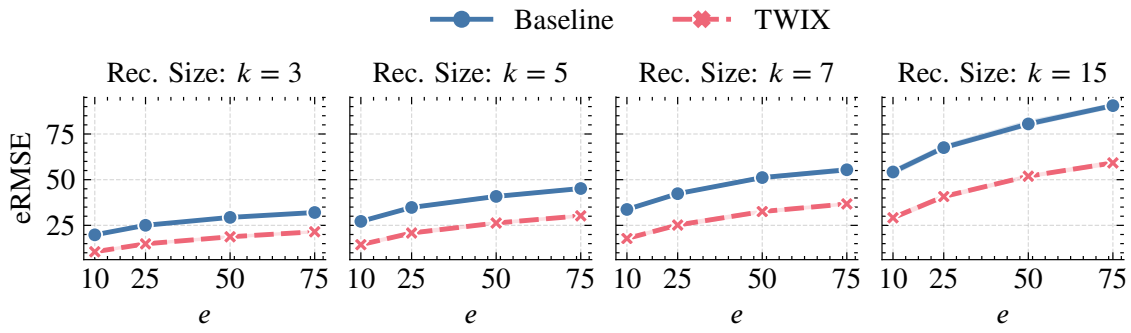

Figure 3. Fairness comparison between baseline and TWIX for MovieLens dataset.

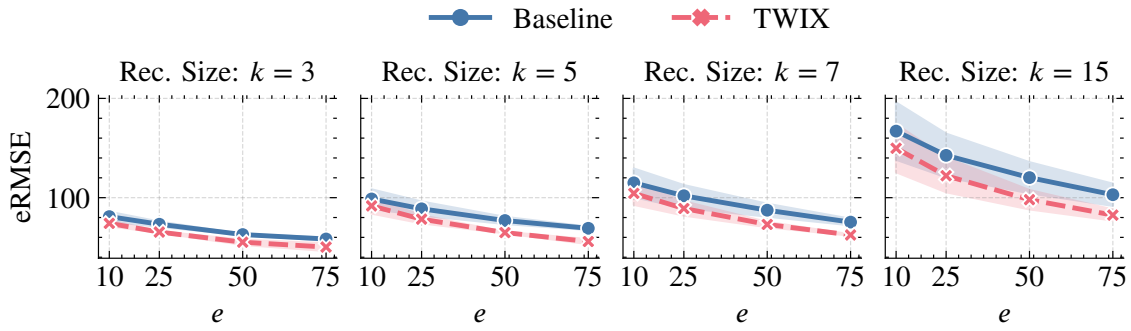

Figure 4. Fairness comparison between baseline and TWIX for Amazon dataset.

Figure 4 presents eRMSE results for the Amazon dataset, where TWIX shows both superior performance to the baseline and a decreasing trend with larger latent dimensions. This pattern, however, is not consistently observed across all datasets, as eRMSE exhibits varying dependencies on latent size in other cases. Figure 5 reveals that eRMSE for $k = 3$ and $k = 5$ follows the characteristic U-shaped pattern, illustrating the classic bias-variance trade-off and the transition between overfitting and underfitting regimes, e.g., smaller latent sizes suggest overfitting and larger latent sizes indicate underfitting. Additionally, Figure 3 demonstrates that increasing the latent dimension size can paradoxically lead to higher eRMSE values, suggesting potential underfitting in the model's performance.

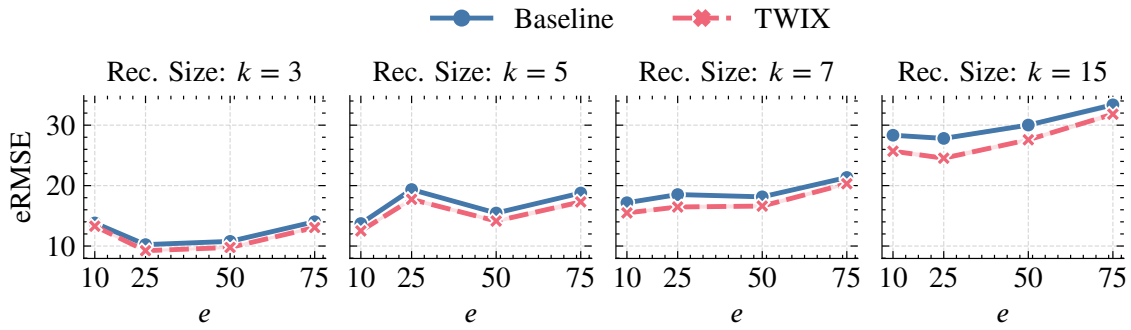

Figure 5. Fairness comparison between baseline and TWIX for LastFM dataset.

Our proposed method, TWIX, also outperformed the baseline regarding recommendation

utility, as shown in Table 5. The NDCG metric indicates that our approach consistently achieved higher scores across all datasets, demonstrating its effectiveness in improving recommendation quality while maintaining fairness. Our method exhibited a significant improvement in NDCG for the LastFM dataset, where the baseline achieved an NDCG of 0.026486 ± 0.019744, while our method reached 0.062210 ± 0.065152. Howerver TWIX suffers from higher variance than baseline. This suggests that while our method can achieve better performance, it may also be more sensitive to the specific characteristics of the dataset and the hyperparameter settings.

Dataset	Method	NDCG
Amazon	Baseline	0.829983 ± 0.115985
	TWIX	$\mathbf{0.831973} \pm 0.133127$
LastFM	Baseline	0.026486 ± 0.019744
	TWIX	$\mathbf{0.062210} \pm 0.065152$
MovieLens	Baseline	0.646303 ± 0.134496
	TWIX	$\mathbf{0.654868} \pm 0.137016$

Table 5. Utility comparison between baseline and TWIX.

6. Conclusion

We propose a hybrid method to address the fairness of exposure in recommender systems. While favoring high-quality items, our approach mitigates issues arising from limited interaction data, such as the cold start problem and unreliable quality estimation. Ideally, item exposure should be proportional to true quality. However, for items with quality estimates below a certain threshold, we treat these values as underestimated. We apply a quality adjustment for such items in our static recommendation setting. As a result, exposure is computed proportionally to the item's true quality if it exceeds the threshold, or proportionally to the threshold itself if it does not. This adjustment effectively reshapes the long-tail distribution, mitigating popularity bias by enforcing a minimum exposure level for underrepresented items.

We compare our approach against a Matrix Factorization baseline optimized for utility. Using widely adopted benchmark datasets, we empirically show that our method preserves recommendation utility while significantly improving exposure fairness.

Future Direction. A promising direction for future work is to extend our method to an online setting. In a dynamic scenario, item quality can be continuously updated and recalculated to better estimate true quality across both the short-head and long-tail of the exposure distribution. By ensuring that items with few interactions receive greater exposure, and thus more opportunities for user feedback, their quality estimates can become more accurate over time.

Acknowledgments

This research was partially funded by CAPES under grant number 88887.694847/2022-00. It was also partially funded by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) (grant # 316729/2021-3).

References

[Abdollahpouri et al. 2017] Abdollahpouri, H., Burke, R., and Mobasher, B. (2017). Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the eleventh ACM conference on recommender systems*, pages 42–46.

[Abdollahpouri et al. 2019] Abdollahpouri, H., Mansoury, M., Burke, R., and Mobasher, B. (2019). The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286*.

[Biega et al. 2018] Biega, A. J., Gummadi, K. P., and Weikum, G. (2018). Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 405–414.

[Boratto et al. 2021] Boratto, L., Fenu, G., and Marras, M. (2021). Connecting user and item perspectives in popularity debiasing for collaborative recommendation. *Information Processing & Management*, 58(1):102387.

[Celma 2010] Celma, O. (2010). *Music Recommendation and Discovery in the Long Tail*. Springer.

[Do et al. 2021] Do, V., Corbett-Davies, S., Atif, J., and Usunier, N. (2021). Two-sided fairness in rankings via lorenz dominance. *Advances in Neural Information Processing Systems*, 34:8596–8608.

[Frank et al. 1956] Frank, M., Wolfe, P., et al. (1956). An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110.

[Harper and Konstan 2015] Harper, F. M. and Konstan, J. A. (2015). The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19.

[Järvelin and Kekäläinen 2002] Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

[Khenissi and Nasraoui 2020] Khenissi, S. and Nasraoui, O. (2020). Modeling and counteracting exposure bias in recommender systems. *arXiv preprint arXiv:2001.04832*.

[Koren et al. 2009] Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

[Merton 1968] Merton, R. K. (1968). The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810):56–63.

[Morik et al. 2020] Morik, M., Singh, A., Hong, J., and Joachims, T. (2020). Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 429–438.

[Natarajan et al. 2020] Natarajan, S., Vairavasundaram, S., Natarajan, S., and Gandomi, A. H. (2020). Resolving data sparsity and cold start problem in collaborative filtering recommender system using linked open data. *Expert Systems with Applications*, 149:113248.

[Ni et al. 2019] Ni, J., Li, J., and McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197.

[Sang et al. 2024] Sang, S., Ma, M., and Pang, H. (2024). Weighted matrix factorization recommendation model incorporating social trust. *Applied Sciences*, 14(2):879.

[Singh and Joachims 2018] Singh, A. and Joachims, T. (2018). Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2219–2228.

[Usunier et al. 2022] Usunier, N., Do, V., and Dohmatob, E. (2022). Fast online ranking with fairness of exposure. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2157–2167.

[Wu et al. 2021] Wu, Y., Cao, J., Xu, G., and Tan, Y. (2021). Tfrom: A two-sided fairness-aware recommendation model for both customers and providers. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1013–1022.

[Yin et al. 2012] Yin, H., Cui, B., Li, J., Yao, J., and Chen, C. (2012). Challenging the long tail recommendation. *arXiv preprint arXiv:1205.6700*.