

# Deconfounded Recommendation for Alleviating Bias Amplification

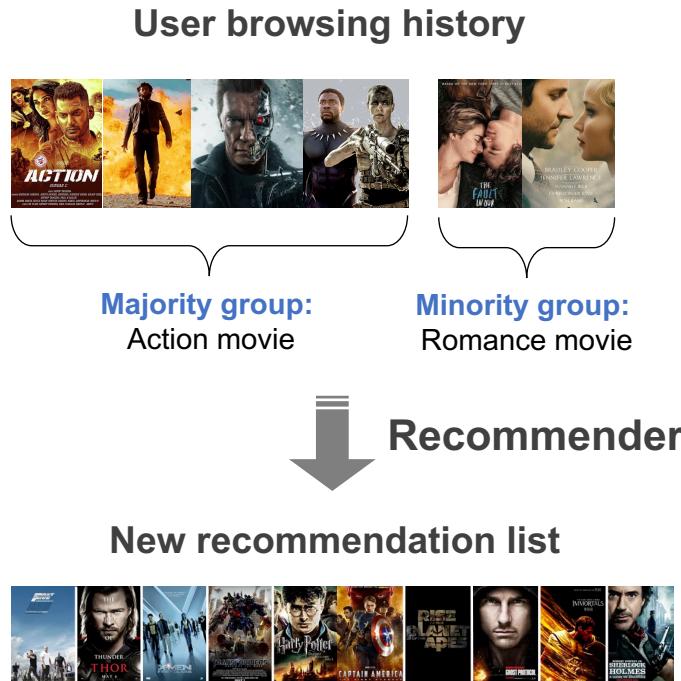
Wenjie Wang<sup>1</sup>, Fuli Feng<sup>12\*</sup>, Xiangnan He<sup>3</sup>, Xiang Wang<sup>12</sup>, and Tat-Seng Chua<sup>1</sup>

<sup>1</sup>National University of Singapore, <sup>2</sup>Sea-NExT Joint Lab, <sup>3</sup>University of Science and Technology of China  
`{wenjiewang96,fulifeng93,xiangnanhe}@gmail.com, xiangwang@u.nus.edu, dcscts@nus.edu.sg`

**Speaker: Fuli Feng**  
Jun 2021

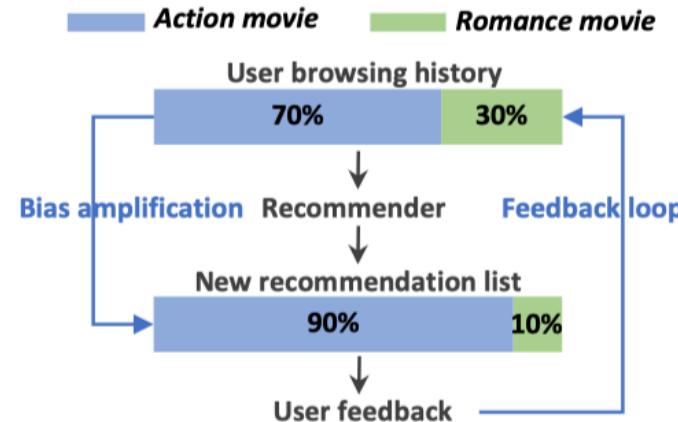
- 1. Bias amplification**
- 2. Related work**
- 3. Deconfounded RS**
- 4. Experiments**
- 5. Conclusion**

- Bias amplification: over-recommend items in the majority group



- More action movies!

- Bias is continually amplified due to the feedback loop.



(a) An example of bias amplification.

## • Problems:

1. Low-diversity: limit users' view and narrow down user interests.
2. Possible reason of Filter bubbles and echo chambers.
3. Unfairness: unfair to the high-quality items in minority groups.

## 1. Fairness

- Pursue **equal exposure opportunities** for items of different groups.
- e.g., discounted cumulative fairness (*Yang et al. 2017*), and fairness of exposure (*Singh et al. 2018*).

## 2. Diversity

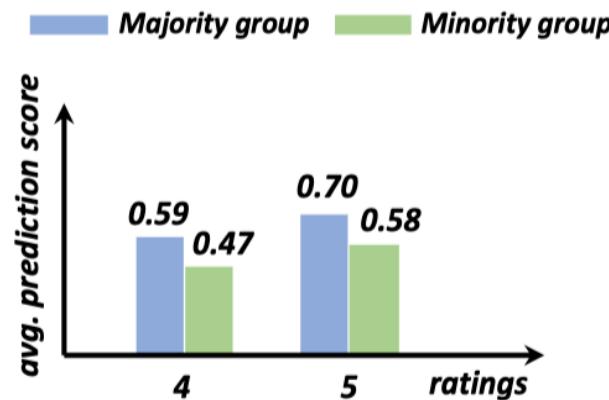
- **Decrease the similarity** of the recommended items
- e.g., re-ranking via the intra-list similarity (*Ziegler et al. 2005*).

## 3. Calibration

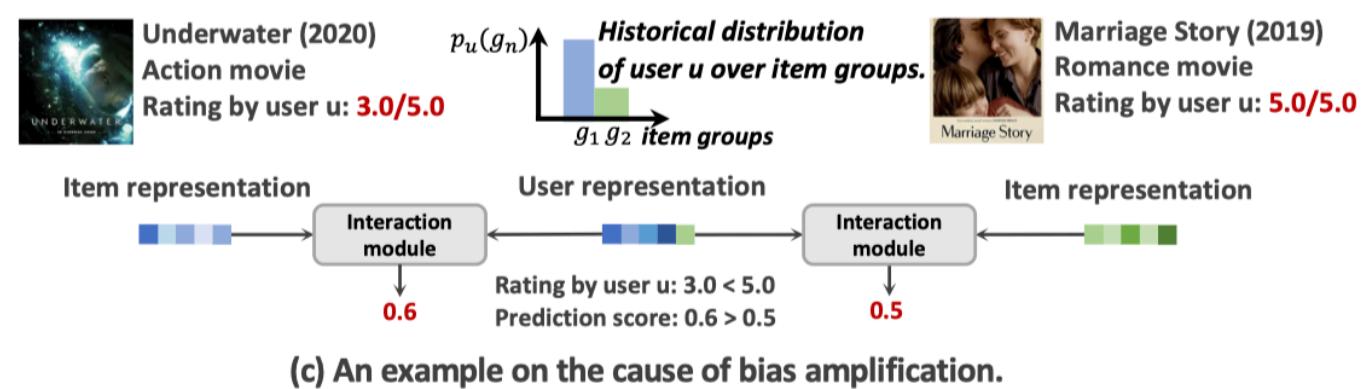
- Encourage **the distribution** of recommended item groups to **follow that of the browsing history**.
- e.g., calibrated recommendation: re-ranking based on KL-divergence (*Steck et al. 2018*).
- **Common drawback:** inevitably **sacrifice recommendation accuracy**.

- *Yang et al. 2017. Measuring fairness in ranked outputs. In SSDBM.*
- *Singh et al. 2018. Fairness of exposure in rankings. In KDD.*
- *Steck et al. 2018. Calibrated recommendations. In RecSys.*
- *Ziegler et al. 2005. Improving recommendation lists through topic diversification. In WWW.*

- What is the **root reason** for bias amplification?
- An example of bias amplification.

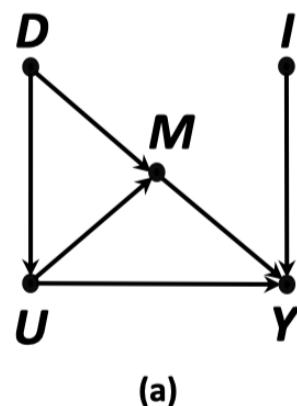


(b) Prediction score difference between the items in the majority and minority groups over ML-1M.



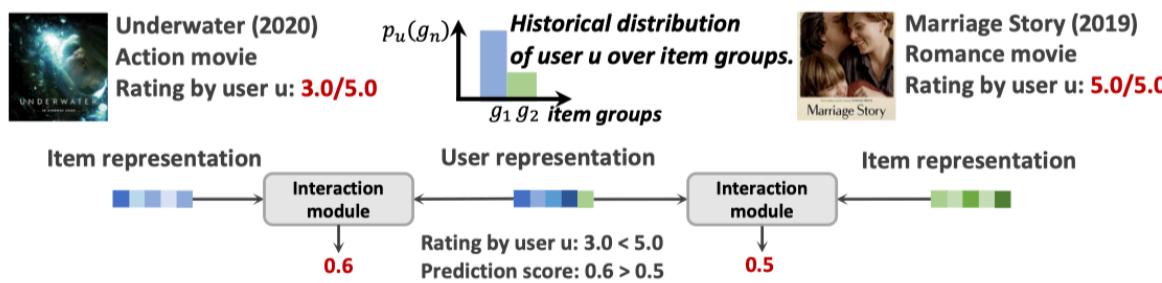
- An item with **low rating receives a higher prediction score** because it belongs to the majority group.
- Intuitively, we can know that the user representation **shows stronger preference** to majority group.

- A Causal View of Bias Amplification**



**U** *User representation*  
**I** *Item representation*  
**D** *User historical distribution over item groups*  
**M** *Group-level user representation*  
**Y** *Prediction score*

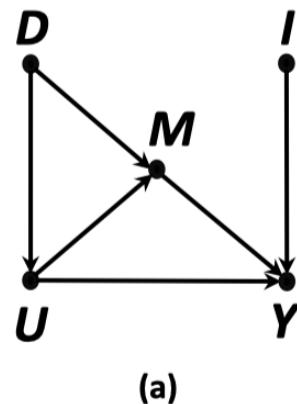
(a)



(c) An example on the cause of bias amplification.

- $D$ : user historical distribution over item groups. Given  $N$  item groups  $\{g_1, \dots, g_N\}$ ,  $d_u = [p_u(g_1), \dots, p_u(g_N)] \in R^N$  is a particular value of  $D$ . e.g.,  $d_u = [0.8, 0.2]$ .
- Use  $M$  to describe how much the **user likes different item groups**; decided by  $D$  and  $U$ .
- The prediction score  $Y$  is affected by  $U$  and  $M$ , implying that:  
an item  $i$  can have a high prediction score because 1) **user's pure preference over the item** ( $U \rightarrow Y$ ) or 2) the **user shows interest in the item group** ( $U \rightarrow M \rightarrow Y$ ).
- $M$  is a confounder between  $U$  and  $Y$ : opens the backdoor path ( $U \leftarrow D \rightarrow M \rightarrow Y$ ).
- Cause **spurious correlation** when estimating the effect of  $U$  on  $Y$ : given the item  $i$  in a group  $g$ , the more items in group  $g$  the user  $u$  has clicked in the history, the higher the prediction score  $Y$  becomes.
- i.e., the high prediction scores are purely caused by the users' historical interest in the group instead of the items themselves.

- A Causal View of Bias Amplification



**U** User representation  
**I** Item representation  
**D** User historical distribution over item groups  
**M** Group-level user representation  
**Y** Prediction score

$$P(Y|U = \mathbf{u}, I = \mathbf{i})$$

$$= \frac{\sum_{\mathbf{d} \in \mathcal{D}} \sum_{\mathbf{m} \in \mathcal{M}} P(\mathbf{d}) P(\mathbf{u}|\mathbf{d}) P(\mathbf{m}|\mathbf{d}, \mathbf{u}) P(\mathbf{i}) P(Y|\mathbf{u}, \mathbf{i}, \mathbf{m})}{P(\mathbf{u}) P(\mathbf{i})} \quad (1a)$$

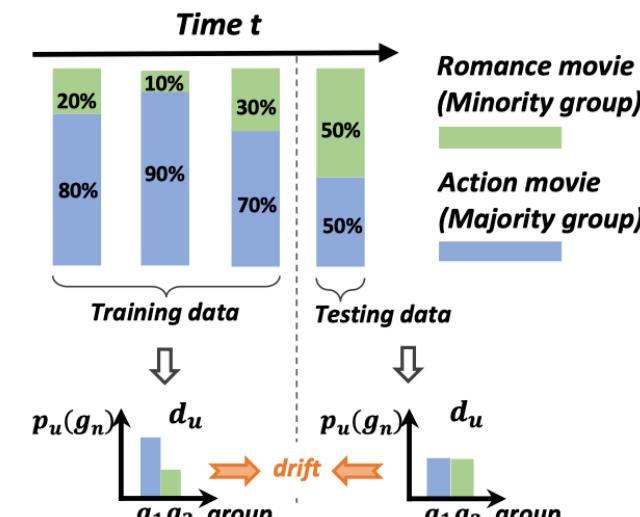
$$= \sum_{\mathbf{d} \in \mathcal{D}} \sum_{\mathbf{m} \in \mathcal{M}} P(\mathbf{d}|\mathbf{u}) P(\mathbf{m}|\mathbf{d}, \mathbf{u}) P(Y|\mathbf{u}, \mathbf{i}, \mathbf{m}) \quad (1b)$$

$$= \sum_{\mathbf{d} \in \mathcal{D}} P(\mathbf{d}|\mathbf{u}) P(Y|\mathbf{u}, \mathbf{i}, M(\mathbf{d}, \mathbf{u})) \quad (1c)$$

$$= P(\mathbf{d}_u|\mathbf{u}) P(Y|\mathbf{u}, \mathbf{i}, M(\mathbf{d}_u, \mathbf{u})), \quad (1d)$$

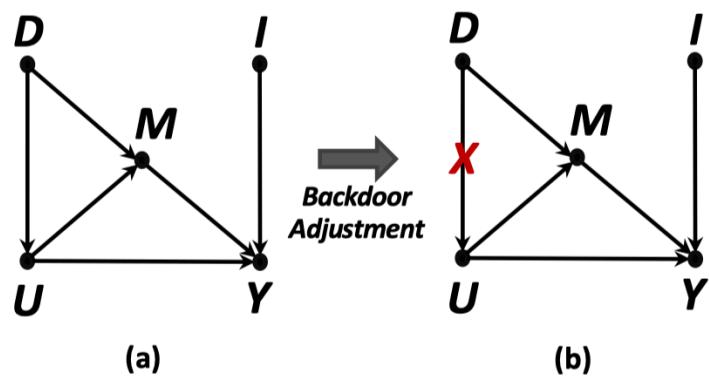
- Impact of the spurious correlation:

- 1) **Bias amplification**: the items in the majority group, even including the low-quality ones, are easy to have high ranks.
- 2) **User interest drift**. The user representation heavily relies on the user historical distribution over item groups, e.g.,  $d_u = [0.8, 0.2]$ . Once users' future interest in item groups changes (i.e., OOD settings), the recommendations will be dissatisfying.



(a) User interest is changing over time. (b)

## • Backdoor Adjustment



- ## • Deconfounded Recommender System (DecRS)

- Use backdoor adjustment to achieve  $P(Y|do(U=\mathbf{u}), I=\mathbf{i})$  where  $do(U=\mathbf{u})$  can be intuitively seen as cutting off the edge  $D \rightarrow U$  in the causal graph and blocking the effect of  $D$  on  $U$ .
  - DecRS estimates the prediction score  $Y$  by considering every possible value  $d$  of  $D$  subject to the prior  $P(d)$ , rather than the only  $d_u$  in Eq. (1d).
  - The items in the majority group will not receive high prediction scores purely because of a high click probability in  $d_u$ . => **alleviate bias amplification**.

$$P(Y|U=\mathbf{u}, I=\mathbf{i}) = \frac{\sum_{\mathbf{d} \in \mathcal{D}} \sum_{\mathbf{m} \in \mathcal{M}} P(\mathbf{d})P(\mathbf{u}|\mathbf{d})P(\mathbf{m}|\mathbf{d}, \mathbf{u})P(\mathbf{i})P(Y|\mathbf{u}, \mathbf{i}, \mathbf{m})}{P(\mathbf{u})P(\mathbf{i})} \quad (1a)$$

$$= \sum_{\mathbf{d} \in \mathcal{D}} \sum_{\mathbf{m} \in \mathcal{M}} P(\mathbf{d}|\mathbf{u})P(\mathbf{m}|\mathbf{d}, \mathbf{u})P(Y|\mathbf{u}, \mathbf{i}, \mathbf{m}) \quad (1b)$$

$$= \sum P(\mathbf{d}|\mathbf{u})P(Y|\mathbf{u}, \mathbf{i}, M(\mathbf{d}, \mathbf{u})) \quad (1c)$$

$$\equiv P(\mathbf{d}_{\text{obs}} | \mathbf{u}) P(Y | \mathbf{u}, \mathbf{i}, M(\mathbf{d}_{\text{obs}}, \mathbf{u})). \quad (1d)$$

## Conditional probability

$$\begin{aligned}
 & P(Y|do(U = \mathbf{u}), I = \mathbf{i}) && \text{law of total probability \& Bayes rule} \\
 &= \sum_{\mathbf{d} \in \mathcal{D}} P(\mathbf{d}|do(U = \mathbf{u}))P(Y|do(U = \mathbf{u}), \mathbf{i}, M(\mathbf{d}, do(U = \mathbf{u}))) && (2a) \\
 & && \text{insertion/deletion of actions} \\
 &= \sum_{\mathbf{d} \in \mathcal{D}} P(\mathbf{d})P(Y|do(U = \mathbf{u}), \mathbf{i}, M(\mathbf{d}, do(U = \mathbf{u}))) && (2b) \\
 &\equiv \sum_{\mathbf{d} \in \mathcal{D}} \boxed{P(\mathbf{d})}P(Y|\mathbf{u}, \mathbf{i}, M(\mathbf{d}, \mathbf{u})), && \text{action/observation exchange} && (2c)
 \end{aligned}$$

Force U, I to incorporate every d fairly during training.  
Cut off the relation between D and U.

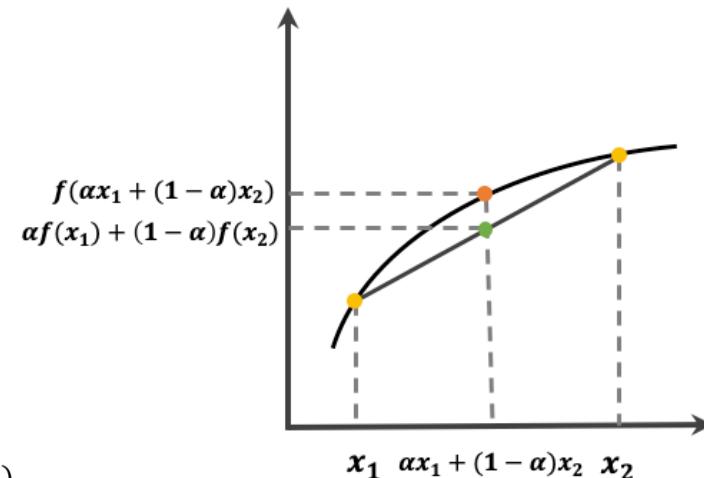
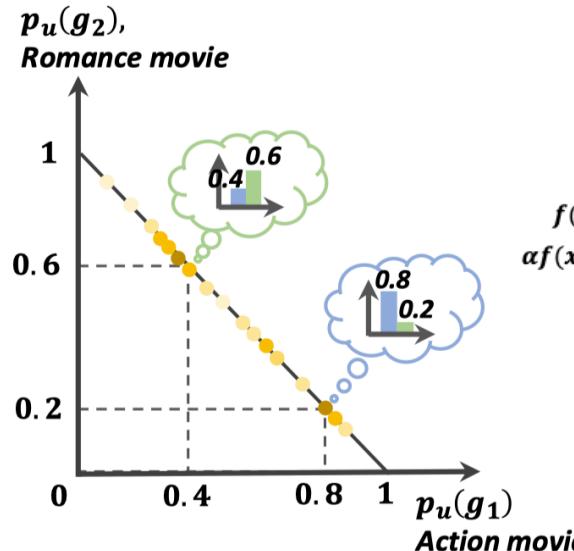
- Backdoor Adjustment

$$P(Y|do(U = \mathbf{u}), I = \mathbf{i})$$

$$= \sum_{\mathbf{d} \in \mathcal{D}} P(\mathbf{d}|do(U = \mathbf{u}))P(Y|do(U = \mathbf{u}), \mathbf{i}, M(\mathbf{d}, do(U = \mathbf{u}))) \quad (2a)$$

$$= \sum_{\mathbf{d} \in \mathcal{D}} P(\mathbf{d})P(Y|do(U = \mathbf{u}), \mathbf{i}, M(\mathbf{d}, do(U = \mathbf{u}))) \quad (2b)$$

$$= \sum_{\mathbf{d} \in \mathcal{D}} P(\mathbf{d})P(Y|\mathbf{u}, \mathbf{i}, M(\mathbf{d}, \mathbf{u})), \quad (2c)$$



- Deconfounded Recommender System (DecRS)

- Challenge:** the sample space of  $D$  is infinite.

- Backdoor Adjustment Approximation:

- Sample users' historical distributions over item groups in the training data to estimate the distribution of  $D$ ; Use function  $f(\cdot)$  (FM) to calculate  $P(Y|\mathbf{u}, \mathbf{i}, M(\mathbf{d}, \mathbf{u}))$ .

$$\begin{aligned} P(Y|do(U = \mathbf{u}), I = \mathbf{i}) &\approx \sum_{\mathbf{d} \in \tilde{\mathcal{D}}} P(\mathbf{d})P(Y|\mathbf{u}, \mathbf{i}, M(\mathbf{d}, \mathbf{u})) \\ &= \sum_{\mathbf{d} \in \tilde{\mathcal{D}}} P(\mathbf{d})f(\mathbf{u}, \mathbf{i}, M(\mathbf{d}, \mathbf{u})), \end{aligned} \quad (4)$$

- Approximation of  $E_d[f(\cdot)]$ .

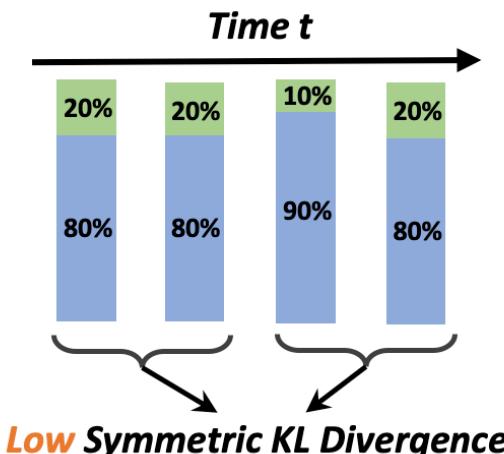
- Expectation of function  $f(\cdot)$  of  $\mathbf{d}$  in Eq. 4 is hard to compute because we need to calculate the results of  $f(\cdot)$  for each  $\mathbf{d}$ .
- Jensen's inequality:** take the sum into the function  $f(\cdot)$ .

$$P(Y|do(U = \mathbf{u}), I = \mathbf{i}) \approx f(\mathbf{u}, \mathbf{i}, M(\sum_{\mathbf{d} \in \tilde{\mathcal{D}}} P(\mathbf{d})\mathbf{d}, \mathbf{u})). \quad (5)$$

- **Inference Strategy**
  - Some users might only like several majority groups, i.e., enjoy the over-recommending of majority groups.
  - A **user-specific inference strategy** to regulate the impact of backdoor adjustment dynamically.
    - 1) Divide the historical interactions into two parts according to the timestamps.
    - 2) Calculate the symmetric KL divergence to measure stability.
    - 3) Use KL divergence to balance  $P(Y|U=\mathbf{u}, I=\mathbf{i})$  and  $P(Y|do(U=\mathbf{u}), I=\mathbf{i})$ .
  - Users with low KL divergence will rely more on the conditional probability.

$$\begin{aligned}\eta_u &= KL(\mathbf{d}_u^1 \mid \mathbf{d}_u^2) + KL(\mathbf{d}_u^2 \mid \mathbf{d}_u^1) \\ &= \sum_{n=1}^N P_u^1(g_n) \log \frac{P_u^1(g_n)}{P_u^2(g_n)} + \sum_{n=1}^N P_u^2(g_n) \log \frac{P_u^2(g_n)}{P_u^1(g_n)},\end{aligned}\quad (10)$$

$$Y_{u,i} = (1 - \hat{\eta}_u) * Y_{u,i}^{RS} + \hat{\eta}_u * Y_{u,i}^{DE}, \quad \hat{\eta}_u = \left( \frac{\eta_u - \eta_{min}}{\eta_{max} - \eta_{min}} \right)^\alpha$$



- To summarize, DecRS has three main differences from the conventional RS:
  - 1) DecRS **models the causal effect**  $P(Y|do(U = u), I = i)$  instead of the conditional probability  $P(Y|U = u, I = i)$ .
  - 2) DecRS equips the recommender models with **a backdoor adjustment operator**.
  - 3) DecRS makes recommendations with **a user-specific inference strategy** instead of the simple model prediction (e.g., a forward propagation).

## Experimental Settings

- **Datasets:** ML-1M and Amazon-Book.
- **Baselines:**
  - 1) **Unawareness** (*Kusner et al. 2017*) removes the features of item groups (e.g., movie genre in ML-1M).
  - 2) **FairCo** (*Morik et al.. 2020*) introduces one error term to control the exposure fairness across groups.
  - 3) **Calibration** (*Steck et al. 2018*) uses a calibration metric  $C_{KL}$  to re-rank items.
  - 4) **Diversity** (*Ziegler et al. 2005*) aims to decrease the intra-list similarity.
  - 5) **IPS** (*Saito et al. 2020*) is a classical causal method to reduce bias.
- **Evaluation Metrics.**
  - 1) Recall@K and NDCG@K
  - 2) **A calibration metric  $C_{kl}$**  (*Steck et al. 2018*): quantifies the distribution drift over item groups between the history and the new recommendation list (comprised by the top-20 items). Higher  $C_{kl}$  scores suggest a more serious issue of bias amplification.

- *Kusner et al. 2017. Counterfactual Fairness. In NeurIPS.*
- *Morik et al.. 2020. Controlling Fairness and Bias in Dynamic Learning-to-Rank. In SIGIR.*
- *Steck et al. 2018. Calibrated recommendations. In RecSys.*
- *Ziegler et al. 2005. Improving recommendation lists through topic diversification. In WWW.*
- *Saito et al. 2020. Unbiased Recommender Learning from Missing-Not-At-Random Implicit Feedback. In WSDM.*

# Experiments

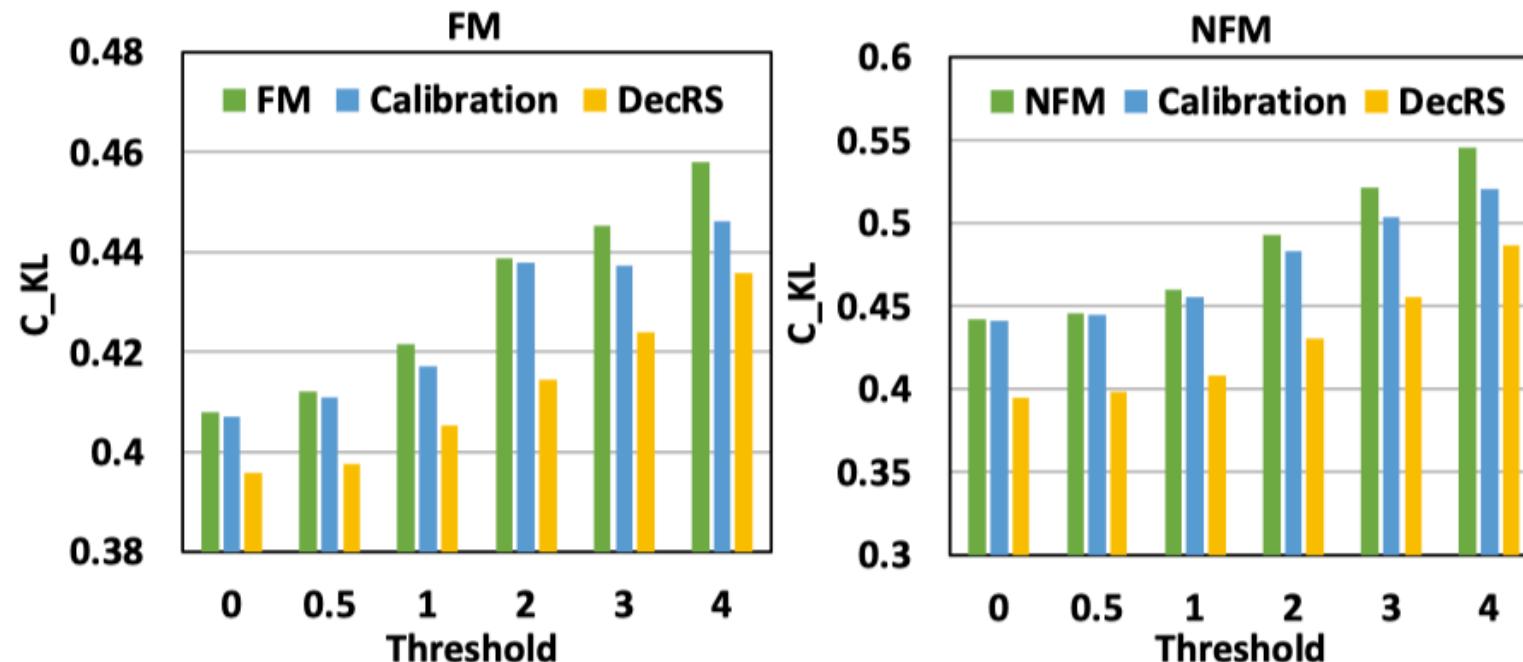
**Table 3: Overall performance comparison between DecRS and the baselines on ML-1M and Amazon-Book. %improv. denotes the relative performance improvement achieved by DecRS over FM or NFM. The best results are highlighted in bold.**

Method	FM								NFM							
	ML-1M				Amazon-Book				ML-1M				Amazon-Book			
	R@10	R@20	N@10	N@20												
FM/NFM [16, 29]	0.0676	0.1162	0.0566	0.0715	0.0213	0.0370	0.0134	0.0187	0.0659	0.1135	0.0551	0.0697	0.0222	0.0389	0.0144	0.0199
Unawareness [15]	0.0679	0.1179	0.0575	0.0730	0.0216	0.0377	0.0138	0.0191	0.0648	0.1143	0.0556	0.0708	0.0206	0.0381	0.0133	0.0190
FairCo [21]	0.0676	0.1165	0.0570	0.0720	0.0212	0.0370	0.0135	0.0188	0.0651	0.1152	0.0554	0.0708	0.0219	0.0390	0.0142	0.0199
Calibration [32]	0.0647	0.1149	0.0539	0.0695	0.0202	0.0359	0.0129	0.0181	0.0636	0.1131	0.0526	0.0682	0.0194	0.0335	0.0131	0.0178
Diversity [47]	0.0670	0.1159	0.0555	0.0706	0.0207	0.0369	0.0131	0.0185	0.0641	0.1133	0.0540	0.0693	0.0215	0.0386	0.0140	0.0197
IPS [30]	0.0663	0.1188	0.0556	0.0718	0.0213	0.0369	0.0135	0.0187	0.0648	0.1135	0.0544	0.0692	0.0213	0.0370	0.0137	0.0189
DecRS	<b>0.0704</b>	<b>0.1231</b>	<b>0.0578</b>	<b>0.0737</b>	<b>0.0231</b>	<b>0.0405</b>	<b>0.0148</b>	<b>0.0205</b>	<b>0.0694</b>	<b>0.1218</b>	<b>0.0580</b>	<b>0.0742</b>	<b>0.0236</b>	<b>0.0413</b>	<b>0.0153</b>	<b>0.0211</b>
%improv.	4.14%	5.94%	2.12%	3.08%	8.45%	9.46%	10.45%	9.63%	5.31%	7.31%	5.26%	6.46%	6.31%	6.17%	6.25%	6.03%

**Table 4: Performance comparison across different user groups on ML-1M and Amazon-Book. Each line denotes the performance over the user group with  $\eta_u >$  the threshold. We omit the results of threshold  $> 4$  due to the similar trend.**

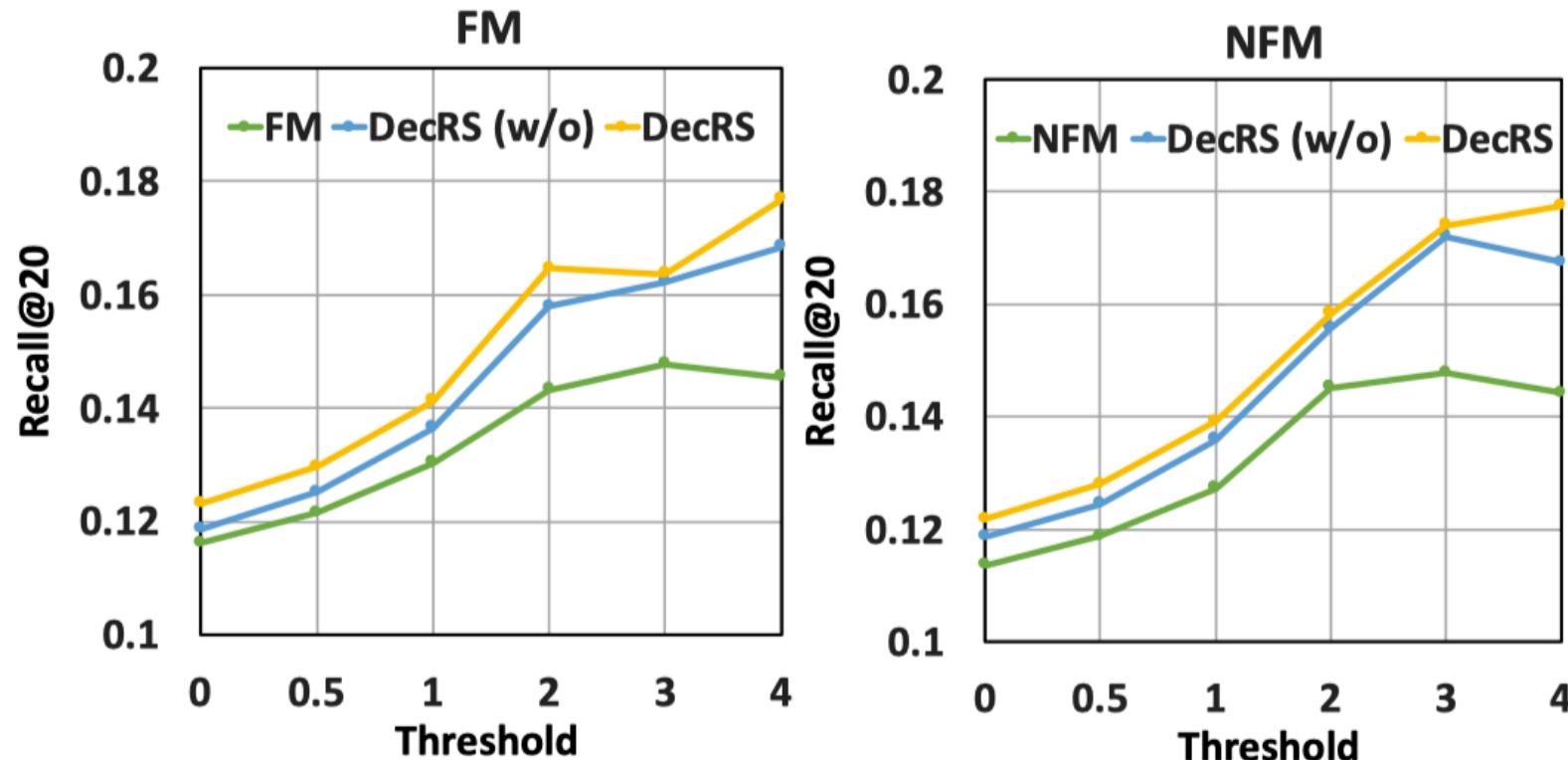
Threshold	ML-1M						Amazon-Book					
	R@20			N@20			R@20			N@20		
	FM	DecRS	%improv.	FM	DecRS	%improv.	FM	DecRS	%improv.	FM	DecRS	%improv.
0	0.1162	0.1231	5.94%	0.0715	0.0737	3.08%	0.0370	0.0405	9.46%	0.0187	0.0205	9.63%
0.5	0.1215	0.1296	6.67%	0.0704	0.0730	3.69%	0.0383	0.0424	10.70%	0.0192	0.0213	10.94%
1	0.1303	0.1412	8.37%	0.0707	0.0741	4.81%	0.0430	0.0479	11.40%	0.0208	0.0232	11.54%
2	0.1432	0.1646	14.94%	0.0706	0.0786	11.33%	0.0518	0.0595	14.86%	0.0231	0.0274	18.61%
3	0.1477	0.1637	10.83%	0.0620	0.0711	14.68%	0.0586	0.0684	16.72%	0.0256	0.0318	24.22%
4	0.1454	0.1768	21.60%	0.0595	0.0737	23.87%	0.0659	0.0793	20.33%	0.0284	0.0362	27.46%

- Effectiveness of alleviating bias amplification



**Figure 4: The performance comparison between the baselines and DecRS on alleviating bias amplification.**

- Effectiveness of the inference strategy



**Figure 5: Ablation study of DecRS on ML-1M.**

---

Thank you !

