南京航空航天大學
Nanjing University of Aeronautics and Astronautics

General Debiasing for Graph-based Collaborative Filtering via Adversarial Graph Dropout

An Zhang
National University of Singapore
anzhang@u.nus.edu

Wenchang Ma
National University of Singapore
e0724290@u.nus.edu

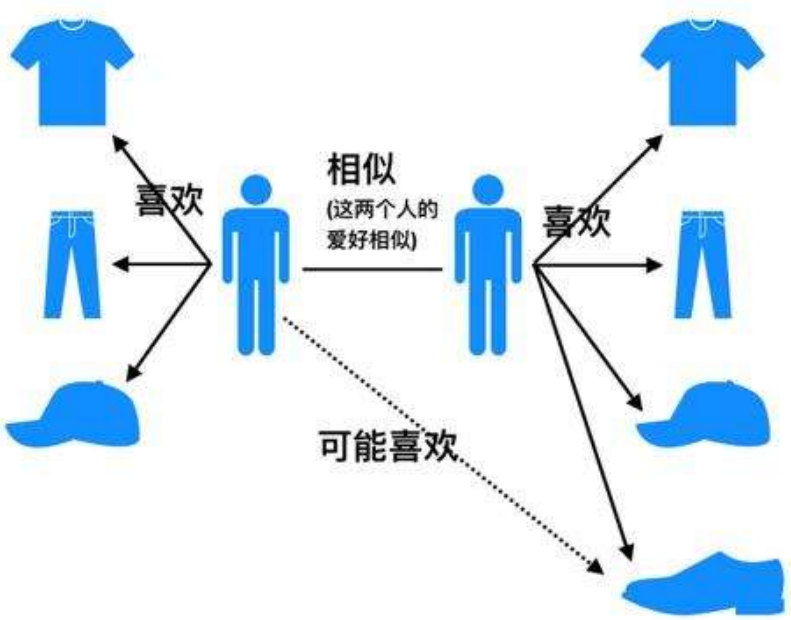
Pengbo Wei
National University of Singapore
pengbo.wei@u.nus.edu

Leheng Sheng
Tsinghua University
chenglh22@mails.tsinghua.edu.cn

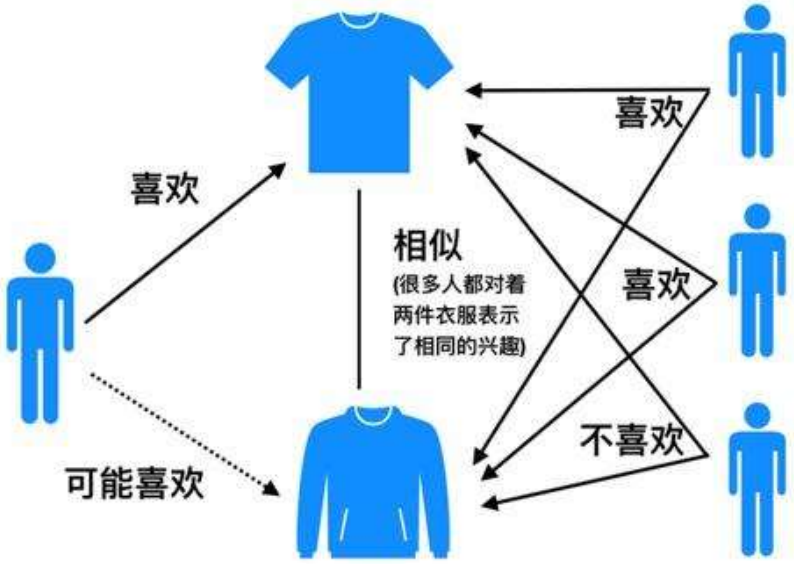
Xiang Wang*
University of Science and Technology of China
xiangwang1223@gmail.com

WWW 2024

Introduction



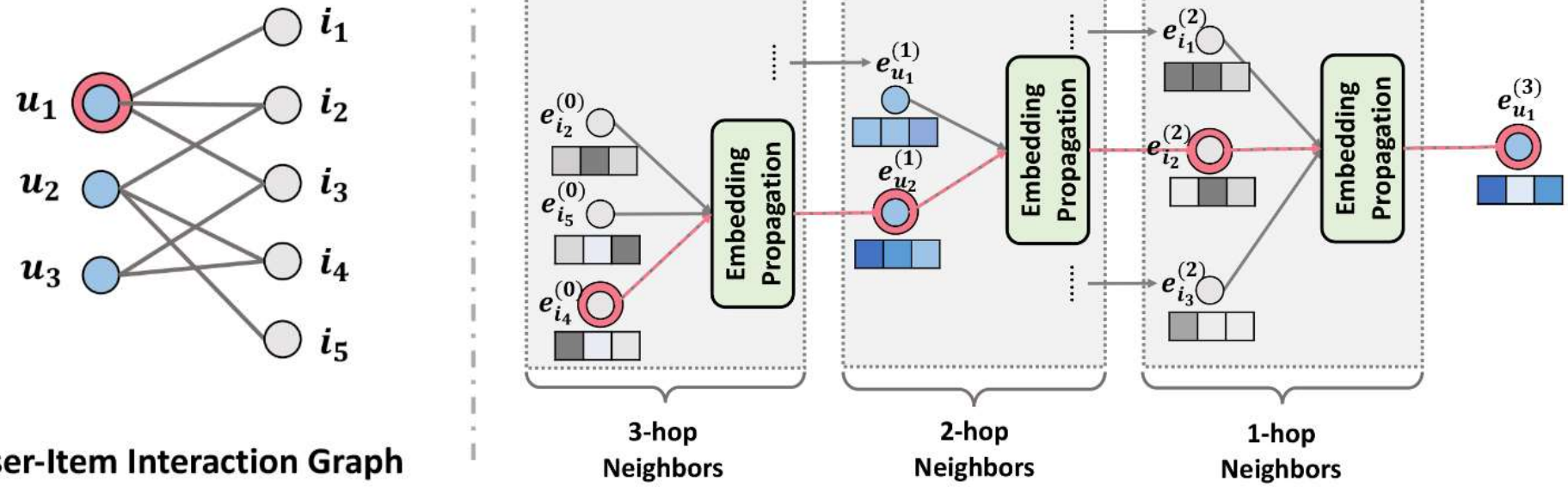
“人以群分”的基于用户的协同过滤



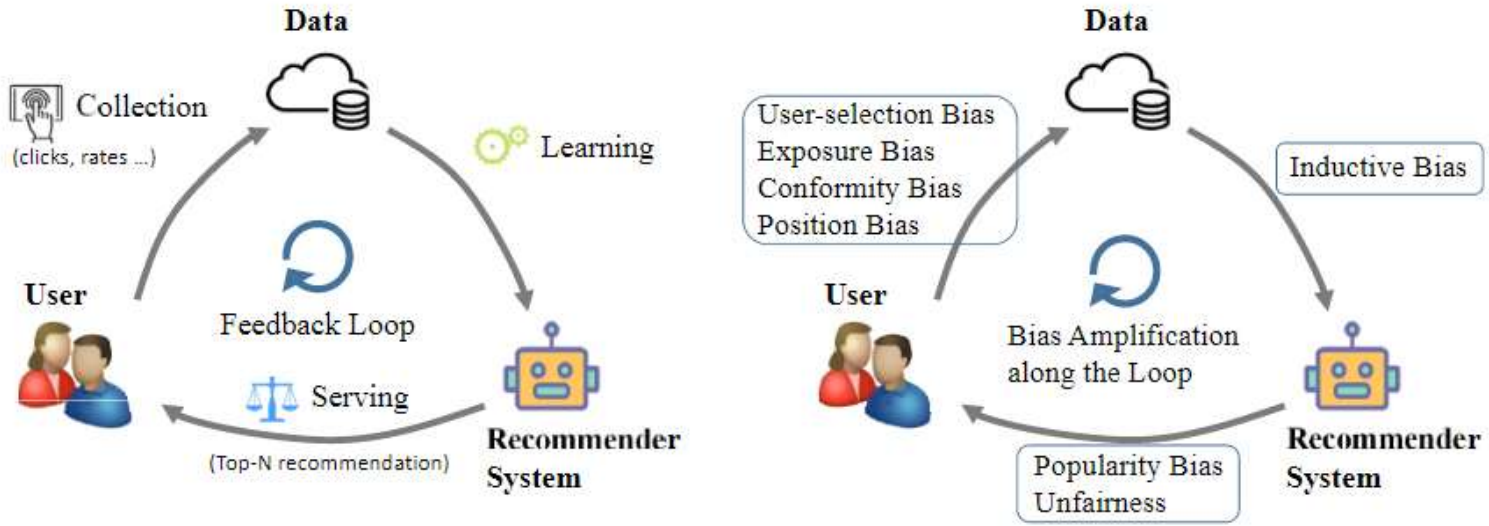
“物以类聚”的基于物品的协同过滤

Introduction

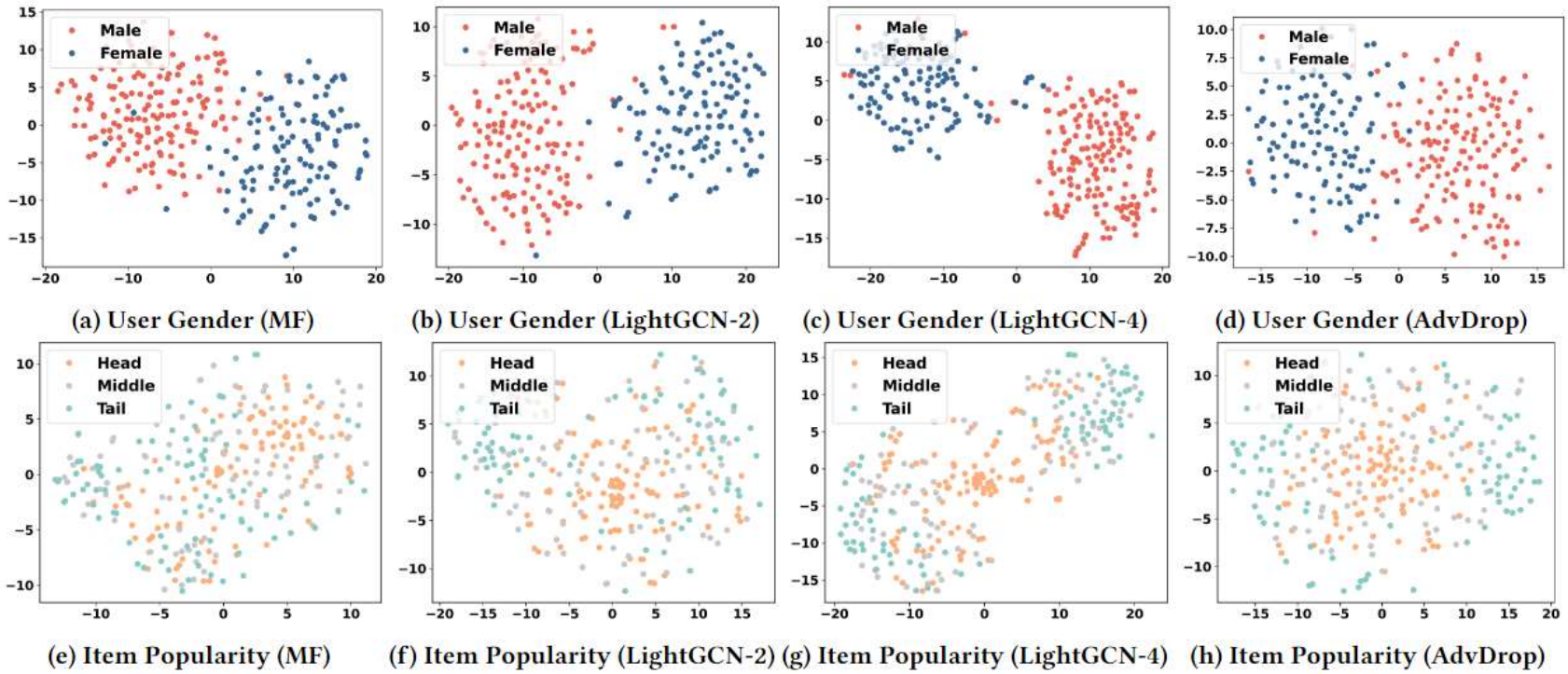
南京航空航天大學
Nanjing University of Aeronautics and Astronautics



Introduction



Introduction



(a) User Gender (MF) (b) User Gender (LightGCN-2) (c) User Gender (LightGCN-4) (d) User Gender (AdvDrop)

(e) Item Popularity (MF) (f) Item Popularity (LightGCN-2) (g) Item Popularity (LightGCN-4) (h) Item Popularity (AdvDrop)

Figure 1: T-SNE [46] visualizations of user and item representations learned by MF [40], LightGCN [26], and our proposed AdvDrop. Note that MF, LightGCN-2, and LightGCN-4 are specialized with zero, two, and four graph convolutional layers, respectively. Subfigures 1a-1d show the representation distribution *w.r.t.* two groups of user gender (*i.e.,* female, male), while Subfigures 1e-1h depict the representation distribution *w.r.t.* three groups of item popularity (*i.e.,* head, middle, tail).

Introduction

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Motivation:

(1) Data Bias Perspective: Typically, only one specific factor is considered to be mitigated, such as item popularity, user conformity, and sensitive attributes. However, interaction data is often riddled with various biases, whether predefined or arising from latent confounders.

(2) Bias Amplification Perspective. The debiased GNN mechanism is tailor-made for a particular bias. For instance, tailored specifically for popularity bias, they randomly sample interaction subgraphs and alter the information propagation scheme, but might fail in other biases. Here we argue that, for graph-based CF models, effective debiasing should comprehensively address both biases present in interaction data and those amplified by the GNN mechanism.

Method

Adversarial Graph Dropout (AdvDrop)

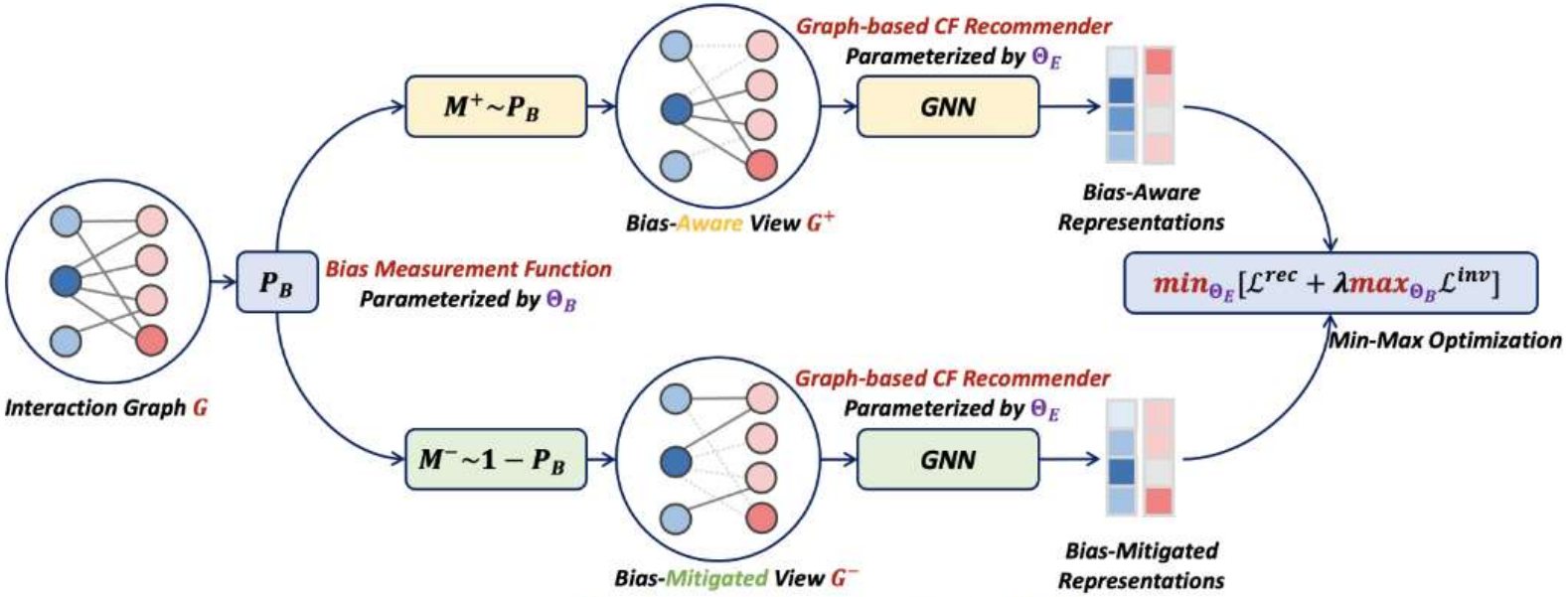


Figure 3: The overall framework of AdvDrop.

Method

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Stage 1: Debiased Representation Learning

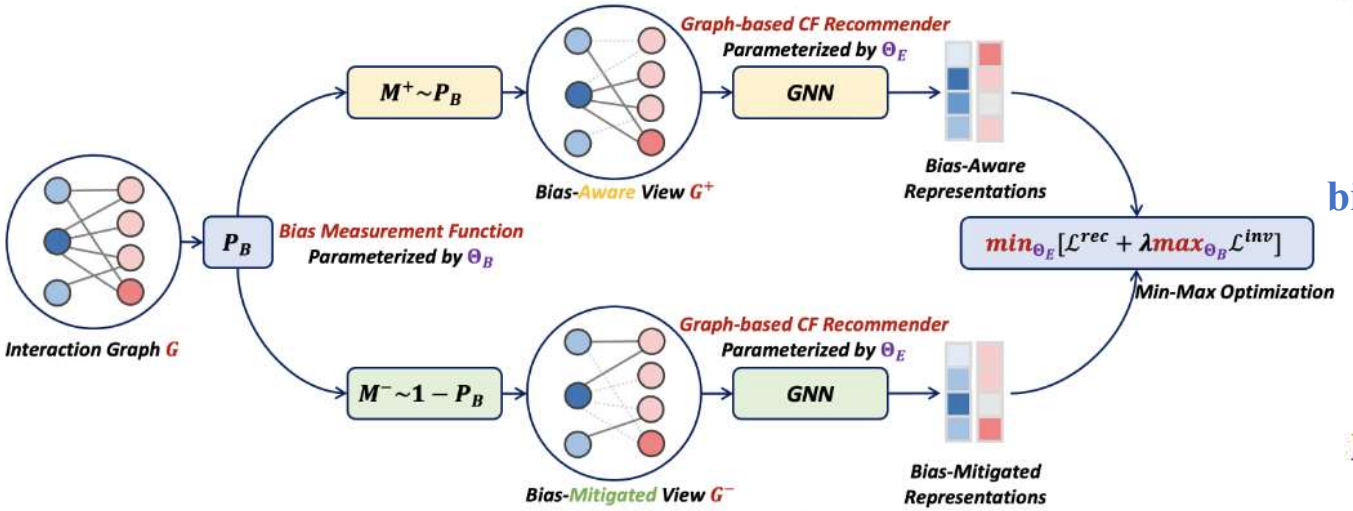


Figure 3: The overall framework of AdvDrop.

the level of bias:

$$b_{ui} = P_B(u,i). \in [0,1]$$

- 0, genuine preference
- 1, biased

bias-aware G^+ and bias-mitigated G^-:

$$\mathbf{A}^+ = \mathbf{A} \odot \mathbf{M}^+,$$
$$\mathbf{A}^- = \mathbf{A} \odot \mathbf{M}^-,$$

$$\mathbf{M}^+ \sim P_B(\mathcal{G}^+) = \prod_{\{(u,i)|A_{ui}=1\}} \text{Bern}(m^+_{ui}; P_B(u,i)),$$

$$\mathbf{M}^- \sim P_B(\mathcal{G}^-) = \prod_{\{(u,i)|A_{ui}=1\}} \text{Bern}(m^-_{ui}; 1 - P_B(u,i)).$$

Method

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Stage 1: Debiased Representation Learning

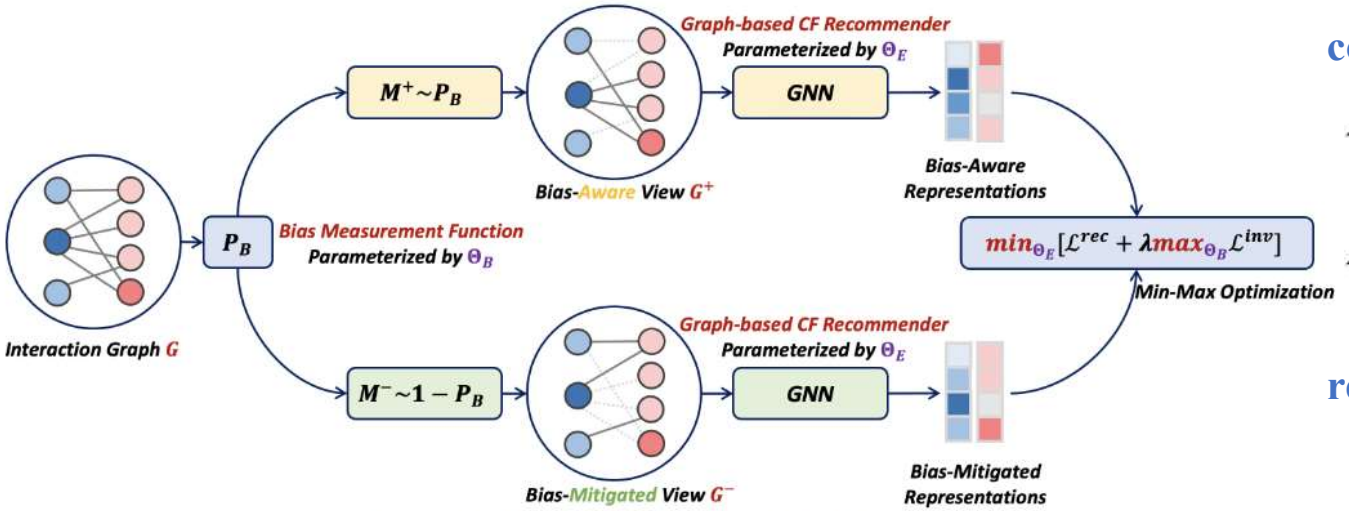


Figure 3: The overall framework of AdvDrop.

view centric representations

$$Z_U^+, Z_I^+ = \text{GNN}(\mathcal{G}^+|\Theta_E),$$
$$Z_U^-, Z_I^- = \text{GNN}(\mathcal{G}^-|\Theta_E).$$

contrastive loss

$$\mathcal{L}_U^{\text{inv}} = \sum_{u \in \mathcal{U}} -\log \frac{\exp(s(z_u^+, z_u^-/\tau))}{\sum_{u' \in \mathcal{U}} \exp(s(z_u^+, z_{u'}^-))/\tau))},$$

$$\mathcal{L}_I^{\text{inv}} = \sum_{i \in \mathcal{I}} -\log \frac{\exp(s(z_i^+, z_i^-/\tau))}{\sum_{i' \in \mathcal{I}} \exp(s(z_i^+, z_{i'}^-))/\tau))},$$

recommendation loss

$$\mathcal{L}^{\text{rec}} = \mathcal{L}_{\text{BPR}}^+ + \mathcal{L}_{\text{BPR}}^-$$

Stage1 objective: $\min_{\Theta_E} \mathcal{L}^{\text{rec}} + \lambda \mathcal{L}^{\text{inv}},$

Method

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Stage 2: Bias Identification

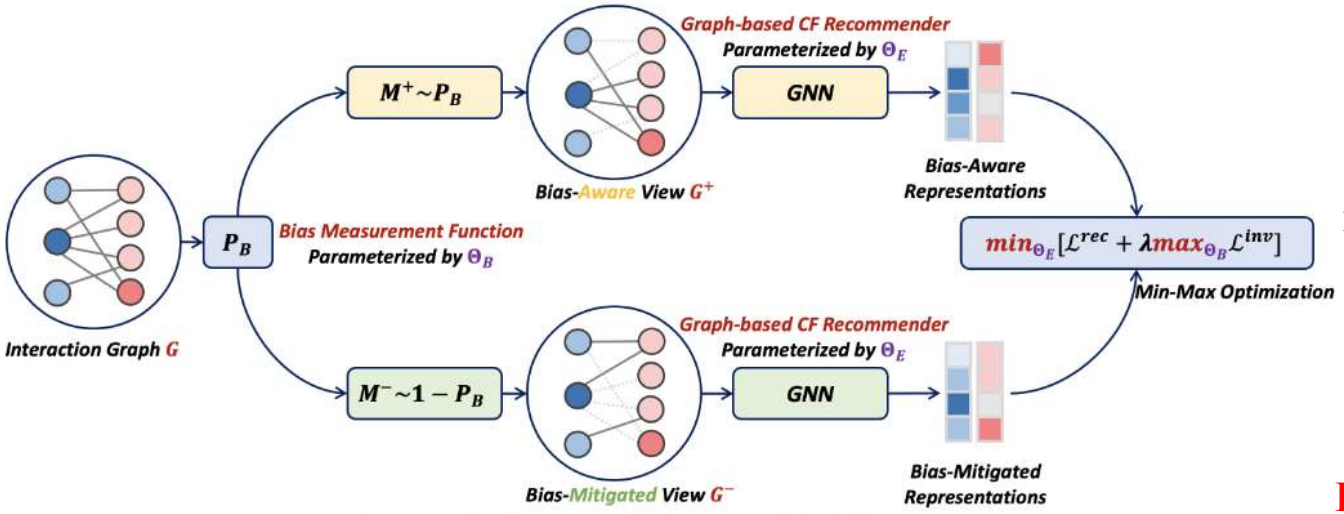


Figure 3: The overall framework of AdvDrop.

measurement function

$$P_B(u,i|\Theta_B) = \sigma(f_B(\mathbf{Z}|\Theta_B)) = \sigma(\mathbf{W}_B[\mathbf{z}_u^{(0)}||\mathbf{z}_i^{(0)}] + b_B),$$

$$\Theta_B = (\mathbf{W}_B, b_B)$$

maximize the contrastive loss(stage2 objective)

$$\max_{\Theta_B} \mathcal{L}^{inv}.$$

Final objective: $\min_{\Theta_E}[\mathcal{L}^{rec} + \lambda \max_{\Theta_B} \mathcal{L}^{inv}].$

Method

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Algorithm 1 AdvDrop Algorithm for General Debias

Input: Training dataset $\mathcal{D}$ consist of (u, v) pairs, the number of stage 1 epochs K_{stage1}, and number of stage 2 epochs K_{stage2}.
Output: Debiased user representations $\mathbf{Z}_U$ and debiased item representations $\mathbf{Z}_I$.
Initialize: initialize Θ_E and Θ_B.
while not converged **do**
  Stage 1: Debiased Representation Learning:
  Fix bias measurement function parameters Θ_B.
  for $k_1 \leq K_{stage1}$ **do**
    Compute $\mathcal{L}^{rec}$ and $\mathcal{L}^{inv}$ by Equations (7) and (13).
    Update Θ_E by Equation (14).
    $k_1 \leftarrow k_1 + 1$
  end for
  Stage 2: Bias Identification:
  Fix graph neural networks parameters Θ_E.
  for $k_2 \leq K_{stage2}$ **do**
    Recompute P_B and resample $\mathcal{G}_-$ and $\mathcal{G}_+$.
    Compute $\nabla_{f_B}\mathcal{L}^{inv}$ by Equation (19) and Corollary 3.2.
    Obtain gradients by back-propagation and update Θ_B.
    $k_2 \leftarrow k_2 + 1$
  end for
end while
Compute $\mathbf{Z}_U$ and $\mathbf{Z}_I$ by Equation (20).
return $\mathbf{Z}_U$ and $\mathbf{Z}_I$.

Experiment

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

RQ1: How does Advdrop perform compared with other baseline models on general debiasing datasets?

Table 1: General debiasing performance on Coat, Yahoo, KuaiRec, and Douban. The improvements achieved by AdvDrop are statistically significant (p-value $\ll 0.05$).

	Coat		Yahoo		KuaiRec		Douban	
	NDCG@3	Recall@3	NDCG@3	Recall@3	NDCG@20	Recall@20	NDCG@20	Recall@20
LightGCN	0.499	0.394	0.610	0.640	0.334	0.073	0.059	0.030
IPS-CN	0.516 +3.41%	0.406 +3.05%	0.598 −1.97%	0.628 −1.88%	0.014 −95.81%	0.002 −97.26%	0.056 −5.08%	0.031 +3.33%
DR	0.506 +1.40%	0.416 +5.58%	0.611 +0.16%	0.637 −0.47%	0.037 −88.92%	0.010 −86.30%	0.060 +1.69%	0.033 +10.00%
CVIB	0.488 −2.20%	0.386 −2.03%	0.597 −2.13%	0.632 −1.25%	0.342 +2.40%	0.079 +8.22%	0.062 +5.08%	0.032 +6.67%
InvPref	0.365 −26.85%	0.293 −25.63%	0.594 −2.62%	0.621 −2.97%	-	-	0.060 +1.69%	0.029 −3.33%
AutoDebias	0.502 +0.60%	0.401 +1.78%	0.601 −1.48%	0.627 −2.03%	0.327 −2.10%	0.072 −1.37%	0.058 −1.69%	0.030 0.00%
AdvDrop	**0.532*** +6.61%	**0.418*** +6.09%	**0.617*** +1.15%	**0.643*** +0.47%	**0.362*** +8.38%	**0.089*** +21.92%	**0.068*** +15.25%	**0.036*** +20.00%

Experiment

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

RQ2: Can Advdrop successfully address various specific biases?

Table 3: Performance of mitigating popularity bias on Yelp2018.

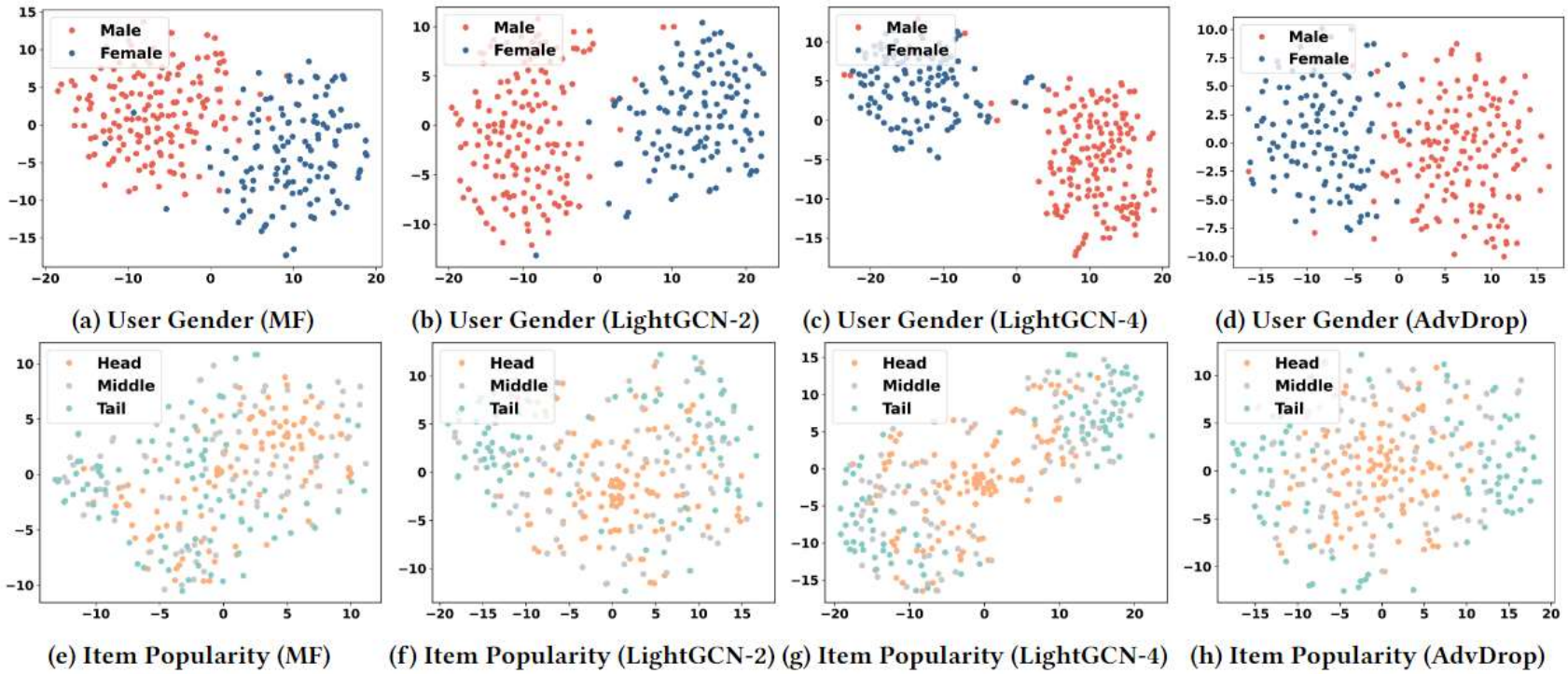
Test Split	Test ID		Test OOD	
Metrics	NDCG@20	Recall@20	NDCG@20	Recall@20
LightGCN	0.0371	0.0527	0.0028	0.0026
IPS-CN	0.0337	0.0470	0.0033	0.0030
CDAN	0.0496	0.0703	0.0037	0.0037
sDRO	0.0492	0.0702	0.0035	0.0034
AdvDrop	**0.0608**	**0.0817**	**0.0066**	**0.0073**

Table 2: Performance of mitigating attribute unfairness on Coat.

Evaluation Metrics	Performance Metrics				Fairness Metric (Prediction Bias)		
	NDCG@3	NDCG@5	Recall@3	Recall@5	user gender	item colour	item gender
MF	0.473	0.508	0.349	0.501	0.102	0.175	0.091
LightGCN	0.499	0.523	0.394	0.519	0.420	0.589	0.468
AdvDrop	**0.532**	0.553	0.418	0.540	0.142	0.053	0.062
+Embed info	0.518	**0.560**	0.418	0.578	0.052	0.032	0.046
+Mask info	0.512	0.554	**0.425**	**0.581**	0.045	0.024	0.039
CFC	0.485	0.513	0.395	0.517	**0.012**	**0.011**	**0.012**

Experiment

南京航空航天大學
Nanjing University of Aeronautics and Astronautics



(a) User Gender (MF) (b) User Gender (LightGCN-2) (c) User Gender (LightGCN-4) (d) User Gender (AdvDrop)

(e) Item Popularity (MF) (f) Item Popularity (LightGCN-2) (g) Item Popularity (LightGCN-4) (h) Item Popularity (AdvDrop)

Figure 1: T-SNE [46] visualizations of user and item representations learned by MF [40], LightGCN [26], and our proposed AdvDrop. Note that MF, LightGCN-2, and LightGCN-4 are specialized with zero, two, and four graph convolutional layers, respectively. Subfigures 1a-1d show the representation distribution *w.r.t.* two groups of user gender (*i.e.,* female, male), while Subfigures 1e-1h depict the representation distribution *w.r.t.* three groups of item popularity (*i.e.,* head, middle, tail).

Experiment

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

RQ3: Within AdvDrop, what pivotal insights does the adversarial learning framework extract, and how do these influence the learned representations?

Table 4: Ablation study of AdvDrop on Yelp2018.

Test Split	Test ID		Test OOD	
Metrics	NDCG@20	Recall@20	NDCG@20	Recall@20
AdvDrop	**0.0608**	**0.0817**	**0.0066**	**0.0073**
w/o P_B	<u>0.0575</u>	<u>0.0781</u>	<u>0.0060</u>	<u>0.0060</u>
w/o P_B & $\mathcal{L}_{inv}$	0.0371	0.0528	0.0027	0.0024

Experiment

RQ3: Within AdvDrop, what pivotal insights does the adversarial learning framework extract, and how do these influence the learned representations?

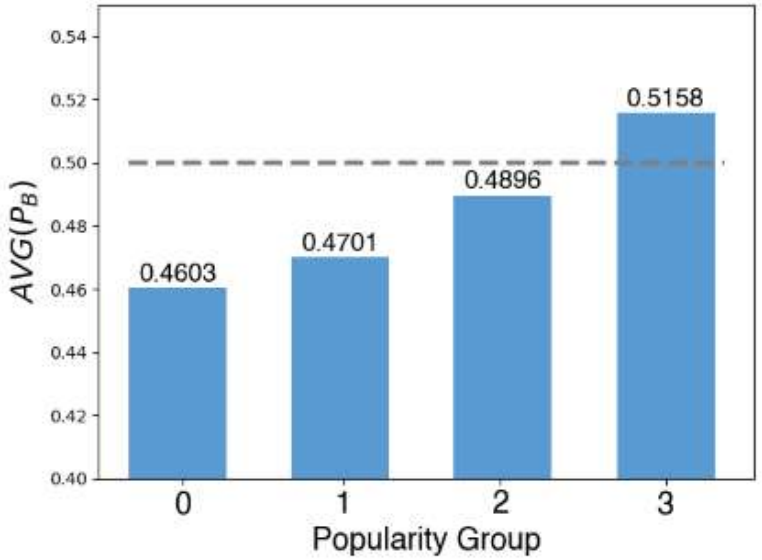
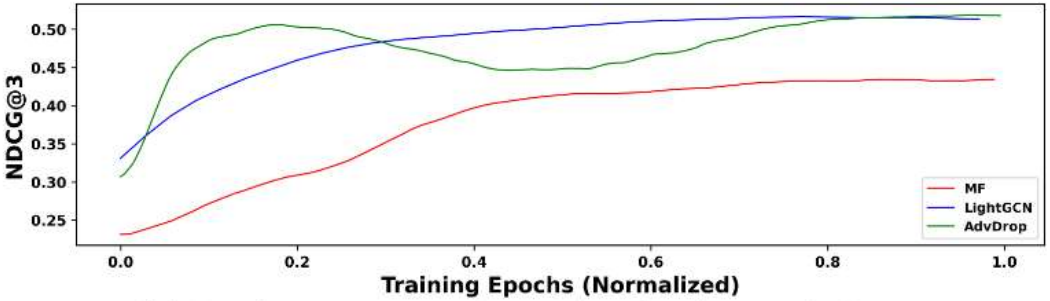


Figure 4: Visualization of learned bias measurement function P_B *w.r.t.* item popularity.

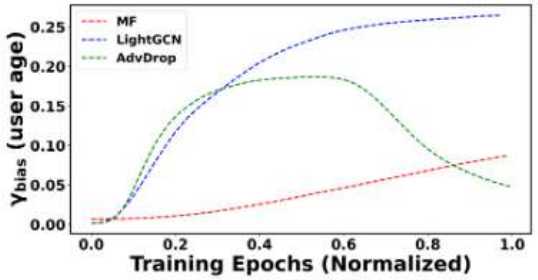
Experiment

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

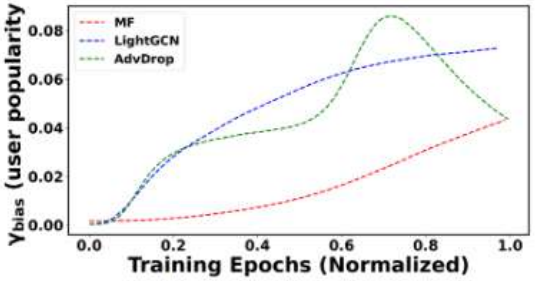
RQ3: Within AdvDrop, what pivotal insights does the adversarial learning framework extract, and how do these influence the learned representations?



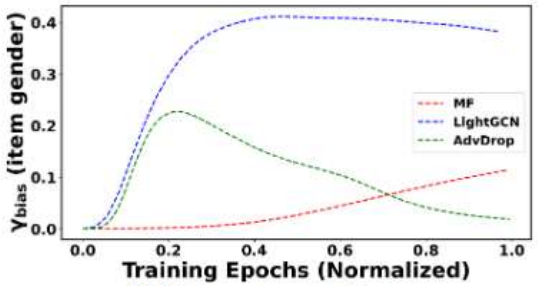
(a) Performance trajectories during training



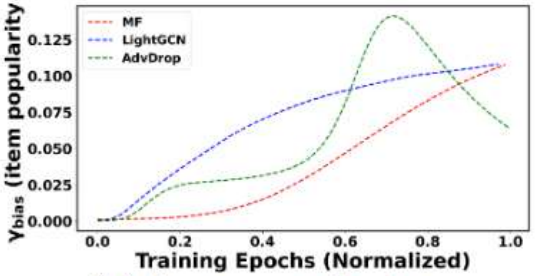
(b) User age



(c) User activities



(d) Item gender



(e) Item popularity

Figure 5: (a) The overall recommendation performance v.s. epochs during training. (b) ~ (e) The debiasing performance *w.r.t.* epochs on both user and item attributes.