Fairness in Collaborative Filtering Recommender Systems: A Comparative Analysis of Trade-offs Across Model Architectures

[1] EEMCS, Delft University of Technology, The Netherlands

Jeeyoon Kang[1] Supervisor: Masoud Mansoury[1]

1. Introduction

Background: Collaborative Filtering (CF) learns from past user–item interactions to make personalized recommendations. Widely used by platforms like **Netflix**, **Spotify**, and **Amazon**.

Problem: Fairness in Recommender Systems

- **User-side:** Some groups get lower-quality or less relevant recommendations.
- **Item-side:** Long-tail items are under-recommended (popularity bias).
- **Bias amplification:** Demographic and behavioral imbalances are reinforced.
- **Impact:** Lower exposure diversity; marginalized users and content.

Gap: Few studies conduct **standardized, controlled model comparisons**. Fairness interventions add complexity → hard to isolate models' **inherent behavior**.

2. Research Questions

- **RQ1:** How do different CF models perform on **accuracy**, **user fairness**, and **item fairness**?
- **RQ2:** What **trade-offs** arise between accuracy and fairness across model architectures?
- **RQ3:** How do trade-offs vary across datasets with different **sparsity**, **popularity bias**, and **user activity imbalance**?

3. Methodology

We compare **6 CF models** to analyze how their **architectural design** influences **accuracy–fairness trade-offs** on two real-world datasets. **Fairness is assessed at the group level.**

Datasets

- **Both:** Skewed popularity and activity distributions.
- **MovieLens 1M (ML-1M):**
  - Denser dataset; stronger popularity bias.
  - User groups: **Gender**, **Age**, **Activity**
  - Item group: **Popularity**
- **Book-Crossing (BX):**
  - Sparse dataset; stronger activity bias.
  - User groups: **Activity**, **Preference (popularity-based)**
  - Item group: **Popularity**

Models Compared

- **Baselines:** Popularity, Random
- **Matrix Factorization:** BPR
- **Linear:** SLIMElastic
- **Neural:** NeuMF
- **Graph-based:** LightGCN

Evaluation Setup

- **Framework:** RecBole, W&B
- **Split:** 80/10/10 train/val/test, grouped by user
- **Negative Sampling:** Uniform, $n = 1$
- **Ranking Protocol:** Full ranking over all items
- **Interaction Order:** Randomized
- **Tuning Objective:** Max NDCG@10 on validation set

Evaluation Metrics

- **Accuracy:** Recall, Precision, NDCG, MAP, Hit Rate
- **User Fairness:** Group-wise accuracy, dispersion (MAD, STD)
- **Item Fairness:** Coverage (IC), Tail%, Entropy, Gini, Avg. Popularity

Dataset	Users	Items	Interactions	Sparsity	User Skew	Item Skew
ML-1M	6,040	3,706	1,000,209	95.5%	2.74	2.81
BX	6,851	9,085	114,978	99.8%	39.13	6.39

Table 1. Dataset characteristics. Skewness is the third standardized moment (via `scipy.stats.skew`).

4. Key Results

RQ1: Accuracy and Fairness Across Models

Model	Accuracy	Item Fairness	User Fairness
SLIMElastic	★★★★★	★☆☆☆☆	★☆☆☆☆
LightGCN	★★★★☆	★★★★☆	★★☆☆☆
BPR	★★★☆☆	★★★☆☆	★★★☆☆
NeuMF	★★☆☆☆	★★☆☆☆	★★★☆☆
Pop	★☆☆☆☆	☆☆☆☆☆	★★★★☆
Random	☆☆☆☆☆	★★★★★	★★★★★

Table 2. *Overall model performance. More stars = better.*

- **Accuracy:** Rankings are **consistent** across metrics.
- **Item-side fairness:** **Generally stable** across metrics, but **diverges** on BX (Tail%, Avg. popularity).
- **User-side fairness:** Rankings **vary by metric, group, and dataset** — less stable than item-side fairness.

→ **All models** favor head items; some improve tail exposure slightly.
→ **Male users** and **bestseller-preferring users** get higher accuracy.
→ **Active users** are generally better served:
- **LightGCN:** Favors cold users (both datasets).
- **SLIM:** Favors cold users (ML-1M), active users (BX).
- **BPR:** Favors active users (ML-1M), cold users (BX).

RQ2: Trade-offs Between Accuracy and Fairness

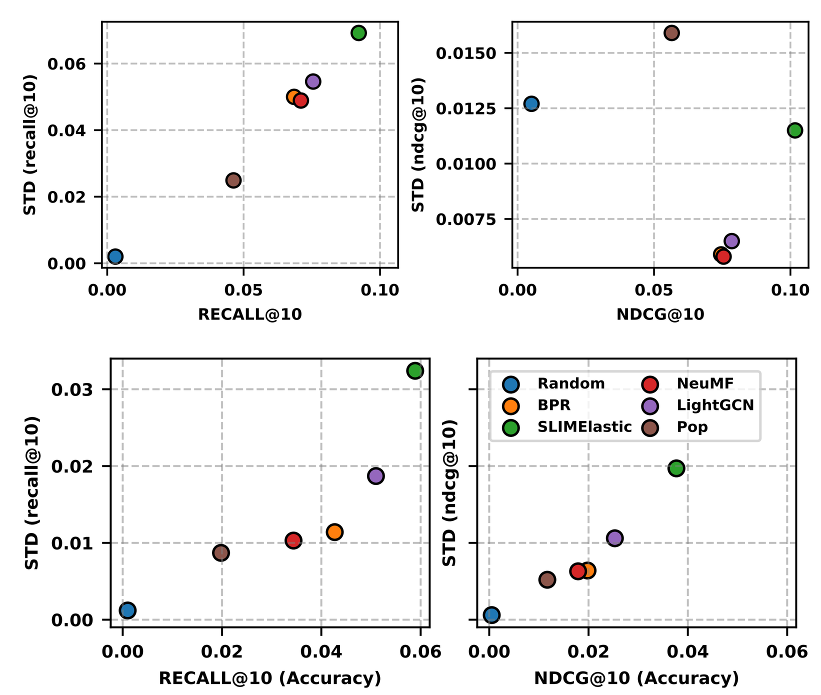

Figure 1. **User-side fairness trade-offs.** **Top:** ML-1M (by activity); **Bottom:** BX (by popularity preference).

Overall trend: Higher accuracy ⇒ greater unfairness (esp. user-side).

Item-side fairness:
- **LightGCN:** Best accuracy–fairness trade-off.
- **SLIMElastic:** Prioritizes accuracy over fairness.

User-side fairness:
- Trade-offs are stronger; no clear winner.
- **SLIMElastic:** Worst fairness despite best accuracy.

→ Fairness–accuracy trade-offs depend on model architecture.

→ Item fairness ⇏ user fairness — they can diverge.

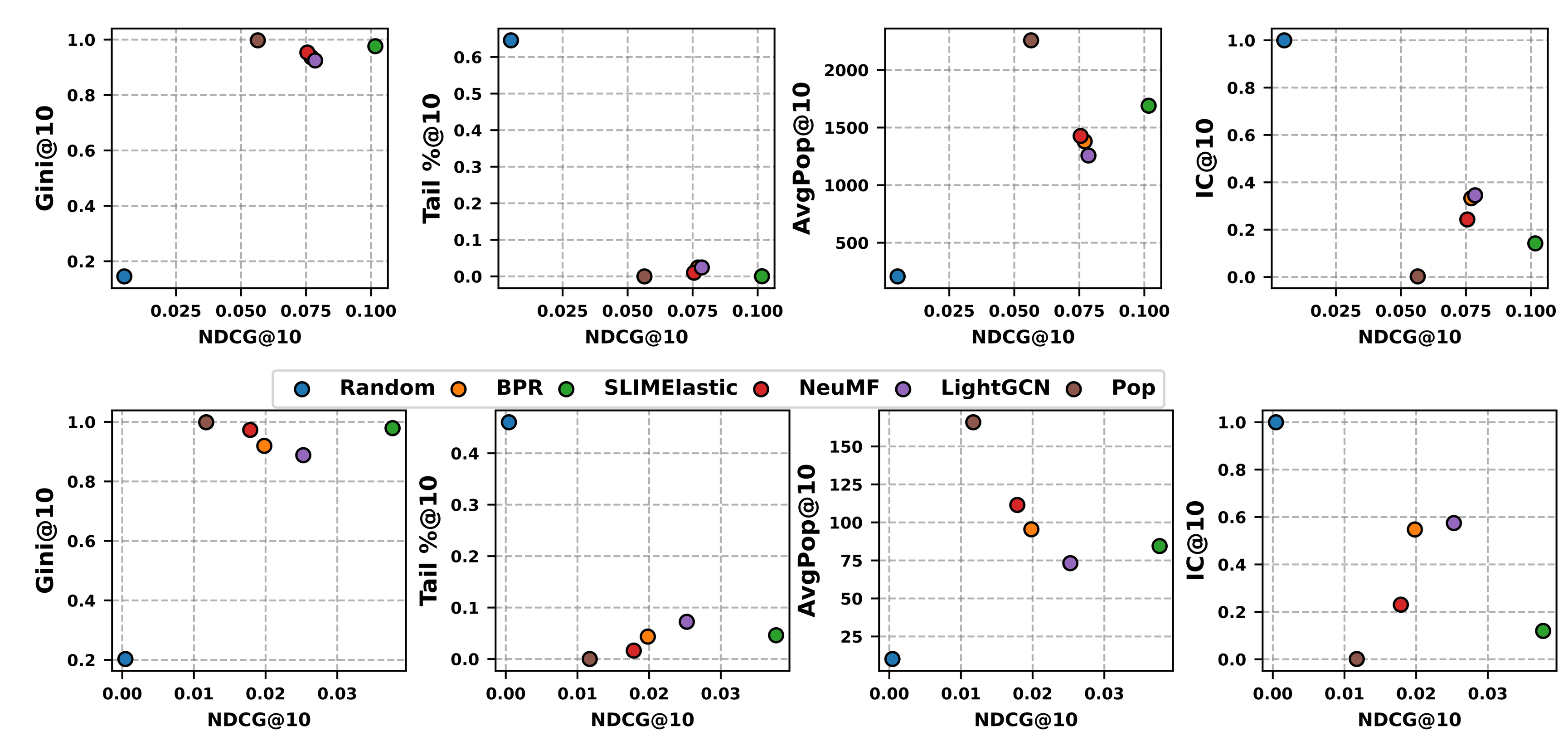

Figure 2. **Item-side fairness Trade-offs. Top:** ML-1M; **Bottom:** BX.

RQ3: Generalization Across Datasets

Model performance varies by dataset:

- **BX (sparse):** Amplifies performance gaps across models.
- **NeuMF:** Lower accuracy and item fairness on BX.
- **SLIMElastic:** Higher Tail% and Avg. Popularity exposure on BX.
- **LightGCN, BPR:** Nearly double item coverage on BX.

User group effects are dataset-specific:

- **ML-1M:** *Activity group* shows highest dispersion; varies by metric.
- **BX:** *Preference group* shows largest, most consistent disparities.
- **Insight: Skewness ⇏ Dispersion**, **Activity ⇏ Low Quality**
  BX has higher user skewness, but preference groups show greater disparities.
  ⇒ Popularity bias may matter more than activity level.

Additional insights:

- **Sparsity:** Reduces accuracy, may improve item fairness (weakening popularity bias).
- **Pairwise models:** More robust to sparsity; fairer to low-activity users.
- **Complexity ⇏ Better:** Simpler models (e.g., SLIMElastic) could outperform deeper ones.
- **Group imbalance:** Distorts fairness — e.g., male-heavy ML-1M (4331 vs. 1709) → models favor **male** users.

5. Key Limitations

- **User fairness hard to generalize:** Varies by group, metric, and overlapping factors (e.g., gender effects may stem from activity imbalance).
- **Suboptimal model tuning:** Some models (e.g., NeuMF) may improve with better hyperparameter search.
- **Limited dataset scale:** Small datasets may not support deep models effectively.
- **No statistical significance testing.**

6. Conclusions and Future Work

Key Conclusions

- **Model architecture** shapes both accuracy and fairness outcomes.
- **LightGCN:** Best trade-off across datasets.
- **Item fairness ⇏ user fairness** — both must be evaluated separately.
- **Complexity ⇏ better results** — simpler models often perform better.
- **Sparsity** strongly affects both accuracy and fairness.
- **Skewness effects** are nuanced — not always predictive of fairness dispersion.

Future Work

- Add **statistical significance testing**.
- Use more **robust user fairness metrics**.
- Explore **individual-level and counterfactual fairness**.
- Extend to **broader model families and domains**.

j.kang-2@student.tudelft.nl