NDBI021, Lecture 6

User preferences, 2/1 ZK+Z,

Wed 12:20 – 13:50 S8

Wed 14:00 – 15:30 SW2 (odd weeks)

https://www.ksi.mff.cuni.cz/~peska/vyuka/ndbi021/2022/

matfyz

siret research group

https://ksi.mff.cuni.cz

Fairness in Recommender Systems

RUTGERS
UNIVERSITY | NEW BRUNSWICK

Tutorial on Fairness of Machine Learning in Recommender Systems

Yunqi Li
Rutgers University
yunqi.li@rutgers.edu

Yingqiang Ge
Rutgers University
yingqiang.ge@rutgers.edu

Yongfeng Zhang
Rutgers University
yongfeng.zhang@rutgers.edu

Tutorial from: https://fairness-tutorial.github.io/

Fairness issues in RecSys and IR

- News recommendation/social networks
  - Does the suggested articles close me into some opinion bubble?
    - Fairness of the presented opinions on controversary subjects
- Job matching & marketplaces
  - Am I omitted from the list of possible applicants just because [black/old/female…]
  - Is one content provider favored over others?
- Finance domain
  - Why am I not recommended for loan? Why is my credit score lower/higher?
- E-commerce
  - Is this product being recommended because it is the best for me… or because the provider earns the most from it?

What if these features are learned indirectly?

Fairness in General

- Equality of opportunities
  - „You should not be disqualified /mistreated based on generic statistics that should not affect the outcome“
    - „You will not get the job because you are female“
    - What about already biased inputs?
- Equality of outcome
  - „Submission vs. acceptance ratio for male/female authors should not differ, if they differ, countermeasures should be taken“
    - Is this still fair?
    - Someone may be in „higher need“ of getting help vs. Someone had been mistreated in the past.
- Fairness vs. proportionality

RUTGERS

More on biases in RS soon...

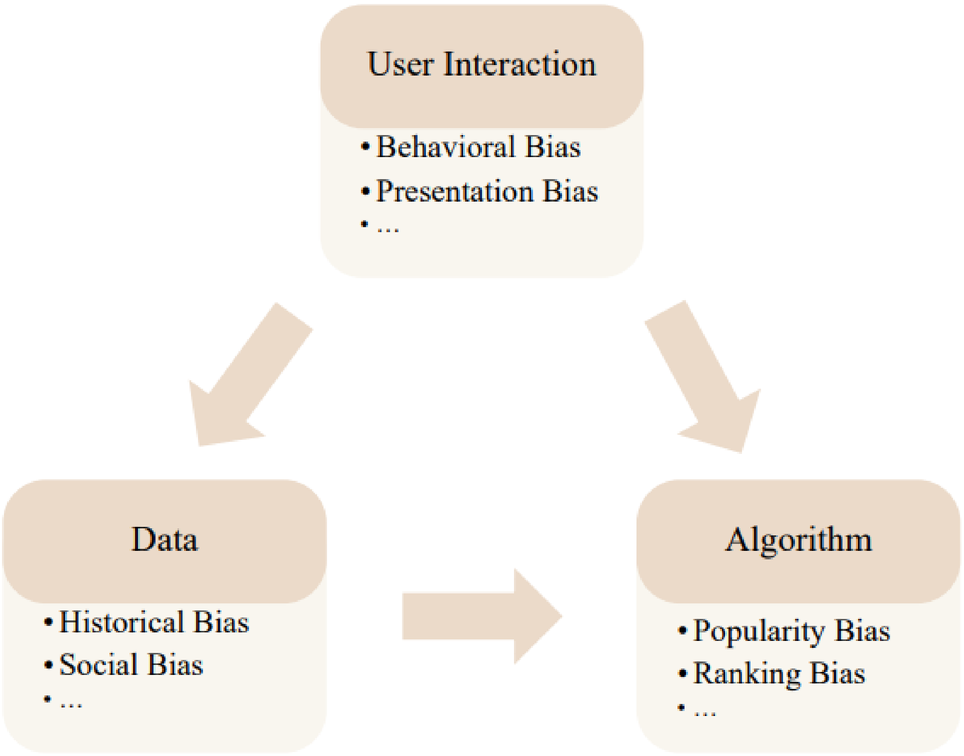
Fairness in Machine Learning — Causes

Data Bias

- Statistical Bias: non-random sample; record error
- Historical Bias: biased decision
- …

Algorithmic Bias

- Ranking Bias: exposure allocation
- Evaluation Bias: inappropriate benchmarks
- …



Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *arXiv preprint arXiv:1908.09635* (2019).
Castelnovo, Alessandro, et al. "The zoo of Fairness metrics in Machine Learning." *arXiv preprint arXiv:2106.00467* (2021).

Fairness in Machine Learning — Methods

Pre-processing	In-processing	Post-processing
Try to transform the data so that the underlying discrimination is removed.	Try to modify the learning algorithms to remove discrimination during the model training process.	Perform after training by accessing a holdout set which was not involved during the training of the model.

Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *arXiv preprint arXiv:1908.09635* (2019)

Fairness in Machine Learning – Basic tasks

Fairness in Classification

Fairness in Ranking

Fairness in Ranking – Introduction

List-wise definitions for fairness: depend on the entire list of results for a given query

Unsupervised criteria: the average **exposure** near the top of the ranked list to be **equal for different groups** [71][72][75]

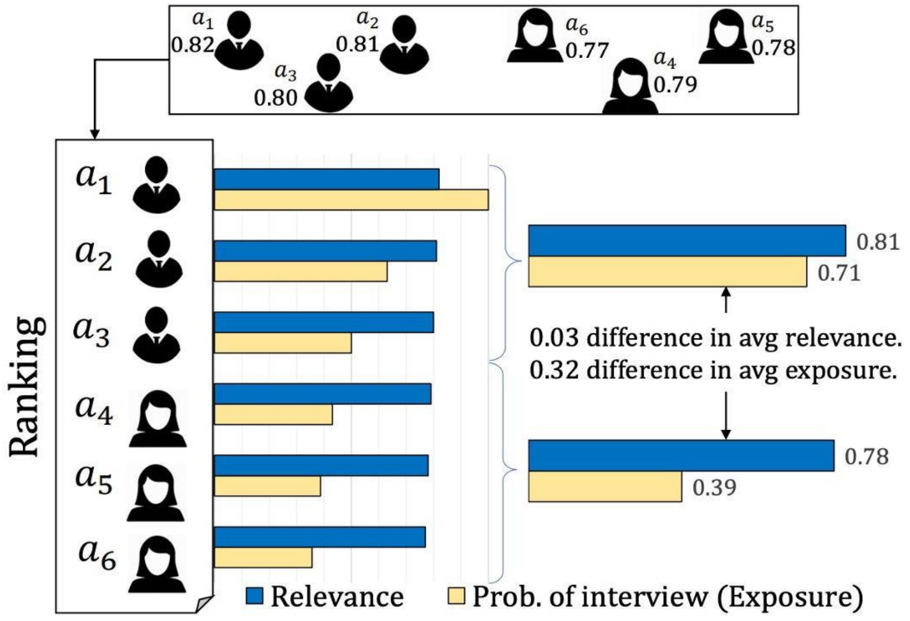
Supervised criteria: the average **exposure** for a group to be proportional to the average **relevance** of that group's results to the query [65][67]

Fairness in Ranking

- **Fairness Concerns**: A conceptual and computational framework that allows the formulation of fairness constraints on rankings in terms of **exposure allocation.**

- Job seeker example: a small difference in **relevance** can lead to a large difference in **exposure** (an opportunity) for the group of females.

Reasonable if relevance has direct probabilistic interpretation



Singh, Ashudeep, and Thorsten Joachims. "Fairness of exposure in rankings." *SIGKDD*’2018.

Fairness in Ranking

- **Method:** $r = \text{argmax}_r\, \text{U}(r|q)$ s.t. r is fair

- **Exposure** for a document d_i under a probabilistic ranking P as:

$$\text{Exposure}(d_i|\mathbf{P}) = \sum_{j=1}^{N} \mathbf{P}_{i,j}\mathbf{v}_j \quad \text{Exposure}(G_k|\mathbf{P}) = \frac{1}{|G_k|} \sum_{d_i \in G_k} \text{Exposure}(d_i|\mathbf{P})$$

- **Demographic Parity Constraints**:

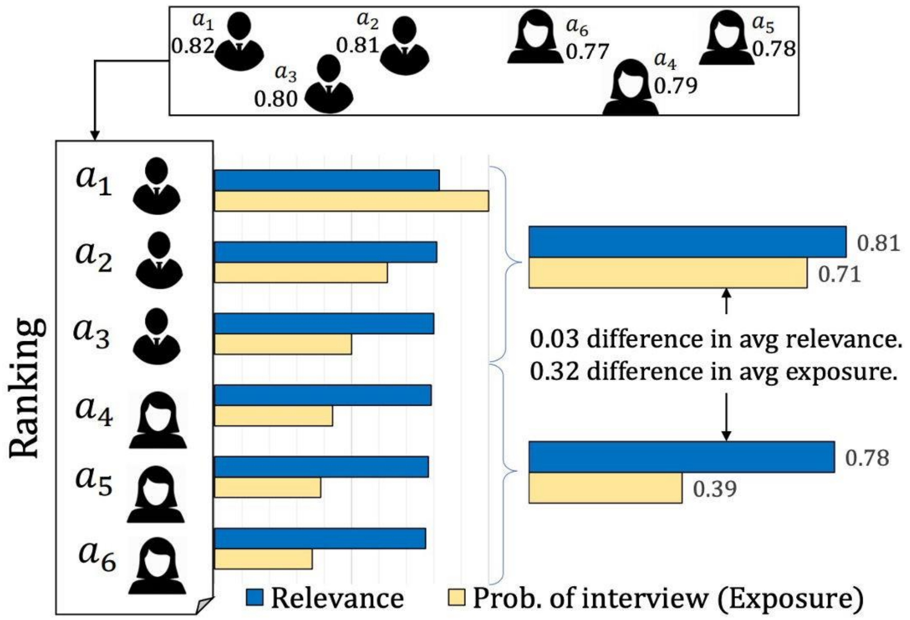
$$\text{Exposure}(G_0|\mathbf{P}) = \text{Exposure}(G_1|\mathbf{P}) \Leftrightarrow \mathbf{f}^T P\mathbf{v} = 0$$

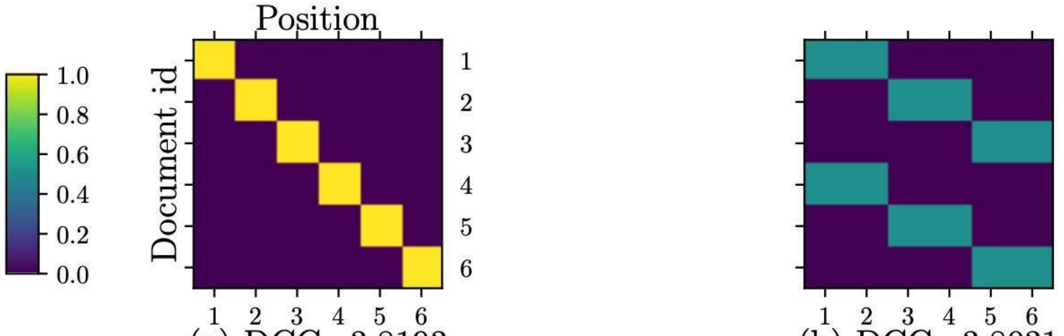
$$\left(\text{with } \mathbf{f}_i = \frac{\mathbb{1}_{d_i \in G_0}}{|G_0|} - \frac{\mathbb{1}_{d_i \in G_1}}{|G_1|}\right)$$

Singh, Ashudeep, and Thorsten Joachims. "Fairness of exposure in rankings." *SIGKDD*’2018

Fairness in Ranking

- Figure (a) is optimal unfair ranking that maximizes DCG.
- Figure (b) is optimal fair ranking under demographic parity.
- Compared to the DCG of the unfair ranking, the optimal fair ranking has slightly **lower utility** with a DCG.





(a) DCG=3.8193
(b) DCG=3.8031

Singh, Ashudeep, and Thorsten Joachims. "Fairness of exposure in rankings." *SIGKDD*'2018

Fairness in Ranking

Based on P_i,j = probability of document i being recommended at position j (for some query q)

- Linear programming

We will see in Section § 3.4 that not only does R imply a doubly stochastic matrix $\mathbf{P}$, but that we can also efficiently compute a probabilistic ranking R for every doubly stochastic matrix $\mathbf{P}$. We can, therefore, formulate the problem of finding the utility-maximizing probabilistic ranking under fairness constraints in terms of doubly stochastic matrices instead of distributions over rankings.

$$
\begin{aligned}
\mathbf{P} = \operatorname{argmax}_{\mathbf{P}} \ & \mathbf{u}^T \mathbf{P} \mathbf{v} && \text{(expected utility)} \\
\text{s.t. } & \mathbb{1}^T \mathbf{P} = \mathbb{1}^T && \text{(sum of probabilities for each position)} \\
& \mathbf{P}\mathbb{1} = \mathbb{1} && \text{(sum of probabilities for each document)} \\
& 0 \leq \mathbf{P}_{i,j} \leq 1 && \text{(valid probability)} \\
& \mathbf{P} \text{ is fair} && \text{(fairness constraints)}
\end{aligned}
$$

Note that the optimization objective is linear in N^2 variables $\mathbf{P}_{i,j}$, $1 \leq i, j \leq N$. Furthermore, the constraints ensuring that $\mathbf{P}$ is doubly stochastic are linear as well, where $\mathbb{1}$ is the column vector of size N containing all ones. Without the fairness constraint and for any $\mathbf{v}_j$ that decreases with j, the solution is the permutation matrix that ranks the set of documents in decreasing order of utility (conforming to the PRP).

Now that we have expressed the problem of finding the utility-maximizing probabilistic ranking, besides the fairness constraint, as a linear program, a convenient language to express fairness constraints would be linear constraints of the form

$$\mathbf{f}^T \mathbf{P} \mathbf{g} = h.$$

Individual fairness variants can be expressed via f & g vectors

3.5 Summary of Algorithm

The following summarizes the algorithm for optimal ranking under fairness constraints. Note that we have assumed knowledge of the true relevances $u(d|q)$ throughout this paper, whereas in practice one would work with estimates $\hat{u}(d|q)$ from some predictive model.

(1) Set up the utility vector $\mathbf{u}$, the position discount vector $\mathbf{v}$, as well as the vectors $\mathbf{f}$ and $\mathbf{g}$, and the scalar h for the fairness constraints (see Section § 4).
(2) Solve the linear program from Section § 3.3 for $\mathbf{P}$.
(3) Compute the Birkhoff-von Neumann decomposition $\mathbf{P} = \theta_1 \mathbf{P}_1 + \theta_2 \mathbf{P}_2 + \cdots + \theta_n \mathbf{P}_n$.
(4) Sample permutation matrix $\mathbf{P}_i$ with probability proportional to θ_i and display the corresponding ranking r_i.

of exposure in rankings." SIGKDD 2018

Fairness in Recommendation – Challenges

More Perspectives

Multiple Models And Goals

Extreme Data Sparsity

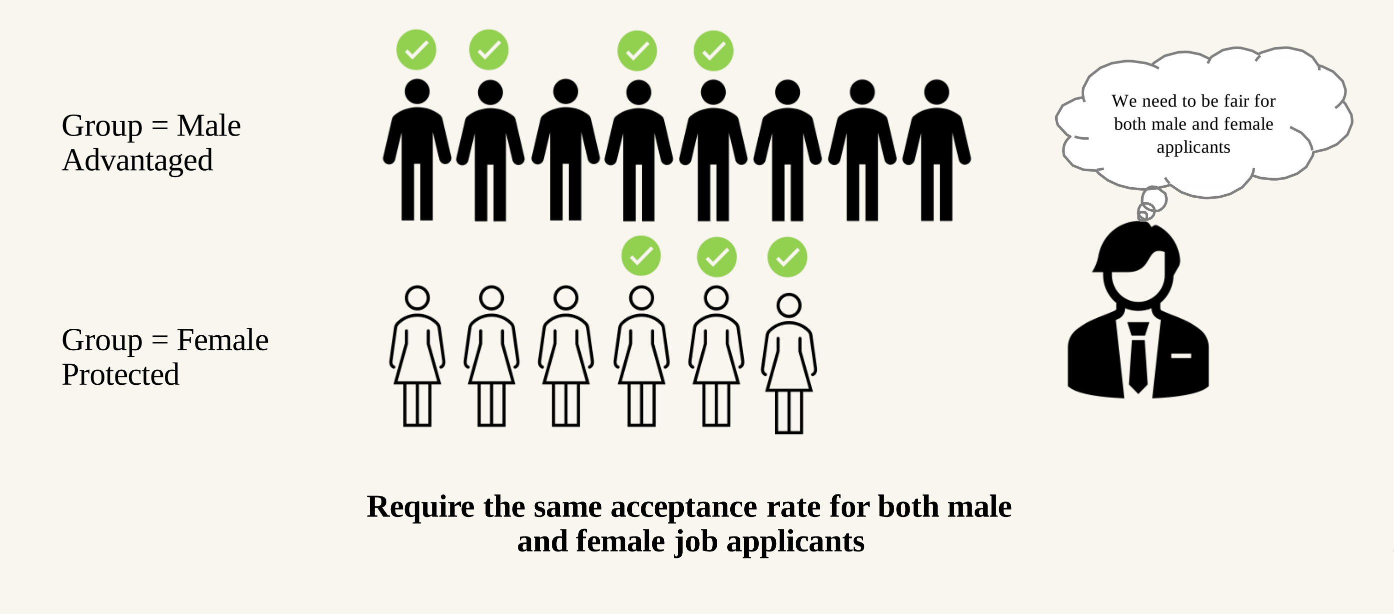
Dynamics

Taxonomies

- Group vs. Individual
- User vs. Item
- Association vs. Causality
- Single-sided vs. Multi-sided
- Static vs. Dynamic

Group Fairness vs. Individual Fairness

Group fairness requires that the protected groups should be treated similarly to the advantaged group.



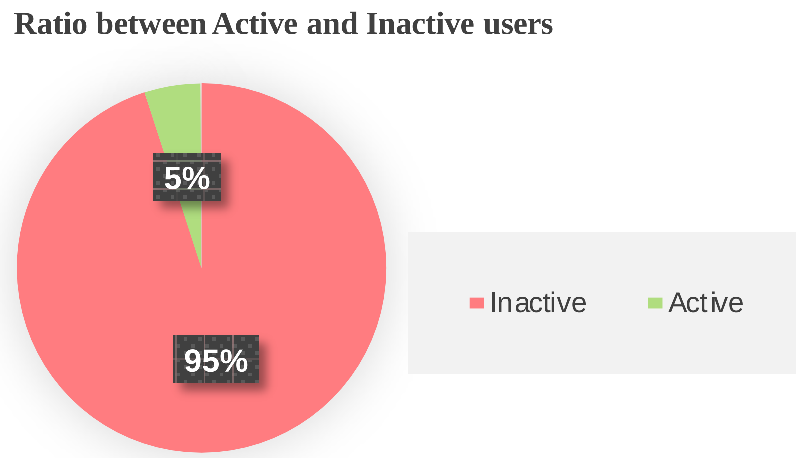
Group Fairness vs. **Individual Fairness**

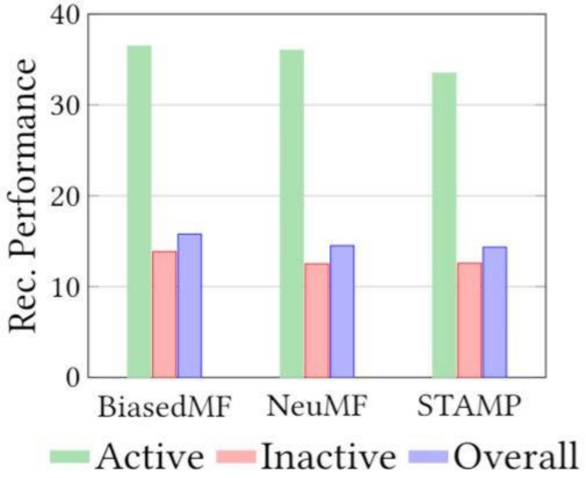
- Individual fairness requires that the similar individual should be treated similarly.

Image source: https://mitibmwatsonailab.mit.edu/wp-content/uploads/2020/04/Leadspace-GettyImages-598952582.jpg

Group Fairness in Recommendation

- **Fairness concerns**: The **unfair recommendation quality** between **user groups** with different activity levels, e.g., number of interactions.
- Unfairness of current recommender systems:
  - Active users only account for a **small** proportion of users.
  - The average recommendation quality on the small group (*active*) is **significantly better** than that on the remaining majority of users (*inactive*) for all baselines.

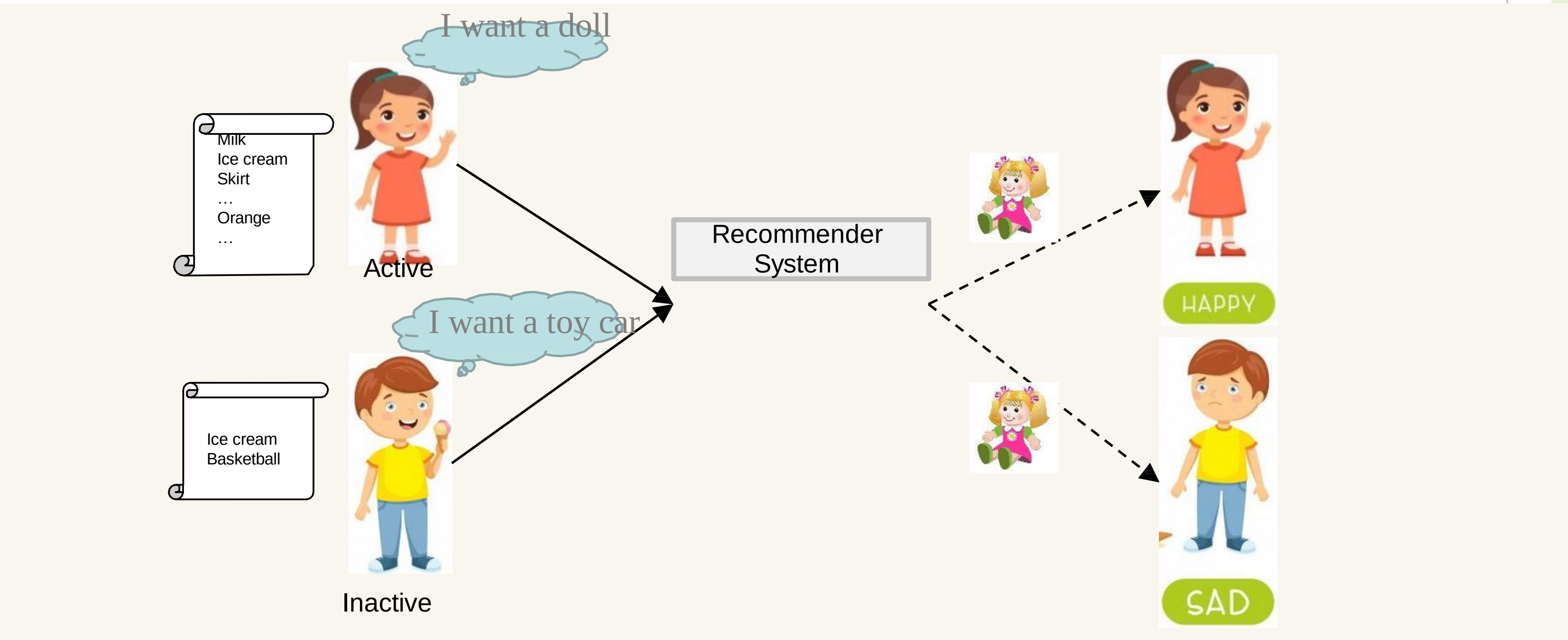




Li.Y et al. "User-oriented Fairness in Recommendation" WWW'21.

Active vs. Inactive groups

Fairness on user side: Fairness requirements in recommender systems may come from users.



Group Fairness in Recommendation

Fairness-aware Algorithm: A re-ranking method with user-oriented group fairness constrained on the recommendation lists generated from any base recommender algorithm.

$$\max_{\mathbf{W}_{ij}} \sum_{i=1}^{n}\sum_{j=1}^{N} \mathbf{W}_{ij}\boxed{S_{i,j}}$$

Preference of user i in terms of item j

$$\text{s.t.} \quad \boxed{\mathrm{UGF}(Z_1, Z_2, \mathbf{W}) < \varepsilon}$$

Fairness constraint

$$\boxed{\sum_{j=1}^{N} \mathbf{W}_{ij} = K, \mathbf{W}_{ij} \in \{0, 1\}}$$

Top-K list

Experiment Results: Improve fairness; Improve recommendation quality of overall and disadvantaged users. However, the performance of advantaged users is reduced to satisfy our fairness requirement.

			Beauty			
			Overall	Adv.	Disadv.	UGF
BiasedMF	F1	Orig.	14.27	30.68	12.77	17.91
		Fair	**15.06**	19.18	**14.68**	**4.50**
	NDCG	Orig.	43.25	67.79	41.00	26.79
		Fair	**43.97**	52.51	**43.19**	**9.32**

Improvement of overall accuracy

Disadv. ⇧

Adv. ⇩

Improvement of fairness

Li.Y et al. "User-oriented Fairness in Recommendation" WWW'21.

Individual Fairness in Recommendation

- **Fairness concerns**: the position bias which leads to disproportionately: less attention being paid to low-ranked subjects (position bias).

- No single ranking can achieve individual attention fairness.

- **Equity of Amortized Attention**: A sequence of rankings $\{1,2, \ldots m\}$ offer equity of amortized attention if each subject u receives cumulative attention proportional to her cumulative relevance:

$$\frac{\boxed{\sum_{l=1}^{m} a_{i1}^{l}}}{\boxed{\sum_{l=1}^{m} r_{i1}^{l}}} = \frac{\sum_{l=1}^{m} a_{i2}^{l}}{\sum_{l=1}^{m} r_{i2}^{l}}, \forall u_{i1}, u_{i2}$$

attention

relevance

Biega, A. J. et al. "Equity of Attention: Amortizing Individual Fairness in Rankings" SIGIR'18.

Individual Fairness in Recommendation

- **Method (Offline optimization):**

$$\text{minimize} \quad \boxed{\sum_i |A_i - R_i|}$$

▶ Fairness (L1 norm over distributions)

$$\text{subject to} \quad \boxed{NDCG\text{-}quality@k(\rho^j, \rho^{j*}) \geq \theta, \; j = 1, \ldots, m.}$$

▶ Ranking quality

- **Experiment Results:**
  - **Improving equity of attention is crucial**: the discrepancy between the attention received and the deserved attention can be substantial.
  - Improving equity of attention can often be done **without sacrificing much quality** in the rankings.

Integer linear programming (re-ranking) https://en.wikipedia.org/wiki/Integer_programming

Biega, A. J. et al. "Equity of Attention: Amortizing Individual Fairness in Rankings" SIGIR'18.

Associative Fairness vs. Causal Fairness

Find the **discrepancy of statistical metrics** between individuals or sub-populations.

In **binary classification**, fairness metrics can be represented by regularizing the classifier's positive or negative rates over different protected groups.

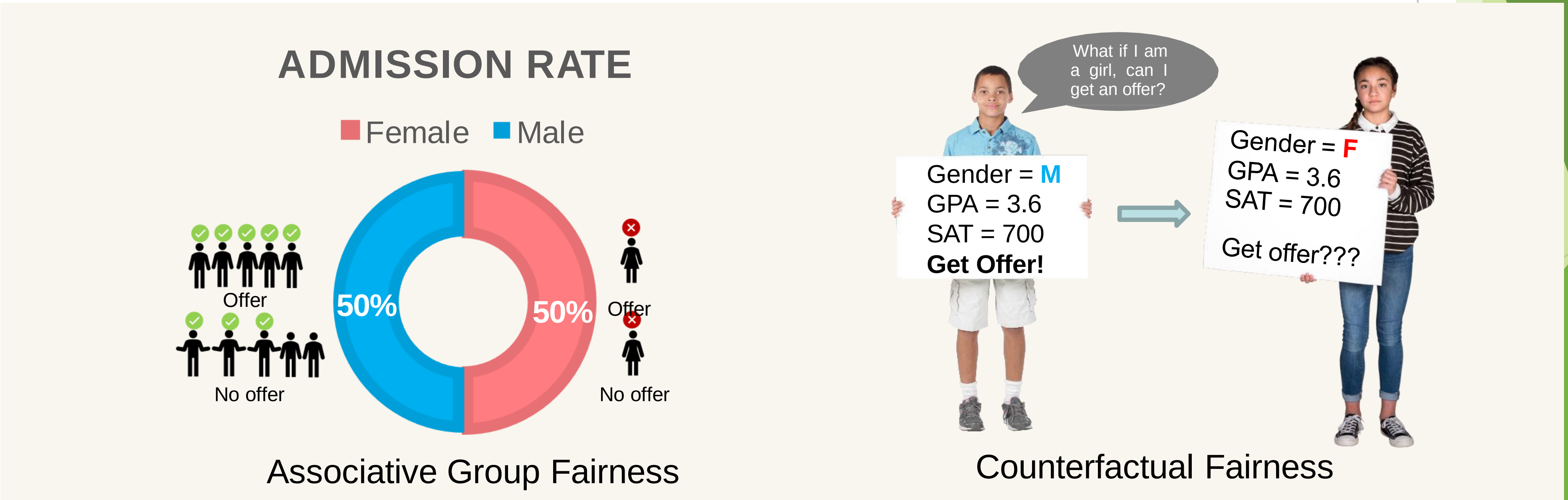
Associative Fairness vs. **Causal Fairness**

- Fairness cannot be well assessed only based on association notions [46-49].
- Difference:
  - Reason about the **causal relations** between the protected features and the model outcomes.
  - Leverage **prior knowledge** about the world structure in the form of causal models, help to understand the propagation of variable changes in the system.

https://en.wikipedia.org/wiki/Causal_model

Counterfactual fairness

- Counterfactual fairness is an individual-level causal-based fairness notion. It requires that for any possible individual, the predicted result of the learning system should be the **same** in th**e counterfactual world** as in the **real world.**



Associative Fairness in Recommendation

- **Method**:

$$\min_{\boldsymbol{P},\boldsymbol{Q},\boldsymbol{u},\boldsymbol{v}} \boxed{J(\boldsymbol{P},\boldsymbol{Q},\boldsymbol{u},\boldsymbol{v})} + \boxed{U}$$

Loss for recommender model

Fairness constraint

- **Experiment Results**: the experiments on synthetic and real data show that minimization of these forms of unfairness is possible with no significant increase in reconstruction error.

Unfairness	Error	Value	Absolute	Underestimation	Overestimation	Non-Parity
None	0.887 ± 1.9e-03	0.234 ± 6.3e-03	0.126 ± 1.7e-03	0.107 ± 1.6e-03	0.153 ± 3.9e-03	0.036 ± 1.3e-03
Value	0.886 ± 2.2e-03	**0.223 ± 6.9e-03**	0.128 ± 2.2e-03	**0.102 ± 1.9e-03**	**0.148 ± 4.9e-03**	0.041 ± 1.6e-03
Absolute	0.887 ± 2.0e-03	0.235 ± 6.2e-03	**0.124 ± 1.7e-03**	0.110 ± 1.8e-03	0.151 ± 4.2e-03	0.023 ± 2.7e-03
Under	0.888 ± 2.2e-03	0.233 ± 6.8e-03	0.128 ± 1.8e-03	**0.102 ± 1.7e-03**	0.156 ± 4.2e-03	0.058 ± 9.3e-04
Over	**0.885 ± 1.9e-03**	0.234 ± 5.8e-03	**0.125 ± 1.6e-03**	0.112 ± 1.9e-03	**0.148 ± 4.1e-03**	0.015 ± 2.0e-03
Non-Parity	0.887 ± 1.9e-03	0.236 ± 6.0e-03	0.126 ± 1.6e-03	0.110 ± 1.7e-03	0.152 ± 3.9e-03	**0.010 ± 1.5e-03**

Yao, Sirui, and Bert Huang. “Beyond Parity: Fairness Objectives for Collaborative Filtering” NIPS’17

Causal Fairness in Recommendation

- **Fairness Concerns**: Counterfactual fairness for users in recommendations.

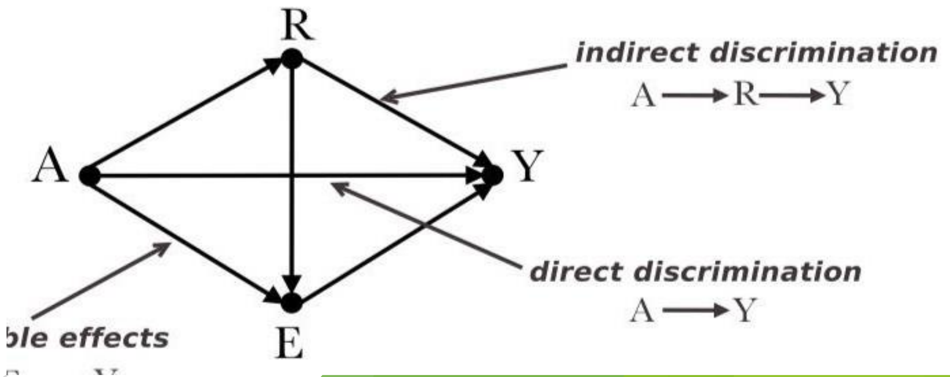
- **Definition:** A recommender model is *counterfactually fair* if for any possible user u with features $X = x$ and $Z = z$, for all L, and for any value z' attainable by Z:

$$P(\boxed{L_z} \mid \boxed{X} = x, \boxed{Z} = z) = P(L_{z'} \mid X = x, Z = z)$$

Top-N recommendation list for user u with sensitive features z

Insensitive features

Sensitive features



Li. Y et al. "Towards Personalized Fairness based on Causal Notion" SIGIR'21

Causal Fairness in Recommendation

- **Fairness Concerns**: Counterfactual fairness for users in recommendations.
- **Definition:** A recommender model is *counterfactually fair* if for any possible user u with features $X = x$ and $Z = z$, for all L, and for any value z' attainable by Z:

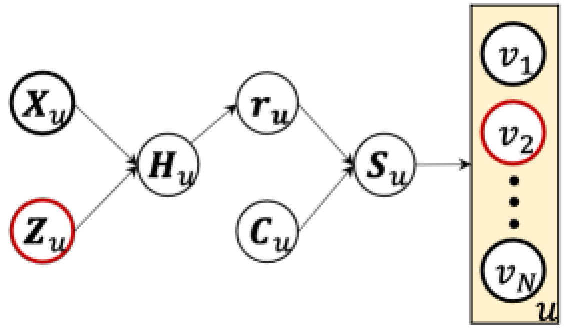


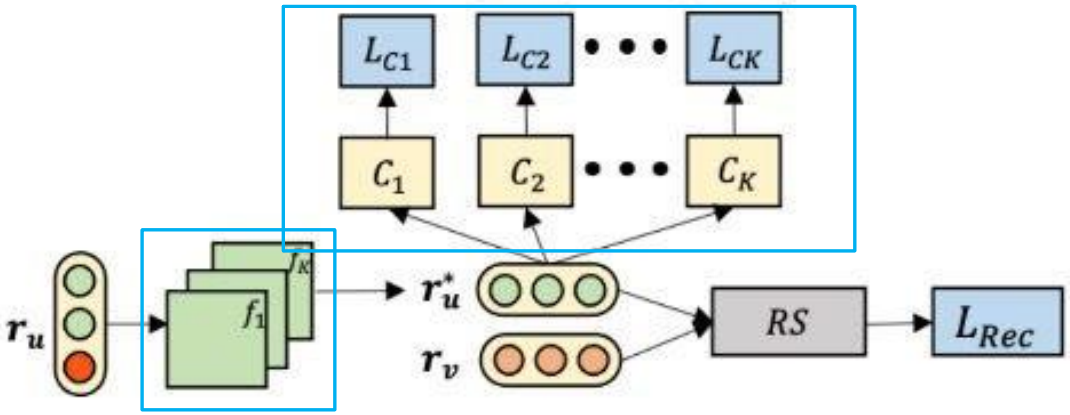
Figure 2: Causal relations for general recommendation models. For a given user u, X_u and Z_u are insensitive and sensitive features of u, respectively. H_u is the user interaction history. r_u is the user embedding. C_u is the candidate item set for u. S_u are the predicted scores over the candidate items. The red circled nodes are used to emphasize the impact of the sensitive features on the final recommendation list.

Li. Y et al. "Towards Personalized Fairness based on Causal Notion" SIGIR'21

Causal Fairness in Recommendation

- **Method:** Generate feature independent user embeddings through *adversary learning*.
  - **Filter Module**: filter the information about sensitive features from user embeddings
  - **Discriminator module:** predict the sensitive features from the learned user embeddings.
- **Experiment Results**:
  - Improve fairness
  - A little sacrifice on recommendation performance

Discriminator module



Filter module

		MoiveLens		
		AUC-G	AUC-A	AUC-O
PMF	Orig.	0.7697	0.8428	0.6024
	SM	**0.5389**	**0.5560**	**0.5289**
	CM	0.5532	0.5951	0.5396

Li. Y et al. "Towards Personalized Fairness based on Causal Notion" SIGIR'21

Fairness in Group Recommendation

- **Group Recommendation**: recommend items to groups of users whose preferences can be different from each other.
- **Fairness Concerns**: maximize the satisfaction of each group member while minimizing the unfairness (the imbalance of user utilities inside the group) between them.
- **Fairness Definitions**:
  - Least Misery: $F_{LM}(g, I) = \min\{\boxed{U(u, I)}, \forall u \in g\}$
  - Variance: $F_{Var}(g, I) = 1 - Var(\{U(u, I), \forall u \in g\})$
  - Jain's Fairness: $F_J(g, I) = \dfrac{\left(\sum_{u \in g} U(u, I)\right)^2}{|U| \cdot \sum_{u \in g} U(u, I)^2}$
  - Min-Max Ratio: $F_M(g, I) = \dfrac{\min\{U(u, I), \forall u \in g\}}{\max\{U(u, I), \forall u \in g\}}$

The individual utility of user *u* in group *g* when a set of items *I* are recommended to the group.

Xiao, Lin, et al. "Fairness-aware group recommendation with pareto-efficiency." Recsys'17

Fairness in Group Recommendation

- **Group Recommendation**: recommend items to groups of users whose preferences can be different from each other.
- **Why not just aggregate individual preferences of users?**

Xiao, Lin, et al. "Fairness-aware group recommendation with pareto-efficiency." Recsys'17

Fairness in Group Recommendation

- **Method**:
  - The **Social Welfare** ($SW(g, I)$): overall utility of all users inside the group g given a group recommendation I.
  - The **Fairness** ($F(g, I)$): a function of $U(u, I), \forall u \in g, \forall I$.
  - Multi-Objective Optimization: $\lambda \cdot SW(g, I) + (1 - \lambda) \cdot F(g, I)$
- **Experiment Results**: The results indicate that considering fairness can improve the quality of group recommendation.

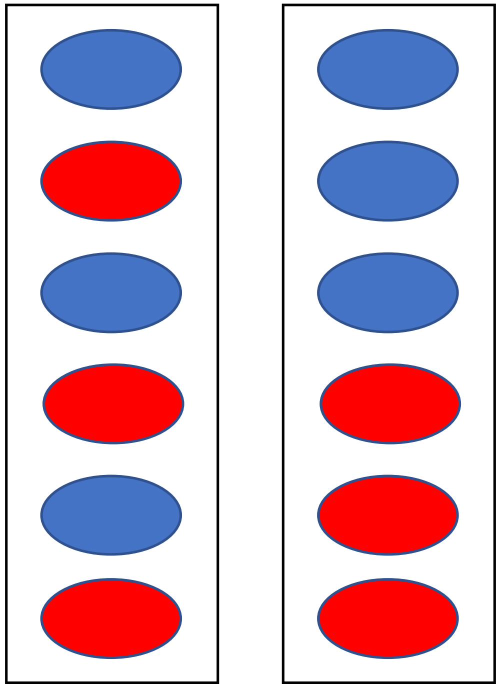
λ, RG	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
F@K	0.0260	0.0817	0.0877	0.0953	0.1019	0.1041	0.1046	0.1053	0.1058	0.1062	0.1062
NDCG@K	0.0697	0.2200	0.2287	0.2334	0.2394	0.2423	0.2440	0.2421	0.2459	0.2478	0.2476

Xiao, Lin, et al. "Fairness-aware group recommendation with pareto-efficiency." Recsys'17

Fairness in Group Recommendation

Possible issues:

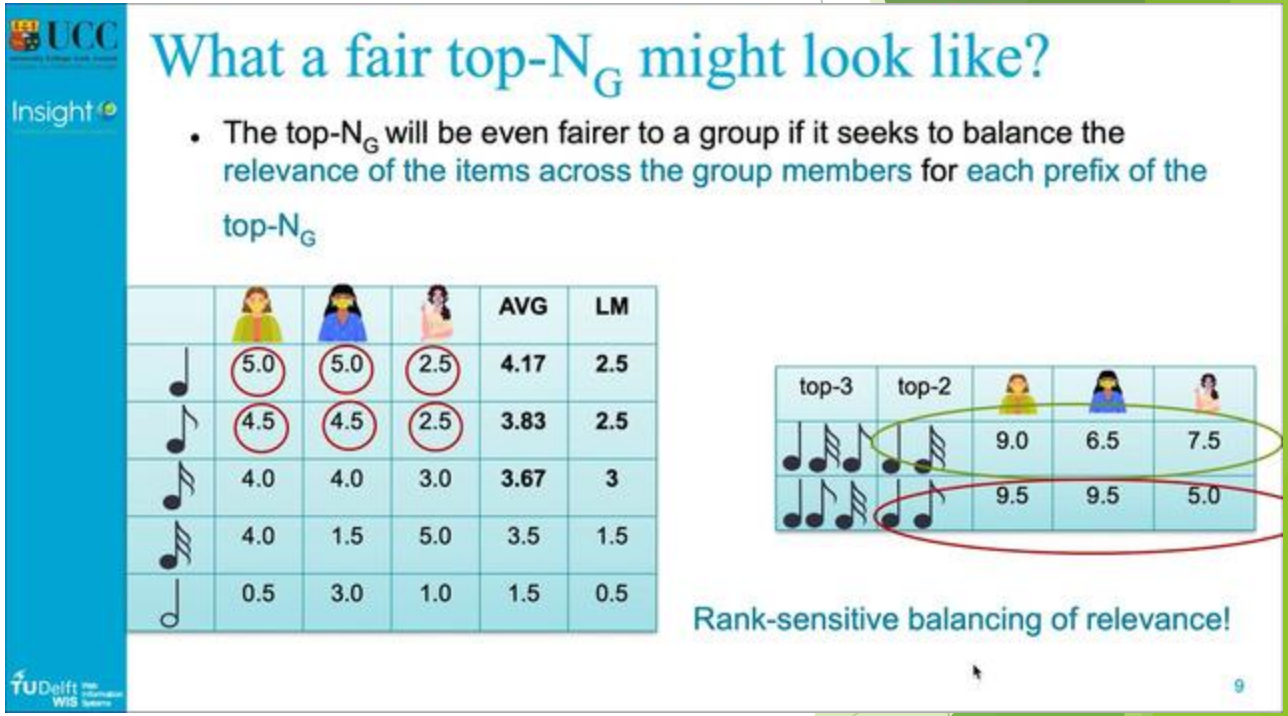
- Fairness metrics does not consider ranking



Fairness in Group Recommendation

Possible issues:

- Ranking aware fairness
- Greedy algorithm GFAR



https://slideslive.com/38934807/ensuring-fairness-in-group-recommendations-by-ranksensitive-balancing-of-relevance?ref=speaker-41949
https://dl.acm.org/doi/10.1145/3383313.3412232

Fairness in Group Recommendation

3.1 GFAR's definition of fairness

For a group member $u \in G$, let $p(\mathrm{rel}\,|u,i)$ be the probability that item i is relevant to u. We estimate $p(\mathrm{rel}\,|u,i)$ as:

$$p(\mathrm{rel}\,|u,i) = \frac{\text{Borda-rel}(u,i)}{\sum_{j \in \text{top-}N_u} \text{Borda-rel}(u,j)} \quad (1)$$

Following Xiao et al. [18], we define Borda-rel$(u,i) = |\{j : \mathrm{rank}(j, \text{top-}N_u) > \mathrm{rank}(i, \text{top-}N_u), \forall_j \in \text{top-}N_u\}|$, where, from above, $\mathrm{rank}(i, \text{top-}N_u)$ is the rank of item i in u's top-N candidate items, which are obtained using the $s(u,i)$ scores predicted by the underlying recommender algorithm.[2]

Let also $p(\neg\,\mathrm{rel}\,|u,S)$ be the probability that none of the items in set S are relevant to user u. Then, we derive the probability that at least one item within S is relevant to u, $p(\mathrm{rel}\,|u,S)$, as follows:

$$\begin{aligned} p(\mathrm{rel}\,|u,S) &= 1 - p(\neg\,\mathrm{rel}\,|u,S) \\ &= 1 - \prod_{i \in S}(1 - p(\mathrm{rel}\,|u,i)) \end{aligned} \quad (2)$$

Now, from $p(\mathrm{rel}\,|u,S)$ for each group member $u \in G$, we define $f(S)$ as the sum of each group member's probability of finding at least one relevant item within the set S:

$$f(S) = \sum_{u \in G} p(\mathrm{rel}\,|u,S) = \sum_{u \in G}\left(1 - \prod_{i \in S}(1 - p(\mathrm{rel}\,|u,i))\right) \quad (3)$$

[2] A more obvious definition is $p(\mathrm{rel}\,|u,i) = s(u,i)/\sum_{j \in C} s(u,j)$, where $C \subseteq I$ are the candidate items. Compared to Eq. 1, this did not work well in our experiments. The probable explanation is that it relies too heavily on the actual $s(u,i)$ values, whereas Eq. 1 uses their ordering.

Eq. 3 shows how to 'balance' relevance across the group members for a set. It is not yet rank-sensitive. To make it rank-sensitive, we define the marginal gain in function f that arises when we add a new item to the set S, $f(i,S)$, as:

$$f(i,S) = f(S \cup \{i\}) - f(S) \quad (4)$$

Using Eq. 3 and Eq. 4, we can obtain the following:

$$f(i,S) = \sum_{u \in G}[p(\mathrm{rel}\,|u,i) \prod_{j \in S}(1 - p(\mathrm{rel}\,|u,j))] \quad (5)$$

Then, we can define an ordered set to be fair if there is balance in each prefix of the set. In other words, the first item in the set should, as far as possible, balance the interests of all group members; the first two items taken together must do the same; also the first three; and so on up to N:

$$\mathrm{fair}(OS) = \sum_{k=1}^{|OS|} f(OS[k], \{i \in OS : \mathrm{rank}(i, OS) < k\}) \quad (6)$$

A natural alternative is to find an approximation of OS^* using a greedy algorithm. The GFAR greedy algorithm starts with an empty set, $OS = \{\}$. At each iteration, it inserts into the ordered result set the item i^* from the remaining candidates (i.e. $C \setminus OS$) that gives the highest marginal gain:

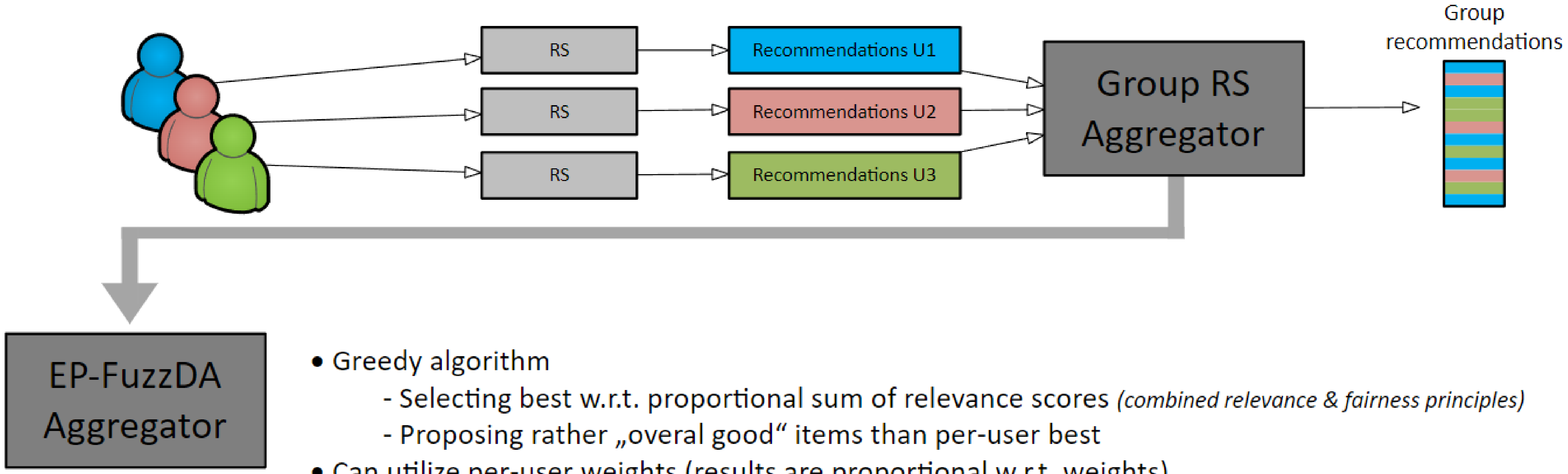
$$i^* = \arg\max_{i \in C \setminus OS} f(i, OS) \quad (8)$$

Drastic decrease of relevance w.r.t. order of items

Finding one item per user is sufficient

Fairness-preserving Group Recommendations With User Weighting

Ladislav Malecek, Ladislav Peska



- Greedy algorithm
  - Selecting best w.r.t. proportional sum of relevance scores *(combined relevance & fairness principles)*
  - Proposing rather „overal good“ items than per-user best
- Can utilize per-user weights (results are proportional w.r.t. weights)
  - Suitable for long-term fairness of permanent groups

See the paper for algorithm details & evaluation

Source codes: https://github.com/LadislavMalecek/UMAP2021

Fairness in Group Recommendation

We consider group recommending strategy as fair if all users receive items with approx. the same sum of estimated relevance scores. This statement should hold for all prefixes of the recommendations list.

Input: group members $u \in \mathcal{G}$, andidate items $c \in C$, relevance scores $r_{u,c} \in \mathbf{R}$, number of intems k, user's weights v_u; $\sum v_u = 1$

Output: ordered list of group recommendations L_G^k

$L_G = []$; $TOT = 0$; $\forall u : r_u = 0$
for $i \in [0, ..., k]$ **do**
  for $c \in C \setminus L_G$ **do**
    $TOT_c = TOT + \sum_{\forall u} r_{u,c}$
    $\forall u : e_u = max(0, TOT_c * v_u - r_u)$
    $gain_c = \sum_{\forall u} min(r_{u,c}, e_u)$
  end for
  $c_{best} = \text{argmax}_{\forall c}(gain_c)$; append c_{best} to L_G
  $\forall u : r_u = r_u + r_{u,best}$; $TOT = \sum_{\forall u} r_u$
end for
return L_G

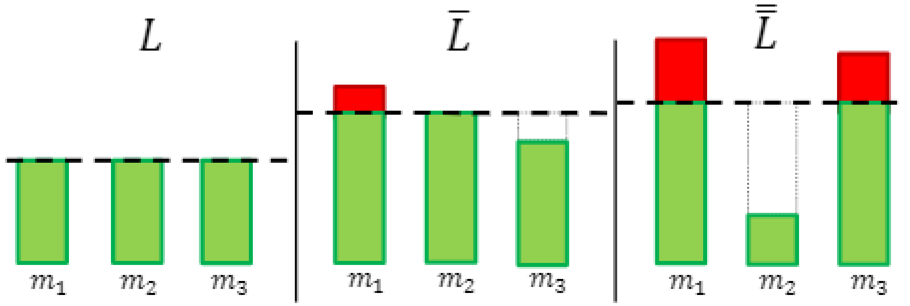


Figure 1: Example of proportionality vs. utility tradeoff. Figure depicts three lists ($L, \bar{L}, \bar{\bar{L}}$) and their score w.r.t. evaluation metrics m_1, m_2, m_3. For simplicity, consider that all metrics has the same weight ($w_1 = w_2 = w_3$). Dashed line denotes exactly proportional fraction of the total utility (i.e. mean utility in the case of equal weights), green bars denote proportional part of metric's utility, while red bar denotes excess over it. L provides perfectly proportional results, but its overall utility is inferior. $\bar{\bar{L}}$ has the highest mean utility (dashed line), but it is highly disproportional. We consider $\bar{L}$ to be the best option as the sum proportional fractions of metric's utility (sum of green bars) is largest.

TOT_c: total relevance of so far recommended objects (plus the relevance of the considered one)

e_u: not yet accounted relevance share of the current user (how much did we ignored this user in the past?)

- **v_u:** weight of individual user. Can e.g. adjust the lack of fairness in previous recommendation sessions

$gain_c$: sum of per-user relevances of considered item (but only the fair portion of per-user relevances are considered)

- for example, if some user is over-represented and his/her $e_u = 0$, relevance w.r.t. this user is completely ignored when calculating the best next object.

Fairness in evaluation

- Popularity bias (more popular => much more attention)
- Biased historical data (missing not at random) => (unbiased) learning algorithm => biased recommendations
- => biased off-line evaluation (same bias vector => better results)
- => discrepancy between off-line and on-line evaluation

- How to evaluate methods fairly?

Fairness in evaluation

- Inverse propensity score
- Weight results by the inverse to the propensity score
  - (probability of being noticed by the user)
  - Definitions may vary on available information
    - Based on general item’s popularity
    - Based on recommended positions
    - Based on user’s actions within the page

De-biasing Off-line Evaluation

- https://dl.acm.org/doi/pdf/10.1145/3240323.3240355

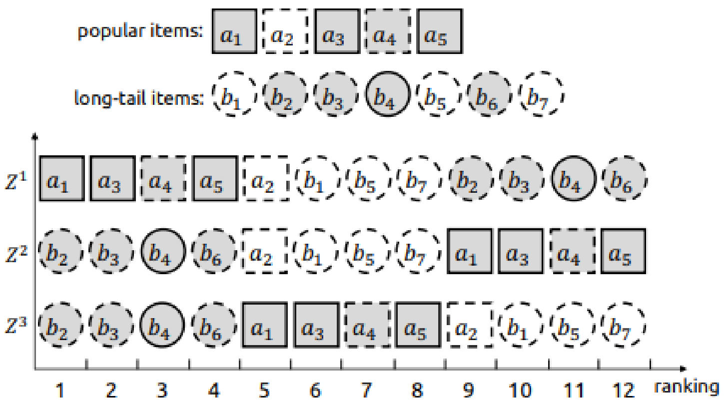


Figure 1: A hypothetical example to illustrate the evaluation bias that results from use of the AOA evaluator. Three recommenders generated distinct lists of recommendations, Z^1, Z^2 and Z^3, for the same user. Among the shaded items that were preferred by the user, the ones with a solid border were observed by recommenders. The performance was measured by DCG, and the results are presented in Table 1.

Table 1: The true and estimated DCG values for three recommenders in Fig. 1. $R(\hat{Z})$ denotes the ground truth, and $\hat{R}_{\text{AOA}}(\hat{Z})$ denotes the AOA estimations. The AOA estimator outputs larger values when popular items are ranked higher.

Estimator	Z^1	Z^2	Z^3
$R(\hat{Z})$	0.463	0.463	0.494
$\hat{R}_{\text{AOA}}(\hat{Z})$	0.585	0.340	0.390

3.1 Average-over-all (AOA) evaluator

In prior literature, $R(\hat{Z})$ was estimated by taking the average over all observed user feedback $\mathcal{S}_u^*$:

$$\begin{aligned}\hat{R}_{\text{AOA}}(\hat{Z}) &= \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\frac{1}{|\mathcal{S}_u^*|}\sum_{i\in\mathcal{S}_u^*} c(\hat{Z}_{u,i}) \\ &= \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\frac{1}{\sum_{i\in\mathcal{S}_u} O_{u,i}}\sum_{i\in\mathcal{S}_u} c(\hat{Z}_{u,i})\cdot O_{u,i}\end{aligned} \tag{6}$$

nDCG, AUC, MAP,...

3.2 Unbiased evaluator

To conduct unbiased evaluation of biased observations, we leverage the IPS framework [16, 22] that weights each observation with the inverse of its propensity, where the term *propensity* refers to the tendency or the likelihood of an event happening. The intuition is to down-weight the commonly observed interactions, while up-weighting the rare ones. In the context of this paper, the probability $P_{u,i}$ is treated as the pointwise propensity score. Therefore, the IPS unbiased evaluator is defined as follows:

$$\begin{aligned}\hat{R}_{\text{IPS}}(\hat{Z}|P) &= \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\frac{1}{|\mathcal{S}_u|}\sum_{i\in\mathcal{S}_u^*}\frac{c(\hat{Z}_{u,i})}{P_{u,i}} \\ &= \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\frac{1}{|\mathcal{S}_u|}\sum_{i\in\mathcal{S}_u}\frac{c(\hat{Z}_{u,i})}{P_{u,i}}\cdot O_{u,i}\end{aligned} \tag{7}$$

Propensity score

$$\hat{P}_{*,i} \propto (n_i^*)^{\gamma}\cdot n_i, \tag{13}$$

where $n_i = \sum_{u\in\mathcal{U}} \mathbf{1}[i \in \mathcal{S}_u]$ and $n_i^* = \sum_{u\in\mathcal{U}, i\in\mathcal{S}_u^*} O_{*,i}$.

However, empirically, n_i is not directly observable. To addres this problem, we observe that n_i^* is sampled from a binomial distri bution[4] parameterized by n_i, that is, $n_i^* \sim \mathcal{B}(n_i, P_{*,i})$. Therefore, a relationship between n_i and n_i^* can be built by bridging the genera tive model (eqn. 13) with the following unbiased estimator:

$$\hat{P}_{*,i} = \frac{n_i^*}{n_i} \propto (n_i^*)^{\gamma}\cdot n_i \tag{14}$$

Therefore, $n_i \propto (n_i^*)^{\frac{1-\gamma}{2}}$. We use this as a replacement for the unobserved n_i in eqn. 13, which results in an unbiased $\hat{P}_{*,i}$ estimato that is determined by only the empirical counts of items:

$$\hat{P}_{*,i} \propto (n_i^*)^{\left(\frac{\gamma+1}{2}\right)} \tag{15}$$

Fairness in RS, further reading

- https://link.springer.com/article/10.1007/s11257-020-09285-1
- https://dl.acm.org/doi/pdf/10.1145/3383313.3411545
- https://www.sciencedirect.com/science/article/pii/S0306457321001503
- https://arxiv.org/abs/1908.06708
- https://dl.acm.org/doi/pdf/10.1145/3450614.3461685
- https://arxiv.org/abs/2006.05255
- https://dl.acm.org/doi/pdf/10.1145/3184558.3186949