Can we trust recommender system fairness evaluation?

The role of fairness and relevance

Theresia Veronika Rampisela, Tuukka Ruotsalo, Maria Maistro, Christina Lioma

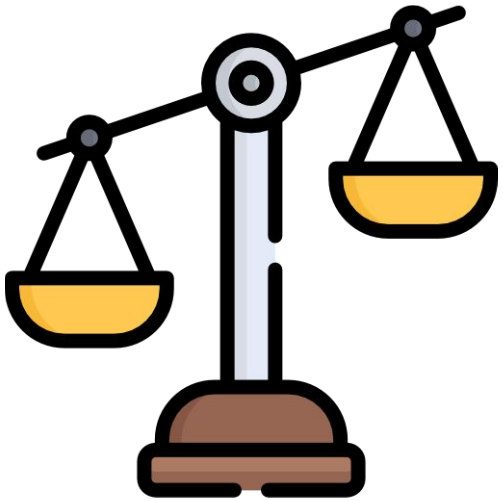
SIGIR 2024
Washington, D.C.

UNIVERSITY OF COPENHAGEN

This work is funded by:
ADD algorithms data & democracy

fairness

How do we **know** if a scale is ‘**broken**’?



fairness

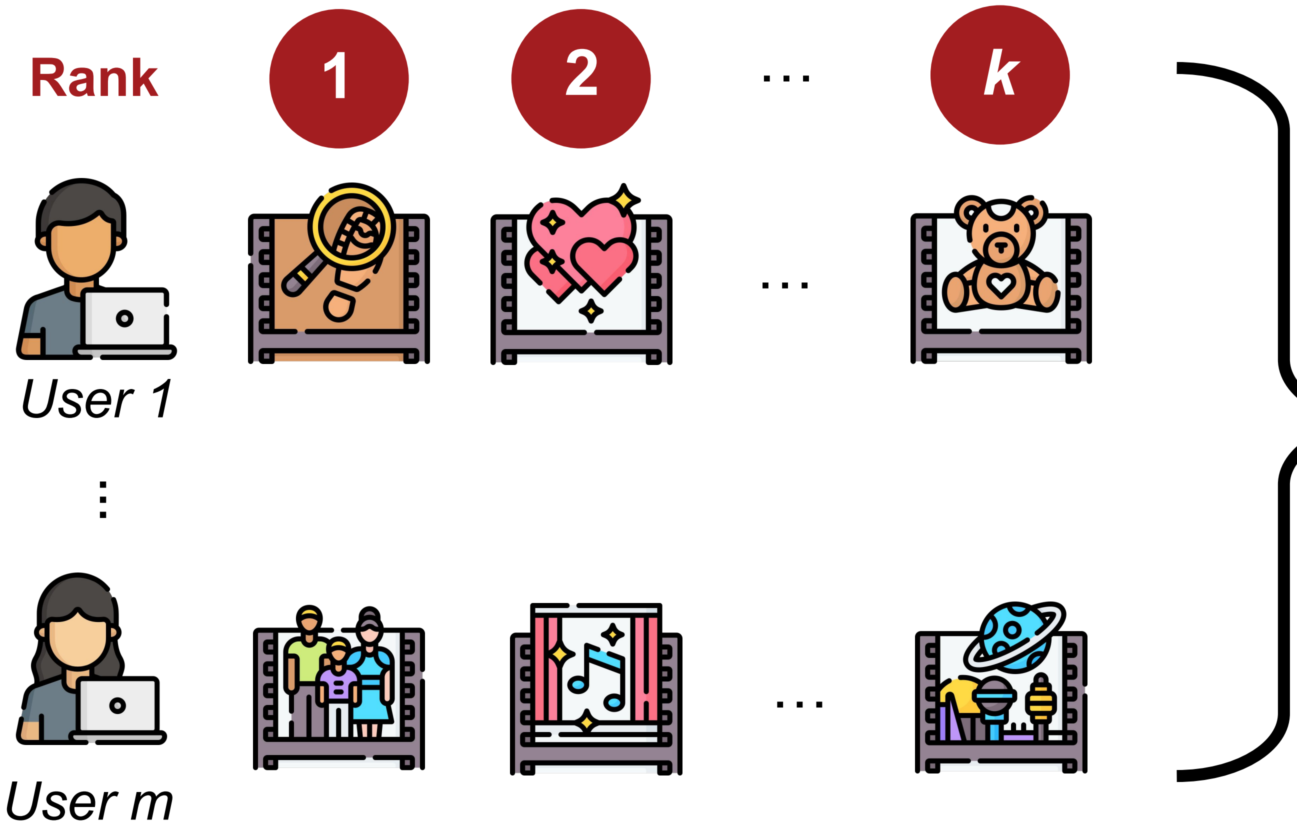
When a scale is broken,

can we trust its measurement? 🤔

Image credit:
Icons created by Freepik - Flaticon

Fairness in recommender systems (RecSys)

Given the top *k* item recommendations across *m* users



Are the recommendations *fair* ?

What is “fair”?

similar individuals receive similar treatments*

Fair towards **whom**?

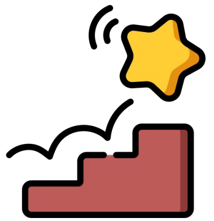
granularity: individual/group

stakeholder: user/item

*Disclaimer: simplified/common definition

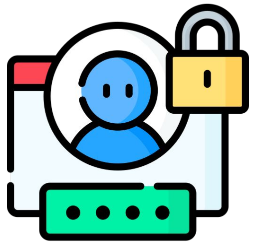
Image credit:
Icons created by Freepik - Flaticon

Why individual item fairness?



Popularity bias causes some items to be recommended more often
→ promoting item fairness may be helpful for new item discovery

Assess distribution **across all individuals** in the population
→ evaluation of individual fairness gives broader view



Sensitive attributes (e.g., gender, age) to identify protected groups often unavailable due to legal/privacy reasons

→ evaluation of individual item fairness does not always require this

Individual item fairness in RecSys

Main terminologies and definitions*

Exposure

Item appearance in
the top k recommendations
(and at which rank position)

Relevance

Whether the user will
find the item relevant
(*interact* with it)

Given recommendations across all users, **individual item fairness** means:

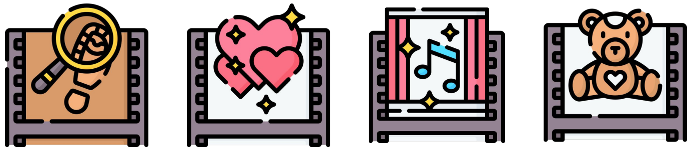
1. all items having equal **exposure** (regardless relevance); or
2. all items receive **exposure** w.r.t. its **relevance** to users

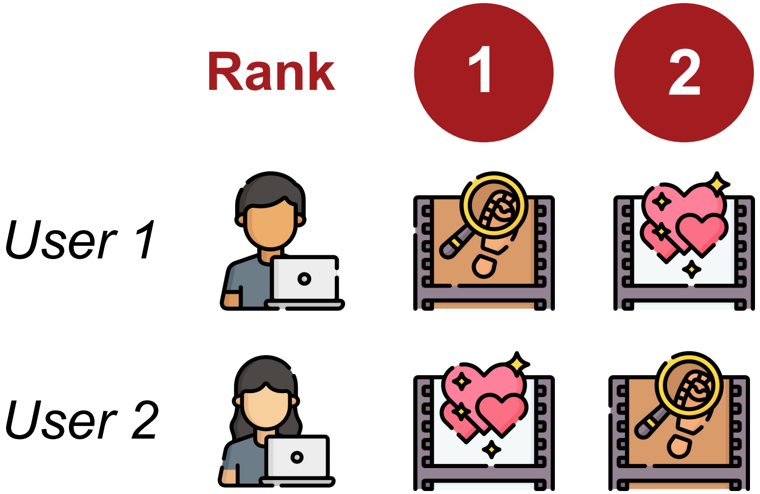
*Disclaimer: simplified/common definition

Image credit:
Icons created by Freepik - Flaticon

Intuitive example: **individual item fairness in RecSys**

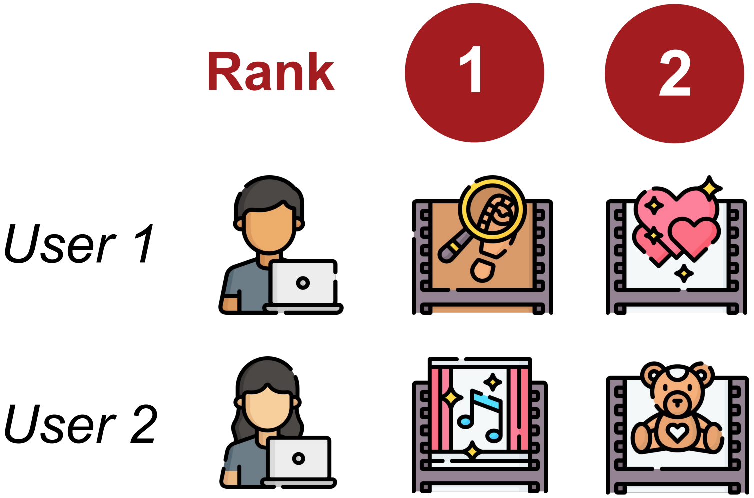
We recommend *k*=2 items from a pool of 4 items to two users

Items in the dataset: 



Only 2/4 unique items are exposed

Case 1



All 4 items are exposed

Case 2

Def.1

"all items have equal **exposure**"

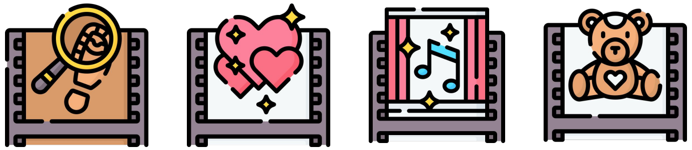
More unique items exposed in Case 2 → Case 2 is fairer

Image credit:
Icons created by Freepik - Flaticon

Intuitive example: **individual item fairness in RecSys**

We recommend *k*=2 items from a pool of 4 items to two users

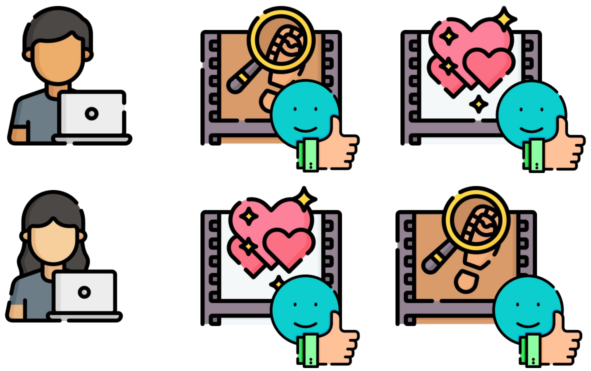
Items in the dataset:



Rank **1** **2**

User 1

User 2



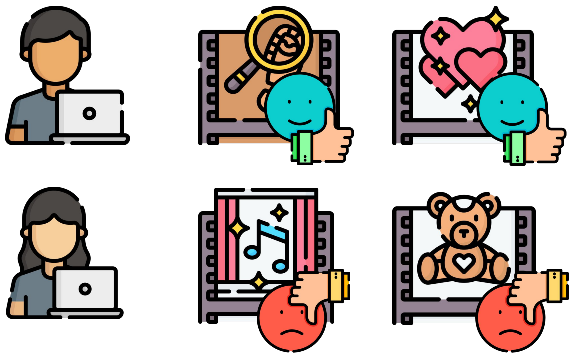
All exposed items are relevant

Case 1

Rank **1** **2**

User 1

User 2



Not all items are relevant

Case 2

Def.1

"all items have equal **exposure**"

More unique items exposed in Case 2 → Case 2 is fairer

Def.2

"**exposure** w.r.t **relevance**"

Items get exposure (more) proportionally to their relevance → Case 1 is fairer

Image credit:
Icons created by Freepik - Flaticon

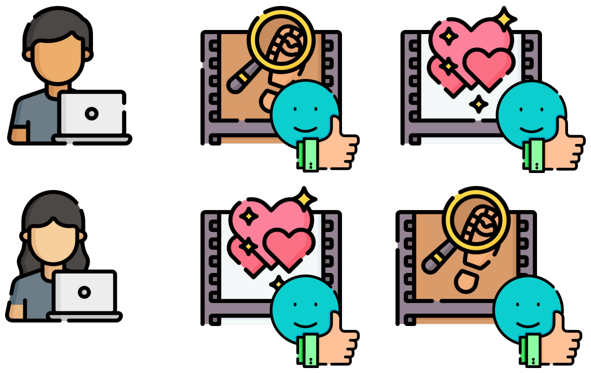
Intuitive example: **individual item fairness in RecSys**

… which case is **fairer** depends on the
fairness definition and the ***evaluation measure***

Rank 1 2

User 1

User 2



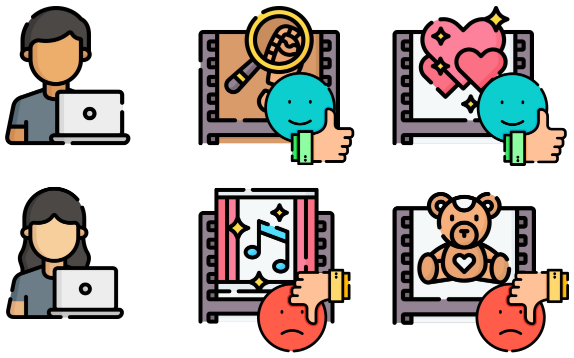
All exposed items are relevant

Case 1

Rank 1 2

User 1

User 2



Not all items are relevant

Case 2

Def.1

“all items have equal **exposure**”

More unique items exposed in Case 2
→ Case 2 is fairer

Def.2

“**exposure** w.r.t **relevance**”

Items get exposure (more) proportionally to their relevance
→ Case 1 is fairer

Image credit:
Icons created by Freepik - Flaticon

Types of individual item fairness measures

Following the two broad individual item fairness definitions:

FAIR measures — measures fairness **only based on exposure**

→ Our previous work investigated the theoretical and empirical limitations of these measures

Evaluation Measures of Individual Item Fairness for Recommender Systems: A Critical Study

THERESIA VERONIKA RAMPISELA, University of Copenhagen, Denmark
MARIA MAISTRO, University of Copenhagen, Denmark
TUUKKA RUOTSALO, University of Copenhagen, Denmark and LUT University, Finland
CHRISTINA LIOMA, University of Copenhagen, Denmark

Accepted to ACM Transactions on Recommender Systems (2023)

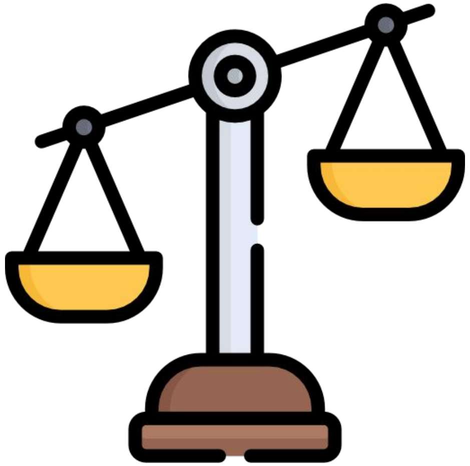
FAIR+REL measures — **'joint'** fairness measures that **consider exposure w.r.t. relevance**

→ This work!

Can the FAIR+REL **(joint)** measures be **trusted**?

RQ1:
… between **FAIR+REL** measures &
- **FAIR** (fairness-only) measures
- **REL** (relevance) measures

RQ2:
… between **FAIR+REL** measures



RQ3:
… across decreasing rank positions

RQ4:
… given increasingly fair and relevant recommendations?

Agreement

Sensitivity

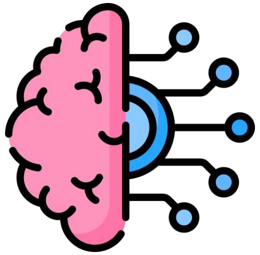
Image credit:
Icons created by Freepik - Flaticon

Experimental setup

4 real-world datasets

Lastfm
ML-10M
Amazon (luxury beauty)
Tenrec (QK-video)

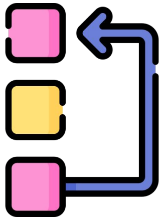
4 recommenders



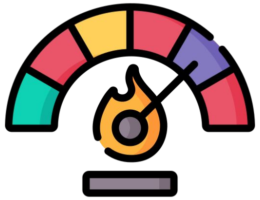
ItemKNN
BPR
MultiVAE
NCL

1 fair reranker

CombMNZ
(based on item coverage and predicted relevance)



20 evaluation measures



6 relevance (REL) measures
5 fairness-only (FAIR) measures
9 joint (FAIR+REL) measures

All evaluated at *k=10* unless otherwise stated

Image credit:
Icons created by Freepik - Flaticon

Evaluation results of all measures

	model	ItemKNN		BPR		MultiVAE		NCL	
	re-ranker	-	CM	-	CM	-	CM	-	CM
REL	↑ HR	0.765	0.581	0.773	0.587	0.778	0.523	**0.793**	0.571
	↑ MRR	0.484	0.270	0.492	0.280	0.476	0.232	**0.503**	0.260
	↑ P	0.172	0.089	0.178	0.092	0.176	0.076	**0.184**	0.087
	↑ MAP	0.137	0.053	0.141	0.058	0.138	0.045	**0.148**	0.050
	↑ R	0.218	0.114	0.224	0.119	0.224	0.098	**0.234**	0.110
	↑ NDCG	0.245	0.119	0.252	0.126	0.247	0.102	**0.261**	0.115
FAIR	↑ Jain	0.042	0.094	0.058	0.140	0.097	**0.222**	0.082	0.215
	↑ QF	0.474	**0.679**	0.362	0.528	0.517	0.678	0.453	0.657
	↑ Ent	0.589	0.735	0.610	0.740	0.707	**0.826**	0.671	0.810
	↑ FSat	0.129	0.216	0.147	0.228	0.202	**0.321**	0.178	0.286
	↓ Gini	0.904	0.790	0.910	0.818	0.839	**0.696**	0.872	0.728
FAIR+REL	↑ IBO	0.209	0.256	0.208	0.253	0.261	0.278	0.242	**0.292**
	↓ IWO	0.791	0.744	0.792	0.747	0.739	0.722	0.758	**0.708**
	↓ IAA	0.004	0.004	0.004	0.004	0.004	0.004	**0.004**	0.004
	↓ $IFD_{\div}$	0.074	0.053	0.075	0.054	0.073	**0.049**	0.076	0.052
	↓ $IFD_{\times}$	0.000	0.000	0.000	0.000	0.000	**0.000**	0.000	0.000
	↓ HD	0.099	0.177	0.104	0.174	0.095	0.203	**0.092**	0.177
	↓ MME	0.001	0.001	0.001	0.001	0.001	0.001	0.001	**0.001**
	↓ II-F	0.001	0.002	0.001	0.002	0.001	0.002	**0.001**	0.002
	↓ AI-F	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.000**

(1) **Extremely small scores for several joint measures** ($\leq 10^{-3}$)

Hard to distinguish across models per dataset!

Evaluation results of all measures

	model	ItemKNN		BPR		MultiVAE		NCL	
	re-ranker	-	CM	-	CM	-	CM	-	CM
REL	↑ HR	0.765	0.581	0.773	0.587	0.778	0.523	**0.793**	0.571
	↑ MRR	0.484	0.270	0.492	0.280	0.476	0.232	**0.503**	0.260
	↑ P	0.172	0.089	0.178	0.092	0.176	0.076	**0.184**	0.087
	↑ MAP	0.137	0.053	0.141	0.058	0.138	0.045	**0.148**	0.050
	↑ R	0.218	0.114	0.224	0.119	0.224	0.098	**0.234**	0.110
	↑ NDCG	0.245	0.119	0.252	0.126	0.247	0.102	**0.261**	0.115
FAIR	↑ Jain	0.042	0.094	0.058	0.140	0.097	**0.222**	0.082	0.215
	↑ QF	0.474	**0.679**	0.362	0.528	0.517	0.678	0.453	0.657
	↑ Ent	0.589	0.735	0.610	0.740	0.707	**0.826**	0.671	0.810
	↑ FSat	0.129	0.216	0.147	0.228	0.202	**0.321**	0.178	0.286
	↓ Gini	0.904	0.790	0.910	0.818	0.839	**0.696**	0.872	0.728
FAIR+REL	↑ IBO	0.209	0.256	0.208	0.253	0.261	0.278	0.242	**0.292**
	↓ IWO	0.791	0.744	0.792	0.747	0.739	0.722	0.758	**0.708**
	↓ IAA	0.004	0.004	0.004	0.004	0.004	0.004	**0.004**	0.004
	↓ IFD÷	0.074	0.053	0.075	0.054	0.073	**0.049**	0.076	0.052
	↓ IFD×	0.000	0.000	0.000	0.000	0.000	**0.000**	0.000	0.000
	↓ HD	0.099	0.177	0.104	0.174	0.095	0.203	**0.092**	0.177
	↓ MME	0.001	0.001	0.001	0.001	0.001	0.001	0.001	**0.001**
	↓ II-F	0.001	0.002	0.001	0.002	0.001	0.002	**0.001**	0.002
	↓ AI-F	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.000**

(1) Extremely small scores for several joint measures ($\leq 10^{-3}$)

(2) Scale mismatch between single-aspect and joint measures

REL scores differ by ~0.16

FAIR scores differ by ~0.14

non-negligible differences!

FAIR+REL scores differ by $\leq 10^{-3}$

the difference seems negligible? 🤔

Explanation for the small scores

Example with IFD:

$$\mathrm{IFD}_{\times}(u) = \frac{1}{n(n-1)} \sum_{i \in I} \sum_{i' \in I \setminus i} \left[\boxed{J_{\times}(u,i)} - J_{\times}(u,i')\right]^2$$

$$J_{\times}(u,i) = \frac{1}{W} \sum_{w=1}^{W} \boxed{r_{u,i,w} \cdot} \boxed{\mathbb{1}_{L_{u,w}}(i) \cdot e_{\mathrm{DCG}}(u,i,w)}$$

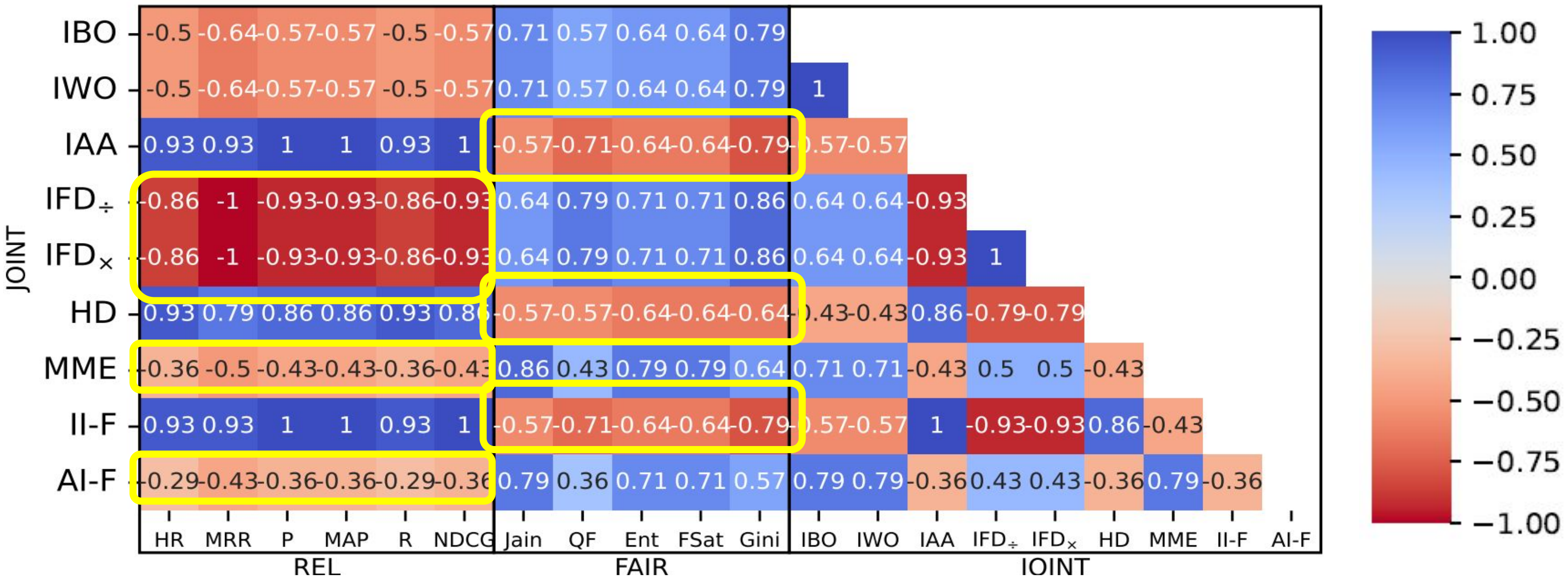
item relevance

item exposure

This term is often 0 due to low number of relevant items per user (in the test set)

RQ1 & RQ2. Agreement between measures

Kendall's Tau correlation between ranking of models, from best to worst, based on different measures



Three groups of similar joint measures:

- IBO/IWO has inconsistent relationships with single-aspect and joint measures (across 4 datasets)
- IAA/HD/II-F do not align with fairness
- IFD/MME/AI-F tend to disagree with relevance

→ no FAIR+REL measures reliably account for both relevance and fairness

Explanation for the grouping of measures

Three groups of measures: (i) IBO/IWO, (ii) IAA/HD/II-F, (iii) IFD/MME/AI-F

Similar formulations

- **IBO/IWO**: fractions of items with an impact score greater/lower than a threshold
- **MME/AI-F** aggregate exposure across users prior to computing the exposure difference (**IAA/HD/II-F** do not)
- **MME/IFD** are pairwise measures.

RQ3. Measure sensitivity at different ranks: setup

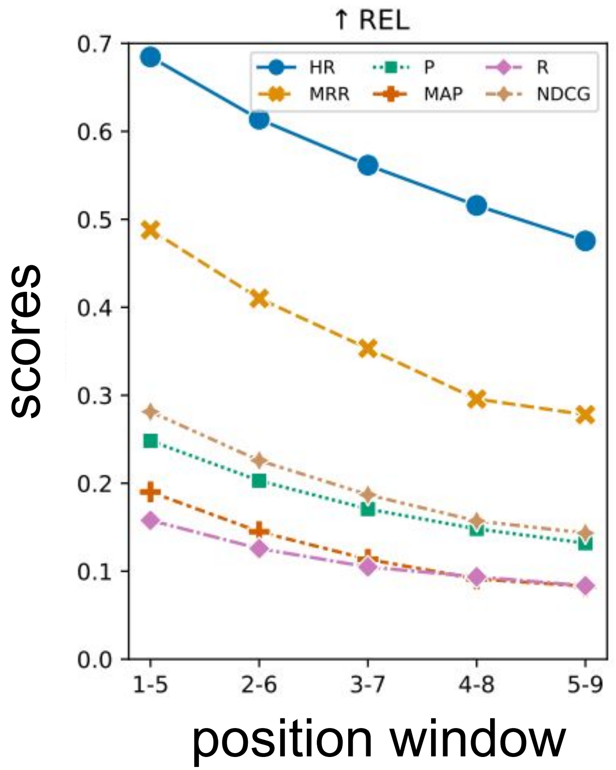
We study **how sensitive the joint measures are** at **decreasing rank positions** compared to relevance- and fairness-only measures

Rank

- Use the runs from the NCL model
- Recommend 5 items from these decreasing rank positions
- Compute all measures at *k=5*

RQ3. Measure sensitivity at different ranks: results

REL measures



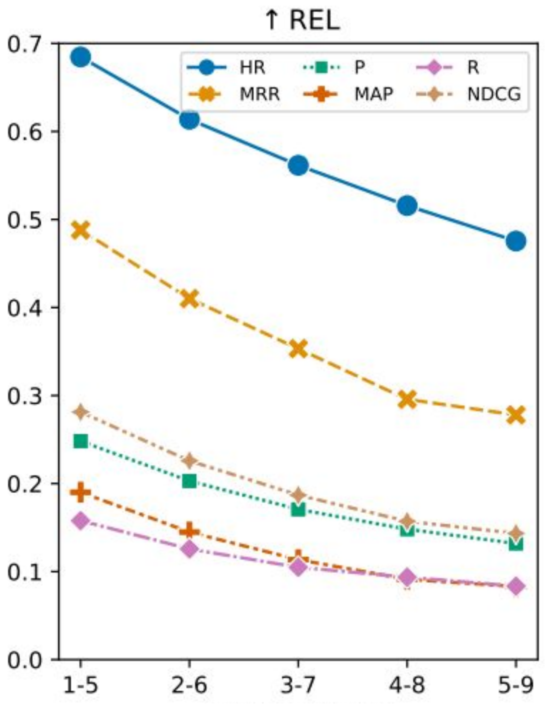
Moving down the rank,
relevance worsens

RQ3. Measure sensitivity at different ranks: results

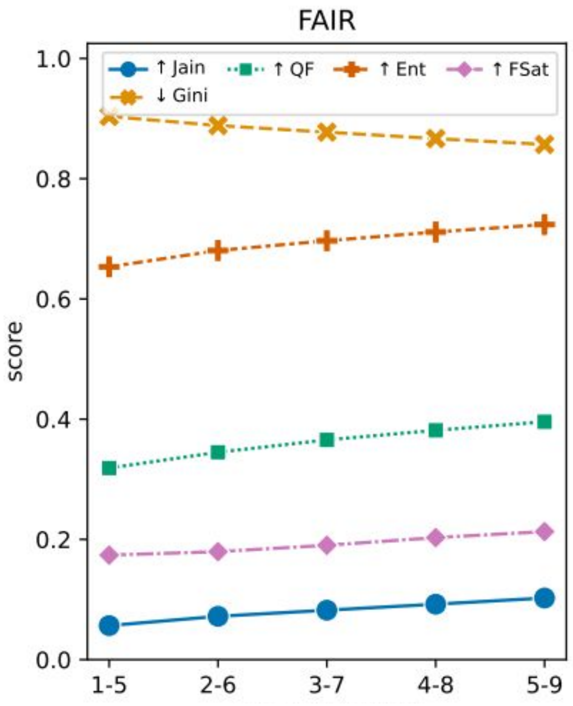
REL measures

FAIR measures

scores



position window



position window

Moving down the rank,
relevance worsens, exposure-based fairness improves

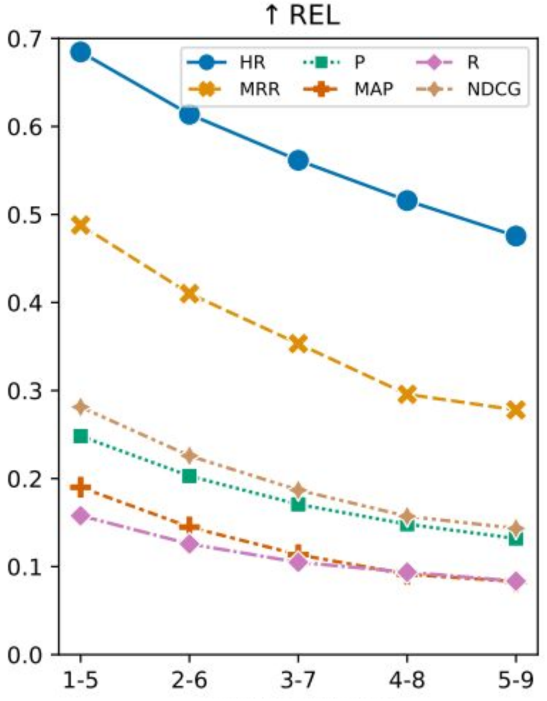
RQ3. Measure sensitivity at different ranks: results

REL measures

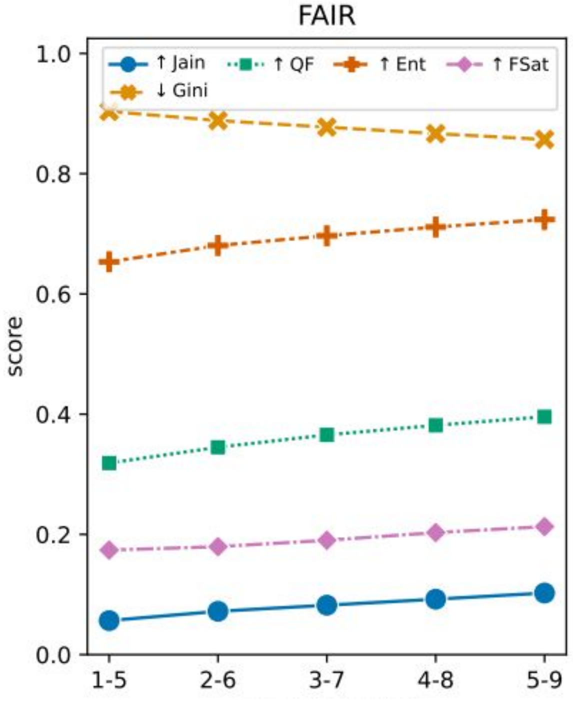
FAIR measures

FAIR+REL (joint) measures

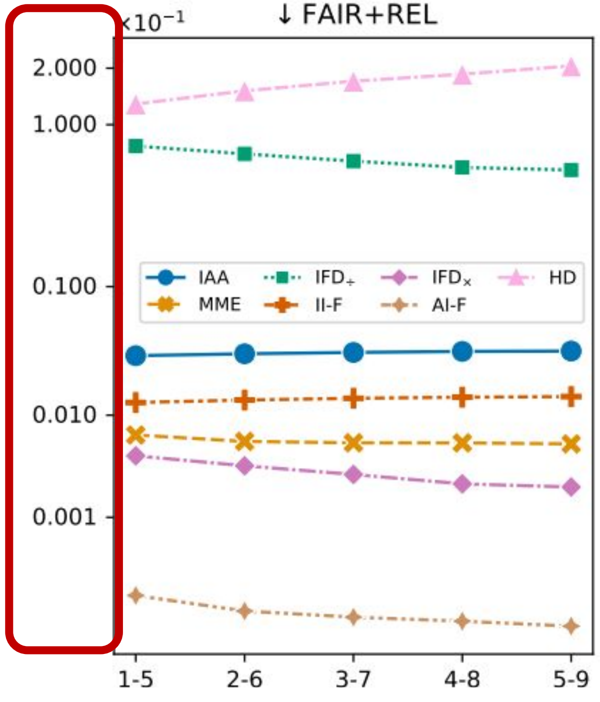
scores



position window



position window



position window

y-axis: scores x10^{-1}

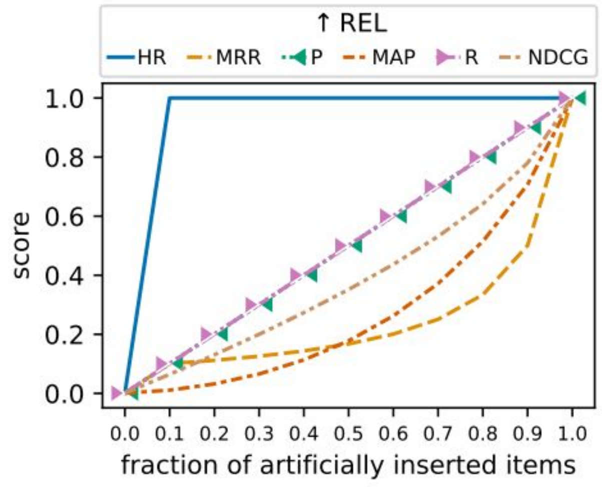
Moving down the rank,
relevance worsens, exposure-based fairness improves
but the **joint measures do not reflect these changes to the same scale**

RQ4. Sensitivity given increasingly fair & relevant recommendations

Idea:

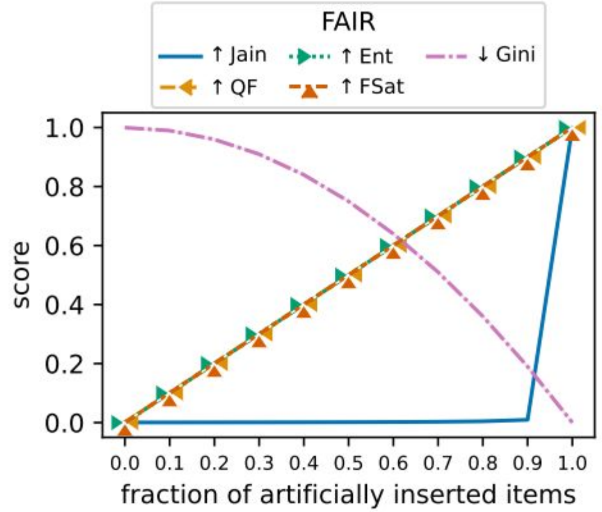
Gradually increase both relevance & fairness:

- increase the proportion of **relevant items**
- distribute **exposure more equally**



Setup:

- Synthetic dataset, artificial recommendation.
- Start by recommending the same *k=10* items that are irrelevant to all users (except for one user where the items are relevant).
- Replace the item at *k* with a less exposed item that is relevant to the user.
  Recompute the measures.
- Repeat the previous step for rank positions *k-1*, …, 1.



RQ4. Sensitivity given increasingly fair & relevant recommendations

Expected result:

FAIR+REL scores

(1) start from the unfairest and reach the fairest

(2) if not, at least, they should become fairer

Actual results:

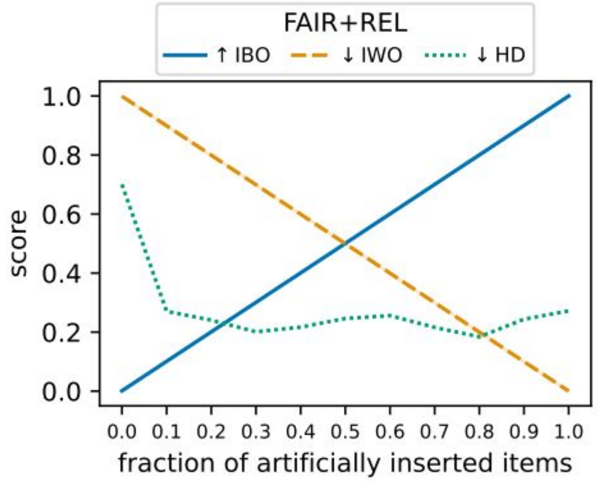
Only IBO and IWO fulfill (1)

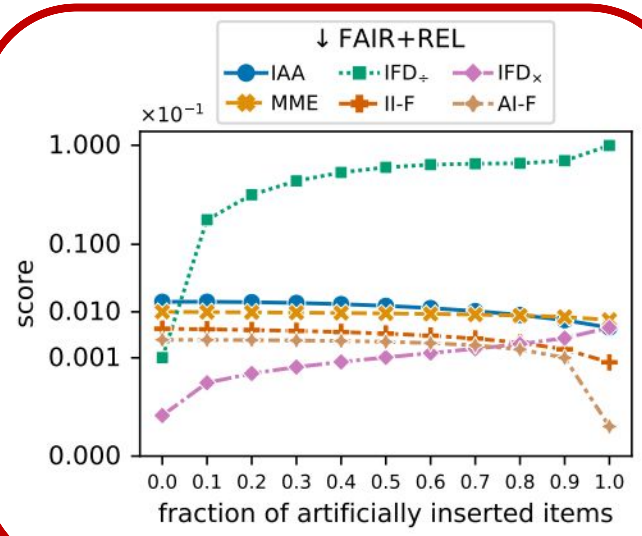
All joint measures slightly improve (except IFD)

Most joint measures are not very sensitive to changes in REL and FAIR scores

the range of these measures: (0, 0.0015)

the range of the single-aspect scores: [0,1]





y-axis: scores x10^{-1}

Explanation

Why did IFD become less fair?

$$\mathrm{IFD}_\times(u) = \frac{1}{n(n-1)} \sum_{i \in I} \sum_{i' \in I \setminus i} \boxed{\left[J_\times(u,i) - J_\times(u,i')\right]}^2$$

IFD: pairwise difference in the combined value of exposure and relevance (J)

When the relevant items start to be moved into the top k:

- the gap between the exposure weight of the relevant items in and outside the top k increases
- thus, unfairness increases

Key Takeaways

1 **Avoid using similar joint measures.**
Three groups: (i) IBO/IWO, (ii) IAA/HD/II-F, (iii) IFD/MME/AI-F
Use only one measure per group to avoid redundancy

2 **Be aware** of the **unintuitive/inconsistent behaviour** and **insensitivity of the joint measures.**

3 **Avoid score misinterpretation in measures with small empirical scales.**
two models differing in scores by 0.001 can be interpreted to perform similarly, yet this difference is due to the nature of the measure empirical range

4 **Measure fairness separately from relevance.**
∵ compressed empirical range, insensitivity, inconsistent alignment to single-aspect measures

Thank you!