# A Study on Accuracy, Miscalibration, and Popularity Bias in Recommendations

Dominik Kowald, Gregor Mayr, Markus Schedl, Elisabeth Lex

ECIR 2023 - BIAS Workshop
2 - 6 April 2023

# Motivation

- Recommender systems suffer from an **inconsistency in recommendation performance** across different user groups [AMBM19, ETA+18]
- **Two examples:**
  - Varying recommendation accuracy across different user groups → **unfair treatment of users** whose preferences are not in the mainstream of a community [KSL20, KMZ+21]
  - Inconsistencies between input data and recommendations generated → recommendations that are either popularity-biased (**popularity lift**) or not match the users' interests (**miscalibration**)
- **Research objectives:**
  - **O1:** Investigate relationship between popularity lift, miscalibration and accuracy for different **users**
  - **O2:** Inspect recommendation inconsistency for different **genres**

# Defining Recommendation Inconsistency

- Accuracy differences **across user groups** [KSL20]
  - **Mean Absolute Error (MAE)**: rating prediction (lower is better)
  - **Recall and Precision**: top-$n$ recommendation (higher is better)
- **Miscalibration (MC)** [Ste18, LSMB20]
  - Kullback-Leibler (KL) divergence between **genre distributions** in profiles $p(c|u)$ and recommendations $q(c|u)$
  - $KL(p||q) = \sum_{c \in C} p(c|u) \log \frac{p(c|u)}{q(c|u)}$
  - 1 means **miscalibrated** and 0 means **calibrated** recommendations
- **Popularity lift (PL)** [AMBM19]
  - Compare **group average popularity** between profiles $(GAP_p(g))$ and recommendations $(GAP_q(g))$
  - $PL(g) = \frac{GAP_q(g) - GAP_p(g)}{GAP_p(g)}$
  - $PL(g) > 0$ means **too popular** recommendations for $g$ and $PL(g) < 0$ means **too unpopular** recommendations, 0 is perfect

# Datasets

- Three datasets from [KL22] extended with genre information:
  - **Last.fm (LFM)**: LFM-1b [Sch16] dataset provided by JKU Linz
    - In case of Last.fm, we need to map user-generated tags assigned to artists to **genres in the AllMusic database**
  - **MovieLens (ML)**: Movielens 1M dataset provided by GroupLens
  - **MyAnimeList (MAL)**: provided by Kaggle
    - For ML and MAL, the datasets already contain genres
- User groups
  - 1k users with lowest (**LowPop**), with medium (**MedPop**) and with highest (**HighPop**) inclination to popularity (i.e., fraction of popular items in the user profile)
  - Available via Zenodo: https://doi.org/10.5281/zenodo.7428435

| Dataset | $\|U\|$ | $\|I\|$ | $\|R\|$ | $\|C\|$ | $\|R\|/\|U\|$ | $\|R\|/\|I\|$ | Sparsity | $R$-range |
|---|---|---|---|---|---|---|---|---|
| LFM | 3,000 | 131,188 | 1,417,791 | 20 | 473 | 11 | 0.996 | $[1-1,000]$ |
| ML | 3,000 | 3,667 | 675,610 | 18 | 225 | 184 | 0.938 | $[1-5]$ |
| MAL | 3,000 | 9,450 | 649,814 | 44 | 216 | 69 | 0.977 | $[1-10]$ |

# Recommendation Algorithms and Evaluation Protocol

- Python-based open-source framework **Surprise**
- **Rating prediction** → predict listening counts in Last.fm
- **Top-n** → 10 items with highest predicted ratings
- 5 recommendation algorithms:
  - 1 **rating**-prediction approach: UserItemAvg [Hug20]
  - 2 **knn**-based approaches: UserKNN, UserKNNAvg [KSL20]
  - 1 **matrix factorization**-based approach: NMF [LZXZ14]
  - 1 scalable **co-clustering**-based approach: CoClustering [GM05]
- Evaluation protocol
  - Random **80/20** train-test split
  - **Five-fold** cross validation
  - Pairwise **t-test** between LowPop and MedPop / LowPop and HighPop
- Available via Github:
  https://github.com/domkowald/FairRecSys

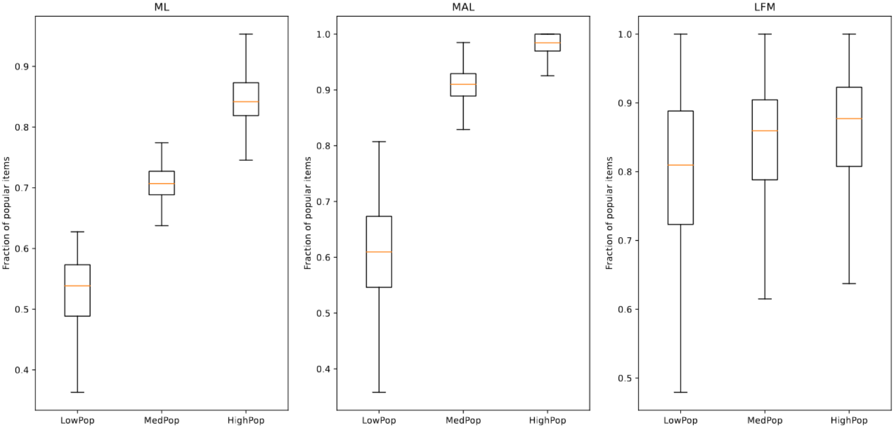
surprise

# O1: MAE, MC, and PL for Different Users

| | Data | LFM | | | ML | | | MAL | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Algorithm** | **Metric** | *MAE* | *MC* | *PL* | *MAE* | *MC* | *PL* | *MAE* | *MC* | *PL* |
| **UserItemAvg** | *LowPop* | **48.02*** | **0.52*** | 1.28 | **0.74*** | **0.78*** | **0.70*** | **0.99*** | **0.95*** | **1.12*** |
| | *MedPop* | 38.48 | 0.48 | **1.61** | 0.71 | 0.71 | 0.42 | 0.96 | 0.73 | 0.42 |
| | *HighPop* | 45.24 | 0.42 | 1.35 | 0.69 | 0.63 | 0.24 | 0.97 | 0.64 | 0.15 |
| **UserKNN** | *LowPop* | **54.32*** | **0.51*** | 0.52 | **0.80*** | **0.75*** | **0.64*** | **1.37*** | **0.92*** | **0.74*** |
| | *MedPop* | 46.76 | 0.50 | **0.82** | 0.75 | 0.69 | 0.37 | 1.34 | 0.72 | 0.22 |
| | *HighPop* | 49.75 | 0.45 | 0.80 | 0.72 | 0.62 | 0.20 | 1.31 | 0.63 | 0.08 |
| **UserKNNAvg** | *LowPop* | **50.12*** | **0.49*** | 0.35 | **0.76*** | **0.78*** | **0.49*** | **1.00*** | **0.90*** | **0.54*** |
| | *MedPop* | 40.30 | 0.47 | 0.61 | 0.73 | 0.70 | 0.33 | 0.95 | 0.73 | 0.24 |
| | *HighPop* | 46.39 | 0.42 | **0.64** | 0.70 | 0.61 | 0.20 | 0.95 | 0.64 | 0.11 |
| **NMF** | *LowPop* | **42.47*** | **0.54*** | 0.10 | **0.75*** | **0.78*** | **0.57*** | **1.01*** | **0.91*** | **0.87*** |
| | *MedPop* | 34.03 | 0.52 | 0.17 | 0.72 | 0.71 | 0.37 | 0.97 | 0.72 | 0.35 |
| | *HighPop* | 41.14 | 0.48 | **0.33** | 0.70 | 0.63 | 0.22 | 0.95 | 0.63 | 0.13 |
| **Co-Clustering** | *LowPop* | **52.60*** | **0.52*** | 0.68 | **0.74*** | **0.77*** | **0.70*** | **1.00*** | **0.90*** | **1.10*** |
| | *MedPop* | 40.83 | 0.51 | **1.04** | 0.71 | 0.70 | 0.43 | 0.96 | 0.72 | 0.42 |
| | *HighPop* | 47.03 | 0.45 | 0.99 | 0.68 | 0.62 | 0.25 | 0.98 | 0.63 | 0.16 |

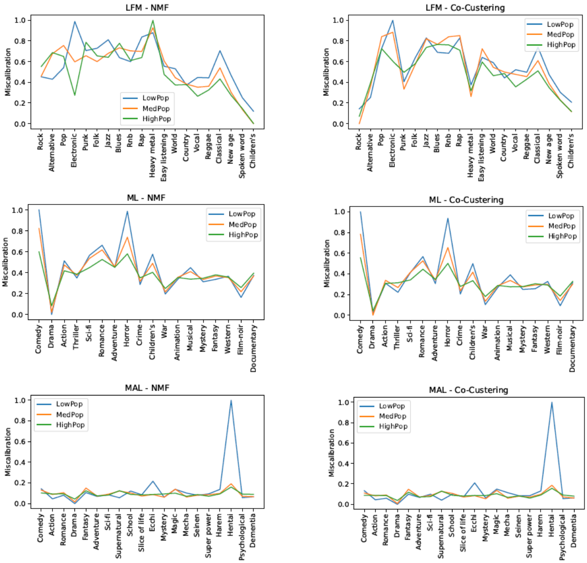
- MAE (Recall/Precision) aligned with MC & PL, except PL for LFM

# O1: Popular Items in the User Profiles Across Groups

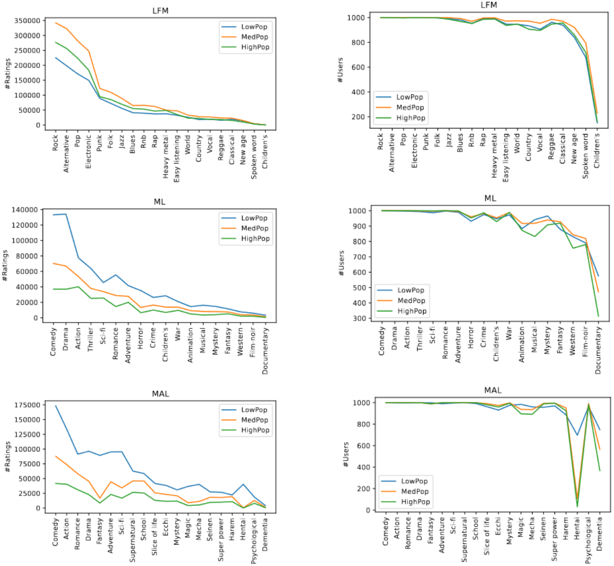



- Repeat consumption patterns in LFM [KSL20, KLS18]

# O2: Recommendation Inconsistency (MC) on Genre Level



# O2: MAL "Hentai" Genre Leads to LowPop Inconsistency



# Conclusion and Future Work

- **O1:** LowPop users get **least accurate, most miscalibrated and most popularity-biased recommendations**
- **O2: Particular genres** contribute to inconsistency in recommendation performance ("Hentai" for LowPop in MAL)
- **We find** a connection between our recommendation inconsistency definitions of accuracy, miscalibration and popularity lift
- **Future Work**
  - Use insights for **popularity bias mitigation** strategies, e.g.,
    - Calibration-based re-ranking for genres that contribute to miscalibration [AMB+21]
    - Personalized re-ranking for users of groups with high popularity lift [ABM19, AK11]
  - Investigate further **popularity bias evaluation metrics** for repeat consumption patterns, e.g., weighted popularity lift
  - Study inconsistency in other **domains** (e.g., e-commerce) using novel **algorithms** (e.g., deep learning)

# Thank you! Questions?

**Contact:**
dkowald [AT] know-center [DOT] at

**Data:**
https://doi.org/10.5281/zenodo.7428435

**Code:**
https://github.com/domkowald/FairRecSys

**Paper:**
https://arxiv.org/pdf/2303.00400.pdf

**Poster/demo on Tuesday → "Uptrendz: API-Centric Real-Time Recommendations in Multi-Domain Settings"**

# References I

Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher, *Managing popularity bias in recommender systems with personalized re-ranking*, The thirty-second international flairs conference, 2019.

Gediminas Adomavicius and YoungOk Kwon, *Improving aggregate recommendation diversity using ranking-based techniques*, IEEE Transactions on Knowledge and Data Engineering **24** (2011), no. 5, 896–911.

Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse, *User-centered evaluation of popularity bias in recommender systems*, Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, 2021, pp. 119–129.

# References II

Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher, *The impact of popularity bias on fairness and calibration in recommendation*, arXiv preprint arXiv:1910.05755 (2019).

Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera, *All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness*, Conference on fairness, accountability and transparency, PMLR, 2018, pp. 172–186.

Thomas George and Srujana Merugu, *A scalable collaborative filtering framework based on co-clustering*, Fifth IEEE International Conference on Data Mining (ICDM'05), IEEE, 2005, pp. 4–pp.

# References III


- Nicolas Hug, *Surprise: A python library for recommender systems*, Journal of Open Source Software **5** (2020), no. 52, 2174.
- Dominik Kowald and Emanuel Lacic, *Popularity bias in collaborative filtering-based multimedia recommender systems*, Advances in Bias and Fairness in Information Retrieval (Cham) (Ludovico Boratto, Stefano Faralli, Mirko Marras, and Giovanni Stilo, eds.), Springer International Publishing, 2022, pp. 1–11.
- Dimitrios Kotzias, Moshe Lichman, and Padhraic Smyth, *Predicting consumption patterns with repeated and novel events*, IEEE Transactions on Knowledge and Data Engineering **31** (2018), no. 2, 371–384.

# References IV


- Dominik Kowald, Peter Muellner, Eva Zangerle, Christine Bauer, Markus Schedl, and Elisabeth Lex, *Support the underground: characteristics of beyond-mainstream music listeners*, EPJ Data Science **10** (2021), no. 1, 14.
- Dominik Kowald, Markus Schedl, and Elisabeth Lex, *The unfairness of popularity bias in music recommendation: A reproducibility study*, European conference on information retrieval, Springer, 2020, pp. 35–42.
- Kun Lin, Nasim Sonboli, Bamshad Mobasher, and Robin Burke, *Calibration in collaborative filtering recommender systems: A user-centered analysis*, Proceedings of the 31st ACM Conference on Hypertext and Social Media (New York, NY, USA), HT '20, Association for Computing Machinery, 2020, p. 197–206.

# References V


Xin Luo, Mengchu Zhou, Yunni Xia, and Qingsheng Zhu, *An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems*, IEEE Transactions on Industrial Informatics **10** (2014), no. 2, 1273–1284.

Markus Schedl, *The lfm-1b dataset for music retrieval and recommendation*, Proceedings of the 2016 ACM on international conference on multimedia retrieval, 2016, pp. 103–110.

Harald Steck, *Calibrated recommendations*, Proceedings of the 12th ACM Conference on Recommender Systems (New York, NY, USA), RecSys '18, Association for Computing Machinery, 2018, p. 154–162.