Learning Fair Representations for Recommendation via Information Bottleneck Principle

Junsong Xie[1†] , **Yonghui Yang**[1†] , **Zihan Wang**[1] and **Le Wu**[1,2*]
[1]Hefei University of Technology
[2]Institute of Dataspace, Hefei Comprehensive National Science Center
{jsxie.hfut, yyh.hfut, zhwang.hfut, lewu.ustc}@gmail.com

Abstract

User-oriented recommender systems (RS) characterize users' preferences based on observed behaviors and are widely deployed in personalized services. However, RS may unintentionally capture biases related to sensitive attributes (e.g., gender) from behavioral data, leading to unfair issues and discrimination against particular groups (e.g., females). Adversarial training is a popular technique for fairness-aware RS, when filtering sensitive information in user modeling. Despite advancements in fairness, achieving a good accuracy-fairness trade-off remains a challenge in adversarial training. In this paper, we investigate fair representation learning from a novel information theory perspective. Specifically, we propose a model-agnostic _**Fair**_ recommendation method via the _**I**_nformation _**B**_ottleneck principle (*FairIB*). The learning objective of *FairIB* is to maximize the mutual information between user representations and observed interactions, while simultaneously minimizing it between user representations and sensitive attributes. This approach facilitates the capturing of essential collaborative signals in user representations while mitigating the inclusion of unnecessary sensitive information. Empirical studies on two real-world datasets demonstrate the effectiveness of the proposed FairIB, which significantly improves fairness while maintaining competitive recommendation accuracy, either in single or multiple sensitive scenarios. The code is available at https://github.com/jsxie9/IJCAI_FairIB.

1 Introduction

In the modern era, recommender systems have become essential allies, seamlessly delivering personalized content across diverse domains [Covington *et al.*, 2016; Chen *et al.*, 2020]. Fueled by sophisticated algorithms and user-centric models [Tan *et al.*, 2021; Yang *et al.*, 2023; Wu *et al.*, 2023], these systems shine in anticipating user preferences, elevating overall user experience and engagement. While recommender systems achieve remarkable success in tailoring recommendations to individual users, they frequently grapple with fairness challenges stemming from biased historical interactions, particularly concerning sensitive attributes such as gender and race [Ekstrand *et al.*, 2018b; Shao *et al.*, 2022]. For example, news recommender systems may disproportionately recommend certain political ideologies over others, which may manipulate user's opinions [Li *et al.*, 2023]. Job recommender systems may display racial or gender discrimination by disproportionately suggesting low-paying jobs to protected user groups [Dong *et al.*, 2023].

Due to the progressive advancement of trustworthy AI, fairness-aware recommendations are capturing the increasing attention of researchers [Li *et al.*, 2023]. Learning fair recommendation representations from user-item interactions is one of the promising methods [Bose and Hamilton, 2019; Xie *et al.*, 2017]. To achieve the goal of fair representations without sensitive information, researchers have proposed various methods [Yao and Huang, 2017; Zhu *et al.*, 2018; Wu *et al.*, 2021a; Wu *et al.*, 2021b]. Among them, adversarial training is the mainstream technique for recommendation fairness, and a series of adversarial-based works have been proposed [Bose and Hamilton, 2019; Dai and Wang, 2021; Wu *et al.*, 2021a]. While learning user representations from observed interactions, adversarial-based methods encourage that sensitive attributes cannot be inferred from the representations. These methods operate within a two-player minimax game framework, effectively safeguarding against the disclosure of such attributes within the learned representations. Despite advancements in fairness, achieving an efficient accuracy-fairness trade-off remains a challenge in adversarial training. Due to the instability of adversarial training [Salimans *et al.*, 2016; Arjovsky *et al.*, 2017], a commonly employed solution is to pre-train recommendation representations based on observed interactions and subsequently optimize sensitive attribute filters through adversarial training [Wu *et al.*, 2021a]. However, this solution is hard to balance accuracy-fairness trade-off, leading to unsatisfied recommendations.

In this paper, we revisit the fairness-aware recommendation from an information theory perspective, and propose a novel *FairIB* via the information bottleneck principle. As shown in Figure 1, we illustrate the optimization objectives of the traditional recommendation and our propose *FairIB*.

[*] Corresponding author.

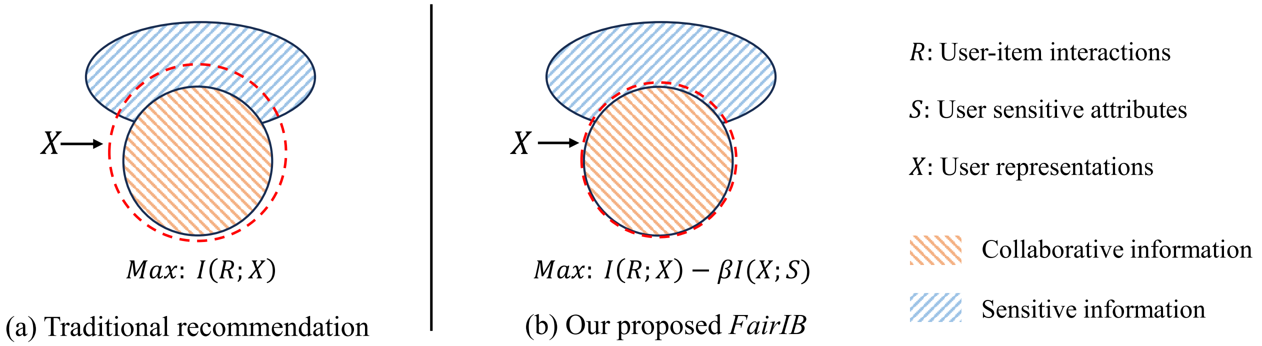

X
$Max: I(R;X)$
(a) Traditional recommendation
X
$Max: I(R;X) - \beta I(X;S)$
(b) Our proposed *FairIB*
R: User-item interactions
S: User sensitive attributes
X: User representations
Collaborative information
Sensitive information

Figure 1: Information diagram of user representations $\mathbf{X}$, user sensitive attributes $\mathbf{S}$, and user-item interactions $\mathbf{R}$. (a) Traditional recommendation methods focus on maximizing the mutual information between representations $\mathbf{X}$ and user-item interactions $\mathbf{R}$. As users' behavior exposes their sensitive attributes, maximizing $I(\mathbf{R};\mathbf{X})$ will lead to representations involving sensitive information. (b) Our proposed *FairIB* via Information Bottleneck principle. It not only maximizes $I(\mathbf{R};\mathbf{X})$, meanwhile minimizing the mutual information between representations $\mathbf{X}$ and sensitive attributes $\mathbf{S}$. Therefore, *FairIB* acquires fair representations $\mathbf{X}$ through IB learning, successfully achieving an efficient trade-off of recommendation accuracy and fairness.

Traditional recommender systems focus on the accuracy task, the optimal user representations $\mathbf{X}$ are obtained by maximizing the mutual information between $\mathbf{X}$ and user-item observed interactions $\mathbf{R}$ ($I(\mathbf{R};\mathbf{X})$). However, empirical studies indicate that users' behaviors can expose their sensitive attributes [Kosinski *et al.*, 2013; Wu *et al.*, 2021a], simply maximizing $I(\mathbf{R};\mathbf{X})$ will introduce sensitive information to representations, leading to unfair recommendation results. To tackle this issue, our proposed *FairIB* acquires fair representations $\mathbf{X}$ based on information bottleneck learning. Besides maximizing the mutual information $I(\mathbf{R};\mathbf{X})$, *FairIB* simultaneously minimizes the mutual information $I(\mathbf{X};\mathbf{S})$. Thus, *FairIB* acquires fair yet precise representations in an IB manner, successfully achieving an efficient trade-off of recommendation accuracy and fairness.

However, directly optimizing the learning objective of $Max : I(\mathbf{R};\mathbf{X}) - \beta I(\mathbf{X};\mathbf{S})$ poses three challenges. First, the calculation of $I(\mathbf{R};\mathbf{X})$ is difficult. Each user representation should match all interacted items, as they interact with different items. Second, it's hard to minimize $I(\mathbf{X};\mathbf{S})$ because estimating the upper bound of mutual information is an intractable problem. While current methods use variational techniques for estimation, their heavy reliance on prior assumptions (e.g., Gaussian marginal distribution) can result in the upper bound failing to accurately estimate mutual information with low bias in practical applications [Alemi *et al.*, 2017; Cheng *et al.*, 2020]. Third, users' sensitive attributes are not only exposed in their representations, but also in subgraph structure [Wu *et al.*, 2021a], so it's necessary to filter sensitive information in user-centric sub-graphs.

To address the above challenges, we elaborately implement FairIB as follows. For maximization of $I(\mathbf{R};\mathbf{X})$, we establish that the goal is equivalent to maximizing the inner product between a user and her interacted items representations. This aligns with the typical optimization objective of general recommendation systems. For minimization of $I(\mathbf{X};\mathbf{S})$, we utilize the Hilbert-Schmidt Independence Criterion (HSIC) regularizer [Ma *et al.*, 2020; Wang *et al.*, 2023] to approximate the optimization, without prior assumptions. Besides, we extend the HSIC regularizer to user-centric sub-graphs, further eliminating the sensitive information. In summary, our main contributions are listed as follows:

- In this paper, we revisit fairness-aware recommendations from an information theory lens, and propose a novel model-agnostic fair representation learning method *FairIB*.
- Technically, *FairIB* introduces the HSIC-based bottleneck to improve recommendation fairness, effectively eliminating the sensitive information from user and subgraph perspectives.
- Extensive experiments demonstrate the effectiveness of our proposed *FairIB*, which can effectively achieve recommendation accuracy-fairness trade-off.

2 Related Work

2.1 Fairness in Recommendation

Recommendation is a widely deployed user-centric application [Gao *et al.*, 2023], and an increasing number of researchers are focusing on the issue of fairness [Xiao *et al.*, 2017; Li *et al.*, 2021; Shao *et al.*, 2024]. In recommendation systems, there is the potential for unequal treatment of sensitive user groups [Ekstrand *et al.*, 2018a]. For instance, an unfair job recommendation system may exhibit preferences towards user groups with specific sensitive attributes, inevitably resulting in biased recommendations [Dong *et al.*, 2023]. Many fairness definitions are proposed to measure potential unfairness [Mehrabi *et al.*, 2021; Biega *et al.*, 2018; Wu *et al.*, 2021b]. For example, individual fairness requires that a model produces similar decisions for similar individuals [Biega *et al.*, 2018]. Group fairness advocates against discrimination of a specific user group based on sensitive attributes [Wu *et al.*, 2021b]. In this work, our emphasis is on group fairness because it can quantitatively measure the differences in how sensitive groups are treated. [Hardt *et al.*, 2016]. In response to fairness concerns, researchers have designed numerous models [Yao and Huang, 2017;

Wu *et al.*, 2021a; Dai and Wang, 2021; Wu *et al.*, 2021b]. [Yao and Huang, 2017] proposes five fairness metrics as regularization terms to measure the discrepancy between the prediction behavior for the female and male groups. Based on adversarial learning, FairRec [Wu *et al.*, 2021a] decomposes adversarial learning and orthogonality regularization. FairGo [Wu *et al.*, 2021b] further develops a more sophisticated graph-based adversarial learning for fairness modeling. FairGNN [Dai and Wang, 2021] proposes a sensitive attribute estimator and incorporates adversarial debiasing and covariance constraints to regularize the GNN for fair node representations and predictions. Recently, FairMI [Zhao *et al.*, 2023] attempts to mitigate the impact of sensitive attributes on the final recommendation representation by applying two aspects of mutual information. Different from these approaches, our approach utilizes the information bottleneck principle for fair representation learning, effectively achieving the best tradeoff between recommendation accuracy and fairness.

2.2 Information Bottleneck Principle

The Information Bottleneck (IB) principle seeks to encapsulate the balance in the to-be-learned representation between the essential information required for decision and the information retained from the input [Tishby *et al.*, 2000]. IB has been employed to enhance the interpretability and disentangle representation in deep learning-based tasks [Bao, 2021; Jeon *et al.*, 2021]. However, precise calculation of the mutual information between two high-dimensional random variables poses a significant challenge. To address this challenge, researchers use neural networks to approximate and estimate mutual information [Alemi *et al.*, 2017]. For instance, InfoNCE [Oord *et al.*, 2018] and variational based method [Alemi *et al.*, 2017] have been introduced to estimate the lower bound of mutual information. Recently, CLUB is proposed to estimate the upper bound of mutual information based on the log-ratio contrastive learning [Cheng *et al.*, 2020], which heavily relies on prior assumptions. In addition to directly optimizing mutual information objectives, researchers employ the Hilbert-Schmidt Independence Criterion (HSIC) as an alternative to mutual information estimation, which can assess the independence of two variables [Ma *et al.*, 2020]. Given the challenge of estimating the upper bound of mutual information, we opt for HSIC as an approximation to minimize the mutual information between the learned representations and sensitive attributes.

3 Preliminary

3.1 Problem Statement

In a recommender system, there are two entity sets: a user set $U(|U| = M)$ and an item set $V(|V| = N)$. The interactions between users and items are represented by the matrix $\mathbf{R} \in \mathbb{R}^{M\times N}$. Specifically, in scenarios involving common implicit feedback, the notation $r_{ai} = 1$ signifies that user a has interacted with item i, otherwise $r_{ai} = 0$. The interactions form a user-item bipartite graph, which can be formulated as $\mathcal{G} =< U \cup V, \boldsymbol{A} >$. The adjacency matrix $\boldsymbol{A}$ can be formulated as:

$$\boldsymbol{A} = \left[\begin{array}{cc} \mathbf{0}^{M\times M} & \mathbf{R} \\ \mathbf{R}^T & \mathbf{0}^{N\times N} \end{array} \right]. \tag{1}$$

$\boldsymbol{X} \in \mathbb{R}^{M\times D}$ and $\boldsymbol{Y} \in \mathbb{R}^{N\times D}$ denote the learned user and item representations, where D refers to the dimension of representation. The goal of recommendation is to predict the potential preference $\hat{r}_{ai}$ of user a to item i, and can be calculated with $\hat{r}_{ai} = \boldsymbol{x}_a^T \boldsymbol{y}_i$, where $\boldsymbol{x}_a$ is user a's representation, $\boldsymbol{y}_i$ is item i's representation. Given a binary sensitive attribute $k \in \{0, 1\}$, s_a refers to user a's attribute value. Then, we can partition the user set U into two subsets:U_0 and U_1, where U_0 represents the set of users with a sensitive attribute value of 0, and U_1 represents the set of users with a sensitive attribute value of 1. Our goal is to learn fair user representations $\mathbf{X}$ while maintaining competitive recommendation accuracy.

3.2 Hilbert-Schmidt Independence Criterion (HSIC)

HSIC is a statistical method employed to quantify the independence between two variables. Followed by [Ma *et al.*, 2020], $HSIC(A, B)$ can be formulated as:

$$HSIC(A, B) = ||C_{AB}||_{hs}^2, \tag{2}$$

where C_{AB} is the cross-covariance operator between the Reproducing Kernel Hilbert Spaces (RKHSs) of A and B, $||\cdot||_{hs}^2$ refers to Hilbert-Schmidt norm. Given the sampled instances $(a_i, b_i)_{i=1}^n$ from the batch training data, we estimate HISC as:

$$H\hat{S}IC(A, B) = \frac{Tr(\mathbf{K}_A \mathbf{H} \mathbf{K}_B \mathbf{H})}{(n-1)^2}, \tag{3}$$

where $\mathbf{K}_A$ and $\mathbf{K}_B$ are kernel Gram matrices [Ham *et al.*, 2004] of A and B, $\mathbf{H} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is a centering matrix, and $Tr(\cdot)$ refers to the trace of the matrix. In our work, we use the commonly employed radial basis function (RBF) [Vert *et al.*, 2004] as the kernel function K_A and K_B, as follows:

$$K(a_i, a_j) = exp(-\frac{||a_i - a_j||^2}{2\sigma^2}), \tag{4}$$

where σ is a hyperparameter that controls the sharpness of RBF.

4 The Proposed *FairIB*

4.1 Overall Objective of *FairIB*

Here, we first present the overall optimization objective of our proposed *FairIB*. The goal of *FairIB* is to learn the optimal user representations $\mathbf{X}$ by $Max : I(\mathbf{R}; \mathbf{X}) - \beta I(\mathbf{X}; \mathbf{S})$. Considering that each user interacts with different items, the optimal user representations $\mathbf{X}$ should match all interacted items. We relax the maximization $I(\mathbf{R}; \mathbf{X})$ to maximization $I(\mathbf{R}; \mathbf{X}, \mathbf{Y})$. Besides, users' sensitive attributes are not only exposed in their representations, but also in user-centric subgraphs [Wu *et al.*, 2021b]. Aiming to eliminate user-sensitive attributes further, we additionally minimize the mutual information of sensitive attributes $\mathbf{S}$ and user-centric sub-graph representations $\mathbf{G}$ of all users. Therefore, the overall objective of *FairIB* is formulated as follows:

$$Max: I(\mathbf{R};\mathbf{X},\mathbf{Y}) - \beta \underbrace{I(\mathbf{X};\mathbf{S})}_{user\ side} - \gamma \underbrace{I(\mathbf{G};\mathbf{S})}_{sub\text{-}graph\ side}, \tag{5}$$

where β and γ are two parameters that control the bottleneck scales, $\mathbf{G}$ denotes user-centric sub-graph representations. Eq.(5) describes that the optimal user representations are acquired in the trade-off of both recommendation accuracy and fairness demand.

4.2 Maximization of $I(\mathbf{R};\mathbf{X},\mathbf{Y})$

We first present how to maximize the mutual information $I(\mathbf{R};\mathbf{X},\mathbf{Y})$, ensuring that the representations of users and items contain sufficient information for recommendation tasks. Based on the properties of mutual information and entropy [Kraskov *et al.*, 2004], we have the following derivation:

$$I(\mathbf{R};\mathbf{X},\mathbf{Y}) = H(\mathbf{R}) - H(\mathbf{R}|\mathbf{X},\mathbf{Y}). \tag{6}$$

As $H(\mathbf{R})$ is a constant, we only need to maximize the second term, as follows:

$$\begin{aligned}
&-H(\mathbf{R}|\mathbf{X},\mathbf{Y}) = -\sum_{a\in U, i\in V} p(\mathbf{x}_a,\mathbf{y}_i) H(\mathbf{R}|\mathbf{X}=\mathbf{x}_a,\mathbf{Y}=\mathbf{y}_i)\\
&= \sum_{a\in U, i\in V} p(\mathbf{x}_a,\mathbf{y}_i) \sum_{r_{ai}=0}^{1} p(r_{ai}|\mathbf{x}_a,\mathbf{y}_i) log(p(r_{ai}|\mathbf{x}_a,\mathbf{y}_i))\\
&= \sum_{a\in U, i\in V} \sum_{r_{ai}=0}^{1} p(r_{ai},\mathbf{x}_a,\mathbf{y}_i) log(p(r_{ai}|\mathbf{x}_a,\mathbf{y}_i))\\
&= \mathbb{E}[log(p(\mathbf{R}|\mathbf{X},\mathbf{Y}))].
\end{aligned} \tag{7}$$

Therefore, maximizing $I(\mathbf{R};\mathbf{X},\mathbf{Y})$ is equivalent to maximizing the log likelihood $p(\mathbf{R}|\mathbf{X},\mathbf{Y})$, which is the typical optimization objective for general recommender systems. In other words, we can employ any recommendation model to realize the goal of maximization $I(\mathbf{R};\mathbf{X},\mathbf{Y})$.

4.3 Minimization of $I(\mathbf{X},\mathbf{S})$ and $I(\mathbf{G},\mathbf{S})$

Sensitive Attribute Encoder. Following the previous works [Wu *et al.*, 2021b; Zhao *et al.*, 2023], we train a sensitive encoder to represent sensitive attributes $\mathbf{S}$. Considering that users' behaviors reflect their attributes [Kosinski *et al.*, 2013; Wu *et al.*, 2021a], we build a graph-based encoder for extraction from the interaction data, followed by a sensitive attribute classifier. After training the encoder, we obtain sensitive representations for all users, i.e., $\{\mathbf{h}_a|a\in U\}$. To further get the sensitive representations for each user group (i.e., U_0 and U_1), we calculate the mean value within each group. This process is calculated as follows:

$$\boldsymbol{e}_k = \frac{1}{|U_k|}\sum_{a\in U_k} \boldsymbol{h}_a, \tag{8}$$

where $k\in\{0,1\}$ refers to different sensitive attribute values, U_k denotes different user groups, and $\boldsymbol{h}_a$ denotes the sensitive representation for user a learned from the encoder. Finally, we obtain the sensitive representations $\mathbf{E} = [\mathbf{e}_0, \mathbf{e}_1]$.

HSIC-based MI Minimization. Given the encoded sensitive representations $\mathbf{E}$, we instead $I(\mathbf{X};\mathbf{S})$ of $I(\mathbf{X};\mathbf{E})$. In the same manner, we instead $I(\mathbf{G};\mathbf{S})$ of $I(\mathbf{G};\mathbf{E})$. However, it's hard to minimize the mutual information because estimating the upper bound of MI is still an intractable problem. Although previous studies use variational based methods to estimate the upper bound of MI, they heavily rely on the prior distribution and the quality of sampling influences the accuracy of the estimation [Alemi *et al.*, 2017; Cheng *et al.*, 2020]. Here we introduce HSIC to replace mutual information for optimization. Therefore, the optimization objective for user side minimization is:

$$H\hat{S}IC(\mathbf{X},\mathbf{E}) = \sum_{a\in U_b} H\hat{S}IC(\boldsymbol{x}_a, \boldsymbol{e}_{s_a}), \tag{9}$$

where U_b denote users in the batch training data, $\mathbf{x}_a$ and $\mathbf{e}_{s_a}$ denote user a's representation and corresponding sensitive attribute representation. For sub-graph side minimization, we first compute user-centric sub-graph representations $\mathbf{G}$ as follows:

$$\begin{gathered}
\mathbf{G} = g(\mathbf{X}^1, ..., \mathbf{X}^L),\\
\begin{bmatrix}\boldsymbol{X}^{l+1}\\ \boldsymbol{Y}^{l+1}\end{bmatrix} = \boldsymbol{D}_A^{-\frac{1}{2}}\mathbf{A}\boldsymbol{D}_A^{-\frac{1}{2}}\begin{bmatrix}\boldsymbol{X}^{l}\\ \boldsymbol{Y}^{l}\end{bmatrix},
\end{gathered} \tag{10}$$

where $\mathbf{X}^l$ and $\mathbf{Y}^l$ denote user and item representations in l^{th} layer, $\mathbf{X}^{l+1}$ and $\mathbf{Y}^{l+1}$ denote user and item representations in $l+1^{th}$ layer. $\boldsymbol{D}_A$ is the degree of the adjacent matrix $\mathbf{A}$ and L is the depth of sub-graph, $g(\cdot)$ is the readout function. Similar to the user side, we implement the minimization of the sub-graph side as follows:

$$H\hat{S}IC(\mathbf{G},\boldsymbol{E}) = \sum_{a\in U_b} H\hat{S}IC(\boldsymbol{g}_a, \boldsymbol{e}_{s_a}), \tag{11}$$

where U_b denote users in the training batch data, $\boldsymbol{g}_a$ is user a's sub-graph representation. Thus, we obtain the minimization objective for a specific sensitive attribute (such as gender):

$$\mathcal{L}_{HSIC} = \beta H\hat{S}IC(\mathbf{X};\mathbf{E}) + \gamma H\hat{S}IC(\mathbf{G};\mathbf{E}). \tag{12}$$

Here we only present the loss function for a single sensitive attribute. In fact, *FairIB* is flexible to extend multiple sensitive attributes, which only performs multiple bottleneck learning for each sensitive attribute. Then, we can get multi-sensitive loss based on HSIC, as follows:

$$\mathcal{L}_{HSIC} = \sum_{t\in T}\mathcal{L}^t_{HSIC}, \tag{13}$$

where T denotes the set of multi-sensitive attributes, $\mathcal{L}^t_{HSIC}$ denotes the loss function of t^{th} sensitive attribute.

4.4 Model Optimization

As we mentioned in section 4.2, we employ common pair-wise ranking loss to optimize the maximization of $I(\mathbf{R};\mathbf{X},\mathbf{Y})$:

$$\mathcal{L}_{rec} = -\sum_{a\in U}\sum_{(i,j)\in D_a} log\sigma(\boldsymbol{x}_a^T\boldsymbol{y}_i - \boldsymbol{x}_a^T\boldsymbol{y}_j) + \alpha||\mathbf{\Theta}||^2, \tag{14}$$

Datasets	Users	Items	Interactions	Attributes
Movielens-1M	6,040	3,952	1,000,209	gender, age
LastFM	48,386	21,711	2,045,305	gender

Table 1: The statistics of two datasets.

where $\sigma(\cdot)$ is the sigmoid function, $\mathbf{\Theta} = [\mathbf{X}, \mathbf{Y}]$ is user and item free embedding matrices and α controls the L_2 regularization coefficient. $D_a = \{(i, j) | i \in R_a \cap j \in V - R_a\}$, R_a denotes user a interacted items. By combining the independence constraints of HSIC-based bottleneck on the user side and the sub-graph side, we obtain the final optimization objective of *FairIB*:

$$\mathcal{L}_{all} = \mathcal{L}_{rec} + \mathcal{L}_{HSIC}. \tag{15}$$

5 Experiments

In this section, we first introduce our experimental settings. Then, we conduct extensive comparisons with SOTA methods to verify the effectiveness of our proposed *FairIB*. Finally, we give a detailed analysis of our method, including ablation studies and parameter sensitivities.

5.1 Experimental Settings

Datasets. To evaluate the effectiveness of our proposed method, we select two real-world recommendation datasets: Movielens-1M [Harper and Konstan, 2015; Wu *et al.*, 2020] and LastFM [Celma Herrada and others, 2009]. Following the previous works [Wu *et al.*, 2021a; Zhao *et al.*, 2023], we split all interactions into training, validation, and test data. For Movielens-1M, we treat gender as the single attribute, and select gender and age as the compositional setting for further experimental study. For LastFM, we only use gender as the sensitive attribute. Detailed statistics of the two datasets are summarized in Table 1.

Evaluation Metrics. As we concentrate on the trade-off between recommendation accuracy and fairness, it is essential to evaluate and report results for both aspects. For measuring the recommendation accuracy, we employ two widely used ranking metrics: NDCG [Järvelin and Kekäläinen, 2017] and Recall [Gunawardana and Shani, 2009]. Larger values of NDCG and Recall indicate superior recommendation accuracy performance. For fairness evaluation, we also employ two popular group fairness metrics: Demographic Parity (DP) [Caton and Haas, 2020] and Equal Opportunity (EO) [Hardt *et al.*, 2016]. Specifically, we calculate DP as follows:

$$\begin{gathered} f^i_{U_0} = \frac{\sum_{a \in U_0} \mathbb{I}_{i \in Q_a}}{|U_0|}, f^i_{U_1} = \frac{\sum_{a \in U_1} \mathbb{I}_{i \in Q_a}}{|U_1|}, \\ \boldsymbol{f}_{U_0} = [f^1_{U_0}, ..., f^i_{U_0}, ..., f^N_{U_0}], \boldsymbol{f}_{U_1} = [f^1_{U_1}, ..., f^i_{U_1}, ..., f^N_{U_1}], \end{gathered} \tag{16}$$

where $Q_a = TopK_a$, and $TopK_a$ is Top-K ranked items for user a; $i \in V$, U_0 and U_1 denote user group with different sensitive attributes; $\mathbb{I}$ is an indicator function, if item i is in the set Q_a, then $\mathbb{I} = 1$, otherwise $\mathbb{I} = 0$. Then we compute Jensen–Shannon divergence to compare two groups:

$$DP = JS(\boldsymbol{f}_{U_0}, \boldsymbol{f}_{U_1})|_{Q_a = TopK_a}. \tag{17}$$

The definition of DP requires that item i be recommended to the two groups with equal probability, regardless of the actual preferences of the population [Caton and Haas, 2020]. EO is proposed by considering user actual preferences, which is calculated as:

$$EO = JS(\boldsymbol{f}_{U_0}, \boldsymbol{f}_{U_1})|_{Q_a = TopK_a \cap R_a}, \tag{18}$$

where R_a denotes the items genuinely liked by user a in the test data. The smaller values of DP and EO, the better recommendation fairness.

Baselines. As our proposed *FairIB* is a model-agnostic method, we implement *FairIB* on two representative recommendation backbones: BPR-MF [Rendle *et al.*, 2012] and LightGCN [He *et al.*, 2020]. These two have emerged as representative recommendation backbones in recent years. We use the full-ranking strategy to evaluate all methods for a fair comparison. We compare *FairIB* with SOTA fairness-aware recommendation methods:

- **Reg** [Yao and Huang, 2017]: is a regularization-based model that incorporates various statistical fairness regularization terms.
- **Adv** [Bose and Hamilton, 2019]: is an adversarial learning method designed to minimize the correlation between sensitive attributes and the learned representation.
- **FairRec** [Wu *et al.*, 2021a]: is a fairness-aware approach with decomposed adversarial learning and orthogonality regularization.
- **FairGo** [Wu *et al.*, 2021b]: develops a more sophisticated adversarial learning approach that takes into account the hidden unfairness within a user-centric graph.
- **FairGNN** [Dai and Wang, 2021]: incorporates adversarial debiasing and covariance constraints to regularize the GNN for fair node representations and predictions. It also includes a sensitive attribute estimator to address the challenge of missing sensitive attribute information.
- **FairMI** [Zhao *et al.*, 2023]: proposes a two-fold mutual information optimization framework, which employs a self-supervised learning approach to maximize the mutual information between pre-trained SOTA model representation and to-be-learned representation. Please note that, we removed the self-supervised learning part for a fair comparison (which is only designed for enhancing accuracy), denoted as FairMI*.

Implementation Details. We conduct experiments on an NVIDIA A40 GPU with Pytorch-2.1.2. For the sensitive attribute encoder, we utilize LightGCN as the backbone and a one-layer fully connected network as the attribute classifier. For model training, we set the latent embedding size as $D = 64$, the batch size is set to 2048 for Movielens-1M and 4096 for LastFM. The regularization parameter α is set to 0.001. We adopt the Adam optimizer with a learning rate of 0.001. We repeat experiments 10 times and report the average results.

5.2 Overall Performance

Table 2 and 3 report the overall experimental results on Movielens-1M and LastFM. From these Tables, we have the following observations:

Models	Methods	NDCG@K↑		Recall@K↑		DP@K↓		EO@K↓	
		10	30	10	30	10	30	10	30
BPR-MF	Base	0.1943	0.2926	0.1437	0.2916	0.2854	0.2412	0.3580	0.3122
	Reg	0.1899	0.2859	0.1402	0.2845	0.1954	0.1575	0.2904	0.2576
	Adv	0.1900	0.2866	0.1404	0.2858	0.1684	0.1214	0.2736	0.2363
	FairRec	0.1896	0.2860	0.1407	0.2847	0.1656	0.1191	0.2714	0.2334
	FairMI*	0.1901	0.2867	0.1410	0.2855	0.1521	0.1145	0.2608	0.2199
	FairIB	**0.1922**	**0.2903**	**0.1428**	**0.2907**	**0.1453**	**0.1140**	**0.2117**	**0.1740**
LightGCN	Base	0.2018	0.3060	0.1511	0.3085	0.2919	0.2449	0.3609	0.3085
	Reg	0.1961	0.2938	0.1422	0.2828	0.2097	0.1545	0.3047	0.2611
	Adv	0.1963	0.2975	0.1469	0.2998	0.1532	0.1068	0.2694	0.2203
	FairRec	0.1950	0.2955	0.1472	0.2986	0.1536	0.1042	0.2590	0.2243
	FairGo	0.1822	0.2741	0.1336	0.2710	0.2728	0.2275	0.3382	0.2921
	FairGNN	0.1964	0.2963	0.1466	0.2969	0.1472	0.1045	0.2608	0.2221
	FairMI*	0.1978	0.2967	0.1480	0.2980	0.1436	0.1046	0.2560	0.2148
	FairIB	**0.2003**	**0.3013**	**0.1502**	**0.3061**	**0.1408**	**0.1033**	**0.2045**	**0.1686**

Table 2: Recommendation accuracy and fairness performances on the Movielens-1M dataset. We report comparisons across different Top-K values. We compare all fairness-aware methods, the best results are highlighted in **bold** and the second best results are displayed in underline.

Models	Methods	NDCG@K↑		Recall@K↑		DP@K↓		EO@K↓	
		10	30	10	30	10	30	10	30
BPR-MF	Base	0.1959	0.2743	0.1564	0.2943	0.2664	0.2376	0.3345	0.3122
	Reg	0.1830	0.2477	0.1445	0.2819	0.2072	0.1728	0.2953	0.2710
	Adv	0.1813	0.2518	0.1443	0.2791	0.1335	0.1101	0.2618	0.2501
	FairRec	0.1876	0.2644	0.1502	0.2853	0.1484	0.1169	0.2712	0.2592
	FairMI*	0.1884	0.2595	0.1505	0.2827	0.1273	0.0958	0.2580	0.2476
	FairIB	**0.1897**	**0.2665**	**0.1517**	**0.2861**	**0.1243**	**0.0930**	**0.2474**	**0.2403**
LightGCN	Base	0.1971	0.2762	0.1572	0.2964	0.2860	0.2569	0.3508	0.3247
	Reg	0.1863	0.2619	0.1484	0.2865	0.2201	0.1965	0.3012	0.2867
	Adv	0.1887	0.2652	0.1499	0.2853	0.1382	0.1163	0.2682	0.2576
	FairRec	0.1892	0.2627	0.1505	0.2837	0.1397	0.1157	0.2700	0.2596
	FairGo	0.1693	0.2389	0.1371	0.2627	0.2626	0.2338	0.3282	0.3059
	FairGNN	0.1879	0.2651	0.1501	0.2859	0.1372	0.1171	0.2690	0.2592
	FairMI*	0.1888	0.2662	0.1506	0.2866	0.1381	0.1058	0.2653	0.2548
	FairIB	**0.1900**	**0.2678**	**0.1520**	**0.2876**	**0.1313**	**0.0998**	**0.2325**	**0.2314**

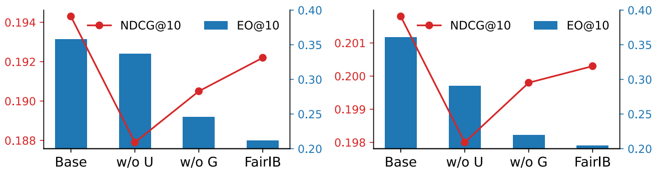
Table 3: Recommendation accuracy and fairness performances on the LastFM dataset. We report comparisons across different Top-K values. We compare all fairness-aware methods, the best results are highlighted in **bold** and the second best results are displayed in underline.

- Compared with base models (BPR-MF, LightGCN), all fairness-aware methods present better recommendation fairness but a slight accuracy drop. This is caused by the natural data distribution, accuracy-fairness is a trade-off process. Furthermore, we find that graph-based backbone (LightGCN) has higher recommendation accuracy but worse recommendation fairness. The reason is that graph convolutions strengthen the collaborative signals, also amplify the sensitive attributes.
- Compared with fairness regularization, adversarial-based methods achieve better fairness performances, which demonstrates that learning fair representations is a more effective technique rather than a simple statistic-based constraint.
- Our proposed *FairIB* consistently outperforms all baselines on fairness evaluation. Specifically, *FairIB* improves LightGCN $w.r.t$ DP@30 by 57.82%, 61.15% on Movielens-1M and LastFM, respectively. Besides, compared with the strongest fairness baseline, i.e., FairMI, *FairIB* also presents better fairness performances, demonstrating the effectiveness of our designed HSIC-based bottleneck learning.
- Besides fairness metrics, we observe that *FairIB* also obtains the best recommendation accuracy results except for the base model. Experiments show the superiority of *FairIB* in producing the fairest recommendation results with the fewest accuracy sacrifice. All indicate that *FairIB* acquires the efficient accuracy-fairness trade-off in the IB manner.

Multi-sensitive Setting. Here, we present comparisons of all methods under multiple sensitive scenarios. As illustrated in Table 4, we conduct experiments on Movielens-1M, and analyze model performances on both gender and age fairness. For the age attribute, we divide users into two groups based on binarization. From the Table 4, we have the following observations. First, *FairIB* achieves the best accuracy and multiple

Models	NDCG↑	Recall↑	DP↓		EO↓	
			Gender	Age	Gender	Age
Base	0.2018	0.1511	0.2919	0.1548	0.3609	0.2237
Reg	0.1953	0.1425	0.2120	0.1483	0.3120	0.2210
Adv	0.1940	0.1448	0.1477	<u>0.1231</u>	0.2588	0.2013
FairRec	0.1920	0.1432	0.1511	0.1327	0.2535	0.1996
FairGo	0.1817	0.1304	0.2730	0.1410	0.3256	0.2178
FairGNN	0.1943	0.1454	0.1503	0.1313	<u>0.2440</u>	<u>0.1905</u>
FairMI*	<u>0.1969</u>	<u>0.1469</u>	<u>0.1428</u>	0.1523	0.2585	0.2180
FairIB	**0.2002**	**0.1487**	**0.1397**	**0.1137**	**0.2046**	**0.1763**

Table 4: Comparisons on multiple sensitive attributes (Top-K=10).



(a) BPR-Based (b) LightGCN-Based

Figure 2: Ablation study with Top-K=10 on Movielens-1M.

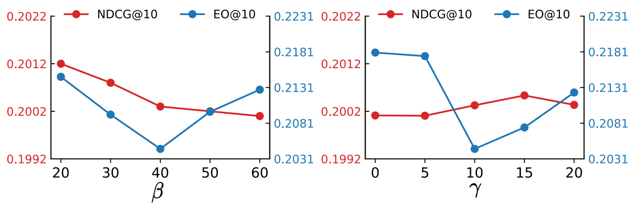


Figure 3: *FairIB* based on LightGCN with different β and γ.

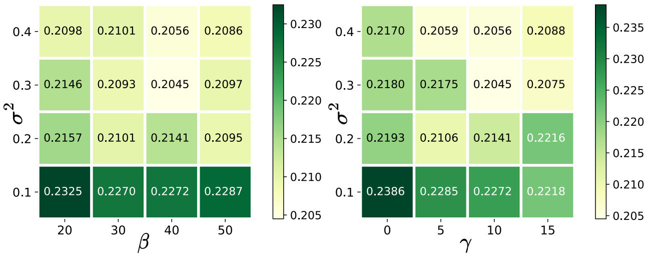


Figure 4: Impact of different IB loss parameters (β, σ^2) and (γ, σ^2).

sensitive attributes fairness performance among all fairness-aware baselines. This demonstrates that our proposed *FairIB* also acquires an efficient accuracy-fairness trade-off under multi-sensitive attributes. Besides, under the multi-sensitive scenario, *FairIB* is more convenient compared with adversarial methods as we don't need multiple discriminator learning. Overall, either single-sensitive or multi-sensitive attributes, *FairIB* is effective and efficient in achieving recommendation accuracy-fairness trade-off.

5.3 Model Analysis

Ablation Studies. In this part, we conduct ablation studies on Movielens-1M to verify the effectiveness of each component in *FairIB*. As illustrated in Figure 2, we conduct experiments on BPR-MF and LightGCN backbones. Among them, "w/o U" and "w/o G" denote the variants of *FairIB* without user and sub-graph sides bottleneck learning (Eq.(5)), respectively. From Figure 2, we have the following observations. First, "w/o U" shows better performance both in accuracy and fairness compared to "w/o G". This indicates that bottleneck learning on the user side is effective in eliminating more unnecessary sensitive information while capturing essential collaborative signals more effectively. Second, by equipping bottleneck learning on both the user side and sub-graph side, our method can further improve recommendation accuracy and fairness.

Parameter Sensitivity Analyses. We investigate the impact of user and sub-graph bottleneck learning scales β and γ in Eq. (12) on Movielens-1M. Since the optimal parameters for *FairIB* on the Movielens-1M dataset are β=40 and γ=10, we fix β=40 and analyze the sensitivity of γ. Similarly, we fix γ=10 and analyze the sensitivity of β. We have several observations from Figure 3. First, with the increase of β, this implies the enhancement of on user-side bottleneck learning, user representations will contain less sensitive information, which achieves a better fairness performance accompanied by a decrease in accuracy. However, a larger β will impact the model's training, resulting in a deterioration of both accuracy and fairness. The optimal balance is achieved when β equals 40. Second, after reaching the optimal value for β, a moderate γ can simultaneously improve both fairness and accuracy. We speculate that this might be a proper bottleneck learning on the sub-graph side helping to capture noise-sensitive information in the graph structure, thereby enhancing both model accuracy and fairness simultaneously. Finally, *FairIB* achieves an efficient trade-off of recommendation accuracy and fairness when β is set to 40 and γ to 10.

Furthermore, we also delve into the impact of kernel function parameter σ^2 in Eq. (4) on Movielens-1M. We investigated various combinations involving σ^2 in $[0.1, 0.2, 0.3, 0.4]$, β in $[20, 30, 40, 50]$ and γ in $[0, 5, 10, 15]$. As shown in Figure 4, *FairIB* achieve the best fairness performance (EO@10) when $(\beta = 40, \sigma^2 = 0.3)$ and $(\gamma = 10, \sigma^2 = 0.3)$. Comprehensive experiments have verified that a moderate value for σ^2 (i.e., 0.3) significantly contributes to improving the model's fairness performance.

6 Conclusion

In this paper, we researched fairness-aware recommender systems from the information theory perspective. Motivated by the information bottleneck principle, we proposed a novel model-agnostic fair representation method *FairIB* to eliminate the sensitive information from the learned representations. Specifically, *FairIB* maximizes the mutual information between learned representations and observed interactions, meanwhile minimizing it between representations and user sensitive attributes. To achieve this goal, we introduced HSIC-based bottleneck to recommender systems, and applied to both the user and sub-graph sides. Extensive experiments on two real-world datasets demonstrated *FairIB* is effective in balancing recommendation accuracy-fairness trade-off, either in single or multiple sensitive scenarios. In the future, we would like to extend the proposed method by mining attribute correlations to address the practical scenario of missing values for sensitive attributes.

Acknowledgments

This work was supported in part by grants from the National Key Research and Development Program of China (Grant No.2021ZD0111802), the National Natural Science Foundation of China (Grant No.U23B2031, 72188101).

Contribution Statement

Junsong Xie and Yonghui Yang contribute equally to this work.

References

[Alemi *et al.*, 2017] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR*, 2017.

[Arjovsky *et al.*, 2017] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223. PMLR, 2017.

[Bao, 2021] Feng Bao. Disentangled variational information bottleneck for multiview representation learning. In *Artificial Intelligence: First CAAI International Conference, CICAI 2021, Hangzhou, China, June 5–6, 2021, Proceedings, Part II 1*, pages 91–102. Springer, 2021.

[Biega *et al.*, 2018] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *SIGIR*, pages 405–414, 2018.

[Bose and Hamilton, 2019] Avishek Bose and William Hamilton. Compositional fairness constraints for graph embeddings. In *ICML*, pages 715–724. PMLR, 2019.

[Caton and Haas, 2020] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *CSUR*, 2020.

[Celma Herrada and others, 2009] Òscar Celma Herrada et al. *Music recommendation and discovery in the long tail*. Universitat Pompeu Fabra, 2009.

[Chen *et al.*, 2020] Lei Chen, Le Wu, Richang Hong, Kun Zhang, and Meng Wang. Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach. In *AAAI*, volume 34, pages 27–34, 2020.

[Cheng *et al.*, 2020] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *ICML*, pages 1779–1788. PMLR, 2020.

[Covington *et al.*, 2016] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *RecSys*, pages 191–198, 2016.

[Dai and Wang, 2021] Enyan Dai and Suhang Wang. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *WSDM*, pages 680–688, 2021.

[Dong *et al.*, 2023] Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li. Fairness in graph mining: A survey. *TKDE*, 2023.

[Ekstrand *et al.*, 2018a] Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Conference on fairness, accountability and transparency*, pages 172–186. PMLR, 2018.

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

[Gao *et al.*, 2023] Chongming Gao, Shiqi Wang, Shijun Li, Jiawei Chen, Xiangnan He, Wenqiang Lei, Biao Li, Yuan Zhang, and Peng Jiang. Cirs: Bursting filter bubbles by counterfactual interactive recommender system. *TOIS*, 42(1):1–27, 2023.

[Gunawardana and Shani, 2009] Asela Gunawardana and Guy Shani. A survey of accuracy evaluation metrics of recommendation tasks. *JMLR*, 10(12), 2009.

[Ham *et al.*, 2004] Jihun Ham, Daniel D Lee, Sebastian Mika, and Bernhard Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *ICML*, page 47, 2004.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NeurIPS*, volume 29, 2016.

[Harper and Konstan, 2015] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *TIIS*, 5(4):1–19, 2015.

[He *et al.*, 2020] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *SIGIR*, pages 639–648, 2020.

[Järvelin and Kekäläinen, 2017] Kalervo Järvelin and Jaana Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *SIGIR*, volume 51, pages 243–250. ACM New York, NY, USA, 2017.

[Jeon *et al.*, 2021] Insu Jeon, Wonkwang Lee, Myeongjang Pyeon, and Gunhee Kim. Ib-gan: Disentangled representation learning with information bottleneck generative adversarial networks. In *AAAI*, volume 35, pages 7926–7934, 2021.

[Kosinski *et al.*, 2013] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 110(15):5802–5805, 2013.

[Kraskov *et al.*, 2004] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[Li *et al.*, 2021] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. User-oriented fairness in recommendation. In *WWW*, pages 624–632, 2021.

[Li *et al.*, 2023] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng

Zhang. Fairness in recommendation: Foundations, methods, and applications. *TIST*, 14(5):1–48, 2023.

[Ma *et al.*, 2020] Wan-Duo Kurt Ma, JP Lewis, and W Bastiaan Kleijn. The hsic bottleneck: Deep learning without back-propagation. In *AAAI*, volume 34, pages 5085–5092, 2020.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *CSUR*, 54(6):1–35, 2021.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv*, 2018.

[Rendle *et al.*, 2012] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv*, 2012.

[Salimans *et al.*, 2016] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, volume 29, 2016.

[Shao *et al.*, 2022] Pengyang Shao, Le Wu, Lei Chen, Kun Zhang, and Meng Wang. Faircf: Fairness-aware collaborative filtering. *Science China Information Sciences*, 65(12):222102, 2022.

[Shao *et al.*, 2024] Pengyang Shao, Le Wu, Kun Zhang, Defu Lian, Richang Hong, Yong Li, and Meng Wang. Average user-side counterfactual fairness for collaborative filtering. *TOIS*, 2024.

[Tan *et al.*, 2021] Yanchao Tan, Carl Yang, Xiangyu Wei, Yun Ma, and Xiaolin Zheng. Multi-facet recommender networks with spherical optimization. In *ICDE*, pages 1524–1535. IEEE, 2021.

[Tishby *et al.*, 2000] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv*, 2000.

[Vert *et al.*, 2004] Jean-Philippe Vert, Koji Tsuda, and Bernhard Schölkopf. A primer on kernel methods. *Kernel methods in computational biology*, 47:35–70, 2004.

[Wang *et al.*, 2023] Zifeng Wang, Zheng Zhan, Yifan Gong, Yucai Shao, Stratis Ioannidis, Yanzhi Wang, and Jennifer Dy. Dualhsic: Hsic-bottleneck and alignment for continual learning. In *ICML*, 2023.

[Wu *et al.*, 2020] Le Wu, Yonghui Yang, Kun Zhang, Richang Hong, Yanjie Fu, and Meng Wang. Joint item recommendation and attribute inference: An adaptive graph convolutional network approach. In *SIGIR*, pages 679–688, 2020.

[Wu *et al.*, 2021a] Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. Fairness-aware news recommendation with decomposed adversarial learning. In *AAAI*, volume 35, pages 4462–4469, 2021.

[Wu *et al.*, 2021b] Le Wu, Lei Chen, Pengyang Shao, Richang Hong, Xiting Wang, and Meng Wang. Learning fair representations for recommendation: A graph-based perspective. In *WWW*, pages 2198–2208, 2021.

[Wu *et al.*, 2023] Chenwang Wu, Xiting Wang, Defu Lian, Xing Xie, and Enhong Chen. A causality inspired framework for model interpretation. In *KDD*, pages 2731–2741, 2023.

[Xiao *et al.*, 2017] Lin Xiao, Zhang Min, Zhang Yongfeng, Gu Zhaoquan, Liu Yiqun, and Ma Shaoping. Fairness-aware group recommendation with pareto-efficiency. In *RecSys*, pages 107–115, 2017.

[Xie *et al.*, 2017] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *NeurIPS*, volume 30, 2017.

[Yang *et al.*, 2023] Yonghui Yang, Zhengwei Wu, Le Wu, Kun Zhang, Richang Hong, Zhiqiang Zhang, Jun Zhou, and Meng Wang. Generative-contrastive graph learning for recommendation. In *SIGIR*, pages 1117–1126, 2023.

[Yao and Huang, 2017] Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. In *NeurIPS*, volume 30, 2017.

[Zhao *et al.*, 2023] Chen Zhao, Le Wu, Pengyang Shao, Kun Zhang, Richang Hong, and Meng Wang. Fair representation learning for recommendation: a mutual information perspective. In *AAAI*, volume 37, pages 4911–4919, 2023.

[Zhu *et al.*, 2018] Ziwei Zhu, Xia Hu, and James Caverlee. Fairness-aware tensor-based recommendation. In *CIKM*, pages 1153–1162, 2018.