



PAPER



CODE

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

Email: thra@di.ku.dk
X: @theresia_v_r

Theresia Veronika Rampisela¹ Maria Maistro¹ Tuukka Ruotsalo^{1, 2} Christina Lioma¹
¹University of Copenhagen, Denmark ²LUT University, Finland



Women in RecSys Journal Paper of the Year 2024 (Junior)

Background

- **Individual item fairness:** ensures that each item receives **similar amount of exposure** (e.g., similar recommendation frequency) → promotes **new item discovery**
- Two types of individual item fairness **evaluation measures:**
 - (i) **exposure-only:** considers only disparity in item exposure, regardless its relevance to users
 - (ii) **relevance-aware:** considers item exposure and relevance
- Many of such measures exist, but it's **unclear in which cases they can or cannot be used, or if their equation is flawed**

This work **analyses the limitations** of the exposure-only measures



All exposure-only individual item fairness measures have limitations!



Legend	Source	9 Fairness Measures								
		Jain	QF	Ent	Gini	Gini-w	FSat	VoCD	II-D	AI-D
●: we fully resolve the limitation										
○: the limitation is unresolvable										
✓: another measure resolves the limitation										
non-realisability: cannot reach max/min score (cause number denoted by C)										
C1. Most unfair score is only given to an impossible scenario	us	●	●	●	●	●	●	●	●	●
C2. Fewer recommendation slots compared to number of items	us	●	●	●	●	●	●	○	○	○
C3. Number of recommendation slots is indivisible by number of items	us	●	●	●	○	○	○	○	○	○
C4. Non-realisability due to unknown formulation of max/min score	us	●	●	○	○	○	○	○	○	○
quantity-insensitivity: ignores frequency of item recommendation	[1]		✓							
undefinedness: cannot be computed (undefined value)	us			●						
always-fair: gives fairest score regardless of recommendation contents	[2]						○			
item-representation-dependence: depends on how items are represented	us							○		

5 limitations; 3 identified by us

[1] Zhu et al. 2020. FARM: A fairness-aware recommendation method for high visibility and low visibility mobile APPs. IEEE Access 8 (2020)
[2] Patro et al. 2020. FairRec: Two-Sided Fairness for Personalized Recommendations in Two-Sided Platforms. WWW'20.

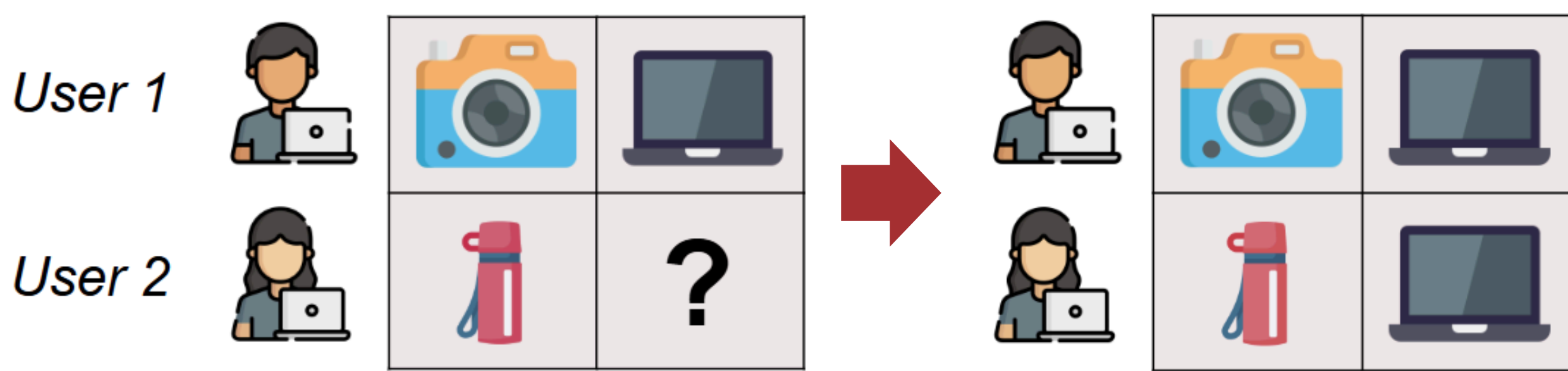
Non-realisability

Limitation: The measure cannot reach its max/min score (C3)

Items in the dataset



How to get the **fairest recommendation**?
→ recommend each item the same amount



Impossible to recommend each item exactly the same frequency!

One possible **fairest recommendation** for this case

Example with Jain's Index (Jain)

- Jain score ranges between **0 (unfairest)** and **1 (fairest)**
- Is Jain = 1 for this **fairest possible recommendation**?
 - No, **Jain = 0.89** for this case
 - **Theoretically impossible** to get Jain > 0.89 here
 - Cannot reach the maximum score of 1!

→ **Illusion that fairness can still be improved**

All 9 exposure-only individual item fairness measures have this limitation!

Solution: Resolve non-realisability with min-max normalisation

Our corrected measure:

$$Jain_{our} = \frac{Jain_{ori} - \frac{k}{n}}{Jain_{max} - \frac{k}{n}}$$

original measure

unfairest possible score: recommend the same k items to all users

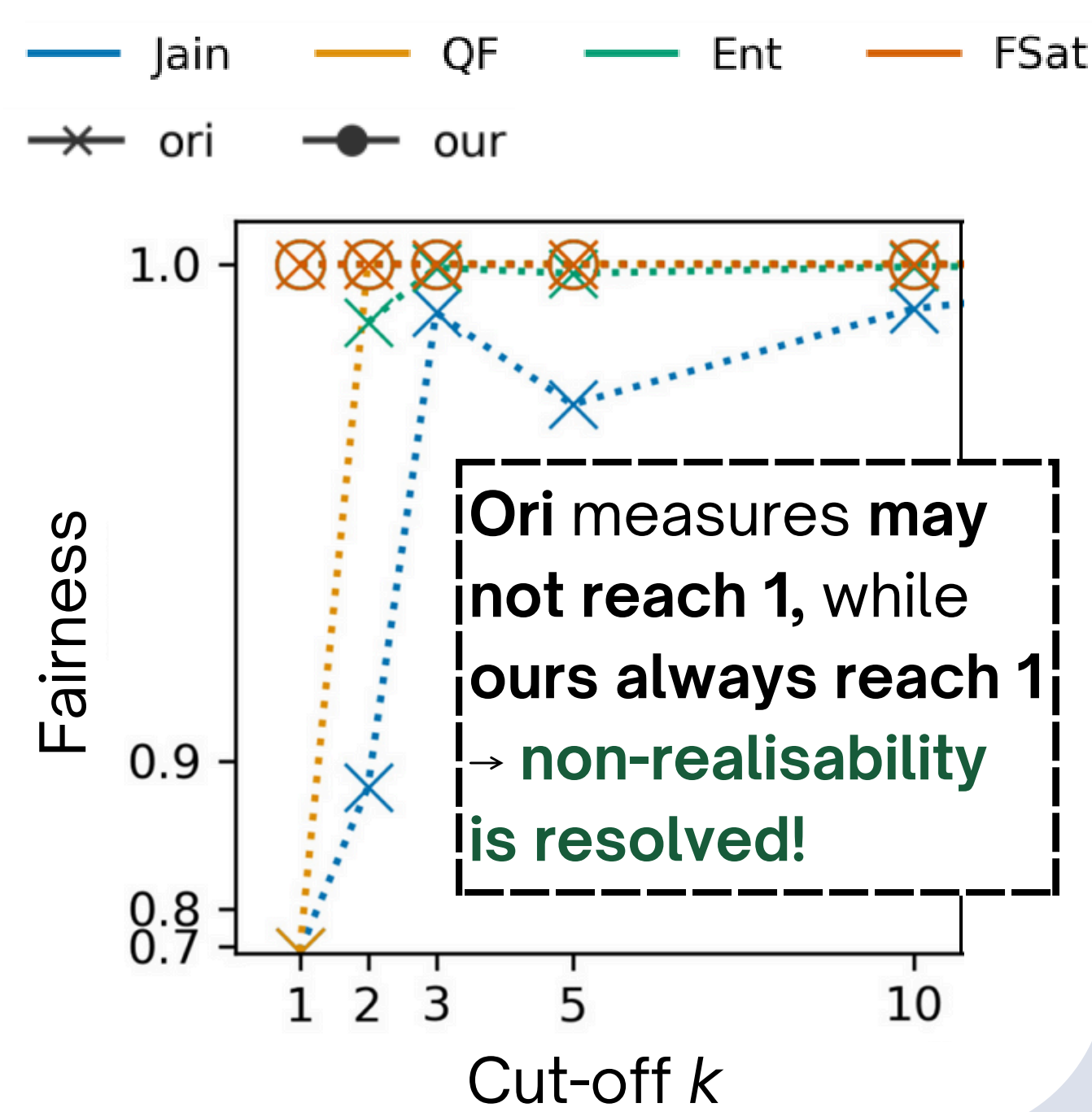
fairest possible score: recommend each item such that they receive exposure as equal as possible

$$Jain_{max} = \frac{(km)^2}{n \left(n \left[\frac{km}{n} \right]^2 + (km \bmod n) \left(2 \left[\frac{km}{n} \right] + 1 \right) \right)}$$

full derivation in the paper!

k: cut-off m: # users n: # items in the dataset

Given the fairest recommendation, can the measures reach 1?



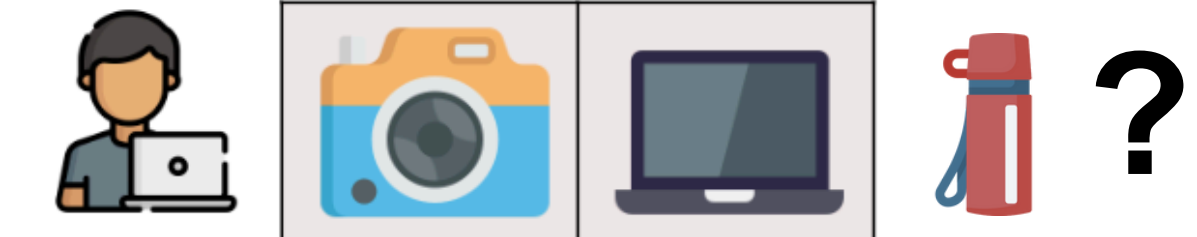
Undefinedness

Limitation: The measure cannot be computed if any item is not recommended

Items in the dataset



Recommendation



$$Ent_{ori} = - \sum_{i \in I} p(i) \log p(i)$$

do for all items

- $p(\text{water bottle}) = 0$ because water bottle is not recommended
- $\log(0)$ is undefined → **Ent can't be computed**

Solution: Resolve undefinedness by redefining the measure

$$Ent_{def} = - \sum_{i \in R} p(i) \log p(i)$$

do only for recommended items

Always-fair

Limitation: The measure gives the fairest score, regardless of the recommendation content

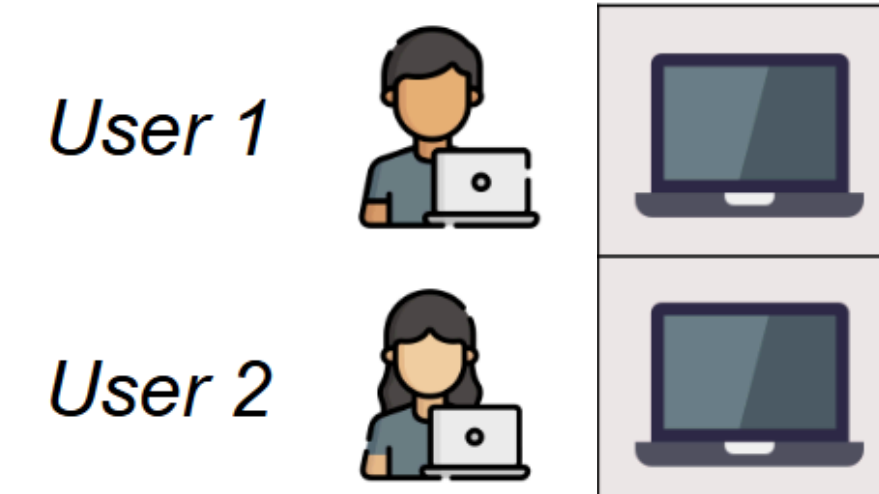
FSat ranges between **0 (unfairest)** and **1 (fairest)**

When there are **fewer slots than items**, FSat is always 1

Items in the dataset

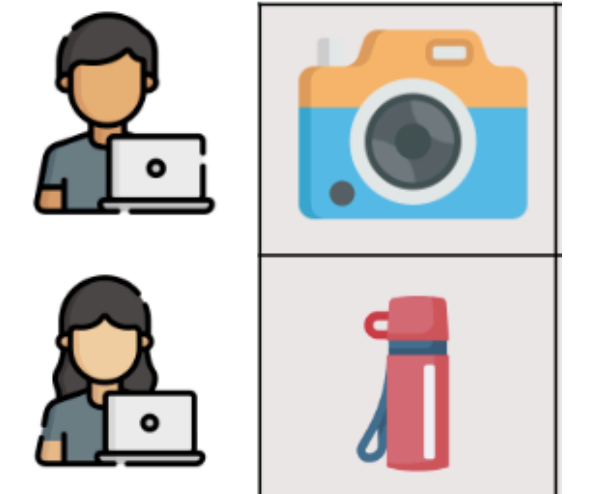


Unfairest Recommendation



FSat = 1, even if this is the **unfairest recommendation** (only 1 unique item exposed)

Fairest recommendation



FSat = 1

FSat cannot distinguish the fairest from unfairest case → **FSat is unusable** in this case!

Solution: use another measure when there are fewer slots than items

Guidelines for Measure Usage

- 1 Use the **original measures** for assessing to what extent a recommendation is fairer than another
- 2 Use **our corrected measures** to evaluate how close a recommendation is to the fairest case



Follow-up Work

Limitations of relevance-aware individual item fairness measures

[SIGIR'24] Can We Trust Recommender System Fairness Evaluation? The Role of Fairness and Relevance

Empirical Limitations

[TORS'25] Relevance-aware Individual Item Fairness Measures for Recommender Systems: Limitations and Usage Guidelines

Theoretical + Empirical Limitations

This work is supported by:



VILLUM FONDEN
VELUX FONDEN



Presented at:

