Group-level Item Fairness

Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems

Ziwei Zhu, Jianling Wang, and James Caverlee

Texas A&M University

Group-level Item Fairness --motivation

- Most of previous works focus on measuring fairness/bias based on **score distributions** across item groups
- Most of previous works focus on **demographic parity/statistical parity** based fairness definition

Group-level Item Fairness

- Propose the **ranking-based statistical parity (RSP)** metric;
- Propose the **ranking-based equal opportunity (REO)** metric;
- Propose the **Debiased Personalized Ranking (DPR)** model;

Two metrics -- notations

$\mathcal{U} = \{1, 2, \ldots, N\}$ — A set of users.

$\mathcal{I} = \{1, 2, \ldots, M\}$ — A set of items.

$\mathcal{I}_u^+ = \{i, j, \ldots\}$ — For each user u, there is a set of items she has 'clicked' before, as training data.

$L_u = [L_{u,1}, L_{u,2}, \ldots, L_{u,K}]$ — For each user u, the RecSys provides a ranked list of K items as recommendation result.

$y_{u,i}$ — A binary variable to show whether user u likes item i in the test set (ground-truth during testing time)

$\mathcal{G} = \{g_1, g_2, \ldots, g_A\}$ — A set of group labels, each item belongs to one or more groups.

$G_{g_a}(i)$ — A function to identify whether the given item i belongs to the group g_a or not. Output '1' for yes, '0' for no.

Ranking-based Statistical Parity (RSP)

RSP measures the difference of recommendation probability (probability to be ranked in top-k) across different item groups.

$$P(rank@K|g = g_a) = \frac{\sum_{u=1}^{N} \sum_{k=1}^{K} G_{g_a}(L_{u,k})}{\sum_{u=1}^{N} \sum_{i \in \mathcal{I} \setminus \mathcal{I}_u^+} G_{g_a}(i)}$$

The probability of being ranked in top-K given the item belongs to group g_a.

$$RSP@K = \frac{std(P(rank@K|g = g_1), \ldots, P(rank@K|g = g_A))}{mean(P(rank@K|g = g_1), \ldots, P(rank@K|g = g_A))}$$

The **relative standard deviation** of the ranking probabilities across groups.

Ranking-based Statistical Parity (RSP)

Higher *RSP@K* means more severe unfairness.

RSP@K=0 (fair) when the ranking probability is the same across groups:

$$P(rank@K|g=g_1)=P(rank@K|g=g_2)=\ldots=P(rank@K|g=g_A)$$

Ranking-based Statistical Parity (RSP)

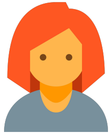
RSP is especially important when the item groups are determined by **sensitive attributes** (for example, gender or race when people are recommended) because low recommendation probability for specific sensitive groups will result in **social unfairness issues**.

RSP – motivating example

Example: Recommend job candidates to companies

$$\boldsymbol{P(recommend|\,\male\,) = 0.6}$$

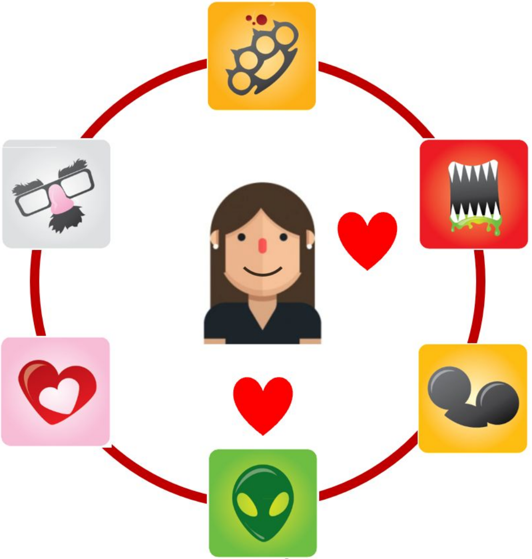
$$\boldsymbol{P(recommend|\,\female\,) = 0.2}$$



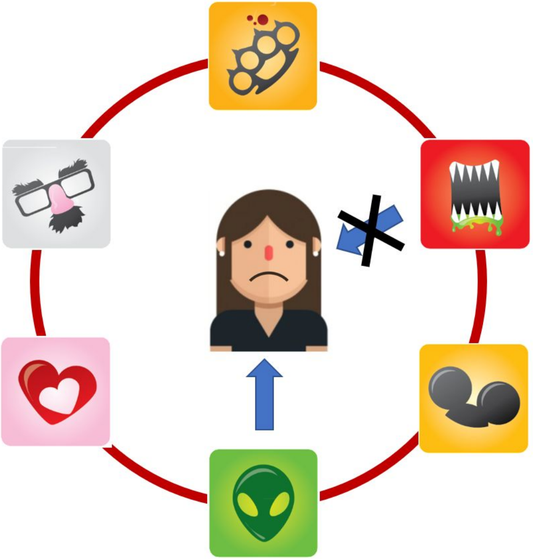
Unfair for female candidates.

Ranking-based Equal Opportunity (REO)

For a **more general RecSys**, we do not require exact the same exposure for different groups. Instead, we want the RecSys to be driven by **user preference** and the user has the same chance to see items from different groups as long as she likes them (the **same true positive rate** across item groups).



true preference



recommendation

Ranking-based Equal Opportunity (REO)

REO measures the true positive rate difference across item groups.

$$P(rank@K|g = g_a, y = 1) = \frac{\sum_{u=1}^{N}\sum_{k=1}^{K} G_{g_a}(L_{u,k}) \cdot y_{u,L_{u,k}}}{\sum_{u=1}^{N}\sum_{i \in \mathcal{I}\setminus\mathcal{I}_u^+} G_{g_a}(i) \cdot y_{u,i}}$$

The probability of being ranked in top-K given the item belongs to group g_a and is liked by a user in the test set.

$$REO@K = \frac{std(P(rank@K|g = g_1, y = 1), \ldots, P(rank@K|g = g_A, y = 1))}{mean(P(rank@K|g = g_1, y = 1), \ldots, P(rank@K|g = g_A, y = 1))}$$

REO – motivating example

Example: Recommend movies to users

$$p(recommend|horror\&liked) = 0.3$$

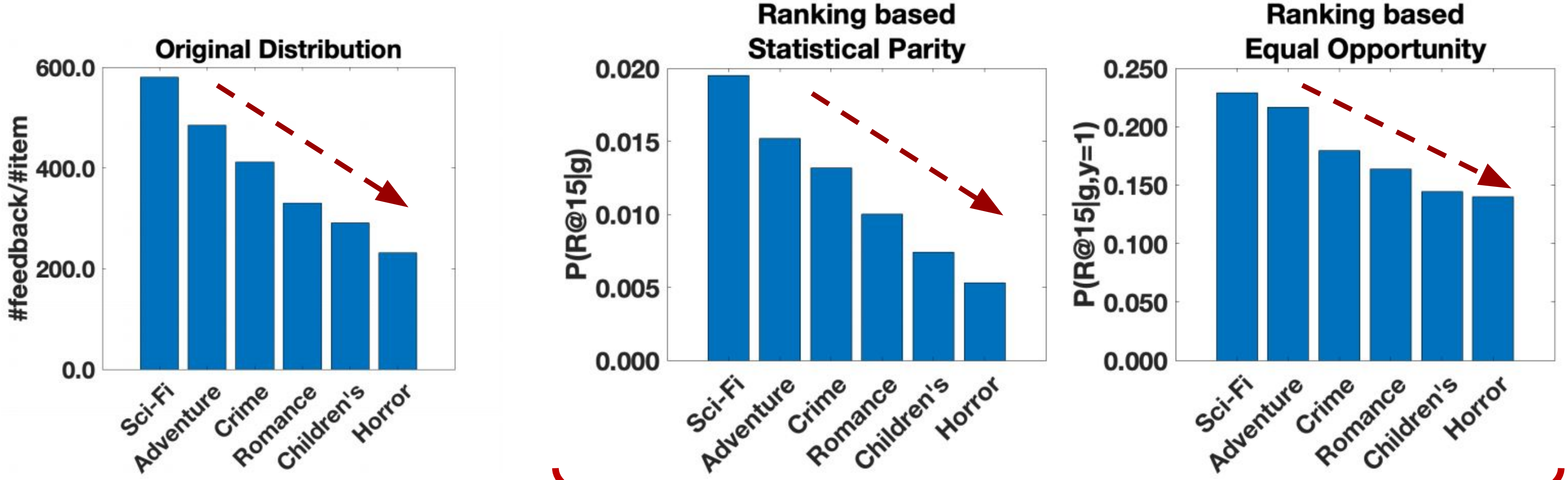
horror and sci-fi movies lover

$$p(recommend|sci-fi\&liked) = 0.9$$

In long term, horror movies will get **fewer and fewer feedback**, which is harmful for both horror movie lovers and movies providers.

Data-driven study - MovieLens

BPR generates RSP and REO based bias



Results by Bayesian Personalized Ranking (BPR)

Debiased Personalized Ranking (DPR) Model

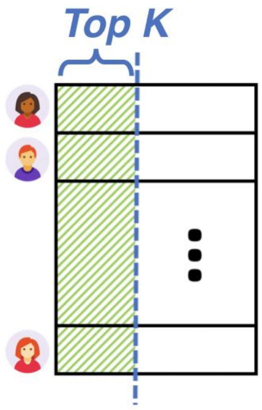
- An **adversarial learning** based method;
- An **in-processing** method, but is not coupled with any specific RecSys model;
- Flexible to be used to mitigate **RSP or REO based bias**;
- Can work for **multi-group** case.

Debiased Personalized Ranking (DPR) Model -- RSP

To mitigate RSP based bias:

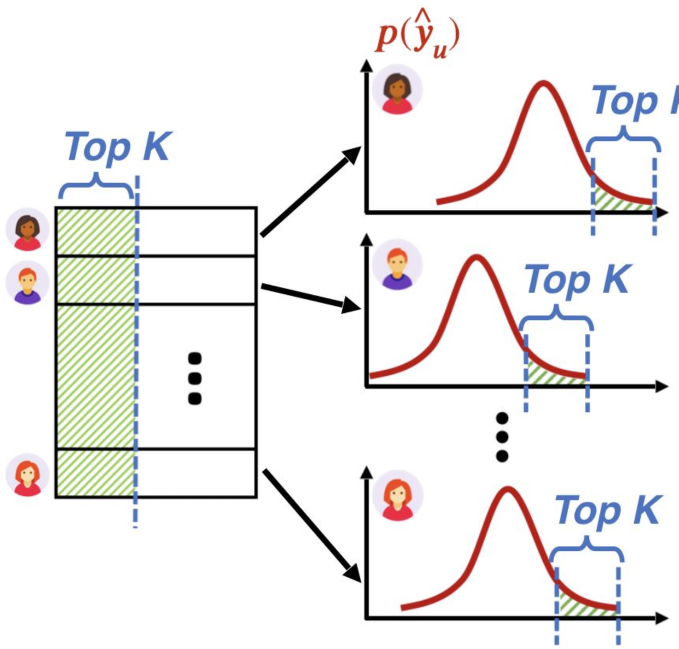
- Decouple the predicted score with group attribute;
- Normalize the score distribution for each user so that every user has the same score distribution;

Debiased Personalized Ranking (DPR) Model -- RSP



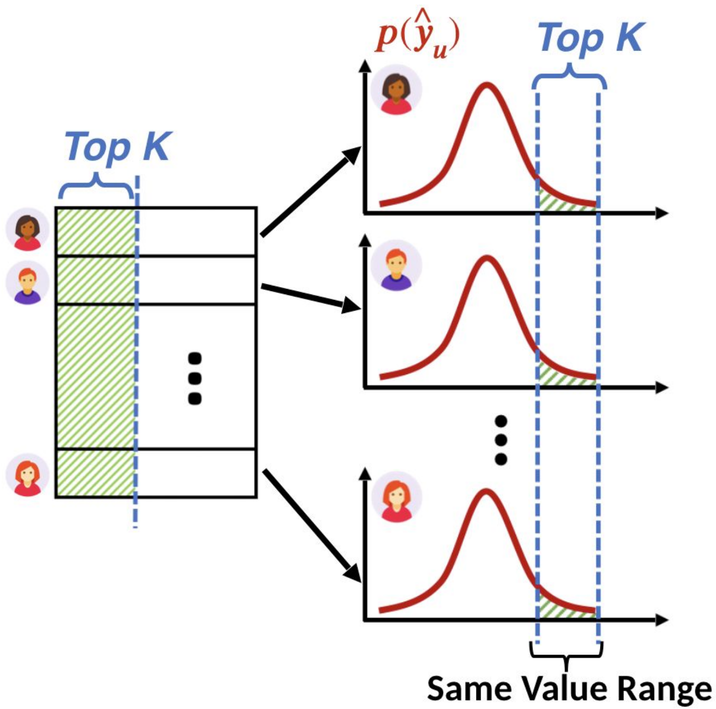
Rank items based on predicted scores for users.

Debiased Personalized Ranking (DPR) Model -- RSP



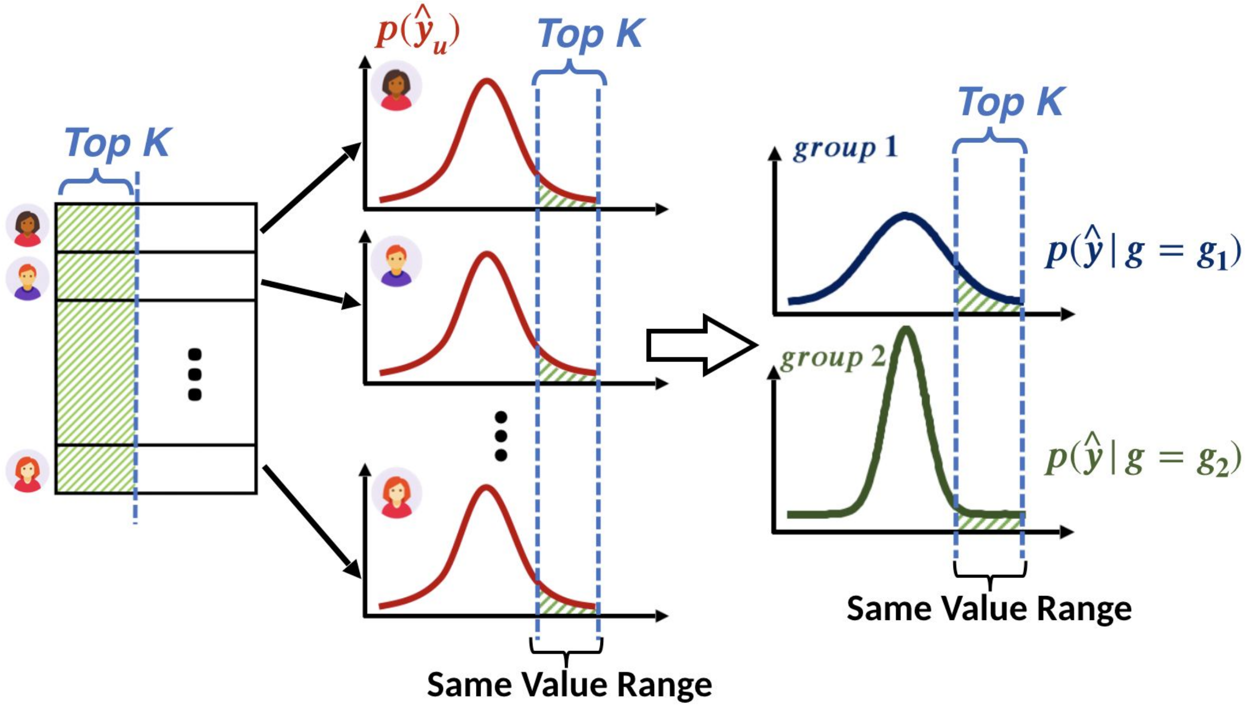
The top-K items for each user will lay in the most right part in user score distribution.

Debiased Personalized Ranking (DPR) Model -- RSP



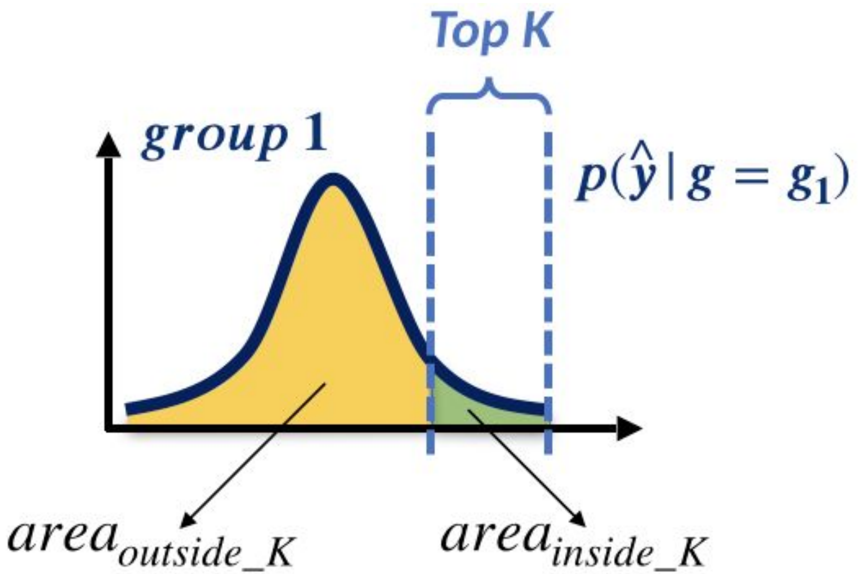
Normalize score distribution for each user so that all users have the same score distribution.

Debiased Personalized Ranking (DPR) Model -- RSP



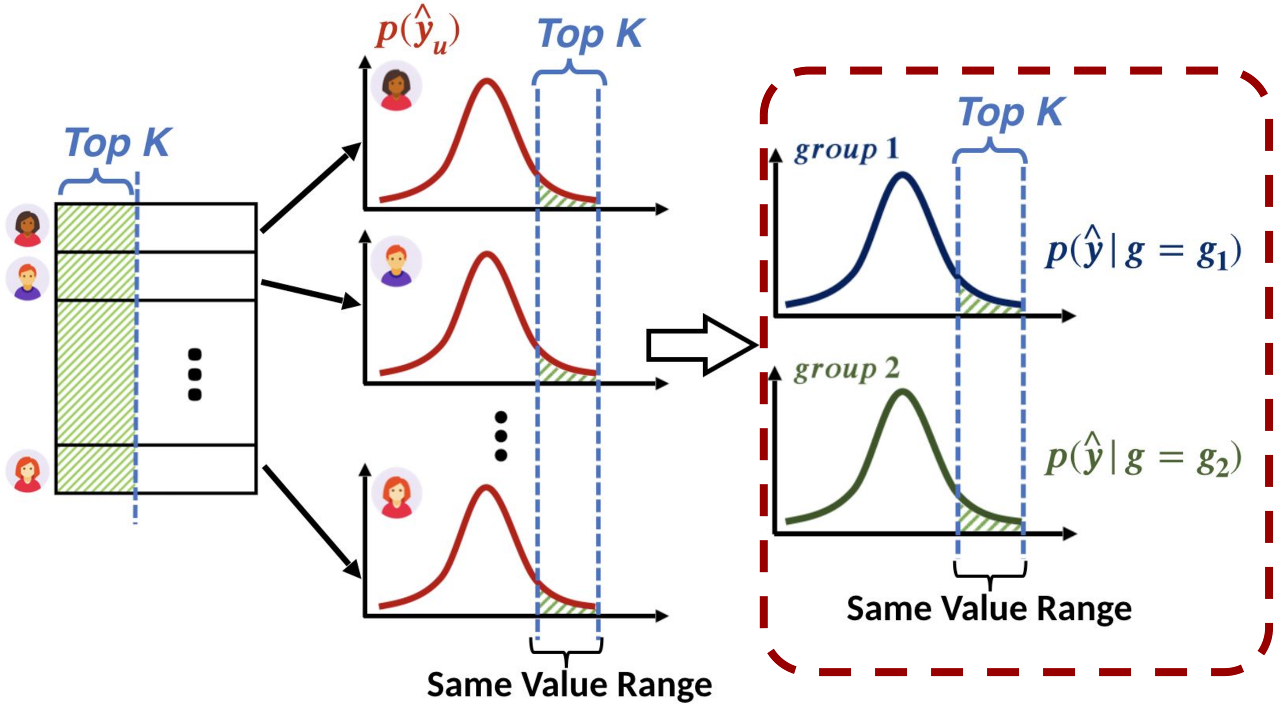
Plot the score distribution for item groups, scores for recommended items in different group lay in the same score range.

Debiased Personalized Ranking (DPR) Model -- RSP



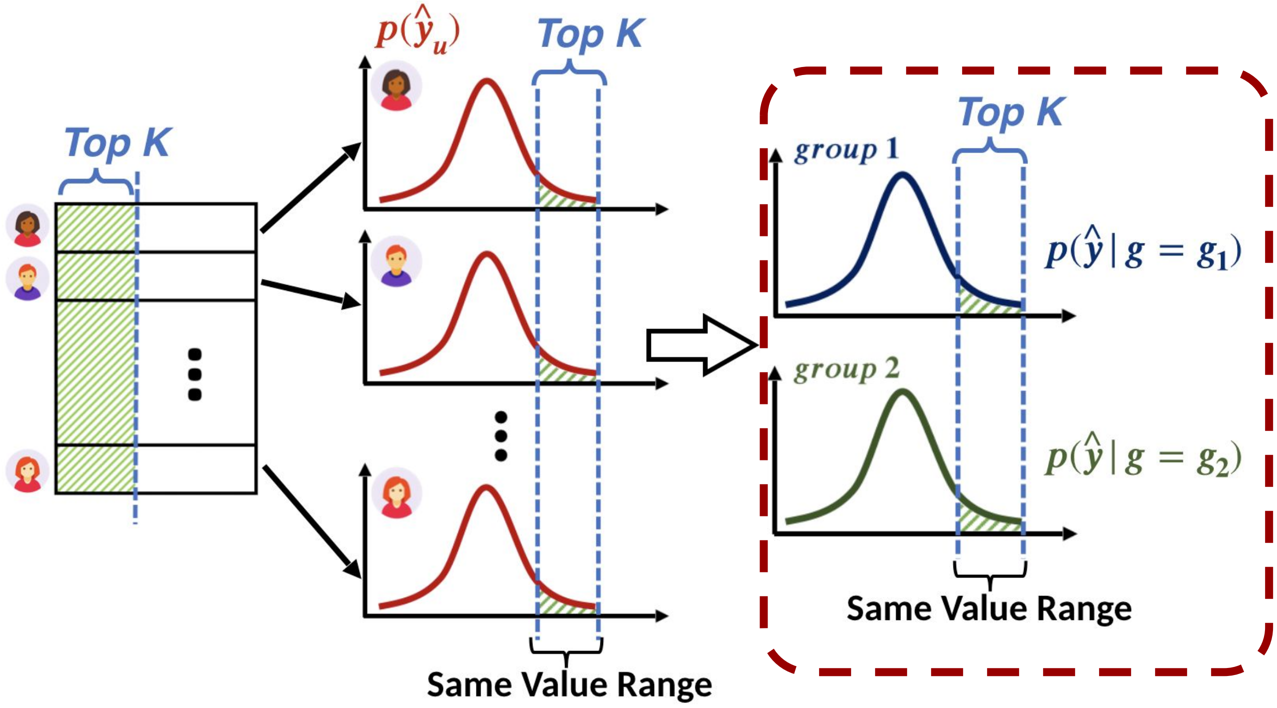
$$P(rank@K \,|\, g = g_1) = \frac{area_{inside\_K}}{area_{outside\_K} + area_{inside\_K}}$$

Debiased Personalized Ranking (DPR) Model -- RSP



Force the same score distribution for different item groups.

Debiased Personalized Ranking (DPR) Model -- RSP



$$P(rank@K|g = g_1) = P(rank@K|g = g_2)$$

Debiased Personalized Ranking (DPR) Model -- RSP

To mitigate RSP based bias:

- Decouple the predicted score with group attribute;
- Normalize the score distribution for each user to have the same score distribution for all users.

Debiased Personalized Ranking (DPR) Model -- RSP

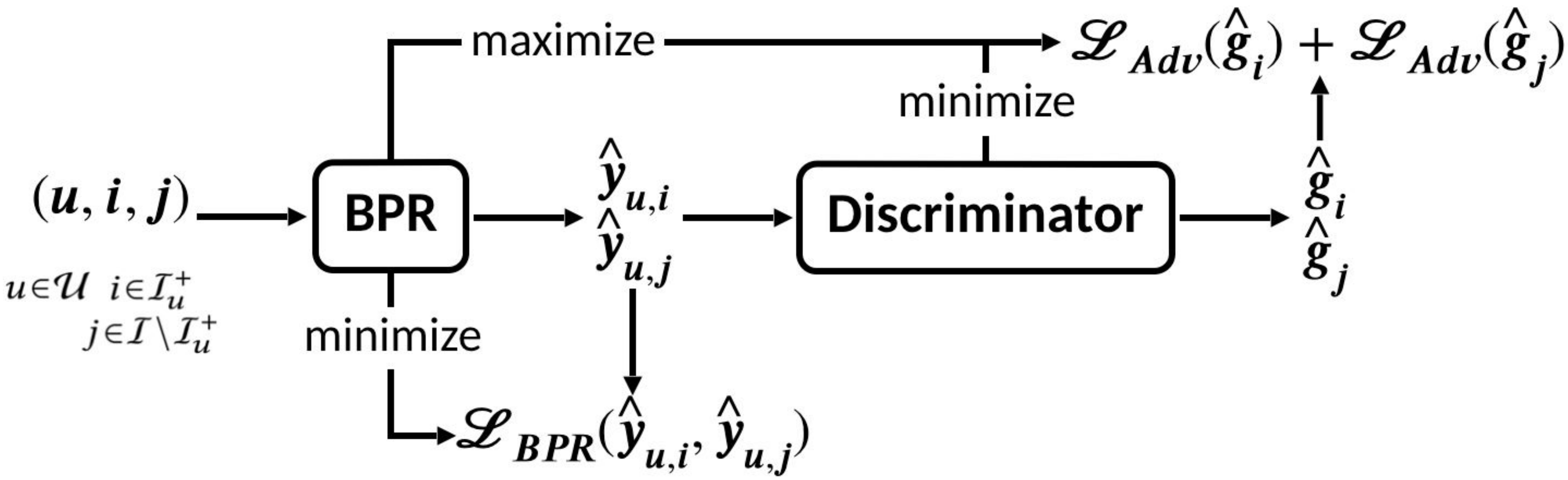
To mitigate RSP based bias:

- ➔ **Decouple the predicted score with group attribute;**
- Normalize the score distribution for each user to have the same score distribution for all users.

Debiased Personalized Ranking (DPR) Model -- RSP

To mitigate RSP based bias:

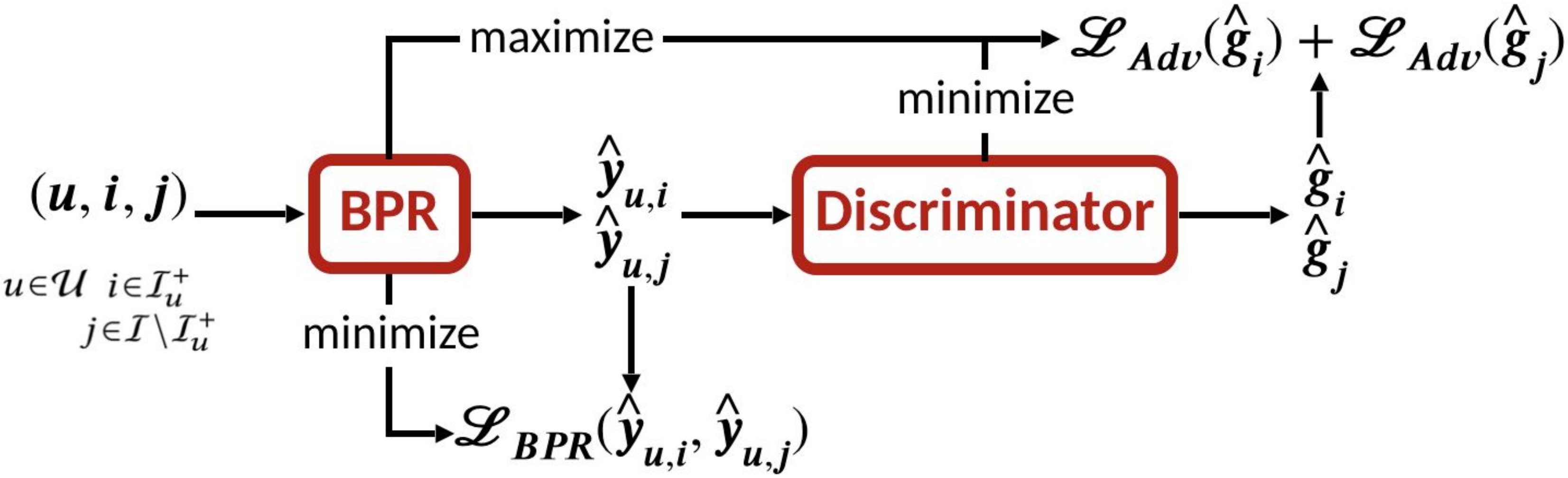
➔ **Decouple the predicted score with group attribute;**



Debiased Personalized Ranking (DPR) Model -- RSP

To mitigate RSP based bias:

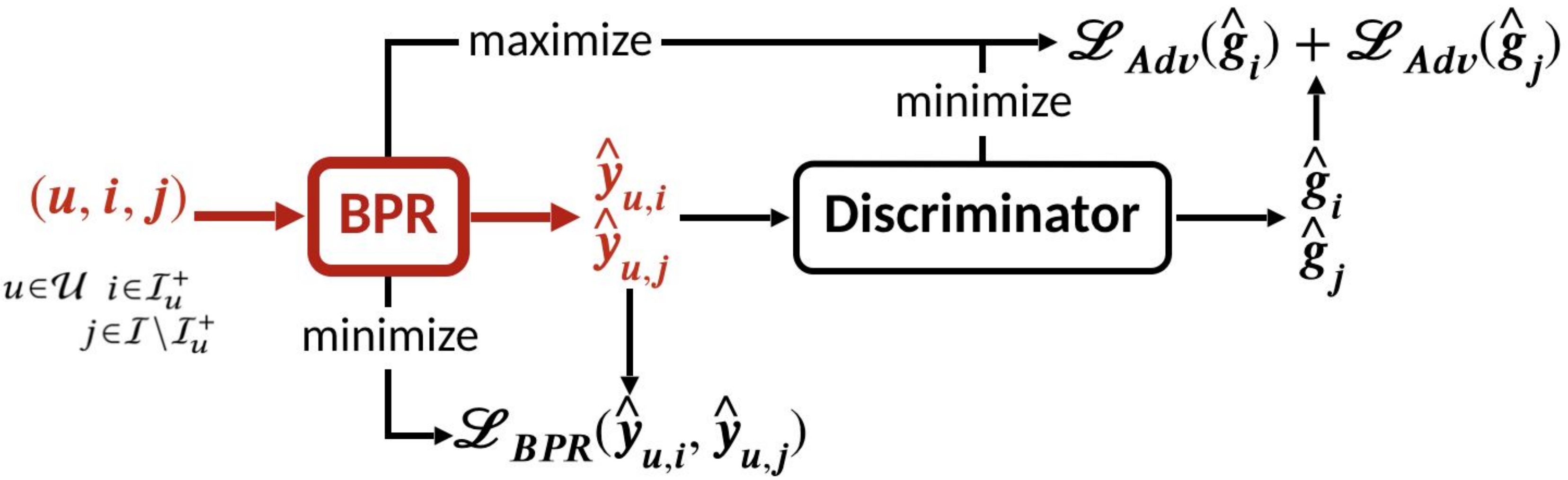
➔ **Decouple the predicted score with group attribute;**



Debiased Personalized Ranking (DPR) Model -- RSP

To mitigate RSP based bias:

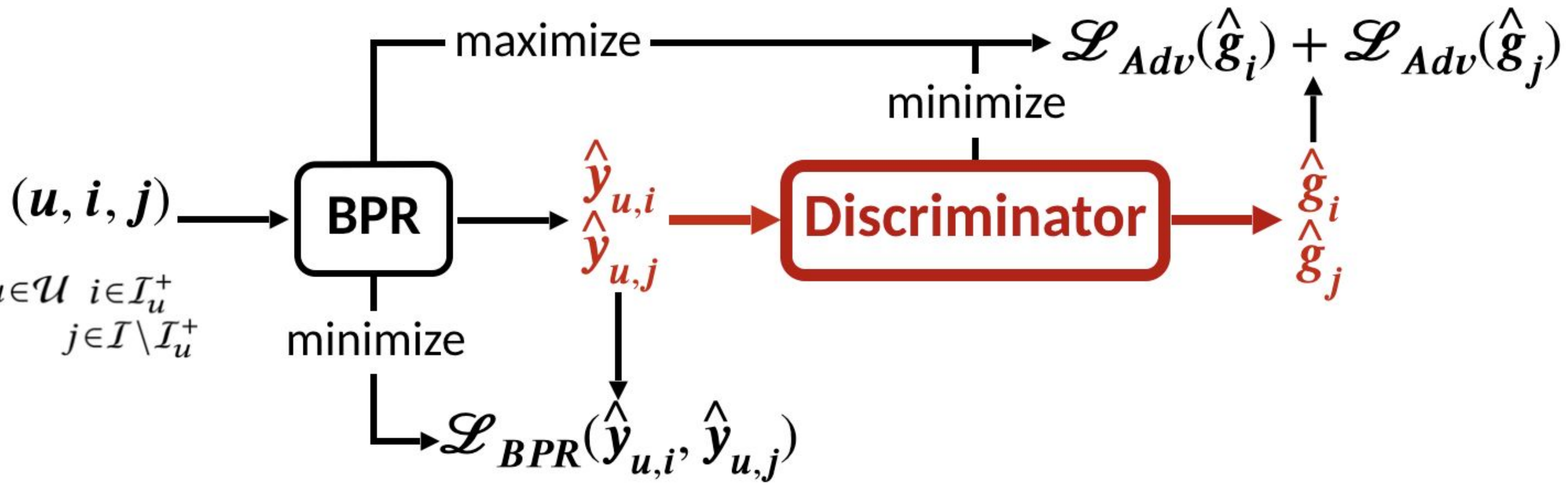
➔ **Decouple the predicted score with group attribute;**



Debiased Personalized Ranking (DPR) Model -- RSP

To mitigate RSP based bias:

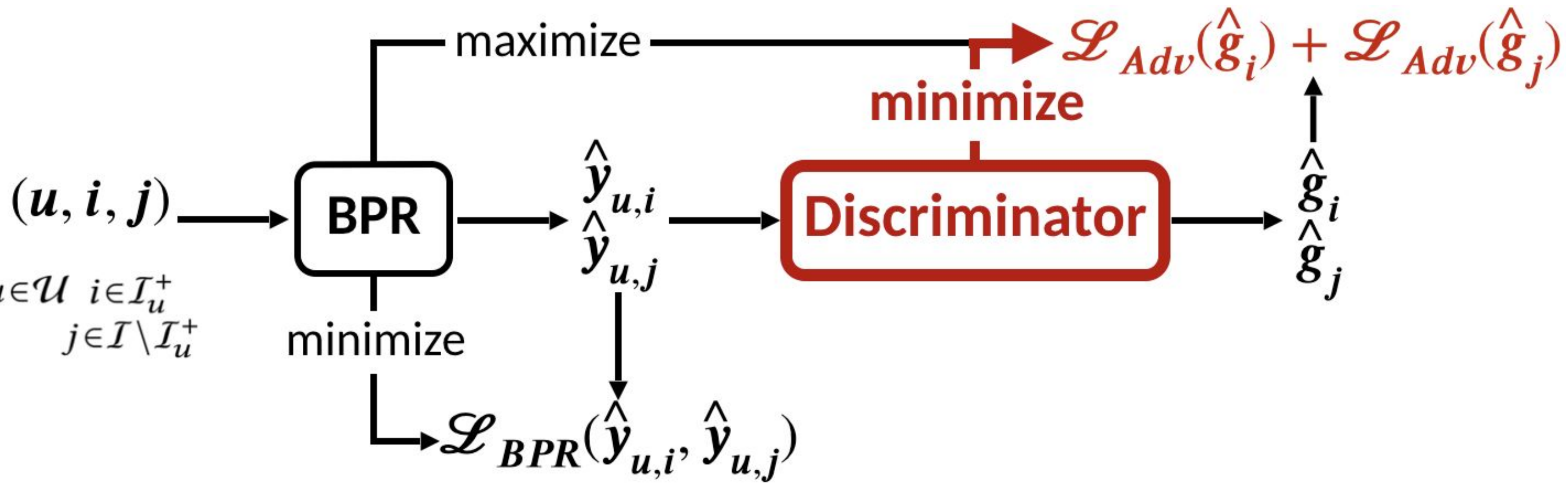
➔ **Decouple the predicted score with group attribute;**



Debiased Personalized Ranking (DPR) Model -- RSP

To mitigate RSP based bias:

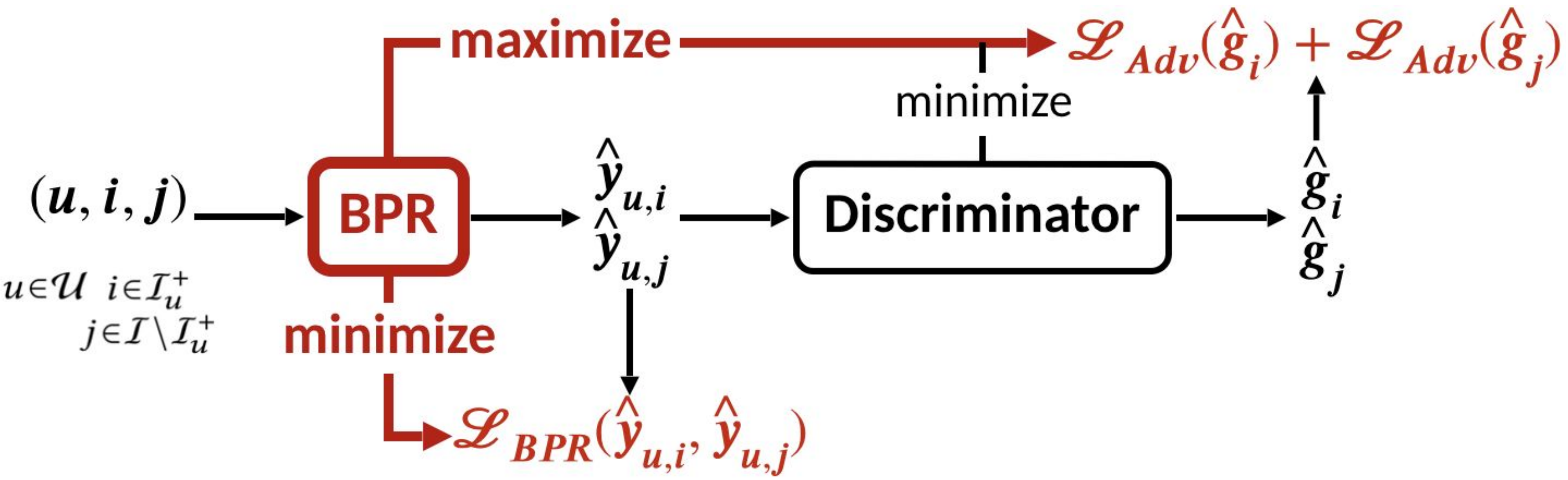
➔ **Decouple the predicted score with group attribute;**



Debiased Personalized Ranking (DPR) Model -- RSP

To mitigate RSP based bias:

➔ **Decouple the predicted score with group attribute;**



Debiased Personalized Ranking (DPR) Model -- RSP

To mitigate RSP based bias:

➔ **Decouple the predicted score with group attribute;**

$$\min_{\Theta} \max_{\Psi} \sum_{u \in \mathcal{U}} \sum_{\substack{i \in \mathcal{I}_u^+ \\ j \in \mathcal{I} \setminus \mathcal{I}_u^+}} (\mathcal{L}_{BPR}(u, i, j) + \alpha(\mathcal{L}_{Adv}(i) + \mathcal{L}_{Adv}(j))) + \beta \mathcal{L}_{KL}$$

Debiased Personalized Ranking (DPR) Model -- RSP

To mitigate RSP based bias:

➔ **Decouple the predicted score with group attribute;**

$$\boxed{\min_{\Theta} \max_{\Psi}} \sum_{u \in \mathcal{U}} \sum_{\substack{i \in \mathcal{I}_u^+ \\ j \in \mathcal{I} \setminus \mathcal{I}_u^+}} (\mathcal{L}_{BPR}(u, i, j) + \alpha(\mathcal{L}_{Adv}(i) + \mathcal{L}_{Adv}(j))) + \beta \mathcal{L}_{KL}$$

Play a minimax game between the BPR component (with parameter set $\mathbf{\Theta}$) and the adversarial component (with parameter set $\mathbf{\Psi}$).

Debiased Personalized Ranking (DPR) Model -- RSP

To mitigate RSP based bias:

➔ **Decouple the predicted score with group attribute;**

$$\min_{\Theta}\max_{\Psi}\sum_{u\in\mathcal{U}}\sum_{\substack{i\in\mathcal{I}_u^+\\ j\in\mathcal{I}\setminus\mathcal{I}_u^+}}\left(\boxed{\mathcal{L}_{BPR}(u,i,j)}+\alpha(\mathcal{L}_{Adv}(i)+\mathcal{L}_{Adv}(j))\right)+\beta\mathcal{L}_{KL}$$

Conventional BPR loss for a user *u* with one positive item *i* and one negative item *j*:

$$\mathcal{L}_{BPR}(u,i,j) = -ln\,\sigma(\widehat{y}_{u,i}-\widehat{y}_{u,j}) + \frac{\lambda_\Theta}{2}\|\Theta\|_F^2$$

Debiased Personalized Ranking (DPR) Model -- RSP

To mitigate RSP based bias:

➔ **Decouple the predicted score with group attribute;**

$$\min_{\Theta} \max_{\Psi} \sum_{u \in \mathcal{U}} \sum_{\substack{i \in \mathcal{I}_u^+ \\ j \in \mathcal{I} \setminus \mathcal{I}_u^+}} \left(\mathcal{L}_{BPR}(u,i,j) + \boxed{\alpha(\mathcal{L}_{Adv}(i) + \mathcal{L}_{Adv}(j))}\right) + \beta \mathcal{L}_{KL}$$

The adversarial component takes predicted score as input and predict the group label of the given item. Train the adversarial component by:

$$\max_{\Psi} \mathcal{L}_{Adv}(i) = \sum_{a=1}^{A} (\mathrm{g}_{i,a} log\, \widehat{\mathrm{g}}_{i,a} + (1 - \mathrm{g}_{i,a}) log\,(1 - \widehat{\mathrm{g}}_{i,a}))$$

Debiased Personalized Ranking (DPR) Model -- RSP

To mitigate RSP based bias:

- Decouple the predicted score with group attribute;

➔ **Normalize the score distribution for each user to have the same score distribution for all users.**

Debiased Personalized Ranking (DPR) Model -- RSP

To mitigate RSP based bias:

➔ **Normalize the score distribution for each user to have the same score distribution for all users.**

$$\min_{\Theta} \max_{\Psi} \sum_{u \in \mathcal{U}} \sum_{\substack{i \in \mathcal{I}_u^+ \\ j \in \mathcal{I} \setminus \mathcal{I}_u^+}} (\mathcal{L}_{BPR}(u, i, j) + \alpha(\mathcal{L}_{Adv}(i) + \mathcal{L}_{Adv}(j))) + \boxed{\beta \mathcal{L}_{KL}}$$

Minimize the KL divergence between the score distribution of each user and the standard normal distribution to normalize score distribution for users:

$$\mathcal{L}_{KL} = \sum_{u \in \mathcal{U}} D_{\mathrm{KL}}(q_{\Theta}(u) || \mathcal{N}(0, 1))$$

Debiased Personalized Ranking (DPR) Model -- REO

REO considers the **true positive rate** across groups

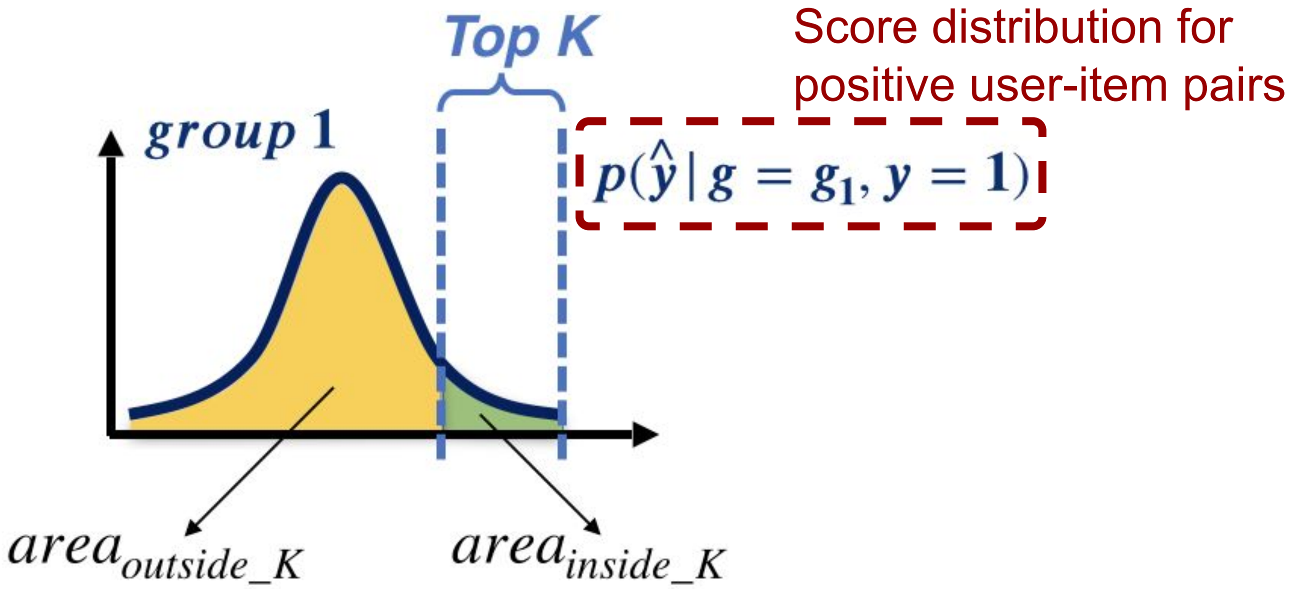
$$P(rank@K \mid g = g_1, y = 1)$$

Debiased Personalized Ranking (DPR) Model -- REO

To mitigate REO based bias:

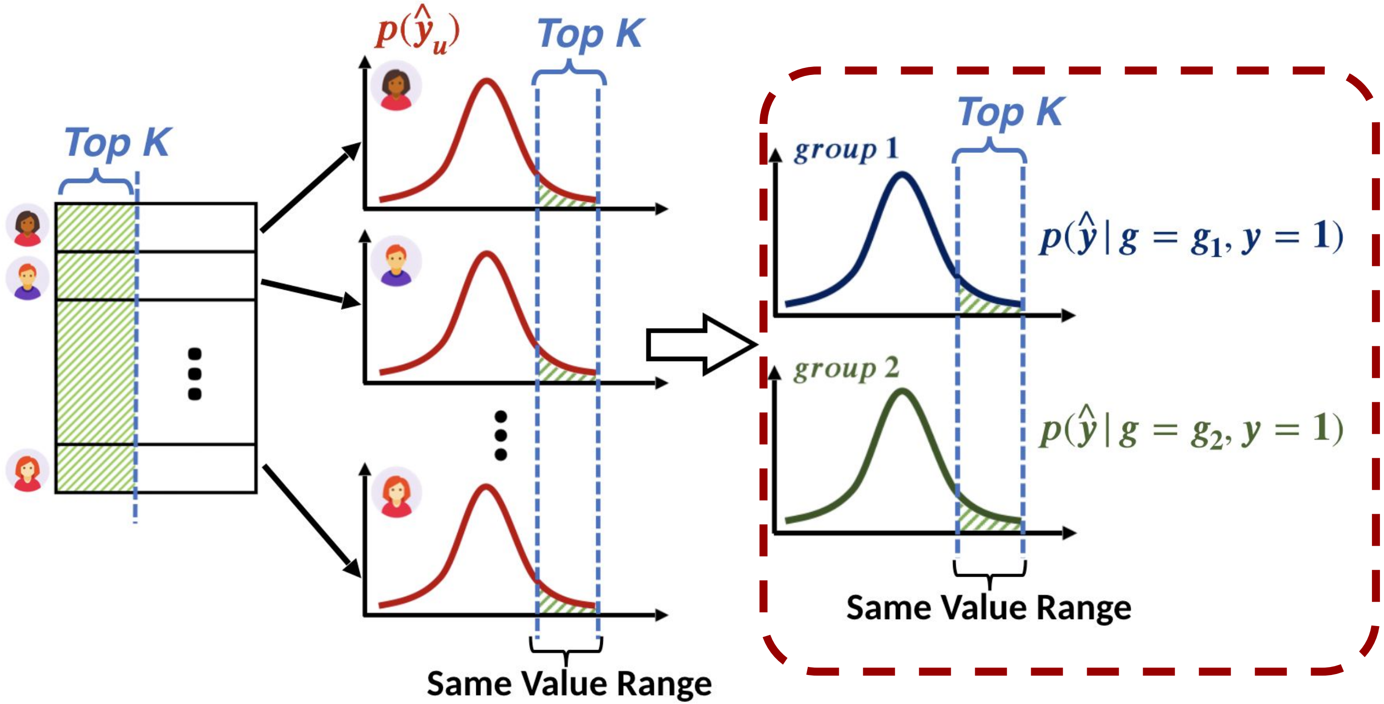
- Decouple the group attribute with the predicted score for **positive user-item pairs**;
- Normalize the score distribution for each user to have the same score distribution for all users.

Debiased Personalized Ranking (DPR) Model -- REO



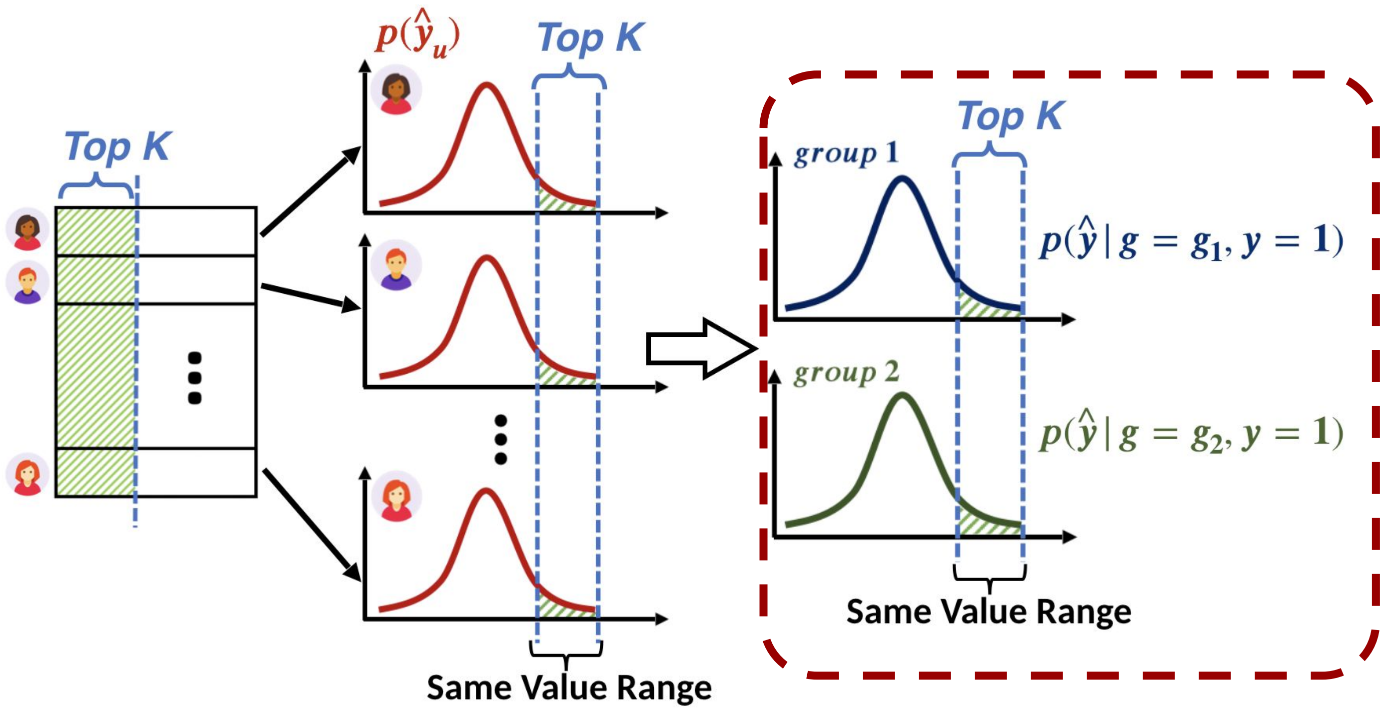
$$P(rank@K \mid g = g_1, y = 1) = \frac{area_{inside\_K}}{area_{outside\_K} + area_{inside\_K}}$$

Debiased Personalized Ranking (DPR) Model -- REO



Force the same score distribution for **positive user-item pairs** for different item groups.

Debiased Personalized Ranking (DPR) Model -- REO



$$P(rank@K|g = g_1, y = 1) = P(rank@K|g = g_2, y = 1)$$

Debiased Personalized Ranking (DPR) Model -- REO

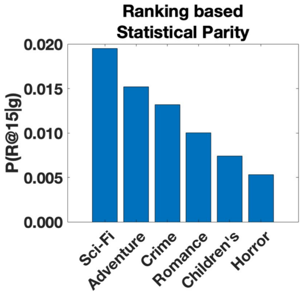
To mitigate REO based bias:

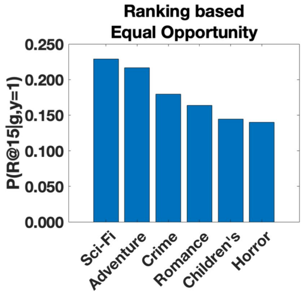
- Decouple the group attribute with the predicted score for **positive user-item pairs**;

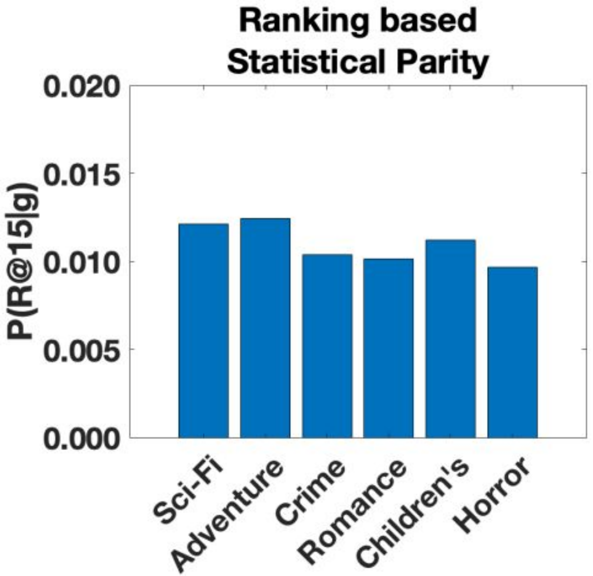
$$\min_{\Theta} \max_{\Psi} \sum_{u \in \mathcal{U}} \sum_{\substack{i \in \mathcal{I}_u^+ \\ j \in \mathcal{I} \setminus \mathcal{I}_u^+}} \left(\mathcal{L}_{BPR}(u,i,j) + \boxed{\alpha \mathcal{L}_{Adv}(i)}\right) + \beta \mathcal{L}_{KL}$$

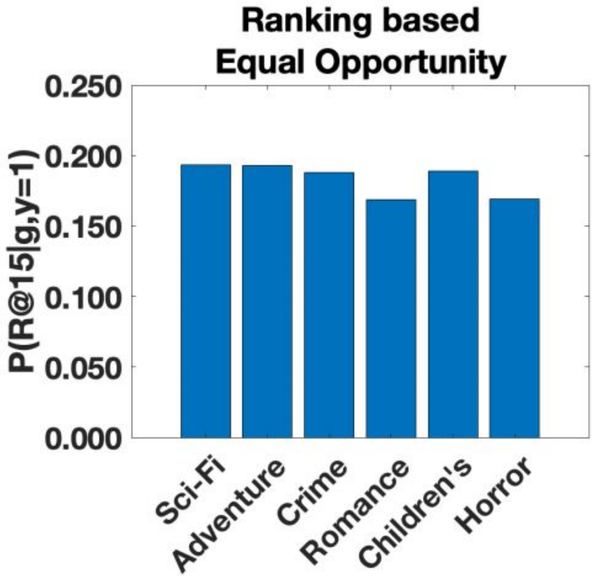
Only input scores for positive user-item pairs to the adversarial component.

Experiments – visualize debiased results



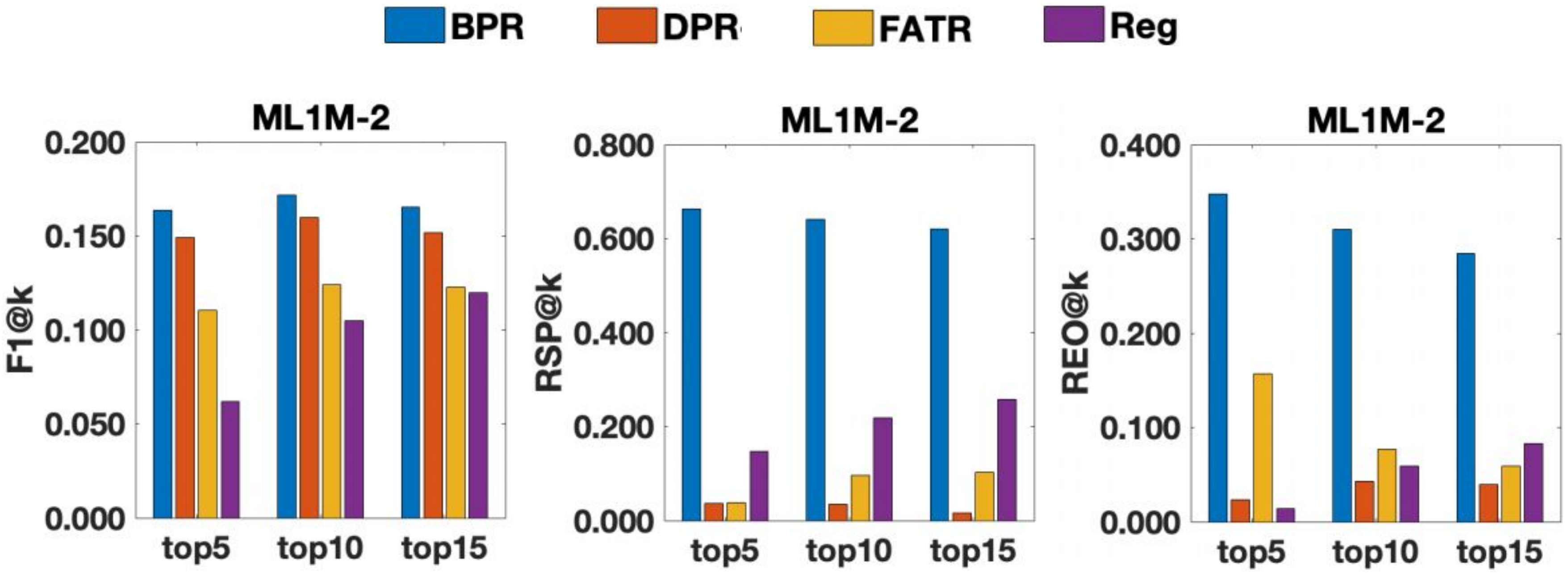




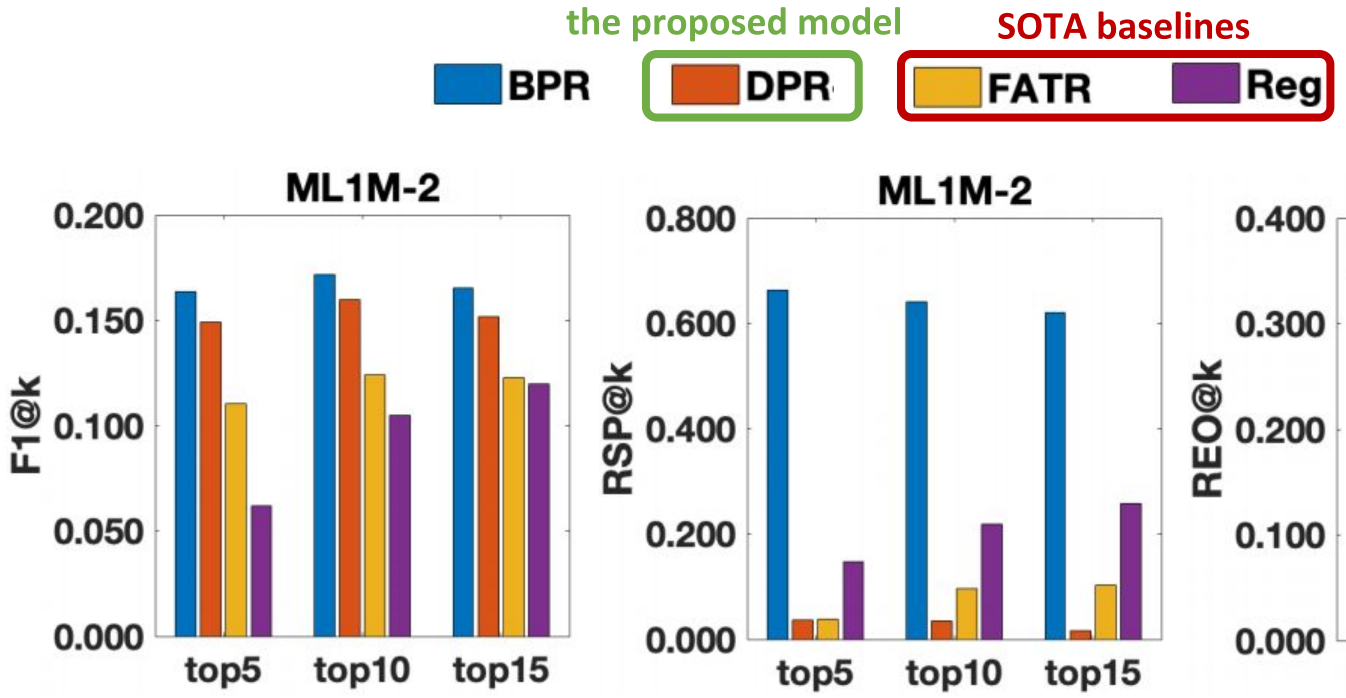


by the proposed DPR

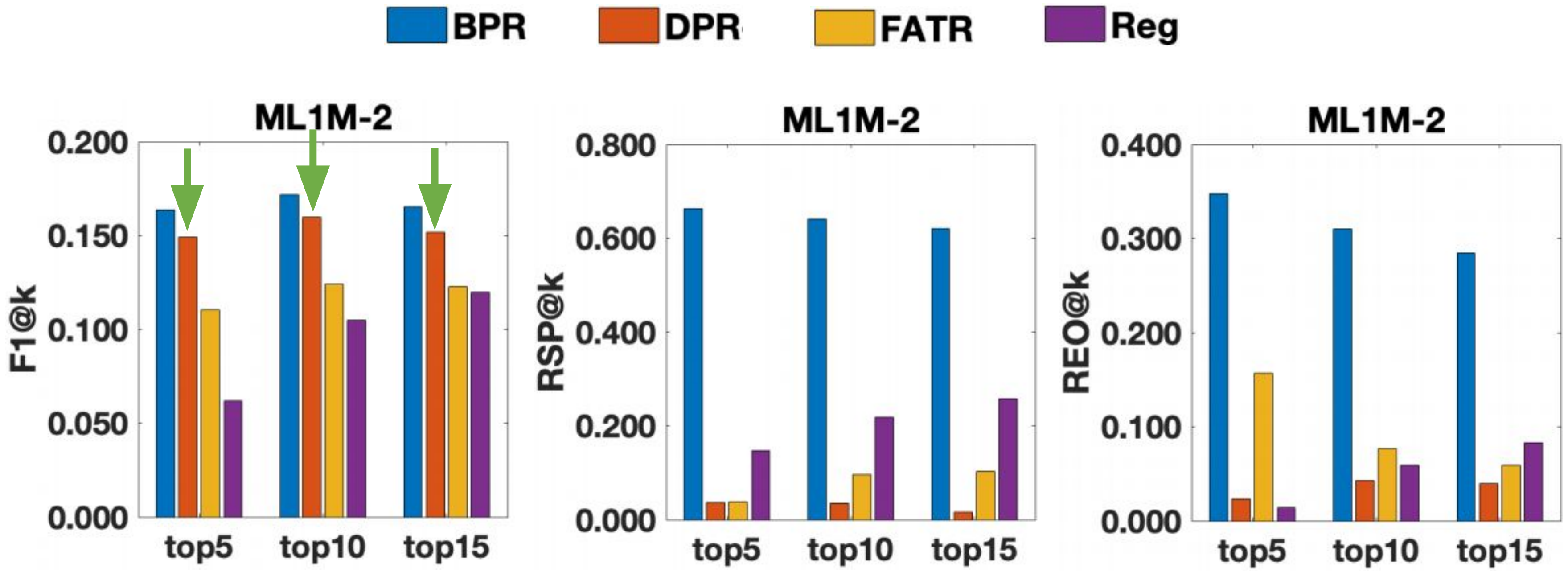
Experiments – compare with baselines



Experiments – compare with baselines

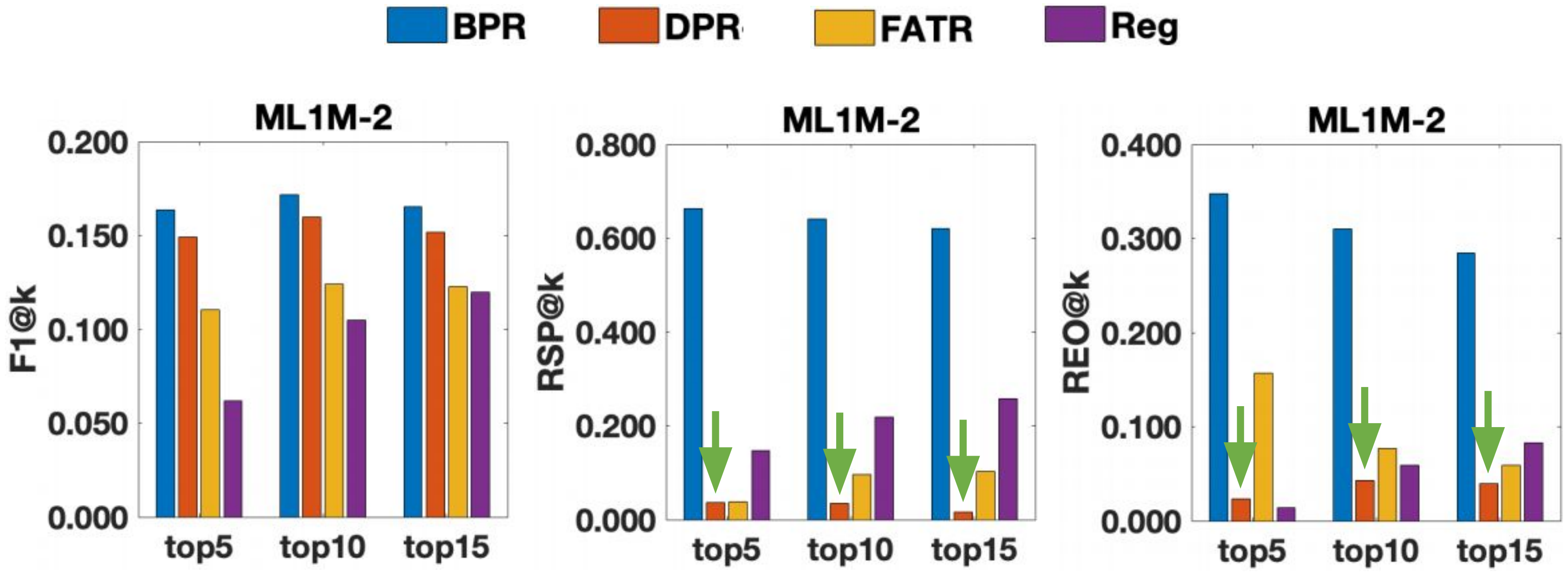


Experiments – compare with baselines



Proposed model **preserves high recommendation utility.**

Experiments – compare with baselines



And enhance **RSP and REO fairness** effectively!

Experiments – more in the paper

More experimental details and results can be found in the paper, including:

- Detailed experiment setup;
- Experiments on other datasets;
- Experiments for ablation study;
- Experiments for hyper-parameter study;