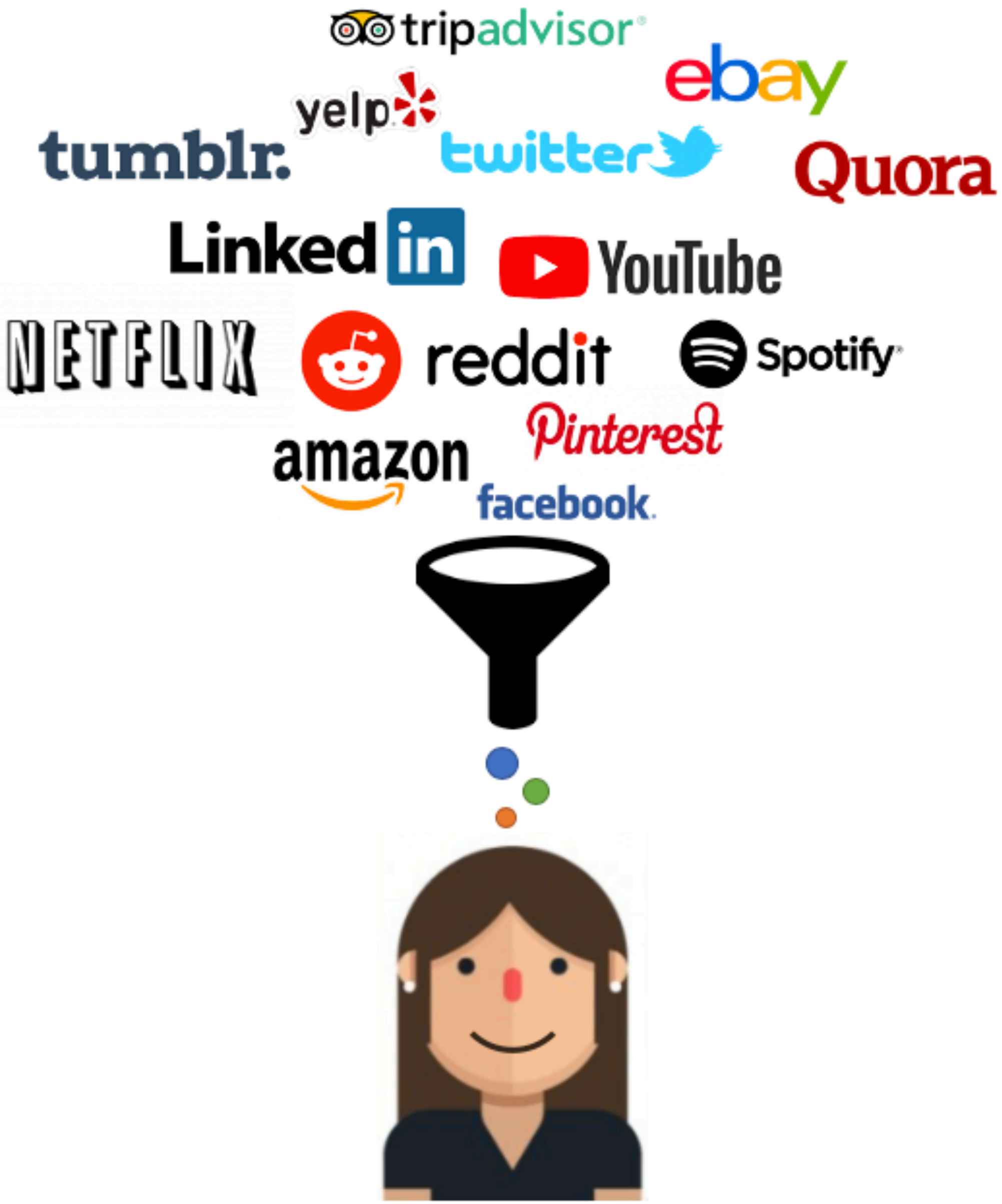
Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems

Ziwei Zhu, Jianling Wang, and James Caverlee

Texas A&M University

SIGIR 2020
XI'AN·CHINA

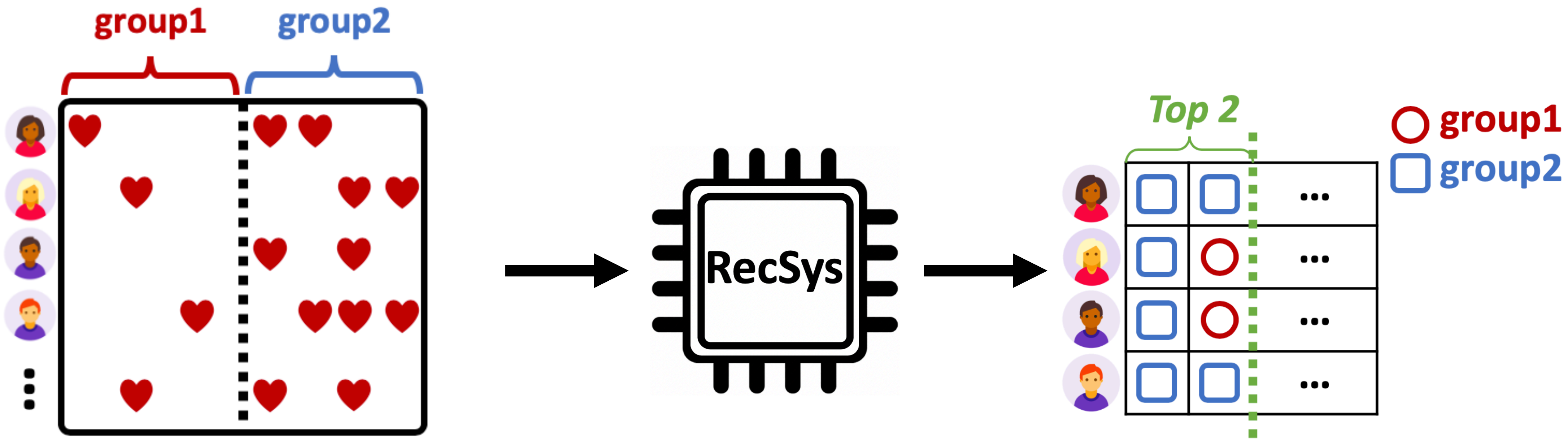
Recommenders – essential conduits



Algorithmic bias in recommenders

Item groups are under-recommended

Due to i) the imbalanced distribution of feedback for different item groups;
ii) the unawareness of bias in recommendation algorithm;
Items from some groups will be under-recommended compared to other popular item groups.



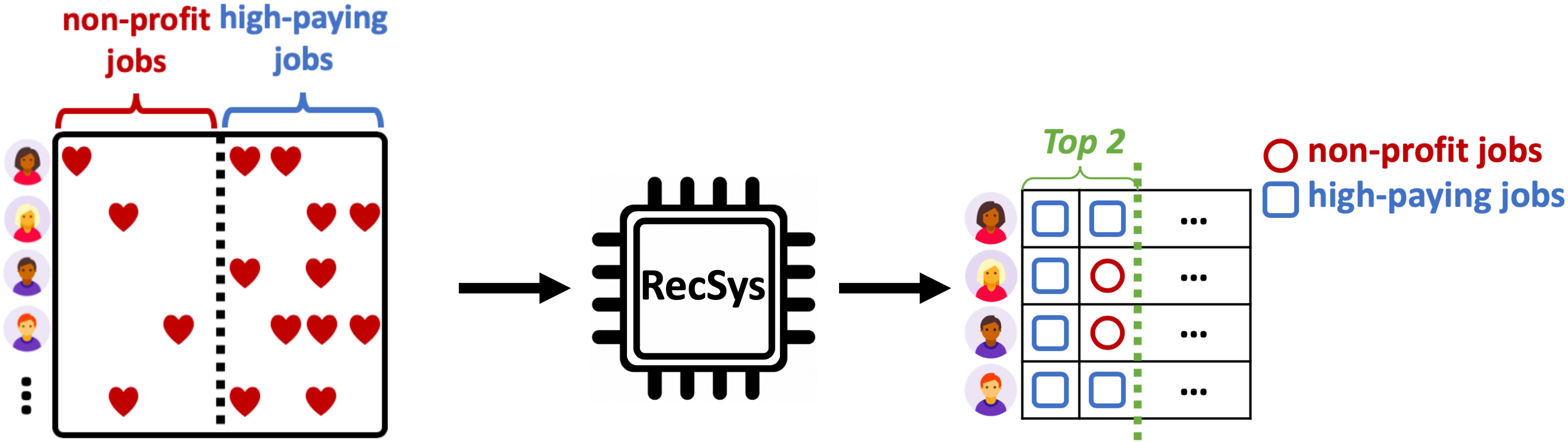
Imbalanced distribution of feedback for item groups.

Model without awareness of bias.

Items in group1 are under-recommended

Item groups are under-recommended

Example: when recommend jobs to users, **non-profit jobs** are under-recommended compared with **high-paying jobs**.



Imbalanced distribution of feedback for item groups.

Model without awareness of bias.

Non-profit jobs are under-recommended

Previous works

- Measure the bias on predicted scores of item groups.

- Measure the bias based on the concept of statistical parity.

- No bias: $P(score|group1) = P(score|group2) = \cdots = P(score|groupA)$

Previous works

- ➢ Measure the bias based on predicted scores of item groups.
- ➢ Predicted score is the intermedia step towards the rankings, thus, unbiased scores do not necessarily lead to unbiased recommendation.

- Measure the bias based on the concept of statistical parity.

- No bias: $P(score|group1) = P(score|group2) = \cdots = P(score|groupA)$

Previous works

- Measure the bias based on predicted scores of item groups.
- Predicted score is the intermedia step towards the rankings, thus, unbiased scores do not necessarily lead to unbiased recommendation.

- Measure the bias based on the concept of statistical parity.
- Statistical Parity is too strict for scenarios where there is no sensitive attributes for items (like books or movies).

- No bias: $P(score|group1) = P(score|group2) = \cdots = P(score|groupA)$

Previous works

- Measure the bias based on predicted scores of item groups.
- Predicted score is the intermedia step towards the rankings, thus, unbiased scores do not necessarily lead to unbiased recommendation.

- Measure the bias based on the concept of statistical parity.
- Statistical Parity is too strict for scenarios where there is no sensitive attributes for items (like books or movies).

➢ No bias: $P(score|group1) = P(score|group2) = \cdots = P(score|groupA)$
➢ Therefore, bias measurements based on **ranking** and other **bias concepts** are in need.

Contributions

- Propose the **ranking-based statistical parity (RSP)** measurement;
- Propose the **ranking-based equal opportunity (REO)** measurement;
- Propose the **Debiased Personalized Ranking (DPR)** model;
- Empirically demonstrate that the fundamental recommendation model – Bayesian Personalized Ranking (BPR) – is vulnerable to the under-recommendation bias, and show the effectiveness of the proposed DPR.

Ranking-based Statistical Parity (RSP)

$$P(score|group1) = P(score|group2) = \cdots = P(score|groupA)$$

Predicted scores are intermedia steps towards rankings, which serve as the final recommendation results. Thus,
unbiased predicted scores $\neq$ unbiased rankings

Ranking-based Statistical Parity (RSP)

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

$$P(topk|group1) = P(topk|group2) = \cdots = P(topk|groupA)$$

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**.

$$P(topk|group1) = P(topk|group2) = \cdots = P(topk|groupA)$$

RSP – motivating example

Example: Recommend job candidates to companies

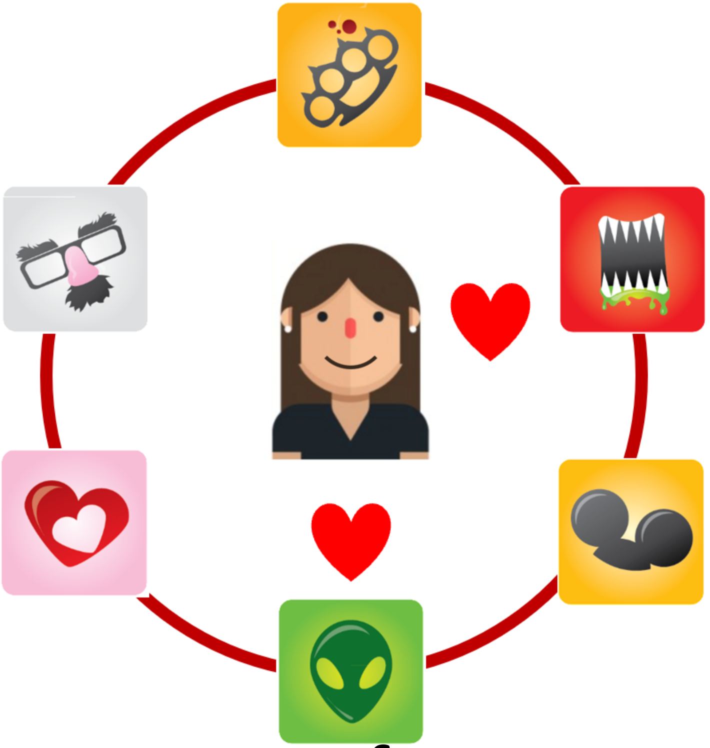
$\boldsymbol{P(recommend|\text{♂}) = 0.6}$

$\boldsymbol{P(recommend|\text{♀}) = 0.2}$

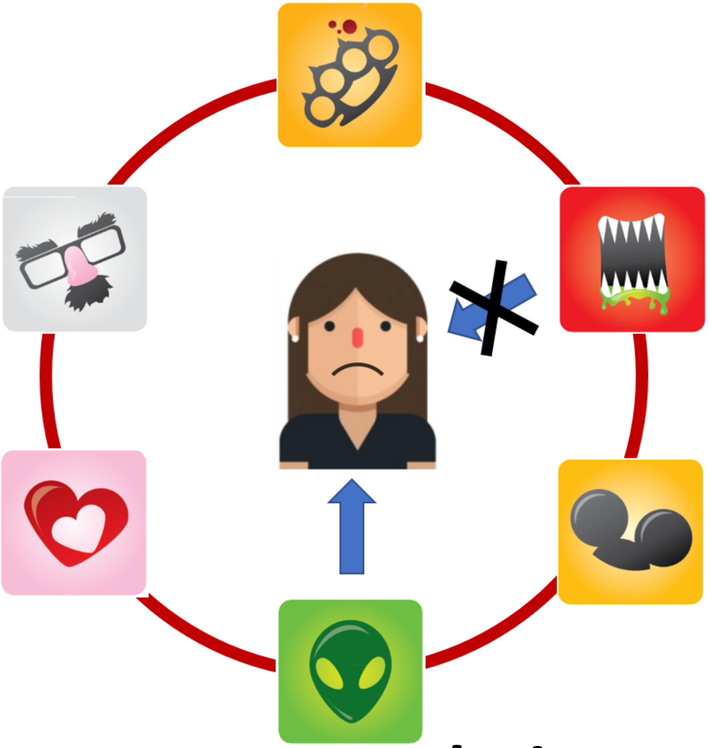
Unfair for female candidates.

Ranking-based Equal Opportunity (REO)

For a **more general RecSys**, we do not require statistical parity, but 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.

$$P(topk|group1\&\boldsymbol{liked}) = \cdots = P(topk|groupA\&\boldsymbol{liked})$$

REO – motivating example

Example: Recommend movies to users

horror and sci-fi movies lover

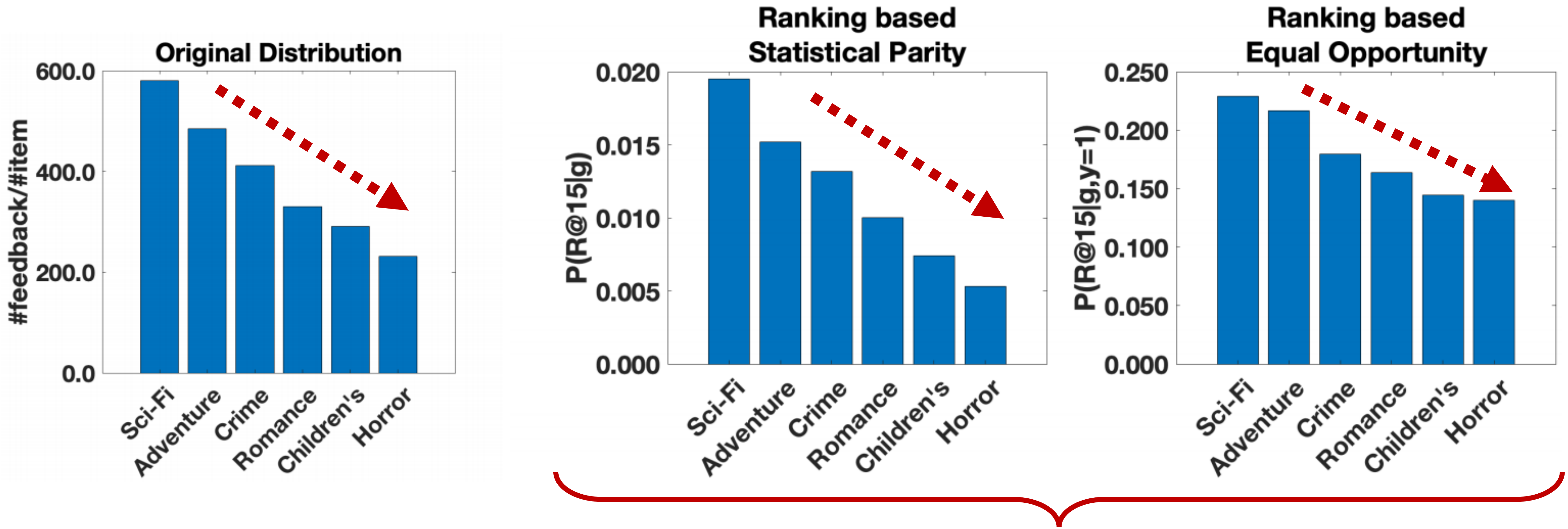
$$\boldsymbol{p(recommend|horror\&liked) = 0.3}$$

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

For a long time, 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

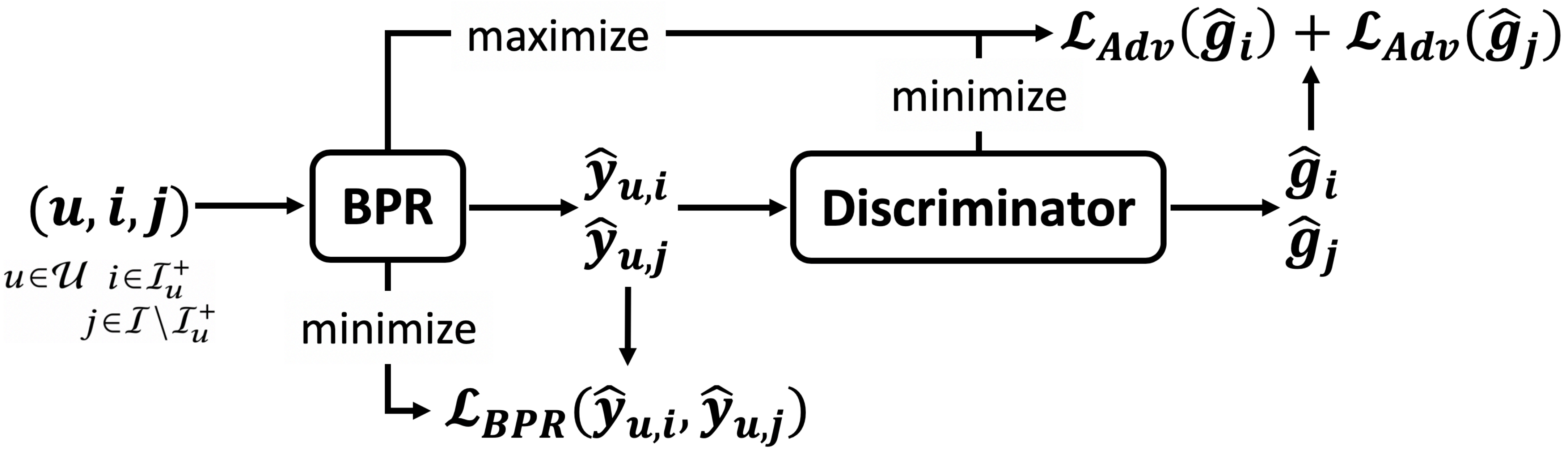
To mitigate RSP based bias:

- Decouple the predicted score with group attribute;
- Normalize the score distribution for each user to align predict score with ranking position.

Debiased Personalized Ranking (DPR) Model

To mitigate RSP based bias:

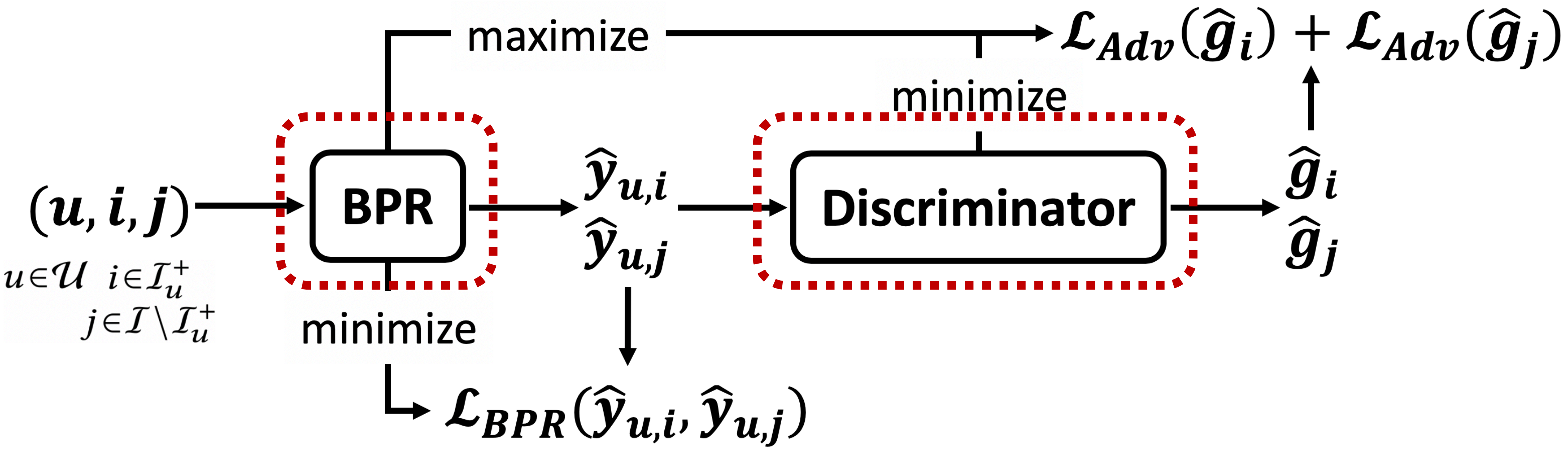
- ➢ **Decouple the predicted score with group attribute;**
- Normalize the score distribution for each user to align predict score with ranking position.



Debiased Personalized Ranking (DPR) Model

To mitigate RSP based bias:

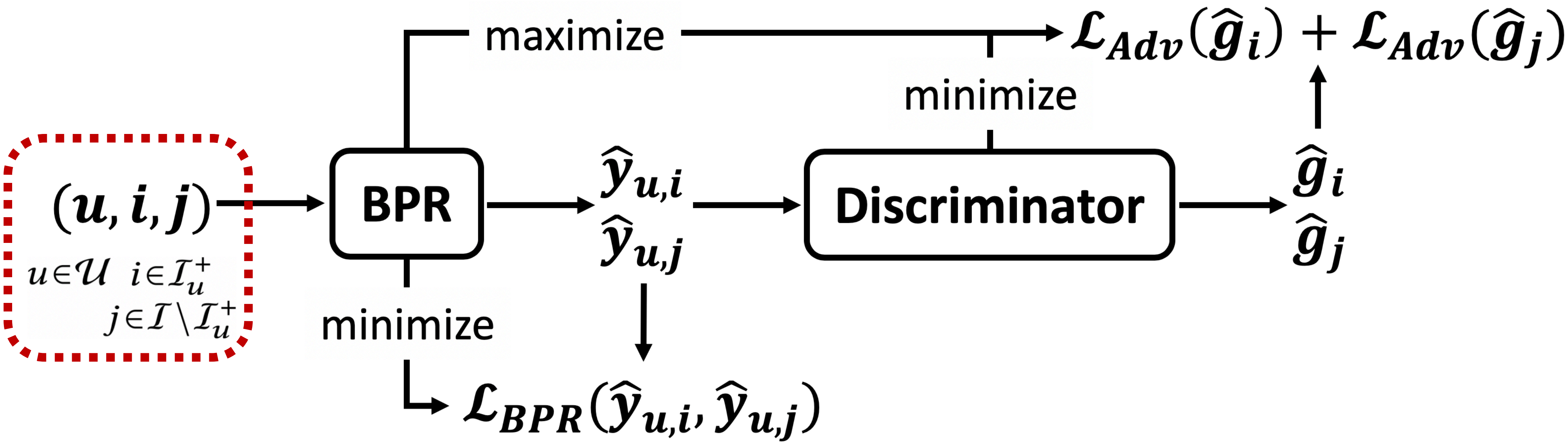
- ➢ **Decouple the predicted score with group attribute;**
- Normalize the score distribution for each user to align predict score with ranking position.



Debiased Personalized Ranking (DPR) Model

To mitigate RSP based bias:

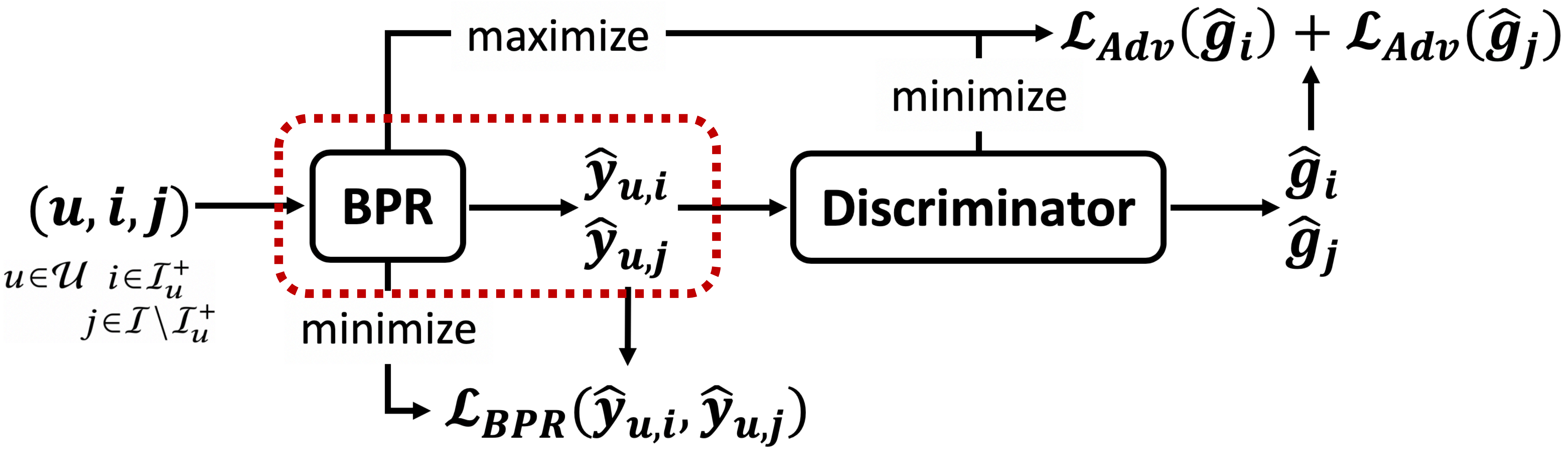
- ➢ **Decouple the predicted score with group attribute;**
- Normalize the score distribution for each user to align predict score with ranking position.



Debiased Personalized Ranking (DPR) Model

To mitigate RSP based bias:

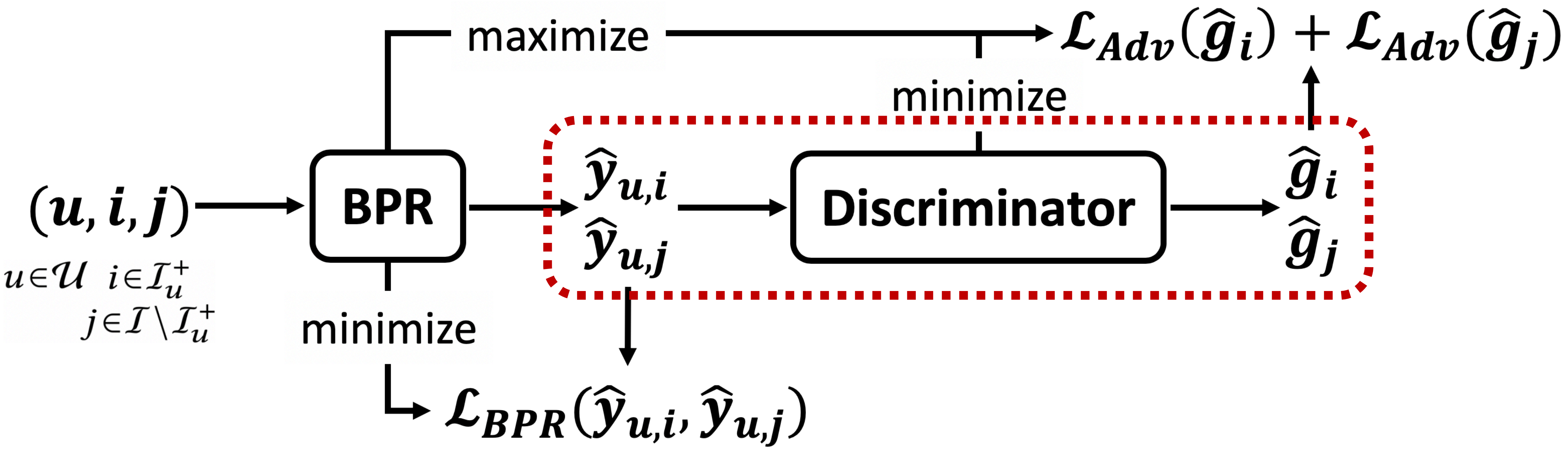
- ➢ **Decouple the predicted score with group attribute;**
- Normalize the score distribution for each user to align predict score with ranking position.



Debiased Personalized Ranking (DPR) Model

To mitigate RSP based bias:

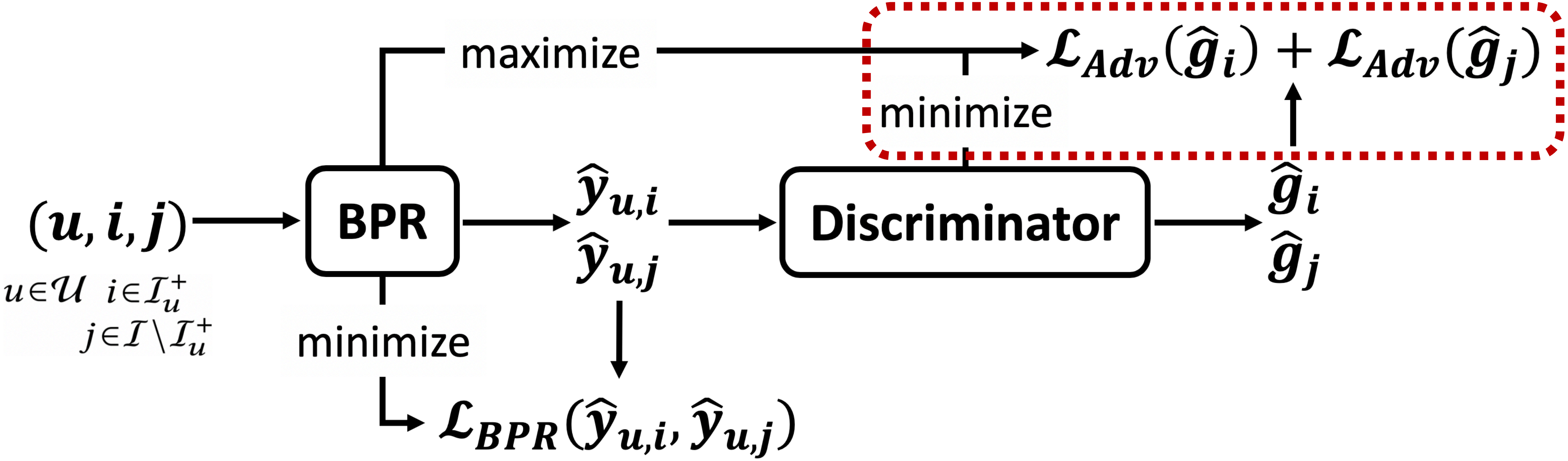
- ➢ **Decouple the predicted score with group attribute;**
- Normalize the score distribution for each user to align predict score with ranking position.



Debiased Personalized Ranking (DPR) Model

To mitigate RSP based bias:

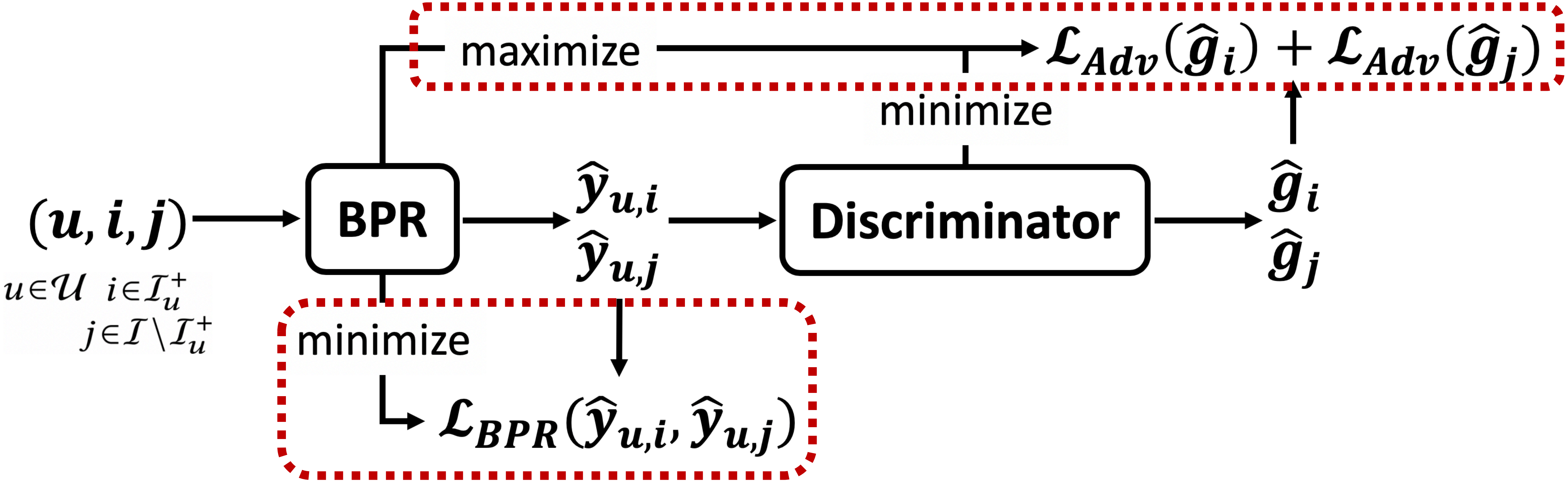
- ➢ **Decouple the predicted score with group attribute;**
- Normalize the score distribution for each user to align predict score with ranking position.



Debiased Personalized Ranking (DPR) Model

To mitigate RSP based bias:

- ➢ **Decouple the predicted score with group attribute;**
- Normalize the score distribution for each user to align predict score with ranking position.



Debiased Personalized Ranking (DPR) Model

To mitigate RSP based bias:

- ➢ **Decouple the predicted score with group attribute;**
- Normalize the score distribution for each user to align predict score with ranking position.

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

Debiased Personalized Ranking (DPR) Model

To mitigate RSP based bias:

- ➢ **Decouple the predicted score with group attribute;**
- Normalize the score distribution for each user to align predict score with ranking position.

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

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

To mitigate RSP based bias:

- ➢ **Decouple the predicted score with group attribute;**
- Normalize the score distribution for each user to align predict score with ranking position.

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

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

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

Debiased Personalized Ranking (DPR) Model

To mitigate RSP based bias:

- ➢ **Decouple the predicted score with group attribute;**
- Normalize the score distribution for each user to align predict score with ranking position.

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

The adversarial component takes predicted score as input and predict the group 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

To mitigate RSP based bias:

- Decouple the predicted score with group attribute;

➢ **Normalize the score distribution for each user to align predict score with ranking position.**

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

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

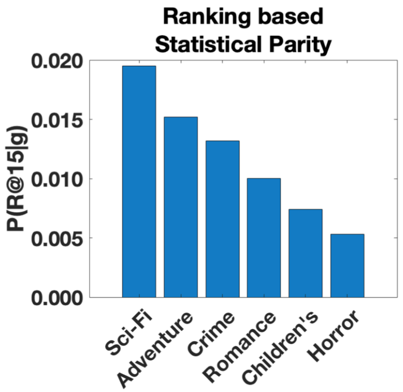
To mitigate REO based bias:

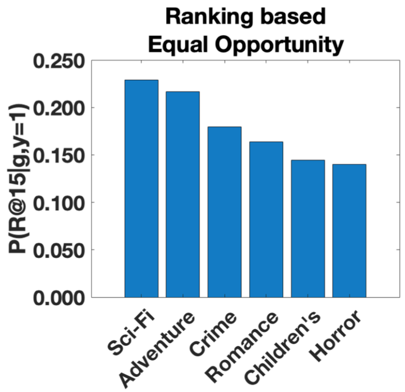
- Decouple the group attribute with the predicted score for **positive user-item pair**;
- Normalize the score distribution for each user to align predict score with ranking position.

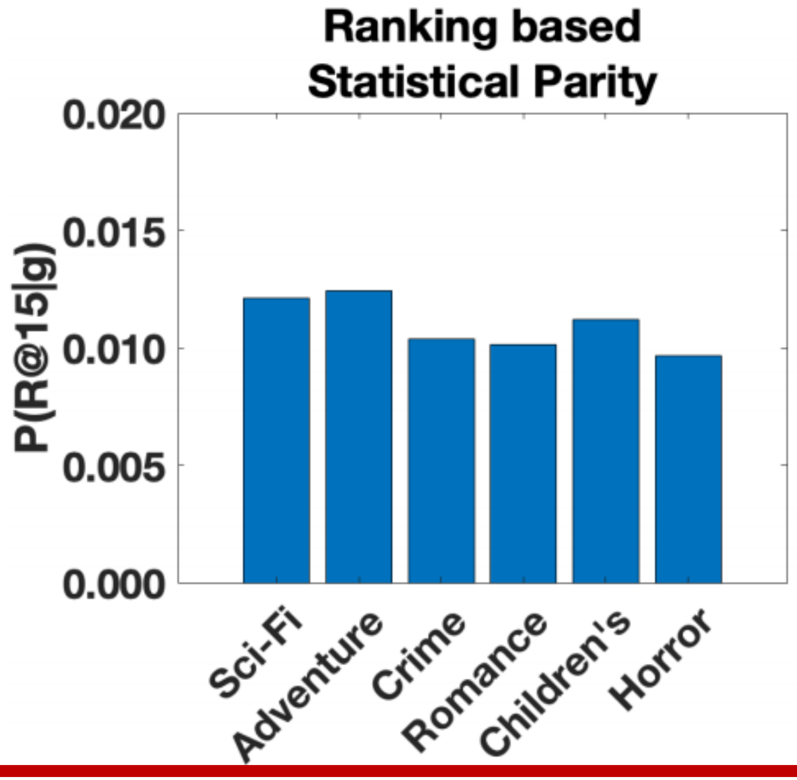
$$\min_{\Theta}\max_{\Psi}\sum_{u\in\mathcal{U}}\sum_{\substack{i\in\mathcal{I}_u^+\\ j\in\mathcal{I}\setminus\mathcal{I}_u^+}}\left(\mathcal{L}_{BPR}(u,i,j)+\boxed{\alpha\mathcal{L}_{Adv}(i)}\right)+\beta\mathcal{L}_{KL}$$

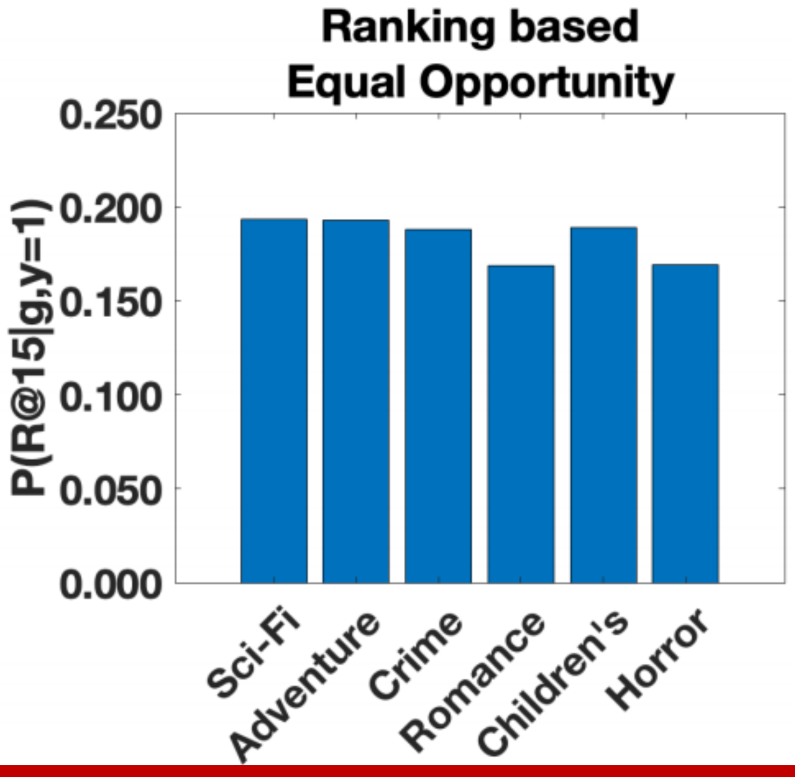
Only input scores for positive user-item pairs to the adversarial component.

Experiments – visualize debiased results



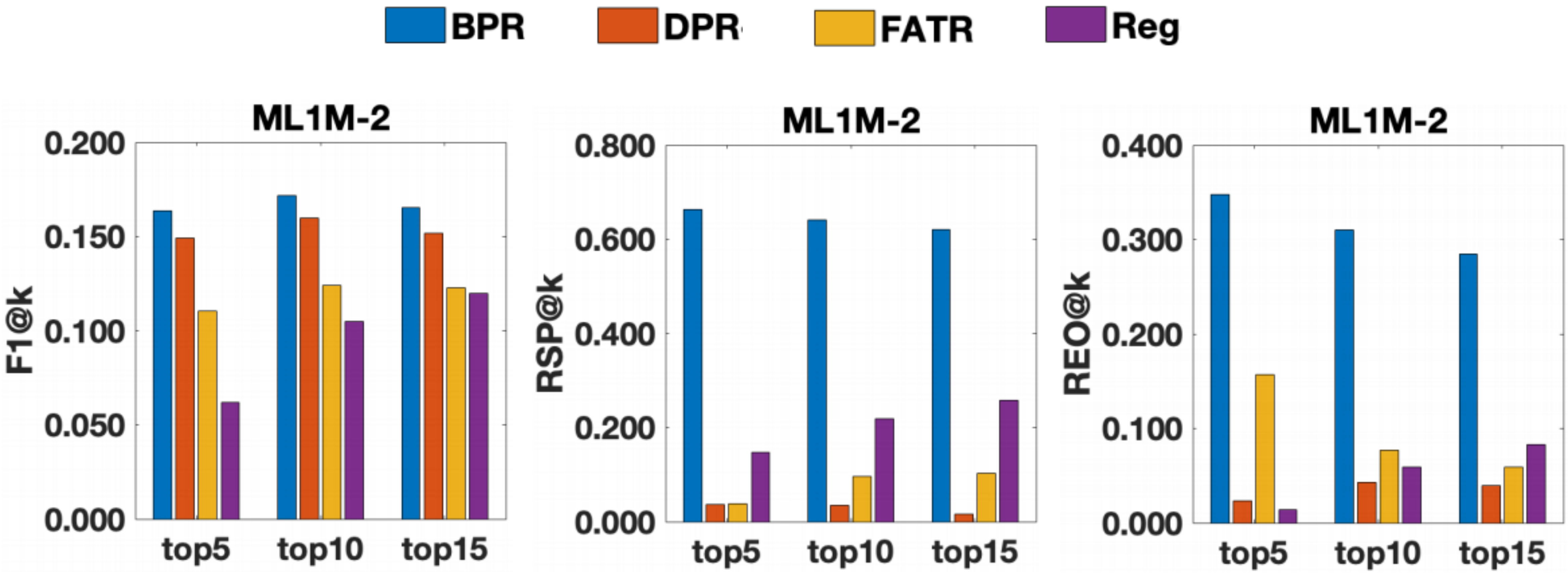






by the proposed DPR

Experiments – compare with baselines

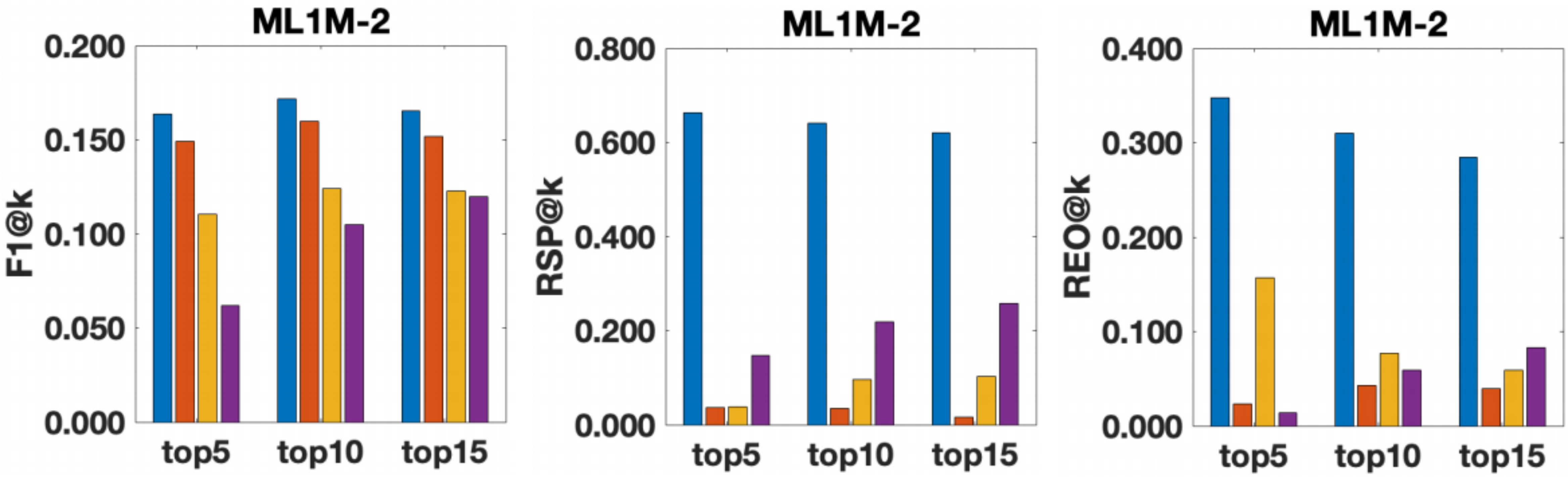


Experiments – compare with baselines

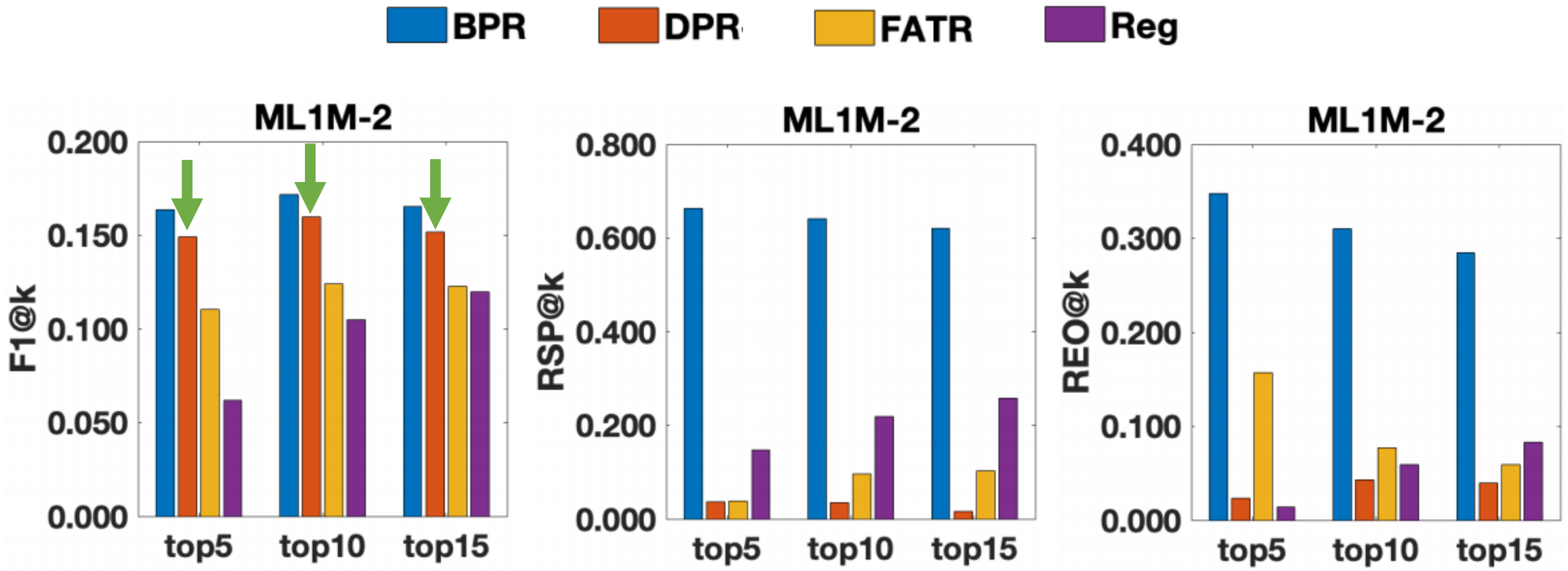
the proposed model

SOTA baselines

BPR DPR FATR Reg

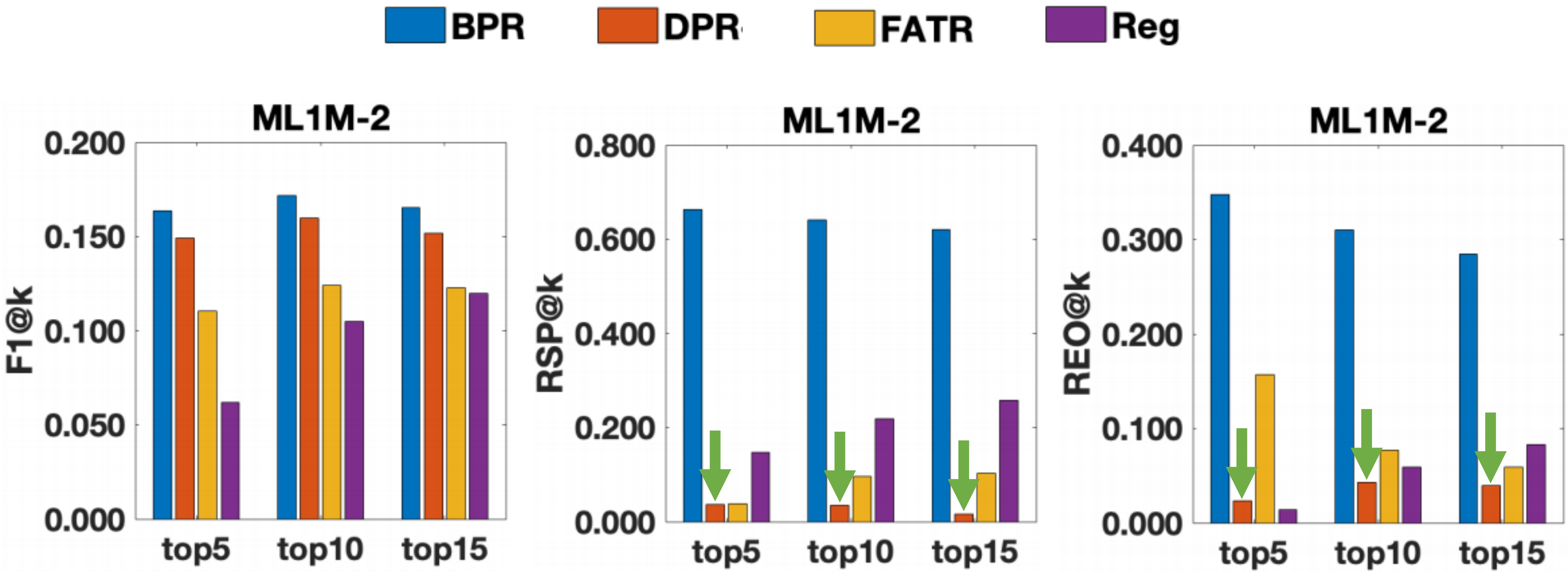


Experiments – compare with baselines



Proposed model **preserves high recommendation quality**, and enhance **RSP and REO fairness** effectively!

Experiments – compare with baselines



Proposed model **preserves high recommendation quality**, 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;
- Experiments with multi-group datasets;

Conclusions

- Propose two **ranking-based** under-recommendation bias **metrics**;
- Propose an **adversarial learning based model** which can mitigate the two studied recommendation bias;
- Experiments show the existence of bias in widely used BPR model, and show the **effectiveness** of the proposed debiasing model.

Thank You!

Ziwei Zhu, Jianling Wang, and James Caverlee

Texas A&M University

TEXAS A&M UNIVERSITY

SIGIR 2020
XI'AN·CHINA