



PAPER

Stairway to Fairness: Connecting Group and Individual Fairness

Theresia Veronika Rampisela¹ Maria Maistro¹ Tuukka Ruotsalo^{1, 2} Falk Scholer³ Christina Lioma¹¹University of Copenhagen, Denmark ²LUT University, Finland ³RMIT University, Australia

Email: thra@di.ku.dk

X: @theresia_v_r

UNIVERSITY OF
COPENHAGENLUT
UniversityRMIT
UNIVERSITY

CODE



Recommender systems that are **highly fair for groups** can be **very unfair for individuals!**

We study the **relationship between evaluation measures** of **user-side group fairness** and **individual fairness**

Group fairness for users: **equitable outcome** across user groups; the groups can be formed based on one or more **sensitive attributes** (e.g., age)

→ Example: similar effectiveness for users from different age groups

Individual fairness for users: **equitable outcome** for (similar) users

→ Example: similar effectiveness for all users

Background

1

- Fairness is an important aspect of Recommender Systems (RSs), and can be evaluated for **groups** or for **individuals**
- Prior work discusses **conceptually** how RSs can be fair to groups and at the same time unfair to individuals, or vice versa
- However, **no work has empirically studied this**
- Prior work either:
 - Evaluates fairness only for groups or for individuals
 - Evaluates both, but with two **different measures** or for **different subjects** (users/items) or **objectives**

→ **Hard to compare properly!**

To address this gap, we evaluate user-side group and individual fairness with **measures that can quantify both**

Experimental Setup

2

Datasets (3):

- MovieLens-1M (ML-1M), Job Recommendation (JobRec), LFM-1B
- Each dataset has 3 sensitive attributes

LLM-Based Recommenders (4):

GLM-4-9B, Llama-3.1-8B, Ministral-8B, Qwen2.5-7B

Prompt Types (2):

 Both use in-context learning (train+val items as input)

- Non-Sensitive (NS): does not contain user's sensitive attributes
- Sensitive (S): contains user's sensitive attributes

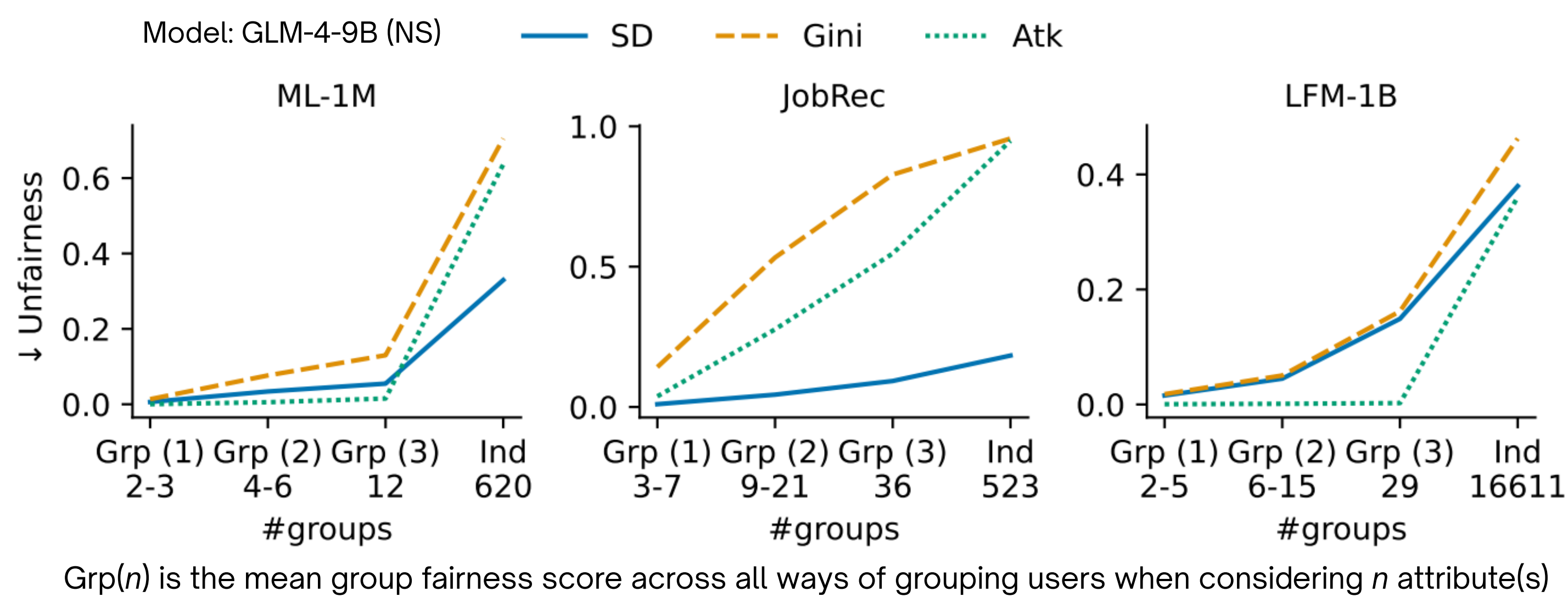
Evaluation: all at $k=10$

- Effectiveness: HR@k, MRR@k, P@k, NDCG@k
- Fairness: 10 group fairness measures + 3 individual fairness measures

Intersectional Fairness

4

Finding #2: Fairness worsens as more attributes are used to form groups → **important to consider intersectionality of user identities!**

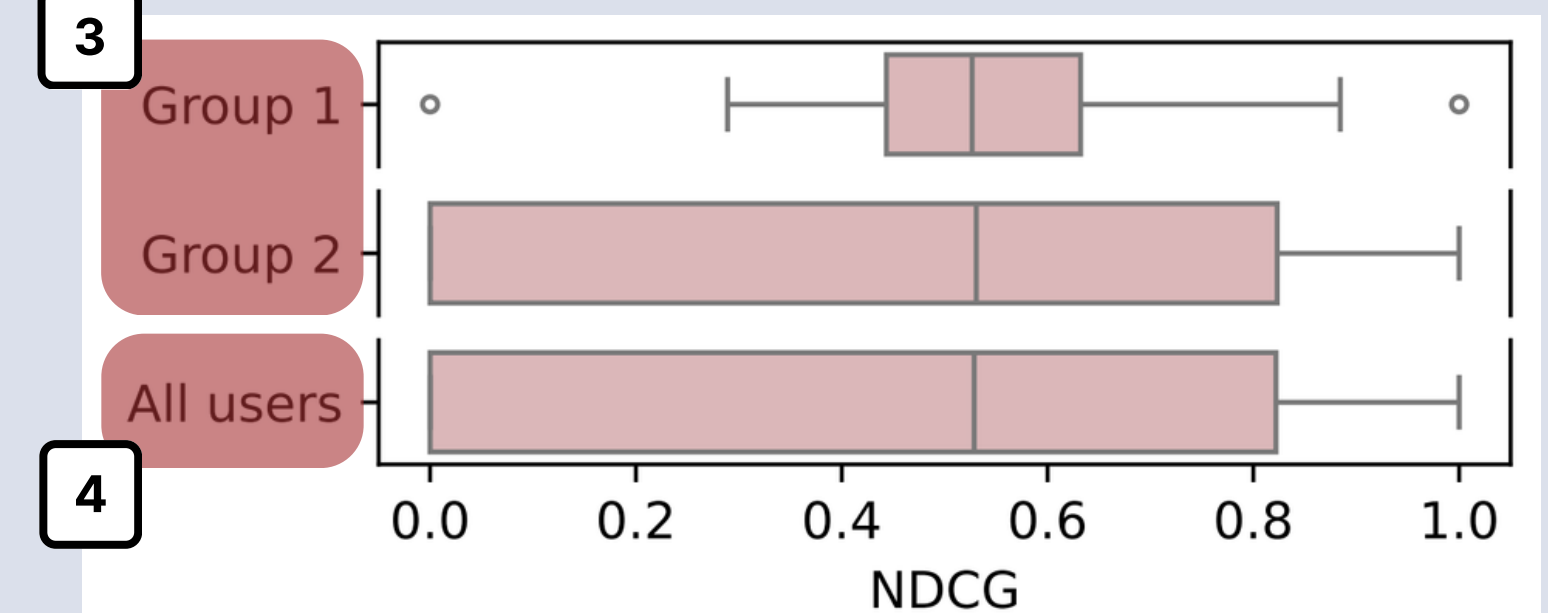


Takeaway: evaluate for individual and within-group fairness alongside group fairness!

Effectiveness (NDCG) and Fairness (Gini) for user groups and individual users.

	#user	↑NDCG	↓Gini
Group 1	11	0.546	0.037
Group 2	901	0.471	0.037
Individual	912	0.472	0.446

Effectiveness (NDCG) score distribution per group and across all users. Circles = outliers.



Fairness is **better between groups** than across all individuals!¹ Both groups have **similar average NDCG**², but **within-group variance is high**,³ which means that **recommendation quality varies widely across users**!⁴