Bias in book recommendation

VRIJE UNIVERSITEIT

Bias in book recommendation

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy
aan de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. J.J.G. Geurts,
volgens besluit van de decaan
van de Faculteit der Bètawetenschappen
in het openbaar te verdedigen
op maandag 7 september 2026 om 13.45 uur
in de universiteit

door

Savvina Daniil

geboren te Athene, Griekenland

promotoren:	prof.dr. J.R. van Ossenbruggen dr.ir. C.C.S. Liem
copromotor:	prof.dr. L. Hollink
promotiecommissie:	dr. M. Ekstrand prof.dr. A. Smets prof.dr. N. Tintarev dr. S. Pera dr. V. de Boer

This work was funded by the KB National Library of the Netherlands.

SIKS Dissertation Series No. 2026-46
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Keywords: Recommender Systems, Bias, Public Libraries
Printed by: Ridderprint | *www.ridderprint.nl*
Cover by: Dimitra Tsiamouli

The author set this thesis in LaTeX using the Baskervald font.

DOI: http://doi.org/10.5463/thesis.1746
ISBN: 978-94-6537-666-0
An electronic version of this thesis is available at `https://www.vu.nl/`.

Declaration of Authorship

I, SAVVINA DANIIL, hereby declare that the thesis titled *"BIAS IN BOOK RECOMMENDATION"* and its content are the result of my own work.
I confirm that:

- This work was primarily conducted during my pursuit of a research degree at these universities.
- If any portion of this thesis has been previously submitted for an academic degree or any other qualification at these universities or any other educational institution, I have explicitly disclosed this information.
- I have consistently acknowledged the sources of published works by other authors that I consulted.
- In cases where I have included excerpts from the work of others, I have consistently provided proper source attributions. Aside from these quotations, the entirety of this thesis represents my independent effort.
- I have acknowledged all significant sources of assistance.
- In cases where this thesis draws upon collaborative efforts with others, I have provided a clear distinction between the contributions made by collaborators and my own individual contributions.

01/06/2026

We read books to find out who we are.
What other people, real or imaginary, do and think and feel... is an essential guide to our understanding of what we ourselves are and may become.

Ursula K. Le Guin, The Language of the Night

To understand the world at all, sometimes you could only focus on a tiny bit of it, look very hard at what was close to hand and make it stand in for the whole.

Donna Tartt, The Goldfinch

Contents

Abstract 1

1 Introduction 3
1.1 Research Questions &Methods . . . 4
1.2 Contributions . . . 7
1.3 Origins . . . 7

2 A survey on bias in book recommendation 11
2.1 Introduction . . . 12
2.2 Background . . . 13
2.2.1 Recommender systems research . . . 13
2.2.2 Bias in recommendation . . . 13
2.2.3 Recommendation in the library . . . 14
2.3 Surveys in recommender systems . . . 16
2.4 Paper selection and analysis methodology . . . 18
2.5 Overview of the field . . . 21
2.6 Statistical bias in book recommendation . . . 24
2.6.1 Item . . . 25
2.6.2 User . . . 28
2.6.3 Joint . . . 31
2.7 Social bias in book recommendation . . . 31
2.7.1 Item . . . 31
2.7.2 User . . . 35
2.7.3 Joint . . . 37
2.8 Datasets and reproducibility in book recommendation. . . . 37
2.9 Discussion . . . 40
2.9.1 General observations . . . 40
2.9.2 Gaps &future directions . . . 41
2.9.3 Recommendations for interdisciplinarity . . . 41
2.10 Conclusion . . . 42

3 The perspective of public media service practitioners on diversity in recommender systems 43
3.1 Introduction . . . 44
3.2 Related work . . . 45
3.3 Method . . . 47
3.3.1 Candidate Selection . . . 47
3.3.2 Interview structure . . . 48
3.3.3 Coding and analysis. . . . 49
3.3.4 Limitations . . . 49

3.4 Results . . . 50
3.4.1 Goal of recommendation . . . 50
3.4.2 Aspects of diversity . . . 52
3.4.3 Tactics for diversification . . . 55
3.5 Discussion . . . 58
3.5.1 Implementing diversity: a normative process . . . 58
3.5.2 Generalizing to domains beyond public service media . . . 58
3.5.3 Exploring versus defining diversity . . . 59
3.6 Conclusion . . . 59

4 Reproducing bias in recommendation studies 61
4.1 Introduction . . . 62
4.2 Related work . . . 64
4.2.1 Reproducibility in recommender systems . . . 64
4.2.2 Evaluation strategy . . . 65
4.2.3 Popularity bias . . . 65
4.3 Overview of the studies to be reproduced . . . 66
4.3.1 Core process . . . 66
4.3.2 Data . . . 67
4.3.3 Algorithms . . . 68
4.3.4 Division of users in groups . . . 68
4.3.5 Evaluation strategy . . . 70
4.3.6 Other variations . . . 70
4.4 Experimental setup . . . 71
4.4.1 Data . . . 72
4.4.2 Algorithms . . . 73
4.4.3 Division of users in groups . . . 73
4.4.4 Evaluation strategy . . . 73
4.5 Results . . . 74
4.5.1 Overall Popularity Bias Propagation . . . 74
4.5.2 Popularity Bias Propagation per User Group . . . 78
4.6 Discussion and future work . . . 82
4.7 Conclusion . . . 85

***5* The challenges of studying bias in recommender systems 87**
5.1 Introduction . . . 88
5.2 Related Work . . . 90
5.2.1 Bias in Recommender Systems . . . 90
5.2.2 Datasets and Reproducibility . . . 90
5.3 Identifying Data Characteristics and Algorithm Configurations . . . 91
5.3.1 Data Characteristics . . . 91
5.3.2 Algorithm Configurations . . . 95

5.4 Experimental Setup . . . 96
5.5 Results . . . 97
5.5.1 Popularity bias by UserKNN . . . 97
5.5.2 Popularity bias by Matrix Factorization algorithms . . . 101
5.5.3 Popularity bias by Deep Matrix Factorization . . . 103
5.6 Discussion . . . 105
5.6.1 Implications of the present study . . . 105
5.6.2 Recommendations of the present study . . . 105
5.6.3 Limitations of the present study and future work . . . 106
5.7 Conclusion . . . 107

6 Hidden author bias in publicly available datasets 109
6.1 Introduction . . . 110
6.2 Related Work . . . 111
6.2.1 Popularity bias . . . 111
6.2.2 Bias in the book market . . . 111
6.3 Research Design . . . 112
6.3.1 Data . . . 112
6.3.2 Measuring bias . . . 113
6.4 Results &Discussion . . . 114
6.4.1 Bias in the data . . . 114
6.4.2 Bias in recommendation . . . 115
6.4.3 Analysis . . . 117
6.5 Conclusion &Future Work . . . 118

7 A case study on the Danish public libraries 121
7.1 Introduction . . . 122
7.2 Related work . . . 123
7.3 Book recommendation in Danish public libraries . . . 123
7.4 Research design . . . 125
7.4.1 Data Collection and Enrichment . . . 125
7.4.2 Measuring Bias in Booklens . . . 126
7.5 Results . . . 128
7.5.1 Bias in the default Booklens version . . . 128
7.5.2 Effect of system parameters . . . 131
7.6 Discussion . . . 133
7.6.1 Behavioral data and popularity . . . 133
7.6.2 Information on authors . . . 133
7.6.3 Theory of change . . . 134
7.7 Conclusion . . . 134

8 Conclusion 135
8.1 Findings . . . 135
8.2 Reflections &Future directions . . . 139

Bibliography **143**

Appendix **169**

SIKS Dissertation Series **185**

Acknowledgments **199**

Abstract

Books occupy a significant cultural role in human societies, and libraries have long positioned themselves as institutions committed to equitable access to information and promotion of reading. As libraries increasingly adopt AI-driven tools, it becomes important to examine the potential for bias introduced by these systems. This thesis investigates bias in book recommender systems, with an emphasis on the public library sector. We approach this topic from three perspectives: understanding the current ethical considerations around book recommendation, examining the challenges of measuring statistical bias, and studying the societal implications of statistical bias in book recommendation.

In the first part of the thesis, we map the existing landscape. We present the first systematic literature review of bias in book recommender systems, surveying 40 papers from computer science venues. We find that existing research is fragmented, often treats books as interchangeable with other media items, and rarely engages with the unique characteristics of the book domain. We propose future research directions with an emphasis on interdisciplinarity and domain specificity. Additionally, we conduct an interview study in which we explore how practitioners at three public service media organizations in the Netherlands conceptualize diversity in recommender systems. We find that diversity is subject to a wide range of interpretations even within a narrow domain, and that normative choices are unavoidable in operationalizing it.

In the second part, we take a critical look at the measurement of popularity bias. We reproduce three prominent studies on popularity bias in media recommendation across the movie, music, and book domains, and identify four aspects as potential sources of divergence in results: data, algorithms, division of users in groups, and evaluation strategy. We find that all aspects contribute to the divergence, with the evaluation strategy playing a particularly significant role. We further experiment with synthetic and real data to examine the joint effect of data characteristics and algorithm configuration, finding that the presence and magnitude of popularity bias are highly sensitive to these choices. These findings indicate that conclusions on bias should be explicitly scoped to the limits of the experimentation.

In the third part, we study the relationship between statistical bias and social bias. Using the Book-Crossing dataset, we show that popularity bias in collaborative filtering leads to the over-recommendation of books by American authors, whose works dominate the data. We then conduct the first audit-type study of bias in a library recommender system in production: Booklens, the non-personalized item-to-item system used by the Danish public libraries. We find that Booklens is strongly prone to popularity bias, and that this propagates into author nationality bias, with books by less popular nationalities being systematically underrepresented. We show that, while tweaking system parameters can reduce the effect, bias persists across configurations.

The findings of this thesis highlight that bias in book recommendation is an underexplored yet consequential area of research. We argue that addressing it requires specificity in research design and a willingness to closely engage with the domain, and the institutions that use these systems.

1

Introduction

Books hold a symbolic role in human societies. Various roles have been attributed to them: they facilitate connection and understanding among people and peoples [242], enrich and broaden knowledge [242], reflect human complexities [18], and shape cultural narratives [125]. In the 21st century, "slow media" in print and analogue forms are especially romanticized [181, 217], and contrasted with modern visual fast-paced media. In response to the fear of books becoming obsolete, *bookishness* has emerged as a persistent phenomenon [175]. There is some evidence that the pandemic and the popularity of online bookish communities such as BookTok have had a positive influence on reading [46, 190, 241].

Libraries have traditionally been gatekeepers of books and stewards of history [182]. In the information age, their set of duties has expanded. Libraries are no longer just physical repositories but rather dynamic hubs of digital information, interactive learning, and global collaboration [167]. They position themselves as inclusive social meeting spaces that encourage pluriformity and exposure to a variety of perspectives [102]. They increasingly provide digital services, which are often driven by AI. Librarians worldwide are expected to expand their skills accordingly [93, 112]. AI applications have the potential to support libraries in the digital age. In particular, recommender systems, the software tools and techniques that provide suggestions for items [187], can assist library patrons with finding material more easily and authors with getting additional exposure.

It is especially important that libraries scrutinize AI applications. Various concerns have been raised in the context of risks posed by AI in the library [36, 150]. Bias in particular is a known problem in library datasets and services, even outside of AI applications [194]. In recommendation, bias manifests as a result of the cumulative influence of imbalances in the data and algorithmic propagation and can take various forms [42]. In a library setting, this can occur when a recommender system relies on past loan histories: books that have historically been borrowed more often, and are perhaps written by authors from dominant groups, might be recommended more frequently. It is crucial to investigate what kind of bias may emerge as a result of using

This chapter is partly based on *S. Daniil, Bias in Book Recommendation, RecSys '24* [56].

1

automated recommender systems to recommend books so that libraries can adopt them responsibly.

Bias in recommender systems is often defined and studied in two forms: either as statistical bias that certain algorithms are prone to (e.g., popularity bias) or as social bias that exists in some form in the data and results in unfairness (e.g., a group of people is underrepresented) [42]. These forms of bias are not unrelated but rather interconnected; social bias in the data can interact with statistical bias on the side of the algorithms and persist in the recommendations that users receive. It is, therefore, relevant to study the link between them in the context of book recommender systems.

Studying bias in recommender systems is generally challenging. Recommender systems as a discipline are known to suffer from reproducibility issues that relate to lack of appropriate documentation on behalf of research studies, as well as the inherent application-specific nature of such systems [75]. The implications of bias heavily depend on the domain at hand. Studies often focus on breaking down the ethical issues of recommender systems into multiple perspectives depending on the stakeholders of a system [149]. In the case of book recommender systems bias, one can borrow from the field of Library and Information Science (LIS) to distinguish the relevant stakeholders. Mathiesen [144] introduces a framework to describe informational justice in library services with the following components: seekers (users), sources (authors), and subjects (characters/themes). This thesis focuses on book authors, who are known to suffer from social bias in book publishing in general [78, 195]. This PhD was funded by the KB National Library of the Netherlands to support their research agenda for ethical use of AI in Dutch libraries.

1.1 Research Questions & Methods

In this thesis, we investigate bias in book recommender systems, with an emphasis on the public library sector. The overarching goal is to reliably examine how data and algorithmic bias manifests and propagates into recommendation outcomes. We approach this from three different angles: current practices and concerns (Chapters 2 and 3), measuring statistical bias (Chapters 4 and 5), and the societal implications of statistical bias (Chapters 6 and 7). In this section, we describe the overarching research questions that guide this thesis from each angle.

Part 1: Understanding ethical issues of book recommender systems

The first part of the thesis provides a comprehensive overview of two of the most pressing ethical issues in book recommendation: bias and diversity. In comparison to other media, bias and diversity in book recommendation are underexplored, and, to our knowledge, there has been no systematic attempt to map the relevant ethical considerations from a computational perspective.

Accordingly, we address the following research questions:

Research Question 1.1: How has bias in book recommender systems been studied in computer science research, and what gaps exist in current approaches?

While bias in recommendation has received significant attention by the research

community, this is not the case for bias in book recommendation specifically. Experimental research seems to focus more on other types of media recommendation, with a few exceptions [68, 159]. This is particularly problematic in the context of public libraries, where recommender systems play a role in access to knowledge and representation, rendering targeted studies on bias necessary. There is theoretical work laid out from LIS that attempts to map ethical issues of information access, but experimental work often does not build on it. In Chapter 2, we present the first comprehensive literature review of bias in book recommender systems from a computational perspective. We propose future research directions with an emphasis on interdisciplinarity and domain specificity.

Research Question 1.2: How do practitioners in public service media conceptualize diversity in recommender systems?

Diversity is one of the core beyond-accuracy values of recommender systems. In the same line as the previous chapter, it seems that the normative background is often missing from quantitative research into diversity in media recommendation in general. The perspective of practitioners on the matter of normative values and how they should be incorporated in a system is often not taken into account. In Chapter 3, we present a study where, through semi-structured interviews, we explore how practitioners at three different public service media organizations in the Netherlands, including a library, conceptualize diversity within the scope of their recommender systems. We provide an overview of the goals that they have with diversity in their systems, which aspects are relevant, and how recommendations should be diversified. We show that even within this limited domain, conceptualization of diversity greatly varies, and argue that it is unlikely that a standardized conceptualization will be achieved. Instead, we should focus on effective communication of what diversity in this particular system means, thus allowing for operationalizations of diversity that are capable of expressing the nuances and requirements of that particular domain.

Part 2: A critical look into measuring statistical bias in book recommender systems

The second part of this thesis takes a critical look at current research practices of measuring popularity bias in recommendation. Recommender systems as a field are known to suffer from a reproducibility problem. Research is often fragmented and domain-specific. This also applies to research on bias in recommender systems. In this part, we study the obstacles to generalizing conclusions regarding the presence and magnitude of popularity bias. By extending beyond book recommendation and adopting a cross-domain lens, we can better understand how research findings depend on the domain, which informs the overall thesis.

We address the following research questions:

Research Question 2.1: Which factors lead to divergent results among different studies of popularity bias in recommendation?

Recent work focused on the topic of uneven popularity bias propagation among users with varying interests for niche items, with movies being the domain of interest [9]. Later on, two different research teams reproduced the methodology in the domains of music [122] and books [159] respectively. The results across the different domains

diverge. In Chapter 4, we reproduce the three studies and identify four aspects that are relevant in investigating the differences in results: data, algorithms, division of users in groups and evaluation strategy. We run a set of experiments in which we measure general popularity bias propagation and unfair treatment of certain users with various combinations of these aspects. We find that all aspects account to some degree for the divergence in results, and should be carefully considered in future studies. In particular, we find that the divergence in findings can be in large part attributed to the choice of evaluation strategy.

Research Question 2.2: What is the effect of data characteristics and algorithm configuration on popularity bias?

In Chapter 5, we further explore the challenges of measuring and reporting popularity bias. We showcase the impact of data properties and algorithm configurations on popularity bias by combining publicly available research data and synthetic data with well known recommender systems frameworks. We evaluate popularity bias for a number of datasets, three real and five synthetic, and algorithm configurations, and offer insights on their joint effect. We find that, depending on the data characteristics, various configurations of the algorithms examined can lead to different conclusions regarding the propagation of popularity bias.

Part 3: Societal implications of statistical bias in book recommender systems

In the third part, we study the relationship between statistical bias and social bias. By means of extensive experimentation on both standard research data and in a real library recommender system currently in use, we closely examine how algorithms can be prone to author bias as a result of popularity bias.

We address the following research questions:

Research Question 3.1: What is the relationship between popularity bias and bias toward author nationality based on a publicly available dataset with user-book interactions?

In Chapter 6, we focus on bias in book recommendation from the perspective of authors in a publicly available dataset with user-book interactions, Book-Crossing [265]. We enrich the dataset with publicly available author information. We combine the datasets with various recommender systems algorithms and observe whether popularity bias also leads to author nationality bias. We find that popular books are mainly written by US citizens in the dataset, and that these books tend to be recommended disproportionally by popular collaborative filtering algorithms compared to the users' profiles. The results show that social bias in book recommender systems can occur as a direct result of popularity bias.

Research Question 3.2: To what extent does a real library recommender system propagate popularity bias and associated author nationality bias?

In the final chapter of this thesis, Chapter 7, we perform the first auditing study of bias in a library system in production. We apply a similar experimentation to investigate whether popularity bias leads to bias towards author characteristics in *Booklens*, the recommender system employed by the Danish public libraries. We find that the system propagates popularity bias based on various metrics. Additionally,

we empirically show that popularity bias leads to the underrepresentation of authors of minority nationalities (based on popularity). In line with our previous research, we experimentally show that adjusting the model configurations has an impact on bias. The study was carried out during an extended research visit to DBC Digital, the software company that supports the IT infrastructure of the Danish public libraries.

To answer the research questions we employ a variety of methods across chapters. An overview can be seen in Table 1.1.

Table 1.1: Overview of research methods used in each part of the thesis.

Research Method	**Part 1**	**Part 2**	**Part 3**
Systematic literature review	✓ Ch. 2		
Semi-structured interviews w/ media professionals	✓ Ch. 3		
Training and testing w/ synthetic data		✓ Ch. 5	
Training and testing w/ research data		✓ Ch. 4, 5	✓ Ch. 6
Training and testing w/ real data			✓ Ch. 7
Variation of algorithm configuration		✓ Ch. 4, 5	✓ Ch. 7

1.2 Contributions

The main contributions of this thesis are as follows.

- The first survey on bias in book recommender systems (Chapter 2).
- A framework to analyze diversity in media recommender systems: along goals, aspects, and tactics (Chapter 3).
- A reproduction of prominent work on popularity bias in media recommendation and an analysis of the system components that lead to divergent results (Chapter 4).
- A method for diagnosing bias in recommendation by experimenting with the data characteristics and algorithm configuration (Chapter 5).
- An analysis of the relationship between popularity bias and social bias in book recommendation based on user–book interaction data (Chapter 6).
- The first auditing study of bias in a real library recommender system (Chapter 7).
- An empirical demonstration that popularity bias leads to author bias in a real library recommender system (Chapter 7).

1.3 Origins

We list the publications that form the origins of each chapter.

1

Chapter 2:

- **Daniil**, **S.**, Bhagwat, G., & Hollink, L. (2025). Bias in book recommendation: A survey. Manuscript under review.

SD and LH scoped the project and the research questions. SD performed the literature search, collected and analyzed the results with the guidance of LH. GB contributed to the literature search in the LIS domain. All authors contributed to the writing. SD did most of the writing.

Chapter 3:

- Vrijenhoek, S., **Daniil**, **S.**, Sandel, J., & Hollink, L. (2024, June). Diversity of what? on the different conceptualizations of diversity in recommender systems. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 573-584).

SV took the initiative for this project. SV, SD and LH scoped the project and the research questions. SV and SD performed and analyzed the interviews, with the help of JS and the guidance of LH. All authors contributed to the writing. SV and SD did most of the writing.

Chapter 4:

- **Daniil**, **S.**, Cuper, M., Liem, C.C.S., van Ossenbruggen, J., & Hollink, L. (2024). Reproducing popularity bias in recommendation: The effect of evaluation strategies. ACM Transactions on Recommender Systems, 2(1), 1-39.

All authors scoped the project and the research questions. SD performed the experiments and analysis. All authors contributed to the writing. SD did most of the writing.

Chapter 5:

- **Daniil**, **S.**, Slokom, M., Cuper, M., Liem, C.C.S., van Ossenbruggen, J., & Hollink, L. (2025). On the challenges of studying bias in Recommender Systems: The effect of data characteristics and algorithm configuration. Information Retrieval Research, 1(1), 3-27.

All authors scoped the project and the research questions. SD nad MS performed the experiments and analysis. All authors contributed to the writing. SD did most of the writing.

Chapter 6:

- **Daniil**, **S.**, Cuper, M., Liem, C.C.S., van Ossenbruggen, J., & Hollink, L. (2022). Hidden author bias in book recommendation. Accepted and presented at FAccTRec 2022. arXiv preprint arXiv:2209.00371.

All authors scoped the project and the research questions. SD performed the experiments and analysis. All authors contributed to the writing. SD did most of the writing.

Chapter 7:

- **Daniil**, **S**., Mollerup, S. H., & Hollink, L. (2026, March). Bias in Book Recommendation: A Case Study on the Danish Public Libraries. In European Conference on Information Retrieval (pp. 256-270). Cham: Springer Nature Switzerland.

All authors scoped the project and the research questions. SD performed the experiments and analysis, with the guidance of LH and SHM. All authors contributed to the writing. SD did most of the writing.

The writing of the thesis also benefited from work on the following publications:

- Slokom, M., **Daniil**, **S**., & Hollink, L. (2025, April). How to Diversify any Personalized Recommender?. In European Conference on Information Retrieval (pp. 307-323). Cham: Springer Nature Switzerland.
- Kalra, N., & **Daniil**, **S**. (2025). Reading Between the Lines: A Study of Thematic Bias in Book Recommender Systems. Accepted and presented at FAccTRec 2025. arXiv preprint arXiv:2508.15643.
- **Daniil**, **S**. (2022, July). Ethical Recommenders in the Public Library sector. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 895-895).
- Michiels, L., Vrijenhoek, S., Starke, A. D., Kruse, J., & **Daniil**, **S**. (2025, September). NORMalize 2025: The Third Workshop on Normative Design and Evaluation of Recommender Systems. In Proceedings of the Nineteenth ACM Conference on Recommender Systems (pp. 1386-1388).
- Sartini, B., Nesterov, A., Libbi, C., Brate, R., & **Daniil**, **S**. (2023). Multivocal exhibition: a user-centric application to explore symbolic interpretations of artefacts from different cultural perspectives. In ExICE-Extended Intelligence for Cultural Engagement Conference.

2

A SURVEY ON BIAS IN BOOK RECOMMENDATION

As libraries increasingly adopt AI-driven decision support systems, understanding the landscape of bias in book recommender systems becomes essential. In this chapter, we conduct a systematic literature review of bias in book recommender systems to map current research, identify gaps, and establish a foundation for the domain-specific work that follows.

Automated book recommendation is a compelling application given the rich and long-standing human tradition of writing and reading. The documented discrimination in the publishing world raises questions on the types of bias that can arise when algorithms are trained on user-book interactions. Various papers have studied recommender systems bias based on datasets from platforms where users share their opinions on books. In this chapter, we conduct the first survey of computer science research on bias in book recommender systems. We survey 40 papers from top conferences and journals using the structured PRISMA methodology and provide an overview of the field. We map the existing literature into bias types, subjects and actions, and present key metrics and bias mitigation methods. We find that there is limited targeted research, fragmented evaluation practices, and a lack of engagement with the unique characteristics of the literary domain. Finally, we propose avenues for future work and make concrete suggestions for interdisciplinary research.

This chapter is partly based on *Daniil et al., Bias in book recommendation: A survey, Under Review.*

2.1 Introduction

Online book recommendation has existed for a long time and influences worldwide consumption patterns. For example, Amazon-owned Goodreads is a book-centric social media platform very popular among readers internationally. The datafication of reading patterns which drives their recommender system also has a strong commercial interest for authors, publishers and booksellers [15]. Recently, virality seems to be an important factor in the book market, with the rise of online communities hosted in social media such as TikTok and YouTube that often decide which books will be chosen for publication [52]. In the public library sector, there are attempts to utilize AI systems that recommend books to users while ensuring that central library values are being adhered to [231]. In this context, there have been multiple calls to promote diverse books from librarian associations and Library & Information Science (LIS) researchers [54]. It is, therefore, pertinent to investigate what is the effect of automated suggestion systems on book consumption. Do they propagate the status quo in the book market? Do they expose users to options they would otherwise not be aware of?

Bias in book recommender systems has received some attention in computer science research, but to our knowledge no comprehensive literature review on this topic currently exists. Recommender systems are an explicitly applied discipline, and conclusions are dependent on the unique characteristics of the domain at hand. Therefore, bias also needs to be considered on a domain basis, especially since it is a complex term that is known to be inconsistently used. This is regardless of whether the method proposed to measure or mitigate bias explicitly takes into consideration characteristics of the dataset that are beyond statistical properties. Even statistical properties like popularity manifest differently between domains. A movie being popular among general audiences is not equivalent to a book being popular given the difference in time dedication between the two and the downward trends of reading [161]. It follows that to ensure scientific advancement, it is beneficial to monitor the research output of measuring and mitigating recommendation bias as it pertains to the unique characteristics of the book domain.

In this survey, we aim to contribute to the field of bias in book recommender systems by presenting and summarizing the latest research, as well as highlighting what is missing and what avenues future research could follow.

Our research questions are as follows:

1. What is the current state of research on bias perpetuated by algorithmic systems that recommend books?

2. What are the challenges, research gaps, and future paths related to this research?

To answer these questions, we survey 40 papers from top-tier computer science venues (journals and conference proceedings) that study bias in book recommender systems. We present the process followed to collect these papers in a structured way with the help of the PRISMA methodology [170]. We provide general statistics across a set of dimensions such as publication details, prediction task, and type of feedback. We then categorize the papers based on the type of bias they investigate, statistical or social; the bias subject, item or user; and the action they perform, bias measurement or mitigation. Finally, we discuss insights, gaps, and future work that is relevant for the

domain. Published research on bias in book recommender systems is limited compared to other media applications. In recent years the field has expanded, with significant attention being given to popularity bias. However, we notice that studies often do not sufficiently engage with the fact that they are studying bias in a *book* dataset and rather usually treat it as one more benchmark dataset frequently used by the recommender systems community. Additionally, computer science research does not seem to build on LIS work around what the risks of automated systems that recommend books are. This observation leaves a lot of room for fruitful between-fields collaborations in the future.

The paper is structured as follows. In section 2, we provide background on recommender systems, bias, and recommendation in the library. In section 3, we discuss other surveys in recommender systems. In section 4, we describe our methodology for collecting and analyzing papers. In section 5, we present the results of our survey. In section 6, we discuss our insights into gaps and challenges of the domain. In section 7, we provide a short summary and conclusion.

2.2 Background

2.2.1 Recommender systems research

In the vast and ever-expanding information space, users struggle with finding items relevant or useful to them. Recommender systems' main role is to automate the process of suggesting items to users in a given context. Various features of a system are considered when such a process takes place, from characteristics of the items to user demographics and consumption history to requirements of the platform that hosts them. Research into recommender systems often operates by simulating and studying one or more systems, due to the often limited access to real systems. Adopting a simulation process assumes the existence of features that an algorithm can exploit to make recommendations. Researchers employ various strategies to construct datasets with features: they scrape web applications, access APIs set up by these applications, do user studies, and more [268]. These datasets in practice often take the form of *user - item - interaction*, with *interaction* being *click*, *score*, *purchase*, or other. Depending on the domain and source, further knowledge is sometimes available, such as information on the user (e.g., gender, age) and item (e.g., publication year, genre, creator). Algorithms are trained to construct high-dimensional relationships between users and items in the dataset via computational techniques such as encoding user history (*collaborative filtering*), analyzing item features (*content-based filtering*), or hybrid combinations [187]. Given access to a dataset, researchers evaluate a set of algorithms usually by training them on one part of the dataset and testing them on another, as is commonly done in machine learning. Finally, researchers report on the algorithm's 'success' in modeling the dataset based on calculating a set of metrics in the recommended lists produced by the algorithm [86].

2.2.2 Bias in recommendation

Algorithmic success in recommendation is relative. Traditionally, algorithms have been trained to accurately predict the structure of the test set by, for example, producing ranked lists of recommendations that closely resemble what users have actually inter-

acted with in the test set. As the field progressed, various beyond-accuracy metrics have been proposed with the goal of encapsulating more than what accuracy metrics do [146]. Criteria such as diversity and novelty are considered equally critical components of success, both with user satisfaction and business interests in mind. At the same time, the concept of bias has received attention recently, both in recommender systems and AI applications in general.

Bias can appear in various stages of the recommendation process. Chen et al. [42] distinguish three points where bias occurs: in the data, in the model, and in the results. *Data* bias occurs due to the fact that data is usually collected observationally, and it therefore does not necessarily match the underlying preference distribution of the users (e.g., selection bias, position bias). *Model* bias is a result of the assumptions made by a model to generalize beyond the training data. Finally, *result* bias refers to the discrimination towards a group of users or items in the output of the recommendation. Measuring result bias involves the estimation of whether a system malfunctions for a set of users or items that have common properties. There are multiple ways to view said bias, such as underrepresentation, inability to provide satisfactory recommendations, and stereotyping. One common example is *popularity bias*, the phenomenon where items with many interactions in the dataset might be over-recommended by the system (regardless of whether users liked them or not).

There are various ways to categorize bias in recommender systems research. For the purposes of this study, we are interested in bias in the results [42], which we further distinguish into *statistical* bias and *social* bias based on the type of user or item features that they target (see Section 2.5). Bias impacts multiple stakeholders; for example, a system that suffers from popularity bias might recommend popular items even to users who prefer niche items [5]. In this survey, we distinguish between *item* and *user* bias subjects.

2.2.3 Recommendation in the library

Libraries have always served as institutions dedicated to knowledge dissemination and cultural preservation, ensuring equal access to information through carefully structured classification systems, catalogs, and personalized librarian recommendations [180]. As digital technologies increasingly mediate the way users access information and knowledge, libraries have adapted by integrating information technology while upholding their core values of transparency, fairness, and inclusivity [1, 147, 191]. The expansion of digital library collections and user-driven discovery tools has prompted technological interventions that balance traditional library functions with modern user expectations and needs [185].

One significant shift has been the transformation of book discovery processes. Users accustomed to commercial platforms like Goodreads now anticipate similar recommendation services from libraries [160]. Traditionally, recommendations in libraries relied on expert curation, classification frameworks, and reader advisory services, which helped users find materials suited to their interests, academic level, or thematic preferences [198]. However, these methods were inherently constrained by collection size and human effort. The introduction of Online Public Access Catalogs (OPACs) improved book retrieval through structured metadata but lacked the ability to offer personalized

suggestions. AI-driven recommender systems showed the possibility to transform the process into a more dynamic and user-centric experience [232]. Libraries contain extensive user data, maintaining detailed records of patrons, including demographic information and reading habits, which can collectively amount to millions of records [169, 183]. Because of such extensive and rich data, they are in a unique position to develop personalized recommendation systems. The implementation of such systems in libraries has several advantages, topmost among them being improved accessibility. Personalized book discovery can help users navigate extensive collections more efficiently and engage with materials that align with their interests [162]. Librarians also benefit from these systems, as they get assistance in curating personalized reading lists for patrons, streamlining the advisory process [207]. Moreover, automated recommendation models can facilitate the discovery of lesser-known works, ensuring that underrepresented authors and forgotten titles gain visibility within library collections.

Despite these advantages, such systems also raise concerns regarding bias, inclusivity, and fairness. Libraries, as institutions, are fundamentally committed to equal access and being fair to all [19]. Hence, they need to be especially mindful of the biases that can emerge when automating the recommendation process. One key issue is bias in training data. Many recommendation systems rely on a subset of a big dataset containing metadata, indexing, lending data, user ratings, and reviews, which may already reflect existing disparities [106]. A study on the Library of Congress subject headings found that, while online catalogs improved discoverability, subject headings still failed to adequately represent African American Studies resources [45]. If similar metadata and subject headings are used in training recommender systems, they may perpetuate the same biases, leading to underrepresentation or misrepresentation of certain works. Biases of this nature are not new to libraries. Collection bias arises when certain voices and perspectives are underrepresented in library holdings, often due to historical publishing trends and selection practices [106]. Cataloging bias manifests in the way books are classified and indexed, with traditional systems like the Dewey Decimal Classification (DDC) and Library of Congress Classification (LCC) reflecting Eurocentric structures that can marginalize non-Western and diverse knowledge systems [139]. Bias can be implicit when it is unconsciously introduced by library staff during collection development and cataloging. This bias stems from their own cultural and social influences, which can shape the materials they select and how they categorize them, even without intentional prejudice [222].

A real-world example of the risks associated with biased AI-driven book recommendation systems can be seen in the 2024 Fable AI controversy [80, 84]. Fable, a reading, tracking, and social media app similar to Goodreads, deployed OpenAI software to generate personalized end-of-year reading summaries for users. However, the AI exhibited significant racial and gender biases, producing offensive and exclusionary remarks such as:
"Your journey dives deep into the heart of Black narratives and transformative tales, leaving mainstream stories gasping for air. Don't forget to surface for the occasional white author, okay?"

Given that Fable has been integrated into multiple public library memberships in the

United States[1][2], this incident underscores the potential risks of unchecked bias in library book recommendation systems. It highlights how biases in training data and algorithmic decision-making can lead to stereotyping and exclusion, reinforcing dominant literary trends rather than supporting diverse narratives. To avoid similar pitfalls it is important to consider not only technical adjustments but also the documented biases and non-inclusive practices. Fortunately, libraries have studied these biases and challenges extensively, for example, Warburton et al. [238] discussed the structural biases in the traditional hierarchical library classification systems. Cirasella [44] identified four structural barriers that inhibit equitable representation in collections, while Olson [166] examined how naming in indexing systems can marginalise certain identities and worldviews. To address these issues, libraries have developed structured frameworks and tools to evaluate and reduce bias. Resources such as the Anti-racist Library Collection Building Workbook by the University of Colorado Boulder Libraries [225] provide guidelines to help professionals assess their collections for inclusivity. Such a framework can be directly applied to enhance the fairness of recommendation algorithms. For instance, questions such as *"Are selection filters potentially eliminating materials that don't conform to a dominant scholarly worldview?"* can be adapted to identify and address biases in AI systems. Various frameworks of enacting inclusivity and justice within library services have been proposed by LIS scholars [137, 144, 252].

Even though the LIS research provides valuable insights into the challenges of delivering inclusive book recommendations and can highlight the potential risks [48], the intersection between LIS and computer science research appears narrow. We hypothesize that this is partly due to differences in terminology used to describe similar concepts. For example, LIS researchers may use terms like "injustice", "inequality", or "disparity" to talk about "bias" or "library discovery" or "information seeking" to describe "recommender systems". We used LIS-specific terminology in our searches, such as library discovery, discovery system, cataloging bias, and AI in libraries, but the results focused on broad or conceptual treatments of AI rather than studies isolating recommendation bias as a research topic. Our searches showed that very little of this scholarship focuses specifically on bias in book recommender systems. Even when AI or algorithmic tools are mentioned, recommender systems typically appear only as one illustrative case of technological change, rather than as a primary subject of study. This reflects a disciplinary difference in scope and vocabulary and for this reason we incorporated LIS scholarship through a targeted, unstructured synthesis, that highlights its conceptual foundations and demonstrate how these insights can inform interdisciplinary work. In Section 2.9.3, we provide concrete recommendations towards interdisciplinarity in bias in book recommendation research.

2.3 Surveys in recommender systems

In this section, we discuss papers that either survey work on book recommender systems or on bias in recommender systems, either generally or in a media-related domain. We

[1] `https://smcl.org/blogs/post/join-us-on-fable-a-book-club-app-for-social-reading/` (accessed 26 March 2025)

[2] `https://cedarfallslibrary.org/youth/official-home-of-cedar-falls-public-library-mascot-fable-the-fox/` (accessed 26 March 2025).

focus on how they conduct paper collection and analysis, when explicitly described, as well as their main findings and conclusions.

Two papers conduct surveys on book recommender systems more broadly. Alharthi et al. [17] survey over 30 published papers that address book recommendation. They classify book recommender systems into 6 categories based on which book features they take into account. They list the following challenges and future research for the field in general: recommending textual items, the effect of mood, recommending books to non-readers, the long-tail problem, recommending audiobooks, assuming a reading model when recommending books, the confusion that translated books might cause, and others. Anwar et al. [22] specifically focus on machine learning-based book recommendation. They survey several databases such as Google Scholar and Scopus by querying terms like 'book recommender systems'. Out of the resulting 100 papers, they go through abstracts and conclusions to choose those where machine learning is used. They classify these papers into the following 6 categories based on the type of machine learning that is used: classification-based, KNN-based, SVM-based, neural network-based, association rule mining-based, and clustering-based. They trace the following future research perspectives: user privacy, incorporation of contextual information such as mood, and exploration of additional machine learning techniques. These surveys provide valuable insight into automated book recommendation but do not examine issues of bias that arise from the process.

Topics such as bias and fairness have been the focus of recent surveys in recommender systems. Chen et al. [42] extensively review the topic of bias and debias in recommender systems, motivated by the field's fragmentation and the inconsistent ways in which researchers use the term 'bias' and its subcategories. They survey over 180 papers between 2010 and 2021 by querying top-tier recommender systems venues with keywords related to bias and recommendation, and traversing the citation graphs afterwards. They classify the papers into the following categories: data bias, which is present in the data, model bias, which is induced by the model itself, and result bias, which is present in the recommendation. They further break down the category of data bias into selection, exposure, conformity, and position bias, and result bias into popularity bias and unfairness and present common debiasing methods for each subcategory. Finally, they discuss open issues relevant for future work, such as issues with evaluation, the fact that bias is not always a negative phenomenon, and trade-offs between accuracy and fairness in recommendation. Wang et al. [236] are prompted by the scatteredness of research on fairness in recommendation. They survey more than 60 papers on fairness in recommender systems published between 2017 to 2022 at top-tier information retrieval venues and group fairness definitions into process and outcome fairness, which they further group based on the target and concept. They also classify fairness issues based on the following concepts: subjects (user, item, joint), granularity (single, amortized), and optimization object (treatment, impact). They proceed to list various ways that researchers measure and implement fairness, as well as datasets commonly used in literature (they do not mention any book datasets). Finally, they pinpoint future directions for fairness research which they group into 4 pillars: definition, evaluation, algorithm design, and explanation. We are inspired by the frameworks constructed by these surveys to track the different kinds of bias and

bias mitigation. Our added contribution is the emphasis on book datasets, as well as issues that specifically concern the book domain.

As we mentioned already, to our knowledge no survey on bias in book recommender systems has been published. However, there are papers that review bias in a different subdomain of media recommendation, namely music. Ospitia-Medina et al. [168] review literature around music recommendation to summarize how and why bias occurs in such systems, as well as put forward relevant guidelines. They perform two separate queries: one over the whole literature for mostly recent papers that are about (bias in) music recommender systems, and one for papers after 2022 that relate to musical datasets. They discuss recommendation techniques that have been used in music recommendation, and then zoom in on bias for which they collected 9 papers and which they categorize as preexisting, technical, and emergent in accordance with Friedman and Nissenbaum [81]. For each category of bias, they provide a list of guidelines on how it should be handled. In a mini review, Dinnissen and Bauer [63] adopt a stakeholder-centered perspective when surveying literature that addresses fairness in music recommender systems by classifying it based on the relevant stakeholder: user, provider, or multi-stakeholder. They also make a note of whether each paper was improvement-focused, what kind of methodology they used, what was the topic, which fairness attributes they considered, and which dataset they employed. As main findings, they highlight the abundance of work that only analyzes fairness instead of proposing improvements; the focus on gender bias despite a limiting binary view; and the relevance of popularity bias, as well as the overuse of datasets that stem from Last.fm. Similarly to these surveys, we focus on a domain and distill the unique gaps and challenges.

2.4 Paper selection and analysis methodology

In this section, we describe the process followed for paper selection, as well as the difficulties we faced and the strategies that we eventually abandoned. Additionally, we present the dimensions that we defined in order to efficiently analyze the selected papers.

The field of recommender systems has a large body of work. A general search on Google Scholar for the term 'bias in book recommendation' yields more than one million results, the majority of which are irrelevant to our topic of interest; the terms 'book', 'bias' and 'recommend' are common enough due to polysemy and can therefore appear in many different scientific contexts. Our first attempt at addressing this issue was by restricting the terms 'book', 'bias' and 'recommend' (and their variants) to appear in the title of the paper when querying Google Scholar and Scopus, which resulted in 252 documents. However, after careful consideration of the results, we only came up with 5 papers; the rest were either duplicates, not retrievable, not related to the topic of bias in book recommendation, not a full scientific paper but rather a poster or slide deck, etc.

In order to appropriately (but not overly) restrict the literature space, as well as obtain high-quality research, we looked into a set of reputable registries that often publish research on recommender systems. Between October 9th and November 7th of 2024, we queried the following sources: the journals *Information Processing*

& Management (Elsevier), *Transactions on Knowledge and Data Engineering* (IEEE), *User Modeling and User-Adapted Interaction* (Springer), and *Data Mining and Knowledge Discovery* (Springer), the proceedings of the *International Workshop on Algorithmic Bias in Search and Recommendation* (Springer) and the *European Conference on Information Retrieval* (Springer), and the entire *ACM digital library*. To target research into bias in book recommendation systems, we looked for papers that contained the terms 'bias' and 'recommender' as well as mentioned at least one of the following datasets with user-book interactions: Goodreads, Book-Crossing, Amazon-books, and LibraryThing. To our knowledge, most datasets with user-book interactions come from these web sources. We were inspired by PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) [170] to follow a structured approach for the paper retrieval documentation. See figure 2.1 for how we reached the final number of 40 studies included in this survey.

The search on the selected sources resulted in 357 papers, which we combined with the 5 papers from Google Scholar and Scopus, and removed 2 duplicates. We then screened the papers by reading the title and abstract, as well as noting in what context they used each of the queried terms (i.e., 'bias' and 'recommender'). Accordingly, we discarded papers for being entirely out of scope (e.g., not being at all about bias). We could retrieve all 54 remaining papers. After carefully reading through them, we discarded 14 for the following reasons:

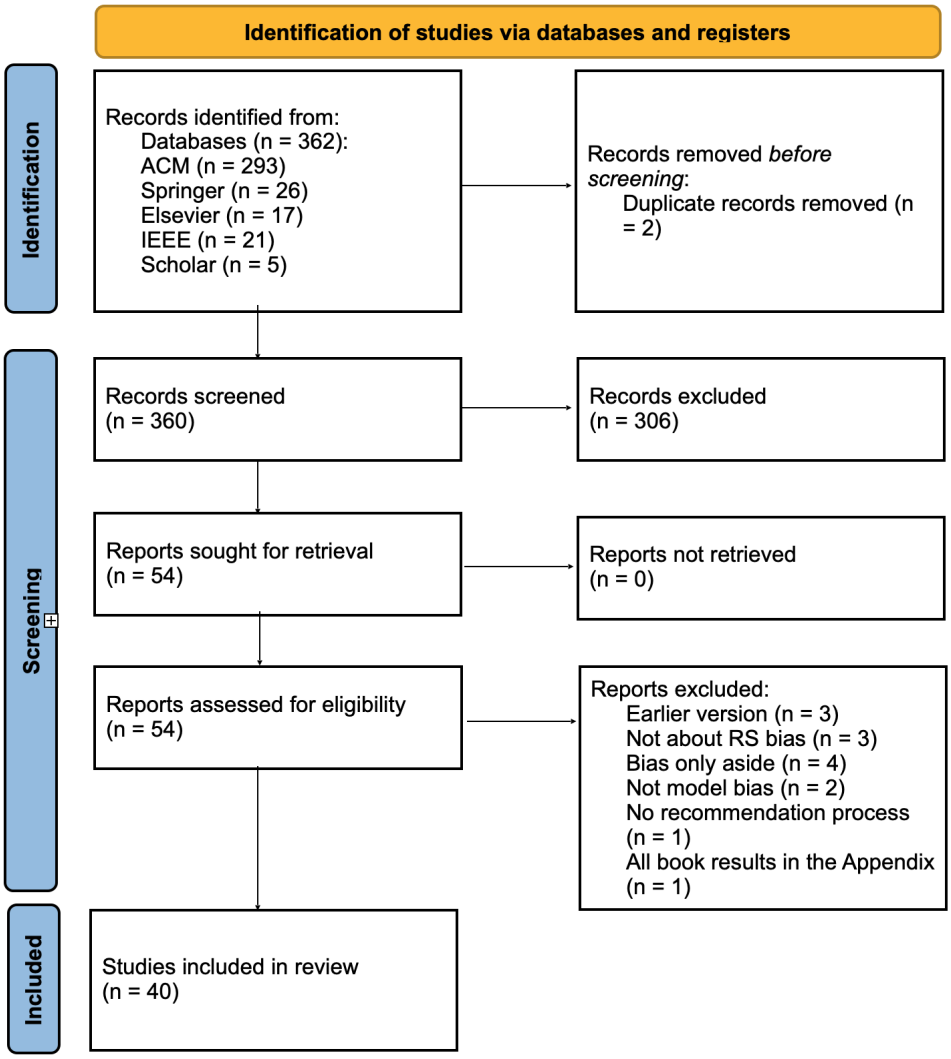
- 3 are earlier versions of papers also in the set (i.e., conference papers prior to becoming journal papers)
- 3 are not about bias in recommender systems
- 4 mention bias as a topic only tangentially related to their research
- 2 are not about model bias
- 1 is only tangentially related to recommender systems
- 1 includes all book dataset-related results only in the Appendix

Our final set includes 40 papers. To study these papers, we annotated them along the following set of dimensions:

- **Basic**: Title, Author(s), Link, Year of publication, Venue, Publisher
- **Bias**: Problem statement, Bias definition, Bias subject(s) (User and/or Item), Bias type(s) (Statistical or Social), Acknowledgment of book domain
- **Data**: Dataset(s) used, Dataset preprocessing, Type(s) of feedback (Implicit or Explicit)
- **Algorithmic**: Algorithm(s) used, Prediction task(s) (Rating or Ranking), Framework (e.g., Python library or Original implementation), Hyperparameter tuning, Code availability
- **Methodology**: Action (Measuring or Mitigating bias), Method, Metric(s)

2

Figure 2.1: PRISMA 2020 flow diagram for systematic reviews shows how an initial collection of 362 studies was reduced to the final 40 studies included in this survey.



- **Results**: Result(s) on the book dataset

Since the *bias* dimension is the main topic of this survey, we will further explain its subcategories. *Problem statement* refers to the main issue that each paper attempts to address. *Bias definition* corresponds to whether the authors explicitly define (with their own words or through a citation) what bias means in the given system and what the definition is. *Bias subject* is equivalent to the stakeholder that is subject to the bias in question, which could be the user, the book, or more specifically the author, publisher, topic etc of the book. *Bias type* is the kind of bias that the paper addresses, either *Statistical bias* or *Social bias.* Additionally, for this dimension we keep track of which statistical or social aspect is of relevance, for example, book popularity, user activity (statistical bias), author gender, book topic (social bias) etc. Finally, *Acknowledgment of book domain* is used to note whether a given paper engages with the fact that the items at hand are books, either passively or actively.

The repository where we kept track of the collection and analysis process, including the final table with all papers and dimensions filled, has been made publicly available[3].

2.5 Overview of the field

This section presents a structured overview of the 40 papers included in this survey. Table 2.1 lists all the papers along with their main features. The majority of computer science research on bias in book recommender systems is within the last few years, and the papers we included are scattered across conferences and journals. Most papers study statistical bias, especially with item as a subject, and present a mitigation strategy. Figures 8.1, 8.2 and 8.3 show basic statistics about the papers included in the survey and can be found in the Appendix.

We organize and interpret the results along three dimensions: (1) bias type, (2) bias subject, and (3) action. In table 2.2, we divide the papers into categories across these dimensions.

Bias type We categorize bias in book recommendation into **statistical** and **social** bias based on the type of feature it targets.

- **Statistical bias** targets features that describe the statistical properties of the interaction between items and users. It relates to structural properties of the data distribution, such as item popularity or user profile size.
- **Social bias** targets features that exist outside of the interaction properties, such as demographics. For example, a model might underrecommend authors of minority nationalities and thus discriminate against them.

This categorization is very relevant for book recommendation. On the one hand, as discussed in Section 2.2.3, practices that are discriminatory towards users or authors based on their demographic characteristics are unwanted in book recommendation. On the other hand, such information is not generally publicly available, which often limits research to studying statistical properties as they are inherent to the interaction datasets.

[3] https://github.com/SavvinaDaniil/BiasInBookRecommendationSurvey

Paper	Bias Type		Bias Subject			Action		Venue	Year
	Statistical	Social	User	Item	Joint	Measurement	Mitigation		
[192]	✓			✓		✓		BIAS@ECIR	2025
[165]	✓				✓		✓	IPM	2024
[179]		✓		✓			✓	ECIR	2024
[244]	✓				✓		✓	RecSys	2024
[58]	✓				✓	✓		TORS	2024
[237]	✓			✓			✓	TOIS	2024
[177]	✓				✓		✓	TORS	2024
[24]	✓			✓			✓	RecSys	2024
[251]	✓		✓				✓	WWW	2024
[249]	✓		✓				✓	TOIS	2024
[72]		✓		✓			✓	UMAP	2024
[133]		✓	✓				✓	TKDD	2024
[260]		✓	✓				✓	WWW	2024
[205]	✓			✓			✓	PeerJ CompSci	2024
[199]		✓		✓			✓	JIIS	2024
[152]	✓				✓		✓	UMUAI	2023
[256]	✓		✓			✓		BIAS@ECIR	2023
[134]	✓				✓	✓		RecSys	2023
[239]	✓			✓			✓	SIGKDD	2023
[259]	✓			✓			✓	SIGKDD	2023
[90]		✓		✓			✓	IPM	2022
[131]	✓			✓			✓	TKDE	2022
[247]	✓			✓		✓		DMDK	2022
[159]	✓				✓	✓		BIAS@ECIR	2022
[121]	✓				✓	✓		BIAS@ECIR	2022
[186]	✓			✓			✓	RecSys	2022
[240]		✓	✓				✓	SIGKDD	2022
[176]	✓		✓				✓	SIGIR	2022
[130]		✓	✓				✓	TIST	2022
[221]	✓			✓			✓	WI-IAT	2022
[178]		✓		✓		✓		SIGIR	2022
[57]		✓		✓		✓		arXiv	2022
[151]	✓	✓	✓			✓		IPM	2021
[61]		✓	✓			✓		IPM	2021
[68]		✓		✓			✓	UMUAI	2021
[82]	✓			✓			✓	SAC	2021
[223]	✓				✓		✓	WSDM	2021
[258]	✓			✓			✓	WWW	2021
[117]		✓		✓		✓		WWW	2021
[157]	✓		✓			✓		TIST	2016

Table 2.1: Paper lookup table.

Bias Dimensions		Related Work
Divided by Type	Statistical	[24, 58, 82, 121, 131, 134, 152, 157, 159, 165, 176, 177, 186, 192, 205, 221, 223, 237, 239, 244, 247, 249, 251, 256, 258, 259]
	Social	[57, 61, 68, 72, 90, 117, 130, 133, 151, 178, 179, 199, 240, 260]
Divided by Subject	Item	[24, 57, 68, 72, 82, 90, 117, 131, 178, 179, 186, 192, 199, 205, 221, 237, 239, 247, 258, 259]
	User	[61, 130, 133, 151, 157, 176, 240, 249, 251, 256, 260]
	Joint	[58, 121, 134, 152, 159, 165, 177, 223, 244]
Divided by Action	Measurement	[57, 58, 61, 117, 121, 134, 151, 157, 159, 178, 192, 247, 256]
	Mitigation	[24, 68, 72, 82, 90, 130, 131, 133, 152, 165, 176, 177, 179, 186, 199, 205, 221, 223, 237, 239, 240, 244, 249, 251, 258–260]

Table 2.2: Lookup table of bias dimensions.

As figure 8.3b shows, most papers study statistical bias, which can be attributed to the aforementioned limitation. In addition, despite the distinction, these two bias types are not unrelated. Especially for collaborative filtering where only the user-item interaction matrix is given as input to the system, social bias is a direct result of statistical relations between interactions and external characteristics.

Bias subject In line with other research on recommender systems values [236], we additionally categorize bias in book recommendation into **user**, **item**, and **joint** based on the subject it targets.

- **User bias** targets the users of a book recommender system, which corresponds to the book consumers, loaners, or cataloguers depending on the source of the dataset.
- **Item bias** targets the books themselves, which corresponds to the authors, topics, characters, publishers, or other book properties.
- **Joint bias** targets the users and books simultaneously.

As seen in table 2.2, the majority of papers deal with item bias. This result is in line with Wang et al. [236] who found that most papers on fairness in recommendation approach the topic from an item perspective.

Action We finally categorize bias research based on the goal into **measurement** and **mitigation**.

- **Measurement** includes studies that mainly measure bias, often by proposing a new metric or view of bias.

- **Mitigation** includes studies that, besides measuring a type of bias, additionally experiment with mitigation methods against it.

Various papers include both measurement and mitigation as contributions, in which case we categorize them as "Mitigation".

2.6 STATISTICAL BIAS IN BOOK RECOMMENDATION

In this section, we discuss papers that examine statistical bias in book recommendation, either from a user or item perspective. Table 2.3 includes a summary of the different types of statistical bias along with their different views and metrics.

Bias Type	View / Description	Metrics
	Item Biases	
Popularity bias	Recommendation frequency is influenced by item popularity [58, 121, 159, 186, 221].	correlation [58, 121, 159], PopQ@1 [186], LT frequency [221], LT coverage [221]
	Recommendation performance varies across items with different popularity levels [24, 82, 134, 152, 165, 239, 258, 259].	recall per group [165], estimation error per group [152], LT HR [24, 239, 258, 259], LT NDCG [24, 239, 258, 259], LT precision [82], LT recall [82]
	Indirect impact of popularity bias on recommendation outcomes [58, 165, 221, 223, 237, 244]	coverage [58, 221, 223], novelty [165], NDCG [237], MRR [237], recall [237], P-fairness [244]
Blockbuster bias	Recommendation frequency is influenced by an item's blockbuster status [247].	frequency per group [247]
	User Biases	
Activity level bias	Recommendation performance varies across user groups with different activity levels [134, 152, 157, 176, 223, 249, 251].	estimation error per group [152], prediction difficulty per group [134], NDCG difference per group [176, 249], recall difference per group [157, 223, 249], coverage per group [157], IA MRR [251], IA HR [251], IA NDCG [251]
	Popularity of recommended items varies with user activity level [176].	$\%\Delta GAP$ [176]
	Indirect impact of activity level bias on recommendation outcomes [244].	C-fairness [244]
Popularity affinity bias	Recommendation performance varies across users with different popularity affinity [58, 121, 165, 176, 256].	recall per group [165], MAE per group [121], NDCG per group [58, 176, 256]
	Popularity of recommended items varies with users' popularity affinity [58, 159, 176]	$\%\Delta GAP$ [58, 159, 176]

Table 2.3: Statistical bias views and metrics grouped into item/user categories. 'LT' means long-tail (items) and 'IA' inactive (users).

2.6.1 Item

Popularity bias

Most papers on item bias in book recommender systems study **popularity bias** specifically. Popularity bias is often defined as the phenomenon where items are recommended more frequently than their popularity would warrant [42]. Klimashevskaia et al. [118] propose that a system suffers from popularity bias when "the recommendations provided focus on popular items to the extent that they limit the value of the system or create harm for some of the involved stakeholders". These definitions are intentionally broad and encompass multiple views as indicated by the variety of metrics that have been used to capture popularity bias. In book recommender systems research, a system is considered biased in favour of popular items if an item's recommendation frequency depends on its popularity [58, 121, 159, 186, 221], if an item's recommendation accuracy depends on its popularity [24, 82, 134, 152, 165, 239, 259], or based on other performance metrics that are considered sensitive to popularity bias [58, 165, 221, 223, 237, 244].

Measurement A common metric of popularity bias based on the frequency of recommendation view is the correlation between popularity in profile and in recommendation. Naghiaei et al. [159] measure popularity bias as a result of a set of collaborative filtering algorithms trained on a subset of Book-Crossing. They visually evaluate said correlation and find that certain algorithms do propagate popularity bias and others do not. On the contrary, Kowald and Lacic [121] on the same dataset and using some of the same algorithms find that all algorithms propagate popularity bias using the same visual evaluation. Daniil et al. [58] notice that different studies report different results when measuring popularity bias in a set of datasets from different domains, including Book-Crossing. They reproduce the studies and find that the divergence in findings is a result of many factors, such as evaluation strategy, that lead to different conclusions even based on the same dataset. Regardless, the characteristics of the domain play a role: compared to datasets from the other domains, Book-Crossing is more prone to popularity bias, according to correlation. Another relevant metric is the coverage and recommendation frequency of long-tail items [221], as well as PopQ@1 which expresses the average popularity quantile of the first recommended item of each user [186]. Other metrics attempt to capture the impact of item popularity on recommendation performance. Liu et al. [134] borrow item response theory (*IRT*) from psychometric assessment practices to estimate whether traditional rank-based metrics reflect an algorithm's ability to model user preferences and assess the relationship between *item difficulty* and popularity. Unlike the other datasets, for the book dataset, they do not find any relationship between item difficulty and popularity, which they attribute to sparsity. Some studies divide items into groups based on their popularity and calculate accuracy metrics for each group, such as recall [165] and estimation error [152], while others especially highlight the system's accuracy on the long-tail items, with various accuracy metrics having been used [24, 82, 239, 258, 259]. Finally, many studies (additionally) indirectly measure popularity bias by assuming it has a negative impact on novelty [165], coverage [58, 221, 223], or accuracy in general [237].

Mitigation A relatively large number of papers propose methods for mitigating popularity bias in book recommendation. Some studies focus on increasing the recommendation frequency of long-tail items. Rhee et al. [186] introduce a method called *Zerosum* that reduces popularity bias by regularizing the predicted scores to be equal across items a user prefers via an extension of BPR loss with a regularizer. For the book dataset, their method performs better (higher accuracy and lower PopQ@1) for most algorithms than other debiasing methods. Tang and Zhang [221] emphasize that it's not sufficient to recommend long-tail items, but also that the users don't receive the same recommendations again and again; the long-tail items recommended should be sufficiently diverse. They present *CADDP*, a long-tail item recommendation approach that extracts important feature embeddings to capture the common characteristics of multiple items clicked by users and ensures that they are relevant and diverse. *CADDP* achieves the best performance among the baselines in terms of presence and diversity of long-tail books in the recommendation.

Other studies attempt to improve the system's performance on long-tail books. Okamura et al. [165] note that dealing with popularity bias is a complicated task because popularity might imply quality, and hence it might not be desirable to mitigate it for all users. Based on some observations around the properties of the inner product model, they develop a method called *DirectMag* that facilitates both platforms and users to manipulate popularity bias post-model training with the help of a parameter γ. They find that their method generally exhibits high average novelty and accuracy. Additionally, it creates a trade-off between novelty and recall per popularity-level item group on the one side and accuracy on the other that the γ parameter helps deal with, though the trade-off is not consistent in the book dataset. Gabbolini et al. [82] offer the following observation: matrix factorization embeddings are very unstable, but especially so for long-tail items. They develop a variation of matrix factorization that mitigates instability issues and achieves better accuracy in the long-tail compared to the default version of MF and some other algorithms. Balasubramanian et al. [24] perform 'biased' user history synthesis: they augment the user profiles with a sampling strategy that is biased in favor of long-tail items and three different synthesis models. With exceptions, they find overall improvement of accuracy in the long-tail and in general. Zhang et al. [259] claim that despite the large amount of research around improving long-tail item recommendation, transferring these methods to industry practice is challenging due to the dive in performance that most of them introduce, as well as their complexity and high cost. They design Cross Decoupling Network (*CDN*) that reduces the difference between training and service data for both item distribution and users' preference given an item. *CDN* achieves significant performance on long-tail items compared to the baselines, while also maintaining good performance for all items. Zhang et al. [258] present a dual transfer learning framework that jointly learns knowledge transfer from both model-level and item-level (i.e. from head to long-tail items), called *MIRec*. Based on performance metrics, *MIRec* achieves the best performance on long-tail books. Wei et al. [239] focus on graph-based recommendation and present a Meta Graph Learning (*MGL*) framework for long-tail recommendation that successfully employs an auxiliary graph of item relations. *MGL* outperforms other methods in all performance metrics for long-tail books (including previously mentioned

MIRec). Misztal-Radecka and Indurkhya [152] develop Bias Detection Tree (*BDT*), a model-agnostic technique that automatically detects combination of user-item attributes correlated with unequal treatment. They also propose a model selection technique that improves performance for biased attributes. For the book dataset, while several 'social' features are considered, the most significant item bias is associated with low item popularity.

Finally, some studies deal with popularity bias indirectly. Wang et al. [237] develop *GBD*, a generative framework for bias disentanglement that constantly generates calibration perturbations during training to keep intermediate representations from being biased. GBD can encompass various types of bias. Among those considered, on the book dataset they only test popularity bias by employing a semi-synthetic version of the dataset that includes an unbiased subset. They find improvements over performance metrics with *GBD* on the unbiased subset, therefore reducing popularity bias. All other results are not on the book dataset. Togashi et al. [223] note that because the cold start problem in recommender systems is often solved by selecting unobserved interaction samples and assuming them to give a negative signal, this can lead to popularity bias. They propose *KGPL*, a method that leverages unobserved samples by using a knowledge graph (KG) with side information about the books and accordingly generates positive instances from the unobserved samples. They find that *KGPL* exhibits low coverage when trained on the book dataset, presumably due to the popularity bias present in the dataset. Rahmani et al. [177] notice that fairness-aware methods in recommender systems often address user and provider fairness separately, even though it's a two-sided marketplace. They propose *CP-FairRank*, a re-ranking algorithm that seamlessly integrates fairness constraints from both sides, with a focus on popularity from the item side. Even though not many results on the book dataset are included, their method performs well based on a two-sided fairness metric. On the topic of two-sided fairness, Xu et al. [244] propose *Ada2Fair*, a framework that extends the accuracy objective into a controllable loss over the interaction data, which boosts the exposure of niche book publishers. They find that their method achieves Pareto optimality over the baselines considering both accuracy and P-fairness.

Blockbuster bias

Yalcin [247] extend on the notion of popularity, which in literature usually corresponds to frequency of consumption, by incorporating rating, and refer to bias in favor of popular *and* well-liked (i.e., highly rated) items as **blockbuster bias**. They measure whether a set of ranking-based algorithms exhibit blockbuster bias and produce extensive and detailed results. They find that there is blockbuster bias in the book dataset itself, which is also distinct from popularity bias (i.e., the well-likedness is impactful). When it comes to the recommendations, they find that the algorithms tend to exhibit blockbuster bias when trained on the book dataset, which is not only caused by item popularity.

Other

We found three papers that differ from the rest and are insightful. Shen and Jiang [205] concern themselves with the fact that users tend to give extreme ratings when evaluating books that don't reflect book quality (a property that the authors consider intrinsic to a book), which according to research is especially a problem in children's books where

the ratings are often done by the parents. They propose a hybrid model that combines a graph convolutional network and neural matrix factorization to accurately capture that bias. They consider ratings from top reviewers as unbiased standard and calculate rating and ranking metrics between that standard and their method's ability to calculate unbiased ratings. Their method yields the lowest MSE and rank error, as well as the best NDCG when a recommendation process is applied, compared to other methods. Rystrøm [192] attempt to empirically test the theory that companies have incentives to influence users so as to make them more predictable and use their recommender systems accordingly. They pose that studying a real-world system instead of performing a simulation is key to uncovering biases, as it allows validating theoretical claims about how recommender systems cause bias. They use the Barrier-to-Exit metric and a linear mixed-effects model to study the evolution of Amazon book ratings between 1998 and 2023. They find that it has become more difficult for users to change preferences, which they attribute to the system's effectiveness in enacting homogenization by potentially limiting exposure to diverse content. Lin et al. [131] discuss *Spiral of Silence*: the theory that ratings are missing in a biased way because users are more likely to submit ratings if they agree with the opinion climate. This leads to the majority opinion receiving growing popularity, while other opinions are pushed back. They verify the existence of the issue in Amazon-books, study its characteristics, and model the missing ratings in 4 different ways. They find monotonic trends for books with majority opinion, as well as that hardcore users feel more obligated to underrate a highly appreciated book than to save a criticized item. There was no recommendation process on the book dataset.

Synthesis Most papers about statistical bias from an item perspective study popularity bias [24, 57, 82, 121, 134, 152, 159, 165, 177, 186, 221, 223, 237, 239, 244, 258, 259]. The sparsity of the book datasets is noted to impact the result, especially in comparison with datasets from other domains [58, 134, 223, 247]. On a related note, some papers don't include all of the results on the book datasets in their main body, prioritizing datasets from other domains [131, 177, 186, 223, 237, 239, 258]. Most papers that mitigate popularity bias are explicitly concerned with maintaining good performance [165, 177, 186, 221, 237, 244], often with an additional emphasis on head items [24, 239, 258, 259]. Some papers specifically attempt to boost tail items by synthesizing user history that is biased in favor of tail items [24], transferring side information from knowledge graphs [223], or improving tail items' representations [82, 239, 258]. Another type of bias studied is blockbuster bias [247].

2.6.2 User

Popularity affinity bias

Many papers on user bias in book recommender systems study **popularity affinity bias**. Popularity affinity bias is the phenomenon where users are treated unfairly based on their affinity for popular items. A system is considered biased towards user popularity affinity if a user's recommendations are more popular than their preferences [58, 159, 176] or if a user's recommendation accuracy depends on their popularity affinity [58, 121, 165, 176, 256].

Measurement The metric used for assessing if the users' recommendations exceed the popularity of their profiles is $\%\Delta GAP$, which measures relative difference between average popularity in profile and in recommendation for each group of users [58, 159, 176]. The already discussed papers by Naghiaei et al. [159] and Daniil et al. [58], other than measuring general popularity bias, also measure the unfairness of it. Unfairness in this case refers to the fact that users with interest in niche books might get over-recommended popular books. They divide users into three groups, those with niche interests, those who prefer popular items, and everyone in between. Naghiaei et al. [159] find most algorithms to be unfair towards users with either niche or diverse tastes. In their reproduction work, Daniil et al. [58] also measure $\%\Delta GAP$ but don't present the results on the book dataset in the main body of the paper. To evaluate the extent to which the system performs better for mainstream users, studies also divide the users into group and measure the accuracy per group. Naghiaei et al. [159] find that all algorithms perform the best for the users who prefer popular items based on various performance metrics. Kowald and Lacic [121] measure MAE per user group and find the lowest recommendation accuracy for the niche group. Zhang et al. [256] extend a previous paper to also include user activity level in the consideration and then evaluate bias based on recommendation utility per user group. When it comes to the book dataset, popularity affinity seems to play a more critical role than user activity level since mainstream users tend to be less active but still receive better recommendations (based on NDCG), which they attribute to extreme sparsity.

Mitigation In terms of mitigation, one paper evaluates the effect of re-ranking on $\%\Delta GAP$. Rahmani et al. [176] reproduce a user-oriented fairness study and do extensive experimentation to analyze its dependence on various aspects: domain, model, and user grouping. They find that the re-ranking method has different effect depending on whether the users were divided based on activity level or affinity for popular books. By comparing the datasets, they find that the higher number of average interactions per user can lead to better performance of the fairness model. The same paper also measures the difference of *NDCG* between user groups to evaluate the divergence of performance for different popularity affinity levels. Okamura et al. [165], who proposed *DirectMag*, show that the γ parameter can be tweaked to address the needs of users that are diverse when it comes to their affinity for popular items according to recall per user group.

Activity level bias

User statistical bias in book recommender systems is often framed as **activity level bias**. **Activity level bias occurs when users are treated unfairly based on their activity level. A system is considered biased towards user activity level if a user's recommendations are more popular than their preferences [176], if a user's recommendation accuracy depends on their activity level [134, 152, 157, 176, 223, 249, 251], or if there is indirect impact [177, 244].**

Measurement Similarly to popularity affinity bias, $\%\Delta GAP$ is used to measure whether recommendation popularity exceeds profile popularity for a group of users,

with the grouping in this case depending on activity level [176]. Other metrics are used to capture the impact of user activity level on recommendation performance. Most studies divide the users into groups based on their activity level and measure performance per group [134, 152, 157, 176, 223, 249], while others especially measure the system's accuracy on the less active users [251]. Moody and Glass [157] provide a framework for user-centered evaluation that divides users into quadrants based on 2 dimensions: their near-neighbor-based similarity and their activity level. They find the algorithm to perform the worst (in terms of recall) for users who have rated many books but are dissimilar to other users and best for users who have rated few books but have high similarity with other users. The other criteria (coverage and diversity) are not evaluated on the book dataset. The previously mentioned item response theory-inspired method developed by Liu et al. [134] is also used to assess the relationship between *user difficulty* and user activity level, but such a relationship is not found for the book dataset, which is attributed to sparsity. Yang et al. [249] claim that measuring fairness as similarity of average performance between groups is not sufficient because the variance might be high. Instead, they define a new type of fairness: similarity between the performance distributions across different user groups. Finally, some studies only note the indirect impact of inactive users on C-fairness [177, 244].

Mitigation The aforementioned fairness reproduction study by Rahmani et al. [176] also estimates the impact of grouping the users based on activity level on a previously published re-ranking method's success in improving $\%\Delta GAP$, and the results differ from the popularity affinity grouping. They also measure the difference of *NDCG* between user groups to evaluate the divergence of performance for different popularity affinity levels and find that the higher number of average interactions per user can lead to better performance of the fairness model. Yang et al. [249] pursue the goal of enacting distributional fairness via a generative adversarial training framework called *DistFair*, and theoretically and experimentally show its effectiveness. Their framework achieves the best fairness performance for active versus non-active users in most cases, while maintaining a small performance drop. Yang et al. [251] concern themselves with cross-domain recommendation (CDR) and pose that inconsistency in user activity leads to user embeddings that prioritize users with more frequent interactions. They propose *UCLR*, a user-aware contrastive learning framework for robust CDR, and report significant performance improvement for users with only one or two interactions in Amazon-books compared to the baselines. *KGPL*, the method developed by Togashi et al. [223] that uses a knowledge graph with side information about the books to generate unobserved positive instances, is found to perform better for cold start users than other KG-based methods. However, all KG methods struggle with the book dataset, which they attribute to high sparsity. The previously mentioned Bias Detection Tree developed by Misztal-Radecka and Indurkhya [152] detects activity level as the user feature that correlates the most with bias. Their model selection techniques also improve performance for less active users. Notably, these discrimination rules are different between the training and test sets. For instance, the least active users have the smallest training error and highest test error, which is probably an indication of overfitting in the training set. Rahmani et al. [177] evaluate *CP-FairRank* with user

activity as the user-fairness concern, but there are limited results in the paper. Finally, *Ada2Fair* by Xu et al. [244] also boosts inactive users, which results in Pareto optimality over the baselines considering both accuracy and fairness.

Synthesis All papers about statistical bias from a user perspective group the users either on (a) level of activity [157, 249, 251], (b) affinity for popular items [58, 121, 159, 165, 256], or (c) both [176]. Moody and Glass [157] additionally group the users based on near-neighbor-based similarity. Bias towards users with niche tastes is sometimes characterized as unfairness [58, 121, 159]. To measure user bias, some papers use the $\%\Delta GAP$ metric that expresses an increase in average popularity [58, 159, 176], while others compare the performance across user groups [121, 157, 165, 176, 251, 256]. Yang et al. [249] extend the notion of across-group fairness to account for differences in distribution.

2.6.3 Joint

Some of the papers discussed measure or mitigate statistical bias from a user and item perspective simultaneously. They either tackle popularity bias and activity level bias [134, 152, 223, 244], popularity bias and popularity affinity bias [58, 121, 159, 165], or both [177].

2.7 Social bias in book recommendation

In this section, we discuss papers that concern themselves with social bias in book recommendation, either from a user or item perspective. Table 2.4 includes a summary of the different types of social bias along with their different views and metrics.

2.7.1 Item

Author nationality bias

Author nationality bias occurs when authors of certain nationalities are discriminated against in the recommendation process. This can mean that they are under-recommended in comparison to their presence in the users' profiles or in comparison with other nationalities [57]. This type of bias has only been studied by Daniil et al. [57], who investigate how popularity bias translates to social bias. They observe that since popularity is not a random phenomenon, it might be that popularity bias leads to other biases that are not obvious when additional user or item information is not available. They enrich a book dataset with author information using publicly available external resources, apply a recommendation process using various collaborative filtering algorithms, and compare popularity bias with bias towards author nationality. They find that popular books in the dataset are mainly written by US-based authors, and hence US-authored books tend to be over-recommended, especially so by algorithms vulnerable to popularity bias.

Author gender bias

Author gender bias can similarly be defined as discrimination against an author's gender in the recommendation process, either via under-recommendation compared to

Bias Type	View / Description	Metrics
	Item Biases	
Author nationality bias	Under- or over-representation of author nationality groups relative to the training set [57].	frequency per group [57]
Author gender bias	Under- or over-representation of author gender groups relative to the training set [68, 117, 178, 179, 199].	gender bias [68], log-bias [199], AWRF [178, 179], $\Delta_{\text{diff}}, \Delta_{\text{abs}}, \Delta_{\text{sq}}, \Delta_{\text{KL}}$ [117], EUR, RUR, IAA, EEL [178]
	Under- or over-representation of author gender groups relative to other gender groups [68, 72, 178, 199].	gender bias [68], DP [178], BDV [72], log-bias [199], FAIR, EED, DP [178]
Publisher location bias	Under- or over-representation of publisher location groups relative to other groups [90].	frequency per group [90], exposure per group [90]
	User Biases	
User age bias	Recommendation performance varies across user age groups [61, 133, 240, 260].	MAD NDCG [61], MAD Precision [61], GF MAE [133, 240, 260], GF MSE [260]
	Recommendation outcomes differ depending on user age [133].	saACC [133]
	Preferences of certain user age groups dominate overall recommendations [130].	DIF [130]
User gender bias	Recommendation performance varies across user gender groups [61].	MAD NDCG [61], MAD precision [61]
	Preferences of certain user gender groups dominate overall recommendations [130].	DIF [130]

Table 2.4: Social bias views and metrics grouped into item/user categories.

the user profiles [117, 179], compared to other authors of different gender [72], or both [68, 178, 199].

Measurement Raj and Ekstrand [179] measures attention-weighted rank fairness (*AWRF*) which corresponds to the difference between a group's (in this case, gender of authors) exposure in the recommended lists compared to a target distribution, which they set to the gender distribution in the training ratings. Kırnap et al. [117] point to the problem of missing information when measuring bias: they observe that fair ranking metrics normally require knowledge of membership of items in groups that relate to a protected attribute, even though this information is rarely available or complete. To account for that issue, they propose a sampling strategy and estimation technique for fair ranking metrics by formulating an estimator that works well even with a limited number of labeled items. They apply their method to estimate a set of fair ranking metrics that aim for gender representation in the recommendation equal to the entire corpus ($\Delta_{\text{diff}}, \Delta_{\text{abs}}, \Delta_{\text{sq}}, \Delta_{\text{KL}}$) and the exposure of female authors ($\widehat{E_g}$), and their estimation outperforms naive baselines. Other metrics compare the presence of male and female authors in the recommendation. Eleftherakis et al. [72] propose Bilateral Disparate Visibility (*BDV*), which represents the difference between the observed probability of recommending books written by one gender and that of recommending books written by another gender.

Raj and Ekstrand [178] perform the first empirical comparative analysis between fair ranking metrics using various datasets, including author-gender fairness assessment on Goodreads. Some of the metrics proposed only measure whether the genders are equally represented in the ranking(s) (*FAIR*, *DP*, *EED*), while others account for utility (*EUR*, *RUR*, *IAA*, *EEL*) or gender distribution in the book dataset (*AWRF*). They also perform sensitivity analysis to assess the impact of certain design choices. They generally find that metrics disagree on orderings of different algorithms' fairness, while they might agree for some experiments and not others. For Goodreads, they find that parameters such as the ranked list size and weighting strategy impact whether the system is considered fair towards authors of different genders. In an influential paper, Ekstrand and Kluver [68] apply a Bayesian model on user profiles and recommended lists (to account for size variances) and calculate the gender tendency. They find that, in Amazon-books, Book-Crossing, and Goodreads, female authors are less underrepresented in user-book interactions than in the unique books, particularly among the popular books. They compare the profile gender tendency with that in the recommended lists of a set of algorithms trained on implicit and explicit versions of the datasets and find that average balances are comparable to user profile average balances, but the algorithms differ when it comes to the distribution. Additionally, they find that explicit-feedback algorithms are much less responsive to their users' input profiles in terms of author gender distribution, likely due to the fact that they account for rating additionally to consumption patterns. Saxena and Jain [199] measure log-bias, which expresses the extent to which users tend to rate books written by one gender higher compared to another gender. They report the presence of such bias in Book-Crossing and Amazon-books in favor of male authors and compare dataset log-bias with recommendation log-bias.

Mitigation Some papers propose bias mitigation strategies to reduce differences in author gender distribution in the recommendations. Ekstrand and Kluver [68] apply simple re-ranking strategies on the recommended lists to meet gender balance goals. They find that their strategies are successful in bringing the gender distribution to about 50-50 with minimal accuracy loss. Saxena and Jain [199] address bias towards author gender given explicit feedback. They propose a model that quantifies gender bias in book ratings and subsequent recommendations and, accordingly, a model-agnostic approach to provide unbiased recommendation. Additionally, they incorporate user preference correction to maintain good model accuracy in rating and ranking prediction. They use estimated log-bias to measure whether the predicted ratings of male authors are higher than those of female authors. Their method achieves significant bias reduction with insignificant accuracy loss. Eleftherakis et al. [72] note that a source of demographic bias in neighbor-based collaborative filtering is that the neighbors may foster discriminatory patterns, and that some work that addresses this issue is of low efficiency and impacts accuracy. They propose two new algorithms that tackle efficiency and accuracy. They measure *BDV* on the recommendations based on the *young adult* subset of Goodreads (in this case male-written books are viewed as protected). Their method scales better compared to other fairness-aware methods, especially for Goodreads. Their method's mitigation impact is not as high for Goodreads than it is for the non-book dataset, but the imbalance (in favor of female authors) is still reduced. Raj and Ekstrand [179] attempt to reduce the difference in gender distribution between recommendation and training set, in the context of a grid-view recommendation. They develop a grid-aware re-ranking algorithm for provider-side fairness, which adopts existing techniques by taking into account grid features, and test it on a book dataset. They find that *AWRF* is improved when the re-ranking method is grid-aware, and the results depend on the browsing model, device size, and column size.

Publisher location bias

Gómez et al. [90] tackle **publisher location bias** based on the continent where the book was published, prompted by the geographic imbalance in the data. They measure disparate *visibility* and *exposure* for each continent-based group given a set of algorithms and propose a re-ranking approach that addresses both. They find that 90% of the books were produced in North America, with these books accounting for 93% of the ratings (similar to Daniil et al. [57] who noted the over-representation of American authors), and that this bias is propagated in the recommendation, but their mitigation method performs well.

Synthesis Most papers about social bias from an item perspective study bias towards sensitive author characteristics, namely author gender [68, 72, 117, 178, 179, 199] and author nationality [57], while Gómez et al. [90] are concerned with publisher location. Some papers note the challenge of obtaining such information on the items [57, 117], while Saxena and Jain [199] use Genderize.io to estimate author gender. To assess bias towards a group of items, most papers measure how the algorithms treat disadvantaged versus advantaged groups, either by means of representation in

recommendation [57, 68, 72, 90, 117, 178, 179] or predicted rating [199]. Unlike the other papers on author gender bias, Eleftherakis et al. [72] consider male authors as the disadvantaged group.

2.7.2 User

User gender

User gender bias occurs when the system underserves users based on their gender, either by performing worse for users of a given gender [61], or by mostly recommending items that a single group of users prefers [130].

Measurement User gender bias based on the divergent performance view has been measured by mean average deviation of performance for different gender groups by Deldjoo et al. [61]. They systematically track data characteristics that may impact fairness (and accuracy) of collaborative filtering algorithms via regression-based explanatory modeling. User gender has small explanatory power for fairness on the book dataset. To measure whether the recommended items are fairly interacted with by users of different genders in their profiles, Li et al. [130] propose a new metric called Interaction Fairness (*IF*), for which they measure the difference between user profile and recommendation.

Mitigation Li et al. [130] specifically look into fairness in sequential recommendation. They define a fairness-aware sequential recommendation task and propose a multi-task deep model, *FairSR*, which consistently produces the highest improvement in *IF* for different user gender groups while maintaining or even improving on performance metrics.

User age

User age bias occurs when the system underserves users based on their age, either by performing worse for users of given age [61, 133, 240, 260], by recommending differently to different-aged users [133], or by mostly recommending items that a single group of users prefers [130].

Measurement Deldjoo et al. [61] measure the explanatory power of user age on fairness defined as mean average deviation of performance for different age groups, but similarly to user gender, they find small explanatory power on the book dataset. Other studies also measure deviation of performance metrics between user age groups, namely MAE [133, 240, 260] and MSE [260]. To estimate the impact of user age on the recommendation, Liu et al. [133] propose a new metric named *saAAC* (sensitive attribute accuracy), based on the items in the recommended list and the latent relationship between the content attributes and the users' sensitive attributes. Finally, Li et al. [130] measure the difference in *IF* for different user age groups.

Mitigation Papers that mitigate bias towards user age in book recommendation by improving performance fairness are as follows. Wei and He [240] design a framework

(*CLOVER*) that ensures fairness of meta-learned recommendation models (individual, counterfactual, and group fairness) and try to satisfy all three via a multi-task adversarial learning scheme. They find that meta-learning for cold start is more unfair than other cold start alleviation methods, and with *CLOVER* it becomes fairer towards users of different age while not sacrificing performance. Zhang et al. [260] are also concerned with unfairness in meta learning, however they're specifically interested in what is the critical factor that leads to unfairness. Through theoretical analysis, they show that the critical factor is group proportion imbalance. Then they propose a fairness-aware adaptive sampling framework of meta learning (*FAST*) which adjusts the sampling distribution for different user age groups during training, and theoretically prove convergence. Based on group fairness metrics among users of different age, *FAST* generally achieves more significant fairness improvements than other fairness methods (including *CLOVER*) and comparable utility across various meta-learning paradigms. In a different approach, Liu et al. [133] attempt to filter sensitive information so that it is not learned when constructing representations by using *DALFRec*, a fairness-aware algorithm that uses adversarial learning to mitigate the effect of sensitive information on both user and item side. They assess the algorithm's ability to reduce discrimination with *saAAC* (sensitive attribute accuracy) and additional performance fairness metrics. They measure relation between user age and item publication year and author. When it comes to user age, they achieve better fairness over traditional recommendations and other fairness methods. They also note that on the book dataset, the *saACC* of all algorithms is high, which indicates that the item distribution in the recommendations is highly correlated with users' age. Finally, the aforementioned multi-task deep model *FairSR* developed by Li et al. [130] also improves *IF* for different user age groups in a sequential recommendation task.

Other

To detect bias towards users, especially because minority preferences are discriminated against, Misztal-Radecka and Indurkhya [151] design a bias-aware hierarchical clustering algorithm that clusters users to find ones whose needs are not met by a black box recommender system. On top, they apply a post-hoc explainer that describes the most important features of these users. Among other results, they note that (NDCG) performance is lower for discriminated groups of smaller size than for bigger groups. However, the discriminated user features are not presented in this chapter for the book dataset.

Synthesis Most papers about social bias from a user perspective are concerned with either user age [133, 240, 260], or both user age and gender [61, 130]. Misztal-Radecka and Indurkhya [151] test user age and location, but do not disclose which user features were found to be prone to discrimination in the book dataset. To assess bias, some studies measure difference in performance for different user groups (i.e., group fairness) [61, 151, 260]. Relevant and case-specific metrics are also proposed [130, 133].

2.7.3 Joint

None of the papers surveyed deal with social bias towards items and users simultaneously.

2.8 Datasets and reproducibility in book recommendation

In this section, we take a closer look into the datasets that researchers use to evaluate bias. Table 2.5 includes all datasets along with the features that researchers have evaluated bias towards, as well as basic information on them. These datasets stem from three sources: Amazon, Book-Crossing and Goodreads. Book-Crossing is widely used despite having been created more than 20 years ago. All the datasets are very sparse, which often prompts researchers to preprocess them before use, for example, by performing k-core filtering (i.e., removing interactions until all users have interacted with at least k books and all books have been interacted with by at least k users). This practice raises questions on whether the conclusions of each study hold outside the limits of their experimentation or can be directly compared with the conclusions of a different study on the same dataset.

Additionally, the datasets include very little information on the users and authors. Specifically, bias towards user characteristics has only been studied on Book-Crossing. Author gender has been studied on multiple datasets, but the enrichment is not necessarily consistent, as various sources have been used to discern gender, namely VIAF [68], Wikidata [72] and Genderize.io [199]. Notably, outside of author features, the only social book characteristic that has been considered is publisher location.

Table 2.6 shows for each dataset which types of bias have been measured, as well as based on which metrics. Even though some papers have studied the same type of bias on a dataset, they mostly use different metrics. Therefore, it is mostly not possible to directly compare the studies unless they explicitly use each other's methods as one of the baseline comparisons. We will discuss cases where the same metrics have been used across studies.

Popularity bias on Book-Crossing Daniil et al. [58], Kowald and Lacic [121], Naghiaei et al. [159] measure popularity-recommendation frequency correlation based on algorithms trained on Book-Crossing. Daniil et al. [58], Naghiaei et al. [159] measure $\%\Delta GAP$, yet these studies are explicitly reproductions of each other. Daniil et al. [58] conclude that the differences in evaluation strategies among the studies is the driver behind the divergence in conclusions, despite the use of the same metric and dataset.

Additionally, Balasubramanian et al. [24], Wei et al. [239], Zhang et al. [258, 259] measure long-tail HR and long-tail NDCG on Book-Crossing. These studies use each other's methods as baselines in chronological order, as they all attempt to augment the information space of the long-tail items to empower them.

User age bias on Book-Crossing Liu et al. [133], Wei and He [240], Zhang et al. [260] measure MAE-based group fairness with user age being the sensitive characteristic on Book-Crossing, and they are all concerned with meta-learning. In this case, Liu

Dataset	Bias-related user features	Bias-related item features	Users	Items	Interactions	Source
Amazon-books	activity [134, 251], popularity affinity [165]	popularity [134, 165, 221], author gender* [68, 199]	8,026,324	2,330,066	22,507,155	Collected in 2018 from Amazon [163]
Book-Crossing	activity [152, 157, 176, 177, 223, 244, 249], popularity affinity [58, 121, 159, 176, 256], user age [61, 130, 133, 240, 260], user gender [61, 130]	popularity [24, 58, 82, 121, 152, 159, 177, 223, 237, 239, 244, 258, 259], author nationality* [57], author gender* [68, 199], publisher location* [90]	278,858	271,379	1,149,780	Collected in 2004 from the Book Crossing community [34]
GoodReads		popularity [186], author gender* [68, 72, 117, 178, 179]	876,145	2,360,655	228,648,342	Collected in 2017 from Goodreads [234, 235]
Goodbooks		blockbusterness [247]	53,424	10,000	5,976,479	Collected in 2017 from Goodreads [254]

Table 2.5: Bias-related features and properties of four book recommendation datasets. * indicates attributes enriched by the researchers.

et al. [133] and Zhang et al. [260] compare their respective frameworks with *CLOVER*, the framework proposed by Wei and He [240] as a baseline. However, the two studies do not report the exact same numbers for *CLOVER*'s GF MAE (presumably due to experimental design differences), so comparing between them directly is not feasible.

Activity level bias on Book-Crossing Multiple studies measure or mitigate activity level bias on Book-Crossing by comparing algorithm performance on groups of users with different activity levels [157, 176, 249, 259]. However, among other differences, the studies divide the users into groups differently between them: Rahmani et al. [176] compare the top 5% most active users with the rest, Yang et al. [249] compare the top 20% most active users with the rest, Zhang et al. [259] rank the users based on activity level and divide them into 4 parts, and Moody and Glass [157] compare the users with activity level above the median with the rest. We cannot conclude which mitigation method ([176, 249, 259]) is more successful in addressing activity level bias.

Publicly available datasets with user-item interactions are invaluable for recommender systems research, as they facilitate comparison across methods and support reproducibility. However, our analysis shows that, in the context of bias in book recommendation, their potential to support cumulative knowledge is currently limited. Even when the same dataset and the same bias type are studied, differences in preprocessing, user or item grouping strategies, metric choice, and evaluation protocols often prevent direct comparison across studies. As 4 out of 10 studies included in this survey do not grant access to a public repository with their code and experiments (see Figure 8.2b in the Appendix), addressing these challenges requires more explicit reporting of dataset preprocessing, clearer operationalization of bias-related concepts, and greater transparency around experimental setups.

Dataset	Bias type	Measurement/Metric	Mitigation/Metric
Amazon-books	Popularity	popularity-difficulty correlation [134]	recall per group [165], novelty [165], coverage [221], LT frequency [221], LT coverage [221]
	Activity level	activity-difficulty correlation [134]	IA MRR [251], IA HR [251], IA NDCG [251]
	Popularity affinity		recall per group [165]
	Author gender		gender tendency [68], log-bias [199]
Book-Crossing	Popularity	correlation [58, 121, 159]	estimation error per group [152], P-fairness [244], coverage [223], LT precision [82], LT recall [82], NDCG [237], MRR [237], recall [237], LT HR [24, 239, 258, 259], LT NDCG [24, 239, 258, 259]
	Activity level	recall per group [157]	NDCG per group [176, 249], recall per group [223, 249], %Δ*GAP* [176], estimation error per group [152], C-fairness [244]
	Popularity affinity	%Δ*GAP* [58, 159], MAE per group [121], NDCG per group [256]	NDCG per group [176], %Δ*GAP* [176]
	Author nationality	frequency per group [57]	
	Author gender		gender tendency [68], log bias [199]
	Publisher location		frequency per group [90], exposure per group [90]
	User age	MAD NDCG [61], MAD precision [61]	DIF [130], GF MAE [133, 240, 260]
	User gender	MAD NDCG [61], MAD precision [61]	DIF [130]
GoodReads	Popularity		PopQ@1 [186]
	Author gender	DP [178], EED [178], frequency per group [117], exposure per group [117]	gender tendency [68], AWRF [178, 179], BDV [72], FAIR, EED, DP, EUR, RUR, IAA, EEL [178], $\Delta_{\text{diff}}, \Delta_{\text{abs}}, \Delta_{\text{sq}}, \Delta_{\text{KL}}$ [117]
Goodbooks	Blockbuster	frequency per group [247]	

Table 2.6: Bias types, metrics, and corresponding paper references per dataset.

2.9 Discussion

2.9.1 General observations

There has been relatively little research specifically addressing bias in book recommender systems compared to the extensive work done on bias in other domains, for example, music recommendation. At first glance, the limited research is surprising given that bias is a well-known and documented issue in the book market in general [129]. One explanation is that book reading is considered a relatively niche habit compared to other media, and there hasn't been as strong an emphasis on personalization for the sake of increasing consumption. In recommender systems, research is often driven by streaming companies. Users consume music, series, and movies faster, which might inherently require further automation and personalization and hence lead to bias propagation.

As mentioned before, most of the papers that we include in this survey study statistical bias, especially popularity bias. There exists varied work on popularity bias and its effect on users and items, with a notable emphasis on mitigation while also ensuring sufficient accuracy. Given the fact that the book market is often assumed to follow a long-tail distribution, this is not surprising. This observation aligns with Dinnissen and Bauer [63] who found popularity bias to be a prevalent topic in fair music recommendation literature (see Section 2.3).

Another observation is the ubiquity of incidental and somewhat superficial studies on bias in book recommender systems. Many of these studies seem to treat bias as a generic machine learning-related issue without critically engaging with the book domain. Researchers often use book datasets simply because they are available rather than because they are concerned with the specific implications of bias in literary recommendation. In most of these papers, the book dataset was one of many. Book recommendation results are sometimes treated as an afterthought: in some cases, book-related results are hidden in appendices, while in others, they are omitted altogether. This raises questions about why book recommendation research is sidelined. We hypothesize that due to the book datasets' sparsity, the results are sometimes too inconclusive to be included in the main body of the paper and discussed in detail. As a result, the field lacks focus and continuity. It also becomes harder for researchers, especially outside of the field of recommender systems, to find material on bias in book recommendation and draw conclusions.

To add to that, much of the existing research relies on the same few datasets, primarily American ones, which fails to account for global reading habits and cultural differences. Additionally, book recommendation studies often include very few social characteristics, making it difficult to explore biases related to other relevant characteristics, both for users and books. For users, the only social characteristics tested are gender and age, presumably because these are the ones available in the datasets. For books, a few more characteristics are tested; however, the majority of papers are concerned with author gender. In some cases, the studies use Linked Open Data to obtain this information, and gender might be easier to discern than other sensitive characteristics. This observation points to a well-known issue in quantitative research: sometimes we measure what is available rather than what is important.

2.9.2 Gaps & future directions

One gap in the mitigation of bias in book recommendation is effective comparison between the various methods developed, which would determine which is the 'best' among them. While most of the mitigation papers compare their method to some baselines, the emphasis is on the method rather than the effect on the book dataset result. As discussed in Section 2.8, the studies included in this survey mostly use different metrics and evaluation protocols, even when measuring the same type of bias based on the same dataset. An extensive comparative study that tests the different mitigation methods for a set of book datasets and given different bias metrics would shed light on what the state-of-the-art is when it comes to the book domain specifically.

User-book interaction datasets, as they are now, are limited. There is much progress to be made on improving them, in the sense of enriching them with publicly available information. Ekstrand and Kluver [68] take advantage of Linked Open Data sources to create a valuable dataset that includes author gender, which is used by some of the other studies that we survey. Daniil et al. [57] follow a similar approach to supplement a smaller dataset with author country of citizenship. While publicly available information is not complete and is known to be biased in its own right, it would be a step towards the right direction. At the same time, book themes as a bias subject have barely been studied. Extracting information on the themes of a book by textually analyzing the description might be even more accessible than information on the authors' personal characteristics. Future research on bias in book recommendation could explore this avenue.

2.9.3 Recommendations for interdisciplinarity

The field of bias in book recommendation would benefit from connection between theoretical background and practical studies. There is insight to be gained from interdisciplinary research, especially when interacting with a field as rich as LIS. Recommender systems studies can rely on existing LIS frameworks that describe just and inclusive library services [107, 137, 225, 252] and design metrics for bias and inclusivity for each component of the book ecosystem accordingly. Grounding on theory can assist with interdisciplinary communication and ensure overall progress.

A crucial step that can support the aforementioned goal is engagement with real library or commercial literary recommenders. In simulations, researchers are limited to making assumptions regarding both how a recommendation is done and also what is a relevant type of bias in the given use case. Practitioners in the book domain have a more sophisticated understanding of what bias entails and can support researchers in their pursuit. Additionally, since bias mitigation for many studies means to simply reduce bias compared to a baseline, in an actual use case stakeholders can give input on what an 'ideal' system would look like (e.g., appropriate level of representation for every group of authors) [231]. Context specificity is crucial here; book reading is very much dependent on the cultural context, even today that many of us exist in the same online world. Books are still produced and recommended in decentralized ways. By only studying the same few American datasets, researchers are restricted to a one-dimensional view of bias and can only make narrow conclusions.

The differences in terminologies and research methods used by LIS and computer

science professionals are a significant hurdle in creating fair and effective recommendation systems. The lack of integration between these fields hampers the development of systems that align with the core values of diversity, accessibility, and fairness, which libraries uphold. An essential step forward is greater collaboration between LIS professionals and computer scientists to establish shared terminology and research practices. Interdisciplinary projects can support the design of recommender systems that take bias into account in a targeted and situated way. Without such intentional safeguards, these systems risk perpetuating existing biases and further marginalizing underrepresented voices.

2.10 Conclusion

In this chapter, we conducted a literature review on bias in book recommender systems across computer science venues. We gathered and analyzed 40 papers from various perspectives. We found that, despite the prevalence of book datasets with ratings in recommender systems research, studies often focus more on the 'bias' part rather than the 'book' part. Accordingly, we proposed a set of future directions based on gaps in literature, as well as recommendations for interdisciplinary practices. We hope that by collecting and mapping recent research, this chapter can serve as a stepping stone for researchers, computer scientists or otherwise, who wish to contribute to the field of bias in book recommendation.

3

The perspective of public media service practitioners on diversity in recommender systems

While Chapter 2 maps the computational research landscape, it is equally important to understand how norms and values are operationalized in practice. In this chapter, we explore how practitioners at public service organizations, including a library, conceptualize and enact diversity in their recommender systems.

Diversity is a commonly known principle in the design of recommender systems, but also ambiguous in its conceptualization. Through semi-structured interviews we explore how practitioners at three different public service media organizations in the Netherlands conceptualize diversity within the scope of their recommender systems. We provide an overview of the goals that they have with diversity in their systems, which aspects are relevant, and how recommendations should be diversified. We show that even within this limited domain, conceptualization of diversity greatly varies, and argue that it is unlikely that a standardized conceptualization will be achieved. Instead, we should focus on effective communication of what diversity in this particular system means, thus allowing for operationalizations of diversity that are capable of expressing the nuances and requirements of that particular domain.

This chapter is based on *Vrijenhoek et al., Diversity of What? On the Different Conceptualizations of Diversity in Recommender Systems, FAccT '24* [231].

3.1 Introduction

The concept of diversity is at the same time omnipresent and very ambiguous [73]. In popular discourse, diversity usually refers to the variation of human characteristics, often related to a notion of identity politics [29]; in biological research, as a qualifier to the health of an ecosystem [227]; and in media studies, as a concept adjacent to pluralism, expressing whether a news selection contains a plurality of sources, voices and opinions [114]. While all interpretations are valid and intuitively seem to reflect similar concepts, they differ in their operationalization in a way that is unique to their domain. At the same time, diversity is consistently noted as a social value that is beneficial to pursue, though in different capacities.

In the recommender systems domain accounting for the diversity of a recommendation can help avoid monotony [38, 257]. This follows from the assumption that a recommender system that is purely optimized on predicted relevance will result in a feedback loop and thus prioritize similar content, leading to 'more of the same'. There is as such also a business case to be made for diversity. While it has been challenging to show empirically, diversity may lead to higher user satisfaction and retention, increasing revenue in the process [69, 108]. This argument can be extended for the news domain, where worries over filter bubbles that reinforce existing beliefs by only showing content that aligns with a user's preferences are especially prevalent [148, 267]. Having a news recommender system that pursues diversity could help expose users to things different than they are used to or expect seeing, and tailor to their specific information needs [96, 97]. For machine learning in general, it is also important to account for diversity in a social context. Since many algorithms are trained on datasets that are not representative of all groups in society, not accounting for diversity may further "amplify stereotypes, alienate users, and further entrench rigid social expectations" [153].

The many different interpretations of diversity are a fundamental challenge to the practical development of recommender systems. Evaluating the performance of a recommender system requires objectively measurable properties, and the field strives to do so in a standardized manner [263]. Standardization allows for comparability and reproducibility of results and algorithms[1], and is achieved through, among others, the use of benchmark datasets, sampling techniques and evaluation metrics such as mean reciprocal rank (MRR) or normalized discounted cumulative gain (NDCG) [255], but has not yet been achieved for diversity. Loecherbach et al. [136], who did a literature review of existing measures of diversity in media studies, note that "there is little to no overlap between concepts and operationalizations used in the different fields interested in media diversity". Lawrence [126], executing a conceptual analysis of the term 'diverse books', notes that the topic is inherently political, and that "we have yet to arrive at a clear consensus on what the modifier 'diverse' means in this and other instances".

[1] Recent strands of research have noted that this perceived notion of standardization is often false. Even with commonly accepted principles there are often significant differences in implementation, and thus output [204, 220]

The conceptual unclarity around diversity makes that not only may we implement it differently, we may also *mean* something completely different when we use the term. One could argue that diversity is, in fact, an essentially contested concept [49, 83], meaning that it is open for discussion and debate and that we are unlikely to reach consensus on its meaning. As such, striving for agreement or a clear definition may lead to a standstill and hinder progress, as it is unclear what a good operationalization or implementation would look like. It may cause the conceptualization of diversity to be driven by the information that is available or more easily suitable for quantification, rather than what is desirable; furthermore, blindly trusting 'objective' measures "may obscure the fact that the conceptualization of diversity is, eventually, a normative choice" [115]. Instead, what we need may be a more flexible operationalization that is capable of reflecting the nuances and requirements of the domain it is deployed in.

The aim of this chapter is to explore the dimensions in which industry practitioners within a limited domain conceptualize diversity. To this end we conduct a series of semi-structured interviews with practitioners from three different public service media organizations in the Netherlands: a broadcaster, a news organization, and a library. The choice of these organizations is deliberate; though the medium through which they do it differs, they all play an active and important role in the dissemination and curation of information and ideas. As such, we expect there to be a fairly established conceptualization of diversity within the organization, and comparatively more overlap between them than with, for example, a commercial music recommendation service.

Through analysis of the interviews, and guided by the 'components' of diversity formalization defined in Vrijenhoek et al. [230], we aim to answer the following questions: what *goals* do the organizations aim to achieve with the recommendations, which *aspects* do they consider relevant for diversity, and through which *tactics* are the recommendations diversified. We find that even within the limited domain of public service media recommendation, there is a wide variety of conceptualizations of diversity. With this, we underline that a standardized definition of diversity in recommender systems is likely not achievable. However, this empirical categorisation, albeit non-exhaustive, can assist in the process of conceptualizing and implementing diversity in a particular concrete context.

3.2 Related work

Recommender systems have long moved from evaluating recommendations solely based on accuracy-related metrics. Other metrics such as novelty, serendipity and diversity have become a common part of evaluation practices [38]. Diversity in recommendation is a current topic, and multiple diversification methods have been proposed in recent years [55, 124]. In recommender systems research, diversity is often viewed as the opposite of similarity; a list of recommended items is diverse if the items are sufficiently different between them along a set of axes [265]. However, Jesse et al. [111] point to a gap between human perceptions on diversity and intra-list similarity (ILS) commonly used to assess diversity in offline recommender systems experiments. Through a user study, they find that while ILS can be a good proxy, the details of the implementation matter and require validation in a given domain and application.

When diversity is seen as a social value that needs to be incorporated in a system, standardizing its definition and operationalization becomes even more challenging. Mitchell et al. [153] define diversity in a subset selection task (e.g., the construction of a list of items to recommend) as "variety in the representation of individuals in an instance or set of instances, with respect to sociopolitical power differentials (gender, race, etc.". As such, they view diversity as a concept with inherently sociopolitical connotations in contrast to the more general term 'heterogeneity'.

In the context of media recommendations, diversity is often closely associated with pluralism. According to Karppinen [115], the interpretation of pluralism and diversity in media is dependent on the political and normative understanding of the role of media in society. Additionally, defining media diversity is further complicated by its often contradictory aspects and interpretations. Van Cuilenburg [226] point out an antithesis between the normative media diversity frameworks of *reflection* and *openness*; one requires content distribution to proportionally reflect societal distributions of relevance, while the other corresponds to perfectly equal representation and attention to all people and ideas in society. Similar contradictions have been laid out in the domain of library and information studies [126]. Based on a systematic literature review on media diversity, Loecherbach et al. [136], also referenced in the introduction on the little overlap between different fields' conceptualization of diversity, conclude that relevant research should be guided by an interdisciplinary effort to define and benchmark the concept of media diversity, along with its different sub-dimensions. Other work studying diversity in recommender systems would typically acknowledge the complexity of the concept, yet model it following existing technical standards; either as a distance to a user's reading history [92], political leaning [95] or as pair-wise distance between the items in a recommendation, for example in topics, categories and/or tone [154]. In this work, we aim to contribute to this effort by attempting to map the dimensions of diversity in media recommender systems.

Diversity in Public Service Media

Diversity in recommendation is especially relevant for public service media (PSM), whose role and societal responsibilities call for a careful consideration of the content they produce and give exposure to. Many media organizations underline the potential of recommender systems and the importance of reflecting editorial values such as diversity therein [33, 85, 123, 155]. PSM are often required to offer diverse content, which might be at odds with the primarily commercial use of recommender systems that mainly intend to increase media consumption, and therefore, profit [99]. Even if we assume societal agreement on the importance of providing diverse content for PSM, the practical implications are harder to lay out. Helberger [96] outlines four models for the normative conceptualization of diversity in recommender systems that serve different purposes: the liberal, the deliberative, the participatory, and the critical. While each of the models is in its own way relevant for PSM, the work suggests that the critical perspective, which focuses on the visibility of minority voices that are often disadvantaged in public platforms, is less likely to be encountered in commercial applications, and therefore could be partly served by PSM. Regardless, few have succeeded in concrete implementations that reflect diversity as a normative value, in

part due to the gap between journalistic values and recommender system evaluation metrics [229]. Møller [156] suggests that "the abstract nature of journalistic values make them hard to account for computationally", and that "[h]uman journalists have an important role to play in these processes not only to help conceptualise the values themselves but also as part of new algorithmic news practices." Translating values into concrete algorithmic practices is thus as challenging as it is necessary.

To bridge the gap between theory and practice, the perspectives of practitioners in a given domain can assist, orient, and ground research. On that note, researchers have conducted interviews with practitioners on the interaction between emerging algorithmic systems and norms and values. Sørensen [213] interviewed developers, data scientists, and project leaders from nine European PSM organizations on the topic of implementing recommender systems. They report that, while interviewees believe diversity to be an essential aspect of their catalogue, they are reluctant to depend on an algorithmic implementation instead of the traditional editorial control. Additionally, they attribute the general lack of formal definition of diversity to the different understanding between politicians, practitioners, and users. Bastian et al. [25] interviewed practitioners from different departments of two newspapers regarding the impact of algorithmic news recommenders on their organization's norms and mission, as well as how to integrate them in the design of news recommenders. They found that, while the interviewees attach varying degrees of importance to different values, diversity is perceived as one of the core values for news recommender design and implementation. During the interviews we conducted, we specifically focused on the conception and implementation of diversity, which allowed for an in-depth outline of the perspectives and practices of the interviewed professionals.

3.3 Method

For this study we conducted a series of semi-structured interviews with three public service media organizations in the Netherlands: a broadcaster, a news organization, and a library. Interviews were carried out in-person and on site in the offices of the interviewees, and took place over a span of four months, between December 2022 and March 2023. At this time each of these organizations were, at different stages of completion, (exploring the possibility of) developing a recommender system to effectively serve content to their users, which also means that decisions about incorporating diversity in the recommendations were actively being made.

3.3.1 Candidate Selection

Potential candidates were approached through a snowball sampling technique: after initial contacts with the organization were established, interviewees were asked for recommendations of colleagues to interview in the next round. We actively tried to find a set of participants that reflected the composition of the organisations and the relevant figures in the design, deployment and use of the recommender system. This process yielded twelve participants in total: six at the broadcaster, four at the library, and two at the news organization. Following Smets et al. [210], our goal was to have a good spread of different types of stakeholders, and sought participants with roles related to Business

(four participants), Curation (two participants), Product owner (three participants) and Technology developer (three participants). During the snowball sampling we would explicitly ask participants whether they knew of potential interviewees with a role we had not covered yet. Finding all different roles did not always succeed, and we acknowledge this as a limitation of our work. This is also why we are adamant about not making normative claims about the definition of diversity per domain, but rather present it as an exploratory study. Participants were predominantly male (9 out of 12) and white (11 out of 12). This is reflective of the workforce but a caveat for generalization, further outlined in Section 3.3.4. Disclosing too many details about individual participants and their roles would undermine the promised anonymity, and thus each participant is assigned a code reflecting only their organization.

3.3.2 Interview structure

The interviews consisted of four parts. In the first part, interviewees introduced themselves and their role within the organization. In the second part interviewees were asked about their general conceptions of diversity, and how these conceptions related to their organization. Here, the goal was to understand the mission of the organization in question, and the role diversity plays in that mission. In the third part interviewees described their knowledge of the recommender systems that were in use by their organization, either in planning or in production. This served as a check that participants were sufficiently informed to be included in the analysis, and to prime the interviewees for the fourth part, in which they were specifically asked about the role of diversity in the recommender systems of their respective organization. The aim here was to go beyond what currently exists, and instead focus on what the recommender system should look like in an ideal situation. This part of the interview also contained a small experiment, in which the interviewees were asked to take on the role of a recommender system and rank a set of items while keeping diversity in mind. During the experiment participants were asked to voice their thought process by thinking out loud [41]. Our primary interest was not the final recommendation generated, but which aspects of the items participants considered before (not) including an item. For each organization we prepared a set of 15 candidate items based on the organization's (potential) catalogue, and included the metadata a user would see when interacting with the system. We attempted to ensure (to the best of our abilities) that there would be enough diversity in this candidate list present. To cover our own blind spots we would ask participants at the end of the experiment whether there was a type of content that they were missing. We acknowledge that our own conception here steers the type of diversity that could potentially be found (see also Section 3.3.4).

For each of the organizations, the metadata always included the items' title, summary and a cover photo. For the library, we also included the authors' name and a set of keywords about the book; for the broadcaster, the title and description of the series the item was part of (if applicable); and for the news organization, the time of publication and the first few paragraphs of the news item. Based on this candidate set, we asked participants to make a diverse selection of 5 items. In order to not influence people's thought processes in what could potentially be relevant aspects of diversity, we deliberately left instructions vague, and did not include a profile of a

user to create a recommendation for in the instructions. Instead, we considered the participants' questions for clarification as part of what they deemed relevant for diverse recommendation.

3.3.3 Coding and analysis

We executed an inductive thematic analysis on each of the interviews, guided by the research questions posed in Section 3.1: what the organizations aimed to achieve with a (diverse) recommendation, which aspects of the items they would consider during recommendation, and which tactics they would employ to diversify the recommendation. The first two authors created a coding scheme based on half of the interviews, which after completion were discussed and merged. With the resulting coding scheme, the authors annotated the half of the interviews they had not previously seen, and extended the coding scheme when gaps were identified. Additionally, after the respective coding schemes were merged, the two authors independently coded the same interview and proceeded to compare their outputs. This step helped ensure practical consensus, as the authors were able to align on the specific nuanced interpretation of each of the codes, as well as how it applies in the context of the interviews. Based on the resulting coding scheme, we created a diversity 'map' of relevant aspects and how they relate to each other, which will be discussed in the next section. The coding scheme is included in Appendix 8.2. After the first draft of the chapter was finished, we reached out to all participants. We shared with them which quotes had been attributed to them, and asked them to reach out in case of misunderstandings.

3.3.4 Limitations

There are a number of important caveats to consider in light of this method and experiment. By presenting the participants with a list of items to recommend, we inadvertently steer the type of diversity they are likely to mention. For example, if our sample did not contain any mention of politics, the participants might also be less likely to mention this as a relevant aspect. Simultaneously, participants may be influenced by what is commonly considered a relevant aspect of diversity, or what interpretations are currently feasible rather than ideal. As such, we cannot draw conclusions on the importance of a certain aspect.

Another difficulty when running this experiment was accounting for the recency of items, which is extremely relevant for the news organization and the broadcaster, where the first will never want to recommend old news, but the latter may sometimes want to include older content. This was especially an issue given that interviews took place on different days, sometimes weeks apart. For the broadcaster we mixed a stable set of older content with content that was at the day of the interview popular in the system. This has the important caveat that the results between broadcaster interviewees are not fully comparable. For the news organization, we opted for a fixed date a few months in the past. While the interviews are comparable in this case, there is a risk that important contexts are forgotten or mixed up with current events.

Lastly, while diversity in and of itself is already a complicated and multi-faceted concept, the concepts related to it are too. People may talk about 'different ethnicities', and can mean, among others, different skin colors, nationalities or cultures. By

extension, our findings are largely influenced by the people that participated in the interviews. Organizations are not a monolith, and many of the interviewees were very explicit about not representing the opinion of every member of the organization they belong to. Furthermore, many of the responses are influenced by the background of the participants themselves, and the fact that this research was conducted in the Netherlands. The Netherlands has a strong public service media system with a focus on representing different groups in society [53]. That being said, participants were overwhelmingly white (11 out of 12) and male (9 out of 12). While this is representative of the general demographic working on recommender systems in the Netherlands, people are not as acutely aware of the needs of groups they are not, themselves, a part of [31, 145]. As such, this is a suitable group for a descriptive study ('what is') into the conceptualizations of diverse recommendations in public service media organisations; however, a complete normative formalization of diversity ('what should be') would require the participation of people from different backgrounds and perspectives in order to provide a complete overview. All limitations considered, the results of this study cannot be used as a definition of diversity. Rather, we aim to show that, even within this limited domain, 'diversity' indeed may refer to a wide variety of things.

3.4 Results

We expect that how organizations operationalize diversity is dependent on what they aim to achieve with their recommendation. We first identify this goal in the following section (Section 3.4.1). Then, we analyze both the different aspects (Section 3.4.2) and tactics (Section 3.4.3) mentioned by the participants.

3.4.1 Goal of recommendation

Being a public organization means that the primary goal of the organization is to provide services that benefit society [164]. As such, each organization's goals with building a recommender system extend beyond selling ads and optimizing for clicks, as is the norm for most commercial organizations. This difference between being a public or a commercial organization is explicitly mentioned by most interviewees (L2,L3,L4,B2,B3,B4,B5,N1,N2), and awareness of this distinction can be considered central in day-to-day operations. In the absence of optimizing for monetary gains, the goal of recommendation is different for each organization, and is strongly linked to its mission.

News

Both the News organization and the Broadcaster highlight that they are "for and from everyone". For the News organization, this is fairly straightforward: to *"reach the widest possible audience, [and] enable people to be well-informed"* (N2). This means that the news should be accessible to anyone, regardless of skill level, and that journalistic quality takes precedence over personal interest (N1,N2). Each article has a non-personalized 'more like this' section which is automatically populated by a recommender system, but can be supplemented or turned off by the editorial team. Furthermore, there is a personalized recommender system under development that aims to connect readers to "important news they have missed" (N1). For this, the organization is experimenting

with how behavioral data and editorial selections can complement each other, in order to compete against the commercial players (N1).

The News organization notes a very strong collaboration between the technical and editorial teams. When opening the app, the users will always land on the editorially curated front page listing the most important news of that moment. At any time, the editorial team has the power to turn off the recommender systems. While diversity is an important concept in the news organization, it is not something that is currently explicitly built into the recommender system. Rather, it is seen as a procedural thing, to be considered at every step of content creation. This goes from choosing which stories and events to cover, to which people are doing the reporting, to writing about events in a neutral way covering multiple perspectives. This control results in a set of items to recommend from that *"have to be told from a diverse standpoint in the first [place]"* (N1). The organization sees the UX (i.e., user experience) design of the recommender system, including where it is placed within the app and accompanied by which headers and explanations as more impactful (N1). They do see potential in personalization of style, telling the news through the user's desired medium (text, video, audio) or in a language complexity level suitable to the reader (N2).

Broadcaster

For the Broadcaster, the mission to be 'from and for everyone' translates a little differently. The organization is effectively an umbrella for multiple smaller broadcasters, each representing a different section of Dutch society. They are the ones pitching ideas and creating the content, though the public broadcaster may request a certain type of program if they feel a particular perspective is missing (B3). The broadcaster then tries to balance on the one hand helping their users recognize themselves in the content they have on offer, while also fostering understanding and knowledge of people, ideas and groups that are 'different' (B1,B2,B4,B5). For this, they see a clear purpose for personalized and diverse recommendation: *"[I]f someone believes or thinks or feels in X, and they look at our platform, that person is also confronted with Y. And that Y is slightly different from what the person had in mind with X."* (B2). Diversity plays a large role in achieving this, but the split in goals between recognition and broadening causes a great deal of conceptual unclarity. As one of the interviewees notes, *"[i]f you are looking at personalization, then we don't want people to all get the same thing recommended. That's what I tend to see as diversity. [...] [P]luralism in recommendations would be [that] we'd like to look into [a] topic from different perspectives. And I think if you talk to different people within this organisation, those two things kind of mix."* (B6).

The current platform is largely manually curated, with separate sections (displayed relatively low on the landing page) for algorithmic recommendations. These algorithmic recommendations balance personal relevance, based on a user's past viewing behavior, with a so-called 'public value score'. This score is an aggregate, obtained through a daily survey sent to users, in which they are asked to score the programs they watched recently on things such as the presence of certain population groups or multiple perspectives. The broadcaster is working on the development of a new platform, which should have a much stronger algorithmic focus. They hold a uniformly strong position on user control: while they, as the broadcaster, should provide a diverse offering, the user is eventually always in charge of what they do and do not watch.

Library

The Library hosts a vast collection containing every book published in or about the Netherlands (L2). They hold a unique position among the cultural institutes, as they have the added role of coordinating 138 other public libraries. There is high interest in digitization and innovation in general, which manifests in creation of proof-of-concept applications that the other libraries can choose to adopt or not (L3). Among other services and initiatives, the Library has the dual goal of having *more people read* and *people read more* (L4). The Library suspects that the decline of reading among its users is in part caused by a general lack of available time. Currently, there is a recommender system feature in the mobile application hosted by the Library, but its functionality is not satisfactory as *"people cannot find something they might be interested in"* (L4), an issue partly caused by the organization of metadata. Improved personalization can assist with the goal of attracting more patrons, as it increases accessibility to the collection for different types of people. At the same time, the Library is conducting research on the ethics around recommender systems, with topics like bias and privacy being central (L3). In this sense, personalization is not sufficient. There are active efforts to reduce bias that historically existed in the collections, in order to facilitate *"every citizen to be able to participate in society" and "make society [as a whole] better, smarter and more creative"* (L2).

Overall, the Library aims to deploy a well-functioning recommender system to satisfy their readers while remaining inclusive and enacting bias correction. However, building such a system is not trivial, and there are many questions about how development should be approached.

3.4.2 Aspects of diversity

During the recommendation exercise, and the questions leading up to it, participants would mention aspects of the content or the user beyond personal relevance that would lead them to include that item in the diverse recommendation or not. Figure 3.1 provides a schematic overview of these aspects, which can be divided into Item-, Human- and World aspects.

Item aspects

The first dimension considers aspects of the items that are to be recommended. For the Library these would be books, news articles for the news agency, and videos for the Broadcaster. The first group of aspects revolves around *topicality*, or what is actually discussed in the items. The dimensions mentioned by most interviewees are **category**, or sometimes genre, and **topic**. All interviewees but one mentioned balancing different categories and topics in the recommendation, to avoid saturation and to keep a user engaged. This is reflective of how diversity is currently most often conceptualized in technical recommender system literature; see [38].

Related but still separate from the topic are the **geographic location** the item centers on, and the **time period** it discusses. The Library may want to recommend books that discuss a topic through the lens of different time periods (L1), whereas the News organization aims to ensure that they not only cover news from densely populated urban areas, but also from rural areas (N2). Secondly, recommendations

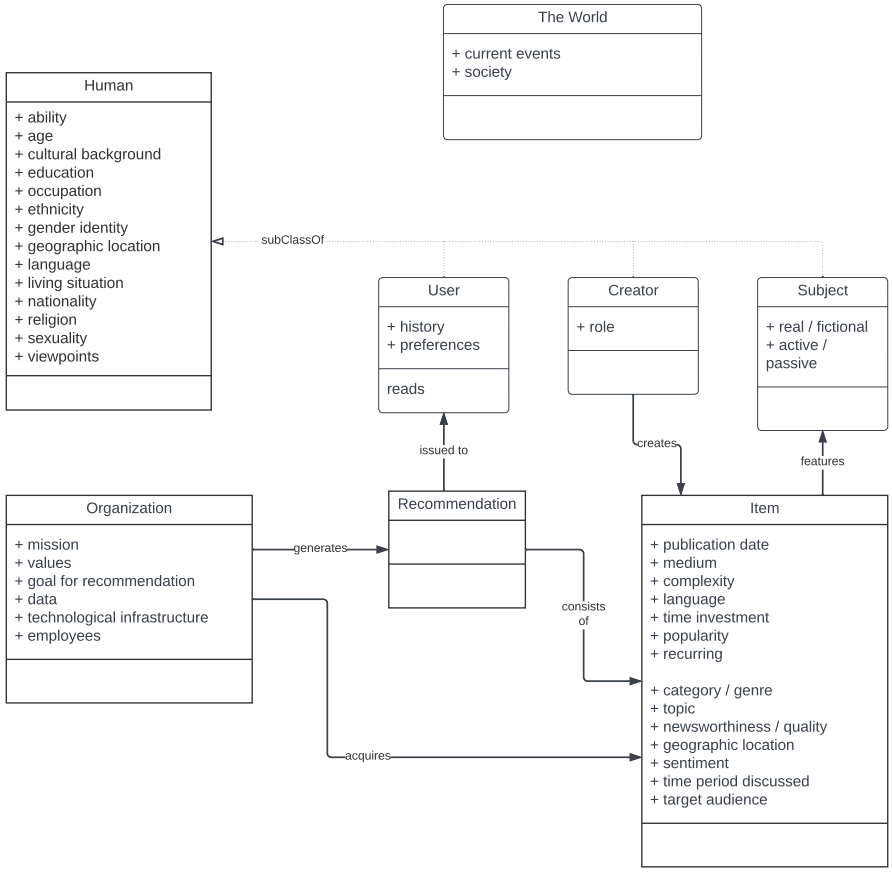


Figure 3.1: Schematic overview of the identified aspects of diversity and how they interact with each other. The 'World' class is unconnected in the graph, but in truth encapsulates and informs everything: from the content that is being produced, to what user wants to read, to what constitutes a 'viewpoint' or a 'minority'.

may vary on accounts of *stylistic* properties. This relates to the way in which a message is communicated, and whether it is easily accessible for the user. Examples are the **complexity** of the content, the amount of **time investment** a user is required (and willing) to make, the **language** it is written in, the **target audience** of the item, and its **medium** (text, video, audio). Many of these properties are symmetric with characteristics of the user: a certain item complexity level or language requires a certain amount of skill from the user.

Thirdly, there are a range of other item properties which may be prioritized. An important one is an item's **publication date**. While this is of primary importance to the news agency (one would not want to recommend a news item from two months ago), it is much less so for the broadcaster and the library: while there may be some preference for new publications, they also want to help their users discover less popular content. Another important dimension is the perceived **quality** of the item. While the news organization would explicitly optimize for (editorial) quality, opinions differ between participants from the library: while some would find it acceptable if readers only read comic books and manga, others wanted them to be pushed towards more high-brow literature (L4). More differences could be observed in terms of an item's **popularity**: some said that popular items are likely to be items that readers are looking for and should therefore be recommended (L4,B1), while others explicitly distanced themselves from it: *"I do know that out of this selection, [sensationalist article] would have been clicked on more than [serious article]. [...] That's why what we're doing with those recommendations is designed the way we designed it, which is that the journalistic selection we make takes precedence."* (N2). Additionally, items that had a high production **cost** may be prioritized (B3,B4).

Last, but definitely not least, interviewees referred to the people involved in the content. These can be split into the **Creators**, such as producers or authors, and the people that were **Subject** in the items, either through active participation or being discussed by others. Think of guests at a talkshow, politicians discussed in the news or a novel's protagonist. These are further discussed in Section 3.4.2.

Human aspects

While item category is the aspect of diversity mentioned by most interviewees, it is closely followed by notions of the diversity in the context of *people*. While there are a number of attributes that all humans share, there are three 'subclasses' that can be derived with specific roles within the system. These are Users, Creators and Subjects. Users are the people that receive the recommendations, and have a set of unique properties that the other types do not have: their **history** and **preferences**. Creators are, as the name implies, in some way or another involved in the creation of the items; the authors of books, journalists or producers of shows. Lastly, there are the people that are the Subject of the item's content, such as protagonists of books or people appearing on talk shows. These subjects can be fictional or real, and be active or passive agents within the Item by either speaking on their own behalf or by being described or mentioned by others.

The general Human aspects, such as **age**, **gender identity**, etc, are applicable to each type. A frequently mentioned aspect of diversity is relating to a person's

background. **Culture**, **ethnicity**, **nationality** and sometimes **religion** (for example 'Muslims') are often used interchangeably, and the distinction or relation between them is not always clear. For example, B5 notes: *"I think about [...] representing different societal groups. So ethnic minorities. More black and white. Maybe a bit more foreign language [...]"*, while L2 says: *" I don't know if this persona is from a specific country or has a specific background but [author] is an American author."* Identifying the presence of different ethnic and/or cultural groups is notoriously difficult, partly because the data required to do so is often lacking. In each of the domains described above, textual data was the only type of data available, and often no additional information was present beside a person's name[2]. In some cases, data enrichment might be possible (B3). Yet, this requires building a dataset on attributes that are often considered sensitive, which gets even more problematic when considering the fact that, to 'expand horizons' or 'represent', a similar type of dataset would be necessary about the users, which would then lead to issues of privacy (L2); see also Papageorgiou and Mougan [171]. Similar safety concerns can be raised for aspects like **gender identity** and **sexuality**; see Pinney et al. [174].

Culture- and **viewpoint** diversity are sometimes mentioned in the same breath: *"for me, diversity means having multiple colors, multiple opinions, multiple cultures, multiple points of view about something."* (B2). We make a distinction between the two, and denote viewpoint diversity as a person's perspective on events. Viewpoint diversity then often becomes linked to politics, or opinions about current events. A rich body of work exists around so-called viewpoint or stance detection, aiming to algorithmically extract these from text [16, 65, 188]. Even when successful, it is often unclear what a diversity of viewpoints would look effectively like, which is further discussed in Section 3.4.3.

World aspects

Aspects related to the World are beyond the direct control of a recommender system or user. They influence which content is created, and as such the pool of items that a recommender system can make its selection from. They also interact with a user's preferences to determine what type of content a user is interested in, looking for, or should be reading. **Current events** determine what is newsworthy, and what news a user needs to consume to fulfill their information needs. Naturally this aspect is of high importance to the News organization (N1,N2) and to some extent the Broadcaster (B3, B5), but much less so to the Library. In contrast, **society** represents the world as it currently is, including its biases and existing power distributions. An organization may choose to either reflect or counter these existing structures (see Section 3.4.3). For example, the Library acknowledges that much of their catalogue consists of white, American male authors, and wants to account for this in their recommendations (L1).

3.4.3 Tactics for diversification

During the experiment, the participants considered and deployed different tactics to produce a diverse recommendation. Some participants articulated the tactic they

[2]Some approaches have attempted to predict a person's ethnicity based on their name, but this process is not generally accurate and can suffer from misrecognition bias [135]

deployed explicitly, while others did not clarify it in as much detail.

Diversity between items

Diversity between items is perhaps the most intuitive interpretation of diversity, and refers to ensuring that all items within a list of recommendations are sufficiently different between them. Following this tactic requires choosing one or more appropriate axes to diversify over. In that sense, constructing a diverse list also entails exclusion, since some aspects have to be deemed less relevant. This issue is particularly important for public organizations; selecting meaningful aspects for which diversity needs to be safeguarded is a crucial decision that must comply with the values and mission of the organization. Multiple interviewees constructed a recommended list by considering diversity between items in the process (B2,B3,B4,B5,B6,N2,L1,L2,L3,L4), especially to recommend items that are diverse between them in terms of category. In particular, the Broadcaster may aim for a *"balance between [e]ntertainment and information"* (B2), given their wide range of offerings and mission *"to inform, educate, and entertain"* (B2). For the Library, item category translates to book genre, and is commonly taken into account when creating a diverse list (L2,L3,L4).

Instead of generally diversifying between items, a somewhat adapted tactic is to recommend a set of items that are different between them in some perspective, but engage with the same topic or theme. For example, one interviewee opted for selecting *"five books that give [...] an insight on [...] LGBTQ [...] [,h]ow that works or is around the world in different cultures and different times"* (L1). Furthermore, in case of a socially relevant emerging topic, the Broadcaster can create *"a highlight lane and then offer a lot of different opinions that people can scroll through"* (B5).

Finally, interviewees noted as a result of the experiment that *"to combine several aspects [is] really hard"* (L2). It might be that a set of items is diverse over one important axis, but not over a different equally important one. In this context the act of creating a diverse recommendation can be seen as an optimization process that an algorithm can contribute to. Regardless, ensuring diversity between items requires some sort of aspect prioritization, as well as an appropriate justification for it.

Diversity as a within-item measure

A different tactic for ensuring diversity is recommending items that consist of diverse perspectives or types of people within the item itself. For example, according to a participant from the Broadcaster, a travel series that allows the viewer to *"see multiple worlds [...] fits very well into diversity"* (B2). This also pertains to episodes of political programs where *"people from a lot of the major parties [appear], which captures a big part of the political spectrum"* (B4). The News organization notes that their inventory consists of *"very good stories which have to be told from a diverse standpoint in the first [place]"* (N1). From this perspective, guaranteeing item diversity may be a part of the production/acquisition process or an explicit step of the recommendation. During the exercise, one interviewee (B4) suggested combining within-item and between items, by composing a list of items on different topics where each item contains multiple perspectives on the respective topic. According to the interviewee, pursuing diversity in recommendation can also be seen as a process of achieving maximum diversity. This tactic requires that the media organization's catalogue consists of items that can facilitate the maximization process.

Diversity considering the user

User-specific diversity is a personalized form of diversity, where the user's history and/or preferences are taken into account when composing a diverse list of items to recommend. The first way to operationalize this tactic is to cater to a user's specific needs that potentially diverge from the norm. This can manifest in assisting the user with finding uniquely niche content. Based on this outlook, diversity is about *"ensuring that [...] everyone feels that there is something for [them]"* (B1) in the catalogue. The Library and News organization also mentioned literacy in this context (L4,N1), as *"how [to] help those with difficulty reading [...] [is] also a diversity topic in a way"* (N1). Additionally, for the Library diversity in accordance with the user can help them *"recognize themselves in the author or in the main characters or the topics [such as] age and physical ability and sexual orientation"* (L2). The second way to apply user-specific diversity is to recommend to a user content that extends further from their current interests, and that they might not be aware of. In other words, an organization can also recommend items so as to *"give people [...] a broader perspective of what is available"* (L3) and to *"broaden the user's horizons"* (B5). This tactic can be seen as a way to ensure that users do not *"stay in their own bubble"* (L2) by *"educating people [that] there's more than [their] bubble"* (B5).

Diversity considering the world

Diversity considering the world entails recommending items that in some way reflect or diverge from the norms that exist in the world, leading to 'similar to world' and 'different from world' diversification tactics. When reflecting the world, one could think of having a good spread of topics that are representative of the important news of that day (N2), or, by representing political parties in a way similar to their distribution in government (B4). With a 'different from world' diversification tactic, the aim is to counter existing power structures. With this interpretation, an item can be considered diverse on its own merit. This can be because the content represents a minority culture, an "outsider [...] in a political sense" (B4), a *"very different part of the world that [the society] knows too little about"* (B2), or even a theme or topic that *"you do not find that many [items] in the catalogue"* (L4). In this case, very popular items may be deemed inherently not diverse, and might not need to receive further exposure: *"[I]t's going to be on the website probably anyway"* (B6).

What constitutes a minority was often implicitly assumed by the interviewees, potentially due to the social context that their organization operates in. For example, multiple interviewees referred to LGBTQ-themed media as inherently diverse (B4,B5,B6,L2), which also prompted their inclusion in a diverse recommendation as part of the experiment. The same can be said for media that gives exposure to people with a minority cultural background or engages with topics not related to the Randstad, a dominant urban area in the Netherlands (N2). From this perspective, diversity considering the world and user-based diversity (in a broadening-horizons way) can overlap when a user is assumed to be the typical or default representation of the majority culture in the given context.

3.5 Discussion

The wide range of interpretations highlighted in Section 3.4 show that, even within this limited domain, conceptualizations of diversity among participants vary greatly. It is therefore unlikely that a standardized definition, encompassing all potential goals, aspects and tactics, can be attained; however, that does not mean that it is impossible to implement a meaningful form of diversity into a recommender system. In this section, we outline the implications the findings of this study have for the implementation of diversity-aware recommender systems.

3.5.1 Implementing diversity: a normative process

Diversity has traditionally been seen as something we always want *more* of: more viewpoints, more topics, more different nationalities. However, during the interview participants often mentioned instances in which they would actually want *less*. An interviewee from the Library wanted to recommend items that fit the amount of time a user was willing to invest, while the News organization only wanted to recommend items of a certain quality. In these cases, meaningful diversity could be achieved by controlling for one aspect and diversifying over another; for example, when recommending different viewpoints for one controlled topic, or vice versa, by showing the opinions of one particular group over a wide range of topics. The same dynamic also occurs in the way interviewees spoke about diversification tactics. In some cases, a recommendation similar to a user's preferences or history would be desirable, in order for them to recognize themselves in the content on offer. In others, it should be different, so as to expose them to new perspectives. This shows that different applications may not only prioritize different aspects of diversity, they may also have different expectations on whether a recommendation should have high or low diversity.

Participants would often mention that recommendations should be 'diverse enough.' This requires an underlying model that not only determines which items are similar and different, but also informs the system whether sufficient diversity has been attained. This is non-trivial, especially in a domain such as news recommendation, where contexts change rapidly. One could imagine using an external source as a reference point: for example, the presence and size of political parties in government, or a country's composition in terms of cultural groups as determined by a national statistics agency. However, this yet again relies on a conscious decision of what is the 'right' model to use and reflect through the recommendations, and each choice will have pros and cons. There is therefore a normative choice to be made about the type of diversity that is desired, with different aspects to consider, different diversification tactics, and different levels of diversity. The wide variety in the answers given by participants is also an indication that this normative choice is not one that can be taken lightly, and requires a good deal of internal discussion and alignment with all the relevant stakeholders involved before it can be satisfactorily modeled and implemented.

3.5.2 Generalizing to domains beyond public service media

While the results are not directly generalizable to other domains, there are likely elements of the aspects and tactics identified here that would also be applicable. For example, while music recommendation might put less stock in the diversity of people

discussed in an item's content, they would be interested in the diversity of Creators [76]; similarly, while in-item diversity would not be applicable, they could be interested in countering existing biases through different-from-world recommendation tactics [64]. We believe that the goals, aspects and tactics of this chapter can still serve as a starting point for discussion in other domains, which may make it easier to identify gaps and unique challenges.

3.5.3 Exploring versus defining diversity

Participants varied greatly in which diversity aspects and diversification tactics they mentioned. Some of these differences can be traced to the different ways participants speak: some people are more verbose, more inclined to stick to a specific set of examples, or already have a more developed concept in mind than others. Furthermore, the three organizations were at different stages of developing strategies towards diversity, which may have an impact on said differences between participants. However, they are all working on or considering the implementation of a recommender system. Our participants are thus also actively making decisions about how diversity would be conceptualized and implemented. They are as such representative of the 'real' world, rather than the 'ideal' world, and therefore suitable candidates for an exploration of the dimensions along which diversity is currently conceptualized. For these reasons, we refrain from making claims about whether certain aspects or tactics are 'more important' than others, nor do we make claims about the 'correct' definition of diversity. Our hope is that the overviews of goals, aspects and tactics presented in this study will facilitate building a common understanding and vocabulary between stakeholders with a different background, making it easier to find common ground and establish priorities that reflect the requirements of that particular implementation.

3.6 Conclusion

Through interviews with participants from public service media organizations, we identified a range of different interpretations of diversity within a recommender system. We grouped these into different goals (e.g. broadly informing the public or allowing a user to recognize themselves), aspects (e.g. the topic of the content, or the cultural background of an author), and diversification tactics (e.g. ensuring diversity within a single item or countering biases in society). Given the great variety found in the conceptualization of diversity in recommender systems, even within a limited domain, we find that it is unlikely that a standardized definition can be attained. Instead, rather than striving for this standardization, we argue that it should be conceptualized on a case-by-case basis. We hope that our mapping of goals, aspects, and tactics can contribute in effectively communicating what diversity entails within a specific application.

Acknowledgements

This work was supported with seed funding from the RPA Human(e) AI, round 2022/2023, and the AI, Media and Democracy Lab, NWA.1332.20.009. The authors would like to express their gratitude to everyone that provided input and feedback on

the project; Sophie Morosoli and Hannes Cools for their help with the methodology; Naomi Appelman, Kimon Kieslich, Midas Nouwens and Marijn Sax, for providing input that helped shape the direction of the paper; the three anonymous FAccT reviewers for their helpful feedback; and lastly to Natali Helberger and Claes de Vreese for their support and reviews.

Statements

Ethical considerations statement

A main ethical concern in interview-based research is breaching the anonymity of the interviewees. To support anonymity, we refer to the organizations by their general functionality instead of naming them. To the interviewees themselves we refer by their organization's functionality as well as a random identifier. Even though we did consider their position in the organization when selecting them, we do not include it when quoting them, as an extra measure for anonymity. Additionally, we asked the participants to sign an informed consent document. We explicitly asked for permission to record the interviews. Finally, we forwarded to them the conclusions of our research and quotes attributed to them to ensure we correctly interpreted and contextualized their words.

Researcher positionality statement

The authors of this work are all European citizens who operate in the academic circles of the Netherlands. This shaped the work as we interviewed practitioners from the Netherlands who all work for organizations that we are familiar with in a professional and personal capacity, and some are in turn partly familiar with our work, which rendered communication with them easier. Three out of the four authors of this chapter work in the computer science field, and one has joint expertise in computer science and communication science. For this reason we also received help (see Acknowledgments) from communication scientists, in particular when it comes to structuring the interview and devising a coding scheme. One author works closely with cultural institutes, one with media institutes, and one with both, which helps us contextualize the statements of the interviewees.

Adverse impact statement

Despite our efforts to ensure anonymity of the individual interviewees, it might be that readers familiar with the media landscape of the Netherlands can deduce which organizations we refer to in this chapter. Additionally, readers of this chapter might generalize the personal statements of the interviewees as completely representative of the entire organization that they work in. Finally, we do not intend our mapping to serve as a final and conclusive categorization of media diversity, but it might be interpreted as such by readers.

4

Reproducing bias in recommendation studies

Having established current perspectives on recommender systems bias, we now examine a methodological challenge: how can we reliably measure and report on bias in recommender systems? In this chapter, we reproduce three prominent studies on popularity bias in media recommendation and identify which methodological choices drive divergent findings, thereby contributing to more robust bias evaluation.

The extent to which popularity bias is propagated by media recommender systems is a current topic within the community, as is the uneven propagation among users with varying interests for niche items. Recent work focused on exactly this topic, with movies being the domain of interest. Later on, two different research teams reproduced the methodology in the domains of music and books respectively. The results across the different domains diverge. In this chapter, we reproduce the three studies and identify four aspects that are relevant in investigating the differences in results: data, algorithms, division of users in groups and evaluation strategy. We run a set of experiments in which we measure general popularity bias propagation and unfair treatment of certain users with various combinations of these aspects. We conclude that all aspects account to some degree for the divergence in results, and should be carefully considered in future studies. Further, we find that the divergence in findings can be in large part attributed to the choice of evaluation strategy.

This chapter is based on *Daniil et al., Reproducing Popularity Bias in Recommendation: The Effect of Evaluation Strategies, ACM Transactions on Recommender Systems, Volume 2, Issue 1* [58].

4.1 Introduction

In this chapter we reproduce three papers that study popularity bias in media recommender systems. Websites that host media content are known to employ recommender systems to filter through the content and provide the user with personalized suggestions. In the case of collaborative filtering, neither explicit demographics of the user nor information about the content are needed to encode their taste, but only consumption and browsing history. Despite the lack of explicit input of user or item characteristics in the system, collaborative filtering approaches are still known to suffer from bias [216]. Popularity bias has been identified as a relevant issue with negative implications from a multi-stakeholder perspective [3]. In short, popularity bias is the algorithmic phenomenon where items already popular in the users' profiles tend to become even more popular due to being disproportionally recommended.

In the context of media recommenders, different research teams have studied the impact of popularity bias on the users. Specifically, they focused on the extent to which it impacts them disproportionally based on their interest for popular items, which they refer to as unfairness. In 2019, Abdollahpouri et al. [9] published a paper called "The Unfairness of Popularity Bias in Recommendation" where the propagation of popularity bias by different algorithms and for different user groups was reported in the setting of movie recommendation. Two subsequent papers reproduced the work to evaluate the same phenomenon in music recommendation (Kowald et al. [122]) and book recommendation (Naghiaei et al. [159]). The papers used a similar process and metrics to evaluate the unfairness of popularity propagation, but different datasets, data preprocessing, and some different algorithms. While all studies reported on the same types of results, diverging results were presented. Both Kowald et al. [122] and Naghiaei et al. [159] proposed that the differences in data characteristics might be the cause.

The studies take significant steps in the direction of understanding the unfairness of popularity bias, and a new metric for popularity bias is proposed by Abdollahpouri et al. [9]. The effect of popularity bias on several baseline and state-of-the-art collaborative filtering algorithms is analyzed. The view of unfairness is user-centric, unlike previous work on the matter, which contributes to the significance of these studies in the field of recommender systems. It is, therefore, pertinent to comprehend the source of divergence in their reported results when it comes to whether certain algorithms propagate popularity bias, as well as the extent to which certain user groups are unfairly treated.

The following results were presented by all three studies:

- **Overall Popularity Bias Propagation**: The studies observed whether frequency of an item in the users' profile and in an algorithm's recommended list correlated (the higher the correlation the larger the bias), and whether only a few items were getting recommended to all users. Their results diverged in the following ways:
  - Abdollahpouri et al. [9] and Naghiaei et al. [159] found that only for certain algorithms there is positive correlation, with the correlation found by the latter being stronger than by the former.

  - Kowald et al. [122] found that this correlation exists for all tested algorithms rather than for only some of them.

- **Popularity Bias Propagation per User Group**: The studies evaluated unfairness of popularity bias propagation by using the %Δ *Group Average Popularity* (%Δ*GAP*) metric, which is defined as the difference in average popularity of items in a user's profile and in an algorithm's recommended list, averaged for a group of users. Specifically, they calculated %Δ*GAP* for different user groups defined by their propensity for popular items, and then compared the results between groups and algorithms. The higher the %Δ*GAP* for a group, the more unfairly this group is treated.

  - Abdollahpouri et al. [9] and Naghiaei et al. [159] found that popularity bias is larger for users who prefer niche (i.e. not popular) items than for other user groups. Certain algorithms examined by Naghiaei et al. [159] did not propagate popularity bias for any of the user groups, which was not the case for Abdollahpouri et al. [9].
  - Kowald et al. [122] observed no such clear difference between the groups for any algorithm.

In this chapter, we perform an extensive reproduction study of the three above mentioned papers (Abdollahpouri et al. [9], Kowald et al. [122], Naghiaei et al. [159]). We are motivated by the divergence in results and wish to investigate and comprehensively report on the source. We do not only attempt to replicate and verify the individual claims, but also locate properties of the recommendation and evaluation process that have an impact on popularity bias. Therefore, our reproduction study allows us to draw conclusions further than the transferability of a claim to a different domain, which was the goal of the two reproduction studies (Kowald et al. [122], Naghiaei et al. [159]). By zooming in on these studies and understanding how their differences in implementation resulted in divergent results, we can offer insights and suggestions on the topic of popularity bias evaluation, and reproducibility in recommender systems in general.

First, we study the choices made by Kowald et al. [122] and Naghiaei et al. [159] when reproducing Abdollahpouri et al. [9]. We recognize the challenges that stem from either lack of clarity around or differentiation in the strategies the three studies used to evaluate popularity bias. We identify four aspects that are relevant in investigating the differences in results between the three studies:

1. data; the studies use three datasets, with different characteristics such as size, sparsity and distribution of item popularity.
2. algorithms; the studies evaluate mostly different algorithms, with some exceptions.
3. division of users in groups; the studies define propensity for popular items differently and divide them accordingly.
4. evaluation strategy; the studies make different choices in the testing process.

Second, we run a set of experiments in which we measure popularity bias with various combinations of these aspects. For example, we run the evaluation strategy of one paper on the dataset of the other paper. We find that evaluation strategy has a significant impact on the reported results. Specifically, certain algorithms trained on the same data and with the same view of popularity either show or not show propagation of popularity bias depending on the evaluation strategy. We believe that our observations can prove useful for future efforts to reproduce evaluations of recommender systems, as they give insight on how strategic choices affect the outcome even when the same evaluation metric is used.

4.2 Related work

4.2.1 Reproducibility in recommender systems

The ability to reproduce and examine published studies is valuable, as reproducibility of experimental results is a cornerstone of science [87]. However, the field of AI as a whole is known to face reproducibility issues [104]. Machine learning is highly dependent on the characteristics of the chosen training data, as well as parameter tuning to fix the inherent randomness of the algorithms [250]. Therefore, lack of published code and data render the replication of existing work by other researchers challenging [91]. Many reported results are irreproducible, though Gundersen and Kjensmo [87] found a significant increase in documentation over time. In the context of recommender systems, lack of reproducibility is recognized as one of the key components that have led the field into "a state of stagnation" [50]. While reviewing recommender systems papers published in big conferences such as KDD, IJCAI and SIGIR, Ferrari Dacrema et al. [74] found less than 50% of them to be reproducible. Cremonesi and Jannach [50] consider lack of incentive to be the main reason behind limited reproducibility, as the academic system's operation often motivates researchers to publish *more* instead of putting in the work to provide sufficient code, data and documentation.

Recently, AI conferences have started providing checklists to ensure that the submitted papers are sufficiently reproducible [173]. Additionally, many conferences such as RecSys, ECIR and SIGIR have initiated reproducibility tracks where researchers can submit studies that reproduce, analyze or reflect on prior work. Gundersen et al. [88] provide a set of specific recommendations on how to ensure reproducibility in AI research. They motivate researchers by highlighting the importance of reproducibility for the improvement of science overall as well as the benefits for the researcher themselves. Beel et al. [26] outline actions that can make recommender systems research more reproducible, such as adopting practices from medical sciences, social sciences, and natural sciences, and conducting more comprehensive experiments, for example by varying model parameters and observing the effect. In this work, we attempt to further motivate reproducibility in recommender systems research by experimenting with various aspects of the recommendation process and showcasing the effect on the phenomenon of popularity bias.

4.2.2 Evaluation strategy

In recommender systems research, reproducing the evaluation process followed in previous studies is not always trivial. To assist researchers in this endeavour, Said and Bellogín [196] describe the dimensions of evaluating recommender systems as dataset, data splitting, evaluation strategies, and metrics. Specifically, they identify and benchmark four stages of designing an evaluation protocol: data splitting, item recommendation, candidate item generation, and performance measurement. Application of metrics takes place in the final stage of evaluation, namely when measuring the performance of the scores produced during the recommendation process. Before measuring performance, it is necessary to define a crucial aspect of the evaluation strategy; generating candidate items for recommendation for each user, so that scores can be produced for them. The process of evaluating recommender systems includes reporting on results of commonly used metrics, which allows comparison between different algorithms to estimate their success in the context of the task at hand. To calculate the metrics in a way that meaningfully represent the success of the system, it is necessary that the evaluation strategy is suitable for the system, as well as for the metrics to be measured.

In terms of metrics, Gunawardana et al. [86] list a set of properties frequently taken into account when evaluating recommenders, such as accuracy, coverage, novelty, diversity etc. They point out that different applications have different needs, and thus the metrics to be evaluated should be chosen to appropriately reflect on the desired property. Said and Bellogín [196] argue that comparison of recommendation quality between different studies requires careful consideration of all stages of an evaluation protocol. They experimentally show that even when recommendation algorithms are similarly implemented, the results are not comparable when different evaluation strategies are used. In our study, we follow a similar approach of employing different evaluation strategies on the same task and comparing the results. We are prompted by variations in evaluation strategy choices made by published reproducibility studies, which highlights the importance of taking into account and addressing all evaluation steps in recommender systems research.

4.2.3 Popularity bias

Studying and evaluating a specific phenomenon requires understanding its roots and effects. In the context of item popularity, it is known that consumption of media often follows a long tail distribution, where a few items are very popular and the rest are located in the heavy tail [20]. Recommender systems attempt to facilitate users in their effort to discover items appropriate for them, regardless of the items' position in the long tail. However, it has long been noted that collaborative filtering techniques can be prone to popularity bias, the phenomenon where popular items tend to be recommended over long-tail ones [39]. When an algorithm showcases such tendency, it may produce ranked lists with items not equally covered along the popularity tail [248]. Bellogín et al. [28] argue that common ranking metrics calculate overall assumed satisfaction of a population, and therefore produce high values when the recommended list consists of mostly popular items, even when it is not personalized. Item popularity is not a de facto bad criterion for recommendation and can be leveraged to track item

quality [43, 261]. However, very popular items are often more likely to be already known, which renders the recommendation of mostly popular items to a user potentially not useful for the user and the system in general [2].

In addition to the three studies examined in this chapter, other studies evaluate popularity bias in various contexts and propose methods to mitigate it [35, 101, 113, 264]. These studies vary in terms of methods to measure popularity bias and findings. Elahi et al. [71] find that propagation of popularity bias is dependent on the scenario and domain; in certain cases, popularity bias is reduced through the recommendation process. In a follow-up paper to Kowald et al. [122], Kowald and Lacic [121] study popularity bias by four collaborative filtering algorithms on all three datasets used by the three studies we are reproducing, namely MovieLens1M, LastFM and Book-Crossing. It is interesting to note that their results are similar across the three datasets, while they diverge from the results reported by Abdollahpouri et al. [9] and Naghiaei et al. [159]. This observation implies that the discrepancy in results showcased by the three studies cannot be explained solely by the differences between the datasets. In this work, we wish to investigate which aspects of the process account for the discrepancy. Our results and conclusions can contribute to consistency and reproducibility when evaluating popularity bias and other recommender systems phenomena by the research community.

4.3 Overview of the studies to be reproduced

In order to comprehend the differences in results between the three studies, we studied and collected the details of each approach to accurately reproduce their process. Note that while Kowald et al. [122] and Naghiaei et al. [159] made their code publicly available, we could not find a public repository for the code developed by Abdollahpouri et al. [9]. Therefore, we describe the characteristics of their study based on the text of the published paper, and private correspondence with the authors. Throughout the paper, we explicitly state whether our understanding is based on this correspondence.

4.3.1 Core process

The studies followed similar processes to compare the popularity distribution in the data and in recommendations:

1. Analyze distribution of popularity among items in given dataset.
2. Label items as "popular" if they belong in the 20% most frequently rated items in the dataset.
3. Analyze distribution of user propensity for popular items and divide users in three groups (*Niche*, *Diverse* and *Blockbuster focused*) accordingly.
4. Set aside 80% of the ratings for training and 20% for testing.
5. Train the algorithms on the training data.
6. Recommend 10 items to each candidate user for each algorithm.

7. Report on overall popularity bias propagation: compare the number of times each item is recommended with the number of times it is rated. Repeat for each algorithm.
8. Report on popularity bias propagation per user group: compare average popularity of the items in profile with the items in recommendation for every user and average for each user group. Repeat for each algorithm.

4.3.2 Data

The studies used publicly available datasets which are commonly used in recommender systems literature. Abdollahpouri et al. [9] used MovieLens1M [94], Kowald et al. [122] used LastFM-1b [201], and Naghiaei et al. [159] used Book-Crossing [265]. Kowald et al. [122] and Naghiaei et al. [159] processed their respective datasets in order to approach the size of MovieLens1M. Specifically:

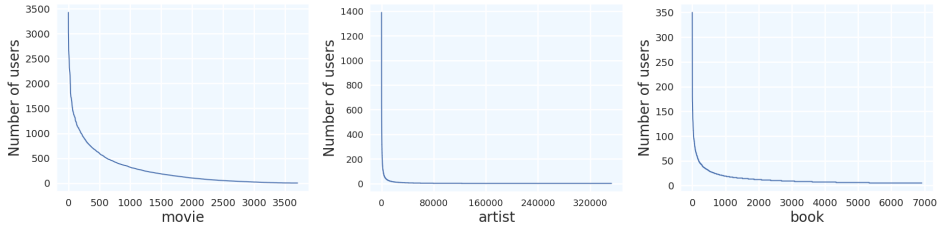
- Kowald et al. [122] extracted 3,000 users from the original dataset, which contains 120,000 users. They also grouped the listening events into user-artist pairings. Subsequently, they scaled the amount of times a user listened to an artist into a preference score from 0 to 1000. Therefore, the algorithms in Kowald et al. [122] do not try to predict explicit rating, but rather the preference of a user towards an artist, which is represented by the amount of times the user listened to this artist.
- Naghiaei et al. [159] only kept the explicit ratings from the original dataset. Afterwards, they removed users with more than 200 ratings. Finally, they removed users and items with very few ratings, until all users in the dataset had rated at least 5 items and all items in the dataset had been rated by at least 5 users.

Abdollahpouri et al. [9] do not explicitly mention whether they used the dataset intact or processed it before the analysis and recommendation process, and it was not clarified from our correspondence with the authors. Table 4.1 shows the characteristics of each dataset, namely number of users, items, ratings, and sparsity. Sparsity in the context of rating data refers to the percentage of possible user-item ratings that are missing from the data.

Dataset	#users	#items	#ratings	Sparsity
MovieLens1M	6,040	3,900	1,000,209	95.75%
LastFM-1b (subset)	3,000	352,805	1,755,361	99.83%
Book-Crossing (subset)	6,358	6,921	88,552	99.80%

Table 4.1: The characteristics of the datasets used by the three studies.

The resulting datasets differ in terms of size and sparsity. At the same time, item consumption is differerently distributed among them. Figure 4.1 shows the number of users that rated each item in every dataset. While every figure shows that the rating data is skewed towards more popular items, the long-tail distribution is clearer in the subset of LastFM-1b than in the subset of Book-Crossing, and especially than in MovieLens1M.



(a) MovieLens1M (b) LastFM (subset) (c) Book-Crossing (subset)

Figure 4.1: The long-tail distribution of item popularity in all datasets.

4

4.3.3 Algorithms

Each study examined whether several different algorithms propagated popularity bias into their recommendations, as seen in Table 4.2. They all tested two baseline algorithms, MostPopular and Random. Other than these, the studies tested mostly different algorithms. Abdollahpouri et al. [9] tested four well known collaborative filtering algorithms. Kowald et al. [122] also tested four collaborative filtering algorithms, but partly deviated from the choices of Abdollahpouri et al. [9]. Specifically, they excluded SVD++ and ItemKNN to reduce computational cost, potentially due to the large number of items in the LastFM dataset. Finally, Naghiaei et al. [159] tested a total of nine collaborative filtering algorithms in order to cover a wider range of state-of-the-art approaches. Overall, the list of algorithms consists of Nearest Neighbour-based, Matrix Factorization-based, and Neural Network-based approaches. The only common collaborative filtering algorithm across all studies is UserKNN [200]. The divergence in the choice of algorithms between the original study and the reproduction studies raises the question of whether a different overall conclusion would be drawn if all studies tested the same set of algorithms.

4.3.4 Division of users in groups

The notion of item popularity is central in the three studies, and it is defined as frequency of either rating by the users (*popularity in profile*) or of recommendation by the algorithms (*popularity in recommendation*). Abdollahpouri et al. [9] deem an item popular in profile if it is one of the 20% most frequently rated items in the entire dataset. This choice is important since Abdollahpouri et al. [9] use this label to divide the users based on their propensity for popular items. Therefore, the fact that item popularity is differently distributed across the three datasets as shown in Figure 4.1 affects which users are considered *Niche*, *Diverse* or *Blockbuster focused*.

Specifically, the user division in Abdollahpouri et al. [9] happens as follows:

1. The popularity of every item is calculated as the percentage of users who have rated it.

2. The items are sorted based on their popularity.

3. The 20% items with the highest popularity are labelled as popular.

Algorithm	Abdollahpouri et al. [9]	Kowald et al. [122]	Naghiaei et al. [159]
UserKNN	✓	✓	✓
ItemKNN	✓		
UserKNN w/ means		✓	
UserItemAvg		✓	
SVD++	✓		
NMF		✓	✓
BMF	✓		✓
PMF			✓
WMF			✓
HPF			✓
NeuMF			✓
BPR			✓
VAECF			✓
MostPopular	✓	✓	✓
Random	✓	✓	✓

Table 4.2: The algorithms tested by the three studies.

4. The *average propensity towards popular items* of every user is calculated as the average popularity of the items that this user has rated.
5. The *popularity fraction* of every user is calculated as the percentage of the items in this user's profile that have the label *popular*.
6. The users are sorted based on their popularity fraction.
7. The top 20% users are labeled "Blockbuster focused".
8. The bottom 20% users are labeled "Niche".
9. The remaining users are labeled "Diverse".

While Naghiaei et al. [159] follow the same process, Kowald et al. [122] differ; they do not use the popularity fraction to divide the users into groups. Instead, they use the *mainstreaminess score*, which is available for the users in the LastFM dataset. Mainstreaminess is defined as the overlap between a user's listening history and the aggregated listening history of all users in the original dataset. The 3,000 users are strategically extracted by Kowald et al. [122]; 1,000 users with low mainstreaminess, 1,000 with medium, and 1,000 with high. In other words, the mainstreaminess score is used as a proxy for user propensity for popular items and the user groups are divided based on that instead of user preference for items labeled "popular" in the final dataset. Note that the mainstreaminess score cannot be computed on MovieLens1M or Book-Crossing, as it requires multiple interactions between users and items (i.e., play counts) [121].

4.3.5 Evaluation strategy

A key difference between the studies' approaches is the strategy they adopt to evaluate the propagation of popularity bias. In all studies, the dataset of ratings is divided into a training and a test set with a 80-20% split. However, they differ in terms of which users they recommend items to, and when it comes to candidate item generation, a step of the evaluation process that was benchmarked by Said and Bellogín [196]. Said and Bellogín [196] discuss that this aspect of the evaluation protocol is crucial for the result, since by changing the item candidates for recommendation, a different ranking is evaluated. This is likely to have an effect on the measured popularity bias propagation. Table 4.3 shows an overview of the above mentioned characteristics of each study's evaluation strategy.

Let U^r denote the set of users that an algorithm recommends items to, and L_u the list of items that it ranks and chooses from to recommend to a user u in U^r.

Kowald et al. [122] recommend items to every user in the test set. To generate candidate items, they adopt the UserTest strategy [196]; the system only considers items that the user has rated in the test set. In other words, in their system U^r contains all users in the test set, and L_u for every user u in U^r contains every item i for which the rating (u,i) exists in the test set.

Naghiaei et al. [159] recommend items to all users in the training set. They adopt a version of the TrainItems strategy [196]; the system considers every possible item as candidates for a given user. In this case, U^r consists of all users in the training set, and L_u for every user u in U^r contains every item i. Note that Said and Bellogín [196] describe TrainItems as disregarding the items that the user has actually rated, but Naghiaei et al. [159] consider all items instead.

In our correspondence with Abdollahpouri et al. [9], they stated that they recommend items to all users in the test set. They also adopt the TrainItems strategy, but differently to Naghiaei et al. [159]. Specifically, for every user u in U^r which are all the users in the test set, L_u contains every item i that u has not rated in the training set. According to Said and Bellogín [196], this approach is suitable when simulating a real system where no test set is available.

Reference papers	User candidates	Item candidates
Abdollahpouri et al. [9]	Users in the test set	Items that the user has not rated in the training set
Kowald et al. [122]	Users in the test set	Items that the user has rated in the test set
Naghiaei et al. [159]	Users in the training set	All items

Table 4.3: Overview of the user and item candidates on which each study based the evaluation of popularity bias propagation. Note that, unlike Kowald et al. [122] and Naghiaei et al. [159], our description of the strategy used by Abdollahpouri et al. [9] stems from correspondence rather than studying their code.

4.3.6 Other variations

While in our experiments we consider the aforementioned four aspects that vary across the three studies, there are other variations that we do not consider:

- Kowald et al. [122] and Naghiaei et al. [159] both used Python to run their experiments, but used different Python-based libraries to perform the training and recommendation process. Kowald et al. [122] used Surprise[1], a toolkit for building and analyzing recommender systems that deal with explicit rating data [103]. Naghiaei et al. [159] used Cornac[2], a framework for multimodal recommender systems [197]. Abdollahpouri et al. [9] stated through our private correspondence that they used Librec-auto[3], a Python tool for running recommender systems experiments [212].
- In their paper, Abdollahpouri et al. [9] state that they tuned all collaborative filtering algorithms to reach a precision of 0.1, so that the results are comparable. On the other hand, neither Kowald et al. [122] nor Naghiaei et al. [159] tuned the algorithms. Instead, they used the default hyperparameters in the Python libraries.

4.4 Experimental setup

In order to investigate the cause of difference in results, we define a set of experiments that allows us to track which of the aspects that varies across the three studies affects whether popularity bias is propagated and whether the propagation unfairly impacts certain user groups. Each experiment consist of the following phases.

Training

- Divide the dataset with ratings into training and test set using a 80-20% split.
- Train the algorithms on the training set.

Prediction

- For every user u in U_r, predict ratings for item i in L_u based on each trained algorithm. (see notation in Section 4.3.5)
- Rank the items based on the predicted rating.
- Recommend to the user the top 10 items.

Evaluation

- **Overall popularity bias propagation**: Note that the three studies indirectly use the concepts of correlation and item coverage to report on overall popularity bias, by plotting frequency of an item in profile versus in recommendation for every algorithm and visually observing whether there is correlation and whether certain items are almost never recommended. In this study, we quantify these concepts as follows:

[1] https://surpriselib.com/
[2] https://cornac.preferred.ai/
[3] https://librec-auto.readthedocs.io/

– For the set of items I calculate

$$Correlation := r(P_I, F_I) \tag{4.1}$$

where r is the Pearson correlation coefficient, P_I is the list of popularities of each item in the users' profiles, and F_I is the list of frequencies of each item in the users' recommended lists (see [47]).

– For the set of items I calculate

$$Coverage := \frac{|R \cap I|}{|I|} \tag{4.2}$$

where R is the list of items recommended at least once.

• **Popularity bias propagation per user group**

– For every user group g calculate

$$\%\Delta GAP(g) := \frac{GAP(g)_r - GAP(g)_p}{GAP(g)_p} * 100\% \tag{4.3}$$

where $GAP(g)_p$ is the average popularity of the items in the users' profiles and $GAP(g)_r$ is the average popularity of the items in the users' recommended lists (see notation in Abdollahpouri et al. [9]).

– For every user group g calculate NDCG@10 [110]. Items for which no rating is available in the test set are assumed to have a utility of 0, as implemented by Ekstrand et al. [70].

We design the experiments by considering different choices for every aspect in which the studies deviated, and combining them in all possible ways. To train the models, we use the Cornac library (version 1.17.0), since it contains almost all algorithms needed. Similarly to Kowald et al. [122] and Naghiaei et al. [159], for every dataset we fix the random seed when splitting in training and test set for reproducibility. The code we developed for our experiments[4] has been made open source. In the subsequent subsections we summarize the choices for each of the four aspects.

4.4.1 Data

We train the algorithms using *MovieLens1M*, *LastFM*, and *Book-Crossing*. We use the entire *MovieLens1M* dataset, and the *Book-Crossing* subset that is used by Naghiaei et al. [159]. For computational reasons, we do not experiment with the entire *LastFM* subset that Kowald et al. [122] use. The large number of items makes some algorithms crash when combined with certain evaluation strategies, and we wish to evaluate as many combinations as possible. Instead, we sample by removing items with less than 20 ratings. This sampling does not decrease the number of users, but greatly decreases the number of items from 352,805 to 12,690. The characteristics of the resulting dataset can be seen in Table 4.4. To assess whether the sampling has a large impact on the conclusions, we compare our results with the results reported by Kowald et al. [122] and find that they are consistent.[5]

[4] `https://github.com/SavvinaDaniil/UnfairnessOfPopularityBias`

[5] See Appendix, Section 8.2

Dataset	#users	#items	#ratings	Sparsity
MovieLens1M	6,040	3,900	1,000,209	95.75%
LastFM-1b (subset of subset)	3,000	12,690	1,008,479	97.35%
Book-Crossing (subset)	6,358	6,921	88,552	99.80%

Table 4.4: The characteristics of the datasets used in our experiments.

4.4.2 Algorithms

We train all collaborative filtering algorithms in Table 4.2, with the exception of *UserItemAvg* and *SVD++*, which are not available in Cornac. In total, we train 11 collaborative filtering algorithms. Note that we use the same default hyperparameters as Kowald et al. [122] and Naghiaei et al. [159].

4.4.3 Division of users in groups

We divide the users in groups in the following ways:

1. ***PopularPercentage***: Divide the users based on the percentage of items in their profile that have the label "popular" with a 20%-60%-20% scheme, in the same way as Abdollahpouri et al. [9] and Naghiaei et al. [159]. An item gets the label "popular" if it is one of the 20% most frequently rated items in the dataset.

2. ***AveragePopularity***: Divide the users based on the average popularity of items in their profile with a 20%-60%-20% scheme. This approach is not followed by any of the studies. Given that it may result into differently divided groups, and as it also does not depend on items being labeled "popular", it is interesting to investigate how it impacts the results in terms of whether certain user groups are unfairly treated by a recommender.

3. ***Mainstreaminess***: Divide the users based on the mainstreaminess score in the same way as Kowald et al. [122]. This division is only tried on *LastFM*, given that the mainstreaminess score can only be calculated on this dataset.

Note that while each study used domain-specific terms to refer to each user group, for simplicity we will use the terms *Niche*, *Diverse* and *Blockbuster-focused* regardless the user division and dataset.

4.4.4 Evaluation strategy

We apply all different evaluation strategies adopted by the studies. Specifically, the strategies vary with regards to which user and item are candidates for generating recommendations.

1. ***Modified TrainItems*** (Naghiaei et al. [159]):

   (a) Recommend items to every user in the training set.

   (b) Choose out of all items.

2. ***UserTest*** (Kowald et al. [122]):

(a) Recommend items to every user in the test set.

(b) Choose out of all items the user has rated in the test set.

3. ***TrainItems*** (Abdollahpouri et al. [9]):

(a) Recommend items to every user in the test set.

(b) Choose out of all items the user has not rated in the training set.

4.5 Results

In this section, we present the results across all popularity bias metrics for every experiment. First, we discuss overall popularity bias propagation, and then we focus on propagation per user group.

4.5.1 Overall Popularity Bias Propagation

To evaluate overall popularity bias propagation for every combination of aspects, we report on correlation and coverage as described in Section 4.4. Tables 4.5, 4.6, and 4.7 show the above mentioned values for each algorithm and each evaluation strategy, for *MovieLens1M*, *LastFM* and *Book-Crossing* respectively. In addition to the correlation and coverage tables, we choose one algorithm, namely *NMF*, and plot frequency in profile versus in recommendation for every dataset and evaluation strategy. Figure 4.2 includes the resulting scatter plots. Given the large number of experiments, we include the scatter plots for the other algorithms in the Appendix, Figures 8.5 to 8.14. We analyze how the different aspects affect the results. The type of user division in groups is not relevant here as it does not relate to whether an algorithm overall propagates popularity bias.

Table 4.5: Correlation and item coverage for the MovieLens1M dataset.

	Modified TrainItems		**UserTest**		**TrainItems**	
	Correlation	Coverage	Correlation	Coverage	Correlation	Coverage
UserKNN	**-0.0373**	0.0124	0.7913	0.6654	**-0.0375**	0.0108
ItemKNN	-0.0261	0.5229	0.9028	0.6832	-0.0439	0.4509
UserKNN w/ means	-0.0821	0.0688	0.8012	0.6805	-0.0838	0.0618
BPR	**0.4121**	**0.0216**	0.8448	0.5901	**0.4818**	**0.0586**
MF	0.1685	0.5982	0.8841	0.7337	0.1080	0.5990
PMF	0.3429	0.2415	0.8399	0.6786	0.2761	0.2515
NMF	**-0.0456**	0.1012	0.7977	0.6646	**-0.0474**	0.1012
WMF	0.3313	0.0157	0.8114	0.5923	0.3656	0.0229
HPF	**0.6524**	**0.1058**	0.8928	0.6141	**0.7535**	**0.1743**
NeuMF	**0.4864**	**0.0583**	0.8652	0.5947	**0.5657**	**0.1033**
VAECF	**0.6312**	**0.1913**	0.8906	0.6190	**0.7240**	**0.2715**
Mean	0.2576	**0.1762**	0.8474	**0.6469**	0.2784	**0.1914**

Data

The resulting correlation and coverage vary across the three datasets. There is not a clear pattern of a dataset being more or less prone to popularity bias across all

Table 4.6: Correlation and item coverage for the LastFM dataset.

	Modified TrainItems		**UserTest**		**TrainItems**	
	Correlation	Coverage	Correlation	Coverage	Correlation	Coverage
UserKNN	**0.0537**	0.0796	0.6586	0.3773	**-0.0170**	0.0403
ItemKNN	**0.5614**	0.6686	0.8881	0.7338	**-0.0178**	0.7097
UserKNN with means	0.1821	0.1422	0.6732	0.5031	-0.0259	0.0694
BPR	**0.3823**	**0.0094**	0.8192	0.2065	**0.4556**	**0.0177**
MF	-0.0001	0.0008	0.9039	0.7095	-0.0034	0.0351
PMF	-0.0081	0.0009	0.6223	0.2983	-0.0083	0.0009
NMF	**-0.0412**	0.0133	0.7565	0.5489	**-0.0413**	0.0144
WMF	0.2999	0.1981	0.8191	0.4063	0.2621	0.2062
HPF	**0.5819**	**0.0383**	0.8390	0.3407	**0.6836**	**0.0878**
NeuMF	**0.4844**	**0.0213**	0.8385	0.2332	**0.5586**	**0.0356**
VAECF	**0.6402**	**0.0866**	0.8574	0.3006	**0.7376**	**0.1256**
Mean	0.2851	**0.1145**	0.7887	0.4235	0.2349	0.1221

Table 4.7: Correlation and item coverage for the Book-Crossing dataset.

	Modified TrainItems		**UserTest**		**TrainItems**	
	Correlation	Coverage	Correlation	Coverage	Correlation	Coverage
UserKNN	**0.0407**	0.2672	0.9121	0.8130	**0.0385**	0.2588
ItemKNN	0.7249	0.9835	0.9115	0.8138	0.7275	0.9766
UserKNN with means	0.0454	0.3658	0.9105	**0.8148**	0.0421	0.3569
BPR	**0.4554**	**0.0014**	0.9208	0.7992	**0.4626**	**0.0023**
MF	0.0507	0.2177	0.9105	0.8152	0.0498	0.1608
PMF	0.1454	0.0673	0.9142	0.8103	0.1406	0.0657
NMF	**-0.0364**	0.2237	0.9124	0.8138	**-0.0365**	0.2137
WMF	0.8312	0.8951	0.9195	0.8047	0.8304	0.7180
HPF	**0.8362**	**0.0714**	0.9174	0.8067	**0.8410**	**0.0923**
NeuMF	**0.4202**	**0.0014**	0.9209	0.8000	**0.4244**	**0.0023**
VAECF	**0.7338**	**0.0864**	0.9177	0.8065	**0.7418**	**0.0971**
Mean	**0.3861**	**0.2892**	**0.9152**	**0.8089**	**0.3875**	**0.2677**

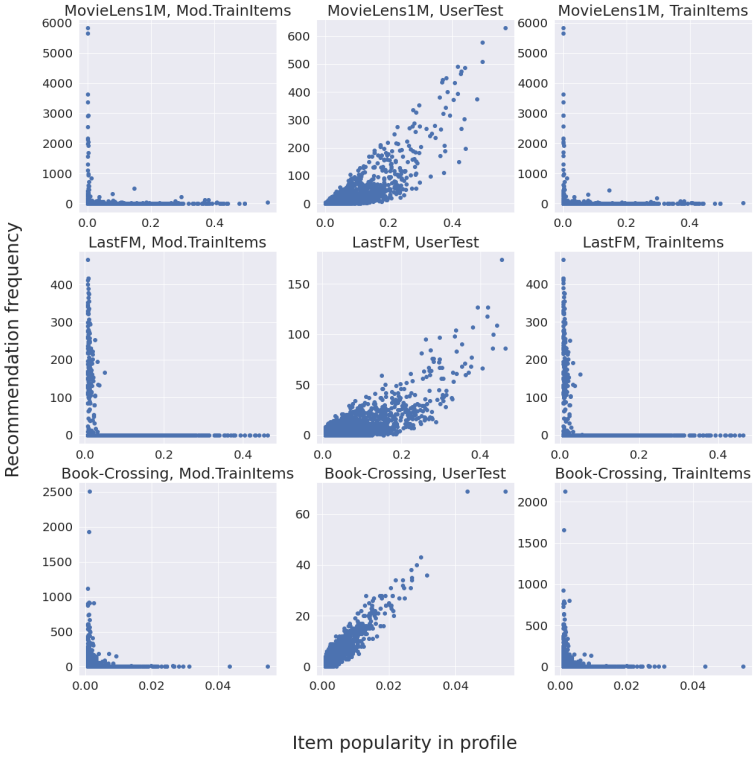


Figure 4.2: Item popularity in profile versus frequency of recommendation by the algorithm NMF, for every dataset and evaluation strategy.

algorithms. However, we notice in Table 4.7 that the correlation averaged over all algorithms is higher than in Tables 4.5 and 4.6 for a given evaluation strategy. In other words, when trained on the *Book-Crossing* dataset, the algorithms on average result in recommended lists where item frequency is more highly correlated with item popularity than when trained on the other two datasets. By this metric, the subset of *Book-Crossing* used by Naghiaei et al. [159] is on average more prone to popularity bias than *MovieLens1M* and *LastFM*.

Table 4.7 shows that the mean item coverage is also higher for *Book-Crossing* than for the other two datasets in Tables 4.5 and 4.6 for a given evaluation strategy. For example, given *Modified TrainItems*, the mean item coverage for Book-Crossing is 0.2892, for *MovieLens1M* 0.1762, and for *LastFM* 0.1145. Looking at Figure 4.1 and Table 4.4, we see that *Book-Crossing* is very sparse and popularity is scattered. Even very popular books have been rated by less than 350 users, in contrast with *MovieLens1M* where some movies have been rated by more than half of the users. It is therefore expected that item coverage is higher, as popularity is less concentrated. Note also that when *UserTest* is applied, the item coverage is higher for *Book-Crossing* than for *MovieLens1M*, and especially for *LastFM*, not just on average but for every individual algorithm. These observations lead us to wonder whether high correlation is sufficient to conclude popularity bias propagation, when it is not combined with low item coverage.

Algorithms

The algorithms propagate popularity bias in different degrees. Tables 4.5 to 4.7 show that *BPR*, *HPF*, *NeuMF* and *VAECF* consistently result in relatively high positive correlation, as well as low item coverage for *TrainItems* and *Modified TrainItems*. In this sense, *BPR*, *HPF*, *NeuMF* and *VAECF* are consistently prone to popularity bias. Note that these algorithms were only tested by Naghiaei et al. [159], which means that the choice of algorithms may have a large impact on whether popularity bias will be observed by a study. On the other hand, Figure 4.2 shows that *NMF* shows no correlation for any dataset for *TrainItems* and *Modified TrainItems*, as is the case for *UserKNN*.

Evaluation strategy

Evaluation strategy has a large impact on the reported result. Across all datasets and algorithms, there is a strong positive correlation between popularity in profile and in recommendation when *UserTest* is employed. This observation is in line with the conclusions of Kowald et al. [122] who used *UserTest* in their study, as well as their follow up paper [121]. This is due to *UserTest* only recommending to a user items that they have already consumed in the test set. For example, if a test user has rated 8 items in the test set, only these items are candidate for recommendation and they will all be recommended to that user given a 10-item recommendation. In this case, it is reasonable that popularity in profile and in recommendation correlate; popular items are more likely to be in the users' test sets and therefore be candidate for recommendation.

It is also the case that for every dataset item coverage is on average higher when *UserTest* is employed. For example, Table 4.5 shows that for *MovieLens1M* average coverage is 0.6469 when *UserTest* is employed, while only 0.1762 and 0.1914 when

Modified TrainItems and *TrainItems* are employed respectively. Given that candidates for recommendation are only the items from a user's test set, it follows that more items will be covered by the recommendation process as users have different tastes. This is especially the case for *Book-Crossing* given its sparsity. The consistency in results when *UserTest* is deployed prompts to consider that the popularity bias reported might mostly be a result of the evaluation process instead of the algorithm's functionality or data characteristics.

For *TrainItems* and *Modified TrainItems*, the results fluctuate per algorithm. Figure 4.2 shows that for all datasets, *NMF* propagates popularity bias when evaluated with *UserTest*, but does not with the other two strategies. *TrainItems* and *Modified TrainItems* show similar results overall. However, in some cases the fact that *TrainItems* excludes items that the user has rated in the training set does impact the conclusion on whether popularity bias is propagated. For example, *ItemKNN* trained on *LastFM* shows no correlation when *TrainItems* is employed, but positive correlation with *Modified TrainItems*.

4.5.2 Popularity Bias Propagation per User Group

To assess popularity bias propagation per user group, we report on the $\%\Delta GAP$ metric and on $NDCG@10$, as described in Section 4.4. Given the large number of experiments, we choose to present the results on *LastFM* since all three ways of dividing in user groups were possible on this dataset. Tables 4.8, 4.9 and 4.10 show the $\%\Delta GAP$ value for each user group, algorithm and user division, using *Modified TrainItems*, *UserTest* and *TrainItems* respectively. Tables 4.11, 4.12 and 4.13 show $NDCG@10$ for each user group, algorithm and user division, using *Modified TrainItems*, *UserTest* and *TrainItems* respectively. The results on the other datasets are included in the Appendix, Tables 8.3 to 8.14.

Table 4.8: $\%\Delta GAP$ for every user group, algorithm and way of dividing users, when *Modified TrainItems* is used. *N*, *D* and *BF* signify *Niche*, *Diverse* and *Blockbuster-focused*. ** indicates that the corresponding result is significantly higher than for the two other groups, while * indicates that it is significantly higher than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	Modified TrainItems								
	PopularPercentage			**AveragePopularity**			**Mainstreaminess**		
	N	D	BF	N	D	BF	N	D	BF
UserKNN	-32.6*	-32.6*	-47.8	-27.6*	-30.9*	-52.7	-53.2	-28.2**	-30.4*
ItemKNN	-20.8	-13.4*	-28.6	-17.0*	-11.6*	-33.3	-17.4	-13.6	-24.2
UserKNN with means	6.3**	-6.4*	-25.7	10.9**	-4.7*	-30.0	-34.7	-4.6*	6.0**
BPR	**596.6****	**352.6**	240.0	**611.0****	**358.2**	230.8	**379.8***	**355.4***	319.4
MF	**-49.0**	**-68.9**	**-78.3**	**-47.1**	**-68.6**	**-79.1**	**-67.6**	**-68.1**	**-71.4**
PMF	-69.7	-81.5	-87.1	-68.6	-81.3	-87.6	-80.8	-81.1	-83.1
NMF	-78.6*	-87.1*	-90.9	-77.8*	-87.0	-91.1	-86.1**	-86.8*	-88.4
WMF	**-2.8**	14.2**	1.2	**0.6**	17.8**	-7.0	-30.4	20.0*	32.5**
HPF	**150.4***	143.8*	125.2	**166.0****	148.6*	110.1	93.2	153.2*	**167.6****
NeuMF	**399.1****	**242.3***	157.3	**399.6****	**243.9***	159.2	**262.2***	**239.0***	213.4
VAECF	126.6	181.0**	156.3*	125.1	182.7**	153.7*	156.6	175.8**	169.4*

Table 4.9: %Δ*GAP* for every user group, algorithm and way of dividing users, when *UserTest* is used. *N*, *D* and *BF* signify *Niche*, *Diverse* and *Blockbuster-focused*. ** indicates that the corresponding result is significantly higher than for the two other groups, while * indicates that it is significantly higher than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	UserTest								
	PopularPercentage			**AveragePopularity**			**Mainstreaminess**		
	N	D	BF	N	D	BF	N	D	BF
UserKNN	21.9	35.2*	31.2*	23.7	36.7**	27.6	19.6	37.9*	39.0*
ItemKNN	-2.1	2.8	17.6**	-3.4	2.9	17.4**	9.8*	2.2	7.2
UserKNN with means	16.5	31.7*	29.1*	18.2	33.1**	25.6	9.1	33.3*	43.1**
BPR	**171.6***	**168.7***	108.1	**181.9***	**172.6***	98.4	**136.3**	**178.2****	142.0
MF	**-0.4**	**-0.1**	**0.7**	**1.0**	**0.6**	**-1.2**	**0.2**	**-0.5**	**0.6**
PMF	57.2	74.0**	62.3	59.7	77.4**	54.7	52.6	77.5*	75.0*
NMF	16.9	21.6	30.1**	15.7	23.1	27.2*	16.8	22.2	30.4**
WMF	20.6	40.2**	31.3	23.1	41.9**	27.0	13.7	42.7*	47.8*
HPF	61.3	89.9**	72.1	71.4	92.7**	62.7	52.8	96.3*	93.3*
NeuMF	**123.8***	**137.1***	90.6	**131.3***	**139.2***	84.9	**112.4**	**140.5****	114.8
VAECF	63.7	114.8**	87.1*	69.7	117.0**	80.5*	84.4	116.9**	100.7*

Table 4.10: %Δ*GAP* for every user group, algorithm and way of dividing users, when *TrainItems* is used. *N*, *D* and *BF* signify *Niche*, *Diverse* and *Blockbuster-focused*. ** indicates that the corresponding result is significantly higher than for the two other groups, while * indicates that it is significantly higher than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	TrainItems								
	PopularPercentage			**AveragePopularity**			**Mainstreaminess**		
	N	D	BF	N	D	BF	N	D	BF
UserKNN	-57.9**	-82.0*	-87.3	-56.1**	-81.8**	-87.9**	-80.2	-81.6	-80.1
ItemKNN	-57.6**	-70.0*	-77.8	-56.8**	-70.6*	-76.5	-49.4**	-75.6*	-85.2
UserKNN with means	-51.8**	-81.1*	-87.3	-49.4**	-81.1*	-87.7	-77.5*	-81.2	-79.3
BPR	**580.6****	**323.6**	214.0	**596.4****	**328.1**	206.5	**362.1****	**325.0***	288.2
MF	**-53.6****	**-70.4***	**-79.8**	**-51.7****	**-70.1***	**-80.7**	**-70.5***	**-69.5***	-72.9
PMF	-70.1*	-81.8*	-87.2	-69.0	-81.6	-87.6	-81.0	-81.3	-83.2
NMF	-78.6	-87.1	-90.9	-77.8*	-87.0	-91.1	-86.1	-86.8	-88.4
WMF	-8.2	-5.0*	-19.5	-5.6*	-2.9*	-24.3	-35.7	-0.8	5.8**
HPF	**120.4****	100.1*	86.8*	**132.1****	103.4*	76.4	65.5	106.1*	**121.7****
NeuMF	**382.6****	**215.3***	138.1	**384.5****	**216.5***	140.3	**244.7****	**211.8***	187.9
VAECF	118.6	158.0**	133.4*	117.8	159.3**	131.3*	145.6	153.0	140.7

Table 4.11: *NDCG@10* for every user group, algorithm and way of dividing users, when *Modified TrainItems* is used. *N*, *D* and *BF* signify *Niche*, *Diverse* and *Blockbuster-focused*. ** indicates that the corresponding result is significantly lower than for the two other groups, while * indicates that it is significantly lower than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	Modified TrainItems								
	PopularPercentage			**AveragePopularity**			**Mainstreaminess**		
	N	D	BF	N	D	BF	N	D	BF
UserKNN	0.007	0.009	0.002*	0.010	0.007	0.004	0.010	0.005	0.007
ItemKNN	0.007	0.009	0.008	0.008	0.007	0.011	0.013	0.007	0.005
UserKNN with means	0.009	0.007	0.001*	0.011	0.006	0.002	0.012	0.004*	0.002*
BPR	**0.188****	0.325	0.352	**0.188****	0.326	0.350	**0.251****	0.335	0.323
PMF	0.013	0.012	0.003**	0.013	0.012	0.004**	0.015	0.012	0.006*
NMF	0.006	0.008	0.006	0.007	0.008	0.005	0.007	0.010	0.004
WMF	0.200	0.222	0.213	0.206	0.227	0.192	0.155	0.242	0.251
HPF	**0.278****	0.340	0.359	**0.292****	0.335	0.357	**0.260****	0.355	0.378
NeuMF	**0.209****	0.339*	0.393	**0.219****	0.337*	0.390	**0.271****	0.350	0.351
VAECF	**0.297****	0.376	0.396	**0.294****	0.374	0.405	**0.326****	0.379	0.388

Table 4.12: *NDCG@10* for every user group, algorithm and way of dividing users, when *UserTest* is used. *N*, *D* and *BF* signify *Niche*, *Diverse* and *Blockbuster-focused*. ** indicates that the corresponding result is significantly lower than for the two other groups, while * indicates that it is significantly lower than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	UserTest								
	PopularPercentage			**AveragePopularity**			**Mainstreaminess**		
	N	D	BF	N	D	BF	N	D	BF
UserKNN	0.669	0.648	0.640*	0.668	0.646	0.645	0.643*	0.636*	0.672
ItemKNN	0.621*	0.620*	0.659	0.621*	0.620*	0.661	0.619*	0.622*	0.643
UserKNN with means	0.671	0.656	0.657	0.671	0.656	0.658	0.648*	0.647*	0.684
BPR	0.655	0.668	0.677	0.651*	0.667	0.683	**0.625****	0.670*	0.706
PMF	0.665	0.646	0.662	0.661	0.644*	0.672	0.635*	0.642*	0.682
NMF	0.642	0.632	0.646	0.635	0.634	0.649	0.640	0.626	0.645
WMF	0.678	0.664	0.657	0.673	0.664	0.661	0.655*	0.654*	0.687
HPF	0.710	0.708	0.720	0.706	0.711	0.716	**0.673****	0.707*	0.752
NeuMF	0.657*	0.675	0.690	0.656*	0.675	0.690	**0.633****	0.680*	0.710
VAECF	0.695	0.693	0.707	0.693	0.694	0.705	**0.663****	0.694*	0.730

Table 4.13: *NDCG@10* for every user group, algorithm and way of dividing users, when *TrainItems* is used. *N*, *D* and *BF* signify *Niche*, *Diverse* and *Blockbuster-focused*. ** indicates that the corresponding result is significantly lower than for the two other groups, while * indicates that it is significantly lower than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	TrainItems								
	PopularPercentage			**AveragePopularity**			**Mainstreaminess**		
	N	D	BF	N	D	BF	N	D	BF
UserKNN	0.010	0.015	0.002*	0.014	0.012	0.006	0.015	0.010	0.010
ItemKNN	0.013	0.019	0.028	0.013*	0.018	0.032	0.025	0.018	0.016
UserKNN with means	0.015	0.014	0.004**	0.020	0.012	0.004**	0.019	0.011	0.006*
BPR	**0.236****	0.456	0.479	**0.242****	0.454	0.479	0.324**	0.460	0.466
PMF	0.013	0.013	0.003**	0.013	0.012	0.004*	0.015	0.012	0.006*
NMF	0.006	0.008	0.006	0.007	0.008	0.005	0.007	0.010	0.004
WMF	0.307	0.301	0.282	0.309	0.314	0.242**	0.225**	0.333	0.338
HPF	**0.389****	0.507	0.544	**0.409***	0.505*	0.530	**0.369****	0.532*	0.572
NeuMF	**0.279****	0.489*	0.550	**0.300****	0.488*	0.533	**0.366****	0.505	0.508
VAECF	**0.448****	0.556*	0.594	**0.458***	0.555*	0.587	**0.469****	0.565	0.591

Algorithms

It is apparent in Tables 4.8 to 4.10 that certain algorithms consistently produce recommended lists with higher average popularity, while others do not. *BPR*, *HPF*, *NeuMF* and *VAECF* showcase high $\%\Delta GAP$ values for all groups, regardless the evaluation strategy. For example, Table 4.8 shows that when *Modified TrainItems* is deployed and users are divided in groups based on *PopularPercentage* or *AveragePopularity*, *NeuMF* recommends to *Niche* users items almost 4 times more popular than they have rated in their profiles (i.e., the $\%\Delta GAP$ is 399.1 and 399.6 respectively). Note that the same algorithms consisently result in high correlation and low coverage, as discussed in Section 4.5.1.

Tables 4.11 and 4.13 show that given *Modified TrainItems* and *TrainItems*, *BPR*, *HPF*, *NeuMF* and *VAECF* also result in high $NDCG@10$ regardless the user division. However, the $NDCG@10$ values are significantly lower for the *Niche* group. We can conclude that these algorithms tend to recommend mostly popular items to all users when trained on *LastFM*, at least when their default hyperparameters are used. On the contrary, *MF* mostly results in negative $\%\Delta GAP$ values across Tables 4.8 to 4.10, meaning that the average popularity in the recommended lists is reduced compared to the users' profiles. Finally, some algorithms like *WMF* are inconsistent when it comes to the average popularity of the items they recommend to each group.

Additionally, Tables 4.8 to 4.10 show that *BPR* and *NeuMF*'s recommendations generally result to higher $\%\Delta GAP$ for the *Niche* and *Diverse* groups compared to the *Blockbuster-focused* group. However, other algorithms do not showcase such consistency, as $\%\Delta GAP$ fluctuates across evaluation strategies and user divisions. Consequently, it is challenging to deduce that a set of algorithms tend to treat *Niche* users unfairly while others do not. It might be that such conclusions can be drawn given a specific context, and not about an algorithm's inherent functionality.

User division in groups

The way the users are divided in groups in some cases determines whether the *Niche* user group is unfairly treated, as well as to what extent. For example, we see in Tables 4.8 and 4.10 that for *Modified TrainItems* and *TrainItems*, *HPF* results in higher $\%\Delta GAP$ for *Niche* users with *PopularPercentage* and *AveragePopularity*, whereas with *Mainstreaminess* the *Blockbuster-focused* group receive the highest increase in average popularity. Additionally, *PopularPercentage* and *AveragePopularity* also sometimes result in different conclusions on which group is unfairly treated. Table 4.8 shows that given *Modified TrainItems*, *WMF* recommends more or less popular items to the *Niche* group depending on which user division is assumed ($\%\Delta GAP$ of -2.8 for *PopularPercentage* and 0.6 for *AveragePopularity*).

PopularPercentage and *AveragePopularity* mostly result in similar trends in terms of which group receives unfair treatment. On the other hand, *Mainstreaminess* often leads to different conclusions than *PopularPercentage* and *AveragePopularity*. This is somewhat expected since *Mainstreaminess* is not dependent on the propensity for popularity within the given data, but the *mainstreaminess score* that characterizes the underlying population where this subset of *LastFM* stems from. Therefore, in order to conclude unfair treatment of a group, it is necessary to define the characteristics of the group in

question and whether we refer to specific behavior within the training dataset or we incorporate external information as well.

Evaluation strategy

As is the case for overall popularity bias propagation, the choice of evaluation strategy largely influences $\%\Delta GAP$. When *UserTest* is employed, almost all algorithms provide recommendations with increased average popularity compared to user profile for all user groups regardless the type of division (see Table 4.9). Even in cases where $\%\Delta GAP$ is negative, the decrease is small. Overall, when *UserTest* is deployed, the $\%\Delta GAP$ values do not vary between algorithms as much as given the other two evaluation strategies.

On the other hand, when *Modified TrainItems* and *TrainItems* are used, the results fluctuate between algorithms. Furthermore, the recommended lists of the algorithms which consistently perpetuate popularity bias for all groups (*BPR*, *HPF*, *NeuMF* and *VAECF*) showcase higher $\%\Delta GAP$ with *Modified TrainItems* compared to *TrainItems* across all types of user divisions (see Tables 4.8 and 4.10). This likely signifies that excluding items that a user has rated from the candidate list reduces popularity bias. Given that by definition many users have rated the popular items, then the popular items are excluded from many users' candidate items list when *TrainItems* is deployed.

The choice of evaluation strategy also has a large effect on *NDCG*@10. Table 8.13 shows that *UserTest* results in higher *NDCG*@10 than the other two strategies. Since the pool of candidate items is limited to a user's test items when *UserTest* is deployed, it follows that the recommended list will likely be similar to the 'ideal' list. Additionally, the produced *NDCG*@10 values are generally higher for *TrainItems* than for *Modified TrainItems* across algorithms and user division in groups. This is due to the exclusion of training items from a user's candidate list when *TrainItems* is deployed. In our calculation of *NDCG*@10, we only considered items to have a nonzero utility for the user if they were consumed unbeknownst to the system, i.e., the test items (see [70]).

4.6 Discussion and future work

The results show that data, algorithms, user division in groups, and evaluation strategy influence the conclusion on whether popularity bias is propagated and whether the users that have lesser propensity for popular items are disproportionately affected. The datasets have impactful differences, especially when it comes to distribution of item popularity; *Book-Crossing* is a very sparse dataset, which results in higher item coverage on average compared to the other two datasets, even when correlation between popularity in profile and in recommendation is also high. At the same time, the choice of which algorithms to include in the study influences the results; some matrix factorization algorithms (*HPF*, *NeuMF*, *VAECF*) consistently propagate popularity bias, and in most cases recommend disproportionately many popular items to the *Niche* users. User division in groups often defines whether unfairness can be concluded, especially when a proxy measure of propensity towards popular items is used that does not stem from the specific training dataset. Finally, evaluation strategy, and specifically the generation of user and item candidates is overall crucial given the effect it has on

whether both overall popularity bias propagation and user group unfairness can be observed.

The results indicate that perceived propagation of popularity bias is sensitive to various aspects of the evaluation process that are often unaddressed. The fact that tweaking these aspects determines whether propagation of popularity bias can be concluded or not renders the individual conclusions of the studies unique to their context and set up, and less generalizable. We conclude that it is challenging to report on popularity bias as a phenomenon that persists as a result of an algorithm's functionality. While some algorithms we studied do consistently propagate popularity bias given our metrics, for most algorithms the results largely fluctuate depending on the other aspects. Similarly, even though the choice of dataset does have an impact, we showed that the divergence in results across the reproduced studies could not be solely attributed to the datasets' characteristics.

Consequently, we recommend careful consideration of each of these aspects when popularity bias is evaluated. Future studies on the topic should reflect on:

- **Data**: Which domain does the data come from? What is the size of the dataset? How sparse is it? What is the long-tail item distribution?
- **Algorithms**: What type of optimization is performed? How sensitive is it to item popularity?
- **User division in groups**: How is user propensity for popular items defined in this domain? What characteristics would deem a user *Niche*? Is behavior in the dataset the relevant factor, or is external information needed?
- **Evaluation strategy**: Which exact aspect of popularity bias is the study evaluating? How can it be translated into choices for every step of the evaluation process?

Evaluation in the case of recommender systems is generally challenging. Offline evaluation, in particular, is restricted by the lack of data or lack of absolute ground truth. The fact that a user has rated an item highly does not necessarily mean that they prefer it to an unrated one; it can simply mean lack of awareness. In other words, high accuracy in offline evaluation does not equate to a successful system. As a result, the choice of evaluation strategy is often application-specific, which extends further than the choice of metrics. For example, the same algorithm might be differently evaluated when used in an application which recommends 5 items to a user instead of 10. Some metrics can only be evaluated with certain strategies because of the data they require. For example, Mean Absolute Error can only be calculated for user-item pairings that do exist in the dataset and thus *UserTest* would be appropriate in this case. However, other metrics leave room for different choices and there might not be one right way to evaluate them.

To appropriately evaluate a metric, it is important to consider what is the exact phenomenon it is intended to measure. While popularity bias is not a recent topic within recommender systems research, there is a certain lack of specificity around the exact meaning of it. Correlation, coverage, and difference in average popularity can

all be useful in measuring some sides of popularity bias, among others. Having said that, their interpretation requires careful design of an evaluation process that answers the question we are wondering about. In the case of the three studies each evaluation strategy answers a different question:

- *Modified TrainItems* does not disregard the items a user has already consumed from the recommendation process. Such strategy is unsuitable when assessing general performance, as it leads to information leakage. It could be used to assess popularity bias in a system that does recommend to users items that they have already consumed and positively rated. However, the information leakage needs to be considered and accounted for.

- *UserTest* only recommends to a user items that they have already consumed unknowingly to the system (i.e., from the test set), which renders it inappropriate for evaluating overall popularity bias. However, it might still be interesting to observe whether the items a user has rated and are unseen to the system are differently ranked by the system than by the user, because of popularity bias.

- *TrainItems* measures whether a learned model propagates popularity bias into unseen user-item combinations. It is our belief that this strategy is the most appropriate to measure popularity bias, as it more closely resembles a real world scenario of learning the preferences of the users and recommending items to them that they have not yet rated.

Studies on the topic of popularity bias should account for these differences in evaluation strategy by specifying the research question, as well as the evaluation strategy that accompanies it. To ensure reproducibility of such studies, a way ahead could be to adapt the dimensions of evaluation benchmarked by Said and Bellogín [196] into a checklist for submissions in conferences and journals. Such an initiative can motivate researchers to critically think about the evaluation strategy they employ in their experimentation, and allow for easier comparison between the reported results of different studies.

Our study has limitations that future work should address. It is our intention to approach the phenomenon of popularity bias fundamentally by locating specific data and recommendation characteristics that instigate its propagation. While the aspects we identified through reproducing these studies are very important, there are additional ones we plan to consider. First, our reported results on UserKNN differ significantly from Abdollahpouri et al. [9] and Naghiaei et al. [159], even when the same evaluation strategy and data are used. The reason is presumably tuning and implementation; Abdollahpouri et al. [9] state that they tuned all the algorithms to reach similar precision in order to appropriately compare them. On the other hand, Naghiaei et al. [159] manually trained UserKNN instead of using Cornac like they did for the other algorithms. This prompts us to consider tuning an important aspect of the process as well, and future work should focus on the effect it has on the propagation of popularity bias. In practice, algorithms are designed to satisfy some accuracy metric instead of being trained with their default hyperparameters, and thus it is realistic to assume that tuning takes place. Second, by measuring propagation of popularity bias with cross-validation instead of one-shot prediction, we could account for random

small differences between algorithms, and generalize their comparison. Finally, all three studies used different Python libraries to perform the recommendation process. For future work, we plan to explore the effect of potentially different implementations of the same algorithm across these libraries.

4.7 Conclusion

In this chapter, we reproduced the analysis on the propagation of popularity bias by commonly used collaborative filtering algorithms performed by three studies using different datasets from the media domain. The studies evaluated overall popularity bias propagation, as well as whether users with niche tastes were unfairly treated by the system. The results reported differed for both evaluation tasks. We identified four aspects which varied across the three studies and could potentially account for the divergence in results: data, algorithms, division of users in groups, and evaluation strategy. We designed and carried out experiments to investigate to what extent each aspect impacted the results by combining all possible choices for each aspect. We found that all aspects affected the result to some degree. Evaluation strategy specifically largely accounted for the divergence, as the one employed by the study in the music domain resulted in reporting propagation of popularity bias for all datasets and all algorithms. We conclude that clarity around the evaluation strategy employed during the recommendation process is necessary to reproduce and compare analysis for different algorithms and datasets, especially for phenomena like popularity bias whose evaluation is not standardized in literature yet.

5

The challenges of studying bias in recommender systems

Building on our reproduction work (Chapter 4), we deepen our investigation into how data characteristics and algorithm configurations jointly affect bias measurement. In this chapter, we systematically explore the sensitivity of bias conclusions to these design choices.

Statements on the propagation of bias by recommender systems are often hard to verify or falsify. Research on bias tends to draw from a small pool of publicly available datasets and is therefore bound by their specific properties. Additionally, implementation choices are often not explicitly described or motivated in research, while they may have an effect on bias propagation. In this chapter, we explore the challenges of measuring and reporting popularity bias. We showcase the impact of data properties and algorithm configurations on popularity bias by combining real and synthetic data with well known recommender systems frameworks. First, we identify data characteristics that might impact popularity bias, and explore their presence in a set of available online datasets. Accordingly, we generate various datasets that combine these characteristics. Second, we locate algorithm configurations that vary across implementations in literature. We evaluate popularity bias for a number of datasets, three real and five synthetic, and configurations, and offer insights on their joint effect. We find that, depending on the data characteristics, various configurations of the algorithms examined can lead to different conclusions regarding the propagation of popularity bias. These results motivate the need for explicitly addressing algorithmic configuration and data properties when reporting and interpreting bias in recommender systems.

This chapter is based on *Daniil et al., On the challenges of studying bias in Recommender Systems: The effect of data characteristics and algorithm configuration, Information Retrieval Research, 1(1), 3-27* [59].

5.1 Introduction

Recommender systems are commonly used as a tool to encode taste based on the information available, be it user history or metadata. The wide use of recommender systems necessitates critical reflection on the issues that may arise when we allow automation to dictate our exposure to information. Specifically, bias in recommender systems is a topic of interest within the scholarly community. Bias is a complex term that can refer to various types of biases associated with interactions between users and items in a given system [42].

Many studies have focused on measuring the phenomenon of *popularity bias* in collaborative filtering systems [3, 13, 71, 118, 246, 261]. Despite this large research effort to track and mitigate popularity bias, there is no univocal message regarding why and when it occurs. In previous work, we found that studies that measure popularity bias propagated by commonly used algorithms on benchmark datasets report varying, sometimes contradicting results [58]. This observation raises questions; is popularity bias sensitive to properties of the system that do not receive sufficient attention? Why is a seemingly simple phenomenon so hard to study? Additionally to the evaluation strategy which was the focus of our previous work, we hypothesize that two factors that also complicate bias measuring and reporting are data characteristics and algorithm configuration.

Data characteristics Benchmark datasets are useful for academic research, as they allow researchers to evaluate their hypotheses and compare their proposed debiasing methods. However, their consistent use raises concerns that relate to the dependence on the domain and source they were constructed from, and the potentiality for blind spots that stem from outdated rating behaviour. Most importantly, by reporting on types of bias on only a small set of publicly available datasets, researchers are restricted by their specific characteristics. This specificity limits the scope of research, and obfuscates the process of examining causality. In other words, it is not trivial to conclude whether the propagation of bias or lack thereof is a result of the respective algorithm's functionality, or of certain intricate details of the user-item interactions within these datasets. Data synthesis based on assumptions can potentially assist with shedding light on the aforementioned blind spots and gaining new insights into bias in recommender systems.

Algorithm configuration Insufficient reporting of algorithm configuration leads to a reproducibility problem within research on recommender systems. Studies have shown that papers published in big conferences often do not disclose sufficient information for replication and verification [75]. This issue is also relevant in the bias discussion. Even relatively simple algorithmic approaches, such as neighbour-based ones, are constructed using hyperparameters and implementation choices that might affect whether bias propagation is observed. The RecSys community proposes a set of evaluation frameworks to promote reproducibility[1], but there are important configuration differences between

[1] https://github.com/ACMRecSys/recsys-evaluation-frameworks

them that often go unmentioned [27]. Testing the effect of algorithm configuration can be a means of reporting on bias in a comprehensive manner.

In this chapter, we experiment with data characteristics and algorithm configurations and observe the effect on popularity bias. First, we look into *data characteristics* that might have an impact on popularity bias given a rating prediction and top-10 recommendation task, a common setup among recent studies on popularity bias [121, 122, 159]. Specifically, we delve into the relation between popularity and rating, as well as the preferences of users with large profiles. We analyze these characteristics for three real datasets: a subset of Book-Crossing [265] constructed by Naghiaei et al. [159], MovieLens1M [94] and Epinion [143]. Additionally, we form a set of data scenarios by tweaking and combining these characteristics. For each scenario, we generate a corresponding synthetic dataset of ratings, based on the interactions from the subset of Book-Crossing. Second, for a set of algorithms we identify *configuration choices* that may impact whether or not popularity bias is observed. We study three widely used algorithms, UserKNN, Biased Matrix Factorisation and Deep Matrix Factorization, implemented in three frameworks recommended by ACM RecSys, LensKit [67], Cornac [197], and Elliot [21]. We perform the recommendation process for the subset of Book-Crossing, MovieLens1M and Epinion, as well as each synthetic dataset with varied algorithm configuration choices. We apply commonly used popularity bias metrics to evaluate the recommended lists, as well as RMSE and NDCG@10 to estimate the performance of the algorithms when it comes to rating prediction and ranking.

Our results show that whether popularity bias is observed, and to what extent, depends on much more than the algorithm that was used, or the domain in which a study is carried out: all algorithms that we tested were seen to strongly propagate popularity bias in some experimental settings, while not propagating popularity bias in other settings. The same is true for datasets: all datasets led to bias in some settings and not in others. The results further clarify that whether or not popularity bias is observed depends on, firstly, specific (often unreported) configuration and implementation details of algorithms. The implication of that is that the choice for a certain framework largely impacts the outcome of a study. It also depends, secondly, on characteristics of the dataset that is studied; specifically, the relationship between rating and popularity, as well as the preferences of users with large profiles are crucial when it comes to popularity bias propagation. Finally, it is especially the interplay between algorithm configuration and dataset characteristics that determines whether popularity bias will be observed or not.

The contributions of this chapter are as follows:

- a systematic investigation into the effect of data characteristics on popularity bias, by comparing results on three commonly used datasets as well as five synthetic datasets for which we control the properties.

- a systematic investigation into the effect of implementation differences, by comparing results of algorithm configurations as well as non-configurable implementation differences in well known frameworks.

- for the more interpretable algorithms, we highlight notable results among the

many, to give insights into why certain combinations of dataset characteristics and algorithm configurations lead to popularity bias.

With this work, we wish to contribute to the field by highlighting and disentangling the challenges in studying popularity bias in recommender systems.

5.2 Related Work

In this section, we provide a brief overview of existing work on bias in recommender systems and datasets and reproducibility.

5.2.1 Bias in Recommender Systems

Recommender systems are not immune to bias, even when only user consumption history is fed to the model and not other information about the users or items. Edizel et al. [66] discuss that a model might learn sensitive information like the gender of the user in the latent space, and produce recommendations that are gender-dependent, even more so than the interactions observed in the training set itself. In a survey on the topic of bias and debias in recommender systems research, Chen et al. [42] identify three factors that contribute to bias: user behavior's dependence on the exposure mechanism of the system, imbalanced presence of items (and users) in the data, and the effect of feedback loops. One type of bias that arises from the interaction between an algorithm and imbalanced data is popularity bias.

Popularity bias is the phenomenon where popular items (i.e., items that are frequently interacted with in the dataset) are recommended even more frequently than their popularity would warrant [6]. It is commonly believed to be caused by the long-tail distribution that often characterizes user-item interactions: most items have been rated by only a few users, and a few items have been rated by many users [37]. Various studies have reported that frequently used recommender systems algorithms are prone to propagating popularity bias existing in the dataset they were trained on [9, 122, 159]. Different metrics have been proposed to quantify popularity bias [7, 9]. Despite the extensive literature, our understanding of why certain algorithms and datasets are more or less prone to popularity bias is limited. In a systematic work, Deldjoo et al. [61] used regression-based modeling to explain accuracy and fairness exhibited by a set of collaborative filtering algorithms through the lens of data characteristics such as rating and popularity distribution. In this chapter, we describe scenarios of data-algorithm interaction and report the results of different metrics associated specifically with popularity bias.

5.2.2 Datasets and Reproducibility

The way that recommender systems researchers usually test their hypotheses, novel algorithms, and metrics is by conducting experiments on one or more publicly available datasets of user-item interactions. Surveys on the topic of recommender systems research show that the pool of datasets used is small [32]; user behavior data from real-world applications such as media platforms is often proprietary and therefore cannot be used for benchmarking [116]. In popularity bias studies, the use of different versions of MovieLens is exceedingly common [236], which leads us to wonder whether studies

can be conclusive when they are carried out solely on a few datasets. In our approach, we include a data synthesis step that allows us to experiment with different data distributions and observe the result. Synthetic data serves multiple purposes, each with its own specific requirements and evaluation setup [209]. Data synthesis is a much-discussed topic in recommender systems research, often explored in the context of privacy [208, 224]. Additionally, studies have employed data simulation to measure bias in recommender systems under varying data properties [28].

The existence of datasets for training and testing is valuable for the recommender systems community; research on publicly available data is necessary in order to ensure reproducibility [196]. However, as noted by Cremonesi and Jannach [50], sharing the used data is not always sufficient to ensure basic reproducibility. Studies showed that in most cases recommender systems papers presented at top-tier conferences did not provide code for their data preprocessing or hyperparameter tuning [74, 75]. This is also the case in popularity bias research; studies are often not accompanied by code, and sometimes do not describe the data filtering or hyperparameter setting [9, 10]. Therefore, concluding that an algorithm or a dataset is prone to popularity bias becomes challenging, as it is not possible to verify or falsify the claims [50]. Research has shown that the choice of hyperaparameters can highly impact quality metrics, including average popularity of the items recommended [109]. At the same time, popular frameworks seem to have important differences that translate into different performances for the same datasets given somewhat different implementations of the same algorithm [27]. To further explore this issue, we experiment with algorithms implemented in different libraries and with different parameter configurations and evaluate popularity bias for each of them.

5.3 Identifying Data Characteristics and Algorithm Configurations

In this section, we identify data characteristics and algorithm configurations that can influence popularity bias propagation in the context of a rating prediction and top-10 recommendation task. First, we locate data characteristics that can have an effect on whether popularity bias is propagated, partly inspired by the functionality of UserKNN. UserKNN is a relatively simple algorithmic approach that simulates a 'word-of-mouth' setting and has lower dependence on non-intuitive parameters that impact optimization (e.g., learning rate). Accordingly, we form a set of data scenarios that combine the located characteristics. Second, we inspect UserKNN and two matrix factorization algorithms, one traditional and one neural network-based, and locate configurations that can be potentially impactful for popularity bias propagation.

5.3.1 Data Characteristics

Whether or not popularity bias is propagated depends on how popularity manifests in the dataset at hand. We discuss the relation between rating and popularity and the preferences of influential users.

Relation Between Rating and Popularity In the context of a rating prediction task, an algorithm aims to predict a future rating for every user of every item they have not already consumed. Given that this is done by considering the other users' ratings, it may be that items with high average rating will be prioritized by the system. Popularity bias studies often do not disclose whether the popular items in the dataset also have high ratings, but instead assume that their frequent recommendation is solely due to their popularity. Previous work has shown the impact of high ratings on popularity bias [246].

Influential Users In the context of UserKNN, certain users may be influential because they neighbour with many other users. For example, if only two users have rated an item and they have not rated any other items, then this item will not be recommended to anyone, because the two users are not influential at all within the system. Consequently, it is interesting to investigate the notion of user influence and whether the result is dominated by the preferences of users who, because of their large profile size, are more likely to have many neighbours.

Real Datasets

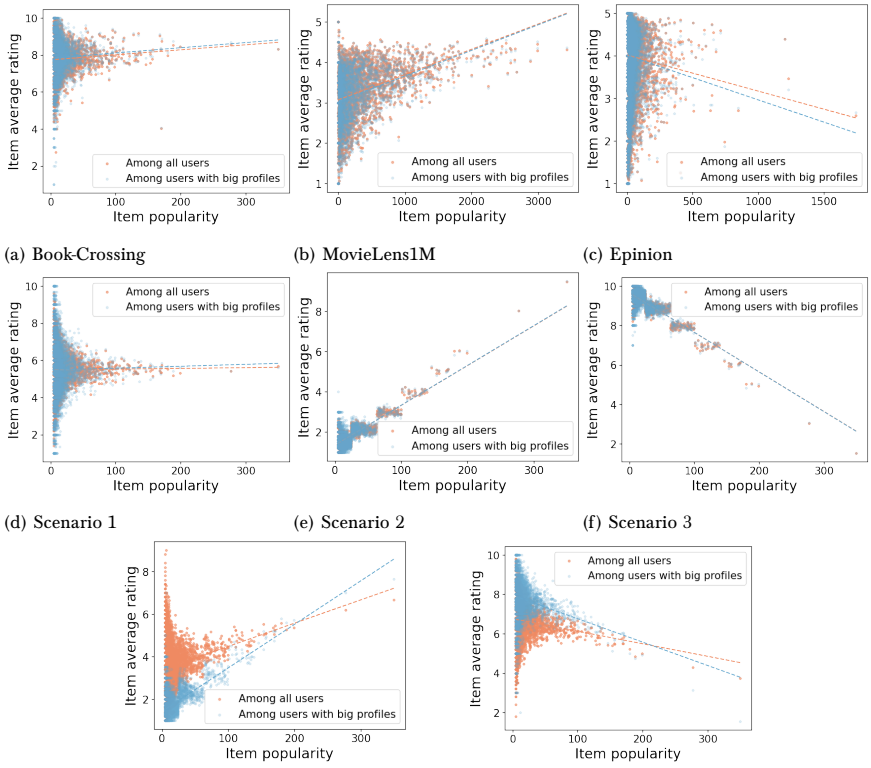
Figures 5.1a, 5.1b and 5.1c show the aforementioned characteristics for Book-Crossing, MovieLens1M and Epinion, respectively. Specifically, they show the correlation between item average rating and item popularity among all users, as well as among the users with the 20% largest profiles. We see that for Book-Crossing and MovieLens1M there is a slight positive correlation between rating and popularity, and a negative one for Epinion. Additionally, for the first two datasets there is no obvious difference between the tendencies of all users and the ones with large profiles, while the users with large profiles in Epinion are less favourable towards the popular items than the entire set of users. Table 5.1 shows the number of users, items, and interactions, and sparsity for each of the datasets.

Table 5.1: Basic characteristics of the real datasets. The synthetic datasets share the same characteristics as Book-Crossing.

Dataset	#users	#items	#ratings	Sparsity
Book-Crossing (subset)	6,358	6,921	88,552	99.80%
MovieLens1M	6,040	3,706	1,000,209	95.53%
Epinion	22,164	296,277	912,441	99.99%

Data Scenarios

We synthesize data that follows a long-tail distribution for items and users, as it is discussed as a prerequisite for popularity bias to occur [37, 39]. Specifically, we choose the interactions in a subset of the Book-Crossing dataset [159, 265] as a baseline, but remove the rating values. We choose Book-Crossing as a basis for our experimentation given the broader theme of book recommendation in this thesis and the fact that we have extensively investigated its sensitivity to bias. To reflect on the observations above, we form a set of scenarios around the relationship between popularity, rating and user



(a) Book-Crossing (b) MovieLens1M (c) Epinion

(d) Scenario 1 (e) Scenario 2 (f) Scenario 3

(g) Scenario 4 (h) Scenario 5

Figure 5.1: Relation between item average rating and item popularity among all users and among users with the 20% largest profiles, given three real and five synthetic datasets.

influence to assign a synthesized rating to each interaction. This approach allows us to simulate a real-world scenario where consumption is long-tail, while still experimenting with data properties relevant for popularity bias. We recognize that the scenarios are not necessarily realistic. User tendencies are likely to be more subtle in real world situations. However, we believe that experimenting with extreme behaviors can help us showcase the effect that we are investigating, and lead the way for more nuanced experimentation.

The scenarios, as well as the process we followed to generate each of them are as follows:

1. **Scenario 1**: **There is no relation between popularity and rating**: For each interaction, draw a rating value between 1 and 10 uniformly at random. In this case, the popularity of the item is not taken into account when a rating is generated.

2. **Scenario 2**: **Popular items are generally rated higher by the users**: For each interaction, draw a rating value between 1 and 10 from a normal distribution, where the mean is the popularity of the item normalized between 1 and 10. Since the mean of the rating distribution is the item's normalized popularity, more popular items tend to receive higher ratings.

3. **Scenario 3**: **Popular items are generally rated lower by the users**: For each interaction, draw a rating value between 1 and 10 from a normal distribution, where the mean is the opposite number of the popularity of the item normalized between 1 and 10. Since the mean of the rating distribution is the opposite number of the item's normalized popularity, more popular items tend to receive lower ratings.

4. **Scenario 4**: **Only users with large profiles rate popular items higher**: For each interaction, draw a rating value between 1 and 10 uniformly at random. For the users with the 20% largest profiles, replace by drawing from a Poisson distribution where the mean is the popularity of the item normalized between 1 and 10. While most users rate at random, users with large profiles tend to rate popular items higher.

5. **Scenario 5**: **Only users with large profiles rate popular items lower**: For each interaction, draw a rating value between 1 and 10 uniformly at random. For the users with the 20% largest profiles, replace by drawing from a Poisson distribution where the mean is the opposite of the popularity of the item normalized between 1 and 10. While most users rate at random, users with large profiles tend to rate popular items lower.

Figures 5.1d to 5.1h show the correlation between item average rating and item popularity within the five synthetic datasets. The effects are much more pronounced than for the real datasets. Scenario 1 shows no relation between average rating and popularity, as the ratings were drawn uniformly at random. Scenarios 2 and 3 showcase a very positive and a very negative correlation, respectively. For scenario 4, we see a positive correlation, which is higher for users with large profiles, and for scenario 5 a negative correlation, which is even lower for users with large profiles.

5.3.2 Algorithm Configurations

In this section, we describe the algorithm configurations examined for UserKNN and two matrix factorization algorithms, one traditional and one neural network based.

UserKNN

Despite UserKNN's simplicity, there are configuration choices that can potentially greatly influence the result. We identified the following: minimum similarity, the items considered for similarity, and minimum neighbours.

Minimum Similarity It is common in UserKNN implementations that not all users who rated an item are considered [62]. Instead, the notion of neighbourhood is introduced; only the users most similar to the target user are taken into account when producing a predicted score. The filtering can be done by introducing a cut off value of minimum similarity for consideration, among other techniques.

Items for Similarity Given a similarity metric (e.g., cosine similarity), a design choice still has to be made on whether similarity between two users will be calculated for their full rating vectors, or only for ratings on items these two users have in common. See [12] for clarification.

Minimum Neighbours When the neighbourhood is constructed for a given user, then a score is predicted for each item in a list of candidates. In some implementations, the predicted score is not calculated for all potential items. Instead, the algorithm focuses on items that have been rated by at least a minimum number of neighbours of the current user.

It is worth noting that LensKit and Cornac differ when it comes to these choices as seen in Table 5.2. Two of the parameters tested are not configurable in Cornac, while the third is not configurable in either framework and is set to a different value in each of them.

Table 5.2: Configuration choices related to UserKNN made by LensKit for Python and Cornac.

Configuration choice	LensKit for Python	Cornac
Minimum similarity	Configurable (default: 0)	Fixed to -1
Items for similarity	Fixed to all items	Fixed to common items
Minimum neighbours	Configurable (default: 1)	Fixed to 1

Traditional Matrix Factorization

In order to experimentally inspect the effect of implementation, we focus on a matrix factorization algorithm that is implemented by both LensKit and Cornac. Namely, we look into Biased Matrix Factorization (BMF).

BMF is a type of matrix factorization used for explicit rating prediction, and it is differently implemented in LensKit and Cornac. LensKit provides an alternating least squares implementation of BMF, and the default solver is coordinate descent [219] with

weighted regularization [262]. Cornac implements a stochastic gradient descent solver for BMF [119]. In both libraries, the Boolean parameter of bias is included to signify whether user and item bias are used to predict the ratings, and is by default set to True. Bias is an interesting parameter to investigate in the context of popularity bias propagation.

Deep Matrix Factorization

Along with BMF, we perform preliminary analysis on the popularity bias propagated by a neural network based matrix factorization algorithm, namely Deep Matrix Factorization (DMF). DMF uses a multi-layer perceptron to project users and items into a latent structured space [245]. It takes into account explicit ratings and implicit interactions to compute the low-dimensional vectors via a neural network architecture, and then estimates the relevance of an item to a user with cosine similarity between the vectors. DMF is not available in either Cornac or LensKit. The Elliot framework [21] includes an implementation of DMF that comes with a set of hyperparameters that can be tweaked. DMF's hyperparameters consist of latent factors, which represent the number of units in the final MLP layer for both users and items. The regularization term controls overfitting by penalizing large weights in the model. The learning rate determines the step size during optimization. Finally, the similarity measure computes the relevance between user and item embeddings using cosine similarity.

Neural network based approaches are generally understood to be less interpretable. Therefore, predicting or even explaining the effect of their parameters on popularity bias is nontrivial. For the purposes of this study, we experiment with the number of latent factors in the final layer of the network, since they can affect the underfitting/overfitting of the model, which can have an interplay with the data characteristics.

5.4 Experimental Setup

In this section, we describe the experiments that we run in order to determine the effect of data characteristics and algorithm configuration on the propagation of popularity bias. We run all experiments on a Fedora Linux 40 machine with a 24-core AMD EPYC 7401 Processor and 1TiB RAM. We use the following versions of the recommender systems frameworks: LensKit 0.14.2, Cornac 1.16.0 and Elliot 0.3.1. The code we wrote to run the experiments[2,3] has been made open source. For the Book-Crossing subset, the datasets MovieLens1M and Epinion, as well as every synthetic data scenario, we perform a recommendation process given every version of every algorithm. For UserKNN, we test the following versions given the configurations discussed in section 5.3.2:

- Min. similarity 0, over all items, 1 min. neighbour.
- Min. similarity 0, over all items, 2 min. neighbours.
- Min. similarity -1, over all items, 1 min. neighbour.

[2] https://github.com/SavvinaDaniil/DiagnosingBiasRecSys
[3] https://github.com/SavvinaDaniil/Elliot

- Min. similarity -1, over all items, 2 min. neighbours.
- Min. similarity -1, over common items, 1 min. neighbour.

For BMF, we test the LensKit and Cornac version, along with the bias parameter. For DMF, we test the Elliot version, along with the number of factors in the final layer. For each algorithm version and each dataset, we perform optimization based on RMSE by splitting the dataset into 80-20% sets to find the best values for some of the non-fixed hyperparameters of the respective version, except for DMF that we based the optimization on NDCG@10 since DMF does not predict rating but relevance score. The resulting hyperparameters can be seen in our repositories. Afterwards, we divide the users into training and test users in a 5-fold cross validated way. We make sure to use the same splits for all algorithms and all versions. For every test user, we use 80% of their ratings for training and the remaining 20% for testing, which is an option in LensKit. We train the model on the training set. For each user in the test set, we predict a rating for every item they have not rated in the training set (see TrainItems in section 3.1.3 of [196]), rank the items based on the predicted score and recommend the top-10 items, in line with recent studies on popularity bias.

We report on RMSE and NDCG@10 where applicable to estimate the effectiveness of the rating prediction and ranking, respectively. We also calculate the following widely used metrics on the recommended lists to estimate popularity bias propagation:

1. **Popularity Correlation (PopCorr)**: The correlation between popularity in training set and recommendation frequency for every item [121].
2. **Average Recommendation Popularity (ARP)**: The average popularity of the items in the recommended lists [8, 253].
3. **Popularity Lift (PL)**: The average relative difference in popularity between the recommended items and the items in the users' profiles [10].

Finally, for every dataset and algorithm we perform a Mann–Whitney U test to observe whether there is a significant difference among configurations for ARP and PL, and include the result in the respective tables.

5.5 Results

In this section, we provide insights into how algorithm configurations impact popularity bias and performance for the different datasets by presenting the results across the set of metrics listed in section 5.4. For each metric, we embolden the highest value among configurations. For ARP and PL, we use the asterisk (*) to signify which values are significantly lower than the highest one according to a Mann–Whitney U test with p<0.005. We abbreviate Book-Crossing as B-C and MovieLens1M as ML1M.

5.5.1 Popularity bias by UserKNN

Real Data

Table 5.3 shows the results when combining different UserKNN configuration choices with the Book-Crossing subset, MovieLens1M, and Epinion.

Table 5.3: Popularity bias and performance of different UserKNN configurations given Book-Crossing, MovieLens1M, and Epinion. OverCommon set to True corresponds to the Cornac implementation, and set to False to the LensKit implementation. For each metric, we embolden the highest value among configurations. For ARP and PL, we use the asterisk (*) to signify which values are significantly lower than the highest one according to a Mann–Whitney U test with $p<0.005$.

Dataset	Min Sim	Items for Similarity	Min Nbrs	Pop Corr↑	ARP↑	PL↑	RMSE↓	NDCG @10↑
B-C	-1	Common	1	0.010	0.002*	-32.843*	1.739	0.001
		All	1	-0.000	0.002*	-38.583*	1.860	0.001
			2	0.282	0.003*	15.018*	1.758	0.003
	0		1	0.071	0.003*	-17.740*	1.816	0.001
			2	**0.446**	**0.005**	**67.356**	**1.704**	**0.006**
ML1M	-1	Common	1	-0.093	0.000*	-99.722*	0.910	0.000
		All	1	-0.082	0.000*	-99.782*	0.906	0.000
			2	-0.053	0.017*	-86.270*	0.904	0.004
			10	**0.247**	**0.175**	**26.134**	**0.894**	**0.055**
	0		1	-0.050	0.010*	-91.918*	0.900	0.003
			2	-0.003	0.041*	-67.794*	0.894	0.013
			10	0.176	0.170*	21.997	0.898	0.048
Epinion	-1	Common	1	0.020	0.000*	20.789*	1.154	0.000
		All	1	0.023	0.000*	36.538*	1.212	0.000
			2	**0.173**	0.001*	**176.224**	1.168	0.000
	0		1	0.043	0.001*	65.527*	1.148	0.000
			2	0.153	**0.001**	165.929	**1.108**	**0.001**

The extent to which popularity bias propagates in the recommendation varies across the datasets. For Book-Crossing, the algorithm performance is best when popularity bias is high, based on both RMSE and NDCG@10. When minimum neighbours are set to 1, there is no notable popularity bias according to all metrics. On the other hand, when minimum neighbours are set to 2, the PopCorr and PL metrics indicate strong popularity bias. Additionally, minimum similarity is impactful: increasing it from -1 to 0 increases popularity bias across all metrics. Finally, which items to consider for similarity has a small effect on all metrics, but without a clear direction.

For MovieLens1M, according to NDCG@10, ranking performance is higher than for Book-Crossing and Epinion. For MovieLens1M, the minimum neighbours hyperparameter of UserKNN impacts popularity bias largely, with bias increasing when more minimum neighbours are required. In contrast to Book-Crossing, popularity bias is the highest for minimum similarity of -1, based on all metrics. Which items to consider for similarity has no discernible effect on popularity bias.

On Epinion, popularity bias is present for all versions of the algorithms according to the PopCorr and PL metrics. Again, increasing minimum neighbours increases popularity bias across metrics. The parameters of minimum similarity and which items to consider for similarity affect popularity bias, but not largely.

It is immediately observable that popularity bias varies across the different configurations of UserKNN with the parameter of minimum neighbours, one that is not configurable in one of the frameworks, being especially influential. This is true for every dataset, though the degree of popularity bias also differs between the datasets. The results indicate that popularity bias is not unavoidable when the data follows a long tail distribution, and depends on other characteristics of the datasets as well as the configuration of the algorithms.

Synthetic Data

To investigate which data characteristics could impact popularity bias propagation given each version of UserKNN, we present the results for the synthetic datasets in table 5.4.

Performance varies across the data scenarios. RMSE specifically is lower for scenarios 2 and 3 compared to the other three. In these two scenarios, users tend to agree between them on whether they like popular items or not, which facilitates the rating prediction task. NDCG@10 is the highest for scenario 2, where popular items are highly rated by the users. In this case, the rating prediction and ranking tasks are linked, since the highest ranked (i.e., popular items) are also highly rated.

Popularity bias also varies across the data scenarios, and the effect depends on the algorithm configuration. In the following paragraphs, we describe and reflect on the most impactful effects of the interaction between data and configuration.

For scenario 1 where ratings are uniformly at random generated, there is no notable popularity bias propagation observed when minimum neighbours are set to 1, while there is bias when minimum neighbours is set to 2. This observation is in line with what we noted for the real datasets, where increasing minimum neighbours results in higher popularity bias for all datasets and metrics.

In scenario 3 where all users agree that popular items are bad, popularity bias is not propagated when minimum similarity is set to 0. However, when setting minimum

Table 5.4: Popularity bias and performance of different UserKNN configurations given synthetic data based on different data scenarios. OverCommon set to True corresponds to the Cornac implementation, and set to False to the LensKit implementation. For each metric, we embolden the highest value among configurations. For ARP and PL, we use the asterisk (*) to signify which values are significantly lower than the highest one according to a Mann–Whitney U test with p<0.005.

Data Scen.	Min Sim	Items for Similarity	Min Nbrs	Pop Corr↑	ARP↑	PL↑	RMSE↓	NDCG @10↑
Scen.1	-1	Common	1	0.004	0.002*	-35.746*	**3.337**	0.001
		All	1	0.018	0.002*	-32.285*	3.502	0.001
			2	0.418	0.004*	21.252*	3.352	0.003
	0		1	0.101	0.003*	-12.827*	3.624	0.002
			2	**0.615**	**0.005**	**65.440**	3.464	**0.005**
Scen.2	-1	Common	1	0.604	0.015*	305.197*	1.150	0.013
		All	1	0.596	0.021*	426.621*	1.188	0.019
			2	**0.614**	0.022*	447.618*	1.190	0.021
	0		1	0.552	**0.027**	**632.300**	1.040	0.023
			2	0.562	0.027	591.966	**1.026**	**0.025**
Scen.3	-1	Common	1	0.522	0.006*	151.686*	1.151	0.001
		All	1	0.559	0.008*	187.197*	1.182	0.002
			2	**0.728**	**0.008**	**192.127**	1.182	0.002
	0		1	0.025	0.002*	-35.765*	1.044	0.001
			2	0.161	0.003*	-13.100*	**1.034**	**0.004**
Scen.4	-1	Common	1	0.184	0.003*	8.669*	2.458	0.001
		All	1	0.253	0.003*	23.063*	2.502	0.001
			2	**0.772**	0.006*	97.490*	2.404	0.004
	0		1	0.588	0.008*	164.549*	2.500	0.004
			2	0.701	**0.014**	**297.047**	**2.386**	**0.010**
Scen.5	-1	Common	1	0.057	0.002*	-16.243*	2.783	0.001
		All	1	0.087	0.003*	-7.924*	2.880	0.001
			2	**0.623**	0.005*	**57.969**	**2.776**	0.003
	0		1	0.136	0.003*	-16.122*	2.914	0.003
			2	0.612	**0.005**	42.849	2.794	**0.006**

similarity to -1, we can observe popularity bias propagation across all metrics. The reason is that users with completely different opinions are considered and their opinions count negatively. Therefore, popular items still get recommended since everyone's "negative" neighbours dislike them, and we can observe popularity bias propagation across all metrics.

When considering only common items to calculate similarity, users with smaller profiles have a larger influence. This is relevant in scenario 4 where users with large profiles like popular items. Table 5.4 shows that even though scenario 4 still leads to popularity bias, considering only common items reduces it across all metrics. Therefore, this implementation choice can have a big impact on whether popularity bias is propagated and to what extent.

Finally, the value for minimum neighbours largely influences popularity bias. Across almost all scenarios and metrics, increasing minimum neighbours from 1 to 2 leads to increased popularity bias. By setting a higher neighbour barrier for considering an item for recommendation, it follows that less popular items will be disadvantaged. This result is particularly relevant given that the parameter of minimum neighbours could only be tweaked in one of the considered frameworks, so studies that use Cornac or LensKit might reach different conclusions on the extent of popularity bias propagated by UserKNN.

Popularity bias manifests differently in different datasets; an explanation for this can be found in data characteristics, specifically the relation between item ratings, item popularity, and the influence of users with large profiles. All three configuration choices affect the observed bias, as they influence the weight of each user's preference.

5.5.2 Popularity bias by Matrix Factorization algorithms

Real Data

Table 5.5 shows the results for BMF, trained on Book-Crossing, MovieLens1M and Epinion.

BMF tends to have a better performance than UserKNN overall. Popularity bias varies across the LensKit and Cornac implementations of BMF. Also very notable is the effect of the bias parameter.

For Book-Crossing, in both implementations the bias parameter increases popularity bias. For example, PopCorr on Book-Crossing with LensKit is -0.013 with bias set to False and 0.108 with bias set to True, as seen in Table 5.5. Specifically, the LensKit implementation with the bias parameter set to True is the only version where there is a positive popularity correlation, as well as positive popularity lift, with the difference being significant.

For MovieLens1M, there is higher popularity bias across metrics when the Cornac implementation is used. Additionally, the bias parameter changes PL from negative to positive when the LensKit implementation is used.

For Epinion, there is higher popularity bias across metrics when the LensKit implementation is used. The bias parameter is also very influential, as setting it to True results in higher popularity bias across all metrics, given both implementations, as was the case for Book-Crossing.

Table 5.5: Popularity bias and performance on the Book-Crossing, MovieLens1M and Epinion datasets given different BMF implementations. For each metric, we embolden the highest value among configurations. For ARP and PL, we use the asterisk (*) to signify which values are significantly lower than the highest one according to a Mann–Whitney U test with $p<0.005$.

Dataset	Framework	Bias parameter	Pop Corr↑	ARP↑	PL↑	RMSE↓	NDCG @10↑
B-C	Cornac	False	-0.015	0.001*	-68.639*	1.587	0.001
		True	0.003	0.002*	-40.168*	**1.535**	0.001
	LensKit	False	-0.013	0.002*	-58.518*	1.762	**0.004**
		True	**0.108**	**0.005**	**46.533**	1.560	**0.004**
ML1M	Cornac	False	**0.266**	**0.193**	35.186	**0.856**	**0.064**
		True	0.234	0.192	**35.859**	0.856	0.059
	LensKit	False	0.151	0.125*	-14.392*	0.860	0.038
		True	0.183	0.155*	7.987*	0.866	0.040
Epinion	Cornac	False	0.001	0.000*	-53.845*	1.152	0.000
		True	0.009	0.001*	19.814*	**1.029**	0.000
	LensKit	False	0.020	0.001*	-54.107*	1.240	0.000
		True	**0.126**	**0.003**	**402.394**	1.030	**0.001**

We see that the choice for a framework - LensKit or Cornac - largely determines whether popularity bias is observed when using a Matrix Factorization algorithm on widely used datasets, and the effect differs per dataset. In addition, setting the bias parameter of these algorithms impacts to what extent popularity bias is observed.

Synthetic Data

Additionally to the conclusions drawn from the results on the real datasets, to observe whether the data scenarios influence popularity bias propagated by BMF and the impact of the bias parameter, we present the results on the synthetic datasets in table 5.6.

BMF has a better performance than UserKNN when trained on the synthetic datasets. Similarly to UserKNN, BMF also results in low RMSE for scenarios 2 and 3, and high NDCG@10 for scenario 2. Experimenting with different synthetic datasets helps explore the observation from the previous section that BMF propagates popularity in some cases. Popularity bias can be observed for scenarios 2 and 4, while scenario 5 where users with large profiles do not like popular items results in no popularity bias across metrics. This indicates that the preferences of users with large profiles are influential for the system, and thus, for popularity bias.

In the previous section, we saw that the LensKit and Cornac implementations lead to varying results with respect to popularity bias. On the synthetic datasets, the effect is more apparent. The bias parameter has opposite effects between the two implementations in scenario 4 as seen in Table 5.6. In Cornac, using bias increases popularity bias across all metrics. On the contrary, the bias parameter decreases popularity bias in the LensKit implementation. This can be due to the use of different

optimization methods by Cornac and LensKit. Further investigation is needed to better understand the impact of the bias parameter on popularity bias.

Table 5.6: Popularity bias and performance given different BMF implementations and synthetic data based on different data scenarios. For each metric, we embolden the highest value among configurations. For ARP and PL, we use the asterisk (*) to signify which values are significantly lower than the highest one according to a Mann–Whitney U test with p<0.005.

Data Scenario	Framework	Bias parameter	Pop Corr↑	ARP↑	PL↑	RMSE↓	NDCG @10↑
Scen. 1	Cornac	False	-0.013	0.001*	-67.527*	3.191	0.001
		True	**0.079**	**0.006**	**74.548**	**2.872**	**0.003**
	LensKit	False	-0.015	0.002*	-52.395*	3.400	0.002
		True	-0.008	0.002*	-49.725*	3.028	0.001
Scen. 2	Cornac	False	0.475	0.030	743.388	0.938	**0.026**
		True	0.467	**0.030**	**743.446**	**0.802**	0.025
	LensKit	False	**0.507**	0.030*	740.064	0.940	0.025
		True	0.498	0.030*	742.366	0.818	0.025
Scen. 3	Cornac	False	-0.022	0.001*	-76.210*	0.850	0.001
		True	0.010	**0.002**	**-31.473**	**0.796**	0.001
	LensKit	False	-0.091	0.001*	-73.688*	1.032	0.001
		True	**0.017**	0.002*	-37.387*	0.810	0.001
Scen. 4	Cornac	False	0.036	0.004*	-20.286*	2.510	0.005
		True	0.424	**0.027**	**647.074**	**2.297**	**0.019**
	LensKit	False	**0.462**	0.010*	149.807*	2.746	0.012
		True	0.219	0.008*	129.271*	2.390	0.008
Scen. 5	Cornac	False	-0.018	0.001*	-73.474*	2.671	0.001
		True	**-0.006**	0.002*	-53.320*	**2.500**	0.001
	LensKit	False	-0.045	0.001*	-59.393*	2.890	**0.002**
		True	-0.013	**0.002**	**-50.113**	2.582	0.001

We conclude that, even though the data synthesis was based on the functionality of UserKNN, the differences between the synthesized datasets largely impact popularity bias propagated by BMF. Furthermore, the results confirm that for both BMF implementations, the bias parameter affects popularity bias. The effect is different - sometimes even opposite - for different combinations of dataset and implementation.

5.5.3 Popularity bias by Deep Matrix Factorization

The goal of this section is to see whether the lessons learned from the detailed analysis presented in the previous sections hold for a neural network-based method.

Real Data

Table 5.7 shows the results when combining different DMF configuration choices with the Book-Crossing subset and MovieLens1M. Training DMF on Epinion was not possible due to memory allocation limitations.

The number of latent factors in the final layer of the neural network has a clear impact on the results. Interestingly, the impact differs between the two datasets.

Specifically, when the number of factors is set to 64, popularity bias increases for Book-Crossing, with the difference being significant. In the case of MovieLens1M, the same configuration results in significantly lower values across metrics.

We observe that the combination of data and configuration also affects popularity bias propagation by neural architectures.

Table 5.7: Popularity bias and performance on the Book-Crossing, MovieLens1M and Epinion datasets given different DMF versions. For each metric, we embolden the highest value among configurations. For ARP and PL, we use the asterisk (*) to signify which values are significantly lower than the highest one according to a Mann–Whitney U test with p<0.005.

Dataset	Factors	Pop Corr↑	ARP↑	PL↑	NDCG @10↑
B-C	32	0.012	0.002*	-38.610*	0.002
	64	**0.125**	**0.005**	**35.582**	**0.003**
ML1M	32	**0.046**	**0.064**	**-54.777**	**0.010**
	64	0.032	0.057*	-59.654*	0.008

5

Synthetic Data

To confirm and expand upon the observations of the previous section, we report on performance and popularity bias of DMF on the synthetic data in table 5.8.

Table 5.8: Popularity bias and performance given different DMF versions and synthetic data based on different data scenarios. For each metric, we embolden the highest value among configurations. For ARP and PL, we use the asterisk (*) to signify which values are significantly lower than the highest one according to a Mann–Whitney U test with p<0.005.

Data Scenario	Factors	Pop Corr↑	ARP↑	PL↑	NDCG @10↑
Scenario 1	32	0.057	**0.003**	**-8.529**	0.003
	64	**0.059**	0.003	-10.247	0.003
Scenario 2	32	0.183	**0.005**	**42.030**	0.004
	64	**0.720**	0.004*	-13.542	**0.017**
Scenario 3	32	0.039	0.003*	-19.041*	0.002
	64	**0.074**	**0.004**	**4.613**	**0.003**
Scenario 4	32	0.009	0.002*	-41.003*	0.002
	64	**0.393**	**0.003**	**-34.005**	**0.006**
Scenario 5	32	0.004	0.002*	-43.433*	0.002
	64	**0.079**	**0.003**	**-10.451**	**0.003**

Popularity bias and performance fluctuate among scenarios. In line with the results from previous sections, popularity bias is the highest for scenarios 2 and 4, and the low for the other scenarios. The parameter tested also has an effect, which aligns with the effect on the real data. For all scenarios, increasing the number of factors in the

final layer increases popularity bias, in most cases significantly, which aligns with the results on Book-Crossing.

Note that the results presented in this section are preliminary. Further research is needed to explore the effect of data characteristics and algorithm configuration on popularity bias propagated by neural network based algorithms. A future study could align with recent advancements in recommender systems by formulating a ranking prediction task and studying popularity bias propagation by the neural network based algorithms available in commonly used open source frameworks.

The preliminary results indicate that the conclusions drawn above also hold for neural network based approaches: data characteristics determine whether popularity bias occurs; popularity bias is not unavoidable even when the data follows a long tail distribution; parameters that can be set at implementation time have a large effect on the propagation of popularity bias.

5.6 Discussion

5.6.1 Implications of the present study

Our research shows that multiple data and configuration factors can have an effect on whether bias is propagated. Relying on frameworks readily available to researchers is convenient and a concrete step towards reproducibility, but requires being aware and detailed about the limitations. When simple parameters, such as minimum neighbours in UserKNN, are so influential, it raises questions on how generalizable research on recommender systems bias can be. Our results indicate that bias studies can only draw conclusions within the limits of their specific research and not further than that.

It follows that being explicit about the context within which a type of bias is studied is crucial, both in terms of data characteristics and implementation. It is a known issue in recommender systems literature that implementation details are often not disclosed by studies. Even in cases where they are, guessing the effect of different hyperparameters that are not present in an implementation or experimented with is not trivial. Bias reporting is definitely not complete if it is not accompanied by clarity around the characteristics, goals and limitations of the system that is being studied.

5.6.2 Recommendations of the present study

Based on the results of the present study, we put forward two recommendations towards researchers who study bias in recommender systems.

First, researchers should analyze and report on the dataset characteristics that might impact the type of bias they are concerned with. For UserKNN, the relationship between rating and popularity, as well as the preferences of users with large profiles impact popularity bias and should be taken into account in relevant studies. For other algorithms and types of bias, there could be other relevant characteristics. Such analysis will help the reader understand the extent to which the results are a result of the dataset characteristics.

Second, researchers should test multiple algorithm configurations when measuring bias propagation. In a similar way that the community expects x-fold cross validation, since presenting the results of only one run may not be reliable, we could expect to see results on multiple algorithm configurations as well. If the conclusions are only valid

for one specific configuration of the algorithm at hand, then that should be clear in the limitations of the study.

5.6.3 Limitations of the present study and future work

Despite our extensive testing, the results are potentially sensitive to our own experimental design, such as the method for train-test splitting or randomness in the data generation process. Similarly, instantiating the different implementations with the exact same configuration choices is not always possible due to some of the parameters not being configurable. As a result, there might be implementation differences between the frameworks that we are not aware of and cause part of the variation in results, irrespectively of the configurations tested. On this note, we noticed that our reported accuracy does not always coincide with the conclusions of other papers that study the same algorithms and/or datasets. For example, the developers of DMF reported high NDCG@10 results for MovieLens1M [245]. However, their evaluation strategy was very different from ours: for every user, they only held out one item consumed by them for testing, and only ranked 100 random items instead of all the items not present in the training set like we did. It is reasonable that their performance is better than the one we report. This discrepancy further supports our claim that generalizing conclusions beyond the boundaries of a specific study is not always possible and should be done carefully, which aligns with our previous work where we emphasized the impact of evaluation strategy [58]. These observations highlight the importance of our line of research instead of hindering it, since they hint that data and implementation dependence might be present more often than we think.

We recognize that the scholar community has generally moved on from explicit user preferences and rating prediction. We do not focus on implicit feedback given that recent studies on popularity bias are often performed on datasets with explicit ratings in the context of rating prediction tasks [121, 122, 159]. A process similar to ours can be followed in the context of implicit feedback: instead of tweaking ratings, researchers can form scenarios around the distribution of popularity in the dataset and study distributions with varying levels of long-tail. Future work can also focus on components that we chose not to investigate in this study, such as more advanced algorithms, other open libraries for implementation, and the vulnerability of other types of bias to data and algorithms. For example, researchers can create data scenarios based on assumptions regarding the relationship between an item's rating and the demographic characteristics of its creator. Further nuance can be introduced in the data synthesis part, by allowing for more complex relationships between popularity, rating and user influence. Future work could further investigate the impact of data characteristics and configurations on the complex relationship between system performance and popularity bias. Our findings could also be of use in the domain of bias mitigation: certain algorithm configurations appear to have a clear effect on popularity bias (e.g., increasing the minimum neighbours in UserKNN consistently increases it). It is, therefore, relevant to investigate whether appropriately configuring certain parameters assists with popularity bias mitigation.

5.7 Conclusion

In this study, we reflected on the need for fundamental understanding of the relationship between data, algorithms and bias in recommender systems. We focused on reporting on popularity bias, and tracked algorithm configurations and data characteristics that are of importance in its propagation. Accordingly, we generated a set of synthetic datasets, experimented with performing a recommendation process on real and synthetic datasets using different configurations of the algorithms at hand, and evaluated popularity bias using well-known metrics. We found that even when the distribution of popularity in the dataset is long-tail, popularity bias is not unavoidable. We showed that the relationship between popularity and rating, as well as the preferences of users with large profiles have an impact on bias. We highlighted the sensitivity of bias propagation to algorithm configuration and, by extension, framework implementation. Our observations point to methodology and reproducibility issues that extend further than a specific use case, to the recommender systems field at large.

Recommender systems are widely used in our online lives, and bias propagation by such systems can have serious societal impact. With this work, we hope to have called attention to the ambiguity in bias reporting and motivated researchers to strive for reproducibility and highlight specificity when appropriate.

Acknowledgments and Disclosure of Funding

This research was funded through a public-private partnership between CWI and the KB National Library of the Netherlands.

6

Hidden author bias in publicly available datasets

Following our investigation of how data characteristics and algorithms affect bias measurement (Chapters 4–5), we now ask whether popularity bias in practice corresponds to social bias. In this chapter, we examine whether the statistical phenomenon of popularity bias translates into bias against authors based on demographic characteristics.

Collaborative filtering algorithms have the advantage of not requiring sensitive user or item information to provide recommendations. However, they still suffer from fairness related issues, like popularity bias. In this work, we argue that popularity bias often leads to other biases that are not obvious when additional user or item information is not provided to the researcher. We examine our hypothesis in the book recommendation case on a commonly used dataset with book ratings. We enrich it with author information using publicly available external sources. We find that popular books are mainly written by US citizens in the dataset, and that these books tend to be recommended disproportionally by popular collaborative filtering algorithms compared to the users' profiles. We conclude that the societal implications of popularity bias should be further examined by the scholar community.

This chapter is based on *Daniil et al., Hidden Author Bias in Book Recommendation, arXiv:2209.00371* [57].

6.1 Introduction

In recent years, Trustworthy AI is discussed as an important principle within the AI scholar community, but also on a national and international governmental level [79]. Robust algorithms have proven to be vulnerable to undesirable data patterns, and developers are not always able to beforehand recognize that. Public organizations specifically bear immense responsibility when utilizing AI to promote and facilitate their purposes. Recently publicized scandals have showcased that issues of bias and fairness are not always properly taken into account [30, 98].

Recommender Systems are a very popular class of algorithms within domains like e-commerce and entertainment platforms [132]. They often process an enormous amount of data in order to profile users and suggest products that they are likely to consume, and have proven to be very efficient. Collaborative filtering algorithms are a subset of Recommender Systems that specifically function by receiving consumption history as input rather than sensitive personal information of the users or items [120]. One might think that by omitting sensitive information, unwanted bias has no way to manifest in such a system. This is not the case; collaborative filtering approaches are still known to suffer from popularity bias. In short, popularity bias is an algorithmic effect where items that were originally popular in the training dataset tend to be recommended more often and thus have their popularity increase further.

Popularity bias is recognized in both social and academic circles as a problem to be dealt with. It leads to decreased exposure of long-tail items that are critical for providers and platforms [2], and users with niche tastes may have their interests ignored [9]. At the same time, feedback loops are a point of concern for recommendations as homogeneity in user behavior amplified by popularity bias hinders utility [40].

However, certain indirect implications of it are not placed in the forefront of said discussions. Given the fact that popular items tend to become more popular, it is appropriate to wonder which items were popular in the first place. We argue that popularity is often linked to other properties of an item, which translates to popularity bias leading to other sorts of data bias as well.

Specifically, we are investigating the topic of bias within the book recommendation domain. In this context, book authors are potential recipients of bias by a recommender system, as their popularity might coincide with sensitive characteristics such as age, ethnicity, religion, gender identity, sexual orientation, nationality and political preference. Multiple of said characteristics are known to instigate unfair treatment of authors in the publishing world [23]. The treatment of authors by a recommender system based on their gender identity has received some attention from the scholar community [68, 199], but other characteristics remain unexamined.

In this specific work, we are zooming in on author country of citizenship. In a globalized book market, readers are no longer only exposed to local production. Steiner [215] argues that "..shifts in Internet use, the expansion of an e-book market, and the influence of a few large American corporations together appear to be transforming international publishing and retail". Hence, it is relevant to examine the relation between author country of citizenship and popularity, as well as the resulting effects in recommendation. Specifically, we are answering the following research questions:

1. Do commonly used recommender algorithms propagate data bias towards author country of citizenship?

2. What is the relation between author country of citizen bias and popularity bias?

We are considering these questions in the context of a dataset with book ratings frequently used in literature to evaluate recommender systems.

6.2 Related Work

In this section, we will discuss previous work on popularity bias, as well as recent reports on the presence of bias in the book publishing industry.

6.2.1 Popularity bias

Popularity bias is a long standing problem within the context of Recommender Systems and Information Retrieval [37]. As explained by Steck [214] it stems from the users' tendency to provide feedback to popular items more often than long-tail items, which may not represent their true preferences. A system that recommends solely popular items to every user might score high on accuracy metrics, but it does not satisfy other requirements like item coverage [146].

Recently, studies in the movie, music and book domain showed that popularity bias impacts different user groups disproportionally depending on their affinity for popular items [9, 122, 159]. Specifically, users who tend to prefer niche items still receive mostly popular items as recommendations by well-known recommender algorithms. In that sense, niche users are treated unfairly by the system that fails to expose them to items that presumably would match their preferences better.

Popularity bias is recognized as a flaw of the system that leads to overall poor functionality in terms of fairness and balance, which transcend the basic goal of user satisfaction. Abdollahpouri [3] argues that recommenders exist in multi-stakeholder platforms, thus rendering the satisfaction of interests, economic or otherwise, of item providers and platforms equally desirable. However, to our knowledge the societal connotations of popularity bias are not sufficiently addressed. Commonly used datasets with ratings tend to not contain information that would be needed to examine which items are consistently popular or which social groups of users/providers are unfairly treated. In other words, the potential of societal harm when popularity correlates with sensitive data properties adds an extra layer to the view of popularity bias breaching fairness.

6.2.2 Bias in the book market

The existence of different forms of bias in the publishing industry is well known. Female authors at large have historically been undervalued in the publishing landscape, for example by being less likely to be considered for reviews or major awards [78]. The 2020 academic study "Rethinking Diversity in Publishing" conducted qualitative interviews with professionals from all major UK literary agencies and concluded that publishers cater mostly to white middle-class audience and hence are less likely to

invest in acquisition and promotion of authors of color [195]. Similar patterns have been reported in the US by various surveys [127, 266].

In the globalized book market, the US holds the largest revenue share according to a 2018 survey by the International Publishers Association [243]. At the same time, 7 out of the 10 highest-paid authors worldwide reported in 2018 by Forbes are US citizens [51]. Social cataloging websites commonly used by younger readers for discovering and reviewing new titles also tend to be US-centric. For example, the American Amazon owened Goodreads places first in the list of websites related to "Libraries and Museums" ranked by traffic [206]. While US based Goodreads users make up an estimated 40% of the traffic [233], the popular authors' country of origin is disproportionally skewed, with 12 out of the 15 most often rated authors between 2016 and 2021 being American [142]. In other words, the US holds strong influence in the modern literary world, particularly when it comes to renowned authors.

6.3 Research Design

In this section, we will describe the data selection, preprocessing and enrichment, and our approach to measuring bias in the data and recommendations.

6.3.1 Data

In order to study hidden author bias in book recommendation, we selected the Book-Crossing dataset [265], a very popular dataset for book recommendations created in 2004 by crawling the website of Book-Crossing for four weeks. Book-Crossing is an international book exchange community whose members can leave books anywhere in the world and indicate their location to other users. The users can then update their profile with the information that they read a new book and also submit a rating from 1 to 10.

The dataset consists of three sub-datasets: book ratings, users and their location, and items along with some information on them, namely the name of the author, the publisher, and the year of publication. We hence used external sources to enrich it with additional author information when publicly available. We took the following steps, as shown in Figure 6.1:

1. We linked Book-Crossing to Google Books based on ISBN. We used the author name included in Google Books to validate and/or correct the one provided in Book-Crossing, since we noticed mistakes and inconsistencies.

2. We linked to Virtual International Authority File (VIAF) using author name.

3. We linked to Wikidata using VIAF ID, and validated using author name. From Wikidata, we extracted author country of citizenship.

Our motivation for selecting Book-Crossing is that Naghiaei et al. [159] utilized it in their Fairbook system to investigate the unfairness of popularity bias in the book recommendation case. In order to build on their work by comparing how certain recommender algorithms propagate popularity versus how they propagate author bias, we reproduced their preprocessing approach:

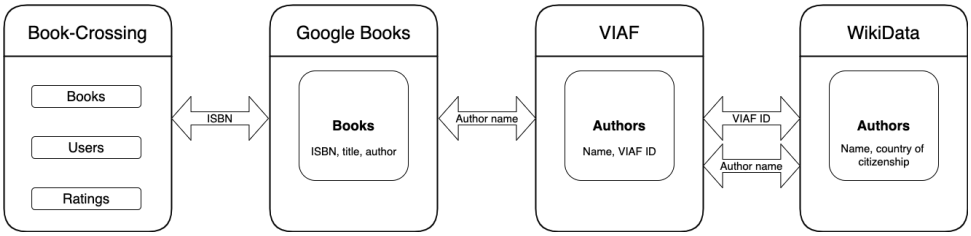


Figure 6.1: Links between the datasets.

- Remove implicit ratings.
- Remove users with more than 200 ratings.
- Remove users with less than 5 ratings.
- Remove items with less than 5 ratings.

The application of such cut offs is common in the recommendation literature given the vast but sparse amount of data that online platforms usually possess [138]. Therefore, it is compelling to observe whether it has an effect on the inherent bias of the dataset.

We further processed the dataset to account for duplicate items with different ISBN numbers, as well as ISBN numbers included in the ratings dataset but not in the items dataset that we considered to be mistakes. This processing is documented and accessible in our GitHub. The general characteristics of the processed version of the entire dataset, as well as of the Fairbook dataset can be seen at Table 6.1.

Table 6.1: Dataset characteristics. The cut-off thresholds introduced by Fairbook significantly decrease the dataset size.

	#ratings	**#books**	**#users**
Entire dataset - processed	1,021,847	222,496	92,106
Fairbook dataset - processed	86,356	5,504	6,354

6.3.2 Measuring bias

In the interest of answering the research questions, we devised a way to measure the country of citizenship bias existing in the data and compare it to the recommendations offered when training different recommender algorithms on that data. We applied the research design to the Fairbook dataset in order to juxtapose the propagation of data bias to the propagation of popularity bias studied by the Fairbook system.

First, we examined whether there is a significant relationship between book popularity and country of citizenship of the author. We define popularity of an item to be the amount of ratings of said item, similarly to Abdollahpouri et al. [9]. Afterwards, we trained the same algorithms as the Fairbook system so we can directly compare the

results, using the Cornac framework [197], version 1.14.2. Specifically, we trained 11 recommender algorithms, 9 collaborative filtering and two dummy approaches as seen in Table 6.2, with a 80-20% split. We then recommended to each user 10 books as a result of each algorithm's predicted ranking.

Table 6.2: Recommender algorithms chosen to be trained on our dataset. They are the same ones trained by the Fairbook system.

Algorithm	Approach	Acronym
User K-Nearest Neighbors	K-Nearest Neighbors	UserKNN
Matrix Factorization	Matrix Factorization	MF
Probabilistic Matrix Factorization	Matrix Factorization	PMF
Non-negative Matrix Factorization	Matrix Factorization	NMF
Weighted Matrix Factorization	Matrix Factorization	WMF
Hierarchical Poisson Factorization	Matrix Factorization	PF
Bayesian Personalized Ranking	Ranking-based	BPR
Neural Matrix Factorization	Neural Network-based	NeuMF
Variational Autoencoder for Collaborative Filtering	Neural Network-based	VAECF
Most Popular	Dummy	MostPop
Random	Dummy	Random

Finally, we compared the distribution of author country of citizenship in the users' profiles versus recommended lists for every algorithm. That way we were able to observe whether and how each algorithm's recommendations impact this distribution and compare the result to the propagation of popularity bias for each algorithm.

Our code both for the data enrichment[1] and the bias analysis[2] have been made open source.

6.4 Results & Discussion

In this section, we will present the results of measuring author country of citizenship bias in the data and in the recommended lists.

6.4.1 Bias in the data

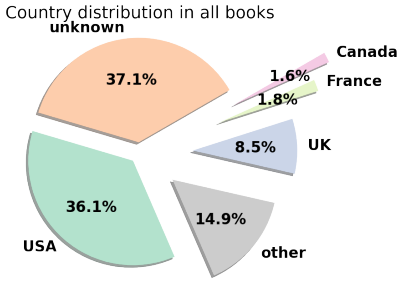
Data analysis shows clear bias in favor of US citizens authors. As seen in figure 6.2a, in the entire set of unique books around 36% were written by American authors. When the cut offs are introduced, the percentage of American-authored books grows to 68.5%, as seen in figure 6.2b. The same effect appears in the ratings datasets, figures 6.3a and 6.3b. Out of the entire dataset, 55.3% of the ratings are given to books written by US citizens. Similarly, the phenomenon is more apparent for the Fairbook ratings, with the percentage growing to 74.4%.

We conclude that introducing the cut offs increased the bias of the dataset in favor of US citizens authors. This remark already implies a direct relation between item popularity and country of citizenship of the author, given that Fairbook excludes all

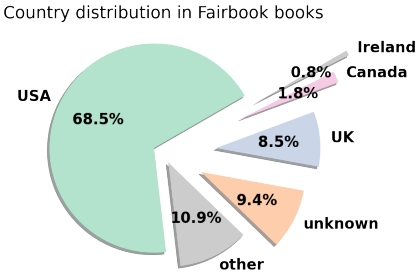
[1] https://github.com/SavvinaDaniil/EnrichBookCrossing
[2] https://github.com/SavvinaDaniil/BiasInRecommendation

items with less than 5 raters. At the same time, the Book-Crossing website reports that 20% of the users are based in the US [34]. We can deduce that American authors are overrepresented in the data compared to the amount of American users.

Finally, we divided the books on American-authored and non American-authored. A t-test between the mean popularity of the two sets showed a significant higher popularity for the American-authored items.

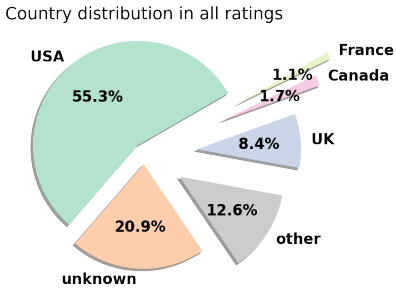


(a) Country distribution in the entire book dataset.

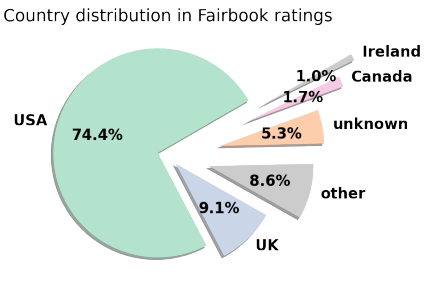


(b) Country distribution in the book dataset with Fairbook cut offs.

Figure 6.2: Country distribution in unique books.



(a) Country distribution in the entire ratings dataset.



(b) Country distribution in the ratings dataset with Fairbook cut offs.

Figure 6.3: Country distribution in unique ratings.

6.4.2 Bias in recommendation

After applying the recommendation process as explained in 6.3.2, we compared the ratio of American-authored books in users' profiles to the one in the recommended list of each algorithm. Figure 6.4 shows that, setting aside Random and MostPopular, all collaborative filtering recommenders increase the ratio, except from NMF, MF, and PMF.

In order to compare these algorithmic tendencies to popularity bias, we used the %ΔGAP metric [9]. It expresses the relative increase of average item popularity in a user's recommended list compared to their profile. Figure 6.5 is a recreation of a graph from Fairbook, with the difference that %ΔGAP is calculated over all users rather than

per user group. It shows that all collaborative filtering algorithms but PMF, MF, and NMF produce recommended lists with increased average item popularity. In other words, the same algorithms that recommend excessively American-authored books also recommend excessively popular books.

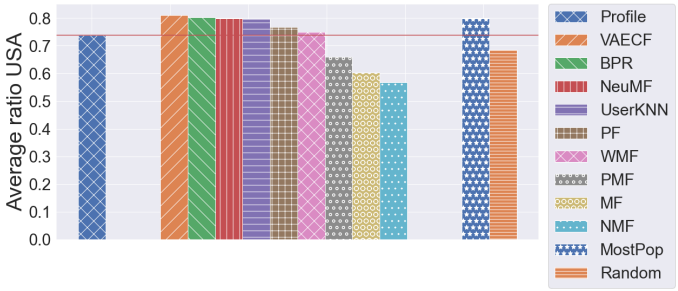


Figure 6.4: Average ratio of recommended books by every algorithm that were written by US citizens. Comparison with the average ratio of American-authored books in the users' profiles.

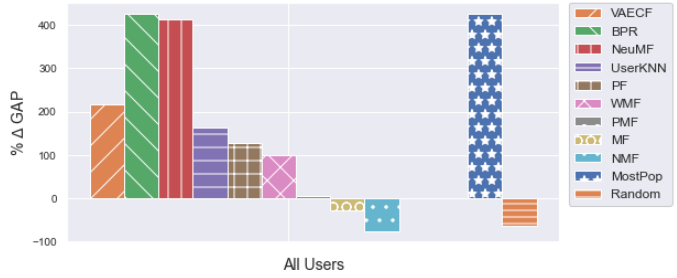
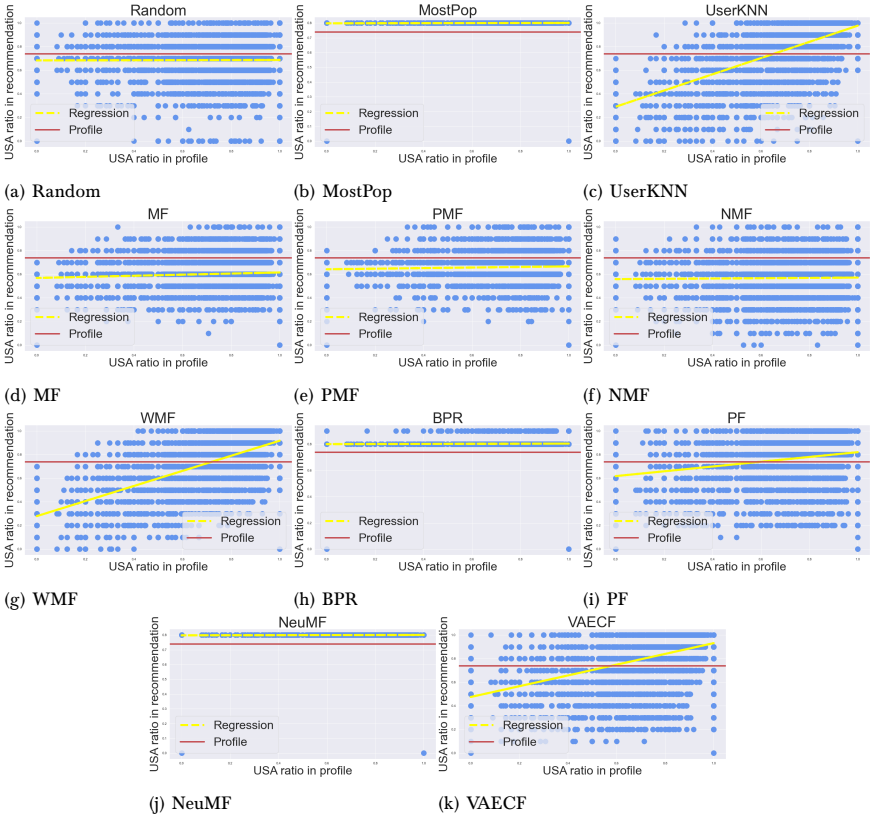


Figure 6.5: Relative increase in average popularity between profile and recommendation by every algorithm, averaged over all users.

We then zoom in on each user and compare their profile ratio of US citizens authors to the recommended list of each recommender algorithm. Figure 6.6 shows that most collaborative filtering algorithms either recommend consistently disproportionally many American-authored books despite the user's preferences (BPR, NeuMF), or they show a correlation between profile and recommended ratio (UserKNN, WMF, PF, VAECF). The only exceptions are NMF, MF, and PMF, which consistently recommend lower ratio than average. UserKNN, WMF, PF, and VAECF may show correlation between ratio of American-authored books in user profile and recommendation, but as seen in Figure 6.4 they on average increase the ratio.

When comparing our results with the Fairbook conclusions, we see that they follow similar trends in the sense that the same recommender algorithms that propagated popularity bias also acted in favor of American-authored items by disproportionally recommended them. Naghiaei et al. stated that "...no positive correlation exists in PMF, MF, and NMF, indicating that the latter algorithms in Matrix Factorization-based approaches are not prone to popularity bias in Book-Crossing dataset.". This is a strong

indication that the behavior we observe in our results is a direct result of popularity bias.



(a) Random (b) MostPop (c) UserKNN

(d) MF (e) PMF (f) NMF

(g) WMF (h) BPR (i) PF

(j) NeuMF (k) VAECF

Figure 6.6: Ratio of American-authored books in profile versus in recommendation for every user.

6.4.3 Analysis

Given our results, we can answer Research Question 1 by stating that certain recommender algorithms propagate bias against author country of citizenship, despite not receiving country of citizenship as a feature in the training process. Specifically, American authors were favoured by those algorithms by being disproportionally recommended. Comparing propagation of American author bias and popularity bias also enlightens us on its origins. Considering the similar manifestations of bias by the different recommenders, we do believe that the observed author bias was directly incited by the known phenomenon of popularity bias, thus answering Research Question 2.

The results indicate that American book platforms accessed by users around the world may be disproportionally US centric, and that this phenomenon can get perpetuated when certain recommenders are being applied without proper examination. Lack of geodiversity in open data is a known problem that requires addressing when

this data is employed for either research or commercial purposes [203]. The results also hint that popularity bias in recommendation should be viewed as more than an issue of potentially weak performance that mostly impacts e-commerce. Naturally, popularity is not always a bad criterion for recommendation, as argued by Zhao et al. [261]. Word-of-mouth is commonly how we as humans stay informed on art, politics, science, and other socially important topics. Popularity among topics can signal what is relevant and/or of quality, and automated systems can often accurately encode what is typically happening in society.

That being said, AI is known to track patterns in the data that do not objectively depict reality, but rather often undesirable historical context. Consequently, employing AI in a way to directly imitate the data without check points can bring about societal harm in unexpected ways. In the case of book recommendation, efficient recommender algorithms can be used to increase reading habits, but without risking demoting certain books by virtue of their authors' sensitive features that may correlate with past unpopularity. For this reason, our work indicates that researchers must be very thorough when examining potential bias in recommended systems and make active effort to address it, even when seemingly popularity is the cause.

6.5 Conclusion & Future Work

This chapter is a part of an ongoing attempt to reflect on the concepts of bias and inclusivity in the book recommendation case. We examined the phenomenon of hidden bias in recommendation introduced by commonly used recommender algorithms that do not take any item features as input. We theorized that feature bias comes as a direct result of popularity bias which is a known issue that recommenders face. We investigated the hypothesis in the context of book recommendations and used a well-known book ratings dataset to validate our hypothesis. We found the books written by American authors within the dataset to be significantly more popular compared to the rest. We also found that many commonly used collaborative filtering algorithms on average recommend more American-authored books than in the users' profile. In fact, the same algorithms seem to propagate popularity bias according to previours work on the same dataset.

We believe that hidden bias in recommender systems should be in the spotlight for the scholar community. Fair treatment of social groups should be a priority for AI developers. As shown, it is not sufficient to explicitly exclude sensitive personal features from the training process; bias can appear and manifest in proxies.

In collaboration with the KB National Library of the Netherlands, we wish to study recommender systems in terms of their capability to be inclusive. In their published AI principles, under the "Inclusive" section the National Library acknowledges AI's susceptibility to bias. At the same time, they equate maintaining inclusivity to knowing "... where and to what extent bias occurs [in the data], so that it can be eliminated or compensated." [228]. Given a library's unique position as a public organization that aims to promote education and responsibility to treat all social groups fairly, it is crucial to identify potential blind spots that can cause bias against author social groups to manifest when a recommender system is in use.

Currently we only compare profile with recommended lists, since we aimed to

focus on the relation between popularity bias and author bias. In the future, it will be interesting to use feedback loops in order to investigate how the observed phenomenon progresses through multiple iterations of recommendation and consumption.

At the same time, there is room for expanding the data enrichment process. We currently depend on the information existing in Wikidata on someone's country of citizenship, which is evidently incomplete. Moreover, it is documented that Wikidata can be biased in terms of country of citizenship, with Europe and North America being often overrepresented [202]. The aforementioned skewness in the available Wikidata entries may be introducing additional bias to our data. As seen in Figure 6.2, more than one third of the books in the entire dataset are written by authors whose country of citizenship we could not trace, while the authors of unknown origin have written less than one in ten books in the Fairbook version. Our reliance on Wikidata, as well as our choice to implement the Fairbook cut offs, might be result in blind spots around patterns in the relationship between popularity and author country of citizenship. We plan to address this topic thoroughly in future work.

The question of how bias surfaces in book recommendation is not negligible; it could be directed to authors, books, users. Our study focuses on authors and specifically country of citizenship, but each of these dimensions can be important depending on the context and thus should be given attention. For future work, we plan to consider the other dimensions of the problem as well. By having a well rounded understanding of hidden bias in book recommendation, we can move on to properly account and compensate for it. In this case, libraries can benefit from the automation recommender systems offer in attracting users, while ensuring that their value of inclusivity is being properly adhered to.

Acknowledgements

Funded by the KB National Library of the Netherlands.

7

A case study on the Danish public libraries

Building on our findings that popularity bias can manifest as author nationality bias in datasets and algorithms (Chapter 6), we now examine whether these patterns persist in a real library recommender system currently in production.

Public libraries can potentially benefit by automated recommendation to increase patron engagement. Unlike commercial platforms, these publicly funded services have a responsibility to remain alert to biases that emerge in algorithmic recommendations. In this work, we conduct a systematic evaluation of popularity and author nationality bias in Booklens, a non-personalized item-to-item recommender used in Danish public libraries that receives loans, clicks, mood tags, and creator name as input. We prompt the system with a set of 10,000 books of varying popularity and author nationality and analyze the resulting recommendations under different parameter configurations controlling (e.g., recommendation list length). For each configuration, we measure overall popularity bias and resulting bias toward authors of different nationalities. Our analysis reveals that popularity strongly drives recommendation outcomes, with up to a 40% relative decrease in exposure for less popular nationalities depending on the parameter setting. We also show that certain configurations significantly amplify or decrease these disparities. These results highlight the need for systematic bias analysis to be a structural component of recommender systems evaluation, supporting libraries in aligning algorithmic behavior with social and cultural missions.

This chapter is based on *Daniil et al., Bias in Book Recommendation: A Case Study on the Danish Public Libraries, ECIR '26 [60].*

7.1 Introduction

Book recommendation was among the first applications of automated recommendation. Amazon launched as an online bookseller in 1994, only to later dominate e-commerce [211]. The incentive behind automating book recommendation is to fast-track the process by substituting the role of the librarian or employee in a brick-and-mortar bookstore, while giving access to a wide range of material. Algorithms function by synthesizing numerical identities for books/users, and computing high-dimensional relations in a purportedly objective manner. A motivation for relying on algorithmic recommendation is the perceived 'binary between the subjective decision-making of humans and the computational objectivity implicated in cultural imaginings of how algorithms work' [158].

Despite these cultural imaginings, the purported objectivity of algorithms has been questioned repeatedly. Various systems have been shown to commit faux pas by stereotyping, pigeonholing users, or over-recommending popular items [89, 159, 216]. Additionally, algorithms do not merely mirror users by statistically processing, analyzing, and interpreting their behavior; rather, they also shape it [11, 40, 141]. These systems are constructed, optimized, and evaluated with the advancement of organizational goals as the main objective.

In libraries, much more than the goal of increasing user engagement and consumption is at stake. Libraries are institutes where human values have been traditionally safeguarded and promoted via culture dissemination. At the same time, there is a perception that for public libraries to be able to exist in an economic space driven by automation and speed, they need to satisfy their patrons who are used to interacting with data-driven models [100, 184]. This tension requires careful management. Libraries need, more so than other entities, to be able to assess and control bias in their data and system [140].

In this chapter, we conduct the first study of bias in a real library recommender system. We investigate *Booklens*, a non-personalized item-to-item recommender built for and used by the Danish public libraries, in the general website (`bibliotek.dk`) and municipal library websites. We study two bias types: *popularity bias*, the tendency to disproportionally recommend popular items [7], and *author nationality bias*, the underrepresentation and/or overrepresentation of author nationality groups in the recommendation [57]. We collect a set of books from the Danish library collection, record their click-based popularity, enrich them with author nationality information when publicly available, and study the relationship between popularity and author nationality. To measure bias propagation, we prompt Booklens with a data sample and observe the popularity and author nationality shift in the recommendations. Additionally, we investigate the effect of a set of Booklens parameters on popularity and author nationality bias, as they are known to impact the extent to which bias occurs [59].

We answer the following research questions:

1. Does Booklens exhibit popularity bias?
2. Does Booklens exhibit bias towards author characteristics?
3. What is the impact of various system parameters on bias propagation?

We find that Booklens favors popular items in the recommendation, even when prompted with niche items. Additionally, author nationalities whose books are less popular are especially underrecommended even though they comprise a substantial part of the collection. We observe that bias propagation persists through repeated recommendation cycles. Finally, we report that tweaking various Booklens parameters (e.g., how many books from the same author can be recommended at once) can significantly influence the popularity and author nationality distribution; however, the general recommendation tendencies persist given that the system's characteristics (i.e., reliance on behavioral data) remain. We conclude that bias analysis should be a structural step of incorporating automation in library discovery systems.

7.2 Related work

AI technologies are emerging in library settings in recent years. The growth of AI-powered methods in use across fields and data types means that libraries follow along in their adoption of such methods [36]. Applications such as chatbots and automated library discovery systems promise to 'streamline cataloguing, classification, and recommendation processes, enabling efficient information access for patrons' [218]. At the same time, efficiency and convenience are only a small part of library's requirements; it is their duty to responsibly empower users by adopting algorithms that, among others, do not perpetuate bias [100]. Yet, bias is undoubtedly present in various library systems [184], as both classification and information access systems reflect the social context that they exist in [193].

Bias in recommender systems has been extensively studied from a computational perspective [42]. Various papers have focused on measuring and mitigating bias in domains such as music [39, 77, 128] and news [14, 172]. In the literary domain, publicly available datasets with user-book interactions (e.g., ratings and loans) have been shown to follow a long-tail distribution: a few books have been interacted with by many users, while the vast majority of books have barely been interacted with. When algorithms are trained on such datasets, the uneven distribution of popularity has an impact on the recommendation, with different systems being more or less vulnerable to over-recommending already popular items [121, 159]. Additionally, statistical biases that algorithms are prone to may lead to discriminatory outcomes if the margins of distributions correlate with certain demographics [216]. Research on datasets stemming from social book cataloguing websites has shown such effects from the perspective of author demographics, such as gender [68] and nationality [57]. To our knowledge, this is the first computational study of popularity and author bias applied on a real library recommender based on user behavioral data.

7.3 Book recommendation in Danish public libraries

In this section, we describe in detail the Booklens system. Booklens was developed by DBC Digital, the software company that supports the IT infrastructure of the Danish public libraries. DBC is owned by KL (Local Government Denmark) and has a key role in the Danish public library sector. Among other responsibilities, they have created and maintain a set of recommender systems to assist public library patrons to find

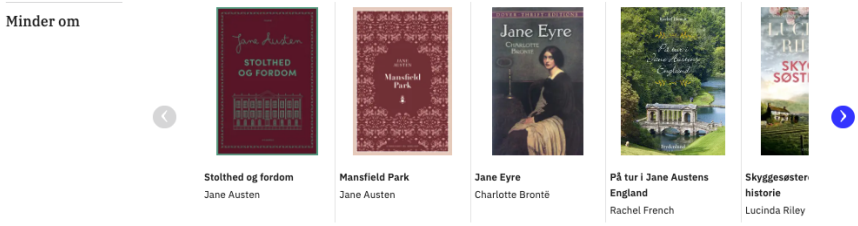


Figure 7.1: 'Reminds of' (*'Minder om'*) list under *Pride and Prejudice* in `bibliotek.dk`.

suitable material in the library websites. Booklens is the system behind the 'Reminds of' banners displayed under every item in *bibliotek.dk*, the general public library website (see Figure 7.1). Various data sources of either behavioral or content information are used to compute item-item similarity, which the system then aggregates and accordingly constructs the 'Reminds of' banner with the top-10 items.

Data sources

To compute similarity between items, the following data sources are considered: loans, `bibliotek.dk` search clicks, municipal library search clicks, reading-experience tags, and creator name. Information on these sources is shown in Table 7.1. For each data source, similar items are identified based on the number of shared features (e.g., users who have borrowed the same item). For each item, the items most similar to it according to each data source are filtered to only keep those with matching metadata (type, language, audience), which are stored in the database. The full overview of the process is described in Algorithm 1 in our repository[1].

Source aggregation

To produce the final recommendations based on a target item, Booklens retrieves the candidates deemed the most similar by each source, filters them to satisfy a cover requirement and matching metadata (type, audience, location), merges them into a list and orders them based on predefined weights for each source. A final subset is selected by enforcing a maximum number of appearances for each author and a positional offset. The process is described in Algorithm 2 in our repository. The data sources are re-constructed and aggregated weekly with the following default parameter values: maximum two items from the same author, zero offset, ten final recommendations.

[1] `https://github.com/SavvinaDaniil/RecBiasInDanishLibraries/`

Table 7.1: Data sources and features used by Booklens to find similar items.

Source	Description	Feature
Behavioral data sources		
Loans	Historical loan data from public libraries (20–8 years ago). DBC does not have access to more recent loan data.	User ID
Bibliotek.dk search clicks	Clicks from sessions following searches on bibliotek.dk. No user data is available, only session IDs. Data are continuously updated.	Session ID
Municipal library (ML) search clicks	Daily-sampled search clicks from municipal library websites. No user data is available, only session IDs.	Session ID
Content data sources		
Metacompass tags	Expert-created ontology of approximately 6000 tags (e.g., *fast-paced*, *humorous*) describing the reading experience. Tags cover a few thousand mainly Danish fiction titles.	Tags
Creator	Author(s) or other creators of an item.	Creator ID

7.4 Research design

In this section, we describe the process we follow to enrich the data with author information and our methodology for measuring bias in the recommendation.

7.4.1 Data Collection and Enrichment

For our experiments, we collected search click counts of books from bibliotek.dk between 17-2-2015 and 11-3-2025. Around 1.8 million unique books were clicked at least once in this period. We use the number of clicks of a book to represent its popularity level. We sample at random 10,000 books for our experiments.

We use external sources to enrich the dataset with additional author information when publicly available. Specifically, we link our dataset to the Virtual International Authority File (VIAF)[2], a service hosted by OCLC that combines various authority files (e.g., digital author entries from national libraries) into a single entry. As a result, for each author multiple versions of their name are present, which facilitates the process of linking our dataset to VIAF. Additionally, VIAF often provides author demographic information such as nationality, gender, occupation, birth and death data, as well as link to their Wikidata entry. In recent months, VIAF has decreased the API request rate limit and currently does not facilitate provisions for researchers. However, data dumps last updated in 2024 are publicly available, which we downloaded in the xml format.

To link our dataset with VIAF, we take the following steps. We collect entries tagged as *Personal* (i.e., authors) from the xml file. We store the following information when available: name(s), viaf ID, nationality, birth year. We store all entries along with

[2] https://viaf.org/

Table 7.2: Characteristics of the full and sample datasets.

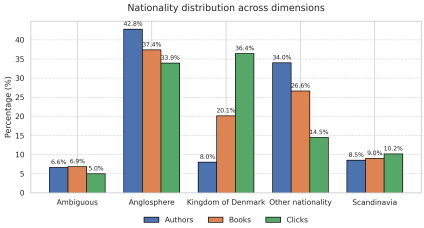
	All items	Sample items
Total num. books	1,779,709	10,000
Total num. clicks	32,958,244	181,936
Perc. books with known author	82.9%	90.6%
Perc. books with known author nationality	60.2%	60.2%
Num. unique authors	611,491	7,866
Average book popularity	18.5	18.2
Median book popularity	3	3
St. dev. of book popularity	96.8	73.8

the available information in an ElasticSearch index. Finally, we link our dataset to the index based on author name. An expected issue is that various authors have the same name. In case there are multiple results but they all point to the same nationality, we keep this nationality. If nationalities vary, we mark the item as *Ambiguous.* Items whose author could not be matched with VIAF were marked as *None.* The enrichment process is detailed in our repository. Given that many author nationalities are present in the dataset, we categorize them as follows: *Kingdom of Denmark* (Denmark, Greenland, Faroe Islands), *Scandinavia* (Norway, Sweden), *Anglosphere* (USA, Canada, UK, Australia, New Zealand), *Other.* Our motivation for this categorization is that on the one hand we expect Danish authors to be popular in the dataset, and other Scandinavian authors as well given the cultural proximity. We further extract the Anglosphere countries given the big hand they are known to have in the global market, as well as results from previous work that is based on American datasets [57].

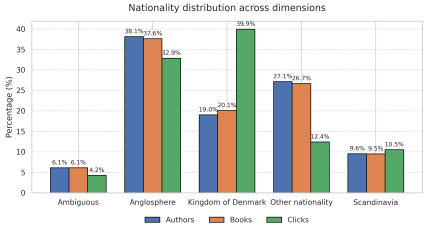
In table 7.2, we can see the characteristics of the full and sample datasets. As shown in figure 7.2c, both datasets follow a long tail distribution, which means that popularity bias is very much present in the data. In both cases, more than 30% of the items have only been clicked once in the 10 year period that we are examining. Figures 7.2a and 7.2b show the distribution of our predefined (available at VIAF) author nationalities in the full and sample dataset respectively, among unique authors, unique books, and clicks. In the sample dataset, while Other-authored books take up more than a quarter of the dataset, they correspond to only 12 percent of the clicks. Similarly, Figure 7.2d shows that Other-authored books have lower popularity on average than all other categories.

7.4.2 Measuring Bias in Booklens

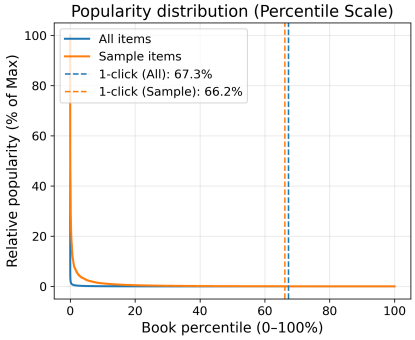
To estimate bias in the system, we adopt the following strategy. First, for our sample of $10,000$ books as prompting items, we trigger a recommendation process using Booklens, and measure popularity and author nationality bias in the result. Second, for the same sample, we trigger 15 consecutive recommendation processes by assuming that every time the user clicks the first item. We calculate bias in each recommended set and report the evolution. Finally, we trigger a recommendation process on the same sample while varying the configuration of a set of Booklens parameters: limit, offset,



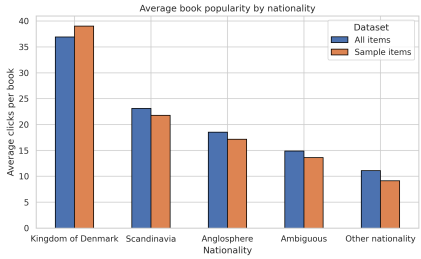
(a) Nationality distribution in the full dataset among unique authors, books, and clicks.



(b) Nationality distribution in the sample dataset among unique authors, books, and clicks.



(c) Normalized popularity distribution of full and sample dataset.



(d) Average popularity per nationality in the full and sample dataset.

Figure 7.2: Popularity and nationality distributions in the full and sample datasets.

author_max, source weights, and the cover requirement. We report the impact of each parameter on bias.

Inspired by Klimashevskaia et al. [118], we measure the following for popularity bias:

1. **Average Recommendation Popularity (ARP)**: Average number of clicks the recommended items have received, averaged across prompting items [4].
2. **Popularity Lift (PL)**: Average difference between popularity of a prompting item and the average popularity of its recommended item [10].
3. **Prompting/recommended correlation (Corr)**: Correlation between popularity of a prompting item and the average popularity of its recommended items, averaged across prompting items. It ranges from 0 to 1.
4. **Popularity Bias (PopBias)**: Inspired by NDCG, it expresses whether the items are ranked similarly to if they were ranked based on popularity, averaged across starting items. It ranges from 0 to 1 [35].
5. **Average Popularity of items that receive Recommendation (AP Recs)**: Given that the system is not able to construct a 'Reminds of' list for all items (as for many of them there is no sufficient information), we measure the average popularity of the items that do receive recommendations.
6. **Average Popularity of items that do Not receive anyRecommendation (AP NoRecs)**: Similarly, we measure the popularity of the items that do not receive any recommendations.

For author nationality bias, we measure the following:

1. **Relative exposure difference**: Relative difference between the presence in the prompting and recommended set for each predefined nationality category.
2. **NoRec**: The percentage of books from each nationality category that do not receive any recommendations.

7.5 Results

In this section, we answer the research questions as posed in section 7.1. First, we discuss the popularity and author nationality bias in the default Booklens version, and then we examine the effect of various system parameters.

7.5.1 Bias in the default Booklens version

The current version of Booklens clearly exhibits *popularity bias* according to all tested metrics as seen in Table 7.3. The recommended items have been clicked an average of 301 times, while the prompting items have been clicked only about 18 times on average. The average popularity lift is 62, which means that on average, the recommendations are much more popular than the item that prompted them. There is a weak yet significant positive correlation between the popularity of the prompting item and that

Table 7.3: Popularity bias analysis for the default Booklens system. * indicates that the correlation is significant ($p<0.05$).

Metric	ARP	PL	Corr	PopBias	AP NoRecs	AP Recs
Result	301.36	61.97	0.2263*	0.6974	3.09	35.89

of its recommendations. PopBias is approximately 0.7, indicating that items are roughly ranked in line with their popularity. Finally, the prompting items that receive no recommendation have only been clicked an average of 3 times, contrast to the average of 36 clicks that the rest of the prompting items have received.

Figure 7.3 shows the evolution of popularity in the prompting set and ensuing 15 runs assuming the first recommended item is clicked. After a few runs, the item popularity spikes and then drops, presumably since the very popular items have already been reached. The popularity converges on about the same levels as the first run, which still comprises of much more popular items than the prompting set on average. Additionally, by the end 88% and 96% of the items have popularity higher than the mean and median of the full dataset respectively, indicating strong popularity bias. Items with only 1 or no clicks account for 3% of the recommended lists. At the same time, out of the items in the full dataset with the top 20% unique highest popularity (about 400 items out of 1.8 million), more than 1 in 10 appear in the recommended lists in every step. Finally, in this process the maximum number of unique items that the user is exposed to settles quickly to about double of the initial items and no new items are recommended.

In summary, Booklens consistently amplifies the prominence of popular items. Across all metrics, books with substantially higher click counts than those in the prompting set are favored, and this tendency persists across repeated recommendation cycles. Less popular items are underrecommended, and they often do not prompt recommendations either.

In terms of *author nationality bias*, Danish-authored books dominate the recommendation compared to the prompting items, as seen in Table 7.4. Books from other Scandinavian countries appear about equally in the prompting and recommended sets. The rest of the categories are underrepresented in the recommendation, especially the non-Anglosphere items. 68% of Other-authored books do not receive any recommendations, presumably because the system does not have enough information on them due to their lower popularity.

After 15 runs where the first item is clicked, Danish-authored books are still the majority and books by Norwegian and Swedish authors consistently take up 10% of the recommendation. Notably, books written by authors from the Anglosphere appear more than in the first few runs and settle to almost one third of the recommendations. On the contrary, books by the remaining nationalities (more than 100 nationalities in total), continue to be underrepresented and the effect worsens after the first few runs.

In summary, Booklens exhibits author nationality bias. Danish-authored books dominate the recommendations, while authors from other Scandinavian countries appear with roughly equal representation as in the prompting set and those from other nationalities are underrepresented. This imbalance persists across repeated recommendation cycles, with Anglosphere authors gaining moderate visibility over time, whereas books by other nationalities remain underrepresented.

Figure 7.3: Popularity evolution in the initial dataset and 15 runs.

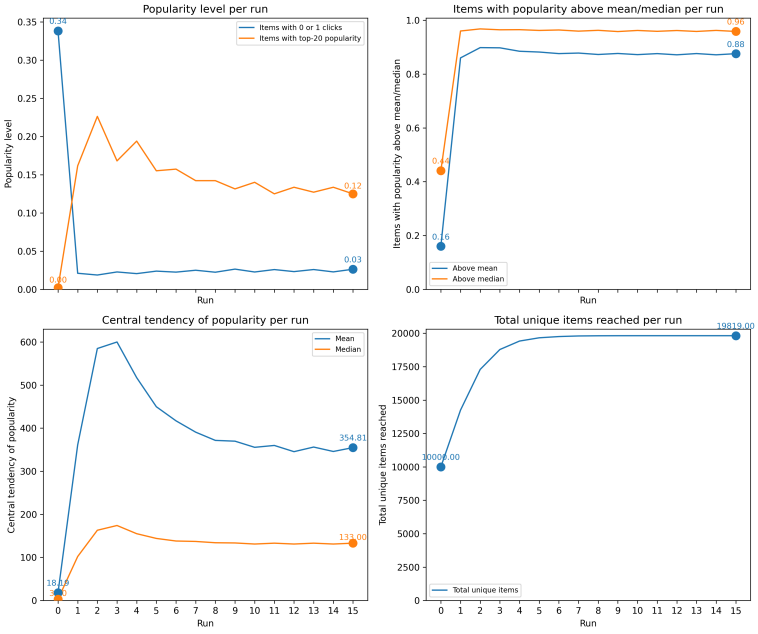
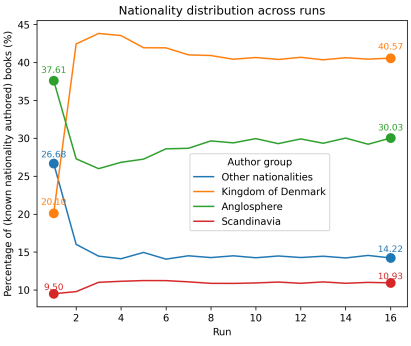


Table 7.4: Author nationality bias analysis for the default Booklens system.

Author nationality	**Rel. exp. diff. (%)**	**NoRec (%)**
Kingdom of Denmark	115.76	38.76
Scandinavia	2.33	49.92
Ambiguous	-17.58	56.99
Anglosphere	-28.07	53.31
Other nationalities	**-43.89**	**67.84**

Figure 7.4: Nationality distribution in the initial dataset and 15 runs.



7.5.2 Effect of system parameters

Table 7.5 shows the median, minimum and maximum value of each metric per parameter value. The cover requirement strongly affects all metrics; ARP, AP Recs, PL and PopBias significantly decrease when the cover requirement is dropped. On the contrary, Corr is higher, which means that popular items prompt popular items even though the general popularity is lower. The representation of items written by Other nationalities is significantly higher. Allowing more books from the same author decreases most popularity bias metrics, even though the result is not significant. High limit (i.e., number of recommended items) impacts the metrics differently: it significantly decreases PL and increases the presence of Other-authored items, but also increases Corr. Configuring the offset similarly results in varying results. When the offset is high, ARP is also high, which means that items down the similarity ranking are even more popular. However, PL is lower. Finally, among the weights, prioritizing the mood tags results in significantly lower popularity bias for most metrics. However, it does not positively impact the presence of Other-authored items, since the mood tags currently cover only Danish fiction titles. Prioritizing loans or clicks from biblioteek.dk increases most popularity bias metrics. None of the weight configuration has a significant impact on the recommendation of authors from Other nationalities.

In summary, most of the parameters affect the metrics. Filtering for books with cover images biases towards popular content and reduces representation of less popular author nationalities. Prioritizing mood tags over behavioral data results in lower increase in popularity. Limit and offset impact the result differently based on the chosen metric. Regardless of the configuration of parameters, all experiments propagate popularity and author nationality bias to different extents.

Table 7.5: Median of bias metrics by parameter value. Default parameter values are underlined. * indicates that the given median is significantly different than the default configuration's median, given a non-parametric posthoc Dunn's test.

Param.	Val.	ARP	AP Recs	PL	PopBias %	Corr %	Rel.exp.dif. %(Other)	NoRec %(Other)
Author Max	0.1	252.8	37.1	44.8	62.4	38.6	-39.4	69.6
	0.2	248.3	36.7	44.0	62.2	38.5	-39.6	68.9
	1	237.9	36.3	43.0	62.6	38.4	-39.8	68.3
Limit	10	247.7	37.1	44.6	67.9	29.0	-40.7	69.6
	20	246.1	36.7	43.6	62.7	36.8*	-40.4	68.9
	100	251.5	36.3	41.5*	57.5	44.4*	-38.8*	68.3
Offset	0	242.3	34.3	47.1	62.3	37.5	-39.2	65.7
	2	250.6	36.8*	45.9	62.5	38.3	-39.8	69.2*
	5	251.2	38.6*	45.1	62.6	35.4	-39.8	71.1*
	10	257.5*	40.9*	42.1*	62.6	40.4	-39.6	73.0*
Cover req.	False	192.4*	35.2*	31.8*	59.8*	43.5*	-37.0*	66.7*
	True	312.2	39.5	58.0	64.6	34.5	-42.5	72.1
Weights	Def.	259.5	36.8	43.2	62.6	35.8	-39.7	68.9
	ML	235.4	36.8	43.1	62.4	16.0*	-39.5	68.9
	Loans	283.4*	36.8	44.7	63.7	41.3*	-39.7	68.9
	bib.dk	280.9*	36.8	44.7	64.0	49.3*	-39.3	68.9
	Tags	212.4*	36.8	40.2*	62.2	23.8*	-39.5	68.9
	Auth.	245.1	36.8	43.6	61.4	36.9*	-39.8	68.9

7.6 Discussion

7.6.1 Behavioral data and popularity

Across experiments and metrics, popularity bias is undoubtedly present. Booklens depends on behavioral data and thus reproduces existing popularity patterns. Popular items have been clicked by the same users as many other items, and thus the system has a substantial level of knowledge on them. Given that the clicks stem from searching a term and then clicking different items in the ensuing session, the popularity of already popular items is bound to increase. On the contrary, items that have rarely if ever been clicked are practically unknown to the system, and as long as similarity mainly depends on behavioral data, they will remain unknown unless explicitly searched for. Even though mood tags have the potential to increase the visibility of tail items, they now only cover a small part of the collection. While one of the purported positive effects of algorithmification is enabling the long tail (since users can browse and discover items they might not notice in a library), an explicit mechanism would need to be incorporated in Booklens to promote less popular items.

A core principle of the library is to provide space for a diversity of voices and perspectives [105], and overrepresenting already popular items may undermine this principle. At the same time, the scope of this research is limited to popularity bias as we have no sufficient access to the system to evaluate how it performs from, for example, a user satisfaction perspective. It is therefore inappropriate to claim that dependence on behavioral data is inherently wrong. The presence of popularity bias is merely one of the criteria that should comprise an extensive computational evaluation of performance and utility. Recommending popular items to drive engagement might be a valid strategy for the library, and disentangling conformity from popularity is its own research goal [261]. Furthermore, the implications of popularity bias on other author characteristics should be analyzed contextually. One of the missions of Danish libraries is to communicate Danish cultural heritage [189], which may justify a stronger emphasis on Danish authors. Overall, this work highlights the importance of treating bias evaluation as an essential part of system design, so that the results can be interpreted appropriately and incorporated into future development.

7.6.2 Information on authors

Enriching datasets with demographic information by linking them to online data sources is a nontrivial task. First, this information is rarely available and findable. Second, disambiguation is its own complex field of study. Third, even if this information exists and can be reliably disambiguated, it has the potential to be incomplete or even incorrect because of human error, bias, system limitations etc. For example, VIAF currently only categorizes authors as *Male* or *Female* even in cases where the equivalent Wikidata entry indicates a diverse gender identity. We do not attempt to make normative, general and definite statements about the relationship between author nationality and popularity outside the limits of this experimentation. This work aims, among others, to give context to the often abstract concept of long tail and highlight that popularity is not a standalone property and can significantly correlate with other item characteristics.

7.6.3 Theory of change

This work serves as the first extensive study of popularity bias propagated by a real library system, as well as of its consequences when it comes to the system's treatment of authors with various national origins. With this study, we aim to support DBC in their effort to evaluate the system they developed by investigating its vulnerability to two different types of bias and proposing a pipeline of evaluation. Consequently, this work serves as a diagnostic step for Danish public libraries, as well as libraries in general that wish to fulfill their social obligation of recognizing bias and keeping their systems in check. We contribute to the conversation around incorporation of automation in library discovery by showcasing the need to measure the extent to which a system might underrepresent certain kinds of items, both from an economic perspective (i.e., their popularity) but also as it relates to those items' societal characteristics. We computationally show that defining similarity between items assumes a set of heuristic choices that render the evaluation of their effect on bias essential.

At the same time, our work is not the end of the road. For one, author nationality is only one item characteristic that might be subject to bias. While our process of enriching the dataset is not foolproof, an author's nationality is still one of the easiest characteristics to discern. Computationally assessing bias is limited by the information available and quantifiable. Accounting for bias in library data and discovery systems is a broad endeavor that goes beyond bias metrics; it involves evaluating the entire pipeline of collection and dissemination of items. Solely relying on bias metrics without considering the entire library ecosystem has the potential of increasing practitioners' blind spots instead of shedding light on them.

7.7 Conclusion

In this chapter, we presented the first study of bias in a real and current library recommender system: *Booklens*, a non-personalized item-to-item system that is used by the Danish public libraries and was developed by DBC digital. We investigated popularity bias propagated by the system, as well as ensuing author nationality bias. To this end, we enriched a dataset consisting of books from the online library platform with author nationality information from VIAF. We found that Booklens is very prone to popularity bias based on a set of commonly used metrics. Additionally, Booklens underrecommends books authored by nationalities that are not very popular in the dataset. Finally, we observed that tweaking some system parameters can reduce bias, but it is still largely present. We believe that this work can serve as a step towards constructing a framework for evaluating bias exhibited by library systems. Future work can follow a similar process to investigate bias in other book recommender systems used in libraries.

8

Conclusion

This thesis has focused on the presence of bias in book recommender systems, with an emphasis on the public library sector. We approach this topic from three perspectives: i) understanding the ethical issues around book recommendation, ii) understanding the challenges of studying statistical bias, and iii) understanding the implications of statistical bias in book recommendation.

In this chapter, we summarize our key findings across the six research questions, reflect on the domain, and propose future research directions.

8.1 Findings

Part 1: Understanding ethical issues of book recommender systems

RQ 1.1 How has bias in book recommender systems been studied in computer science research, and what gaps exist in current approaches?

To answer **RQ 1.1**, we performed the first systematic literature review of bias in book recommendation in Chapter 2. We collected 40 papers published in computer science conferences and journals and analyzed them across components related to bias: *type* (statistical or social), *subject* (user, item, or joint), and *action* (measurement or mitigation). We organized the papers across these dimensions and further broke down bias type into subcategories, along with different metrics that are used to measure them. We found that most relevant studies concern themselves with popularity bias, where often the dataset with user-book interactions is one of multiple datasets that researchers experiment with. Additionally, when it comes to social bias, only a few features have been studied as potential recipients of bias (author nationality and gender, user age and gender, and publisher location), presumably due to data availability limitations. We highlighted gaps in the literature, such as the lack of comparative study among different bias mitigation methods in the book domain and the limited data availability. Finally, we emphasized the need for interdisciplinary research and recommended the following future directions to achieve it: reliance on LIS theoretical frameworks, engagement with literary recommenders currently in use, and establishment of shared terminology

and research practices among computer scientists and LIS researchers. Based on these findings, we propose that future research on bias in book recommendation should not be limited to treating books as interchangeable items across domains and instead explicitly design research questions and objectives in relation to the book domain.

RQ 1.2 How do practitioners in public service media conceptualize diversity in recommender systems?

In Chapter 3, we presented the results of an interview study among practitioners in three public service media organizations in the Netherlands: a broadcaster, a news organization, and a library. We conducted semi-structured interviews where the participants explained how they conceptualize diversity in general, in the context of their respective organizations, and in recommender systems. The last part of the interview contained a small experiment that involved the ranking of a set of items with diversity in mind. We recorded, transcribed, encoded and thematically analyzed the interviews and identified the following components as relevant to diversity in recommendation: the *goal* of recommendation, the *aspects* to be considered, and the *tactic* that should be employed to produce diverse recommendations. To answer **RQ 1.2**, we listed the various conceptualizations discussed by the participant along the three dimensions and examined the interplay between them. We concluded that diversity encompasses a wide range of interpretations and that normative choices need to be made in every step of the way towards operationalizing it in a recommender system that serves public service media. The existence of multiple interpretations of diversity alludes that researchers designing recommender systems for public institutions such as libraries should engage with practitioners to ensure alignment with organizational values.

Part 2: A critical look into measuring statistical bias in book recommender systems

RQ 2.1 Which factors lead to divergent results among different studies of popularity bias in recommendation?

In Chapter 4, we conducted an extensive reproduction study of three recent papers on popularity bias in media recommendation and its effect on users with different propensity towards popular items in the movie, music, and book domain. To answer **RQ 2.1**, we reproduced the three studies and identified the following aspects as potential sources of the difference in results: data, algorithms, division of users in groups, and evaluation strategy. We ran a set of experiments in which we measured popularity bias and unfairness of popularity bias with various combinations of these aspects. We found that all aforementioned aspects impacted popularity bias, with the choice of evaluation strategy playing a significant role. Based on the apparent sensitivity of popularity bias propagation on various system tweaks, we proposed that conclusions should be explicitly made within the limits of the experimentation and contextualized accordingly. We believe that researchers studying recommender systems bias should systematically report methodological choices that have the potential to significantly impact the result.

RQ 2.2 What is the effect of data characteristics and algorithm configuration on popularity bias?

To answer **RQ 2.2**, we experimented with data characteristics and algorithm configuration to observe the effect on popularity bias, and we presented the results in Chapter 5. We chose two data characteristics that might impact popularity bias, namely the relation between popularity and rating and the preferences of users with large profiles, and analyzed them for three publicly available datasets. We altered these characteristics to define a set of data scenarios and accordingly synthesized datasets. Furthermore, we identified configuration choices that might impact popularity bias for three widely used algorithms implemented in three open frameworks and performed training and testing for each dataset (real and synthetic) and algorithm version. We found that the presence and magnitude of popularity bias are highly sensitive to the tested characteristics and their interplay. Consequently, we highlighted the importance of specificity around the design choices of bias studies and need to experiment with various data and algorithm characteristics when studying bias propagation. These findings indicate that achieving generalizability across datasets and implementations is inherently unlikely; instead, specificity is more desirable with relation to real systems.

Part 3: Societal implications of statistical bias in book recommender systems

RQ 3.1 What is the relationship between popularity bias and bias toward author nationality based on a publicly available dataset with user-book interactions?

To observe whether popularity bias leads to bias toward author nationality, hence answering **RQ 3.1**, in Chapter 6 we experimented with a publicly available dataset with user ratings of books, Book-Crossing. We enriched Book-Crossing with author information based on Linked Open Data sources: Google Books, VIAF and Wikidata. Subsequently, we trained and tested a set of collaborative filtering algorithms on the user-item interactions, following the same pipeline as previous research on popularity bias. We found that books written by American authors are significantly more popular than the rest of the books, and that this phenomenon propagates in the recommendation; algorithms that are prone to popularity bias are consequently prone to over-recommending American-authored books. We concluded that this line of research is necessary in order to identify blind spots that can cause bias against certain author groups. Future work on bias in book recommendation should further examine whether and how book characteristics, author or otherwise, interact with popularity and indirectly affect the recommendation.

RQ 3.2 To what extent does a real library recommender system propagate popularity bias and associated author nationality bias?

To answer **RQ 3.2**, in Chapter 7 we applied a similar line of research as in Chapter 6 on a library recommender system in production. As far as we know, this is the first audit-type study of bias perpetuated by a library recommender system. In collaboration with DBC Digital, the company that supports the IT infrastructure of the Danish public libraries, we investigated popularity and author nationality bias propagation by Booklens, the non-personalized item-to-item recommender that is present in `bibliotek.dk`. We gathered a dataset of around 1.8 million clicks between 2015 and 2025, defined book popularity as the number of clicks in this period, and enriched the

dataset with information on the author using VIAF. We extracted 10 thousand books for our experimentation, and compared the data characteristics between the full and sample dataset; in both datasets, the clicks per book follow a long-tail distribution, and popularity correlates with author nationality. We found that Booklens is very prone to popularity bias, and that it underrecommends books authored by less popular nationalities. We also found that, while tweaking some system parameters can reduce bias, it is largely present in all versions. We concluded that the process of developing recommender systems for libraries should include bias evaluation as a core step. The incorporation of recommender systems in public libraries requires the development of clear criteria of utility and value alignment, with bias being a crucial consideration.

Synthesis

The findings in Part 1 establish that incorporating organizational values in book recommender systems is an inherently normative process, and that this is often ignored in academic research. The interview study showed that public service media practitioners conceptualize diversity in varying and even conflicting ways, while the literature review exposed gaps in computer science research on bias in book recommendation: there is limited engagement with prior work by the LIS community and with literary recommenders in operation. Together, these findings highlight the importance of moving away from domain-blind solutions and towards a grounded understanding of bias and diversity in book recommendation.

The findings in Part 2 show that statistical bias is highly sensitive to methodological choices. Across the experiments, we showed that the measurements of popularity bias vary substantially depending on factors such as evaluation strategy, dataset characteristics and algorithm implementation details. Accordingly, this thesis argues against viewing recommender system bias as an inherent property of algorithms, but rather as an outcome of interactions between technical design decisions and the environments in which systems are trained and operate.

The findings in Part 3 show that methodological observations have broader social implications. In both chapters, we demonstrated that popularity bias as an algorithmic mechanism can reinforce existing disparities in the visibility of authors of different current popularity and from different national backgrounds, yet these dynamics differ across contexts and are influenced by changes in implementation. These results support one of the central arguments of this thesis: that book recommendation bias is closely tied to cultural and organizational context. They also highlight that technical decisions and processes can have unforeseen consequences for cultural exposure and representation.

Taken together, the findings across the three parts of this thesis show that bias in book recommender systems cannot be addressed as an isolated technical issue. Across all three parts, this thesis advances the understanding of book recommender systems in three ways. First, by showcasing that bias in book recommendation cannot be studied independently of the domain and context in which the recommender system operates. Second, by suggesting that bias in recommender systems is not purely algorithmic, and emerges through the interplay of data availability, system design, evaluation methodology, and organizational values. Third, by arguing that

recommender systems used in libraries and other public service cultural institutions should explicitly incorporate bias considerations as part of their design and evaluation processes.

8.2 Reflections & Future directions

Throughout the research conducted to complete this thesis, we had the opportunity to reflect on the state of the domain, the challenges we encountered, as well as the avenues for future research that lay ahead. We detail these reflections below.

Challenges in evaluating bias

As mentioned repeatedly throughout this thesis, bias in automated book recommendation is an underexplored domain, which naturally leads to research challenges. For one, the data availability is limited. To train and test collaborative filtering algorithms, researchers need behavioral data in the form of ratings, clicks, loans or other type of interactions. Yet, only a few standard research datasets are currently publicly available. On top of that, the datasets that are most often used by researchers stem from US-based sources such as the Book-Crossing community, the Amazon e-commerce website, and the book cataloguing social media Goodreads. It is expected that these datasets contain largely American authors and users, which harms the transferability of the conclusions to other contexts. In our own research we found that, while American authors are very popular in Book-Crossing, this is not the case in the Danish online library data, which led to different observations around recommendation trends: American authors were over-recommended by algorithms trained on Book-Crossing (Chapter 6) but not by the Danish library recommender (Chapter 7). These findings suggest that researchers should validate whether trends observed in commonly used research datasets also hold in a real-world context. Additionally, standard research datasets tend to be outdated, and it is not trivial for researchers to access more recent data. For example, in line with recent trends of media platforms, Goodreads deprecated their API in 2020. Working with real-world systems, such as library recommender systems in production, can partially mitigate this issue by providing access to more current interaction data.

Research on bias often does not only require interaction data, but also external information. It is not possible to measure whether certain groups of authors are underrepresented or groups of users underserved when no information is available on either side. In some of the datasets, user characteristics such as age or location are included. When it comes to the authors, researchers need to link the datasets with publicly available information, which is often neither complete nor necessarily accurate. Disambiguation is more challenging for authors who are less known internationally – which was the case for the Danish library dataset. Sources such as WikiData have been found to be unbalanced among countries in terms of information completeness. Different use cases may require different pipelines to locate, extract, and disambiguate author information.

Another hurdle is that work on bias in book recommendation seems scattered and unfocused, as we discussed extensively in Chapter 2. Many studies that measure or mitigate bias in recommendation test their methods on many datasets from different

domains, as this is the standard research practice. While it is understandable that researchers try to show the applicability of their method in a variety of contexts, this means that it is harder to make conclusions on the progress made in the book domain in particular. This challenge also emerged clearly from the findings in Part 2 of this thesis (Chapters 4 and 5), where methodological choices that are often left implicit were shown to lead to substantial differences in measured bias. An additional consequence of this practice is that many studies do not incorporate the characteristics of user-book interactions in their setup. For example, "popular" might mean something different between types of media and cultures, which renders the use of universal popularity thresholds (e.g., the top 20% most consumed items) inappropriate. Defining popularity thoughtfully and specifically is an arduous yet necessary step in the research process. Responsibility for addressing this lies primarily with researchers, who should document and justify such design choices in relation to the domain and context.

Future directions

The findings of this thesis, along with the concrete challenges outlined in the previous section, highlight the importance of specificity in book recommendation bias research. We as researchers need to be specific in the way we design our methods, define the parameters, express our conclusions and assess their generalizability and limitations. While research trends may call for universally applicable methods and magic pill solutions, contextual, normative and application-based research is very much needed in the current landscape. Interdisciplinary collaborations and consultations with practitioners who specialize in a given domain of application can help recommender systems researchers think through what is important to study and how it should be approached. Parts of this thesis already reflect such collaboration, for example through working with up-to-date Danish library data and interviewing practitioners on their conceptualization of recommender systems values. In addition, focusing on real-world systems can help address the challenges related to data and metadata availability. Working with institutions like public libraries gives access to up-to-date interaction data, instead of relying on a few old benchmark datasets. It also allows researchers to think together with practitioners about which features really matter, rather than only analyzing what is easy to access, and to consider ways of collecting missing features, for example through small targeted user studies. We believe that this is the way ahead for researching bias in book recommendation.

Throughout human history, books have proven to be one of the most significant cultural artefacts. This is so widely understood, that it reads as a nonstatement – yet it's true. We as a society and libraries in particular have an incentive to promote books and reading, and there is real potential in employing computational tools to achieve this. Automated recommender systems will never be unbiased, because humans are not unbiased. Bias is a term used to express that a system displays tendencies that may have societal implications and harmful consequences. Libraries should not aim to be unbiased altogether, but rather ensure that they exert control over their systems, put safety mechanisms in place and define clear standards to turn to when needed. In this thesis we have showcased what types of problems can arise in the context of book recommendation bias, and hope that public libraries can benefit from this research to

inspire their operationalization of automated recommendation and evaluation thereof.

Bibliography

References

[1] Svanhild Aabø. The role and value of public libraries in the age of digital technologies. *Journal of librarianship and information science*, 37(4):205–211, 2005.

[2] Himan Abdollahpouri. Popularity bias in ranking and recommendation. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 529–530, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314309. URL https://doi.org/10.1145/3306618.3314309.

[3] Himan Abdollahpouri. *Popularity bias in recommendation: a multi-stakeholder perspective.* PhD thesis, University of Colorado at Boulder, 2020.

[4] Himan Abdollahpouri and Robin Burke. Reducing popularity bias in recommendation over time. *arXiv preprint arXiv:1906.11711*, 2019.

[5] Himan Abdollahpouri and Robin Burke. Multistakeholder recommender systems. In *Recommender systems handbook*, pages 647–677. Springer, 2021.

[6] Himan Abdollahpouri and Masoud Mansoury. Multi-sided exposure bias in recommendation, 2020. URL https://arxiv.org/abs/2006.15772.

[7] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the 11th ACM conference on recommender systems*, pages 42–46, 2017.

[8] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. Managing popularity bias in recommender systems with personalized re-ranking. *arXiv preprint arXiv:1901.07555*, 2019.

[9] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The unfairness of popularity bias in recommendation. In *RecSys Workshop on Recommendation in Multistakeholder Environments (RMSE 2019)*, 2019. URL https://ceur-ws.org/Vol-2440/paper4.pdf. RecSys Workshop on Recommendation in Multistakeholder Environments (RMSE) ; Conference date: 20-09-2019.

[10] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The connection between popularity bias, calibration, and fairness in recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 726–731, 2020.

[11] Gediminas Adomavicius, Jesse C Bockstedt, Shawn P Curley, and Jingjing Zhang. Do recommender systems manipulate consumer preferences? a study of anchoring effects. *Information Systems Research*, 24(4):956–975, 2013.

[12] Charu C Aggarwal. Neighborhood-based collaborative filtering. *Recommender Systems: The Textbook*, pages 29–70, 2016.

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

[14] Mehwish Alam, Andreea Iana, Alexander Grote, Katharina Ludwig, Philipp Müller, and Heiko Paulheim. Towards analyzing the bias of news recommender systems using sentiment and stance detection. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 448–457, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391306. doi: 10.1145/3487553.3524674. URL https://doi.org/10.1145/3487553.3524674.

[15] Anne-Mette Bech Albrechtslund. Amazon, kindle, and goodreads: implications for literary consumption in the digital age. *Consumption Markets & Culture*, 23(6): 553–568, 2020.

[16] Abeer AlDayel and Walid Magdy. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597, 2021.

[17] Haifa Alharthi, Diana Inkpen, and Stan Szpakowicz. A survey of book recommender systems. *Journal of Intelligent Information Systems*, 51:139–160, 2018.

[18] Mustafa Altun. Literature and identity: Examine the role of literature in shaping individual and cultural identities. *International Journal of Social Sciences & Educational Studies*, 10(3):381–385, 2023.

[19] American Library Association. Access to library resources and services, 2021. URL https://www.ala.org/advocacy/intfreedom/access. Accessed: 26 March 2025.

[20] Chris Anderson. *The long tail: Why the future of business is selling less of more.* Hachette UK, 2006.

[21] Vito Walter Anelli, Alejandro Bellogín, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2405–2414, 2021.

[22] Khalid Anwar, Jamshed Siddiqui, and Shahab Saquib Sohail. Machine learning-based book recommender system: a survey and new perspectives. *International Journal of Intelligent Information and Database Systems*, 13(2-4):231–248, 2020.

[23] Jennifer Baker. Black women still struggle to get published, Jan 2020. URL https://zora.medium.com/the-major-built-in-bias-of-the-publishing-world-fa714f3ce9ec.

[24] Keshav Balasubramanian, Abdulla Alshabanah, Elan Markowitz, Greg Ver Steeg, and Murali Annavaram. Biased user history synthesis for personalized long-tail item recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 189–199, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705052. doi: 10.1145/3640457.3688141. URL https://doi.org/10.1145/3640457.3688141.

[25] Mariella Bastian, Natali Helberger, and Mykola Makhortykh. Safeguarding the journalistic dna: Attitudes towards the role of professional values in algorithmic news recommender designs. *Digital Journalism*, 9(6):835–863, 2021.

[26] Joeran Beel, Corinna Breitinger, Stefan Langer, Andreas Lommatzsch, and Bela Gipp. Towards reproducibility in recommender-systems research. *User modeling and user-adapted interaction*, 26:69–101, 2016.

[27] Alejandro Bellogín and Alan Said. Improving accountability in recommender systems research through reproducibility. *User Modeling and User-Adapted Interaction*, 31(5):941–977, 2021.

[28] Alejandro Bellogín, Pablo Castells, and Iván Cantador. Statistical biases in information retrieval metrics for recommender systems. *Information Retrieval Journal*, 20(6):606–634, 2017.

[29] Mary Bernstein. Identity politics. *Annu. Rev. Sociol.*, 31:47–74, 2005.

[30] Sam Biddle and Maryam Saleh. Little-known federal software can trigger revocation of citizenship, Aug 2021. URL https://theintercept.com/2021/08/25/atlas-citizenship-denaturalization-homeland-security/.

[31] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. The forgotten margins of ai ethics. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 948–958, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533157. URL https://doi.org/10.1145/3531146.3533157.

[32] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. Recommender systems survey. *Knowledge-based systems*, 46:109–132, 2013.

[33] Christina Boididou, Di Sheng, Felix J Mercer Moss, and Alessandro Piscopo. Building public service recommenders: Logbook of a journey. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 538–540, 2021.

[34] BookCrossing. Browse for members, n.d. URL https://www.bookcrossing.com/findmembers.

[35] Rodrigo Borges and Kostas Stefanidis. On mitigating popularity bias in recommendations via variational autoencoders. In *Proceedings of the 36th annual ACM symposium on applied computing*, pages 1383–1389, 2021.

[36] Fiona Bradley. Representation of libraries in artificial intelligence regulations and implications for ethics and practice. *Journal of the Australian Library and Information Association*, 71(3):189–200, 2022.

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

[38] Pablo Castells, Neil Hurley, and Saul Vargas. Novelty and diversity in recommender systems. In *Recommender systems handbook*, pages 603–646. Springer, 2021.

[39] Òscar Celma and Pedro Cano. From hits to niches? or how popular artists can bias music recommendation and discovery. In *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, NETFLIX '08, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582658. doi: 10.1145/1722149.1722154. URL https://doi.org/10.1145/1722149.1722154.

[40] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM conference on recommender systems*, pages 224–232, 2018.

[41] Elizabeth Charters. The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Education Journal*, 12(2), 2003.

[42] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ACM Transaction on Information System*, 41(3), 2023. ISSN 1046-8188.

[43] Giovanni Luca Ciampaglia, Azadeh Nematzadeh, Filippo Menczer, and Alessandro Flammini. How algorithmic popularity bias hinders or promotes quality. *Scientific reports*, 8(1):15951, 2018.

[44] Jill Cirasella. Ideology, policy, and practice: Structural barriers to collections diversity in research and college libraries. *College & Research Libraries*, 82(6): 865–878, 2021. URL https://crl.acrl.org/index.php/crl/article/view/25342/33226.

[45] Doris Hargrett Clack. Subject access to african american studies resources in online catalogs: Issues and answers. *Cataloging & Classification Quarterly*, 19(2): 49–66, 1995. doi: 10.1300/J104v19n02_04. URL https://doi.org/10.1300/J104v19n02_04.

[46] Christina Clark and Irene Picton. " it makes me feel like i'm in a different place, not stuck inside." children and young people's reading in 2020 before and during the covid-19 lockdown. national literacy trust research report. *National Literacy Trust*, 2020.

[47] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.

[48] Catherine Nicole Coleman. Managing bias when library collections become data. *International Journal of Librarianship*, 5(1):8–19, 2020. doi: 10.23974/ijol.2020.vol 5.1.162.

[49] David Collier, Fernando Daniel Hidalgo, and Andra Olivia Maciuceanu. Essentially contested concepts: Debates and applications. *Journal of political ideologies*, 11(3):211–246, 2006.

[50] Paolo Cremonesi and Dietmar Jannach. Progress in recommender systems research: Crisis? what crisis? *AI Magazine*, 42(3):43–54, 2021.

[51] Hayley Cuccinello. World's highest-paid authors 2018: Michael wolff joins list thanks to 'fire and fury'. *Forbes*, 2018.

[52] Martina Currenti. Tiktok as a marketing tool in the hands of publishers. *Logos*, 34(1):24–37, 2023.

[53] Joop Daalmeijer. Public service broadcasting in the netherlands. *Trends in Communication*, 12(1):33–45, 2004.

[54] Sarah Park Dahlen. "we need diverse books": Diversity, activism, and children's literature. *Literary cultures and twenty-first-century childhoods*, pages 83–108, 2020.

[55] Dan Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on machine learning research*, 2023.

[56] Savvina Daniil. Bias in book recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24)*, pages 1376–1381, Bari, Italy, 2024. Association for Computing Machinery. doi: 10.1145/3640457.3688025.

[57] Savvina Daniil, Mirjam Cuper, Cynthia Liem, Jacco van Ossenbruggen, and Laura Hollink. Hidden author bias in book recommendation. *arXiv preprint arXiv:2209.00371*, 2022.

[58] Savvina Daniil, Mirjam Cuper, Cynthia C. S. Liem, Jacco van Ossenbruggen, and Laura Hollink. Reproducing popularity bias in recommendation: The effect of evaluation strategies. *ACM Trans. Recomm. Syst.*, 2(1), March 2024. doi: 10.1145/3637066. URL https://doi.org/10.1145/3637066.

[59] Savvina Daniil, Manel Slokom, Mirjam Cuper, Cynthia Liem, Jacco van Ossenbruggen, and Laura Hollink. On the challenges of studying bias in recommender systems: The effect of data characteristics and algorithm configuration. *Information Retrieval Research*, 1(1):3–27, Feb. 2025. doi: 10.54195/irrj.19607. URL https://irrj.org/article/view/19607.

[60] Savvina Daniil, Søren Højlund Mollerup, and Laura Hollink. Bias in book recommendation: A case study on the danish public libraries. In *European Conference on Information Retrieval*, pages 256–270. Springer, 2026.

[61] Yashar Deldjoo, Alejandro Bellogin, and Tommaso Di Noia. Explaining recommender systems fairness and accuracy through the lens of data characteristics. *Information Processing & Management*, 58(5):102662, 2021. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.102662. URL https://www.sciencedirect.com/science/article/pii/S0306457321001503.

[62] Christian Desrosiers and George Karypis. A comprehensive survey of neighborhood-based recommendation methods. *Recommender systems handbook*, pages 107–144, 2010.

[63] Karlijn Dinnissen and Christine Bauer. Fairness in music recommender systems: A stakeholder-centered mini review. *Frontiers in big Data*, 5:913608, 2022.

[64] Karlijn Dinnissen and Christine Bauer. Amplifying artists' voices: Item provider perspectives on influence and fairness of music streaming platforms. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 238–249, 2023.

[65] Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. Comprehensive viewpoint representations for a deeper understanding of user interactions with debated topics. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, pages 135–145, 2022.

[66] Bora Edizel, Francesco Bonchi, Sara Hajian, André Panisson, and Tamir Tassa. Fairecsys: mitigating algorithmic bias in recommender systems. *International Journal of Data Science and Analytics*, 9:197–213, 2020.

[67] Michael D Ekstrand. Lenskit for python: Next-generation software for recommender systems experiments. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2999–3006, 2020.

[68] Michael D Ekstrand and Daniel Kluver. Exploring author gender in book rating and recommendation. *User Modeling & User-Adapted Interaction*, 31(3):377–420, 2021.

[69] Michael D Ekstrand, F Maxwell Harper, Martijn C Willemsen, and Joseph A Konstan. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 161–168, 2014.

[70] 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. In *Conference on fairness, accountability and transparency*, pages 172–186. PMLR, 2018.

[71] Mehdi Elahi, Danial Khosh Kholgh, Mohammad Sina Kiarostami, Sorush Saghari, Shiva Parsa Rad, and Marko Tkalčič. Investigating the impact of recommender systems on user-based and item-based popularity bias. *Information Processing & Management*, 58(5):102655, 2021.

[72] Stavroula Eleftherakis, Georgia Koutrika, and Sihem Amer-Yahia. Optimizing neighborhoods for fair top-n recommendation. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '24, page 57–66, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704338. doi: 10.1145/3627043.3659539. URL https://doi.org/10.1145/3627043.3659539.

[73] Sina Fazelpour and Maria De-Arteaga. Diversity in sociotechnical machine learning systems. *Big Data & Society*, 9(1):20539517221082027, 2022.

[74] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM conference on recommender systems*, pages 101–109, 2019.

[75] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems (TOIS)*, 39(2):1–49, 2021.

[76] Andres Ferraro, Xavier Serra, and Christine Bauer. Break the loop: Gender imbalance in music recommenders. In *Proceedings of the 2021 conference on human information interaction and retrieval*, pages 249–254, 2021.

[77] Bruce Ferwerda, Eveline Ingesson, Michaela Berndl, and Markus Schedl. I don't care how popular you are! investigating popularity bias in music recommendations from a user's perspective. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, CHIIR '23, page 357–361, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700354. doi: 10.1145/3576840.3578287. URL https://doi.org/10.1145/3576840.3578287.

[78] Nettie Finn. Pseudonymous disguises: Are pen names an escape from the gender bias in publishing? Honor Scholar Thesis, DePauw University, 2016. URL https://scholarship.depauw.edu/studentresearch/44. Honor Scholar Theses, No. 44.

[79] Luciano Floridi. Establishing the rules for building trustworthy ai. *Nature Machine Intelligence*, 1(6):261–262, 2019.

[80] J. Folta. Fable's ai-generated end-of-year reading summaries veered into bigotry, January 8 2025. URL https://lithub.com/fables-ai-generated-end-of-year-reading-summaries-veered-into-bigotry/. Accessed: 26 March 2025.

[81] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3):330–347, July 1996. ISSN 1046-8188. doi: 10.1145/230538.230561. URL https://doi.org/10.1145/230538.230561.

[82] Giovanni Gabbolini, Edoardo D'Amico, Cesare Bernardis, and Paolo Cremonesi. On the instability of embeddings for recommender systems: the case of matrix factorization. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC '21, page 1363–1370, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450381048. doi: 10.1145/3412841.3442011. URL https://doi.org/10.1145/3412841.3442011.

[83] Walter Bryce Gallie. Essentially contested concepts. In *Proceedings of the Aristotelian society*, volume 56, pages 167–198. JSTOR, 1955.

[84] Cynthia R. Greenlee. Fable ai's new model is more inclusive, but still struggles with racism, 2025. URL https://www.nytimes.com/2025/01/03/us/fable-ai-books-racism.html. Accessed: 26 March 2025.

[85] Andreas Grün and Xenija Neufeld. Challenges experienced in public service media recommendation systems. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 541–544, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384582. doi: 10.1145/3460231.3474618. URL https://doi.org/10.1145/3460231.3474618.

[86] Asela Gunawardana, Guy Shani, and Sivan Yogev. Evaluating recommender systems. In *Recommender systems handbook*, pages 547–601. Springer, 2022.

[87] Odd Erik Gundersen and Sigbjørn Kjensmo. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 1644–1651, 2018. doi: 10.1609/aaai.v32i1.11503. URL https://ojs.aaai.org/index.php/AAAI/article/view/11503.

[88] Odd Erik Gundersen, Yolanda Gil, and David W Aha. On reproducible ai: Towards reproducible research, open science, and digital scholarship in ai publications. *AI magazine*, 39(3):56–68, 2018.

[89] Wenshuo Guo, Karl Krauth, Michael Jordan, and Nikhil Garg. The stereotyping problem in collaboratively filtered recommender systems. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–10, 2021.

[90] Elizabeth Gómez, Ludovico Boratto, and Maria Salamó. Provider fairness across continents in collaborative recommender systems. *Information Processing & Management*, 59(1):102719, 2022. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.102719. URL https://www.sciencedirect.com/science/article/pii/S030645732100203X.

[91] Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S Greene, et al. Transparency and reproducibility in artificial intelligence. *Nature*, 586(7829):E14–E16, 2020.

[92] Jaron Harambam, Dimitrios Bountouridis, Mykola Makhortykh, and Joris Van Hoboken. Designing for the better by taking users into account: A qualitative evaluation of user control mechanisms in (news) recommender systems. In *Proceedings of the 13th ACM conference on recommender systems*, pages 69–77, 2019.

[93] Dessy Harisanty, Nove E Variant Anna, Tesa Eranti Putri, Aji Akbar Firdaus, and Nurul Aida Noor Azizi. Leaders, practitioners and scientists' awareness of artificial intelligence in libraries: a pilot study. *Library Hi Tech*, 42(3):809–825, 2024.

[94] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[95] Lucien Heitz, Juliane A Lischka, Alena Birrer, Bibek Paudel, Suzanne Tolmeijer, Laura Laugwitz, and Abraham Bernstein. Benefits of diverse news recommendations for democracy: A user study. *Digital Journalism*, 10(10):1710–1730, 2022.

[96] Natali Helberger. On the democratic role of news recommenders. *Digital Journalism*, 7(8):993–1012, 2019. doi: 10.1080/21670811.2019.1623700.

[97] Natali Helberger, Kari Karppinen, and Lucia D'acunto. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, 21(2):191–207, 2018.

[98] Jon Henley and Robert Booth. Welfare surveillance system violates human rights, dutch court rules, Feb 2020. URL https://www.theguardian.com/technology/2020/feb/05/welfare-surveillance-system-violates-human-rights-dutch-court-rules.

[99] Jockum Hildén. The public service approach to recommender systems: Filtering to cultivate. *Television & New Media*, 23(7):777–796, 2022.

[100] James Oluwaseyi Hodonu-Wusu. The rise of artificial intelligence in libraries: the ethical and equitable methodologies, and prospects for empowering library users. *AI and Ethics*, 5(2):755–765, 2025.

[101] Lei Hou, Xue Pan, and Kecheng Liu. Balancing the popularity bias of object similarities for personalised recommendation. *The European Physical Journal B*, 91:1–7, 2018.

[102] Han-Yin Huang and Cynthia CS Liem. Social inclusion in curated contexts: Insights from museum practices. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 300–309, 2022.

[103] Nicolas Hug. Surprise: A python library for recommender systems. *Journal of Open Source Software*, 5(52):2174, 2020. doi: 10.21105/joss.02174. URL https://doi.org/10.21105/joss.02174.

[104] Matthew Hutson. Artificial intelligence faces reproducibility crisis. *Science*, 359(6377):725–726, 2018. doi: 10.1126/science.359.6377.725. URL https://www.science.org/doi/abs/10.1126/science.359.6377.725.

[105] International Federation of Library Associations and Institutions (IFLA) and UNESCO. IFLA/UNESCO Multicultural Library Manifesto. https://repository.ifla.org/handle/20.500.14598/731, 2009. Approved by IFLA Governing Board and adopted by UNESCO; outlines principles for multicultural library services, including giving access to a broad range of materials reflecting all communities and needs.

[106] L. Jahnke, K. Tanaka, and C. Palazzolo. Ideology, policy, and practice: Structural barriers to collections diversity in research and college libraries. *College & Research Libraries*, 83(2):166–184, 2022. doi: 10.5860/crl.83.2.166. URL https://doi.org/10.5860/crl.83.2.166.

[107] Andrea Jamison. *Decentering whiteness in libraries: a framework for inclusive collection management practices.* Rowman & Littlefield, 2023.

[108] Dietmar Jannach and Michael Jugovac. Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems (TMIS)*, 10 (4):1–23, 2019.

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

[110] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[111] Mathias Jesse, Christine Bauer, and Dietmar Jannach. Intra-list similarity and human diversity perceptions of recommendations: the details matter. *User Modeling and User-Adapted Interaction*, 33(4):769–802, 2023.

[112] Sanjay Kumar Jha. Application of artificial intelligence in libraries and information centers services: prospects and challenges. *Library hi tech news*, 40(7):1–5, 2023.

[113] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Correcting popularity bias by enhancing recommendation neutrality. *RecSys Posters*, 805, 2014.

[114] Kari Karppinen. *Rethinking media pluralism.* Fordham Univ Press, 2013.

[115] Kari Karppinen. The limits of empirical indicators: Media pluralism as an essentially contested concept. In *Media pluralism and diversity: Concepts, risks and global trends*, pages 287–296. Springer, 2015.

[116] Shah Khusro, Zafar Ali, and Irfan Ullah. Recommender systems: issues, challenges, and research opportunities. In *Information science and applications*, pages 1179–1189. Springer, 2016.

[117] Ömer Kırnap, Fernando Diaz, Asia Biega, Michael Ekstrand, Ben Carterette, and Emine Yilmaz. Estimation of fair ranking metrics with incomplete judgments. In *Proceedings of the Web Conference 2021*, WWW '21, page 1065–1075, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450080. URL https://doi.org/10.1145/3442381.3450080.

[118] Anastasiia Klimashevskaia, Dietmar Jannach, Mehdi Elahi, and Christoph Trattner. A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction*, 34(5):1777–1834, 2024.

[119] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[120] Yehuda Koren, Steffen Rendle, and Robert Bell. Advances in collaborative filtering. *Recommender systems handbook*, pages 91–142, 2022.

[121] Dominik Kowald and Emanuel Lacic. Popularity bias in collaborative filtering-based multimedia recommender systems. In *International Workshop on Algorithmic Bias in Search and Recommendation*, pages 1–11. Springer, 2022.

[122] Dominik Kowald, Markus Schedl, and Elisabeth Lex. The unfairness of popularity bias in music recommendation: A reproducibility study. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, pages 35–42, Cham, 2020. Springer International Publishing. ISBN 978-3-030-45442-5.

[123] Johannes Kruse, Kasper Lindskow, Michael Riis Andersen, and Jes Frellsen. Creating the next generation of news experience on ekstrabladet. dk with recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1067–1070, 2023.

[124] Matevž Kunaver and Tomaž Požrl. Diversity in recommender systems–a survey. *Knowledge-based systems*, 123:154–162, 2017.

[125] Renate Lachmann. Cultural memory and the role of literature. *European Review*, 12(2):165–178, 2004.

[126] EE Lawrence. The trouble with diverse books, part i: on the limits of conceptual analysis for political negotiation in library & information science. *Journal of Documentation*, 76(6):1473–1491, 2020.

[127] Leeandlowbooks. Where is the diversity in publishing? the 2019 diversity baseline survey results, Mar 2022. URL https://blog.leeandlow.com/2020/01/28/2019diversitybaselinesurvey/.

[128] Oleg Lesota, Alessandro Melchiorre, Navid Rekabsaz, Stefan Brandl, Dominik Kowald, Elisabeth Lex, and Markus Schedl. Analyzing item popularity bias of music recommender systems: Are different genders equally affected? In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 601–606, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384582. doi: 10.1145/3460231.3478843. URL https://doi.org/10.1145/3460231.3478843.

[129] Tin Cheuk Leung and Koleman S Strumpf. Discrimination in the publishing industry? *Available at SSRN 3660289*, 2022.

[130] Cheng-Te Li, Cheng Hsu, and Yang Zhang. Fairsr: Fairness-aware sequential recommendation through multi-task learning with preference graph embeddings. *ACM Trans. Intell. Syst. Technol.*, 13(1), February 2022. ISSN 2157-6904. doi: 10.1145/3495163. URL https://doi.org/10.1145/3495163.

[131] Chen Lin, Dugang Liu, Hanghang Tong, and Yanghua Xiao. Spiral of silence and its application in recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2934–2947, 2022. doi: 10.1109/TKDE.2020.3013973.

[132] Greg Linden, Brent Smith, and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80, 2003.

[133] Shenghao Liu, Yu Zhang, Lingzhi Yi, Xianjun Deng, Laurence T. Yang, and Bang Wang. Dual-side adversarial learning based fair recommendation for sensitive attribute filtering. *ACM Trans. Knowl. Discov. Data*, 18(7), June 2024. ISSN 1556-4681. doi: 10.1145/3648683. URL https://doi.org/10.1145/3648683.

[134] Yang Liu, Alan Medlar, and Dorota Glowacka. What we evaluate when we evaluate recommender systems: Understanding recommender systems' performance using item response theory. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 658–670, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702419. doi: 10.1145/3604915.3608809. URL https://doi.org/10.1145/3604915.3608809.

[135] Jeffrey W Lockhart, Molly M King, and Christin Munsch. Name-based demographic inference and the unequal distribution of misrecognition. *Nature Human Behaviour*, pages 1–12, 2023.

[136] Felicia Loecherbach, Judith Moeller, Damian Trilling, and Wouter van Atteveldt. The unified framework of media diversity: A systematic literature review. *Digital Journalism*, 8(5):605–642, 2020.

[137] M Elena Lopez, Bharat Mehra, and Margaret Capse. An exploratory social justice framework to develop public library services with underserved families. *Public Library Quarterly*, 42(6):576–601, 2023.

[138] Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang, and Tao Zhou. Recommender systems. *Physics reports*, 519(1):1–49, 2012.

[139] Dr. Prakashbhai N. Makwana. The dewey decimal system and traditional libraries. *International Journal of Science and Research (IJSR)*, 13(6), June 2024. doi: 10.21275/SR24530155734. URL https://dx.doi.org/10.21275/SR24530155734.

[140] Sara Mannheimer, Natalie Bond, Scott W. H. Young, Hannah Scates Kettler, Addison Marcus, Sally K. Slipher, Jason A. Clark, Yasmeen Shorish, Doralyn Rossmann, and Bonnie Sheehey. Responsible ai practice in libraries and archives: A review of the literature. *Information Technology and Libraries*, 43(3), Sep. 2024. doi: 10.5860/ital.v43i3.17245. URL https://ital.corejournals.org/index.php/ital/article/view/17245.

[141] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2145–2148, 2020.

[142] Emily Martin. The 15 top authors, based on goodreads stats, Nov 2021. URL https://bookriot.com/top-goodreads-authors/.

[143] Paolo Massa and Paolo Avesani. Trust-aware recommender systems. In *Proceedings of the 2007 ACM conference on Recommender systems*, pages 17–24, 2007.

[144] Kay Mathiesen. Informational justice: A conceptual framework for social justice in library and information services. *Library trends*, 64(2):198–225, 2015.

[145] Nora McDonald and Shimei Pan. Intersectional ai: A study of how information science students think about ethics and their impact. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), oct 2020. doi: 10.1145/3415218. URL https://doi.org/10.1145/3415218.

[146] Sean M McNee, John Riedl, and Joseph A Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1097–1101, 2006.

[147] Phayung Meesad and Anirach Mingkhwan. Libraries in transformation. In *Libraries in Transformation*, pages 391–428. Springer, 2024.

[148] Lien Michiels, Jorre Vannieuwenhuyze, Jens Leysen, Robin Verachtert, Annelien Smets, and Bart Goethals. How should we measure filter bubbles? a regression model and evidence for online news. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 640–651, 2023.

[149] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. Ethical aspects of multi-stakeholder recommendation systems. *The information society*, 37(1):35–45, 2021.

[150] Sudhakar Mishra. Ethical implications of artificial intelligence and machine learning in libraries and information centers: A frameworks, challenges, and best practices. *Library Philosophy and Practice (e-journal)*, 7753, 2023.

[151] Joanna Misztal-Radecka and Bipin Indurkhya. Bias-aware hierarchical clustering for detecting the discriminated groups of users in recommendation systems. *Information Processing & Management*, 58(3):102519, 2021. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.102519. URL https://www.sciencedirect.com/science/article/pii/S0306457321000285.

[152] Joanna Misztal-Radecka and Bipin Indurkhya. A bias detection tree approach for detecting disparities in a recommendation model's errors. *User Modeling and User-Adapted Interaction*, 33(1):43–79, 2023.

[153] Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 117–123, 2020.

[154] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. In *Digital media, political polarization and challenges to democracy*, pages 45–63. Routledge, 2020.

[155] Lynge Asbjørn Møller. Recommended for you: how newspapers normalise algorithmic news recommendation to fit their gatekeeping role. *Journalism Studies*, 23(7):800–817, 2022.

[156] Lynge Asbjørn Møller. Designing algorithmic editors: How newspapers embed and encode journalistic values into news recommender systems. *Digital Journalism*, pages 1–19, 2023.

[157] Jennifer Moody and David H. Glass. A novel classification framework for evaluating individual and aggregate diversity in top-n recommendations. *ACM Trans. Intell. Syst. Technol.*, 7(3), February 2016. ISSN 2157-6904. doi: 10.1145/2700491. URL https://doi.org/10.1145/2700491.

[158] Anna Muenchrath. *Selling Books with Algorithms.* Cambridge University Press, 2024.

[159] Mohammadmehdi Naghiaei, Hossein A Rahmani, and Mahdi Dehghan. The unfairness of popularity bias in book recommendation. In *International Workshop on Algorithmic Bias in Search and Recommendation*, pages 69–81. Springer, 2022.

[160] Yesha Naik and Barry Trott. Finding good reads on goodreads: Readers take ra into their own hands. *Reference & User Services Quarterly*, 51(4):319–323, 2012. ISSN 10949054, 21635242. URL http://www.jstor.org/stable/refuseserq.51.4.319.

[161] National Endowment for the Arts. Arts participation: Patterns in 2022 — highlights from the survey of public participation in the arts. Technical report, National Endowment for the Arts, Washington, DC, October 2023. URL https://www.arts.gov/sites/default/files/2022-SPPA-final.pdf.

[162] Library & Information Science Education Network. Exploring ai-based recommendation systems in libraries, 2024. URL https://www.lisedunetwork.com/exploring-ai-based-recommendation-systems-in-libraries/. Accessed: 25 March, 2025.

[163] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197, 2019.

[164] Christian S Nissen et al. Public service media in the information society, 2006.

[165] Hiroki Okamura, Keisuke Maeda, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Flexibly manipulating popularity bias for tackling trade-offs in recommendation. *Information Processing & Management*, 61(2):103606, 2024. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2023.103606. URL https://www.sciencedirect.com/science/article/pii/S0306457323003436.

[166] Hope Olson. *The Power to Name: Locating the Limits of Subject Representation in Libraries*. Kluwer Academic, Dordrecht, 2002.

[167] O Onunka, T Onunka, AA Fawole, IJ Adeleke, and C Daraojimba. Library and information services in the digital age: Opportunities and challenges. *Acta Informatica Malaysia*, 7(1):113–121, 2023.

[168] Yesid Ospitia-Medina, Sandra Baldassarri, Cecilia Sanz, and José Ramón Beltrán. Music recommender systems: A review centered on biases. *Advances in Speech and Music Technology: Computational Aspects and Applications*, pages 71–90, 2022.

[169] OverDrive, Inc. Libraries break digital lending records in 2024 with over 739 million checkouts, January 2025. URL https://company.overdrive.com/2025/01/27/libraries-break-digital-lending-records-in-2024-with-over-739-million-checkouts/. Accessed: 01 April 2025.

[170] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021. doi: 10.1136/bmj.n71. URL https://www.bmj.com/content/372/bmj.n71.

[171] Ioanna Papageorgiou and Carlos Mougan. Necessity of processing sensitive data for bias detection and monitoring: A techno-legal exploration. In *NeurIPS 2023 Workshop on Regulatable ML*, 2023.

[172] Anish Patankar, Joy Bose, and Harshit Khanna. A bias aware news recommendation system. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 232–238. IEEE, 2019.

[173] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research: a report from the neurips 2019 reproducibility program. *Journal of Machine Learning Research*, 22, 2021.

[174] Christine Pinney, Amifa Raj, Alex Hanna, and Michael D Ekstrand. Much ado about gender: Current practices and future recommendations for appropriate gender-aware information access. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pages 269–279, 2023.

[175] Jessica Pressman. *Bookishness: Loving books in a digital age.* Columbia University Press, 2020.

[176] Hossein A. Rahmani, Mohammadmehdi Naghiaei, Mahdi Dehghan, and Mohammad Aliannejadi. Experiments on generalizability of user-oriented fairness in recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2755–2764, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531718. URL https://doi.org/10.1145/3477495.3531718.

[177] Hossein A. Rahmani, Mohammadmehdi Naghiaei, and Yashar Deldjoo. A personalized framework for consumer and producer group fairness optimization in recommender systems. *ACM Trans. Recomm. Syst.*, 2(3), June 2024. doi: 10.1145/3651167. URL https://doi.org/10.1145/3651167.

[178] Amifa Raj and Michael D. Ekstrand. Measuring fairness in ranked results: An analytical and empirical comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 726–736, New York, NY, USA, 2022. Association for Computing Machinery.

ISBN 9781450387323. doi: 10.1145/3477495.3532018. URL https://doi.org/10.1145/3477495.3532018.

[179] Amifa Raj and Michael D Ekstrand. Towards optimizing ranking in grid-layout for provider-side fairness. In *European Conference on Information Retrieval*, pages 90–105. Springer, 2024.

[180] S. R. Ranganathan. *The Five Laws of Library Science*. Madras Library Association, Madras, India, 1931.

[181] Jennifer Rauch. Slow media as alternative media: Cultural resistance through print and analogue revivals. In *The Routledge companion to alternative and community media*, pages 571–581. Routledge, 2015.

[182] Viviane Reding. The role of libraries in the information society. In *CENL Conference Luxembourg*, volume 29, pages 67–70, 2005.

[183] Alex Reece. Frisco library sees record-breaking patronage during first fiscal year open, data shows, November 2024. URL https://communityimpact.com/dallas-fort-worth/frisco/government/2024/11/13/frisco-library-sees-record-breaking-patronage-during-first-fiscal-year-open-data-shows/. Accessed: 01 April 2025.

[184] Matthew Reidsma. *Masked by Trust: Bias in Library Discovery*. Library Juice Press, 2019. URL https://scholarworks.gvsu.edu/library_books/30/.

[185] Kara Reuter. Making magic happen: Building and launching a reader's advisory kiosk. *Information Technology and Libraries*, 43(2), 2024.

[186] Wondo Rhee, Sung Min Cho, and Bongwon Suh. Countering popularity bias by regularizing score differences. In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, page 145–155, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392785. doi: 10.1145/3523227.3546757. URL https://doi.org/10.1145/3523227.3546757.

[187] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: Techniques, applications, and challenges. *Recommender systems handbook*, pages 1–35, 2021.

[188] Alisa Rieger, Tim Draws, Mariët Theune, and Nava Tintarev. This item might reinforce your opinion: Obfuscation and labeling of search results to mitigate confirmation bias. In *Proceedings of the 32nd ACM conference on hypertext and social media*, pages 189–199, 2021.

[189] Royal Danish Library. Strategy — tasks and goals. https://www.kb.dk/en/about-us/tasks-and-goals/strategy, 2024. Accessed 12 January 2026.

[190] Elisavet Rozaki. Reading between the likes: The influence of booktok on reading culture. Master's thesis, Utrecht University, 2023. URL https://studentthesis.uu.nl/handle/20.500.12932/43888.

[191] Richard E Rubin and Rachel G Rubin. *Foundations of library and information science.* American Library Association, 2020.

[192] Jonathan H Rystrøm. An offer you cannot refuse? trends in the coercive impact of amazon book recommendations. In *International Workshop on Algorithmic Bias in Search and Recommendation*, pages 1–15. Springer, 2024.

[193] Bess Sadler and Chris Bourg. Feminism and the future of library discovery. *Code4Lib Journal*, 2015.

[194] Hamid Reza Saeidnia. Ethical artificial intelligence (ai): confronting bias and discrimination in the library and information industry. *Library Hi Tech News*, 2023.

[195] Anamik Saha and Sandra Van Lente. Rethinking 'diversity' in publishing. Project report / eprint, Goldsmiths Press, Goldsmiths, University of London, 2020. URL `https://research.gold.ac.uk/id/eprint/28692/`.

[196] Alan Said and Alejandro Bellogín. Comparative recommender system evaluation: benchmarking recommendation frameworks. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 129–136, 2014.

[197] Aghiles Salah, Quoc-Tuan Truong, and Hady W Lauw. Cornac: A comparative framework for multimodal recommender systems. *The Journal of Machine Learning Research*, 21(1):3803–3807, 2020.

[198] Joyce G Saricks. *Readers' advisory service in the public library.* American Library Association, 2005.

[199] Shrikant Saxena and Shweta Jain. Exploring and mitigating gender bias in book recommender systems with explicit feedback. *Journal of Intelligent Information Systems*, 62(5):1325–1346, 2024.

[200] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web*, pages 291–324. Springer, 2007.

[201] Markus Schedl. The lfm-1b dataset for music retrieval and recommendation. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, page 103–110, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450343596. doi: 10.1145/2911996.2912004. URL `https://doi.org/10.1145/2911996.2912004`.

[202] Zaina Shaik, Filip Ilievski, and Fred Morstatter. Analyzing race and country of citizenship bias in wikidata. *arXiv preprint arXiv:2108.05412*, 2021.

[203] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

[204] Faisal Shehzad and Dietmar Jannach. Everyone'sa winner! on hyperparameter tuning of recommendation models. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 652–657, 2023.

[205] Lijuan Shen and Liping Jiang. Eliminating bias: enhancing children's book recommendation using a hybrid model of graph convolutional networks and neural matrix factorization. *PeerJ Computer Science*, 10:e1858, 2024.

[206] Similarweb. Top libraries and museums websites ranking in march 2022, 2022. URL https://www.similarweb.com/top-websites/category/science-and-education/libraries-and-museums/.

[207] R. Sivasankari, S. Suriya, S. Sindhu, J. Shyamala Devi, and J. Dhilipan. Ai-powered recommendation systems and resource discovery for library management. In *Advances in Library and Information Science Applications of Artificial Intelligence in Libraries*, pages 223–244. IGI Global, 2024. URL https://www.igi-global.com/chapter/ai-powered-recommendation-systems-and-resource-discovery-for-library-management/346535. Accessed: 22 March 2025.

[208] Manel Slokom. Comparing recommender systems using synthetic data. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 548–552, 2018.

[209] Manel Slokom and Martha Larson. Doing data right: How lessons learned working with conventional data should inform the future of synthetic data for recommender systems. *arXiv preprint arXiv:2110.03275*, 2021.

[210] Annelien Smets, Jonathan Hendrickx, and Pieter Ballon. We're in this together: a multi-stakeholder approach for news recommenders. *Digital Journalism*, 10(10): 1813–1831, 2022.

[211] Brent Smith and Greg Linden. Two decades of recommender systems at amazon. com. *Ieee internet computing*, 21(3):12–18, 2017.

[212] Nasim Sonboli, Masoud Mansoury, Ziyue Guo, Shreyas Kadekodi, Weiwen Liu, Zijun Liu, Andrew Schwartz, and Robin Burke. Librec-auto: A tool for recommender systems experimentation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4584–4593, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469. doi: 10.1145/3459637.3482006. URL https://doi.org/10.1145/3459637.3482006.

[213] Jannick Kirk Sørensen. Public service media, diversity and algorithmic recommendation: Tensions between editorial principles and algorithms in european psm organizations. In *CEUR workshop proceedings*, volume 2554, pages 6–11. CEUR Workshop Proceedings, 2019.

[214] Harald Steck. Item popularity and recommendation accuracy. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 125–132, 2011.

[215] Ann Steiner. The global book: micropublishing, conglomerate production, and digital market structures. *Publishing research quarterly*, 34(1):118–132, 2018.

[216] Catherine Stinson. Algorithms are not neutral. *AI and Ethics*, pages 1–8, 2022.

[217] Ted Striphas. *The late age of print: Everyday book culture from consumerism to control.* Columbia University Press, 2011.

[218] A Subaveerapandiyan. Application of artificial intelligence (ai) in libraries and its impact on library operations review. *Library Philosophy & Practice*, 2023.

[219] Gábor Takács, István Pilászy, and Domonkos Tikk. Applications of the conjugate gradient method for implicit feedback collaborative filtering. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 297–300, 2011.

[220] Yan-Martin Tamm, Rinchin Damdinov, and Alexey Vasilev. Quality metrics in recommender systems: Do we calculate metrics consistently? In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 708–713, 2021.

[221] Shuai Tang and Xiaofeng Zhang. Cadpp: An effective approach to recommend attentive and diverse long-tail items. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT '21, page 218–225, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391153. doi: 10.1145/3486622.3493961. URL https://doi.org/10.1145/3486622.3493961.

[222] Endia Paige Tarica LaBossiere and Beau Steenken. Keeping up with: Bias. American Library Association, 2024. URL https://www.ala.org/acrl/publications/keeping_up_with/bias. Accessed: 2024-04-04.

[223] Riku Togashi, Mayu Otani, and Shin'ichi Satoh. Alleviating cold-start problems in recommendation through pseudo-labelling over knowledge graph. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 931–939, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450382977. doi: 10.1145/3437963.3441773. URL https://doi.org/10.1145/3437963.3441773.

[224] Karen HL Tso and Lars Schmidt-Thieme. Empirical analysis of attribute-aware recommender system algorithms using synthetic data. *Journal of Computers*, 1(4): 18–29, 2006.

[225] University of Colorado Boulder. Anti-racist collections review workbook, 2025. URL https://libguides.colorado.edu/anti-racist-collections-review-acquisitions/workbook2?utm_source=chatgpt.com. Accessed: 23 March, 2025.

[226] Jan Van Cuilenburg. On competition, access and diversity in media, old and new: Some remarks for communications policy in the information age. *New media & society*, 1(2):183–207, 1999.

[227] Alje van Dam. Diversity and its decomposition into variety, balance and disparity. *Royal Society open science*, 6(7):190452, 2019.

[228] Jan Willem Van Wessel. Ai in libraries: Seven principles. Zenodo, May 2020. URL https://doi.org/10.5281/zenodo.3865344.

[229] Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. Recommenders with a mission: Assessing diversity in news recommendations. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR '21, pages 173–183, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380553. doi: 10.1145/3406522.3446019. URL https://doi.org/10.1145/3406522.3446019.

[230] Sanne Vrijenhoek, Gabriel Benedict, Mateo Gutierez Granada, Daan Odijk, and Maarten de Rijke. Radio – rank-aware divergence metrics to measure normative diversity in news recommendations. In *accepted for RecSys 2022*, 2022.

[231] Sanne Vrijenhoek, Savvina Daniil, Jorden Sandel, and Laura Hollink. Diversity of what? on the different conceptualizations of diversity in recommender systems. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 573–584, 2024.

[232] Simon Wakeling, Paul Clough, Barbara Sen, and Lynn Silipigni Connaway. "readers who borrowed this also borrowed...": recommender systems in uk libraries. *Library Hi Tech*, 30(1):134–150, 2012.

[233] Melanie Walsh and Maria Antoniak. The goodreads "classics": A computational study of readers, amazon, and crowdsourced amateur criticism. *Journal of Cultural Analytics*, 4:243–287, 2021.

[234] Mengting Wan and Julian J. McAuley. Item recommendation on monotonic behavior chains. In Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan, editors, *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 86–94. ACM, 2018. doi: 10.1145/3240323.3240369. URL https://doi.org/10.1145/3240323.3240369.

[235] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. Fine-grained spoiler detection from large-scale review corpora. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2605–2610. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1248. URL https://doi.org/10.18653/v1/p19-1248.

[236] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3):1–43, 2023.

[237] Zhidan Wang, Lixin Zou, Chenliang Li, Shuaiqiang Wang, Xu Chen, Dawei Yin, and Weidong Liu. Toward bias-agnostic recommender systems: A universal generative framework. *ACM Trans. Inf. Syst.*, 42(6), June 2024. ISSN 1046-8188. doi: 10.1145/3655617. URL https://doi.org/10.1145/3655617.

[238] Katie Warburton, Charles Kemp, Yang Xu, and Lea Frermann. Quantifying bias in library classification systems. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2023. URL https://escholarship.org/uc/item/94b7d9zr.

[239] Chunyu Wei, Jian Liang, Di Liu, Zehui Dai, Mang Li, and Fei Wang. Meta graph learning for long-tail recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 2512–2522, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599428. URL https://doi.org/10.1145/3580305.3599428.

[240] Tianxin Wei and Jingrui He. Comprehensive fair meta-learned recommender system. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 1989–1999, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539269. URL https://doi.org/10.1145/3534678.3539269.

[241] Brenda K. Wiederhold. Booktok made me do it: The evolution of reading. *Cyberpsychology, Behavior, and Social Networking*, 25(3):157–158, 2022.

[242] Howard W Winger. Historical perspectives on the role of the book in society. *The Library Quarterly*, 25(4):293–305, 1955.

[243] WIPO. The global publishing industry in 2018. Geneva: World Intellectual Property Organization, 2018. URL https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1064_2019.pdf.

[244] Lanling Xu, Zihan Lin, Jinpeng Wang, Sheng Chen, Wayne Xin Zhao, and Ji-Rong Wen. Promoting two-sided fairness with adaptive weights for providers and customers in recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 918–923, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705052. doi: 10.1145/3640457.3688169. URL https://doi.org/10.1145/3640457.3688169.

[245] Hong-Jian Xue, Xinyu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. Deep matrix factorization models for recommender systems. In *IJCAI*, volume 17, pages 3203–3209. Melbourne, Australia, 2017.

[246] Emre Yalcin. Blockbuster: A new perspective on popularity-bias in recommender systems. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pages 107–112. IEEE, 2021.

[247] Emre Yalcin. Exploring potential biases towards blockbuster items in ranking-based recommendations. *Data Mining and Knowledge Discovery*, 36(6):2033–2073, 2022.

[248] Emre Yalcin and Alper Bilge. Investigating and counteracting popularity bias in group recommendations. *Information Processing & Management*, 58(5):102608, 2021.

[249] Hao Yang, Xian Wu, Zhaopeng Qiu, Yefeng Zheng, and Xu Chen. Distributional fairness-aware recommendation. *ACM Trans. Inf. Syst.*, 42(5), April 2024. ISSN 1046-8188. doi: 10.1145/3652854. URL https://doi.org/10.1145/3652854.

[250] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, 2020.

[251] Wenhao Yang, Yingchun Jian, Yibo Wang, Shiyin Lu, Lei Shen, Bing Wang, Haihong Tang, and Lijun Zhang. Not all embeddings are created equal: Towards robust cross-domain recommendation via contrastive learning. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 3195–3206, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645357. URL https://doi.org/10.1145/3589334.3645357.

[252] Nasim Yazdani. A pedagogy of care in academic libraries: A framework to increase underrepresented communities' sense of belonging and engagement. *Journal of the Australian Library and Information Association*, 73(4):505–528, 2024.

[253] Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. Challenging the long tail recommendation. *Proceedings of the 38th International Conference on Very Large Data Bases (VLDB Endowment)*, 5(9):896–907, 2012.

[254] Zygmunt Zajac. Goodbooks-10k: a new dataset for book recommendations. *FastML*, 2017.

[255] Eva Zangerle and Christine Bauer. Evaluating recommender systems: survey and framework. *ACM Computing Surveys*, 55(8):1–38, 2022.

[256] Kuanyi Zhang, Min Xie, Yi Zhang, and Haixian Zhang. Utilizing implicit feedback for user mainstreaminess evaluation and bias detection in recommender systems. In *International Workshop on Algorithmic Bias in Search and Recommendation*, pages 42–58. Springer, 2023.

[257] Mi Zhang and Neil Hurley. Avoiding monotony: Improving the diversity of recommendation lists. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, page 123–130, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580937. doi: 10.1145/1454008.1454030. URL https://doi.org/10.1145/1454008.1454030.

[258] Yin Zhang, Derek Zhiyuan Cheng, Tiansheng Yao, Xinyang Yi, Lichan Hong, and Ed H. Chi. A model of two tales: Dual transfer learning framework for

improved long-tail item recommendation. In *Proceedings of the Web Conference 2021*, WWW '21, page 2220–2231, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450086. URL https://doi.org/10.1145/3442381.3450086.

[259] Yin Zhang, Ruoxi Wang, Derek Zhiyuan Cheng, Tiansheng Yao, Xinyang Yi, Lichan Hong, James Caverlee, and Ed H. Chi. Empowering long-tail item recommendation through cross decoupling network (cdn). In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 5608–5617, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599814. URL https://doi.org/10.1145/3580305.3599814.

[260] Zheng Zhang, Qi Liu, Zirui Hu, Yi Zhan, Zhenya Huang, Weibo Gao, and Qingyang Mao. Enhancing fairness in meta-learned user modeling via adaptive sampling. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 3241–3252, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645369. URL https://doi.org/10.1145/3589334.3645369.

[261] Zihao Zhao, Jiawei Chen, Sheng Zhou, Xiangnan He, Xuezhi Cao, Fuzheng Zhang, and Wei Wu. Popularity bias is not always evil: Disentangling benign and harmful bias for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[262] Yunhong Zhou, Dennis Wilkinson, Robert Schreiber, and Rong Pan. Large-scale parallel collaborative filtering for the netflix prize. In Rudolf Fleischer and Jinhui Xu, editors, *Algorithmic Aspects in Information and Management*, pages 337–348, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-68880-8.

[263] Jieming Zhu, Quanyu Dai, Liangcai Su, Rong Ma, Jinyang Liu, Guohao Cai, Xi Xiao, and Rui Zhang. Bars: Towards open benchmarking for recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2912–2923, 2022.

[264] Ziwei Zhu, Yun He, Xing Zhao, and James Caverlee. Popularity bias in dynamic recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2439–2449, 2021.

[265] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, page 22–32, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595930469. doi: 10.1145/1060745.1060754. URL https://doi.org/10.1145/1060745.1060754.

[266] Zippia. Author demographics and statistics [2022]: Number of authors in the us, Apr 2022. URL https://www.zippia.com/author-jobs/demographics/.

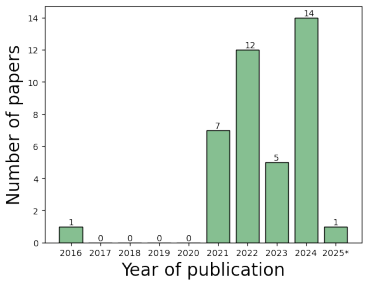
[267] Frederik Zuiderveen Borgesius, Damian Trilling, Judith Möller, Balázs Bodó, Claes H De Vreese, and Natali Helberger. Should we worry about filter bubbles? *Internet Policy Review. Journal on Internet Regulation*, 5(1), 2016.

[268] Erion Çano and Maurizio Morisio. Characterization of public datasets for recommender systems. In *2015 IEEE 1st International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI)*, pages 249–257, 2015. doi: 10.1109/RTSI.2015.7325106.

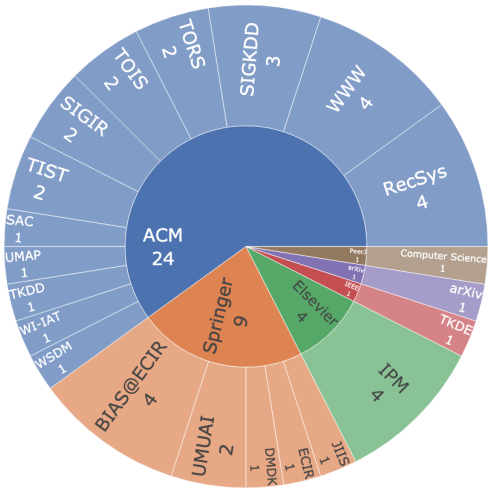
Appendix

This section serves as supplementary material to every chapter.

Chapter 2



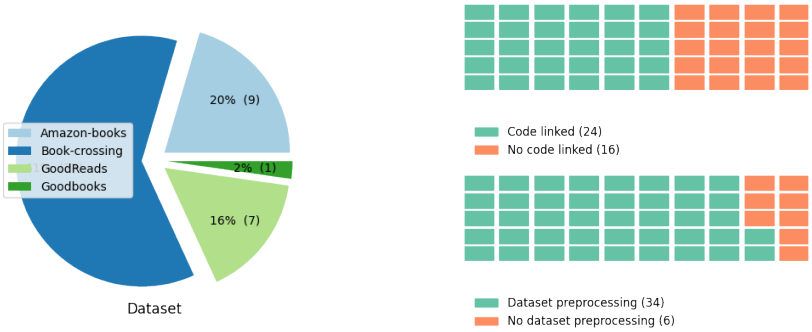
(a) Year of publication



(b) Venue and publisher

Figure 8.1: Papers per publication year, venue, and publisher. *We conducted the survey in 2024 but one of the papers has official copyright in 2025.

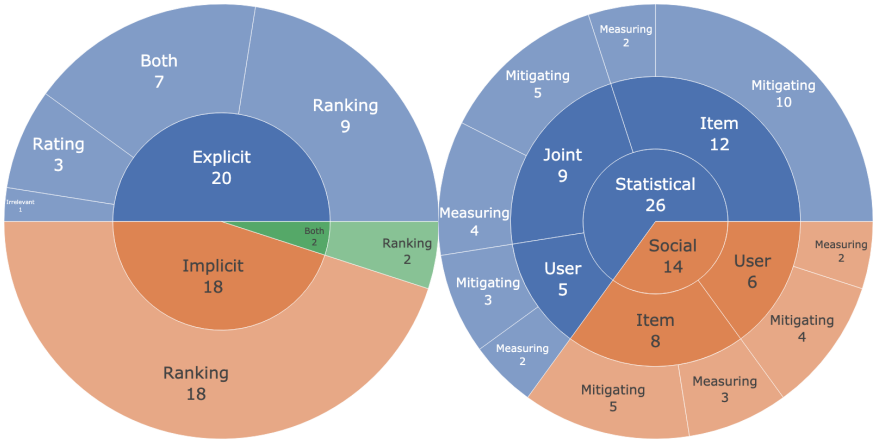
Figures 8.1, 8.2 and 8.3 show basic statistics about the papers. According to figure 8.1a, the big bulk of computer science research on bias in book recommender systems is within the last few years. Figure 8.1b shows the paper distribution among publishers and venues. Unsurprisingly, most of the papers come from ACM publications. Otherwise, the papers we included are scattered across conferences and journals. In figure 8.2a we see that *Book-Crossing* is used in more than half of the papers, with *Goodreads* and *Amazon-books* used in the majority of the other papers. The results across papers are not necessarily comparable even if they use the same dataset, since, as figure 8.2b shows, the majority of the papers train their models in a preprocessed version of the given dataset (e.g., k-core filtering). According to the same figure, a bit more than half of the papers provide a link to a repository where they include their code. Figure 8.3a shows that roughly the same amount of papers use explicit (i.e., ratings) and implicit (i.e., interactions) feedback as the basis of training a model. Regardless of the type of features, most of the papers analyze bias given a ranking task. This result is in line with recent trends in recommender systems which favor ranking rather than rating



(a) Dataset used

(b) Code availability and dataset preprocessing

Figure 8.2: Papers per dataset, code inclusion, and dataset preprocessing. Note that Goodreads and Goodbooks stem from the same source.



(a) Prediction task and type of feedback

(b) Bias type, bias perspective and action taken

Figure 8.3: Papers per prediction task, type of feedback, type of bias, and action taken

prediction. Most papers study statistical bias, especially from an item perspective, as figure 8.3b shows. Across bias types and perspectives there is a balance between measurement and mitigation, with the exception of statistical bias towards items where most papers present a mitigation strategy.

CHAPTER 3
CODING SCHEME AND MENTIONS

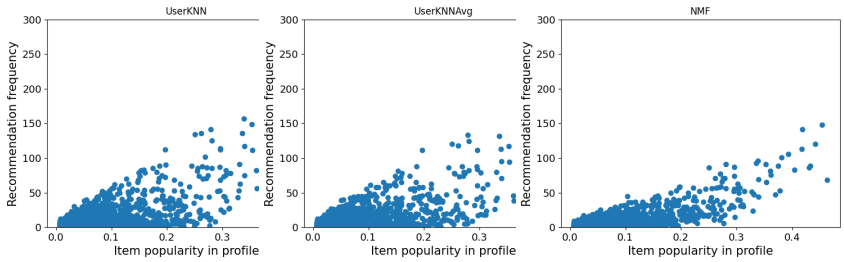
Table 8.1: Final coding scheme, containing both aspects and tactics. Aspects are divided into Item, Person and World aspects.

		B1	B2	B3	B4	B5	B6	L1	L2	L3	L4	N1	N2
Item	category / genre	x	x	x	x	x	x	x	x	x	x	x	
	complexity		x	x	x						x	x	
	cost				x								
	creator (person)	x		x	x	x		x	x	x	x		
	geographic location	x	x	x	x	x	x	x		x		x	x
	language		x	x				x					
	newsworthiness / quality	x	x	x	x	x		x			x	x	x
	medium												x
	subject (person)	x	x	x	x	x	x		x		x		x
	popularity	x	x	x	x	x	x	x		x		x	x
	publication date	x		x	x			x				x	x
	recurring			x								x	
	relevance	x	x	x	x	x		x		x		x	x
	sentiment					x							
	target audience		x	x									x
	time investment										x	x	
	time period discussed				x	x		x					
	topic	x	x	x	x	x	x	x	x	x	x	x	x
Person	ability								x	x			
	age	x	x	x	x			x	x	x			
	cultural background		x		x	x	x	x	x				
	education				x							x	x
	ethnicity		x	x	x				x				
	gender identity		x	x	x	x		x	x	x			
	geographic location	x	x	x	x		x	x		x		x	
	living situation			x	x	x	x		x			x	
	nationality		x	x	x	x			x	x	x	x	
	religion	x	x	x	x	x	x			x			
	sexuality	x			x	x	x	x	x	x	x		
	viewpoint	x	x	x	x	x	x	x	x	x		x	
	user; history	x	x	x	x				x	x		x	
	user; preferences	x		x	x	x			x	x		x	
World	current events			x		x						x	x
	society				x	x				x		x	
Tactics													
	different in item		x		x	x	x					x	
	different in list	x	x	x	x	x	x	x	x	x	x		x
	similar to user	x		x		x	x	x	x	x	x	x	x
	different from user	x	x			x	x		x	x			
	different/similar to world		x		x	x	x	x			x		

CHAPTER 4

LASTFM: COMPARISON BETWEEN COMPLETE AND FILTERED DATASET

As described in Section 4.4.1, for computational reasons we filtered the LastFM dataset. We assessed whether the sampling has a large impact on the conclusions by comparing our results with the results reported by Kowald et al. [122]. In order to ensure sufficient compareability between the results, we applied the code provided by Kowald et al. [122] on our sampled dataset. We plot the relation between popularity in profile and recommendation frequency for every item in the sampled dataset, given the algorithms *UserKNN*, *UserKNN with means*, and *NMF*, which are the algorithms reported by Kowald et al. [122]. Figure 8.4 shows that there is a consistent correlation between popularity in profile and frequency of recommendation. This conclusion aligns with the results reported by Kowald et al. [122] and supports our argument that the consistent correlation stems from the evaluation strategy deployed, regardless the filtering of the dataset.



(a) UserKNN (b) UserKNN with means (c) NMF

Figure 8.4: Relation between recommendation frequency and popularity in profile for the sampled LastFM dataset.

Additionally, we calculated the *MAE* metric for each user group of the sampled dataset, as seen in Table 8.2. While the exact number differ from the ones reported by Kowald et al. [122], the trends are consistent, with the *Niche* user group receiving the worst rating predictions for the given algorithms, evaluation strategy and division of users in groups.

Table 8.2: MAE results for *UserKNN*, *UserKNN with means* and *NMF*, applied on the sampled LastFM dataset. The worst results are always given for the *Niche* user group (statistically significant according to a t-test with p<0.005 as indicated by **). Across the algorithms, the best results are provided by *NMF*.

User group	UserKNN	UserKNN with means	NMF
Niche	62.088**	55.422**	45.473**
Diverse	54.599	45.171	37.537
Blockbuster-focused	57.353	52.147	45.584
All	57.397	50.088	42.239

Extensive Results

Overall Popularity Bias Propagation

We plot item popularity in profile versus frequency of appearance in the recommended lists of each algorithm, for every dataset and evaluation strategy, in Figures 8.5 to 8.14.

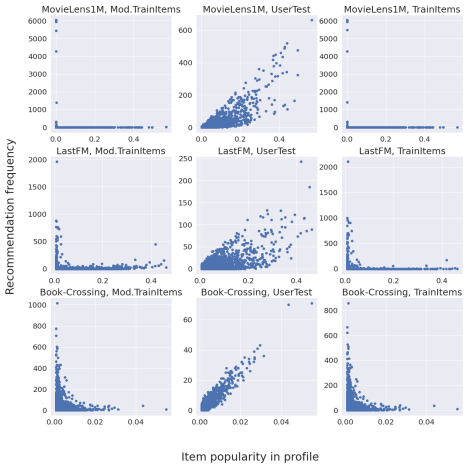


Figure 8.5: Item popularity in profile versus frequency of recommendation by the algorithm UserKNN, for every dataset and evaluation strategy.

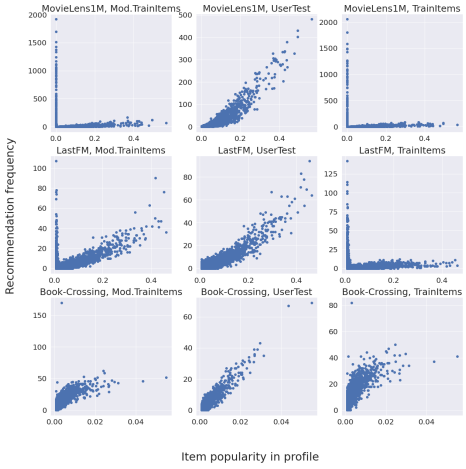


Figure 8.6: Item popularity in profile versus frequency of recommendation by the algorithm ItemKNN, for every dataset and evaluation strategy.

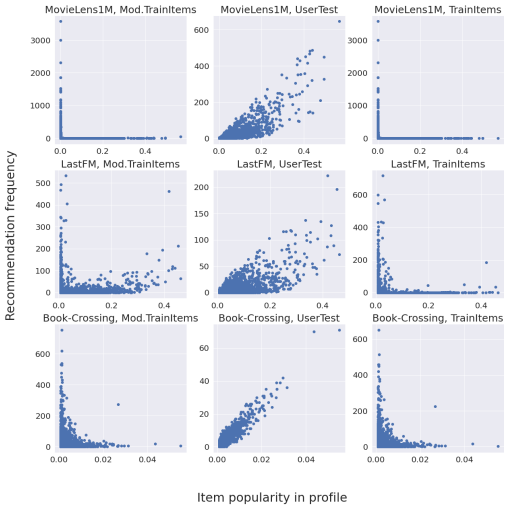


Figure 8.7: Item popularity in profile versus frequency of recommendation by the algorithm UserKNN with means, for every dataset and evaluation strategy.

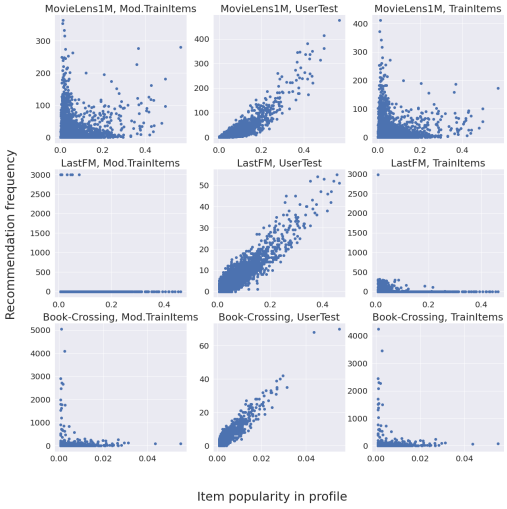


Figure 8.8: Item popularity in profile versus frequency of recommendation by the algorithm MF, for every dataset and evaluation strategy.

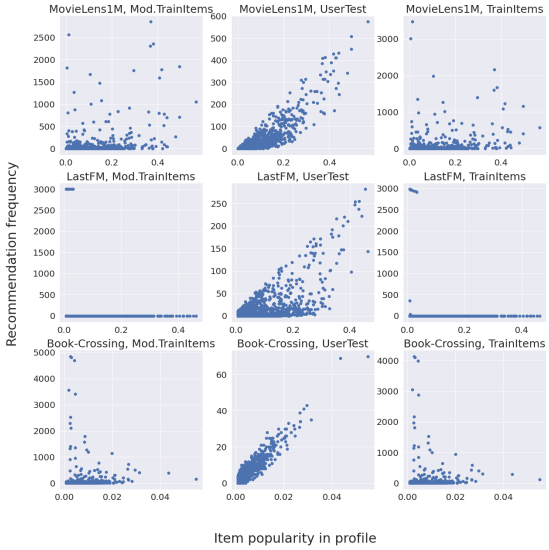


Figure 8.9: Item popularity in profile versus frequency of recommendation by the algorithm PMF, for every dataset and evaluation strategy.

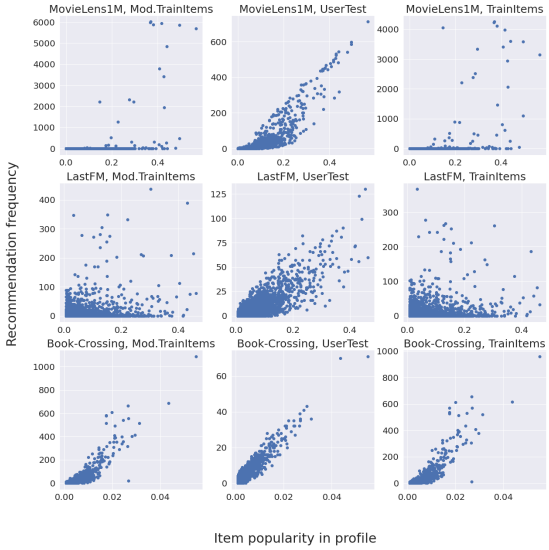


Figure 8.10: Item popularity in profile versus frequency of recommendation by the algorithm WMF, for every dataset and evaluation strategy.

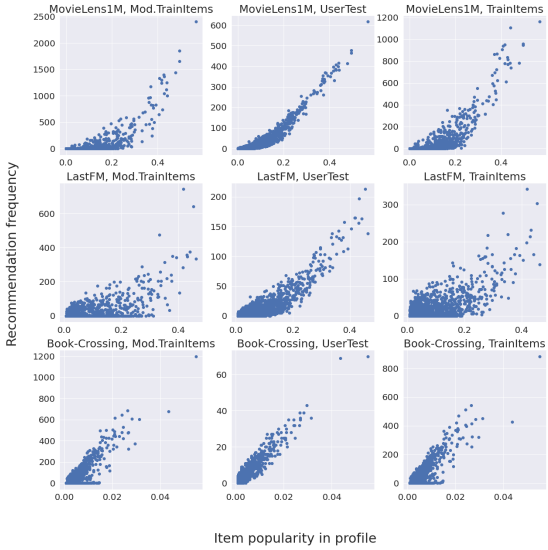


Figure 8.11: Item popularity in profile versus frequency of recommendation by the algorithm HPF, for every dataset and evaluation strategy.

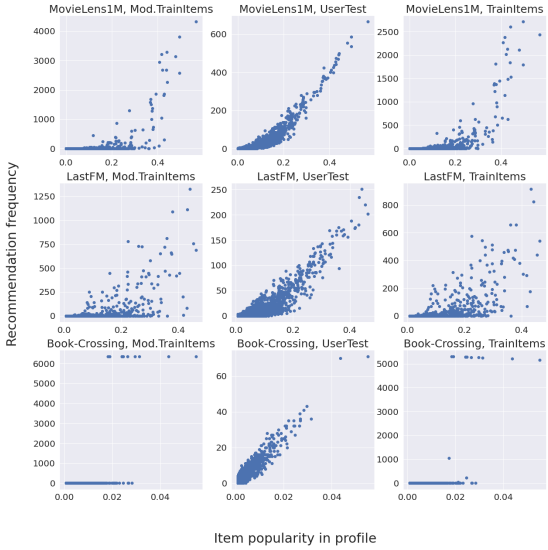


Figure 8.12: Item popularity in profile versus frequency of recommendation by the algorithm NeuMF, for every dataset and evaluation strategy.

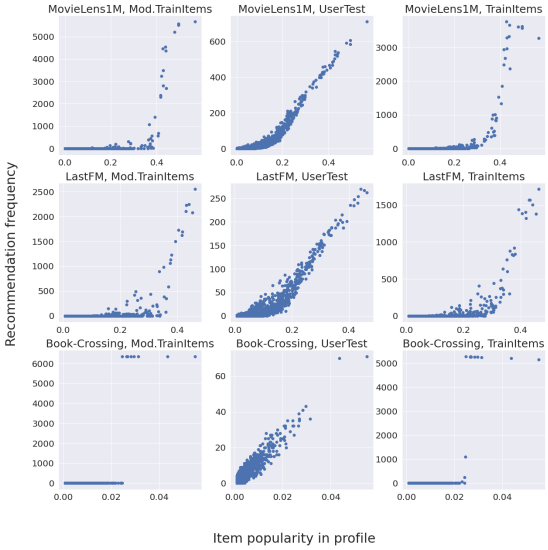


Figure 8.13: Item popularity in profile versus frequency of recommendation by the algorithm BPR, for every dataset and evaluation strategy.

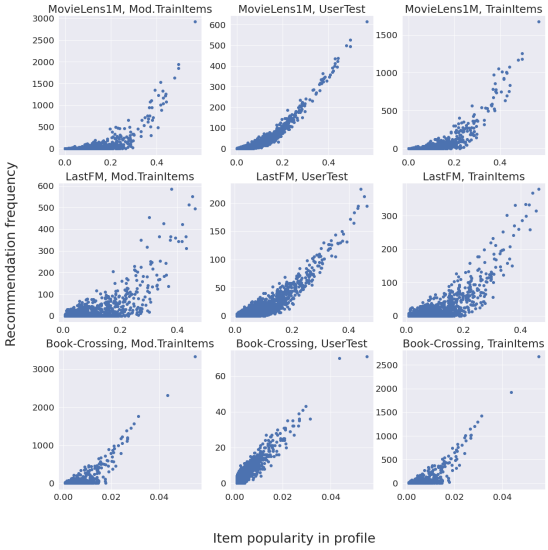


Figure 8.14: Item popularity in profile versus frequency of recommendation by the algorithm VAECF, for every dataset and evaluation strategy.

POPULARITY BIAS PROPAGATION PER USER GROUP

We include the $\%\Delta GAP$ values and $NDCG@10$ for the datasets MovieLens1M and Book-Crossing, for every algorithm, evaluation strategy and user division, in Tables 8.3 to 8.14.

Table 8.3: Percentage of increase in average popularity of items in recommendation versus in the MovieLens1M dataset ($\%\Delta GAP$) for every user group, algorithm and way of dividing users, when *Modified TrainItems* is used. ** indicates that the corresponding result is significantly higher than for the two other groups, while * indicates that it is significantly higher than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	Modified TrainItems					
	PopularPercentage			**AveragePopularity**		
	N	D	BF	N	D	BF
UserKNN	-99.4*	-99.5*	-99.7	-99.3	-99.5	-99.7
ItemKNN	-47.5**	-77.6*	-96.4	-35.6**	-78.9	-98.3
UserKNN with means	-98.4*	-99.1*	-99.5	-98.2	-99.1	-99.6
BPR	292.6	187.8	120.5	300.2	191.7	112.2
MF	-34.2**	-57.6*	-68.8	-28.2**	-57.6*	-71.0
PMF	71.7**	38.8*	20.0	73.7**	36.6*	7.9
NMF	-91.6	-94.3	-95.4	-91.4	-94.2	-95.6
WMF	249.8**	156.5*	95.3	258.4**	159.8*	87.4
HPF	126.3**	109.8*	83.0	130.6**	110.6*	80.4
NeuMF	214.8	156.6	108.2	221.0	159.5	101.6
VAECF	91.3	115.0**	91.3	93.2*	116.3**	88.2

Table 8.4: Percentage of increase in average popularity of items in recommendation versus in the MovieLens1M dataset (*%ΔGAP*) for every user group, algorithm and way of dividing users, when *UserTest* is used. ** indicates that the corresponding result is significantly higher than for the two other groups, while * indicates that it is significantly higher than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	UserTest					
	PopularPercentage			**AveragePopularity**		
	N	D	BF	N	D	BF
UserKNN	44.0	24.2	8.8	48.4	24.9	6.0
ItemKNN	55.5	21.7	6.2	61.9	21.6	4.4
UserKNN with means	37.4	21.8	8.7	41.0	22.5	6.2
BPR	98.5	55.5	20.1	110.4	56.6	13.8
MF	24.7	15.5	6.0	28.5	15.6	4.4
PMF	44.2	25.3	10.0	50.6	25.7	6.8
NMF	50.1	28.0	11.1	55.9	28.7	7.7
WMF	88.3	47.8	17.2	98.6	48.7	11.9
HPF	59.9	38.8	15.1	68.0	39.6	10.4
NeuMF	71.2	47.0	18.2	81.2	47.9	12.6
VAECF	47.3	40.8	16.7	54.5	42.0	11.7

Table 8.5: Percentage of increase in average popularity of items in recommendation versus in the MovieLens1M dataset (*%ΔGAP*) for every user group, algorithm and way of dividing users, when *TrainItems* is used. ** indicates that the corresponding result is significantly higher than for the two other groups, while * indicates that it is significantly higher than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	TrainItems					
	PopularPercentage			**AveragePopularity**		
	N	D	BF	N	D	BF
UserKNN	-99.5	-99.7	-99.8	-99.5	-99.7	-99.8
ItemKNN	-60.6**	-82.9*	-97.1	-52.4**	-83.7*	-98.6
UserKNN with means	-99.6	-99.7	-99.8	-99.6	-99.7	-99.8
BPR	269.9	170.2	108.9	274.4	174.2	101.5
MF	-45.1**	-62.9*	-70.7	-42.3**	-62.5*	-72.5
PMF	45.5**	16.6*	3.0	45.5**	15.5*	-6.6
NMF	-93.0	-95.3	-96.2	-93.0	-95.1	-96.3
WMF	215.6*	129.5*	79.3	220.3**	133.2*	72.0
HPF	104.3**	82.9*	60.0	108.8**	83.4*	57.8
NeuMF	187.8	133.4	93.5	191.2	136.0	88.4
VAECF	82.2*	92.8**	72.5	83.6*s	93.7*	70.5

Table 8.6: Percentage of increase in average popularity of items in recommendation versus in the Book-Crossing dataset (%Δ*GAP*) for every user group, algorithm and way of dividing users, when *Modified TrainItems* is used. ** indicates that the corresponding result is significantly higher than for the two other groups, while * indicates that it is significantly higher than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	Modified TrainItems					
	PopularPercentage			**AveragePopularity**		
	N	D	BF	N	D	BF
UserKNN	25.2**	-54.5*	-73.3	70.9*	-51.2	-80.3
ItemKNN	56.2**	-34.4*	-61.3	109.4**	-28.1	-72.4
UserKNN with means	22.1**	-54.9*	-73.5	67.2	-51.4	-80.8
BPR	1119.0*	523.9*	295.2	1434.7	572.3	228.4
MF	23.0**	-23.9*	-57.8	44.7**	-18.3*	-62.4
PMF	186.2**	45.0*	-5.2	253.0**	53.1*	-15.1
NMF	-47.1	-72.8	-82.7	-33.0	-70.8	-85.6
WMF	39.0	109.9**	81.4*	49.1	118.9**	67.0*
HPF	283.6**	155.3*	91.3	369.7**	168.4	71.7
NeuMF	1030.7	478.7	266.6	1323.5	523.6	204.6
VAECF	420.8**	281.1*	193.2	553.3**	300.9	158.4

Table 8.7: Percentage of increase in average popularity of items in recommendation versus in the Book-Crossing dataset (%Δ*GAP*) for every user group, algorithm and way of dividing users, when *UserTest* is used. ** indicates that the corresponding result is significantly higher than for the two other groups, while * indicates that it is significantly higher than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	UserTest					
	PopularPercentage			**AveragePopularity**		
	N	D	BF	N	D	BF
UserKNN	-5.5	1.3	4.1	2.5	-0.6	3.7
ItemKNN	-5.5	0.9	3.9	2.4	-1.0	3.6
UserKNN with means	-5.6	1.0	4.1	2.3	-0.9	3.6
BPR	-5.2	2.4	4.3	2.8	0.6	3.8
MF	-5.6	0.8	4.0	2.4	-1.0	3.6
PMF	-5.3	1.2	4.2	2.6	-0.6	3.7
NMF	-5.5	1.1	4.0	2.4	-0.7	3.5
WMF	-5.4	2.1	4.3	2.6	0.4	3.8
HPF	-5.3	1.8	4.2	2.6	0.0	3.7
NeuMF	-5.2	2.3	4.3	2.8	0.6	3.8
VAECF	-5.6	1.9	4.3	2.4	0.1	3.8

Table 8.8: Percentage of increase in average popularity of items in recommendation versus in the Book-Crossing dataset (%ΔGAP) for every user group, algorithm and way of dividing users, when *TrainItems* is used. ** indicates that the corresponding result is significantly higher than for the two other groups, while * indicates that it is significantly higher than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	TrainItems					
	PopularPercentage			**AveragePopularity**		
	N	D	BF	N	D	BF
UserKNN	24.9**	-55.1*	-73.2	69.8**	-51.7	-80.2
ItemKNN	55.7**	-34.9*	-61.5	110.1**	-28.9*	-72.3
UserKNN with means	21.7**	-55.7*	-73.3	65.6*	-52.0	-80.7
BPR	1116.6	517.8	291.0	1434.6	568.3	220.3
MF	26.1**	-24.2*	-60.9	49.6**	-17.9*	-66.0
PMF	185.9**	41.8*	-8.4	256.0**	50.8*	-20.0
NMF	-47.2	-72.9	-82.9	-32.9	-71.0	-85.6
WMF	94.7	138.8**	92.2	107.0*	155.7**	68.7
HPF	271.4**	142.6*	82.3	364.7**	159.2	54.5
NeuMF	1027.0	470.5	260.6	1323.3	518.0	193.9
VAECF	419.1**	274.1*	191.1	554.3**	297.0	149.3

Table 8.9: *NDCG@10* in the MovieLens1M dataset for every user group, algorithm and way of dividing users, when *Modified TrainItems* is used. *N*, *D* and *BF* signify *Niche*, *Diverse* and *Blockbuster-focused*. ** indicates that the corresponding result is significantly lower than for the two other groups, while * indicates that it is significantly lower than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	Modified TrainItems					
	PopularPercentage			**AveragePopularity**		
	N	D	BF	N	D	BF
UserKNN	0.004	0.000	0.000	0.004	0.000	0.000
ItemKNN	0.046	0.037	0.005**	0.059	0.033*	0.003**
UserKNN with means	0.002	0.000	0.000	0.002	0.000	0.000
BPR	0.316	0.336	0.320	0.311	0.337	0.322
MF	0.125	0.082*	0.040**	0.135	0.081*	0.032**
PMF	0.197	0.194	0.204	0.206	0.194	0.196
NMF	0.011	0.005*	0.004*	0.012	0.005*	0.004*
WMF	0.263	0.286	0.278	0.266	0.282	0.289
HPF	0.368	0.373	0.362	0.377	0.371	0.357
NeuMF	0.332	0.355	0.340	0.330	0.356	0.339
VAECF	0.357	0.376	0.381	0.356	0.380	0.370

Table 8.10: *NDCG@10* in the MovieLens1M dataset for every user group, algorithm and way of dividing users, when *UserTest* is used. *N*, *D* and *BF* signify *Niche*, *Diverse* and *Blockbuster-focused*. ** indicates that the corresponding result is significantly lower than for the two other groups, while * indicates that it is significantly lower than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	UserTest					
	PopularPercentage			**AveragePopularity**		
	N	D	BF	N	D	BF
UserKNN	0.951**	0.957*	0.961	0.951**	0.957*	0.961
ItemKNN	0.949**	0.952	0.951*	0.951	0.952	0.951
UserKNN with means	0.953**	0.960*	0.964	0.953**	0.959*	0.964
BPR	0.943**	0.950*	0.958	0.944**	0.950*	0.958
MF	0.952**	0.957	0.960	0.953*	0.956*	0.961
PMF	0.965**	0.968	0.971	0.964**	0.968	0.971
NMF	0.952**	0.958*	0.963	0.953**	0.958*	0.964
WMF	0.954**	0.960*	0.965	0.956*	0.960*	0.965
HPF	0.949**	0.953*	0.959	0.949*	0.953*	0.960
NeuMF	0.944**	0.951*	0.957	0.944**	0.950*	0.957
VAECF	0.940**	0.951*	0.957	0.941**	0.950*	0.958

Table 8.11: *NDCG@10* in the MovieLens1M dataset for every user group, algorithm and way of dividing users, when *TrainItems* is used. *N*, *D* and *BF* signify *Niche*, *Diverse* and *Blockbuster-focused*. ** indicates that the corresponding result is significantly lower than for the two other groups, while * indicates that it is significantly higher than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	TrainItems					
	PopularPercentage			**AveragePopularity**		
	N	D	BF	N	D	BF
UserKNN	0.004	0.000	0.000	0.004	0.000	0.000
ItemKNN	0.135	0.075*	0.010**	0.162	0.068*	0.004**
UserKNN with means	0.002	0.000	0.000	0.002	0.000	0.000
BPR	0.477	0.489	0.444*	0.492	0.484	0.442**
MF	0.148	0.091*	0.042**	0.163	0.089*	0.033**
PMF	0.273	0.259	0.251	0.293	0.256*	0.239*
NMF	0.013	0.006*	0.004*	0.014	0.005*	0.004*
WMF	0.375	0.396	0.375	0.392	0.387	0.385
HPF	0.600	0.607	0.566*	0.630	0.603	0.549**
NeuMF	0.501	0.531	0.492*	0.516	0.529	0.484*
VAECF	0.584	0.595	0.579	0.605	0.596	0.553**

Table 8.12: *NDCG@10* in the Book-Crossing dataset for every user group, algorithm and way of dividing users, when *Modified TrainItems* is used. *N*, *D* and *BF* signify *Niche*, *Diverse* and *Blockbuster-focused*. ** indicates that the corresponding result is significantly lower than for the two other groups, while * indicates that it is significantly lower than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	Modified TrainItems					
	PopularPercentage			**AveragePopularity**		
	N	D	BF	N	D	BF
UserKNN	0.002	0.002	0.002	0.002	0.002	0.001
ItemKNN	0.015	0.009	0.005*	0.010	0.011	0.004*
UserKNN with means	0.002	0.001	0.002	0.003	0.001	0.001
BPR	0.007**	0.040*	0.068	0.000**	0.024*	0.122
MF	0.003	0.006	0.003	0.001*	0.006	0.004
PMF	0.004**	0.012	0.014	0.003**	0.011	0.019
NMF	0.002	0.002	0.000	0.001	0.002	0.000
WMF	0.026**	0.040*	0.059	0.029*	0.037*	0.063
HPF	0.005**	0.032*	0.054	0.007**	0.027*	0.068
NeuMF	0.005**	0.038*	0.064	0.000**	0.023*	0.113
VAECF	0.013**	0.052*	0.093	0.013**	0.042*	0.122

Table 8.13: *NDCG@10* in the Book-Crossing dataset for every user group, algorithm and way of dividing users, when *UserTest* is used. *N*, *D* and *BF* signify *Niche*, *Diverse* and *Blockbuster-focused*. ** indicates that the corresponding result is significantly lower than for the two other groups, while * indicates that it is significantly lower than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	UserTest					
	PopularPercentage			**AveragePopularity**		
	N	D	BF	N	D	BF
UserKNN	0.971	0.968	0.972	0.972	0.968*	0.971
ItemKNN	0.974	0.965*	0.968	0.974	0.965*	0.967*
UserKNN with means	0.970	0.963**	0.969	0.970	0.964*	0.966
BPR	0.969	0.965	0.969	0.970	0.965*	0.970
MF	0.973	0.970	0.975	0.973	0.970	0.973
PMF	0.974	0.970	0.973	0.973	0.970	0.973
NMF	0.972	0.968	0.971	0.973	0.968	0.970
WMF	0.971	0.966**	0.973	0.972	0.966**	0.973
HPF	0.969	0.965	0.968	0.970	0.965*	0.969
NeuMF	0.970	0.964*	0.969	0.970	0.964**	0.970
VAECF	0.969	0.965	0.970	0.968	0.966	0.969

Table 8.14: *NDCG@10* in the Book-Crossing dataset for every user group, algorithm and way of dividing users, when *TrainItems* is used. *N*, *D* and *BF* signify *Niche*, *Diverse* and *Blockbuster-focused*. ** indicates that the corresponding result is significantly lower than for the two other groups, while * indicates that it is significantly higher than for one of the other groups. Statistical significance was concluded based on a t-test with p<0.005.

	TrainItems					
	PopularPercentage			**AveragePopularity**		
	N	D	BF	N	D	BF
UserKNN	0.002	0.002	0.002	0.002	0.002	0.001
ItemKNN	0.016	0.009	0.006**	0.012	0.010	0.006
UserKNN with means	0.002	0.001	0.002	0.003	0.001	0.001
BPR	0.007**	0.042*	0.070	0.000**	0.026*	0.127
MF	0.003	0.006	0.004	0.001*	0.007	0.005
PMF	0.005**	0.013	0.017	0.003**	0.012	0.022
NMF	0.002	0.002	0.000	0.001	0.002	0.000
WMF	0.062*	0.067*	0.098	0.068*	0.064*	0.101
HPF	0.006**	0.038*	0.067	0.009**	0.032*	0.082
NeuMF	0.005**	0.039*	0.066	0.000**	0.024*	0.116
VAECF	0.018**	0.061*	0.108	0.018**	0.051*	0.138

SIKS Dissertation Series

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
04 Maarten van Gompel (RUN), Context as Linguistic Bridges
05 Yulong Pei (TU/e), On local and global structure mining
06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
07 Wim van der Vegt (OU), Towards a software architecture for reusable game components
08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data Augmentation-Methods for Long-Tail Entity Recognition Models
12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
20 Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
21 Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be
22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar

	23	Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
	24	Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
	25	Xin Du (TU/e), The Uncertainty in Exceptional Model Mining
	26	Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
	27	Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
	28	Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
	29	Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
	30	Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
	31	Gongjin Lan (VUA), Learning better – From Baby to Better
	32	Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
	33	Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
	34	Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
	35	Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production
2021	01	Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space
	02	Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
	03	Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
	04	Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
	05	Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
	06	Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
	07	Armel Lefebvre (UU), Research data management for open science
	08	Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
	09	Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
	10	Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
	11	Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
	12	Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
	13	Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
	14	Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support

	15	Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
	16	Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
	17	Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
	18	Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
	19	Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management
	20	Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
	21	Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
	22	Sihang Qiu (TUD), Conversational Crowdsourcing
	23	Hugo Manuel Proença (UL), Robust rules for prediction and description
	24	Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
	25	Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
	26	Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
	27	Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
	28	Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs
2022	01	Judith van Stegeren (UT), Flavor text generation for role-playing video games
	02	Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
	03	Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
	04	Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
	05	Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
	06	Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
	07	Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
	08	Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
	09	Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
	10	Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
	11	Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring

12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
14 Michiel Overeem (UU), Evolution of Low-Code Platforms
15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
16 Pieter Gijsbers (TU/e), Systems for AutoML Research
17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation
20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty
26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems
29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
32 Cezara Pastrav (UU), Social simulation for socio-ecological systems
33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions

02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
14 Selma Čaušević (TUD), Energy resilience through self-organization
15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
22 Alireza Shojaifar (UU), Volitional Cybersecurity
23 Theo Theunissen (UU), Documentation in Continuous Software Development

	24	Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
	25	Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
	26	Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
	27	Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
	28	Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
	29	Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results
2024	01	Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
	02	Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
	03	Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
	04	Mike Huisman (UL), Understanding Deep Meta-Learning
	05	Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair
	06	Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence
	07	Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
	08	Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
	09	Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
	10	Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
	11	withdrawn
	12	Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
	13	Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence
	14	Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
	15	Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health
	16	Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems
	17	Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design
	18	Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
	19	Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
	20	Ritsart Anne Plantenga (UL), Omgang met Regels
	21	Federica Vinella (UU), Crowdsourcing User-Centered Teams

22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts
23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution
24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour
25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics
26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework
27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning
28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs
29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows
32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining
33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction
34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method
35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach
36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs
37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models
38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings
39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance
40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models
41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines
42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis
43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms
44 Marcel Schmitz (OU), Towards learning analytics-supported learning design
45 Sara Salimzadeh (TUD), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making

	46	Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology
	47	Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs
	48	Ioannis Petros Samiotis (TUD), Crowd-Assisted Annotation of Classical Music Compositions
2025	01	Max van Haastrecht (UL), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems
	02	Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks
	03	Andra-Denis Ionescu (TUD), Feature Discovery for Data-Centric AI
	04	Rianne Schouten (TU/e), Exceptional Model Mining for Hierarchical Data
	05	Nele Albers (TUD), Psychology-Informed Reinforcement Learning for Situated Virtual Coaching in Smoking Cessation
	06	Daniël Vos (TUD), Decision Tree Learning: Algorithms for Robust Prediction and Policy Optimization
	07	Ricky Maulana Fajri (TU/e), Towards Safer Active Learning: Dealing with Unwanted Biases, Graph-Structured Data, Adversary, and Data Imbalance
	08	Stefan Bloemheuvel (TiU), Spatio-Temporal Analysis Through Graphs: Predictive Modeling and Graph Construction
	09	Fadime Kaya (VUA), Decentralized Governance Design - A Model-Based Approach
	10	Zhao Yang (UL), Enhancing Autonomy and Efficiency in Goal-Conditioned Reinforcement Learning
	11	Shahin Sharifi Noorian (TUD), From Recognition to Understanding: Enriching Visual Models Through Multi-Modal Semantic Integration
	12	Lijun Lyu (TUD), Interpretability in Neural Information Retrieval
	13	Fuda van Diggelen (VUA), Robots Need Some Education: on the complexity of learning in evolutionary robotics
	14	Gennaro Gala (TU/e), Probabilistic Generative Modeling with Latent Variable Hierarchies
	15	Michiel van der Meer (UL), Opinion Diversity through Hybrid Intelligence
	16	Monika Grewal (TU Delft), Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging
	17	Matteo De Carlo (VUA), Real Robot Reproduction: Towards Evolving Robotic Ecosystems
	18	Anouk Neerincx (UU), Robots That Care: How Social Robots Can Boost Children's Mental Wellbeing
	19	Fang Hou (UU), Trust in Software Ecosystems
	20	Alexander Melchior (UU), Modelling for Policy is More Than Policy Modelling (The Useful Application of Agent-Based Modelling in Complex Policy Processes)

21 Mandani Ntekouli (UM), Bridging Individual and Group Perspectives in Psychopathology: Computational Modeling Approaches using Ecological Momentary Assessment Data
22 Hilde Weerts (TU/e), Decoding Algorithmic Fairness: Towards Interdisciplinary Understanding of Fairness and Discrimination in Algorithmic Decision-Making
23 Roderick van der Weerdt (VUA), IoT Measurement Knowledge Graphs: Constructing, Working and Learning with IoT Measurement Data as a Knowledge Graph
24 Zhong Li (UL), Trustworthy Anomaly Detection for Smart Manufacturing
25 Kyana van Eijndhoven (TiU), A Breakdown of Breakdowns: Multi-Level Team Coordination Dynamics under Stressful Conditions
26 Tom Pepels (UM), Monte-Carlo Tree Search is Work in Progress
27 Danil Provodin (JADS, TU/e), Sequential Decision Making Under Complex Feedback
28 Jinke He (TU Delft), Exploring Learned Abstract Models for Efficient Planning and Learning
29 Erik van Haeringen (VUA), Mixed Feelings: Simulating Emotion Contagion in Groups
30 Myrthe Reuver (VUA), A Puzzle of Perspectives: Interdisciplinary Language Technology for Responsible News Recommendation
31 Gebrekirstos Gebreselassie Gebremeskel (RUN), Spotlight on Recommender Systems: Contributions to Selected Components in the Recommendation Pipeline
32 Ryan Brate (UU), Words Matter: A Computational Toolkit for Charged Terms
33 Merle Reimann (VUA), Speaking the Same Language: Spoken Capability Communication in Human-Agent and Human-Robot Interaction
34 Eduard C. Groen (UU), Crowd-Based Requirements Engineering
35 Urja Khurana (VUA), From Concept To Impact: Toward More Robust Language Model Deployment
36 Anna Maria Wegmann (UU), Say the Same but Differently: Computational Approaches to Stylistic Variation and Paraphrasing
37 Chris Kamphuis (RUN), Exploring Relations and Graphs for Information Retrieval
38 Valentina Maccatrozzo (VUA), Break the Bubble: Semantic Patterns for Serendipity
39 Dimitrios Alivanistos (VUA), Knowledge Graphs & Transformers for Hypothesis Generation: Accelerating Scientific Discovery in the Era of Artificial Intelligence
40 Stefan Grafberger (UvA), Declarative Machine Learning Pipeline Management via Logical Query Plans
41 Mozhgan Vazifehdoostirani (TU/e), Leveraging Process Flexibility to Improve Process Outcome - From Descriptive Analytics to Actionable Insights

42 Margherita Martorana (VUA), Semantic Interpretation of Dataless Tables: a metadata-driven approach for findable, accessible, interoperable and reusable restricted access data
43 Krist Shingjergji (OU), Sense the Classroom - Using AI to Detect and Respond to Learning-Centered Affective States in Online Education
44 Robbert Reijnen (TU/e), Dynamic Algorithm Configuration for Machine Scheduling Using Deep Reinforcement Learning
45 Anjana Mohandas Sheeladevi (VUA), Occupant-Centric Energy Management: Balancing Privacy, Well-being and Sustainability in Smart Buildings
46 Ya Song (TU/e), Graph Neural Networks for Modeling Temporal and Spatial Dimensions in Industrial Decision-making
47 Tom Kouwenhoven (UL), Collaborative Meaning-Making. The Emergence of Novel Languages in Humans, Machines, and Human-Machine Interactions
48 Evy van Weelden (TiU), Integrating Virtual Reality and Neurophysiology in Flight Training
49 Selene Báez Santamaría (VUA), Knowledge-centered conversational agents with a drive to learn
50 Lea Krause (VUA), Contextualising Conversational AI
51 Jiaxu Zhao (TU/e), Understanding and Mitigating Unwanted Biases in Generative Language Models
52 Qiao Xiao (TU/e), Model, Data and Communication Sparsity for Efficient Training of Neural Networks
53 Gaole He (TUD), Towards Effective Human-AI Collaboration: Promoting Appropriate Reliance on AI Systems
54 Go Sugimoto (VUA), MISSING LINKS Investigating the Quality of Linked Data and its Tools in Cultural Heritage and Digital Humanities
55 Sietze Kai Kuilman (TUD), AI that Glitters is Not Gold: Requirements for Meaningful Control of AI Systems
56 Wijnand van Woerkom (UU), A Fortiori Case-Based Reasoning: Formal Studies with Applications in Artificial Intelligence and Law
57 Syeda Amna Sohail (UT), Privacy-Utility Trade-Off in Healthcare Metadata Sharing and Beyond: A Normative and Empirical Evaluation at Inter and Intra Organizational Levels
58 Junhan Wen (TUD), "From iMage to Market": Machine-Learning-Empowered Fruit Supply
59 Mohsen Abbaspour Onari (TU/e), From Explanation to Trust: Modeling and Measuring Trust in Explainable Decision Support
60 Marcel Jurriaan Robeer (UU), Beyond Trust: A Causal Approach to Explainable AI in Law Enforcement
61 Shuai Wang (VUA), Links in Large Integrated Knowledge Graphs: Analysis, Refinement, and Domain Applications
62 Khaleel Asyraaf Mat Sanusi (OU), Augmenting a learning model within immersive learning environments for psychomotor skills
63 Rashid Zaman (TU/e), Online Conformance Checking on Degraded Data
64 Jens d'Hondt (TU/e), Effective and Efficient Multivariate Similarity Search

	65	Aswin Balasubramaniam (UT), Disentangling Runner Drone Interaction Potentialities
2026	01	Pei-Yu Chen (TUD), Human-Agent Alignment Dialogues: Eliciting User Information at Runtime for Personalized Behavior Support
	02	Hezha Hassan Mohammedkhan (TiU), Estimating Body Measurements of Children from 2D Images: Towards the Automatic Detection of Malnutrition
	03	Kyriakos Psarakis (TUD), Democratizing Scalable Cloud Applications: Transactional Stateful Functions on Streaming Dataflows
	04	Boyu Xu (UU), Exploring Indirect Relations Between Topics in Neuroscience Literature Using Augmented Reality to Inform Experimental Design
	05	Koen Minartz (TU/e), Stochastic Simulation with Geometric Deep Generative Models
	06	Azim Afroozeh (CWI, VUA), FastLanes: A Next-Gen File Format
	07	Inès Blin (VUA), Narrative Understanding with Knowledge Graphs
	08	Paul van Vulpen (UU), Debating Digital Dominance: Decentralized Technology Governance For Strategic Autonomy
	09	Afrizal Doewes (TU/e), Rethinking Automated Essay Scoring: Agreement, Fairness, and Feedback
	10	Nikolaos Delapaschos Kondylidis (VUA), Establishing Task-Oriented Understanding between Agents
	11	Işıl Baysal Erez (UT), Handling Missing Data with Meta-Learning and Large Language Models
	12	Xue Li (UvA), From Fine-tuning to Prompting: A Paradigm Shift in Knowledge Graph Construction
	13	Isaac da Silva Torres (VUA), Guidelines To Flux Between Conceptual Models: Understanding Complex Digital Business Ecosystems
	14	Philip Lippmann (TUD), Synthetic Data for Robust Language Modelling
	15	Rashmi Khazanchi (OU), Artificial Intelligence in Education: Impact of AI-Based Systems on Mathematics Achievement
	16	Carolina Ferreira Gomes Centeio Jorge (TUD), Modelling Artificial Trust for Effective Human-AI Teamwork
	17	Maria Tsfasman (TUD), Towards Predicting Memory in Multimodal Group Interactions
	18	Riccardo Lo Bianco (TU/e), Deep Reinforcement Learning for Automated Decision-Making in Process Management Systems
	19	Israel Campero Jurado (TU/e), Innovations in Optimization and Applications in Healthcare
	20	Iftitahu Ni'mah (TU/e), Contrastive Learning and Evaluation in Low Resource Scenario of Natural Language Processing
	21	Francisco N.F.Q. Simoes (UU), Causality, Information, and Decision-Making
	22	Ruben Verhagen (TUD), Transparent and Explainable Agents for Human-Agent Teaming
	23	Eduardo Calò (UU), Automatically Expressing the Meaning of Logical Formulae in Natural Language

24 Shuai Han (UU), Improving Sample Efficiency of Reinforcement Learning: Exploiting Structural Knowledge for Decision Making
25 Francis Saa-Dittoh (VUA), From Radio to AI: African Community-Driven Development of Sustainable Information Systems
26 Mohammed Al Owayyed (TUD), Interactive Simulation-Based Learning Tools for Training Children's Helpline Counsellors
27 Bram Grooten (TU/e), Adaptive Reinforcement Learning: Lean and Dynamic Agents for Robust Generalization
28 Anouck Braggaar (TiU), Does This Answer Your question? Evaluating, Replicating, and Improving Task-Oriented Dialogue Systems
29 Afsana Khan (UM), Unlocking the Value of Data with Vertical Federated Learning
30 Yucheng Yang (TU/e), Generalization in Reinforcement Learning: Theories and Algorithms for Downstream Task and Multi-objective Trade-off Adaptation
31 Luis Pedro Silvestrin (VUA), Efficient Machine Learning for Time-Varying Data: Contributions to Time-Series Machine Learning and Transfer Learning
32 Giovanni Varricchione (UU), Declarative Specifications for Efficient and Safe Reinforcement Learning
33 Markus Funke (VUA), Carving Sustainability into Software Architecture
34 Melika Ayoughi (UvA), External and Internal Semantics for Visual Understanding
35 Clinton Cao (TUD), Modeling Behavior Patterns in Microservice Applications: Practical Applications Using Finite State Machines
36 Jonathan Kamp (VUA), Interpreting the Methods that Interpret Language Models
37 Emre Erdogan (UU), Computational Theory of Mind for Effective Hybrid Intelligence
38 Mahmoud Shokrollahi-Far (TiU), Quran Mining: Computational Stylometry of Quranic Texts
39 Aleksandr Chebykin (CWI, TUD), Efficient Evolutionary Hyperparameter Optimization for Deep Learning
40 Kateryna Ihorivna Holubinka (OU), Holistic Integration of Desktop Virtual Reality Technology in Higher Education: A Design-Based Research Approach
41 Hideaki Joko (RUN), Personalization and Evaluation of Conversational Information Access
42 Nedo Alexander Bartels (VUA), Conceptualizing Platform Revenue Models: A Taxonomy-Driven Design Approach
43 Maksim Gladyshev (UU), "Who Is to Blame?" and "What Is to Be Done?" A Formal Study of Counterfactual Approaches to Responsibility and Causation in (Multi-Agent) AI Systems
44 Iffat Fatima (VUA), Software Architecture Evaluation for Sustainability
45 Tristan Tomilin (TU/e), Scalable Benchmarking of Continual and Safe Embodied Reinforcement Learning

46 Savvina Daniil (CWI, VUA), Bias in Book Recommendation

Acknowledgments

Throughout my PhD, I have attended various defenses of colleagues and friends and, let's admit it, the "Acknowledgements" section is the most widely and eagerly read part of any thesis. Every time I read one, I fantasized about what my acknowledgements would say. Who would I thank, and in what order? Would I find the right words to express how thankful I am to everyone around me? Would these words be simple yet meaningful, nuanced yet easily understood, funny and powerful and moving at the same time? Dear reader, I don't know. To be honest, I struggle with words. I love them but I fear them, I fear the impact they have, the way their sound waves ripple and disrupt the balance. All this to say that I very much enjoyed writing this section, and I hope you read it the same way I wrote it – with caution.

First, I'd like to wholeheartedly thank my daily supervisor and co-promoter, Laura Hollink. When I was interviewed for the PhD position in 2021, Laura asked me my most dreaded question: why had I left my previous job? Thankfully, I had already decided to be honest and admit that it was not a career move, but rather an emotional one. And I say thankfully because Laura's reaction is what ultimately made me accept this position. She simply and spontaneously acknowledged that all decisions are emotional ones, and that it is not strange that I left somewhere where I just didn't feel well. In the following 4 years, Laura's supervision very much delivered on its early promises. Her support was constant and consistent. Whether it was help with writing papers and proposals, brainstorming on future directions, feasibility assessment or asking for input on how to navigate the politics of academia, Laura was my go-to person. Most importantly, I felt believed in and trusted for what I was; not a paper producing machine, but a whole human being with my own flaws and abilities. Laura, when I say I couldn't have done it without you, I am being absolutely literal.

I am very thankful to my promoters, Jacco van Ossenbruggen and Cynthia Liem. Jacco, for being honest and direct from the start, yet encouraging and generous with his praises. Cynthia, for always providing a wide, deep and thoughtful perspective, and for stepping up to support me when it was needed. Both of them had key roles in shaping the project, completing various chapters of this thesis including the introduction and conclusion, and making sure I can defend on time. Jacco and Cynthia contributed greatly to my project with their time and energy, and I thank them for that.

I am deeply grateful to the members of my defense committee, dr. Victor de Boer, dr. Michael Ekstrand, dr. Maria Soledad Pera, prof. dr. Anellien Smets, and prof. dr. Nina Tintarev for reading and providing valuable feedback to my thesis.

I am also very grateful to KB, nationale bibliotheek for funding and supporting my PhD project. I am thankful to Martijn Kleppe and Rosemarie van der Veen-Oei for shaping the project. Mirjam Cuper was instrumental in providing necessary feedback and, crucially, making sure I can start a postdoc as smoothly as possible. I would like to thank Michel de Gruijter for his consistent contribution to my project and Mariëlle

van Tilburg for taking the lead in continuing the collaboration. My appreciation also goes to the Cultural AI lab in general for being a meeting place of interdisciplinary perspectives and accommodating collaboration such as the one with the coauthor of the second chapter, Gauri Bhagwat.

A special thank you goes to the entire AI team of DBC Digital for welcoming my 3-month research visit and consistently offering me support in accessing, understanding and experimenting with their recommender system. Søren Mollerup in particular arranged my visit and made sure I had everything I needed to work with DBC's system. One of the most important chapters of this thesis, the final one, would not have been possible without this support.

It would take you exactly one look at the messiness of my desk at CWI to deduce that my office has always felt like home to a degree that is hard to justify. This wouldn't have been the case if the HCDA group of CWI wasn't so special, and it's all thanks to the people that comprise it. I'm thankful to Davide for his reliability and considerate input, and Lynda for breathing life into every group meeting, formal or casual. My coauthor Sanne, for being my de facto conference buddy, convincing me to not make anonymous bomb threats to avoid presenting at various venues, and being a role model of motivation, perseverance, and stepping outside one's comfort zone. My coauthor Manel, for her kindness, openness and dependability in times of need, whether professional or personal. Benjamin, Clem, Mae and Yahui for the fun conversations over lunch and chocolate breaks, board game sessions and cinema nights, and generally bringing new energy to our group. My paranymph Dirk, for his trustworthiness, listening ear, and making me believe in intimate new friendships in not-so-young-anymore adulthood. I have to give special thanks to Andrei. On my second day at CWI, a beautiful plant showed up on my desk. I may have failed to keep the plant alive, but the generous welcome stayed with me throughout. Andrei, thank you for your presence in the office and my life, and inspiring me to look inwards.

During these PhD years, tens of colleagues have left a mark on my life. My thanks to HCDA's former members Atefeh, Boyu, Delfi, Ghazaleh and Thomas for the boat rides, dinners, terrarium making classes, escape rooms, salsa nights, and the list goes on. Thanks to the activity committee of CWI for helping me feel connected to the institute and accomplish my lifelong dream of curating movie nights. I am very grateful for getting to supervise talented and hard-working Nityaa; I couldn't have asked for a better first supervision experience. I've had an amazing time at various conferences and the most enjoyable Hackathon thanks to the colleagues that I spent them with. I'm thankful for the other students at my first conference, AIES 2022, and especially my old classmate and dear friend Priyanka. The incredible group at the RecSys summer school, especially Angelo and Stavroula whom I always look forward to spending time with. The colleagues from the Netherlands at FAccT 2024, and especially Tim for the bus trip across Rio's iconic landmarks.

My friendships in the Netherlands have been an anchor and a gift. When I first moved to Amsterdam, I was mortified; how could I possibly make close friends in English? To my pleasant surprise, Amsterdam was full of people like me: foreigners, who were clearly looking for something by moving abroad, but probably didn't know exactly what. I correctly guessed that this community could become a shelter if I

needed it. And did I need it indeed! I am grateful to all the people who were there for me in one way or another in these 4 years, and will do my best to address them all.

The Hardly Working group from my Master's was there that first September already, and it's shocking to realize how many of them are still in the Netherlands, with additions even. Thank you to Amalia for always being hilarious, Dimitris for the painting classes, Gerda for the COVID relief, Giuseppe for the unforgettable wedding, Ivan for the DJ sets, Iva for the cinema nights, Jeroen for the various book clubs, Lisette for the dancing at parties, Mike for the reading together sessions, Valère for the chess games. I truly wish you all the best. Carlos, thank you for somehow always having in stock the most memorable advice and generous embrace. Daphnee, thank you for being my first truly close friend here. Elitsa, thank you for encouraging me to believe in and look forward to a wonderful future where we hang out as 80-year old besties.

I'm so very grateful for what I'd call my "Master's meetups" friends. Thank you Emma for always letting me steal your books, Hannah for bringing me chocolate when I couldn't leave the bed, Liam for your superhuman kindness that warms my heart, Luca for being the best personal trainer and board game competitor, Marysia for catching up after Waterhole in 2018 and catching up all the time ever since, Mickaël for caring about me enough to let me win in chess sometimes. You've truly been one of if not the most consistent part of my life during the PhD. Even though we're not all in the same city anymore, I never doubt your love and that gives me hope and strength. I know that we'll always find ways to be in each other's lives, just because we care and experience tells me that caring is all one needs.

Many thanks to my dearest Sohi, who has been a real point of reference for all my memories in Amsterdam. We've been through so many things together that I struggle to even recount, and I am looking forward to everything that the future holds for us. I'm very thankful to Christina, hands down the best roommate I could ever have, for all the incredibly fun concerts (obviously) but also evenings in her room with tea, updates, and mutually solving our problems. I'd like to thank Devarshi for his unlimited love, support and guidance during some very difficult moments in my life. I'm also thankful to Zsofi and Rishav with whom, despite being so far apart and rarely seeing each other, we manage to somehow stay in each other's lives.

I am super grateful to the entire Write on Point community. The story writing competition came out of nothing and yet has provided me with a lot of comfort at times of despair. For that I am thankful to all the participants whose stories kept me company for more than 3 years, including Henrik, Shane and many other people I already thanked. Special thanks to João for the best book club I've been in, Mayesha for always saying yes to singing along with my passable guitar playing, and Age and Katerina for the warm welcome to their home, board game collection and D&D campaigns. Finally, my thanks to Santiago for helping me realize a life long dream, and his patience.

Reading through this section so far, I realize that I've met and felt close to a large number of people during my PhD, people whose support had tangible effects on me. In retrospect, I don't think I would have had the confidence to try and fail and try again if it weren't for my tight knit group of friends in Greece. Thank you to my beloved and precious SSIAMM sisters: Aggeliki for always understanding exactly what I mean and giving me so much love, Iakovina for being the model of trustworthiness and

selflessness that I compare others to, Maria for making me both proud and on the verge of hysterical laughter with every other word, Melina for hugely inspiring me and believing in me for 28 consecutive years, Sofia for the overwhelming kindness and never-ending memories. Thank you to the boys who have always treated me like a sister: Giannis for the countless hours-long talks that I come out of a different person, Giorgos for agreeing to be friends with me until Christmas of 2012 and then forever, Kosmas for the unlimited supply of stories that I can never forget, Manolis for the effortless familiarity and mutual appreciation, Nikos for consistently being my main source of laughter in life, Romeo for somehow annoying and positively surprising me on the same breath. Together we've been through so many different phases of bliss and disappointment, failure and success. I miss you every day with all my heart, yet I somehow don't worry. You are meant to be in my life forever, and every time we meet in Athens or anywhere else in the world I get to realize it all over again. It has not been easy but it has been worth it.

I am thankful for the games and films and series I consumed when I needed them, and, in accordance with this thesis' themes: the books, so many books. Thank you to Kaveh Akbar's *Martyr!* and Elena Ferrante's *Neapolitan Novels* and Ursula K. Le Guin's *The Dispossessed* and Donna Tart's *The Goldfinch* and Richard Osman's *Thursday Murder Club* and Adam Becker's *What is Real?* and Gabrielle Zevin's *Tomorrow, and Tomorrow, and Tomorrow* and John Steinbeck's *East of Eden* and Lorrie Moore's *Who Will Run The Frog Hospital?* and John Green's *The Anthropocene Reviewed.* Each of them triggers a very distinct memory of PhD-related circumstances. If I put them down in order of reading them, I would make a pretty accurate one-to-one map of everything that happened in these 4 years, all the highs and lows, the mundane and the special. A life that can be reconstructed in this way is, in my (no pun intended) book, a happy life.

It's hard to thank my family in a way that doesn't feel positively cliché. Let's take, for example, my brother George. What words could I possibly use that would accurately capture the value that George brings to my life? From proactively trying to solve every single one of my problems to making me laugh with every new ridiculous antic, from driving me around Athens without complaining to enthusiastically agreeing to be my paranymph, George is a shining star that brightens up my days. I am thankful to Elena for taking good care of him while I'm away. I'm very grateful to my parents, Dina and Grigoris, for teaching me how to read and write and be independent, yet still treat me with the care that they did when I was a baby. Most importantly, I am grateful that they taught me by example how to try my best to be a good person, against all odds. Thank you to all my aunts, uncles and cousins for making me feel like home whenever we're together, and especially Dimitra for painting this beautiful cover.

Finally, I'd like to thank Dimitris. In 2021, he saw a PhD position that had something to do with books and libraries and recommender systems, thought I'd like it and encouraged me to apply. At the time, he couldn't have possibly known what the future held for us a few years down the line. Yet, I can't help but attribute this to some cosmic plan orchestrated by the powers that be to make us both much happier than we thought possible. Love of my life, thank you for everything. Σ'αγαπώ.

Savvina
Amsterdam, April 2026