Fair Sequential Recommendation without User Demographics

Huimin Zeng
huiminz3@illinois.edu
University of Illinois, Urbana-Champaign
Champaign, USA

Zhankui He
zhh004@ucsd.edu
University of California, San Diego
San Diego, USA

Zhenrui Yue
zhenrui3@illinois.edu
University of Illinois, Urbana-Champaign
Champaign, USA

Julian McAuley
jmcauley@ucsd.edu
University of California, San Diego
San Diego, USA

Dong Wang
dwang24@illinois.edu
University of Illinois, Urbana-Champaign
Champaign, USA

ABSTRACT

Much existing literature on fair recommendation (i.e., group fairness) leverages users' demographic attributes (e.g., gender) to develop fair recommendation methods. However, in real-world scenarios, due to privacy concerns and convenience considerations, users may not be willing to share their demographic information with the system, which limits the application of many existing methods. Moreover, sequential recommendation (SR) models achieve state-of-the-art performance compared to traditional collaborative filtering (CF) recommenders, and can represent users **solely** using user-item interactions (user-free). This leaves a wrong impression that SR models are free from group unfairness by design. In this work, we explore a critical question: how can we build a fair sequential recommendation system without even knowing user demographics? To address this problem, we propose **Agnostic FairSeqRec (A-FSR)**: a model-agnostic and demographic-agnostic debiasing framework for sequential recommendation without requiring users' demographic attributes. Firstly, A-FSR reduces the correlation between the potential stereotypical patterns in the input sequences and final recommendations via Dirichlet neighbor smoothing. Secondly, A-FSR estimates an under-represented group of sequences via a gradient-based heuristic, and implicitly moves training focus towards the under-represented group by minimizing a distributionally robust optimization (DRO) based objective. Results on real-world datasets show that A-FSR achieves significant improvements on group fairness in sequential recommendation, while outperforming other state-of-the-art baselines.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

recommender systems, group fairness, sequential recommendation, model agnostic, demographic agnostic

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657703

ACM Reference Format:
Huimin Zeng, Zhankui He, Zhenrui Yue, Julian McAuley, and Dong Wang. 2024. Fair Sequential Recommendation without User Demographics. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3626772.3657703

1 INTRODUCTION

Recommendation systems help users to efficiently select contents that meet their needs, such as online commerce [37], social media [6], and web recommendations [26]. However, concerns have been raised about group fairness in recommendations. Recently, significant disparities of recommendation performance are observed across different demographic groups [22, 30, 34], which results in group unfairness. In contrast, a fair recommender shall achieve the same or comparable recommendation performance for user groups with different demographic attributes. Otherwise, discrimination in recommendation violates ethical regulations [7, 24].

Moreover, many recommenders are user-free in that they do not build a user model in their recommendations. As such, users' demographic information is not usually available to develop fair methods for such systems. As one of the important user-free recommenders, Sequential Recommendation (SR) systems capture user dynamics and show superior performance over traditional recommenders (e.g., content-based filtering) [10, 14]. More importantly, a user-free SR model can accurately recommend desired items without requiring additional user information (e.g., user profile or demographics) [18, 20]. However, SR still suffers from a non-trivial performance bias across different demographic groups [19].

Efforts have been made to improve group fairness in recommendations [31]. Under group fairness, the recommender shall achieve comparable performance for different demographic groups. However, much existing literature on fair recommendation (i.e., group fairness) assumes that demographic attributes (e.g., gender) are present in the user datasets, and relies upon such an assumption to develop fair methods [19, 23, 31, 36]. This assumption could be rather impractical in real-world applications. For instance, GDPR [1] imposes constraints on collecting and using sensitive human features for decision making. These privacy constraints indicate that collecting user data with demographic information could be

[1] General Data Protection Regulation:https://gdpr-info.eu/

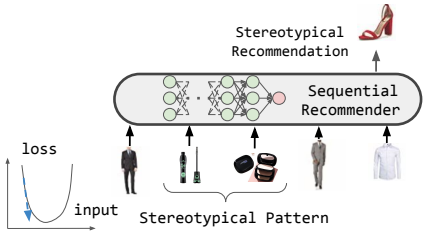

Figure 1: An illustrative example of demographic-stereotypical patterns. Larger training penalty (e.g., gradients) could be incurred w.r.t. such demographic-stereotypical patterns (cosmetic-related items). In comparison, the training penalty could be smaller on the items that are more related to the expected recommendation (i.e., the male dressing shoes), because suit, male dressing shoes, shirts belong to a roughly same category of male clothes.

problematic when building recommendation systems. In addition, many users may choose to not share their sensitive demographic information with the system due to privacy concerns and convenience considerations [11]. Therefore, we study the question of **how to build a *fair* sequential recommendation system *without* knowing user demographics.**

In this work, we aim to develop a model-agnostic and demographic-agnostic method, which can be deployed for arbitrary sequential recommender systems (e.g., BERT4Rec [27], SASRec [18] or NARM [20]). On one hand, our method is motivated by the fact that user demographics may not be available in all datasets, yet fair recommendations are often needed. On the other hand, existing fair SR solutions (e.g., [19]) focus on designing specific model architectures, which are not suitable for general SR models (e.g., transformer-based or RNN-based models). The challenges of our setting are three-fold: (1) No user demographic information: as a consequence, it is infeasible to directly measure and reduce model bias during training as in existing literature. (2) Besides fairness, the framework shall maintain a satisfactory overall performance since trivial fair solutions (e.g., reducing recommendation performance for all groups) should be avoided. (3) The debiasing framework is expected to be applicable for arbitrary sequential recommender architectures.

To better illustrate our method, we present an example in Figure 1. In this example, assume a male shopping a suit for an event. However, the cosmetic-related items in his browse history might mislead the system to recommend high-heels instead of dress shoes. Since the cosmetic-related items could be more related to female products, the system makes such stereotypical recommendation. In practice, men could also use cosmetics for important events. In this example, the cosmetic-related items in his browse history could be regarded as stereotypical patterns. More formally, the demographic-stereotypical patterns are certain sub-sequences within user data that likely mislead the recommender to return stereotypical recommendations. Such stereotypical recommendations lead to performance bias (i.e. recommendations based on the guessed demographic information instead of the true browse history or user-item interactions). Therefore, our method debiases the model by mining such stereotypical patterns from user data.

Model	Dataset	Stats.	@Male/@Female
BERT4Rec (Biased)	ML-100K	NDCG@3	0.1985/0.1437
		Loss ($\mathcal{L}$)	6.6787/6.7967
		Grad. Norm ($\|\nabla\|$)	0.0722/0.1691
	ML-1M	NDCG@3	0.4045/0.3629
		Loss ($\mathcal{L}$)	6.7759/6.8361
		Grad. Norm ($\|\nabla\|$)	0.1246/0.3314
	LastFM	NDCG@3	0.6659/0.5480
		Loss ($\mathcal{L}$)	9.0142/9.6229
		Grad. Norm ($\|\nabla\|$)	0.2720/0.4138

Table 1: Pilot Study: the performance, training loss and the gradient norm of BERT4Rec without any debiasing methods. The female group is the under-represented group.

In the example shown in Figure 1, we hypothesize that the performance bias of a recommender is related to its training gradients w.r.t. the items. In fact, existing studies [5, 28] have also found that gradients w.r.t. the inputs are efficient sensitivity measurements of the model for each input element. Investigating gradients is a primary approach to detecting bias and shortcut learning (i.e., learning superficial data features) of the model [2]. In terms of recommendation, such superficial data features are more likely to be stereotypical patterns [25]. To further demonstrate the relationship between the performance bias and training gradients in our specific SR setting, we conduct a pilot study for BERT4Rec on three different datasets. In the pilot study, we train the BERT4Rec model using the biased datasets without any debiasing methods. After training, we calculate three key statistics (i.e., NDCG@3, the loss, and norm of gradients w.r.t. item embeddings) for the trained biased BERT4Rec on different demographic groups (i.e., male and female)[2]. The results are reported in Table 1. We observe that the under-represented female group experiences worse recommendation performance compared to the male group. More importantly, by comparing three statistics between male group and female group, it is clear that larger training penalty (e.g., gradients) was always incurred on the under-represented group (i.e., female in the example above), which supports our hypothesis.

Motivated by the above observations, our framework firstly detects demographic-stereotypical patterns in the input user data without users' demographic information. To achieve this, we leverage the training gradients w.r.t. the input sequences, and treat items with the largest gradients as stereotypical patterns as motivated in Figure 1 and Table 1. Then, after detecting the stereotypical patterns, we present a novel Dirichlet neighbor smoothing (DNS) module to debias the recommender. Our proposed DNS is a randomized method and blurs the correlation between the stereotypical patterns and predictions. DNS reduces the bias of the model and preserves the temporal transition dynamics within the sequences, which helps maintain an overall high recommendation performance of the debiased model. This is also the rationale that we only focus on the sub-sequences with largest gradients instead of using the gradients of the entire input sequences: we still want to preserve the original user dynamics (interacted items) in the history. Finally, we estimate under-represented users with the detected stereotypical patterns,

[2]The detailed experiment settings are discussed in Section 5. In this pilot study, we used gender labels for evaluation, but they are not used during training.

and improve the worst-case performance on them to enhance the robustness of the fairness improvements on the training data. The rationale behind the worst-case optimization is to address the discrepancy between the estimated under-represented group and the true under-represented one. Eventually, the recommendation performance on the potential minority group is improved, achieving the desired group fairness in recommendation.

We summarize our contributions as follows[3]:

(1) To the best of our knowledge, we are the first to propose a universal debiasing framework without requiring user demographics for sequential recommendation.
(2) Specifically, we present two novel modules: (1) Dirichlet Neighbor Smoothing, a randomized method that blurs the correlation between the stereotypical patterns and predictions; (2) Worst-case Performance Optimization: a Distributionally Robust Optimization (DRO) based module that enhances the robustness of the fairness improvements on potential under-represented users.
(3) We demonstrate the effectiveness of our method with extensive experiments over multiple real-world datasets. In terms of group fairness, our results suggest that the proposed A-FSR consistently outperforms baseline methods when user demographics are not available for training.

2 RELATED WORK

2.1 Fair Recommendation

Fairness has increasingly become a critical objective when developing modern recommender systems [29, 35]. Wang et al. [30] systematically investigate various notions of fairness in recommendation from different perspectives, such as user perspective, item perspective and system perspective. Among all the fairness notions, a large group of studies, including our work, focus on the group fairness of recommendation for its ethical implications in real-world applications. For instance, Beutel et al. [1] proposed new metrics to quantify fairness in recommendation, and added corresponding regularization to improve fairness. Li et al. [21] studied the fairness between active and inactive users, and presented a constrained re-ranking approach. Wei and He [31] studied group fairness in recommendation using adversarial learning and meta learning techniques. However, the above methods assume the demographic attributes of the users are present in the datasets. In practice, due to privacy concerns and legal regulations, collecting user demographic information could be infeasible [11]. Moreover, only limited solutions (e.g., FairSR [19]) are developed to address the bias issue in sequential recommendation. More importantly, FairSR still leverages demographic information of the users, and only works on the specific model architecture (i.e., FairSR) designed in Li et al. [19]. Compared to existing literature, our proposed A-FSR is a demographic-agnostic fair solution and could be deployed for any existing sequential recommenders. And, we note that many current fair recommendation methods are not applicable for sequential recommenders. Therefore, upon evaluation, we only select baseline methods that could be modified to sequential recommenders and exclude methods that require unique model architectures. Finally, we highlight that this work focuses on group fairness, whereas the item fairness (i.e., popularity fairness) [30] is not within this scope of this work.

[3]We adopt publicly available datasets in our experiments and will release the code upon acceptance.

2.2 Sequential Recommendation

Sequential recommendation (SR) achieves state-of-the-art performance and is user-free [14, 16, 38, 39]. SR models take a sequence of interacted items as input, and make recommendations for users by generating items that meet users' interests [18, 20, 27]. For instance, several Recurrent Neural Network (RNN) based SR models are proposed in [15, 20]. Inspired by natural language modeling, He et al. [14], Kang and McAuley [18], Sun et al. [27] proposed transformer based architectures to build sequential recommenders. Moreover, Graph Neural Network (GNN) based models have also been proven to be efficient in sequential recommendation [4, 33]. However, many of these studies overlooked the group unfairness issue of such models. A fair SR solution that could be deployed for arbitrary sequential models is still missing despite the existence of FairSR [19]. More importantly, user-free SR models make recommendations without additional user information (e.g., user profile or demographics) [18, 20]. Such design of sequential recommenders creates a misleading impression that SR models are automatically free from the group unfairness issue, which is not true as shown in [19]. The question of debiasing sequential recommender systems without user demographics still remains.

3 PRELIMINARIES

3.1 Data

An SR model takes a sequence of interacted items x (sorted by timestamps) as input, and recommends items that meet users' needs. A sequence x is a list of items $[x_1, x_2, ..., x_l]$ of length l. Each element in x belongs to the item scope $\mathcal{I}$ that contains all items: $x_i \in \mathcal{I}$. The next user-item interaction $x_{l+1} \in \mathcal{I}$ after x is used as ground truth y (i.e., $y = x_{l+1}$) in our sequential recommendation setting.

Assume a dataset $\mathcal{D}$, containing $|\mathcal{D}|$ sequences of interacted items. $x^{(m)}$ denotes the m-th sequence in $\mathcal{D}$. Each x corresponds to a user. Users who share the same demographic attribute form a demographic group. Our goal is to develop a fair recommender that achieves comparable performance for different demographic groups. However, we assume demographic attributes are not available to train the model, but we will use them for evaluation.

3.2 Model

A sequential recommender is defined as a function f, mapping input sequences into next-item recommendations. Given an input sequence x, f computes a probability distribution over the item scope $\mathcal{I}$, and makes recommendations of items with highest probabilities: $y = \arg\max f(x)$. Moreover, for a better understanding of our framework, we highlight that f consists of an embedding function f_e and a sequential model f_m, where $f(x) = f_m(f_e(x))$ as in [39]. The embedding function f_e transforms the discrete item ids into continuous item embeddings.

3.3 Group Fairness

Under group fairness [8, 31], the recommender should achieve the same recommendation performance for users from different groups (e.g., demographic groups). For instance, although the shopping pattern (e.g., purchasing gender-associated items) could be different between male and female users, a fair recommender is expected to show similar and satisfactory performance for both gender groups. As in Wei and He [31], to measure the **group fairness**, we compute the performance gap between two different demographic groups:

$$\Phi = |\frac{1}{|A_1|}\sum_{x\in A_1} R(x,y) - \frac{1}{|A_2|}\sum_{x\in A_2} R(x,y)|, \tag{1}$$

where A_1 and A_2 refer to two different demographic groups (e.g., A_1 represents the female user group, and A_2 represents the male user group). R could be any metric that evaluates the recommendation performance (e.g., Recall or NDCG [17]). Intuitively, a lower Φ represents a better fairness performance in recommendation. Note that Φ in Equation 1 is defined over two demographic groups. Natural extensions of Equation 1 over multiple demographic groups could be developed by computing the sum of performance differences over all possible demographic groups or by comparing the best-performance group and the worst-performance group.

3.4 Optimization

Traditionally, an optimal recommender is obtained by minimizing the empirical loss over $\mathcal{D}$:

$$f^* = \min_f \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f(x),y)], \tag{2}$$

where $\mathcal{L}$ represents the training loss function. Note that Equation 2 does not take group fairness into consideration and the resulting recommender could be severely biased.

4 ALGORITHM

We first show how to locate the stereotypical patterns in the input data. Then, we present the two novel modules of A-FSR, Dirichlet neighbor smoothing and worst-case performance optimization.

4.1 Demographic-Stereotypical Patterns

The demographic-stereotypical patterns are sub-sequences within user sequences, which likely cause the shortcut learning of the model and mislead it to return biased recommendations. We leverage gradients w.r.t. input items to locate potential demographic-stereotypical patterns without requiring user demographic attributes. For the sake of simplicity, we propose to locate such patterns in the continuous and differentiable embedding space instead of the discrete item space. Since gradients w.r.t. the inputs are efficient sensitivity measurements of the model as shown in Table 1, our framework treats the embedding with the largest gradients and its adjacent embeddings as potential demographic-stereotypical patterns. Embeddings with higher gradients suggest a stronger penalty on such input items, which are more likely to trigger shortcut learning of the recommender. Thus, along with their adjacent embeddings, the embedding with the largest gradient is more likely to form the stereotypical pattern in sequential recommendation, compared to the remaining item embeddings.

Assume an input sequence of items $x = [x_1, x_2, ..., x_l]$, we firstly locate the embedding z_{i^*} with the largest gradient:

$$z_{i^*} = [f_e(x)]_{i^*},$$
$$\text{where} \quad i^* = \arg\max_{i\in[l]} \{||\frac{\partial\mathcal{L}(f(x),y)}{\partial z_1}||, ..., ||\frac{\partial\mathcal{L}(f(x),y)}{\partial z_l}||\}. \tag{3}$$

In Equation 3, $z_i = [f_e(x)]_i$ represents the embedding of the i-th item in x. With z_{i^*}, we further include the adjacent embeddings of z_{i^*} to construct the demographic-stereotypical pattern in the embedding space. That is, the demographic-stereotypical pattern z_{stereo} in the embedding space for x is a span of embeddings with size $2s+1$ centered at z_{i^*}:

$$z_{stereo} = [z_{i^*-s}, z_{i^*-s+1}, ..., z_{i^*}, z_{i^*+1}, ..., z_{i^*+s}]. \tag{4}$$

Note that z_{stereo} is a sub-sequence of z. Correspondingly, in the item space, the stereotypical pattern x_{stereo} is derived by mapping embeddings into the item space, which is also a sub-sequence of the original sequence x:

$$z = [\underbrace{z_1, ..., \overbrace{z_{i^*-s}, ..., z_{i^*}, ..., z_{i^*+s}}^{\textbf{located stereotypical embeddings}}, ..., x_l}_{\textbf{item embeddings of the input sequence } x}]. \tag{5}$$

4.2 Dirichlet Neighbor Smoothing

The demographic-stereotypical patterns lead to stereotypical recommendations. To reduce the correlation between such patterns and the recommendations, we propose Dirichlet Neighbor Smoothing (DNS), a randomized approach to replace the stereotypical embeddings with the embeddings of other items. To begin with, to determine the neighbors for any item x_i, we compute the cosine similarity between all item pairs in the embedding space. Next, we formally define the neighborhood $\mathcal{N}_k$ of an item x_i as follows:

Definition 4.1 (Neighborhood $\mathcal{N}_k$ of an item x_i). The neighborhood $\mathcal{N}_k$ of an item x_i is a set of items, whose elements have top-k cosine similarity with item x_i in the embedding space. The parameter k controls the threshold of becoming a neighbor of x_i:

$$\mathcal{N}_k(x_i) = \arg\max_{\mathcal{N}\subset\mathcal{I},|\mathcal{N}|=k,x_i\notin\mathcal{N}} \sum_{x_j\in\mathcal{N}} CosSim(f_e(x_i), f_e(x_j)). \tag{6}$$

We refer to $\mathcal{N}_k(x_i)$ as the 1-hop neighborhood of x_i, since the items in $\mathcal{N}_k$ are the most similar to x_i in the embedding space. However, $\mathcal{N}_k(x_i)$ could still be demographic-stereotypical. Therefore, we propose to use a further and less-similar neighborhood of the 1-hop neighbors by taking neighbors of neighbors: the 2-hop neighborhood of x_i, constructed using all 1-hop neighbors of x_i's neighbors. To differentiate the 1-hop neighborhood and 2-hop neighborhood of x_i, we add superscripts in $\mathcal{N}_k^{(1)}(x_i)$ and $\mathcal{N}_k^{(2)}(x_i)$.

With defined neighbors, we then compute a convex hull in the embedding space for each item x_i (Figure 2). This convex hull is the smallest convex polygon that encloses all items in the neighborhood, and is spanned by x_i's multi-hop neighbors. Since we aim to reduce the bias incurred by the stereotypical patterns, we propose to substitute each item embedding in the stereotypical pattern with embeddings of other items, which are sampled from its multi-hop neighborhoods. Herein, the convex hull is used as

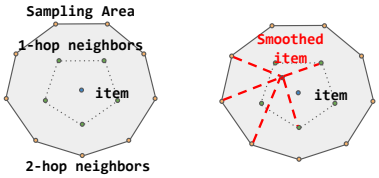

Figure 2: Dirichlet Neighbor Smoothing. Left: we construct the multi-hop neighbors for each item using cosine similarity among items, and sample from the neighborhood according to a Dirichlet distribution. Right: we compute the smoothed item embedding for each item embedding within the located stereotypical pattern, where the smoothing item is a linear combination of sampled neighbors.

the sampling space, and we sample substitutions from the convex hull according to a Dirichlet distribution. The Dirichlet distribution is defined by the vertices (i.e., neighbors of a given item) in the convex hull. As in Figure 2 (right), by sampling multiple neighbors in the neighborhood, we compute 'the expected item' to replace the items in the stereotypical pattern. Formally, for item x_i and the set of its multi-hop neighbors $\mathcal{N}(x_i) = \mathcal{N}_k^{(1)}(x_i) \cup ... \cup \mathcal{N}_k^{(C)}(x_i)$, we represent its neighboring items with η_i, which are sampled from a Dirichlet distribution:

$$\eta_i = [\eta_{i,1}, \ldots, \eta_{i,|\mathcal{N}(x_i)|}] \sim Dirichlet(\beta_1, \ldots), \tag{7}$$

where βs are parameters for the Dirichlet distribution.

We first look up multi-hop neighbors for the items in the stereotypical pattern: $\mathcal{N}(x_i)$ for $x_i \in x_{stereo}$. Then, we sample η_i from $\mathcal{N}(x_i)$ according to the Dirichlet distribution. Finally, we compute the expected embedding vector using η_i and all elements in $\mathcal{N}(x_i)$ to update the original stereotypical pattern of embeddings within each sequence as follows:

$$\forall x_i \in x_{stereo} = \{x_{i^*-s}, x_{i^*-s+1}, ..., x_{i^*}, x_{i^*+1}, ..., x_{i^*+s}\}$$
$$\forall n_{i,j} \in \mathcal{N}(x_i), \quad \tilde{z}_i = \sum_{j=1}^{|\mathcal{N}(x_i)|} \eta_{i,j} \cdot f_e(n_{i,j})$$

where $n_{i,j}$ represents the j-th item in $\mathcal{N}(x_i)$, namely the j-th neighbor of x_i. Eventually, after performing DNS for each item embedding of the stereotypical pattern within a sequence, we obtain a smoothed sequence embedding $\tilde{z}$:

$$\tilde{z} = [\underbrace{z_1, ..., \overbrace{\tilde{z}_{i^*-s}, \tilde{z}_{i^*-s+1}, ..., \tilde{z}_{i^*}, \tilde{z}_{i^*+1}, \tilde{z}_{i^*+s}}^{\textbf{Dirichlet Smoothed Embeddings}}, ..., z_l}_{\textbf{item embeddings of the entire input sequence}}]$$

Note that we select Dirichlet distribution in our method, because it assigns higher importance weights to the top-place values and lower importance weights to the tail-values in a controllable fashion. In our sequential recommendation, it matches the intuition of our smoothing step: preserving the user dynamics by assigning more importance weights to similar items in the closest neighborhood, and using further and less similar neighborhoods to reduce the bias of the stereotypical patterns. Moreover, Dirichlet distribution automatically serves as a normalized re-weighting mechanism to compute the expected items: the $\sum \eta_{i,j} = 1$. This indicates that the smoothed embedding $\tilde{z}_i$ is a linear combination of embeddings of neighbors in the convex hull as shown in Figure 2.

4.3 Worst-case Performance Optimization

Besides DNS, A-FSR improves the model performance for the potential under-represented group under a worst-case scenario, which further improves model's group fairness.

However, without user demographics, it is extremely challenging to identify the under-represented group. To overcome this challenge, we observe that if a recommender is biased against a demographic group, then the poor recommendation performance usually incurs a larger loss on this group during training. The penalty strength (i.e. gradient) on such user data is prone to be larger as well. To this end, we leverage gradients w.r.t. the input data again to estimate the under-represented group. Now, let $x^{(m)}$ denote the m-th input sequence of a dataset $\mathcal{D}$. Therefore, for $x^{(m)}$, its training loss w.r.t. its item embeddings is computed as:

$$\begin{aligned}\nabla_{z^{(m)}} \mathcal{L}(f(x^{(m)}), y^{(m)}) &= \frac{\partial \mathcal{L}(f(x^{(m)}), y^{(m)})}{\partial z^{(m)}} \\ &= \frac{\partial \mathcal{L}(f(x^{(m)}), y^{(m)})}{\partial f_e(x^{(m)})}.\end{aligned} \tag{8}$$

We use the largest gradient of an item embedding as the group indicator for $x^{(m)}$. Mathematically, the group indicator $\hat{a}$ of $x^{(m)}$ is defined as the largest gradient norm of input item embeddings:

$$\hat{a}(x^{(m)}) = \max\{||\nabla_{z_1^{(m)}} \mathcal{L}(f(x^{(m)}), y^{(m)})||, ...\}. \tag{9}$$

Since the true demographic attribute $a^{(m)}$ of sequence $x^{(m)}$ is unknown during training, we propose to use $\hat{a}(x^{(m)})$ as the proxy of $x^{(m)}$'s demographic attribute. To this end, we construct an **estimated** under-represented demographic group $\hat{A}_{under}$, by selecting sequences that have top-M largest group indicators:

$$\hat{A}_{under} = \arg \max_{\hat{A} \subset \mathcal{D}, |\hat{A}|=M} \sum_{x^{(m)} \in \hat{A}} \hat{a}(x^{(m)}). \tag{10}$$

Herein, M is a hyperparameter and is chosen empirically to determine the size of the estimated under-represented group. Note that the sequences (or users) in $\hat{A}_{under}$ are associated with the stereotypical patterns on their user-item histories. We consider the recommendation over $\hat{A}_{under}$ as the performance on the under-represented users. However, applying Equation 2 to $\hat{A}_{under}$ does not necessarily guarantee a generalized performance improvement on the test data. This is because there exists discrepancy between the estimated under-represented group $\hat{A}_{under}$ and the true under-represented demographic group. If M is too large, then $\hat{A}_{under}$ contains sequences (or users) from the true majority demographic group. Minimizing $\hat{\mathcal{L}}_{erm}$ under this case could even further amplify the group unfairness. In contrast, if M is too small, $\hat{A}_{under}$ is then too conservative to cover all under-represented sequences, minimizing the loss is not effective either.

To overcome the generalization issue caused by the discrepancy between $\hat{A}_{under}$ and the true under-represented demographic group, we propose to improve the worst-case performance via a distributionally robust optimization (DRO) approach [13]. That is, instead of computing and minimizing the averaged training loss over $\hat{A}_{under}$, we minimize a robust loss to enhance the robustness

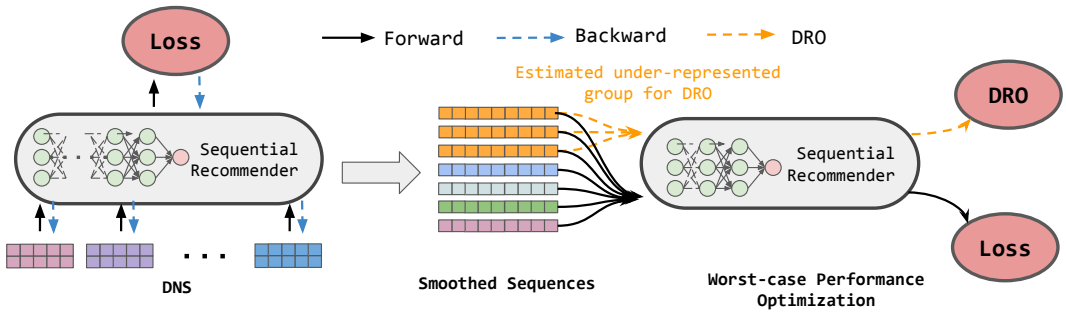

Figure 3: Overview of our proposed Agnostic-FairSeqRec (A-FSR.)

of the performance improvements against such potential distribution shifts. Specifically, we consider an uncertainty set Q, which encodes all possible distributions within a divergence ball center at $\hat{A}_{under}$. Our DRO objective minimizes the loss over the worst-case distribution in Q:

$$\mathcal{L}_{dro} = \sup_{Q \in \mathcal{B}(\hat{A}_{under}, r)} \mathbb{E}_{(x,y)\sim Q}[\mathcal{L}(f(x), y)], \tag{11}$$

where $\mathcal{B}(\hat{A}_{under}, r)$ denotes the divergence ball centered at the empirical distribution of $\hat{A}_{under}$, and r is the radius of the ball. For simplicity, we approximate Equation 11 by adding importance weights q to the sequences in $\hat{A}_{under}$:

$$\begin{aligned} \hat{\mathcal{L}}_{dro} &= \sup_{q \in \hat{\mathcal{B}}} \mathbb{E}_{(x,y)\sim \hat{A}_{under}}[q\mathcal{L}(f(x), y)] \\ &= \sup_{q \in \hat{\mathcal{B}}} \sum_{m=1}^{M} q^{(m)} \cdot \mathcal{L}(f(x^{(m)}), y^{(m)}), \end{aligned} \tag{12}$$

where $\hat{\mathcal{B}}$ is the approximated chi-squared ball for Equation 11. Upon implementation, we use $\hat{\mathcal{B}} = \mathcal{B}(\mathcal{U}(1, M), r) := \{q | D_{\chi^2}([q^{(1)}, ..., q^{(M)}] \| [\frac{1}{M}, ..., \frac{1}{M}]) \leq r, \sum_{m=1}^{M} q^{(m)} = 1\}$. By minimizing Equation 12, the training focus will be implicitly moved towards the potential under-represented users that suffer from the stereotypical patterns. Moreover, the design of Equation 12 also accommodates the discrepancy between the estimated under-represented group and the true under-represented group, so that the performance improvements are robust against potential distribution shifts.

Finally, we highlight that Equation 12 is fundamentally different from directly using DRO for improving model fairness in [13]. The key reason is that in the training stage of sequential recommenders, **different sequences (training samples) might belong to the same user**. This means that directly applying DRO for all training sequences could incompatibly increase and reduce the training penalty on the same user at the same time during training. Instead, we estimate under-represented group first, and then use DRO to overcome the generalization issue caused by the discrepancy of the actual under-represented group and the estimated one. Moreover, the worst-case performance optimization is only used on a small portion of the users (i.e., estimated under-represented groups) instead of the whole population. This mechanism is deliberately designed to avoid the overly-conservative debiasing for all users.

4.4 Overall Framework

Our proposed A-FSR consists of all modules introduced above. The recommender is trained by jointly optimizing the training loss over the dataset $\mathcal{D}$ and the robust loss over the identified under-represented group $\hat{A}_{under}$:

$$\begin{aligned} \mathcal{L}_{A-FSR}(f, \mathcal{D}) &= \mathcal{L}_{rec}(f, \mathcal{D}) + \lambda \cdot \hat{\mathcal{L}}_{dro}(f, \hat{A}_{under}) \\ &= \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \mathcal{L}(f_m(\tilde{z}^{(i)}), y^{(i)}) \\ &\quad + \lambda \cdot \Big[\sup_{q \in \hat{\mathcal{B}}} \sum_{m=1}^{M} q^{(m)} \cdot \mathcal{L}(f_m(\tilde{z}^{(m)}), y^{(m)}) \Big], \\ s.t. \quad \hat{\mathcal{B}} &= \mathcal{B}(\mathcal{U}(1, M), r) \\ \tilde{z} &= [z_1, ..., \tilde{z}_{i^*-s}, ..., \tilde{z}_{i^*}, \tilde{z}_{i^*+s}, ..., z_l] \\ z_i &= f_e(x_i). \end{aligned} \tag{13}$$

Note that in Equation 13, a tunnable trade-off factor λ is introduced to adjust the penalty of the robust loss $\mathcal{L}_{dro}$.$\tilde{z}$ is the item embeddings of x after DNS. Finally, we highlight that A-FSR is different from the DRO-fairness or Maxmin-fairness in [9, 32] in the sense that we debias the model by considering stereotypical patterns in the data and cover all users, whereas in [9, 32], fairness is achieved by merely improving worst-off individuals.

5 EXPERIMENTS

5.1 Experimental Setup

Datasets. We select ML-100K [12], ML-1M [12] and LastFM [3] for experiments. These three datasets contain annotated demographic attributes for us to evaluate the group fairness, but we do **not** use demographic attributes during training. We use gender as the demographic attribute to evaluate the bias and repeat the experiments for 5 times as in [31]. The results are reported with the mean value and the standard deviation.

Baseline Sequential Recommenders. Without loss of generality, we adopt the state-of-the-art NARM [20], SASRec [18] and BERT4Rec [27] as baseline models in our experiments.

Baseline Fair Recommendation Methods. Our setting of no user demographics largely limits our selection of baselines from existing fair recommendation methods. Moreover, we need to select model-agnostic algorithms and exclude baseline methods that require specific model design (e.g., FairSR) for a fair comparison. Therefore, we selected IPW [23], Reg [36] and Adv [31] as baselines.

Model	Dataset		ML-100K	ML-1M	LastFM
	Baselines	topk	$\Phi_R/\Phi_N \downarrow$	$\Phi_R/\Phi_N \downarrow$	$\Phi_R/\Phi_N \downarrow$
BERT4Rec	Naive	@3	$0.0629_{\pm 0.0256}$ / $0.0504_{\pm 0.0194}$	$0.0341_{\pm 0.0051}$ / $0.0404_{\pm 0.0032}$	$0.0644_{\pm 0.0020}$ / $0.0702_{\pm 0.0027}$
		@5	$0.0472_{\pm 0.0167}$ / $0.0439_{\pm 0.0179}$	$0.0275_{\pm 0.0034}$ / $0.0376_{\pm 0.0026}$	$0.0452_{\pm 0.0150}$ / $0.0621_{\pm 0.0079}$
	IPW	@3	$0.0823_{\pm 0.0090}$ / $0.0648_{\pm 0.0073}$	$0.0349_{\pm 0.0053}$ / $0.0396_{\pm 0.0055}$	$0.0529_{\pm 0.0173}$ / $0.0625_{\pm 0.0136}$
		@5	$0.0534_{\pm 0.0132}$ / $0.0519_{\pm 0.0052}$	$0.0243_{\pm 0.0042}$ / $0.0352_{\pm 0.0046}$	$\mathbf{0.0411}_{\pm 0.0071}$ / $0.0577_{\pm 0.0071}$
	Reg	@3	$0.0608_{\pm 0.0352}$ / $0.0440_{\pm 0.0225}$	$0.0504_{\pm 0.0035}$ / $0.0433_{\pm 0.0018}$	$0.0346_{\pm 0.0147}$ / $0.0497_{\pm 0.0144}$
		@5	$0.0519_{\pm 0.0250}$ / $0.0398_{\pm 0.0181}$	$0.0450_{\pm 0.0061}$ / $0.0412_{\pm 0.0023}$	$0.0449_{\pm 0.0063}$ / $0.0541_{\pm 0.0083}$
	Adv	@3	$0.0622_{\pm 0.0220}$ / $0.0532_{\pm 0.0156}$	$\mathbf{0.0223}_{\pm 0.0049}$ / $0.0292_{\pm 0.0052}$	$0.0528_{\pm 0.0218}$ / $0.0682_{\pm 0.0134}$
		@5	$0.0517_{\pm 0.0070}$ / $0.0492_{\pm 0.0098}$	$0.0112_{\pm 0.0037}$ / $0.0248_{\pm 0.0038}$	$0.0413_{\pm 0.0142}$ / $0.0631_{\pm 0.0091}$
	A-FSR (Ours)	@3	$\mathbf{0.0465}_{\pm 0.0238}$ / $\mathbf{0.0357}_{\pm 0.0151}$	$\mathbf{0.0223}_{\pm 0.0097}$ / $\mathbf{0.0290}_{\pm 0.0078}$	$\mathbf{0.0336}_{\pm 0.0137}$ / $\mathbf{0.0342}_{\pm 0.0173}$
		@5	$\mathbf{0.0404}_{\pm 0.0069}$ / $\mathbf{0.0331}_{\pm 0.0072}$	$\mathbf{0.0086}_{\pm 0.0078}$ / $\mathbf{0.0232}_{\pm 0.0059}$	$0.0451_{\pm 0.0009}$ / $\mathbf{0.0392}_{\pm 0.0120}$
SASRec	Naive	@3	$0.1119_{\pm 0.0277}$ / $0.0936_{\pm 0.0178}$	$0.1800_{\pm 0.0107}$ / $0.1636_{\pm 0.0083}$	$0.0731_{\pm 0.0110}$ / $0.0870_{\pm 0.0068}$
		@5	$0.1222_{\pm 0.0235}$ / $0.0977_{\pm 0.0160}$	$0.1748_{\pm 0.0048}$ / $0.1614_{\pm 0.0067}$	$0.0485_{\pm 0.0199}$ / $0.0768_{\pm 0.0110}$
	IPW	@3	$0.0796_{\pm 0.0140}$ / $0.0666_{\pm 0.0051}$	$0.1247_{\pm 0.0030}$ / $0.1125_{\pm 0.0018}$	$0.0745_{\pm 0.0058}$ / $0.0956_{\pm 0.0056}$
		@5	$0.0882_{\pm 0.0206}$ / $0.0702_{\pm 0.0081}$	$0.1142_{\pm 0.0054}$ / $0.1083_{\pm 0.0019}$	$0.0626_{\pm 0.0125}$ / $0.0907_{\pm 0.0115}$
	Reg	@3	$0.0594_{\pm 0.0206}$ / $0.0488_{\pm 0.0230}$	$0.0823_{\pm 0.0090}$ / $0.0774_{\pm 0.0106}$	$0.0860_{\pm 0.0217}$ / $0.0992_{\pm 0.0083}$
		@5	$0.0787_{\pm 0.0247}$ / $0.0564_{\pm 0.0209}$	$0.0697_{\pm 0.0115}$ / $0.0724_{\pm 0.0108}$	$0.0646_{\pm 0.0206}$ / $0.0904_{\pm 0.0105}$
	Adv	@3	$0.0664_{\pm 0.0272}$ / $0.0533_{\pm 0.0228}$	$0.1341_{\pm 0.0190}$ / $0.1241_{\pm 0.0155}$	$\mathbf{0.0523}_{\pm 0.0374}$ / $\mathbf{0.0625}_{\pm 0.0370}$
		@5	$0.0716_{\pm 0.0274}$ / $0.0544_{\pm 0.0211}$	$0.1250_{\pm 0.0197}$ / $0.1204_{\pm 0.0158}$	$0.0375_{\pm 0.0251}$ / $\mathbf{0.0569}_{\pm 0.0318}$
	A-FSR (Ours)	@3	$\mathbf{0.0522}_{\pm 0.0231}$ / $\mathbf{0.0440}_{\pm 0.0184}$	$\mathbf{0.0473}_{\pm 0.0064}$ / $\mathbf{0.0489}_{\pm 0.0059}$	$0.0597_{\pm 0.0210}$ / $0.0799_{\pm 0.0139}$
		@5	$\mathbf{0.0568}_{\pm 0.0207}$ / $\mathbf{0.0464}_{\pm 0.0172}$	$\mathbf{0.0344}_{\pm 0.0055}$ / $\mathbf{0.0435}_{\pm 0.0062}$	$\mathbf{0.0364}_{\pm 0.0086}$ / $0.0703_{\pm 0.0083}$
NARM	Naive	@3	$0.0627_{\pm 0.0058}$ / $0.0522_{\pm 0.0105}$	$0.0289_{\pm 0.0095}$ / $0.0309_{\pm 0.0097}$	$0.0550_{\pm 0.0117}$ / $0.0566_{\pm 0.0122}$
		@5	$0.0460_{\pm 0.0267}$ / $0.0457_{\pm 0.0151}$	$0.0242_{\pm 0.0073}$ / $0.0290_{\pm 0.0077}$	$0.0361_{\pm 0.0361}$ / $0.0490_{\pm 0.0151}$
	IPW	@3	$0.0832_{\pm 0.0169}$ / $0.0690_{\pm 0.0170}$	$0.0381_{\pm 0.0080}$ / $0.0447_{\pm 0.0052}$	$0.0458_{\pm 0.0156}$ / $0.0505_{\pm 0.0121}$
		@5	$0.0699_{\pm 0.0210}$ / $0.0635_{\pm 0.0168}$	$0.0349_{\pm 0.0111}$ / $0.0435_{\pm 0.0055}$	$0.0354_{\pm 0.0039}$ / $0.0462_{\pm 0.0083}$
	Reg	@3	$0.0691_{\pm 0.0251}$ / $0.0625_{\pm 0.0249}$	$0.0360_{\pm 0.0058}$ / $0.0369_{\pm 0.0056}$	$0.0634_{\pm 0.0274}$ / $0.0632_{\pm 0.0210}$
		@5	$0.0610_{\pm 0.0246}$ / $0.0590_{\pm 0.0253}$	$0.0286_{\pm 0.0110}$ / $0.0340_{\pm 0.0070}$	$0.0478_{\pm 0.0227}$ / $0.0567_{\pm 0.0189}$
	Adv	@3	$0.0427_{\pm 0.0172}$ / $0.0407_{\pm 0.0124}$	$0.0206_{\pm 0.0061}$ / $0.0240_{\pm 0.0071}$	$0.0331_{\pm 0.0022}$ / $0.0444_{\pm 0.0016}$
		@5	$0.0333_{\pm 0.0171}$ / $0.0372_{\pm 0.0114}$	$0.0097_{\pm 0.0037}$ / $0.0194_{\pm 0.0057}$	$0.0301_{\pm 0.0069}$ / $0.0427_{\pm 0.0029}$
	A-FSR (Ours)	@3	$\mathbf{0.0359}_{\pm 0.0234}$ / $\mathbf{0.0249}_{\pm 0.0164}$	$\mathbf{0.0032}_{\pm 0.0008}$ / $\mathbf{0.0026}_{\pm 0.0016}$	$\mathbf{0.0296}_{\pm 0.0093}$ / $\mathbf{0.0367}_{\pm 0.0131}$
		@5	$\mathbf{0.0154}_{\pm 0.0125}$ / $\mathbf{0.0166}_{\pm 0.0130}$	$\mathbf{0.0073}_{\pm 0.0029}$ / $\mathbf{0.0041}_{\pm 0.0019}$	$\mathbf{0.0217}_{\pm 0.0117}$ / $\mathbf{0.0338}_{\pm 0.0138}$

Table 2: Evaluation of Group Fairness. Each row represents a fair recommendation method, including baseline methods and A-FSR. Each column stands for one dataset. The performance gap in terms of Recall (Φ_R) and NDCG (Φ_N) are reported in the same column split by /. The best results are highlighted in bold, and the second-best results are highlighted with underlines.

Note that these baselines originally require demographic attributes to debias the model. Thus, upon implementation, random gender labels were assigned to them based on the distribution of the datasets. We also add a naive baseline, where no-debiasing method is applied when training the recommenders.

Evaluation Metrics. Following [18, 38, 39], we perform leave-last-out evaluation. In terms of the evaluation metrics, we use normalized discounted cumulative gain (*NDCG*) and *Recall*. To evaluate group fairness, we compute the performance gap between different demographic groups using Equation 1 by plugging in R with either NDCG (Φ_N) or Recall (Φ_R).

Implementation Details. All models are trained without warmup using an Adam optimizer with a learning rate of 0.001, weight decay 0.01 and batch size of 64. Following [18, 27, 38], we set the maximum sequence lengths of ML-1M to be 200 and 50 for the other two datasets. For our method, we empirically set the span size of the identified stereotypical pattern as 3. When performing DNS, we empirically pick the top-6 closest neighbors of each item in the located stereotypical pattern. Finally, regarding DRO, the size of the under-represented group is 5, and we keep the radius of the divergence ball the same as λ for the sake of simplicity[4].

5.2 Performance Evaluation (Table 2 and 3)

The first set of experiments is conducted to evaluate the efficacy of A-FSR. Table 2 reports the results of model fairness, and Table 3 reports the results of model performance. According to Table 2, we observe that A-FSR outperforms the baselines on all datasets in

[4]During our experiments, we observed that A-FSR is sensitive to λ, but less sensitive to the number of neighbors and the size of the estimated under-represented group. Due to space limit, we could not include these results in this submission.

Model	Dataset		ML-100K	ML-1M	LastFM
	Baselines	topk	*Recall* / *NDCG* ↑	*Recall* / *NDCG* ↑	*Recall* / *NDCG* ↑
BERT4Rec	Naive	@3	$0.2567_{\pm 0.0058}$ / $0.1929_{\pm 0.0021}$	$0.5691_{\pm 0.0032}$ / $0.4771_{\pm 0.0038}$	$0.6032_{\pm 0.0066}$ / $0.5408_{\pm 0.0046}$
		@5	$0.3657_{\pm 0.0110}$ / $0.2377_{\pm 0.0044}$	$0.6585_{\pm 0.0037}$ / $0.5140_{\pm 0.0038}$	$0.6650_{\pm 0.0020}$ / $0.5663_{\pm 0.0027}$
	IPW	@3	$0.2492_{\pm 0.0070}$ / $0.1888_{\pm 0.0051}$	$0.5121_{\pm 0.0050}$ / $0.4221_{\pm 0.0049}$	$0.5929_{\pm 0.0108}$ / $0.5313_{\pm 0.0094}$
		@5	$0.3637_{\pm 0.0148}$ / $0.2358_{\pm 0.0054}$	$0.6094_{\pm 0.0048}$ / $0.4622_{\pm 0.0048}$	$0.6654_{\pm 0.0085}$ / $0.5614_{\pm 0.0075}$
	Reg	@3	$0.2347_{\pm 0.0163}$ / $0.1760_{\pm 0.0127}$	$0.3923_{\pm 0.0172}$ / $0.3091_{\pm 0.0170}$	$0.5900_{\pm 0.0010}$ / $0.5236_{\pm 0.0032}$
		@5	$0.3357_{\pm 0.0158}$ / $0.2175_{\pm 0.0125}$	$0.4990_{\pm 0.0151}$ / $0.3530_{\pm 0.0161}$	$0.6536_{\pm 0.0051}$ / $0.5497_{\pm 0.0016}$
	Adv	@3	$0.2535_{\pm 0.0107}$ / $0.1908_{\pm 0.0072}$	$0.5690_{\pm 0.0057}$ / $0.4764_{\pm 0.0051}$	$0.6157_{\pm 0.0076}$ / $0.5510_{\pm 0.0028}$
		@5	$0.3696_{\pm 0.0103}$ / $0.2386_{\pm 0.0070}$	$0.6587_{\pm 0.0042}$ / $0.5133_{\pm 0.0045}$	$0.6747_{\pm 0.0065}$ / $0.5755_{\pm 0.0019}$
	A-FSR (Ours)	@3	$0.2552_{\pm 0.0092}$ / $0.1916_{\pm 0.0078}$	$0.5657_{\pm 0.0080}$ / $0.4730_{\pm 0.0069}$	$0.6005_{\pm 0.0072}$ / $0.5332_{\pm 0.0065}$
		@5	$0.3741_{\pm 0.0128}$ / $0.2403_{\pm 0.0096}$	$0.6564_{\pm 0.0074}$ / $0.5104_{\pm 0.0066}$	$0.6714_{\pm 0.0040}$ / $0.5624_{\pm 0.0058}$
SASRec	Naive	@3	$0.2545_{\pm 0.0084}$ / $0.1932_{\pm 0.0095}$	$0.5219_{\pm 0.0039}$ / $0.4312_{\pm 0.0021}$	$0.6205_{\pm 0.0082}$ / $0.5469_{\pm 0.0069}$
		@5	$0.3579_{\pm 0.0136}$ / $0.2357_{\pm 0.0116}$	$0.6216_{\pm 0.0015}$ / $0.4723_{\pm 0.0018}$	$0.6802_{\pm 0.0045}$ / $0.5715_{\pm 0.0056}$
	IPW	@3	$0.2254_{\pm 0.0147}$ / $0.1671_{\pm 0.0112}$	$0.5063_{\pm 0.0021}$ / $0.4086_{\pm 0.0020}$	$0.6139_{\pm 0.0088}$ / $0.5451_{\pm 0.0053}$
		@5	$0.3316_{\pm 0.0111}$ / $0.2104_{\pm 0.0095}$	$0.6162_{\pm 0.0060}$ / $0.4538_{\pm 0.0015}$	$0.6851_{\pm 0.0052}$ / $0.5745_{\pm 0.0040}$
	Reg	@3	$0.2219_{\pm 0.0116}$ / $0.1633_{\pm 0.0078}$	$0.4787_{\pm 0.0031}$ / $0.3812_{\pm 0.0026}$	$0.6177_{\pm 0.0022}$ / $0.5510_{\pm 0.0034}$
		@5	$0.3313_{\pm 0.0126}$ / $0.2081_{\pm 0.0075}$	$0.5955_{\pm 0.0050}$ / $0.04293_{\pm 0.0030}$	$0.6836_{\pm 0.0026}$ / $0.5782_{\pm 0.0052}$
	Adv	@3	$0.2239_{\pm 0.0090}$ / $0.1679_{\pm 0.0045}$	$0.5126_{\pm 0.0041}$ / $0.4168_{\pm 0.0059}$	$0.5675_{\pm 0.0340}$ / $0.4923_{\pm 0.0339}$
		@5	$0.3302_{\pm 0.0067}$ / $0.2115_{\pm 0.0030}$	$0.6205_{\pm 0.0047}$ / $0.4612_{\pm 0.0049}$	$0.6439_{\pm 0.0206}$ / $0.5228_{\pm 0.0284}$
	A-FSR (Ours)	@3	$0.2204_{\pm 0.0161}$ / $0.1632_{\pm 0.0123}$	$0.5127_{\pm 0.0061}$ / $0.4157_{\pm 0.0043}$	$0.6231_{\pm 0.0071}$ / $0.5547_{\pm 0.0029}$
		@5	$0.3242_{\pm 0.0205}$ / $0.2057_{\pm 0.0145}$	$0.6206_{\pm 0.0054}$ / $0.4602_{\pm 0.0040}$	$0.6853_{\pm 0.0048}$ / $0.5803_{\pm 0.0020}$
NARM	Naive	@3	$0.3662_{\pm 0.0094}$ / $0.2873_{\pm 0.0058}$	$0.6180_{\pm 0.0032}$ / $0.5313_{\pm 0.0026}$	$0.5826_{\pm 0.0068}$ / $0.5186_{\pm 0.0055}$
		@5	$0.4832_{\pm 0.0054}$ / $0.3354_{\pm 0.0037}$	$0.7032_{\pm 0.0031}$ / $0.5663_{\pm 0.0016}$	$0.6474_{\pm 0.0134}$ / $0.5452_{\pm 0.0086}$
	IPW	@3	$0.3560_{\pm 0.0077}$ / $0.2768_{\pm 0.0045}$	$0.6112_{\pm 0.0025}$ / $0.5247_{\pm 0.0019}$	$0.5715_{\pm 0.0114}$ / $0.5082_{\pm 0.0080}$
		@5	$0.4772_{\pm 0.0104}$ / $0.3265_{\pm 0.0066}$	$0.6943_{\pm 0.0026}$ / $0.5589_{\pm 0.0016}$	$0.6418_{\pm 0.0084}$ / $0.5369_{\pm 0.0065}$
	Reg	@3	$0.3448_{\pm 0.0146}$ / $0.2670_{\pm 0.0090}$	$0.6153_{\pm 0.0021}$ / $0.5255_{\pm 0.0008}$	$0.5851_{\pm 0.0036}$ / $0.5121_{\pm 0.0072}$
		@5	$0.4624_{\pm 0.0070}$ / $0.3153_{\pm 0.0062}$	$0.6978_{\pm 0.0019}$ / $0.5595_{\pm 0.0010}$	$0.6490_{\pm 0.0056}$ / $0.5385_{\pm 0.0075}$
	Adv	@3	$0.3585_{\pm 0.0041}$ / $0.2813_{\pm 0.0019}$	$0.6184_{\pm 0.0042}$ / $0.5321_{\pm 0.0045}$	$0.5814_{\pm 0.0147}$ / $0.5122_{\pm 0.0120}$
		@5	$0.4699_{\pm 0.0056}$ / $0.3270_{\pm 0.0040}$	$0.7025_{\pm 0.0033}$ / $0.5667_{\pm 0.0042}$	$0.6483_{\pm 0.0096}$ / $0.5399_{\pm 0.0100}$
	A-FSR (Ours)	@3	$0.3506_{\pm 0.0057}$ / $0.2758_{\pm 0.0068}$	$0.6238_{\pm 0.0021}$ / $0.5357_{\pm 0.0011}$	$0.6092_{\pm 0.0092}$ / $0.5407_{\pm 0.0072}$
		@5	$0.4711_{\pm 0.0090}$ / $0.3253_{\pm 0.0080}$	$0.7099_{\pm 0.0027}$ / $0.5711_{\pm 0.0012}$	$0.6657_{\pm 0.0180}$ / $0.5561_{\pm 0.0110}$

Table 3: Evaluation of Recommendation Performance. A-FSR maintains a satisfactory performance compared to other baselines.

terms of improving group fairness. In the most biased setting, where the un-debiased SASRec achieves 0.1800 on Φ_R, A-FSR reduces its bias by 73.7% and achieves 0.0473. Moreover on both transformer-based models (i.e., BERT4Rec and SASRec) and RNN-based models (NARM), A-FSR is consistently effective, indicating that A-FSR could be applied to different kinds of sequential recommendation systems. In addition to fairness, we also report the overall recommendation performance of all methods to verify that A-FSR is not a trivial fair solution that reduces performance for all demographic groups. It is observed from Table 3 that A-FSR can maintain similar overall recommendation performance compared to the baseline methods.

5.3 Universal Fairness (Table 4)

Recall A-FSR does not use the user demographic attributes to debias the model. As such, in addition to the gender fairness, we expect that A-FSR could also demonstrate a universal fairness improvements w.r.t. other demographic attributes. To test the universal fairness, we evaluate the trained models w.r.t. another demographic attribute of the users: the *occupations*. That is, we evaluate the fairness for the occupation-based groups. The results are reported in Table 4. As expected, A-FSR also improves the group fairness w.r.t. users' occupations. We only compare A-FSR against the naive baseline, because other fair recommendation baseline methods could not automatically achieve the universal fairness w.r.t. different demographic attributes. Instead, the baseline methods require to re-train the models with specific loss or regularization terms for the new demographic attributes. Due to space limit, we could only report results for BERT4Rec and SASRec on ML-100K and ML-1M.

5.4 Ablation Study (Table 5)

Finally, we perform an ablation study to understand the contribution of each component to A-FSR. In particular, we still train and test the sequence recommenders using A-FSR, but we mask out each component individually: (1) A-FSR without DNS, where we

Model	Dataset	Metrics	Naive	**A-FSR (Ours)**
BERT4Rec	ML-100K	$\Phi_R@3/\Phi_N@3\downarrow$	$0.3310_{\pm 0.0531}$ / $0.2561_{\pm 0.0333}$	$\mathbf{0.2105}_{\pm 0.0431}$ / $\mathbf{0.1062}_{\pm 0.0219}$
		$\Phi_R@5/\Phi_N@5\downarrow$	$0.3952_{\pm 0.0342}$ / $0.2827_{\pm 0.0202}$	$\mathbf{0.3119}_{\pm 0.0487}$ / $\mathbf{0.1851}_{\pm 0.0243}$
	ML-1M	$\Phi_R@3/\Phi_N@3\downarrow$	$0.3673_{\pm 0.0234}$ / $0.3254_{\pm 0.0168}$	$\mathbf{0.2284}_{\pm 0.0705}$ / $\mathbf{0.1839}_{\pm 0.0626}$
		$\Phi_R@5/\Phi_N@5\downarrow$	$0.2859_{\pm 0.0248}$ / $0.2927_{\pm 0.0111}$	$\mathbf{0.1984}_{\pm 0.0669}$ / $\mathbf{0.1709}_{\pm 0.0574}$
SASRec	ML-100K	$\Phi_R@3/\Phi_N@3\downarrow$	$0.2792_{\pm 0.0815}$ / $0.2460_{\pm 0.0558}$	$\mathbf{0.2056}_{\pm 0.0466}$ / $\mathbf{0.1479}_{\pm 0.0350}$
		$\Phi_R@5/\Phi_N@5\downarrow$	$0.3528_{\pm 0.0626}$ / $0.2812_{\pm 0.0514}$	$\mathbf{0.2110}_{\pm 0.0379}$ / $\mathbf{0.1654}_{\pm 0.0388}$
	ML-1M	$\Phi_R@3/\Phi_N@3\downarrow$	$0.3251_{\pm 0.0346}$ / $0.3124_{\pm 0.0261}$	$\mathbf{0.2165}_{\pm 0.0524}$ / $\mathbf{0.1759}_{\pm 0.0431}$
		$\Phi_R@5/\Phi_N@5\downarrow$	$0.2364_{\pm 0.0307}$ / $0.2771_{\pm 0.0243}$	$\mathbf{0.2252}_{\pm 0.0378}$ / $\mathbf{0.2264}_{\pm 0.0372}$

Table 4: Universal Fairness w.r.t. occupations of the users. The best fairness results are highlighted in bold.

Model	Dataset	Metrics	**A-FSR (ours)**	w/o DNS	w/o DRO
BERT4Rec	ML-100K	$\Phi_R@5/\Phi_N@5\downarrow$	$\mathbf{0.0404}_{\pm 0.0068}$ / $\mathbf{0.0331}_{\pm 0.0072}$	$\underline{0.0476_{\pm 0.0063}}$ / $0.0474_{\pm 0.0077}$	$0.0482_{\pm 0.0326}$ / $\underline{0.0437_{\pm 0.0201}}$
		Recall@5/NDCG@5 ↑	$0.3741_{\pm 0.0128}$ / $0.2403_{\pm 0.0096}$	$0.3769_{\pm 0.0058}$ / $0.2442_{\pm 0.0063}$	$0.3881_{\pm 0.0045}$ / $0.2510_{\pm 0.0041}$
	ML-1M	$\Phi_R@5/\Phi_N@5\downarrow$	$\mathbf{0.0086}_{\pm 0.0078}$ / $\mathbf{0.0232}_{\pm 0.0059}$	$\underline{0.0163_{\pm 0.0058}}$ / $0.0277_{\pm 0.0049}$	$0.0172_{\pm 0.0063}$ / $\underline{0.0270_{\pm 0.0038}}$
		Recall@5/NDCG@5 ↑	$0.6564_{\pm 0.0074}$ / $0.5104_{\pm 0.0066}$	$0.6375_{\pm 0.0054}$ / $0.4922_{\pm 0.0026}$	$0.6160_{\pm 0.0036}$ / $0.4686_{\pm 0.0035}$
SASRec	ML-100K	$\Phi_R@5/\Phi_N@5\downarrow$	$\mathbf{0.0568}_{\pm 0.0207}$ / $\mathbf{0.0464}_{\pm 0.0172}$	$0.0957_{\pm 0.0197}$ / $0.0792_{\pm 0.0097}$	$\underline{0.0829_{\pm 0.0276}}$ / $\underline{0.0613_{\pm 0.0237}}$
		Recall@5/NDCG@5 ↑	$0.3242_{\pm 0.0205}$ / $0.2115_{\pm 0.0030}$	$0.3299_{\pm 0.0083}$ / $0.2129_{\pm 0.0046}$	$0.3339_{\pm 0.0123}$ / $0.2160_{\pm 0.0081}$
	ML-1M	$\Phi_R@5/\Phi_N@5\downarrow$	$\mathbf{0.0344}_{\pm 0.0055}$ / $\underline{0.0435_{\pm 0.0062}}$	$0.1220_{\pm 0.0041}$ / $0.1157_{\pm 0.0048}$	$\underline{0.0376_{\pm 0.0079}}$ / $\mathbf{0.0428}_{\pm 0.0053}$
		Recall@5/NDCG@5 ↑	$0.6206_{\pm 0.0054}$ / $0.4602_{\pm 0.0040}$	$0.6126_{\pm 0.0044}$ / $0.4556_{\pm 0.0038}$	$0.6162_{\pm 0.0048}$ / $0.4583_{\pm 0.0030}$

Table 5: Ablation Study: masking out the Dirichlet neighbor smoothing (DNS) and the worst-case performance optimization (DRO). The best fairness results are highlighted in bold and the second-best results are highlighted with underlines.

Size (2s+1)	ML-100K	ML-1M
	$\Phi_R@3\downarrow / Recall@3\uparrow$	$\Phi_R@3\downarrow / Recall@3\uparrow$
3	$0.0465_{\pm 0.0238}$ / $0.2552_{\pm 0.0092}$	$0.0223_{\pm 0.0097}$ / $0.5657_{\pm 0.0080}$
5	$0.0499_{\pm 0.0229}$ / $0.2665_{\pm 0.0133}$	$0.0283_{\pm 0.0029}$ / $0.5112_{\pm 0.0056}$
7	$0.0577_{\pm 0.0245}$ / $0.2689_{\pm 0.0105}$	$0.0354_{\pm 0.0076}$ / $0.5139_{\pm 0.0029}$
9	$0.0278_{\pm 0.0207}$ / $0.2634_{\pm 0.0067}$	$0.0444_{\pm 0.0120}$ / $0.3999_{\pm 0.0107}$
11	$0.0471_{\pm 0.0257}$ / $0.2710_{\pm 0.0081}$	$0.0431_{\pm 0.0086}$ / $0.3932_{\pm 0.0123}$

Table 6: Sensitivity Analysis for BERT4Rec: increasing the span size of the identified stereotypical pattern.

directly compute Equation 13 over the original item embeddings of the input sequence; (2) A-FSR without worst-case performance optimization, where we only perform Dirichlet Neighbor Smoothing and set $\lambda = 0$. The results are reported in Table 5. Due to the space limit, we could only report results for BERT4Rec and SASRec on ML-100K and ML-1M. From Table 5, we observe that both modules are necessary to improve the fairness of the model.

5.5 Sensitivity Analysis (Table 6)

We then perform sensitive analysis w.r.t. the key hyperparameter of A-FSR, i.e., the span size of the stereotypical pattern (i.e., s in Equation 4). In particular, we increase the span size of stereotypical pattern from 3 to 11. The results are reported in Table 6. It is observed that A-FSR shows different behaviors on different datasets. For example, on ML-100K, A-FSR is less sensitive to size of the smoothed sub-sequence, whereas on ML-1M, A-FSR shows a larger sensitivity. On ML-1M, if the size of the smoothed subsequence is too large, then both fairness and recommendation performance degrades. This observation is expected because ML-100K is smaller and simpler than ML-1M. It is more challenging for the recommender to learn user dynamics on ML-1M, and the larger randomness in the smoothing step will further blur the learned user dynamics for the recommender. Due to space limit, we could only report results for BERT4Rec.

6 CONCLUSION

In this work, we study the problem of developing fair sequential recommenders without user demographics. While much existing fair recommendation literature leverages user demographics to develop fair solutions, we highlight the essence of developing fair recommender methods without user demographics: it could be infeasible to collect user demographic information due to privacy concerns or legal regulations. To address this problem, we designed a novel demographic-agnostic and model-agnostic debiasing framework Agnostic FairSeqRec (A-FSR). Our experimental results demonstrate that on multiple model architectures and multiple datasets A-FSR consistently outperforms state-of-the-art baseline methods by a significant margin.

ACKNOWLEDGEMENT

This research is supported in part by the National Science Foundation under Grant No. IIS-2202481, CHE-2105032, IIS-2130263, CNS-2131622, CNS-2140999. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2212–2220.

[2] Alexander Brown, Nenad Tomasev, Jan Freyberg, Yuan Liu, Alan Karthikesalingam, and Jessica Schrouff. 2023. Detecting shortcut learning for fair medical AI using shortcut testing. *Nature Communications* 14, 1 (2023), 4314.

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

[4] Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. 2021. Sequential recommendation with graph neural networks. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 378–387.

[5] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. 2021. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence* 3, 7 (2021), 620–631.

[6] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph neural networks for social recommendation. In *The world wide web conference*. 417–426.

[7] Wenqi Fan, Xiangyu Zhao, Xiao Chen, Jingran Su, Jingtong Gao, Lin Wang, Qidong Liu, Yiqi Wang, Han Xu, Lei Chen, et al. 2022. A Comprehensive Survey on Trustworthy Recommender Systems. *arXiv preprint arXiv:2209.10117* (2022).

[8] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, et al. 2020. Fairness-aware explainable recommendation over knowledge graphs. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 69–78.

[9] David García-Soriano and Francesco Bonchi. 2021. Maxmin-fair ranking: individual fairness under group-fairness constraints. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 436–446.

[10] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*. 299–315.

[11] Monica Grosso, Sandro Castaldo, Hua Ariel Li, and Bart Larivière. 2020. What information do shoppers share? The effect of personnel-, retailer-, and country-trust on willingness to share information. *Journal of Retailing* 96, 4 (2020), 524–547.

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

[13] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*. PMLR, 1929–1938.

[14] Zhankui He, Handong Zhao, Zhe Lin, Zhaowen Wang, Ajinkya Kale, and Julian Mcauley. 2021. Locker: Locally Constrained Self-Attentive Sequential Recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3088–3092.

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

[16] Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022. Towards Universal Sequence Representation Learning for Recommender Systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 585–593.

[17] Kalervo Järvelin and Jaana Kekäläinen. 2017. IR evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*. ACM New York, NY, USA, 243–250.

[18] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 197–206.

[19] Cheng-Te Li, Cheng Hsu, and Yang Zhang. 2022. Fairsr: Fairness-aware sequential recommendation through multi-task learning with preference graph embeddings. *ACM Transactions on Intelligent Systems and Technology (TIST)* 13, 1 (2022), 1–21.

[20] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1419–1428.

[21] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented fairness in recommendation. In *Proceedings of the Web Conference 2021*. 624–632.

[22] Yunqi Li, Yingqiang Ge, and Yongfeng Zhang. 2021. Tutorial on fairness of machine learning in recommender systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 2654–2657.

[23] Dawen Liang, Laurent Charlin, and David M Blei. 2016. Causal inference for recommendation. In *Causation: Foundation to Application, Workshop at UAI. AUAI*.

[24] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender systems and their ethical challenges. *Ai & Society* 35 (2020), 957–967.

[25] Laura Schelenz. 2021. Diversity-aware recommendations for social justice? exploring user diversity and fairness in recommender systems. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 404–410.

[26] Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. Implicit user modeling for personalized search. In *Proceedings of the 14th ACM international conference on Information and knowledge management*. 824–831.

[27] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1441–1450.

[28] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.

[29] Jiayin Wang, Weizhi Ma, Chumeng Jiang, Min Zhang, Yuan Zhang, Biao Li, and Peng Jiang. 2023. Measuring Item Global Residual Value for Fair Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 269–278.

[30] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems* 41, 3 (2023), 1–43.

[31] Tianxin Wei and Jingrui He. 2022. Comprehensive fair meta-learned recommender system. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1989–1999.

[32] Hongyi Wen, Xinyang Yi, Tiansheng Yao, Jiaxi Tang, Lichan Hong, and Ed H Chi. 2022. Distributionally-robust Recommendations for Improving Worst-case User Experience. In *Proceedings of the ACM Web Conference 2022*. 3606–3610.

[33] Chuhan Wu, Fangzhao Wu, Tao Qi, Jianxun Lian, Yongfeng Huang, and Xing Xie. 2020. PTUM: Pre-training User Model from Unlabeled User Behaviors via Self-supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1939–1944.

[34] Le Wu, Lei Chen, Pengyang Shao, Richang Hong, Xiting Wang, and Meng Wang. 2021. Learning fair representations for recommendation: A graph-based perspective. In *Proceedings of the Web Conference 2021*. 2198–2208.

[35] Mengyue Yang, Jun Wang, and Jean-Francois Ton. 2023. Rectifying unfairness in recommendation feedback loop. In *Proceedings of the 46th international ACM SIGIR Conference on Research and Development in Information Retrieval*. 28–37.

[36] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. *Advances in neural information processing systems* 30 (2017).

[37] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. 2018. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 974–983.

[38] Zhenrui Yue, Zhankui He, Huimin Zeng, and Julian McAuley. 2021. Black-Box Attacks on Sequential Recommenders via Data-Free Model Extraction. In *Fifteenth ACM Conference on Recommender Systems*. 44–54.

[39] Zhenrui Yue, Huimin Zeng, Ziyi Kou, Lanyu Shang, and Dong Wang. 2022. Defending Substitution-Based Profile Pollution Attacks on Sequential Recommenders. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 59–70.