Learning Fair Representations for Recommendation via Information Bottleneck Principle

Junsong Xie†, Yonghui Yang†, Zihan Wang and Le Wu†,2,∗
1Hefei University of Technology
2Institute 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 Information Bottleneck 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 min-max 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 unsatisfactory 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.
Traditional recommender systems focus on the accuracy task, the optimal user representations \( X \) are obtained by maximizing the mutual information between \( X \) and user-item observed interactions \( R \) \((I(R;X))\). However, empirical studies indicate that users’ behaviors can expose their sensitive attributes [Kosinski et al., 2013; Wu et al., 2021a], simply maximizing \( I(R;X) \) will introduce sensitive information to representations, leading to unfair recommendation results. To tackle this issue, our proposed FairIB acquires fair representations \( X \) based on information bottleneck learning. Besides maximizing the mutual information \( I(R;X) \), FairIB simultaneously minimizes the mutual information \( I(X;S) \). 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 \( \max : I(R;X) - \beta I(X;S) \) poses three challenges. First, the calculation of \( I(R;X) \) is difficult. Each user representation should match all interacted items, as they interact with different items. Second, it’s hard to minimize \( I(X;S) \) 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 \( I(R;X) \), 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 \( I(X;S) \), 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 categorized graph-based adversarial learning for fairness modeling. 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 trade-off 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 disentanglement in deep learning-based tasks [Bao, 2021; Jeon et al., 2021]. However, precise calculation of the mutual information between 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 \( R \in \mathbb{R}^{M \times N} \). 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 \( G = < U \cup V, A > \). The adjacency matrix \( A \) can be formulated as:

\[
A = \begin{bmatrix}
0^M \times M & R \\
R^T & 0^N \times N
\end{bmatrix}.
\]

\( X \in \mathbb{R}^{M \times D} \) and \( Y \in \mathbb{R}^{N \times D} \) 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 \( \hat{r}_{ai} = x^T_a y_i \), where \( x_a \) is user \( a \)'s representation, \( y_i \) 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 \( X \) 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:

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

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

\[
\hat{HSIC}(A, B) = \frac{Tr(K_A H K_B H)}{(n - 1)^2},
\]

where \( K_A \) and \( K_B \) are kernel Gram matrices [Ham et al., 2004] of \( A \) and \( B \). \( H = I - \frac{1}{n} 11^T \) 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:

\[
K(a_i, a_j) = \exp \left( - \frac{||a_i - a_j||^2}{2\sigma^2} \right),
\]

where \( \sigma \) 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 \( X \) by \( \max : I(R; X) - \beta I(X; S) \). Considering that each user interacts with different items, the optimal user representations \( X \) should match all interacted items. We relax the maximization \( I(R; X) \) to maximization \( I(R; X, Y) \). Besides, users’ sensitive attributes are not only exposed in their representations, but also in user-centric sub-graphs [Wu et al., 2021b]. Aiming to eliminate user-sensitive attributes further, we additionally minimize the mutual information of sensitive attributes \( S \) and user-centric sub-graph
representations $G$ of all users. Therefore, the overall objective of FairIB is formulated as follows:

$$\text{Max} : I(R; X, Y) - \beta I(X; S) - \gamma I(G; S)$$

where $\beta$ and $\gamma$ are two parameters that control the bottleneck scales, $G$ 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 $I(R; X, Y)$

We first present how to maximize the mutual information $I(R; X, Y)$, 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:

$$I(R; X, Y) = H(R) - H(R|X, Y).$$

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

$$- H(R|X, Y) = - \sum_{a \in U, i \in V} p(x_a, y_i)H(R|X = x_a, Y = y_i)$$

$$= \sum_{a \in U, i \in V} \sum_{r_{ai} = 0}^{1} p(x_a, y_i) \log(p(r_{ai}|x_a, y_i))$$

$$= \sum_{a \in U, i \in V} \sum_{r_{ai} = 0}^{1} \log(p(r_{ai}|x_a, y_i))$$

$$= \mathbb{E} \log(p(R|X, Y)).$$

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

4.3 Minimization of $I(X, S)$ and $I(G, S)$

Sensitive Attribute Encoder. Following the previous works [Wu et al., 2021b; Zhao et al., 2023], we train a sensitive encoder to represent sensitive attributes $S$. 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., $\{h_a|a \in U\}$. 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:

$$e_k = \frac{1}{|U_k|} \sum_{a \in U_k} h_a,$$

where $k \in \{0, 1\}$ refers to different sensitive attribute values, $U_k$ denotes different user groups, and $h_a$ denotes the sensitive representation for user $a$ learned from the encoder. Finally, we obtain the sensitive representations $E = [e_0, e_1]$.

4.4 Model Optimization

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

$$\mathcal{L}_{\text{rec}} = - \sum_{a \in U \cap \{i, j\} \in D_a} \log p(x_a^T y_i - x_a^T y_j) + \alpha ||\Theta||^2,$$
where \( \sigma(\cdot) \) is the sigmoid function, \( \Theta = [X, Y] \) is user and item free embedding matrices and \( \alpha \) 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 \( \text{FairIB} \):

\[
\mathcal{L}_{all} = \mathcal{L}_{rec} + \mathcal{L}_{HSIC}.
\]

### 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 \( \text{FairIB} \). Finally, we give a detailed analysis of our method, including ablating 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 Konst, 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, 2007] 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:

\[
f_U^i = \frac{\sum_{a \in U_a} I_i \in Q_a}{|U_a|}, \quad f_1^i = \frac{\sum_{a \in U_1} I_i \in Q_a}{|U_1|},
\]

\[
f_U = [f_U^1, ..., f_U^N], \quad f_1 = [f_1^1, ..., f_1^N],
\]

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

\[
DP = JS(f_U^1, f_1),
\]

The definition of \( EO \) 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:

\[
EO = JS(f_U^1, f_1)_{Q_a = TopK_a \cap R_a},
\]

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 \( \text{FairIB} \) is a model-agnostic method, we implement \( \text{FairIB} \) on two representative recommendation backbone: 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 \( \text{FairIB} \) 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 \( \alpha \) 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:
We compare all fairness-aware methods, the best results are highlighted in Table 3: Recommendation accuracy and fairness performances on the LastFM dataset. We report comparisons across different Top-K values.

- 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 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.

We compare all fairness-aware methods, the best results are highlighted in bold and the second best results are displayed in underline.

<table>
<thead>
<tr>
<th>Models</th>
<th>Methods</th>
<th>NDCG@K</th>
<th>Recall@K</th>
<th>DP@K</th>
<th>EO@K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>30</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>BPR-MF</td>
<td>Base</td>
<td>0.1943</td>
<td>0.2926</td>
<td>0.1437</td>
<td>0.2916</td>
</tr>
<tr>
<td></td>
<td>Reg</td>
<td>0.1899</td>
<td>0.2859</td>
<td>0.1402</td>
<td>0.2845</td>
</tr>
<tr>
<td></td>
<td>Adv</td>
<td>0.1900</td>
<td>0.2866</td>
<td>0.1404</td>
<td>0.2858</td>
</tr>
<tr>
<td></td>
<td>FairRec</td>
<td>0.1896</td>
<td>0.2860</td>
<td>0.1407</td>
<td>0.2847</td>
</tr>
<tr>
<td></td>
<td>FairMI</td>
<td>0.1901</td>
<td>0.2867</td>
<td>0.1410</td>
<td>0.2855</td>
</tr>
<tr>
<td></td>
<td>FairIB</td>
<td>0.1922</td>
<td>0.2903</td>
<td>0.1428</td>
<td>0.2907</td>
</tr>
</tbody>
</table>

| 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.
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 $\beta$ and $\gamma$ in Eq. (12) on Movielens-1M. Since the optimal parameters for FairIB on the Movielens-1M dataset are $\beta=40$ and $\gamma=10$, we fix $\beta=40$ and analyze the sensitivity of $\gamma$. Similarly, we fix $\gamma=10$ and analyze the sensitivity of $\beta$. We have several observations from Figure 2. First, with the increase of $\beta$, 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 $\beta$ will impact the model’s training, resulting in a deterioration of both accuracy and fairness. The optimal balance is achieved when $\beta$ equals 40. Second, after reaching the optimal value for $\beta$, a moderate $\gamma$ 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 $\beta$ is set to 40 and $\gamma$ to 10.

Furthermore, we also delve into the impact of kernel function parameter $\sigma^2$ in Eq. (4) on Movielens-1M. We investigated various combinations involving $\sigma^2$ in $[0.1, 0.2, 0.3, 0.4]$, $\beta$ in $[20, 30, 40, 50]$ and $\gamma$ 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 $\sigma^2$ (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


