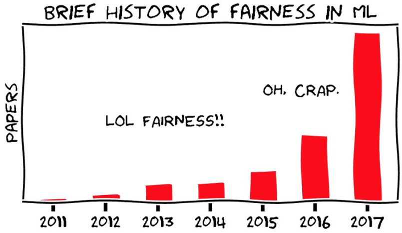
Fairness and Popularity Bias in Recommender Systems

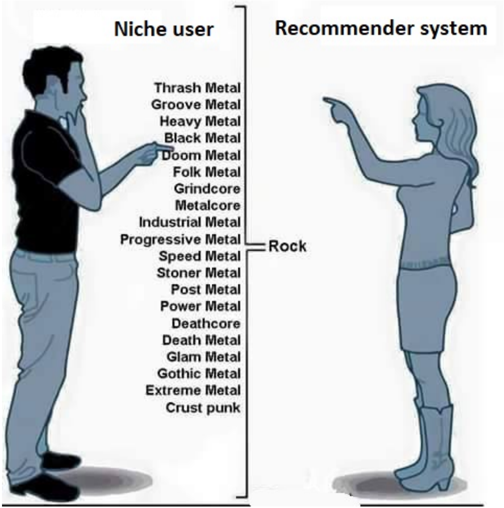
Dominik Kowald, Social Computing, Know-Center Graz

Know-Center and TUG-ISDS Phd retreat



[towardsdatascience.com]

Motivational Example: Music Recommender Systems



[knowyourmeme.com]

Motivation (more formal)

- Popularity bias $\rightarrow$ underrepresentation of unpopular items in recommendation lists
- The group of Prof. Robin Burke [AMBM19] has shown that this also leads to unfair treatment of users with less interest in popular items
- We reproduced this study (small Movie dataset) in a larger setting (large Music dataset) $\rightarrow$ ECIR'2020 reproducibility track [KSL20]
- Investigated research questions
  - **RQ1**: To what extent are recommendation algorithms biased towards popular items?
  - **RQ2**: Is recommendation quality correlated with a user's inclination to popular items?

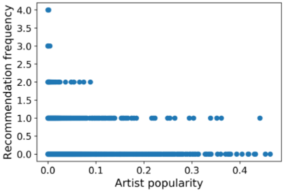
Dataset

- LFM-1b dataset [Sch16]
  - 120k users, 3.1M artists, 1.1B listening events
  - Metadata, e.g., mainstreaminess scores, for users [BS19]
- LFM-1b user groups
  - 1k users with lowest (LowMS), with medium (MedMS) and with highest mainstreaminess (HighMS) → M_global_R_APC measure
- Available via Zenodo: https://zenodo.org/record/3475975

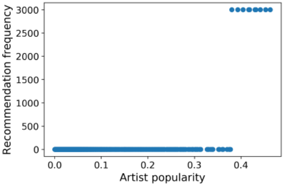
Experimental Setup

- Python-based open-source framework Surprise
- Rating prediction $\rightarrow$ number of listening events of user for artist
- Recommend top-10 artists with highest predicted preferences to user
- Evaluation protocol [AMBM19]
  - Random 80/20 train-test split
  - 3 baselines: Random, MostPopular, UserItemAvg
  - 2 knn-based approaches: UserKNN, UserKNNAvg ($k = 40$)
  - 1 matrix factorization-based approach: NMF ($dim = 15$)
- Available via Github: https://github.com/domkowald/LFM1b-analyses

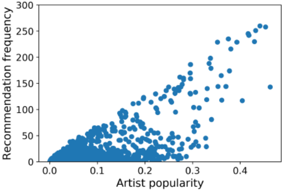
RQ1: Artist Popularity and Recommendation Frequency



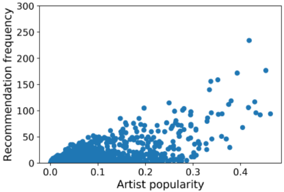
(a) Random.



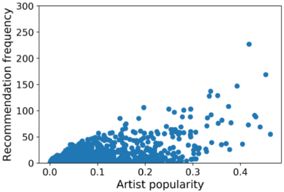
(b) MostPopular.



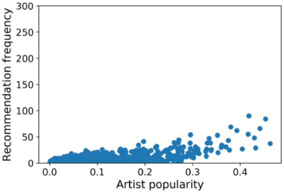
(c) UserItemAvg.



(d) UserKNN.



(e) UserKNNAvg.



(f) NMF.

RQ2: Recommendation Accuracy for User Groups

- Mean Average Error (MAE) metric $\rightarrow$ the lower the better
- LowMS group receives worse recommendations than MedMS and HighMS for all algorithms
- Statistically significant according to t-test with $p < .005$ as indicated by ***
- Best results across user groups by MedMS (in *italic*)

User group	UserItemAvg	UserKNN	UserKNNAvg	NMF
LowMS	42.991***	49.813***	46.631***	38.515***
MedMS	*33.934*	*42.527*	*37.623*	*30.555*
HighMS	40.727	46.036	43.284	37.305
All	38.599	45.678	41.927	34.895

Next steps: Why does accuracy differ?

- Popularity bias
  - If popularity bias is the only reason: HighMS $\rightarrow$ best results, but MedMS $\rightarrow$ best results
- Calibration
  - Are recommendations miscalibrated [LSMB20] for LowMS?
  - If yes, why are they miscalibrated, and how can we ensure calibrated recommendations?
- Diversity
  - Diversity correlated with accuracy?
  - [KMZ+21] $\rightarrow$ across-group diversity ("openness") leads to higher accuracy - CF gets "'distracted" by other users for LowMS?
- Other ideas / interested in collaborations?
  - Please contact dkowald@know-center.at - thank you!

[me.me]

References I

Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher, *The unfairness of popularity bias in recommendation*, Workshop on Recommendation in Multi-stakeholder Environments (RMSE'19), in conjunction with the 13th ACM Conference on Recommender Systems, RecSys'19, 2019.

Christine Bauer and Markus Schedl, *Global and country-specific mainstreaminess measures: Definitions, analysis, and usage for improving personalized music recommendation systems*, PloS one **14** (2019), no. 6, e0217389.

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, 1–26.

References II

- Dominik Kowald, Markus Schedl, and Elisabeth Lex, *The unfairness of popularity bias in music recommendation: A reproducibility study*, 42nd European Conference on IR Research, ECIR 2020, 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, 2020, pp. 197–206.
- Markus Schedl, *The LFM-1B Dataset for Music Retrieval and Recommendation*, Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (New York, NY, USA), ICMR '16, ACM, 2016, pp. 103–110.