Investigating Popularity Bias Amplification in Recommender Systems Employed in the Entertainment Domain

DOMINIK KOWALD, Know Center Research GmbH & Graz University of Technology, Austria

Recommender systems have become an integral part of our daily online experience by analyzing past user behavior to suggest relevant content in entertainment domains such as music, movies, and books. Today, they are among the most widely used applications of AI and machine learning. Consequently, regulations and guidelines for trustworthy AI, such as the European AI Act, which addresses issues like bias and fairness, are highly relevant to the design, development, and evaluation of recommender systems. One particularly important type of bias in this context is popularity bias, which results in the unfair underrepresentation of less popular content in recommendation lists. This work summarizes our research on investigating the amplification of popularity bias in recommender systems within the entertainment sector. Analyzing datasets from three entertainment domains, music, movies, and anime, we demonstrate that an item's recommendation frequency is positively correlated with its popularity. As a result, user groups with little interest in popular content receive less accurate recommendations compared to those who prefer widely popular items. Furthermore, this work contributes to a better understanding of the connection between recommendation accuracy, calibration quality of algorithms, and popularity bias amplification.

Keywords: Recommender Systems, Fairness, Popularity Bias, Bias Amplification, Collaborative Filtering

Reference Format:
Dominik Kowald. 2025. Investigating Popularity Bias Amplification in Recommender Systems Employed in the Entertainment Domain. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 7 pages.

1 Introduction and Motivation

Recommender systems have become one of the most prevalent applications of machine learning and AI, shaping our daily online experiences. They play a crucial role in helping users navigate vast and complex information spaces by identifying relevant content [8, 20]. Since their early implementations [41], these systems have relied on analyzing past user behavior to construct user models and provide recommendations, not only for items like movies, music, and books, but also for social connections in online networks [13, 25, 32].

A variety of techniques have been used to build these user models, including traditional methods such as collaborative filtering (CF) [14], content-based filtering [35], and hybrid approaches [7], as well as more recent techniques based on latent representations (embeddings) and deep learning [10]. The entertainment domain is one of the key areas where recommender systems are widely deployed, assisting users in discovering movies, music, books, and other media. Given the increasing adoption of recommender systems in both research and industry [21], and their inherently human-centric nature, it is essential to consider existing regulations and requirements for trustworthy

Author's Contact Information: Dominik Kowald, Know Center Research GmbH & Graz University of Technology, Graz, Austria, dkowald@know-center.at.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

AI [12]. Various institutions, including the European Commission, have defined trustworthiness through multiple dimensions, leading to regulatory frameworks such as the *EU Artificial Intelligence Act*[11], which emphasizes issues like bias and fairness in AI. These concerns are particularly relevant in the context of recommender systems, as highlighted by recent studies on trustworthy recommendation models[17]. While bias and fairness in AI and machine learning have received significant research attention in recent years [37, 42], the replication and amplification of biases remain open challenges, especially in interactive systems [18] and recommender systems in particular [9]. One of the most prevalent biases in CF-based recommender systems is popularity bias, which results in the underrepresentation of less popular content in personalized recommendations [4, 16].

The present work summarizes our research on investigating the amplification of popularity bias in recommender systems within the entertainment domain. Analyzing datasets from three sectors, music, movies, and animes, we demonstrate that popularity bias disproportionately impacts user groups with little interest in popular content, as they receive less accurate recommendations compared to those who favor popular items [26, 28–30]. Furthermore, this work provides deeper insights into the relationship between recommendation accuracy, algorithmic calibration quality, and the amplification of popularity bias. Finally, we propose directions for future research.

2 Background and Methods

Research has demonstrated that recommendation algorithms, particularly those based on collaborative filtering (CF), exhibit a strong bias toward popular items, resulting in their overrepresentation in recommendation lists [15, 16]. Conversely, this leads to the unfair underrepresentation of less popular, long-tail items [6, 40]. Various metrics have been proposed in the literature to assess and analyze popularity bias from both item and user perspectives [5, 23]. In the following, we describe some of these metrics used in our research, and give an overview of datasets for recommender systems employed in the entertainment domain.

2.1 Metrics to Investigate Popularity Bias Amplification

Our research focuses on three key methods for measuring popularity bias across user groups: (i) differences in recommendation accuracy, (ii) *miscalibration*, and (iii) *popularity lift*. The first approach involves a straightforward comparison of average recommendation accuracy between groups using the mean average error (*MAE*). In contrast, *miscalibration* and *popularity lift* require more complex calculations.

In general, *calibration* measures the alignment between a user profile p and a corresponding recommendation list q in terms of genre distribution [45]. For example, if a user historically consumes 80% rock and 20% pop music, a *calibrated* recommendation list should reflect a similar distribution. While not explicitly a popularity bias metric, calibration is frequently used to assess and interpret popularity bias in recommendations [1, 3]. The concept of *miscalibration* represents the deviation between p and q [34], quantified using the *Kullback-Leibler (KL)* divergence between the genre distributions in p, i.e., $p(c|u)$, and in q, i.e., $q(c|u)$:

$$KL(p||q) = \sum_{c \in C} p(c|u) \log \frac{p(c|u)}{q(c|u)} \tag{1}$$

where C is the set of all genres in the dataset. A value of $KL(p||q) = 0$ indicates perfect *calibration*, while higher values (approaching 1) signify increasingly *miscalibrated* recommendations. These values can be averaged for a given user group g. We term miscalibration as *MC* in this paper.

In contrast, *popularity lift* quantifies the extent to which recommendation algorithms amplify the popularity bias inherent in user profiles [2, 3]. Specifically, it measures the disproportionate recommendation of popular items to a given user group g. The metric is based on the group average popularity $GAP_p(g)$, which represents the mean popularity of items in the user profiles p of group g. Similarly, $GAP_q(g)$ denotes the average popularity of recommended items for users in g. Popularity lift is then defined as:

$$PL(g) = \frac{GAP_q(g) - GAP_p(g)}{GAP_p(g)} \tag{2}$$

A value of $PL(g) > 0$ indicates that recommendations for group g are skewed toward more popular items, whereas $PL(g) < 0$ suggests an overrepresentation of less popular content. The ideal scenario is $PL(g) = 0$, where the popularity distribution remains unchanged. We term popularity lift as *PL* in this paper.

2.2 Datasets and Recommender System Algorithms Employed in the Entertainment Domain

In our research, we analyze three datasets from the entertainment section, namely Last.fm representing the music domain, MovieLens representing the movie domain, and MyAnimeList, representing the anime domain. These datasets are described in more detail in our previous work [26, 28], are summarized in Table 1, and are freely available via Zenodo[1]. As described in [26, 28], we split the users in every dataset in three equally-sized groups (1,000 users) based on their inclination towards popularity. We term the groups *LowPop*, *MedPop*, and *HighPop*.

Table 1. Statistics of the datasets, including the number of users ($|U|$), items ($|I|$), ratings ($|R|$), and distinct genres ($|C|$), as well as sparsity, average interactions per user/item, and the rating range (R-range) [26, 28].

Dataset	$\lvert U\rvert$	$\lvert I\rvert$	$\lvert R\rvert$	$\lvert C\rvert$	$\lvert R\rvert/\lvert U\rvert$	$\lvert R\rvert/\lvert I\rvert$	**Sparsity**	R-range
Last.fm	3,000	131,188	1,417,791	20	473	11	0.996	[1–1,000]
MovieLens	3,000	3,667	675,610	18	225	184	0.938	[1–5]
MyAnimeList	3,000	9,450	649,814	44	216	69	0.977	[1–10]

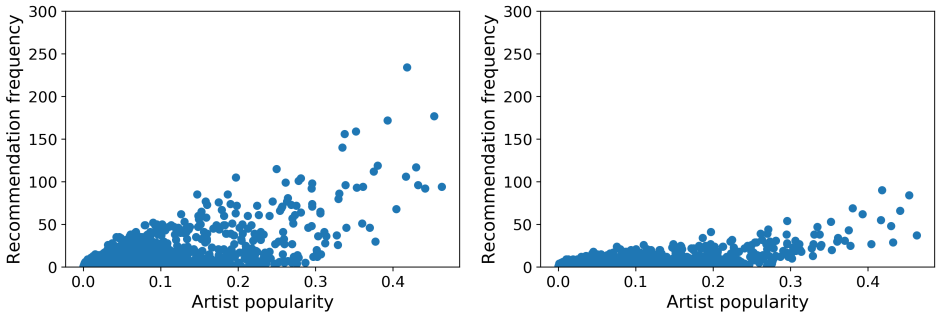
We analyze two well-known personalized recommendation algorithms used in the entertainment section, namely user-based, k-nearest-neighbor CF (UserKNN) [19] and non-negative matrix factorization (NMF) [36]. For the sake of reproducibility [43, 44], the implementation details of these algorithms and the complete source code to reproduce all of our research results are available via GitHub[2].

3 Results and Findings

In Figure 1, we show the correlation between the popularity of music artists and their recommendation frequency in our Last.fm dataset. We see that both algorithms tend to favor popular music artists in their recommendation lists [29, 30]. For the sake of space, we omit the results for MovieLens and MyAnimeList, but similar results can be obtained for these domains as well [26, 28]. This means that the higher the popularity of an item, the higher is also the probability that this item is recommended. Next, we investigate if this item popularity bias also negatively influences user groups with little interest into popularity (i.e., our *LowPop* group).

[1] https://zenodo.org/records/7428435

[2] https://github.com/domkowald/FairRecSys



(a) UserKNN (b) NMF

Fig. 1. Correlation of music artist popularity and recommendation frequency in the Last.fm dataset. Both algorithms investigated tend to favor popular music artists [29, 30]. Similar results can be obtained for the movie and anime domains [26, 28]

Next, Table 2 summarizes our results for the three entertainment datasets across the three user groups and the two algorithms. We see that the *LowPop* user group always receives the statistically significant (according to a t-test with $p < 0.05$) worst accuracy results (*MAE*). In addition, this user group also receives the most miscalibrated (*MC*) and popularity biased (*PL*) results, which helps to better understand the connection between accuracy and these two metrics [26, 28].

Only in the Last.fm dataset, the other user groups reach worse *PL* estimates, which we attribute to the special characteristics of this dataset, namely that the music domain is an entertainment domain with repeatedly consumed items [27, 33, 38]. Thus, in contrast to movies and animes, which are, in most cases, consumed only once or a few items, items in the music domain, such as artists or tracks, are typically consumed repeatedly. The *PL* metric, however, only accounts for unique consumption patterns of items, and therefore this metric provides unintuitive results in the music domain. This suggest, that a weighted variant of *PL* is needed, in which the popularity of an item for a given user is weighted with the number of items the user has consumed this item. We aim to implement and evaluate such an extended *PL* metric in our future research, which we describe in the following section.

Table 2. *MAE*, *MC*, and *PL* results for the *LowPop*, *MedPop*, and *HighPop* user groups. The worst (i.e., highest) results are highlighted in **bold**. Statistical significance (t-test between *LowPop* vs. *MedPop* and *HighPop*) is indicated by * for $p < 0.05$.

	Data	**Last.fm**			**MovieLens**			**MyAnimeList**		
Algorithm	**Metric**	*MAE*	*MC*	*PL*	*MAE*	*MC*	*PL*	*MAE*	*MC*	*PL*
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
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

4 Conclusion, Discussion, and Future Research Directions

In this work, we have summarized our research on investigating the amplification of popularity bias in recommender systems within the entertainment sector. Analyzing datasets from music, movies, and anime, we have shown that recommendation frequency increases with item popularity, disadvantaging users with little interest in popular content. Additionally, we have explored the interplay between recommendation accuracy, algorithmic calibration, and the amplification of popularity bias. We want to further discuss this interplay and its relation to algorithmic fairness in the following paragraph.

Discussion and Relation to Algorithmic Fairness. The metrics of accuracy differences, popularity lift, and miscalibration offer complementary perspectives on how recommender systems treat different user groups, and, together, they reveal important dimensions of algorithmic fairness. Accuracy differences expose disparities in the quality of recommendations received by different user groups. When users with low interest in popular content consistently receive lower accuracy scores, this suggests a systemic bias favoring mainstream preferences. This is not just a performance issue, but a fairness concern, as some users are consistently underserved.

Popularity lift further quantifies this issue by measuring the extent to which recommender systems amplify the popularity bias already present in user profiles. A high popularity lift for certain groups indicates that the system pushes them toward more popular items than they naturally prefer, potentially marginalizing niche tastes and reinforcing majority preferences. Miscalibration offers yet another lens, since it captures the mismatch between the user's historical preferences and the genre distribution in the recommendations. A high miscalibration value implies that the algorithm fails to reflect users' true interests, which often coincides with lower accuracy and higher popularity lift, particularly for users with niche preferences.

Taken together, these three metrics provide a more holistic understanding of fairness. While accuracy differences reveal who is affected, popularity lift and miscalibration help explain how and why certain user groups are disadvantaged. A system that exhibits large disparities across all three dimensions can be considered unfair in its personalization strategy, as it systematically underrepresents the interests of less mainstream users. Thus, improving algorithmic fairness in recommender systems requires not only maximizing overall accuracy but also minimizing popularity lift and miscalibration, especially across diverse user groups. Balancing these metrics is essential for building systems that serve users equitably, beyond the majority preference.

Future Research Directions. Future research should investigate additional domains with respect to popularity bias amplification (e.g., news [31]), and also investigate the usefullness of the different metrics (especially popularity lift) in domains with repeat consumption patterns such as music, as already outlined in Section 3. Furthermore, we plan to work on robust methods for mitigating popularity bias, e.g., based on calibration techniques [22], and evaluating these methods in online user studies (in addition to the offline experiments provided in this work). Finally, we would also like to investigate further aspects of trustworthiness in recommender systems, such as privacy, and study their relation to algorithmic fairness and popularity bias [39].

Acknowledgments

This work was supported by the FFG COMET program. A full version of this article was accepted as a habilitation at Graz University of Technology in June 2024, and is available via Arxiv [24].

References

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

[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The Unfairness of popularity bias in recommendation. In *RMSE Workshop co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019)*.

[3] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2020. The connection between popularity bias, calibration, and fairness in recommendation. In *Fourteenth ACM Conference on Recommender Systems*. 726–731.

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

[5] Abdul Basit Ahanger, Syed Wajid Aalam, Muzafar Rasool Bhat, and Assif Assad. 2022. Popularity bias in recommender systems - a review. In *International Conference on Emerging Technologies in Computer Engineering*. Springer, 431–444.

[6] Erik Brynjolfsson, Yu Jeffrey Hu, and Michael D Smith. 2006. From niches to riches: Anatomy of the long tail. *Sloan Management Review* 47, 4 (2006), 67–71.

[7] Robin Burke. 2002. Hybrid recommender systems: Survey and experiments. *User Modeling and User-adapted Interaction* 12 (2002).

[8] Robin Burke, Alexander Felfernig, and Mehmet H Göker. 2011. Recommender systems: An overview. *AI Magazine* 32, 3 (2011), 13–18.

[9] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.

[10] Xiaocong Chen, Lina Yao, Julian McAuley, Guanglin Zhou, and Xianzhi Wang. 2023. Deep reinforcement learning in recommender systems: A survey and new perspectives. *Knowledge-Based Systems* 264 (2023), 110335.

[11] European Commission. 2021. Proposal for a regulation laying down harmonised rules on Artificial Intelligence (EU AI Act), URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206. Accessed on November 22nd, 2023.

[12] Tommaso Di Noia, Nava Tintarev, Panagiota Fatourou, and Markus Schedl. 2022. Recommender systems under European AI regulations. *Commun. ACM* 65, 4 (2022), 69–73.

[13] Magdalini Eirinaki, Jerry Gao, Iraklis Varlamis, and Konstantinos Tserpes. 2018. Recommender systems for large-scale social networks: A review of challenges and solutions. *Future Generation Computer Systems* 78 (2018), 413–418.

[14] Michael D Ekstrand, John T Riedl, Joseph A Konstan, et al. 2011. Collaborative filtering recommender systems. *Foundations and Trends in Human–Computer Interaction* 4, 2 (2011), 81–173.

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

[16] Mehdi Elahi, Danial Khosh Kholgh, Mohammad Sina Kiarostami, Sorush Saghari, Shiva Parsa Rad, and Marko Tkalčič. 2021. Investigating the impact of recommender systems on user-based and item-based popularity bias. *Information Processing & Management* 58, 5 (2021).

[17] Wenqi Fan, Xiangyu Zhao, Lin Wang, Xiao Chen, Jingtong Gao, Qidong Liu, and Shijie Wang. 2023. Trustworthy recommender systems: Foundations and frontiers. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5796–5797.

[18] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* 14, 3 (1996), 330–347.

[19] Jonathan L Herlocker, Joseph A Konstan, Al Borchers, and John Riedl. 1999. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[20] Dietmar Jannach, Paul Resnick, Alexander Tuzhilin, and Markus Zanker. 2016. Recommender systems — beyond matrix completion. *Commun. ACM* 59, 11 (2016), 94–102.

[21] Dietmar Jannach and Markus Zanker. 2022. Impact and value of recommender systems. *Recommender Systems Handbook* (2022).

[22] Anastasiia Klimashevskaia, Mehdi Elahi, Dietmar Jannach, Christoph Trattner, and Lars Skjærven. 2022. Mitigating popularity bias in recommendation: Potential and limits of calibration approaches. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 82–90.

[23] Anastasiia Klimashevskaia, Dietmar Jannach, Mehdi Elahi, and Christoph Trattner. 2023. A survey on popularity bias in recommender systems. *arXiv preprint arXiv:2308.01118* (2023).

[24] Dominik Kowald. 2024. Transparency, Privacy, and Fairness in Recommender Systems. *arXiv preprint arXiv:2406.11323* (2024).

[25] Dominik Kowald, Sebastian Dennerlein, Dieter Theiler, Simon Walk, and Christoph Trattner. 2013. The social semantic server: A framework to provide services on social semantic network data. In *9th International Conference on Semantic Systems, I-SEMANTICS 2013*. CEUR, 50–54.

[26] Dominik Kowald and Emanuel Lacic. 2022. Popularity bias in collaborative filtering-based multimedia recommender systems. In *Advances in Bias and Fairness in Information Retrieval, BIAS 2022*. Springer, 1–11. https://doi.org/10.1007/978-3-031-09316-6_1

[27] Dominik Kowald, Elisabeth Lex, and Markus Schedl. 2020. Utilizing Human Memory Processes to Model Genre Preferences for Personalized Music Recommendations. In *4th Workshop on Transparency and Explainability in Adaptive Systems through User Modeling Grounded in Psychological Theory*. Association of Computing Machinery. https://doi.org/10.48550/arXiv.2003.10699

[28] Dominik Kowald, Gregor Mayr, Markus Schedl, and Elisabeth Lex. 2023. A study on accuracy, miscalibration, and popularity bias in recommendations. In *Advances in Bias and Fairness in Information Retrieval, BIAS 2023*. Springer, 1–16.

[29] Dominik Kowald, Peter Muellner, Eva Zangerle, Christine Bauer, Markus Schedl, and Elisabeth Lex. 2021. Support the underground: Characteristics of beyond-mainstream music listeners. *EPJ Data Science* 10, 1 (2021), 1–26.

[30] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The unfairness of popularity bias in music recommendation: A reproducibility study. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020*. Springer, 35–42.

[31] Emanuel Lacic, Leon Fadljevic, Franz Weissenboeck, Stefanie Lindstaedt, and Dominik Kowald. 2022. What drives readership? an online study on user interface types and popularity bias mitigation in news article recommendations. In *European Conference on Information Retrieval*. Springer, 172–179. https://doi.org/10.1007/978-3-030-99739-7_20

[32] Emanuel Lacic, Dominik Kowald, Lukas Eberhard, Christoph Trattner, Denis Parra, and Leandro Balby Marinho. 2015. Utilizing online social network and location-based data to recommend products and categories in online marketplaces. In *Mining, Modeling, and Recommending'Things' in Social Media: Revised Selected Papers of MUSE and MSM'2013 Workshops*. Springer, 96–115.

[33] Elisabeth Lex, Dominik Kowald, and Markus Schedl. 2020. Modeling popularity and temporal drift of music genre preferences. *Transactions of the International Society for Music Information Retrieval* 3, 1 (2020). https://doi.org/10.5334/tismir.39

[34] Kun Lin, Nasim Sonboli, Bamshad Mobasher, and Robin Burke. 2020. Calibration in collaborative filtering recommender systems: A user-centered analysis. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media (HT '20)*. 197–206.

[35] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2010. Content-based recommender systems: State of the art and trends. *Recommender Systems Handbook* (2010), 73–105.

[36] Xin Luo, Mengchu Zhou, Yunni Xia, and Qingsheng Zhu. 2014. An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems. *IEEE Transactions on Industrial Informatics* 10, 2 (2014), 1273–1284.

[37] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[38] Marta Moscati, Christian Wallmann, Markus Reiter-Haas, Dominik Kowald, Elisabeth Lex, and Markus Schedl. 2023. Integrating the ACT-R framework with collaborative filtering for explainable sequential music recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. https://doi.org/10.1145/3604915.3608838

[39] Peter Müllner, Elisabeth Lex, Markus Schedl, and Dominik Kowald. 2024. The Impact of Differential Privacy on Recommendation Accuracy and Popularity Bias. In *European Conference on Information Retrieval*. Springer, 466–482.

[40] Yoon-Joo Park and Alexander Tuzhilin. 2008. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM Conference on Recommender Systems*. 11–18.

[41] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. Grouplens: An open architecture for collaborative filtering of NetNews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*. 175–186.

[42] Sebastian Scher, Simone Kopeinik, Andreas Trügler, and Dominik Kowald. 2023. Modelling the long-term fairness dynamics of data-driven targeted help on job seekers. *Scientific Reports* 13, 1 (2023), 1727. https://doi.org/10.1038/s41598-023-28874-9

[43] Harald Semmelrock, Simone Kopeinik, Dieter Theiler, Tony Ross-Hellauer, and Dominik Kowald. 2023. Reproducibility in machine learning-driven research. *arXiv preprint arXiv:2307.10320* (2023).

[44] Harald Semmelrock, Tony Ross-Hellauer, Simone Kopeinik, Dieter Theiler, Armin Haberl, Stefan Thalmann, and Dominik Kowald. 2024. Reproducibility in Machine Learning-based Research: Overview, Barriers and Drivers. *arXiv preprint arXiv:2406.14325* (2024).

[45] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 154–162.