Evaluating The Effects of Calibrated Popularity Bias Mitigation: A Field Study

ANASTASIIA KLIMASHEVSKAIA, MediaFutures, University of Bergen, Norway
MEHDI ELAHI, MediaFutures, University of Bergen, Norway
DIETMAR JANNACH, University of Klagenfurt, Austria
LARS SKJÆRVEN, TV 2, Norway
ASTRID TESSEM, TV 2, Norway
CHRISTOPH TRATTNER, MediaFutures, University of Bergen, Norway

Despite their proven various benefits, Recommender Systems can cause or amplify certain undesired effects. In this paper, we focus on *Popularity Bias*, i.e., the tendency of a recommender system to utilize the effect of recommending popular items to the user. Prior research has studied the negative impact of this type of bias on individuals and society as a whole and proposed various approaches to mitigate this in various domains. However, almost all works adopted offline methodologies to evaluate the effectiveness of the proposed approaches. Unfortunately, such offline simulations can potentially be rather simplified and unable to capture the full picture. To contribute to this line of research and given a particular lack of knowledge about how debiasing approaches work not only offline, but online as well, we present in this paper the results of user study on a national broadcaster movie streaming platform in Norway, i.e., TV 2, following the A/B testing methodology. We deployed an effective mitigation approach for popularity bias, called *Calibrated Popularity (CP)*, and monitored its performance in comparison to the platform's existing collaborative filtering recommendation approach as a baseline over a period of almost four months. The results obtained from a large user base interacting in real-time with the recommendations indicate that the evaluated debiasing approach can be effective in addressing popularity bias while still maintaining the level of user interest and engagement.

CCS Concepts: • **Information systems → Recommender systems**; **Personalization**; • **Computing methodologies → Learning from implicit feedback**; • **Human-centered computing → Field studies**.

Additional Key Words and Phrases: Recommender Systems, Popularity Bias, Long-tail, Online Evaluation, A/B Testing

ACM Reference Format:
Anastasiia Klimashevskaia, Mehdi Elahi, Dietmar Jannach, Lars Skjærven, Astrid Tessem, and Christoph Trattner. 2023. Evaluating The Effects of Calibrated Popularity Bias Mitigation: A Field Study. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3604915.3610637

1 INTRODUCTION AND MOTIVATION

In recent years, the task of finding relevant media content on large streaming platforms to be consumed by online users has become a substantial challenge. This is mainly due to the growing volume and variety of media content produced and shared on these platforms [4]. Hence, users often find it overwhelming to discover exciting content that matches their specific needs and constraints.

Recommender Systems (RSs) are data-driven tools that can address this challenge by offering personalized suggestions tailored to the unique interests and preferences of individual users. However, while these systems offer several benefits

Manuscript submitted to ACM

to users, they can often cause or amplify certain undesired effects [7]. A notable example of such effects is the Popularity Bias, i.e., when a small fraction of popular content items (referred to as "Short Head") continues to gain more popularity, while a larger set of less popular items (referred to as "Long Tail") barely receives any exposure at all. This potentially leads to reduced discovery, diversity and serendipity in recommendation, and it can negatively affect the item producers or the platform itself in decreased exposure rates and sales losses [9, 13, 18].

Even though this phenomenon has been already well-researched, there is a lack of online evaluations through user studies and A/B tests to assess the effectiveness of mitigation strategies in real life scenarios. There appears to be a general lack of evidence of how the offline testing results would correlate with an actual real-life scenario, and how a debiasing method would perform on a real RS platform. In this paper, we address this gap and report the results of an online A/B test, which was conducted on a streaming platform of a national broadcaster in Norway, TV 2 Play [1].

Prior to the A/B test, we first conducted a set of offline experiments to compare several approaches and find the best-performing ones in bias mitigation [15]. We then deployed one of the top-scoring investigated approaches based on Calibrated Popularity [3, 15, 21] on the platform and monitored its performance over a period of months. In the process, we tested out different (cut-off) thresholds for defining the popular/unpopular content which is used by this approach. As a baseline, we considered the recommendation model based on Collaborative Filtering (CF) which is currently implemented and in use on the platform. The results obtained from the A/B study indicate that the evaluated approach can be effective in adjusting recommendation popularity according to user preference without causing a drop in user engagement and interest. This result can be particularly interesting since prior research has reported a trade-off between item relevance and popularity bias mitigation within RSs. On the contrary, our results indicated the possibility of both mitigating the bias and engaging the users.

In summary, this paper offers the following contributions:

- This work is the first one, to our knowledge, that investigates the effects of popularity bias mitigation in a real-life commercial movie recommendation application. We observe the influence of implementing a re-ranking algorithm on the streaming platform by measuring Click-Through Rates (CTRs) for two user groups.
- Additionally, this work is the first one to test generally accepted protocols and parameters for a popularity bias mitigation method and verify their applicability in a realistic scenario.

This paper proceeds as follows: in Section 2 we provide an overview on the previous research on popularity bias mitigation and evaluation; Section 3 we describe the field study design, the setup and different phases of the experiment; afterwards, in Section 3.4 we report the results and statistics collected during the study, with further discussing it in Section 4 and drawing conclusions in Section 5.

2 BACKGROUND

Popularity Bias has been previously researched as "Long Tail problem" of recommendation [19, 23] and the original goal of the research was mostly to improve item coverage to increase sales in e-commerce. Later on, the term *Popularity Bias* was used more frequently, and it became apparent that it has further implications affecting not only the accuracy of recommendations, but also diversity [9], novelty and serendipity [26]. The concept of fairness gets involved in popularity bias research as well, demonstrating how various stakeholders may be affected by popularity bias, which creates unfairness within RS [10]. Many works are especially highlighting how users with different popularity preferences can be affected differently and receive the varying quality of recommendations [2, 6].

[1] https://play.tv2.no/

Prior works in academia and industry have studied this phenomenon and proposed several approaches to mitigate the negative impact on the user and platform side [1, 11, 20, 22, 24, 25]. These methods address popularity bias at different stages of the recommendation process: *pre-processing* approaches (like sampling) treat the bias in the input data; *in-processing* approaches are integrated into the model-building stage and handle the bias within the loss function; *post-processing* approaches deal with the bias by re-ranking the output of the recommender system so that less popular items are included in the front of the recommendation list. Calibrated Popularity (CP) is among the most effective post-processing methods in mitigating bias [3, 15, 21]. This approach can learn user preference towards item popularity and adjust the recommendation popularity accordingly.

Online evaluation in RS is often complicated to set up and thus is less common within the research. To our knowledge, only a few works have described user studies [5, 17, 21, 23] or A/B tests [16] evaluating popularity bias mitigation approaches. While offline testing can give a good understanding of how accurate the produced recommendation can be, only online experiments can give a glimpse of the user perspective and perception of the mitigation effects. Noticing the lack of online evaluation experiments in this field of research, we aim to address this research gap. We believe that this work can provide valuable insights on the effects of popularity bias mitigation and encourage other researchers to consider evaluating biases within RS and mitigation approaches in a real-life settings and scenarios.

3 CALIBRATED POPULARITY BIAS MITIGATION: FIELD STUDY

In this Section we outline the goals and expectations of this field study, providing details of the setup, as well as protocols and procedures followed for various stages of the experiment.

3.1 Study Design and Setup

TV 2 has expressed the interest in diversifying the recommendation they provide to the users to possibly further improve user engagement. Additionally, another potential goal of the platform is to promote a larger part of the item catalogue through recommendation by addressing the problem of popularity bias. The main Key Performance Indicator (KPI) used by TV 2 is the Click-Through Rate (CTR). CTR is aimed to be maximized with the underlying understanding that it represents user interest and engagement. Ultimately, the goal is to test the effects of implementing a post-processing popularity bias mitigation method on a movie streaming platform. Mainly, we aimed to assess the results by measuring CTR as an indication of user engagement. Previous research on popularity bias has demonstrated the trade-off between recommendation quality and popularity bias mitigation [1, 8, 12, 14]. We initiated the study with that consideration in mind, expecting a possibility of a drop in CTR after the implementation of the algorithm, as a result of potentially reduced recommendation quality and user experience. To assess the changes in item catalogue exposure, we also register which items were seen by users in the recommendation row and which items were watched as well from the movie category during the experiments.

We have chosen to implement and experiment with a post-processing re-ranking mitigation method named Calibrated Popularity (CP) that allows a personalized approach to adjusting recommendation popularity level to the popularity preference of each user. The technique adopts a multi-objective task that can be described by the following function:

$$\arg\max(1-\lambda)\cdot Rel(L_u) - \lambda \cdot \mathfrak{J}(P, Q(L_u)) \tag{1}$$

, where L_u is the generated list of the top-K recommendations, $Rel(L_u)$ is the predicted relevancy score for the list L and $\mathfrak{J}(P, Q(L_u))$ is Jensen-Shannon Divergence between two discrete popularity probability distributions P (user history) and Q (generated recommendation list). Following the approach from [3] we use triplets $\{p_1, p_2, p_3\}$ to characterize the

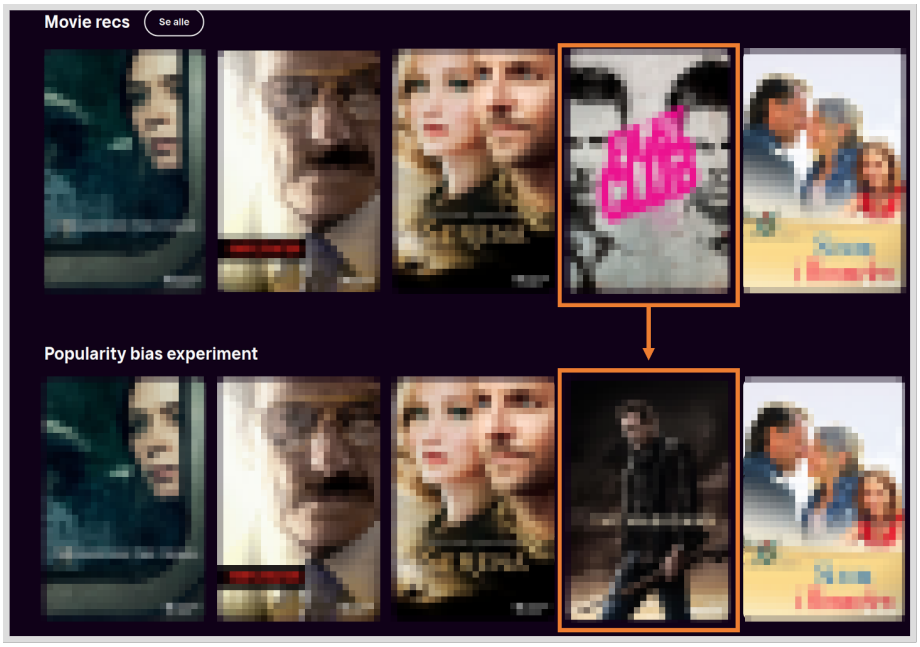

Fig. 1. An example of the re-ranking algorithm exchanging a highly popular item with a less popular alternative for a user less oriented towards popularity (Movie posters are blurred for copyright reasons). If a user has previously expressed an interest in less popular items, judging from their user history, the baseline CF recommendation algorithm would still produce a list with highly popular suggestions. At the same time, CP should be able to detect this user preference and would attempt producing a less popular recommendation, while aiming to still keep the relevancy as high as possible.

proportion of short head, mid-tail and distant tail items within the user history or a recommendation list. For more details on the algorithm we refer the reader to the descriptions [3, 15]. Parameter λ defines the importance of the re-ranking strength over the recommendation relevance and can be adjusted if needed. We set $\lambda = 0.9$ to put more weight on bias mitigation and observe a stronger re-ranking effect for more noticeable changes in recommendations. Previous work on the CP algorithm [3] has explored several λ values, and we select ours according to these insights.

Following the classic A/B testing setup, we split the user base into two groups A and B. We provided the control group A with recommendations made with the current CF-based algorithm used on the platform, while testing group B was exposed to a re-ranked result of that recommendation mechanism. The experiment has been restricted to only movie recommendation, but has been implemented platform-wide for all the active users registered for the service. This means that half of the user base has become a control group, while the other half would be given re-ranked recommendations. See Fig. 1 for a demonstration of how the algorithm alters the recommendation.

In such field studies, various user signals can be tracked during the testing, such as item exposures, clicks, viewing sessions, log-ins, and other analytical data. In this work, we are solely focused on utilizing exposure and click data to calculate CTR values for the whole recommendation row of items, as the main Key Performance Indicator (KPI) on TV 2 Play. An increase in CTR can be a reasonable indicator for at least short-term user interest, engagement, and satisfaction. Potentially it can also imply user's curiosity reaction to the novel, previously unseen items from the long tail of the catalog.

3.2 Preliminary Testing Phase

First, we conducted some short-term preliminary experiments on a smaller set of users to perform sanity checks for the algorithm itself and the parameters used for it. After the checks have been assessed and the required adjustments completed, the main testing phase could be initiated to collect the user data over the agreed period of time.

Adhering to the common approaches in the literature, we tested the algorithm first with a cut-off point $K = 10$ items, i.e., recommending 10 items to every user, and additionally, defined the head items as the top 20% most popular items within the dataset. However, after a short period of pre-testing a few points have become apparent as requiring change and adaptation. Firstly, due to the working principle of the re-ranking algorithm, the items at the end of the top-K list of recommendations have the highest chance of being swapped with less popular alternatives. At the same time the items at the top of this list mostly remain unchanged as a result of the extremely high predicted relevancy. Currently implemented interface layout on desktop and smart-TV versions of TV 2 Play allows the user to see first 6-8 recommended items in personalized row without scrolling, less on mobile (3-5 items). Analyzing user behaviour from item exposure and click data, we quickly realized that not enough users from group B scroll the recommendation row to actually see the new items placed in the list by re-ranking. Thus, we decided to decrease the value of K to 5 and only re-rank the part of the list that is immediately visible to the user in any case.

Secondly, another observation was made regarding setting the threshold for item popularity. Not only it defines the separation of the items into popular and unpopular groups, but also determines which popularity-related category each user ends up in. This way, having more items characterized as highly popular head items results in more users being labeled as mainstream lovers as well. The CP algorithm attempts to adjust the recommendation popularity to the user preference based on this categorization, and often the recommendation (which is generally highly popular) remains unchanged if the user is classified as a mainstream-lover. After quick pre-testing, we realized that within group B the vast majority of users end up being categorized as popularity-oriented and less than 30% of the group actually receives any change in recommendation. Considering this fact and after closely inspecting the popularity distributions of all the items on the platform, we made a decision to claim only top 5% of most popular items as the short head instead of the classic 20%.

3.3 Main Phase

The A/B test was conducted for three consecutive periods: implementing two variations of CP with different popularity thresholds and the "randomized" version of CP. First, from December 19^th 2022 until January 19^th 2023 group B was receiving CP-reranked recommendations with top 5% of most popular items being considered as short head. Afterward, we decided to assess whether this parameter has a lower bound and if setting it too low affects the user experience in a negative way, thus for the period from January 19^{th} 2023 to February 23^{rd} 2023, we have adjusted the popularity threshold to 2.5% instead of 5%. Last but not least, from March 1^{st} 2023 until March 29^{th} 2023, we have implemented a variation that we called *CP-Rand*—a randomized version of the same re-ranking algorithm, which, however, does not consider the relevancy of the items being exchanged in the recommendation in the post-processing step. The tail items to be included in the recommendation are picked randomly and inserted in the list of items suggested to the user in the end. This was performed to analyze the users' perception of the predicted relevancy of the items suggested to them. Both test groups A and B contained more than 10,000 users [2] and each user assigned to a group was supposed to remain in the same group throughout the whole study in all phases.

[2]We were requested by the industry partner to not disclose absolute numbers.

Table 1. Click-Through Rate (Mean) values tracked for each phase of the field study. Highlighted are the CTR values for the better-performing test group. The differences between Group A and B are either statistically significant (in Phase 3) or marginally significant (in Phases 1 and 2).

Phase	Algorithm	CP Condition	Group A CTR	Group B CTR	Change (%)
1	CF vs. CP	5% Top Pop = Head Items	24.87	**25.18**	+1.25
2	CF vs. CP	2,5% Top Pop = Head Items	**18.91**	18.12	-4.18
3	CF vs. CP-Rand	Randomized Re-ranking	22.59	**23.10**	+2.25

Table 2. The number of "unique" item IDs that are overall exposed to the users through recommendations and the number of "unique" items watched by the users of the platform for each phase of the field study.

	Algorithm	Phase 1		Phase 2		Phase 3	
		#Exposed	#Watched	#Exposed	#Watched	#Exposed	#Watched
Group A	CF	3038	2545	2860	2540	2625	2492
Group B	CP	2682	2884	2242	2954	2542	2960

3.4 Results

We report the CTR values registered for each test period in Table 1. In each phase, we collected data from more than 60,000 impressions and user interactions. In the first phase of the testing, with group B receiving CP-reranked recommendations based on the 5% popularity threshold, we observed an increase in CTR of 1.25% compared to the CF baseline in group A. Lowering the threshold to 2.5%, however, led to a decline in CTR for group B—4.18% less compared to group A. Switching to CP-Rand afterward has shown similar results as the first phase, and we again observed an increase in CTR (+2.25%).

In order to check whether the differences between the observed results for the group A and B are statistically significant, we conducted a one-tailed Student's t-test for each phase of the experiment. The outcome of the test showed a marginal significance for the results of Phases 1 and 2 while statistical significance for Phase 3. More particularly, for Phase 1 we observed a p-value of 0.06 (SE=0.0014, t=1.55), and for Phase 2, we observed a p-value of 0.05 (SE=0.0016, t=1.59), respectively. For Phase 3, we observed a p-value of 0.02 (SE= 0.0023, t=2.21).

Additionally, we investigated whether or not more unique items were exposed to the users or watched by them through the recommendation in Group A compared to Group B. We would like to note that, while the users were initially exposed to 6 items in the recommendation row, they could still freely navigate through more recommendations by scrolling through the next items in the list. This can potentially be continued until an item of interest is found (and clicked on to watch) or the end of the recommendation list is reached. Hence, we computed the number of unique item IDs that were exposed to the users through the recommendations and watched by them in different phases of the field study. The results are presented in Table 2. One can observe that group A was exposed to more unique items through the recommendation in all three phases of the experiment. However, at the same time, the consumption of items is higher in group B. This might indicate a higher novelty in the recommendation by CP that required users to scroll less number of items. We discuss these observations further in the next Section.

4 DISCUSSION

Originally, we expected to witness the trade-off between recommendation quality expressed by user engagement and popularity bias mitigation. On the contrary, we instead observed that introducing CP re-ranking to the system seems to at least retain the same relevance, while at the same time exposing the users to more tail items. However, comparing the results of 5%-CP and CP-Rand, it is plausible that the increase in CTR is caused simply by novel items appearing in a recommendation list, regardless of the item relevance. The users might be simply curious about the items they have not seen before, thus having more interest in the whole recommendation list and more likely clicking on it in general. Additionally, this can also lead to a conclusion that predicted relevancy scores do not necessarily correlate with users' perception of a "good" recommendation, thus high relevancy does not necessarily guarantee user satisfaction and engagement.

At the same time, the observed drop in CTR for 2.5%-CP testing phase encourages caution in picking thresholds and parameters for mitigation approaches. Already the pre-testing case of 20% popularity threshold between short head and long tail has demonstrated that commonly accepted practices might not hold up in a real-life scenario. Having the threshold set too high has lead to a relatively weak re-ranking effect, while setting it too low causes a decrease of user engagement, possibly alienating user from the recommendation. One must always perform tests and checks to carefully select correct parameters and thresholds in every particular setting and application to achieve desired effects and not cause any potential harm. We can hypothesize that the most common approach with 20% top items assigned to be the short head is unrealistic due to the common practices of data pre-processing in the literature—the so-called "cold-start" items that have very few or zero interactions are often completely excluded from the datasets for evaluation and training. This can potentially create a false understanding of the power law distribution within the data, which in reality could be even more skewed.

With regards to the observed item exposure vs. item consumption rate between the two experimental groups—due to the design of the user interface on the platform, the user is originally exposed to 6 items in the recommendation row. To see further items the user is required to scroll, this way increasing the number of exposed items. Considering this, we can assume that the users in group A scroll the recommendation list more in search of relevant content. At the same time, the users in group B appear to browse less, but eventually consume more unique items overall, which can be the result of more novel recommendations with re-ranking. This leads us to the conclusion that re-ranking may not necessarily help in overall catalogue exposure, but in exposure of previously unseen, novel and serendipitous items that lead to an increase in consumption. However, these results would require further investigation get a full picture of what stands behind these observations and whether there are any other factors that might be overlooked.

5 CONCLUSIONS AND FUTURE WORK

In this work we presented the results of online popularity debiasing experiments on a streaming platform utilizing a movie recommendation system. We measured CTR values to evaluate the effects of re-ranking calibrated popularity bias mitigation (CP) on user experience. We compared the user engagement under the current CF recommendation approach used by the platform and our re-ranking method in different configurations. Our experiment clearly shows that the well-known trade off between accuracy and debiasing does not necessarily hold up in a real-life scenario, making it possible to promote less popular items while still maintaining recommendation quality demonstrated by overall CTR values.

We would like to acknowledge certain limitations to our experiments. Due to the nature of the real-life scenario testing on a commercial platform, there is no way to control the user demographics and ensure that the same users would keep logging in and using the system for the length of the whole study. This does not significantly affect our goals for this work, but it can potentially influence a more longitudinal analysis of the user behaviour. Such lengthy observations could be invaluable for understanding long-term effects of popularity bias mitigation on user profile and behaviour, which we would be interested in studying in our future work.

ACKNOWLEDGMENTS

This research was supported by industry partners and the Research Council of Norway with funding to MediaFutures: Research Centre for Responsible Media Technology and Innovation, through the Centres for Research-based Innovation scheme, project number 309339. We also would like to thank TV 2 for providing their platform to conduct the A/B testing.

REFERENCES

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the 11th ACM Conference on Recommender Systems*. 42–46.

[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286* (2019).

[3] 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*. 119–129.

[4] Dirk Bollen, Bart P Knijnenburg, Martijn C Willemsen, and Mark Graus. 2010. Understanding choice overload in recommender systems. In *Proceedings of the 4th ACM Conference on Recommender Systems*. 63–70.

[5] Paolo Cremonesi, Franca Garzotto, Roberto Pagano, and Massimo Quadrana. 2014. Recommending Without Short Head. In *Proceedings of the 23rd International Conference on World Wide Web*. 245–246.

[6] 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.

[7] Mehdi Elahi, Dietmar Jannach, Lars Skjærven, Erik Knudsen, Helle Sjøvaag, Kristian Tolonen, Øyvind Holmstad, Igor Pipkin, Eivind Throndsen, Agnes Stenbom, et al. 2022. Towards responsible media recommendation. *AI and Ethics* (2022), 1–12.

[8] Farzad Eskandanian and Bamshad Mobasher. 2020. Using stable matching to optimize the balance between accuracy and diversity in recommendation. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 71–79.

[9] Daniel Fleder and Kartik Hosanagar. 2009. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management science* 55, 5 (2009), 697–712.

[10] Francisco Guíñez, Javier Ruiz, and María Ignacia Sánchez. 2021. Quantification of the Impact of Popularity Bias in Multi-stakeholder and Time-Aware Environments. In *Advances in Bias and Fairness in Information Retrieval: Second International Workshop on Algorithmic Bias in Search and Recommendation, BIAS 2021, Lucca, Italy, April 1, 2021, Proceedings*. Springer, 78–91.

[11] Christian Hansen, Rishabh Mehrotra, Casper Hansen, Brian Brost, Lucas Maystre, and Mounia Lalmas. 2021. Shifting consumption towards diverse content on music streaming platforms. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 238–246.

[12] Elvin Isufi, Matteo Pocchiari, and Alan Hanjalic. 2021. Accuracy-diversity trade-off in recommender systems via graph convolutions. *Information Processing & Management* 58, 2 (2021), 102459.

[13] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction* 25 (2015), 427–491.

[14] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2014. Correcting Popularity Bias by Enhancing Recommendation Neutrality. *RecSys Posters* 10 (2014).

[15] Anastasiia Klimashevskaia, Mehdi Elahi, Dietmar Jannach, Christoph Trattner, and Lars Skjærven. 2022. Mitigating Popularity Bias in Recommendation: Potential and Limits of Calibration Approaches. In *Advances in Bias and Fairness in Information Retrieval: Third International Workshop, BIAS 2022, Stavanger, Norway, April 10, 2022, Revised Selected Papers*. Springer, 82–90.

[16] 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.

[17] Kibeom Lee and Kyogu Lee. 2015. Escaping your comfort zone: A graph-based recommender system for finding novel recommendations among relevant items. *Expert Systems with Applications* 42, 10 (2015), 4851–4858.

[18] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. 2145–2148.

[19] 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.

[20] Sinan Seymen, Himan Abdollahpouri, and Edward C Malthouse. 2021. A Unified Optimization Toolbox for Solving Popularity Bias, Fairness, and Diversity in Recommender Systems.. In *MORS@ RecSys*.

[21] Harald Steck. 2011. Item popularity and recommendation accuracy. In *Proceedings of the 5th ACM Conference on Recommender Systems*. 125–132.

[22] Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. 2021. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1791–1800.

[23] Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. 2012. Challenging the Long Tail Recommendation. *Proc. VLDB Endow.* 5, 9 (2012), 896–907.

[24] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 11–20.

[25] Xiangyu Zhao, Zhendong Niu, and Wei Chen. 2013. Opinion-based collaborative filtering to solve popularity bias in recommender systems. In *Database and Expert Systems Applications: 24th International Conference, DEXA 2013, Prague, Czech Republic, August 26-29, 2013. Proceedings, Part II 24*. Springer, 426–433.

[26] Reza Jafari Ziarani and Reza Ravanmehr. 2021. Serendipity in recommender systems: a systematic literature review. *Journal of Computer Science and Technology* 36 (2021), 375–396.