Exploring Author Gender in Book Rating and Recommendation

Michael D. Ekstrand and Daniel Kluver

Feb. 4, 2021

Abstract

Collaborative filtering algorithms find useful patterns in rating and consumption data and exploit these patterns to guide users to good items. Many of these patterns reflect important real-world phenomena driving interactions between the various users and items; other patterns may be irrelevant or reflect undesired discrimination, such as discrimination in publishing or purchasing against authors who are women or ethnic minorities. In this work, we examine the response of collaborative filtering recommender algorithms to the distribution of their input data with respect to one dimension of social concern, namely content creator gender. Using publicly-available book ratings data, we measure the distribution of the genders of the authors of books in user rating profiles and recommendation lists produced from this data. We find that common collaborative filtering algorithms tend to propagate at least some of each user's tendency to rate or read male or female authors into their resulting recommendations, although they differ in both the strength of this propagation and the variance in the gender balance of the recommendation lists they produce. The data, experimental design, and statistical methods are designed to be reusable for studying potentially discriminatory social dimensions of recommendations in other domains and settings as well.

1 Introduction

The evaluation of recommender systems has historically focused on the accuracy of recommendations [Herlocker et al., 2004, Gunawardana and Shani, 2015]. When it is concerned with other

characteristics, such as diversity, novelty, and user satisfaction [Hurley and Zhang, 2011, Ziegler et al., 2005, Knijnenburg et al., 2012], it often continues to focus on the system's ability to meet traditionally-understood information needs. But this paradigm, while irreplaceable in creating products that deliver immediate value, does not tell the whole story of a recommender system's interaction with its users, content creators, and other stakeholders.

In recent years, public and scholarly discourse has subjected artificial intelligence systems to increased scrutiny for their impact on their users and society. Much of this has focused on classification systems in areas of legal concern for discrimination, such as criminal justice, employment, and housing credit decisions. However, there has been interest in the ways in which more consumer-focused systems, such as matching algorithms [Rosenblat and Stark, 2016, Hannak et al., 2016] and search engines [Magno et al., 2016], interact with issues of bias, discrimination, and stereotyping.

Social impact is not a new concern in recommender systems. *Balkanization* [van Alstyne and Brynjolfsson, 2005] (popularized by Pariser [2011] as the notion of a *filter bubble*), is one example of this concern: do recommender systems enrich our lives and participation in society or isolate us in echo chambers? Understanding the ways in which recommender systems actually interact with past, present, and future user behavior is a prerequisite to assessing the ethical, legal, moral, and social ramifications of their influence.

In this paper, we present experimental strategies and observational results from our investigation into how recommender systems interact with author gender in book data and associated consumption and rating patterns. The direct experimental outcomes of this paper characterize the distribution of author genders in existing book data sets and the response of widely-used collaborative filtering algorithms to that distribution, and assess the accuracy impact of deploying efficient strategies for adjusting the gender makeup of recommendation lists. The data and methods that we have used for this paper, however, extend beyond our immediate questions and we expect them to be useful for much more research on fairness and social impacts of recommender systems. Our data processing, experiments, and analysis are all reproducible from public data sets with the code accompanying this paper.

Our experiments address the following questions:

RQ1 How are author genders distributed in book catalog data?

RQ2 How are author genders distributed in users' book reading histories?

RQ3 What is the distribution of author genders in recommendations generated by common collaborative filtering algorithms? This measures the *overall* behavior of recommender algorithm(s) with respect to author gender.

RQ4 How do individual users' gender distributions propagate into the recommendations that they receive? This measures the *personalized* gender behavior of the algorithms.

RQ5 What control can system developers exert over recommendation distributions, and at what cost?

While we expect recommender algorithms to propagate patterns in their input data, due to the general principle of "garbage in, garbage out", the particular ways in which those patterns do or do not propagate through the recommender is an open question. Recommender systems do not always propagate all patterns from their input data [Channamsetty and Ekstrand, 2017], and it is important to understand how this (non-)propagation relates to matters of social concern.

1.1 Motivation and Fairness Construct

The work in this paper is motivated by our concern for issues of *representation* in book authorship. There are efforts in many segments of the publishing industry to improve representation of women, ethnic minorities, and other historically underrepresented groups. Multiple organizations undertake counts of books and book reviews to assess the representation of women and nonbinary individuals in the literary landscape [Pajović and Vyskocil, 2016, VIDA, 2017].

Our goal is to understand how recommendation algorithms interact with these efforts. Do recommender systems help these authors' work find the audience that will propel them to success? Are they neutral paths, neither helping nor hindering? Or is algorithmic recommendation another hurdle to their success, stacking the deck in favor of well-known authors and the status quo of the publishing industry?

Author representation also has a consumer-facing dimension: what picture does a book service's discovery layer paint of the space of book authorship? When a user is looking for books, do they see books by a diverse range of authors, or are the books that are surfaced focused on certain corners of the authorship space? This is admittedly a complex question, because recommending books that are not relevant to a user's interests or information need just because of their author's demographics does not make for an effective recommendation or information retrieval system. Fairness in recommendation needs to be understood in the context of accuracy and other measures of effectiveness.

We study this in the context of user-provided ratings and interactions collected from three sites widely used by readers. Amazon ratings are provided by Amazon users and are accompanied by textual reviews (not used in the present work) to help prospective purchasers decide whether or not to purchase a book. GoodReads and BookCrossing are reader communities, where readers catalog books they have read or wish to read, rate books, and interact with other readers. GoodReads makes extensive use of a social network, where people can form friendships to see each others' book activities, ask for personal recommendations from friends, and provide reviews to help give other readers insight into a book; the fundamental action is to add a book to a *shelf*, often one of "read", "to-read", or "currently-reading"; when adding a book to the "read" shelf, the user may also provide a rating and a textual review. In addition to the social discovery mechanisms provided by the news feed, the makeup of users' shelves is used as input to GoodReads' recommender algorithms.

The experiments in this paper are focused on *consumer-centered provider fairness*. Our framing is similar to "calibrated fairness" proposed by Steck [2018], in that we are concerned with the makeup of recommendation lists and their connection to users' input profiles. While there are many ways of conceiving of provider fairness, some of which we examine in Section 2.4, list composition seems particularly well-suited to understanding representation as it is experienced by users of

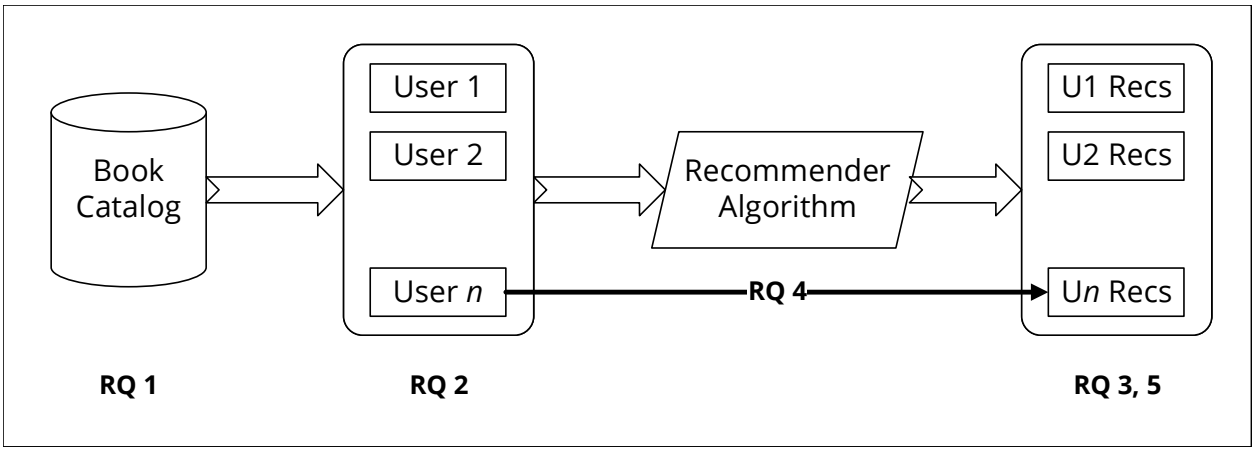

Figure 1: Experiment architecture and data flow.

the system. While our measurements focus on representation, we are measuring representation in the context of recommendation lists that have been optimized for relevance to a user's reading preferences, thus implicitly accounting for accuracy.

The purpose of this paper is not to make any normative claims regarding the distributions we observe, simply to describe the current state of the data and algorithms. We do not currently have sufficient data to determine whether the distributions observed in available data indicate under- or over-representation, or what the "true" values are. We hope that our observations can be combined with additional information from other disciplines and from future work in this space to develop a clearer picture of the ways in which recommender systems interact with their surrounding sociotechnical ecosystems. Our normative claim is that researchers and practitioners should care and seek to understand how their systems interact with these issues. Our methods provide a starting point for such experiments.

1.2 Contributions and Summary of Findings

In the main body of the paper, we provide a detailed and comprehensive account of our data and research methods. Figure 1 shows the stages of the book recommendation pipeline that forms the backbone of our experimental design, with how our research questions map to each stage. In this section, we summarize our contributions and key findings to provide a roadmap for the rest of the paper.

We operationalize gender balance as the fraction of books written by female authors, after discarding books for which we could not determine the author's gender identity. Justification and limitations of this decision are discussed in Section 3.4. This is an observational and correlational study. Our goal here is to understand *what* correlations exist; future work will explore additional variables such as genre to better understand *why* the patterns we observe exist.

1.2.1 Findings

RQ1: Gender in Book Catalogs. 23.2% of books in the Library of Congress for which we could determine the author's gender identity are written by women. Discovery platform book catalogs show higher representation of women: 31.0% of known-gender Amazon books and 38.7% of known-gender GoodReads books are written by women. We therefore see an improvement in the representation of female authors through the early stages of the pipeline as we move from presence in a generic catalog to presence in collaborative filter inputs. Section 3.6 and its supporting figures describe these results.

In the context of a recommendation application, this finding describes the makeup of the set of books that are available to be recommended.

RQ2: Gender in User Profiles. There is high variance between users' author gender balances, but the mean balance is approximately the same as the balance of the underlying set of books. This variation could be the result of many factors beyond the scope of this paper, but it is unsurprising that different users have different rating patterns. It has the benefit of providing a wide range of actual user profiles for which to test the response of components further down the pipeline. Section 5.1 describes these results.

With this finding, we understand more about the individual user histories that make up the training data for the recommender system. Many recommenders, but particularly collaborative filters, will try to learn and replicate the patterns in these profiles.

RQ3: Gender in Recommendations. Recommendation lists were comparable to user profiles in terms of both mean and variance of their gender balances, with a few exceptions. Distribution shapes, however, were markedly different, with some conditions favoring more extreme recommendation outcomes than the input user profiles. Section 5.2 describes these results.

With this we see the makeup of individual recommendation outputs, to understand what view of the book space the recommender is likely to provide to its users on average.

RQ4: Recommender Response to User Profile. Most algorithms we tested propagate users' input profile balances into their recommendation lists, particularly when operating in "implicit-feedback mode" (where we only consider *whether* a user has interacted with a book, not how much they liked it). Users who read more books by women were recommended more books by women. This shows that author gender is correlated with one or more features that drive users' consumption patterns and result in patterns that the collaborative filter captures and reflects, or it is directly one of those features. It also means, however, that a user reading mostly books by authors of one gender will likely receive recommendations that reinforce that tendency unless compensating measures are deployed. Section 5.3 describes these results.

This question gets to how the recommender's personalization capabilities respond to each user's individual tendency towards authors of a particular gender. How much of the patterns that it sees does it replicate?

RQ5: Controlling Recommendation Representation. We designed simple re-ranking strategies to force recommendation lists to meet particular balance goals, such as gender parity or a gender balance that reflects the user's rating profile. These rerankings induce little loss in recommendation accuracy (as measured with mean reciprocal rank in a train-test evaluation). This suggests that, if a system designer wishes, the gender balance of recommended items can be tuned with little cost rather than accepted as the natural consequence of the data and algorithms. Section 6 discusses these results.

1.2.2 Methodological Contributions

Data Integration. We describe an integration of six different public data sources — three datasets of user-book consumption or preference records, and three sources of book and author metadata — to study social issues in book recommendation, cataloging and justifying the data linking decisions we made along the way. We expect this composite data set to be useful for further research on reader-book interactions. Our integration strategy also serves as a case study in obtaining and preparing data for fairness and social impact research, as data collection efforts for similar studies in other domains and applications will need to make similar kinds of decisions. Section 3 describes the data pipeline in detail.

Experimental Methodology. Rigorous, reusable statistical methodologies for analyzing bias in personalization algorithms are still in their infancy. We describe an end-to-end experimental pipeline and statistical analysis for studying representation and list composition in recommendation, and how user patterns do or do not propagate into recommendation outputs. We expect the approach we take to be useful in studying equity in other recommendation and information retrieval settings, and may be more broadly useful as well. Section 4 describes the experimental pipeline.

2 Background and Related Work

Our present work builds on work in both recommender systems and in bias and fairness in algorithmic systems more generally.

2.1 Recommender Systems

Recommender systems have long been deployed for helping users find relevant items from large sets of possibilities, usually by matching items against users' personalized taste [Ekstrand et al., 2010, Adomavicius and Tuzhilin, 2005]. They are deployed for boosting e-commerce sales, supporting music and book discovery, driving continued engagement with news and social media, and in many other contexts and applications. A recommendation problem, in the abstract, usually consists of *items* $i \in I$ and *users* $u \in U$ with recorded user-item interaction $r_{ui} \in R$ often in the form of ratings or some equivalent derived from the user purchasing, consuming, or otherwise expressing interest in the item [Ekstrand and Konstan, 2019]. Each user has a set $R_u \subseteq R$

of the ratings they have provided; for the purposes of this paper, we call this their *user profile*, as it is the data a system such as GoodReads would typically store about a user's consumption history and use as the basis for their recommendations. Recommender system feedback is often divided into two classes: *explicit* feedback, such as 5-star ratings, is provided by the user to express their preference for an item; *implicit* feedback comes from user actions that, in sufficient quantity, indicate preference but are taken for consumption purposes, such as listening to a song or marking a book as "to-read".

Of particular interest to our current work is *collaborative filtering* (CF) systems, which use patterns in user-item interaction data to estimate which items a particular user is likely to find useful. These include both neighborhood-based approaches and latent factor models.

While recommender evaluation and analysis often focuses on the accuracy or quality of recommendations [Herlocker et al., 2004, Gunawardana and Shani, 2015], there has been significant work on non-accuracy dimensions of recommender behavior. Perhaps the best-known is diversity [Ziegler et al., 2005], sometimes considered along with novelty [Hurley and Zhang, 2011, Vargas and Castells, 2011]. Lathia et al. [2010] examined the *temporal* diversity of recommender systems, studying whether they changed their recommendations over time.

Jannach et al. [2015] studied recommendation bias with respect to classes of items, particularly around various levels of item popularity. Their work is similar in its goals to ours, in that it is looking to understand *what* different recommendation techniques recommend, beyond whether or not it seems to match the user's preference. We extend this line of inquiry to the socially-salient dimension of author gender.

2.2 Social Impact of Recommendations

Recommender systems researchers have been concerned for how recommenders interact with various individual and social human dynamics. One example is balkanization or filter bubbles [van Alstyne and Brynjolfsson, 2005, Pariser, 2011], mentioned earlier; recent work has sought to detect and quantify the extent to which recommender algorithms create or break down their users' information bubbles [Nguyen et al., 2014] and studied the effects of recommender feedback loops on users' interaction with items [Hosanagar et al., 2013].

Other work seeks to use recommender technology to promote socially-desirable outcomes such as energy savings [Starke et al., 2017], better encyclopedia content [Cosley et al., 2007], and new kinds of relationships [Resnick, 2001]. Our work provides the exploratory underpinnings for future work that may seek to use recommenders to specifically promote the work of underrepresented authors, and results on a first-pass set of techniques for doing so; Mehrotra et al. [2018] provide an example of pursuing such ends in the music domain.

2.3 Bias and Fairness in Algorithmic Systems

Questions of bias and fairness in computing systems are not new; Friedman and Nissenbaum [1996] considered early on the ways in which computer systems can be (unintentially) biased in their design or impact. In the last several years, there has been increasing interest in the ways that machine learning systems are or are not fair. Dwork et al. [2012] and Friedler et al. [2016] have

presented definitions of what it means for an algorithm to be *fair*. Feldman et al. [2015] provide a means to evaluate arbitrary machine learning techniques in light of *disparate impact*, a standard for the fairness of decision-making processes adopted by the U.S. legal system.

Bias and discrimination often enter a machine learning system through the input data: the system learns to replicate the biases in its inputs. This has been demonstrated in word embeddings [Bolukbasi et al., 2016] and predictive policing systems [Lum and Isaac, 2016, Ensign et al., 2018], among others.

Research has also examined how bias and potential discrimination manifest in the whole sociotechnical system, studying platforms such as TaskRabbit [Hannak et al., 2016] and OpenStreetMap [Thebault-Spieker et al., 2018]. One recent notable study by Ali et al. [2019] found discriminatory patterns in Facebook ad delivery, even when advertisers set neutral budgets and campaign parameters. Bias can also be deployed subtly, as in the decisions of some online dating platforms to reflect presumed latent racial preferences into match recommendations even when users specify that they have no racial preference for their dating partner [Hutson et al., 2018].

2.4 Fair Information Access

Burke [2017] lays out some of the ways in which questions of fairness can apply to recommender systems. In particular, he considers the difference between "C-fairness", in which consumers or users of the recommender system are treated fairly, and "P-fairness", where the producers of recommended content receive fair treatment. Burke et al. [2018] and Yao and Huang [2017] have presented algorithms for C-fair collaborative filtering, and Ekstrand et al. [2018] examine C-fairness in the accuracy of recommendation lists.

Our present study focuses on P-fairness. This dimension is somewhat related to historical concerns such as long-tail recommendation and item diversity [Jannach et al., 2015]. Kamishima et al. [2018] and Beutel et al. [2019] have presented algorithms for P-fair recommendation; calibration [Steck, 2018] can be viewed as another kind of provider fairness.

Biega et al. [2018] and Singh and Joachims [2018] provide metrics for assessing fair *exposure* to providers; this metric assess whether providers are recommended an "appropriate" number of times. Other approaches to assessing the fairness of rankings look at the makeup of the ranking or prefixes thereof [Yang and Stoyanovich, 2017, Sapiezynski et al., 2019, Zehlike et al., 2017]; this is closer to our present work, in which we try to understand how lists are composed from the perspective of gender representation.

A range of approaches are valuable at the present stage of research in fair recommendation and information retrieval, and provide varying perspectives on how to operationalize and assess fairness. In this paper, we present an offline empirical analysis of the calibrated provider fairness of several classical collaborative filtering algorithms and their underlying training data.

2.5 Representation in Creative Industries

As noted in Section 1.1, there are efforts to both improve and audit the representation of women, ethnic minorities, and other historically underrepresented groups [Pajović and Vyskocil, 2016, VIDA, 2017]. In addition to these general representation measurement efforts, Hu [2017] reports

that gender biases in book reviews differ from genre to genre; in particular, "Women are less likely to receive reviews when writing about topics that aren't deemed 'feminine.'". Bucur [2019] found that users on Amazon are more likely to co-purchase books by female authors if they are buying another book by a female author than if their initial book is by a male author, and Thelwall [2019] found that GoodReads users tend to give higher ratings to authors of their own gender.

Beyond books, Epps-Darling et al. [2020] studied gender representation in music streaming and recommendation, finding that female or mixed-gender artists comprise only 20% of organic plays, and a slightly higher fraction of recommender-driven plays. Concurrently with our expanded work, Shakespeare et al. [2020] carried out an experiment similar to ours in music recommendation and found collaborative filtering algorithms also propagating listeners' biases into their recommendations.

3 Data Sources and Integration

Traditional recommender systems experiments typically rely on rating or consumption data. There is a wide range of such data sets publicly available, including movie ratings from MovieLens [Harper and Konstan, 2015], product reviews from Amazon [McAuley et al., 2015], and artist play logs from Last.fm [Celma, 2010]. Sometimes these data sets are augmented with additional data, such as additional sources of item data or text crawled from Web pages. Studying fairness and other social dimensions of recommendation, however, require data that is not commonly provided with rating data [Ekstrand et al., 2018], requiring some creativity.

Investigating how content creator demographics relate to recommendation requires the following classes of data:

- *Consumption data* on books users have read and/or rated, to understand reading patterns and train recommendation algorithms.
- *Book data* describing books and, for our purposes, their authors.
- *Author data* describing the authors themselves, and including demographic characteristics of interest.

Fig. 2 shows how these types of data fit together and the data sets we use for each. Linking the data sets together is not easy, due both to the messiness of the data itself (e.g. malformed ISBNs) and the lack of linking identifiers.

This section provides details on our data integration, justifications of data linking decisions we made, and descriptive statistics of the resulting composite data set.[1]

3.1 User Profiles and Book Ratings

We us three public sources of user-book interactions. For each, we treat it both as an *explicit feedback* data set by consulting rating values, and as an *implicit feedback* data set by ignoring rating

[1] Documentation and code available at `https://bookdata.piret.info`

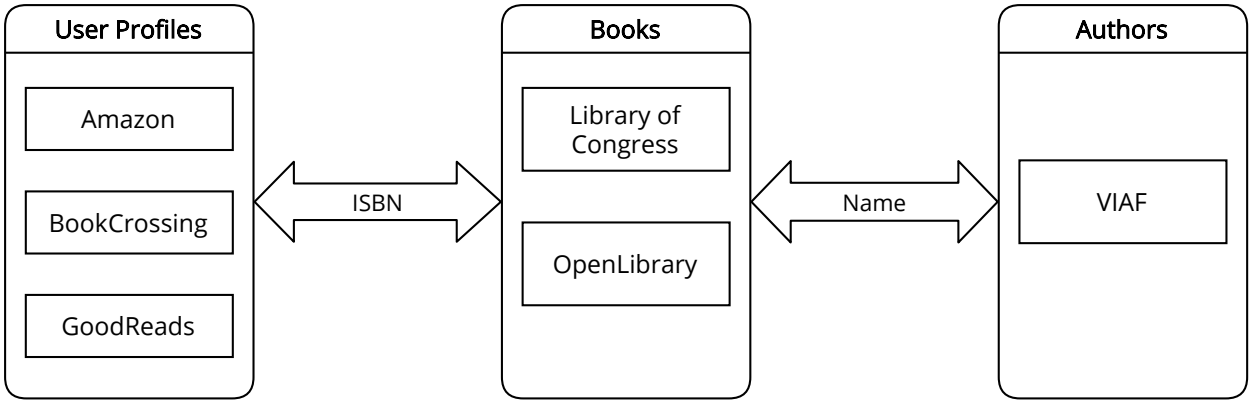

Figure 2: Data set relationships.

	Users	Items	Pairs	Density
AZ	8,026,324	2,268,142	22,460,535	0.0001%
BX-E	77,805	151,670	427,283	0.0036%
BX-I	105,283	279,501	1,129,814	0.0038%
GR-E	808,782	1,080,777	86,537,566	0.0099%
GR-I	870,011	1,096,636	188,943,278	0.0198%

Table 1: Interaction data summaries.

values and considering user-item interactions as positive signals. In implicit-feedback settings, we consider all books a user has interacted with as positive implicit signals, even if they have a low rating: this corresponds to the signal that a bookseller can derive from sales data, as they do not know whether readers actually like the books they purchase once they have read them.

The **BookCrossing** (BX) data set [Ziegler et al., 2005] contains 1.1M book interactions from the BookCrossing reading community. This data set contains both explicit ratings, on a 1–10 scale, and "implicit" actions of unspecified nature. Since not all ratings have rating values, for explicit-feedback settings we exclude implicit actions, resulting in the "BX-E" data set; "BX-I" contains all BookCrossing interactions without rating values.

The **Amazon Books** (AZ) data set [McAuley et al., 2015] contains 22.5M reviews and ratings of books provided by customers on Amazon.com. We use only the rating values, not the review text; since all recorded interactions have rating values, we use the interactions as-is and do not need to subset for explicit feedback.

The **GoodReads** (GR) data set [Wan and McAuley, 2018] contains 189M interactions including ratings, reviews, and "add to shelf" actions from GoodReads, a reading-oriented social network and book discovery service. As with BookCrossing, we extract a rating-only subset ("GR-E") for explicit-feedback analysis, and use all user-book interactions ("GR-I") for implicit feedback.

These data sets provide our historical user profiles (for RQ2) and the training data for our collaborative filtering algorithms. All three are general reading data sets, consisting of user ratings for books across a wide range of genres and styles. Table 1 summarizes these data sets' basic statis-

tics. The "Pairs" column indicates the number of unique user-item pairs that appear in the data set. We resolve multiple editions of the same work into a single item (see Section 3.3), so the item counts we report here may differ slightly from the item counts reported in other uses of these same rating data sets.

3.2 Book Bibliographic Records

We obtain book data, particularly author lists, by pooling records from Open Library[2] and the Library of Congress (LOC) MARC Open-Access Records[3].

We link these book records to rating data by ISBN. Both OpenLibrary and LOC record ISBNs for book entries, and all book rating sources record ISBNs for the books users interact with (in the BookCrossing data, ISBN is the primary key for books; Amazon uses ISBNs as the identification numbers for books that have them).

Unfortunately, ISBN fields in the Library of Congress data are inconsistently formatted and used, including ISBNs in a range of formats as well as text other than ISBNs (many book entries store the cover price in the ISBN field). We use a regular expression to look for sequences of 10 or 13 digits (allowing an X for the last digit in 10-digit sequences), optionally including spaces or hyphens, and treated those as ISBNs. We do not validate check digits, preferring to maximize the ability to match ISBNs in the wild.

3.3 ISBN Grouping

Books are often released in multiple editions, each with their own ISBNs. These can be different formats of the same text — for example, hardcover and paperback editions of the same book will have different ISBNs — or they can be revised and/or translated editions. Each edition, however, is a version of the same creative work. To reduce data sparsity, improve data linking coverage, and reflect a more accurate general-purpose recommendation scenario, we group related ISBNs into a single "item".

To group ISBNs, we form a bipartite graph of ISBNs and record IDs. Library of Congress bibliography records, OpenLibrary "edition" records, and GoodReads book records all constitute records for this purpose. In addition, OpenLibrary and GoodReads each have a concept of a "work"; when an edition or book is linked to a work, we use the work ID instead of the individual edition or book ID. We then find the connected components on this graph, consider each component to be an "item", and assign it a single item identifier.

This process serves a similar purpose as ISBN linking services such as thingISBN [Spalding, 2006] and OCLC's xISBN service, but is completely reproducible using open data sources. One limitation of this technique is that some ISBNs link multiple creative works. This can happen via, for example, in the case of multi-work collections with a single ISBN.

Rarely (less than 1% of ratings) this causes a user to have multiple ratings for a book; we resolve multiple ratings in explicit-feedback settings by taking the median rating value. Taking the most

[2] `https://openlibrary.org/developers/dumps`
[3] `https://www.loc.gov/cds/products/marcDist.php`

recent rating would also be a reasonable option, but BookCrossing does not include timestamps; since multiple ratings appear so infrequently, the precise strategy is unlikely to have significant impact on our results.

3.4 Author Gender Data

We obtain author information from the Virtual Internet Authority File (VIAF)[4], a directory of author information (*Name Authority Records*) compiled from authority records from the Library of Congress and other libraries around the world. Author gender identity (MARC Authority Field 375) is one of the available attributes for many records.

3.4.1 Gender Identity Coding

The MARC21 Authority Record data model [Library of Congress, 1999] employed by the VIAF is flexible in its ability to represent author gender identities, supporting an open vocabulary and begin/end dates for the validity of an identity. The Program for Cooperative Cataloging provides a working group report on best practices for recording author gender identities, particularly for authors who are transgender or have a non-binary gender identity [Billey et al., 2016].

Unfortunately, the VIAF does not use this flexibility — all its gender identity records are "male", "female", or "unknown". The result is that gender minorities are not represented, or are misgendered, in the available data. We agree with Hoffmann [2018] that this is a significant problem. The Library of Congress records better data, and as of August 2019 is in the process of preparing new exports of their linked data servies; we hope this will enable future research to better account for the complex nature of human gender identity and expression.

3.4.2 Linking Author Data

Because OpenLibrary, LOC, and VIAF do not share linking identifiers, we must link books to authority records by author name. Each VIAF authority record can contain multiple name entries, recording different forms or localizations of the author's name. OpenLibrary author records also carry multiple known forms of the author's name. After normalizing names to improve matching (removing punctuation and ensuring both "Last, First" and "First Last" forms are available), we locate all VIAF records containing a name that matches one of the listed names for the first author of any OpenLibrary or LOC records in a book's ISBN group. If all records that contain an assertion of the author's gender agree, we take that to be the author's gender; if there contradicting gender statements, we code the book's author gender as "ambiguous".

We selected this strategy to balance good coverage with confidence in classification. Different authors with the same full name but different genders are unlikely to be a common occurrence. Less than 2.5% of rated books have 'ambiguous' author genders. Table 2 shows relative frequency of link results for the books in our data sets; the columns correspond to the following failure points:

[4] `http://viaf.org/viaf/data/`

Data Set	No Bk	No Auth	No VIAF	Unknown	Ambig.	Male	Female
LOC	—	16.0%	5.6%	23.1%	1.0%	41.7%	12.6%
AZ	41.2%	10.0%	6.9%	10.4%	0.8%	21.2%	9.5%
BX-E	14.9%	4.3%	5.0%	11.1%	2.6%	36.9%	25.1%
BX-I	16.3%	4.6%	5.6%	12.4%	2.4%	34.7%	24.0%
GR-E	—	45.1%	3.3%	8.9%	1.0%	25.5%	16.2%
GR-I	—	45.2%	3.3%	8.9%	1.0%	25.5%	16.1%

Table 2: Summary of gender coverage (% of books with each resolution result).

1. *No Bk* means the rating or interaction could not be linked to a book record of any kind. GoodReads has 100% coverage since it comes with book records, but those records are not used for any data other than record identifiers.

2. *No Auth* means a book record was found, but had no authors listed.

3. *No VIAF* means authors were found, but none could be matched to VIAF.

4. *Unknown* means a VIAF record was found, but there were either no gender identity records or all records said "unknown".

5. *Ambiguous*, *Male*, and *Female* are the results of actual gender identity assertions.

In the remainder of this paper, we group all no-data conditions together as "unlinked"; we present coverage statistics across the pipeline to inform future reuse of the data set.

3.4.3 Coverage and Popularity

To better understand the relationship between coverage and item popularity, we examined the distribution of gender resolution statuses for each item popularity percentile. Fig. 3 shows these results; more popular items are more likely to have gender identity information available. Further, in Amazon and and GoodReads, female author representation seems to be better among the most popular books than among the less-popular ones.

The precise implications of this need further investigation. One immediate implication is that gender label coverage for books in users' profiles is higher than it would be for books selected uniformly at random. This coverage increase also applies to the recommendations from algorithms that tend to recommend more popular books. We expect that this popularity/coverage relationship will be common not just in books but in many other content categories as well, because more popular items are more likely to have broad attention and careful cataloging; items that are known only to a small number of users are also more likely to be unknown to catalogers and metadata curators. This has particular implications for studies looking at the fairness of long-tail recommendations, as the system and experiment's design would be pushing its results into portions of the item space with lower label coverage for the fairness analysis.

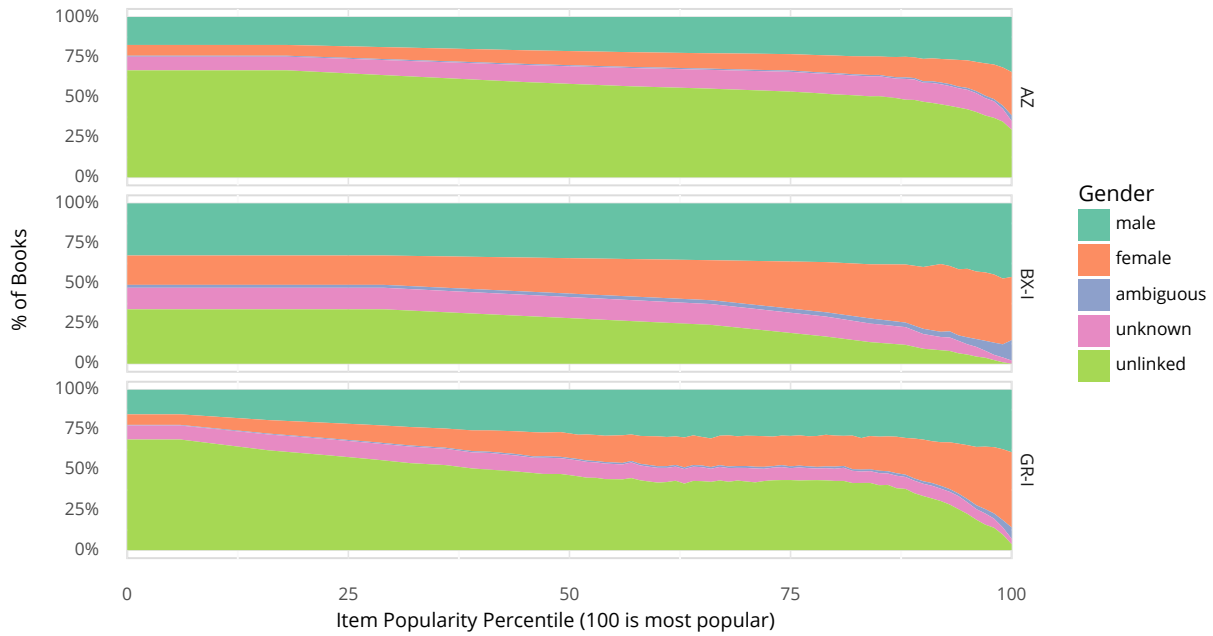

Figure 3: Gender identity coverage by item popularity (as measured by number of interactions).

3.4.4 Alternative Approaches to Author Gender

Other work on understanding the behavior of computing systems with respect to gender and other demographic attributes that have been the basis of historic and/or ongoing discrimination uses various inference techniques to determine the demographics of data subjects. This includes statistical detection based on names [Mislove et al., 2011] and the use of facial recognition technology [Riederer and Chaintreau, 2017].

Such sources, however, have been criticized as reductionistic [Hamidi et al., 2018] and often rely on and reinforce stereotypes regarding gender presentation. Further, even to the extent that face-based gender recognition does work, it is biased in recognizing gender more accurately for lighter-skinned subjects [Buolamwini and Gebru, 2018].

The Program for Cooperative Cataloging working group report specifically discourages inference of gender identity, even when the inference is performed by a human, admonishing catalogers to "not assume gender identity based on pictures or names" [Billey et al., 2016]. Catalogers following the recommendations learn an author's gender from explicit statements from official sources regarding their gender, or from the choice of pronouns or inflected nouns in official sources (such as the author's biography on the book cover).

Given the technical challenges and ethical concerns raised by the prospect of gender inference, and the recommendation of relevant working groups to avoid even human inference of gender, we choose to forego inference techniques in favor of gender identities recorded by professional catalogers.

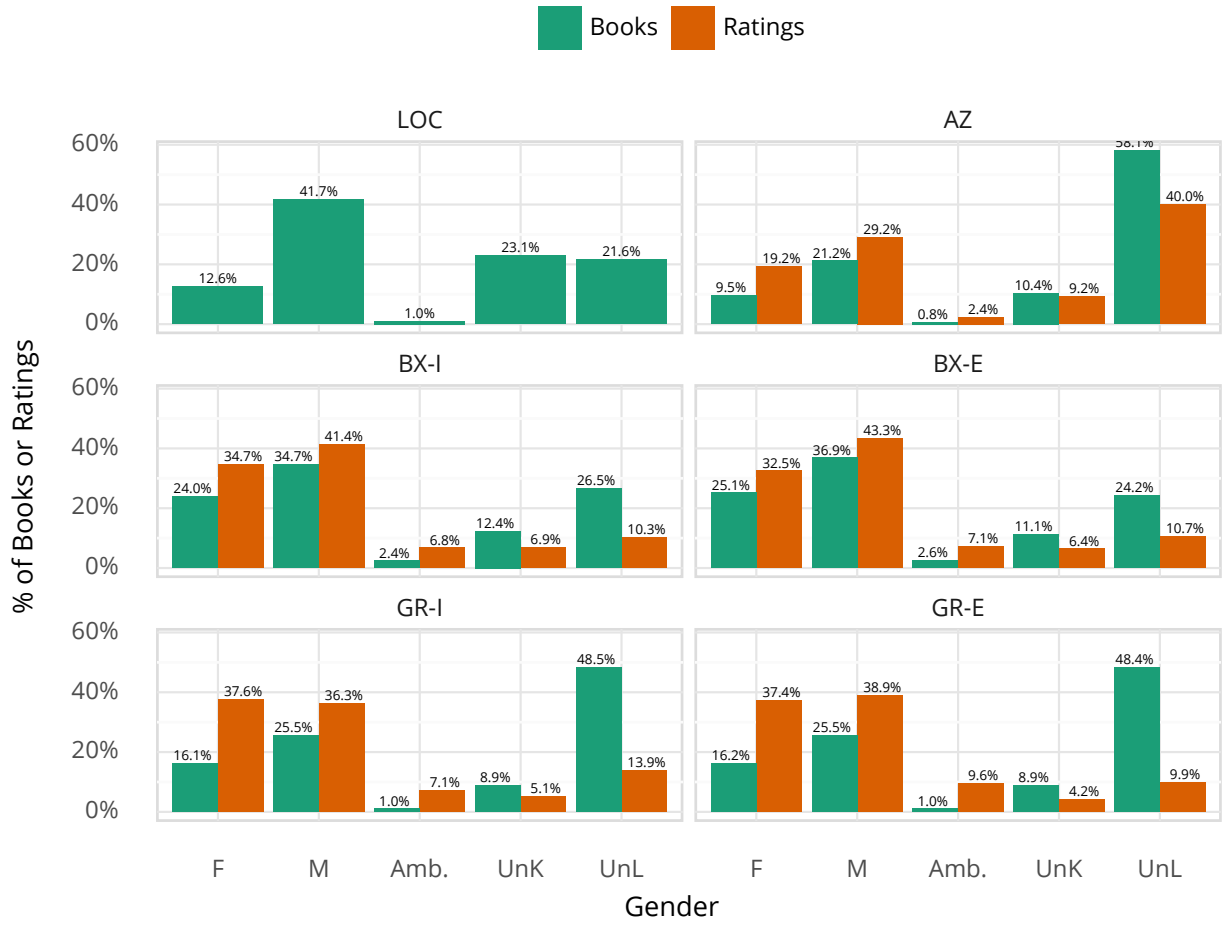

Figure 4: Results of data linking and gender resolution. LOC is the set of books with Library of Congress records; other panes are the results of linking rating data.

3.5 Data Set Statistics

Table 2 and Fig. 4 summarize the results of integrating these data sets. While the data is sparse, it has sufficient coverage for us to perform a meaningful analysis. We also report coverage of the Library of Congress data itself, as a rough approximation of books published irrespective of whether they are rated. Unfortunately, we do not know what biases lie in the coverage rates: are unlinked or unknown books more likely to be written by authors of one gender or another?

Consistent with 3, 4 shows that ratings are concentrated on books with known author genders; while almost 50% of GoodReads books are unlinked, less than 15% of interactions are with unlinked books.

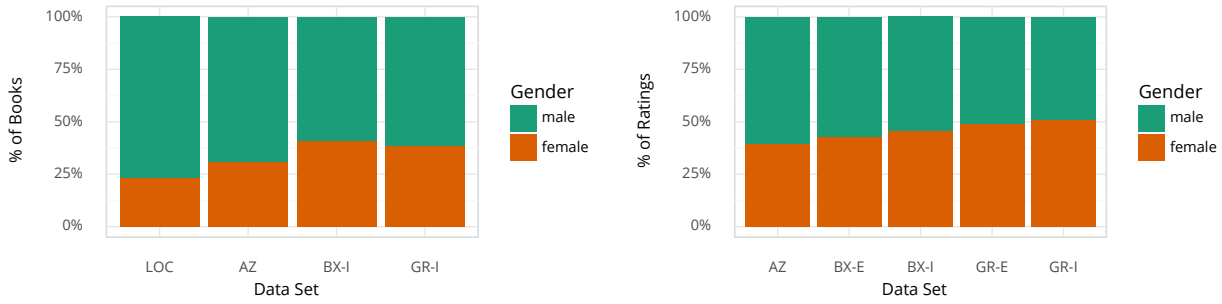

Figure 5: Distribution of known-gender books in each data set.

	Books		Ratings	
	female	male	female	male
DataSet				
LOC	23.2%	76.8%	—	—
AZ	31.0%	69.0%	39.7%	60.3%
BX-E	40.5%	59.5%	42.9%	57.1%
BX-I	40.9%	59.1%	45.6%	54.4%
GR-E	38.8%	61.2%	49.0%	51.0%
GR-I	38.7%	61.3%	50.9%	49.1%

Table 3: Distribution of known-gender books and ratings.

	Female		Male	
	mean	median	mean	median
Data Set				
AZ	20.01	4	13.64	3
BX-I	5.84	2	4.82	1
GR-I	402.35	34	245.28	20

Table 4: Average interactions-per-item by gender.

3.6 RQ1: Baseline Corpus Distribution

This analysis, and the distribution of genders show in in Figs. 4–5 and Table 3, provide our answer to RQ1. Of Library of Congress books with known author genders, 23.2% are written by women. Rating data sets have higher representation of women: 31.0% of books rated on Amazon are written by women, and 40.9% of BookCrossing books. Representation is higher yet when looking at ratings themselves: while 38.7% of known-gender books on GoodReads are written by women, 50.9% of shelf adds of known-gender books are for books by women. On average, books by female authors are interacted with more frequently than books by male authors (on GoodReads, the median interaction count per item is 20 for male-authored books and 34 for female-authored books; Table 4 shows details). As seen in Fig. 3, the most popular books are relatively evenly split between male and female authors in the book community sites (BookCrossing and GoodReads).

In general, we see the following progression in gender balance:

$$\text{Books(LOC)} < \text{Books(platform)} < \text{ratings}$$

Takeaway RQ1

If women are underrepresented in book publishing, they are less underrepresented in book rating data, particularly at the top end of the book popularity scale. The GoodReads community achieves close to gender parity in terms of books rated or added to shelves.

4 Experiment and Analysis Methods

Starting with the integrated book data, our main experiment has several steps:

1. Sample 5000 users, each of whom has rated at least 5 books with known author gender, for analysis.
2. Quantify gender distribution in sample user profiles (RQ2).
3. Produce 50 recommendations for each sample users, using the entire data set for training.
4. Compute recommendation list gender distribution (RQ3) and compare with user profile distribution (RQ4).

This experiment is completely reproducible with scripts available from the authors[5] combined the integrated book data described in Section 3. An end-to-end re-run, not including data integration or hyperparameter tuning, took 48.5 hours (elapsed; 676.5 CPU-hours compute) on a cluster node with two 14-core 2.2GHz Xeon Gold 5120 processors and 512GiB of memory, and produced approximately 200GiB of intermediate and output files.

[5] `https://md.ekstrandom.net/pubs/bag-extended`

4.1 Sampling

We sample 5000 users to keep the final data set tractable. Our statistical analysis methods are computationally intensive, scaling linearly in the number of users. Sampling users for assessing user profile makeup and gender propagation enables this analysis to be done in reasonable time; 5000 users is enough to ensure some statistical validity.

We require each user to have at least 5 books with known author gender so that their profile has enough books to estimate user gender balance, and so that the recommender has history with which to make recommendations.

4.2 Recommending Books

We used the LensKit toolkit [Ekstrand, 2020] to produce 50 recommendations for each of our 5000 sample users using the following algorithms:

- UU, a user-based collaborative filter [Herlocker et al., 1999]. In implicit-feedback mode, it sums user similarities instead of computing a weighted average.
- II, an item-based collaborative filter [Deshpande and Karypis, 2004]. As with UU, in implicit feedback mode, this algorithm sums item similarities instead of computing a weighted average.
- ALS, a matrix factorization model trained with alternating least squares [Pilászy et al., 2010]; we use both implicit and explicit feedback versions.
- BPR, a learning-to-rank algorithm that optimizes pairwise ranking [Rendle et al., 2009]; we use the BPR-MF version.

These algorithms are intended to provide a representative sample of common recommendation paradigms; while there are many different algorithms for doing recommendation, they typically optimize either point-wise recommendation accuracy (like ALS) or ranking loss with a cost function similar to that of BPR. We trained the collaborative filters over all available ratings, even those for books with unknown genders, and only restricted recommendation lists to exclude already-consumed books.

4.2.1 Tuning and Performance

While recommendation accuracy is not the focus of our experiment, we report it for context; it also provides a baseline for our exploration of distribution-constraining rerankers in Section 6. Figure 6 shows the MRR both on the evaluation set and on the tuning set with the best hyperparameters. Nearest-neighbor recommenders performed quite well on implicit-feedback data; we suspect this is partially due to popularity bias [Bellogin et al., 2011], as similarity-sum implicit-feedback k-NN will strongly favor popular items.

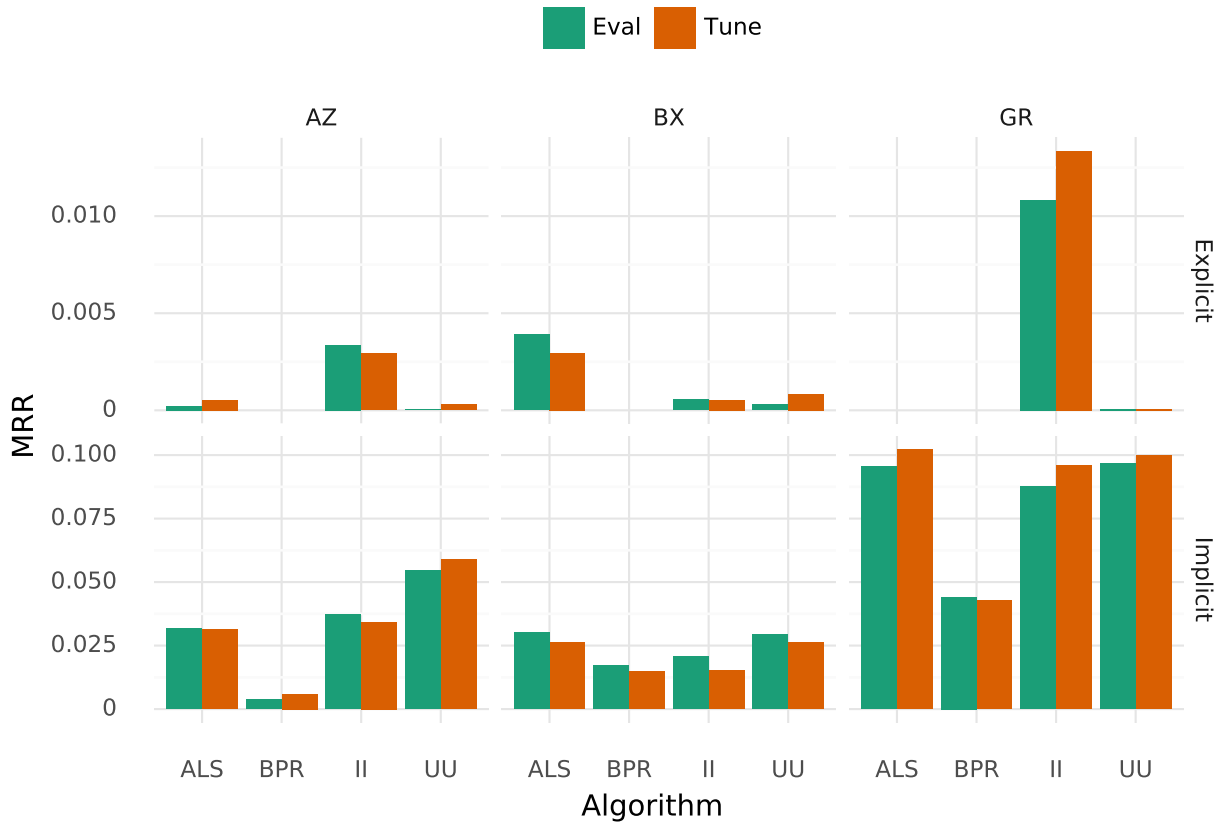

Figure 6: Top-N recommendation accuracy. *Eval* is the accuracy on the evaluation set, and *Tune* is the best accuracy on the tuning set during hyperparameter tuning.

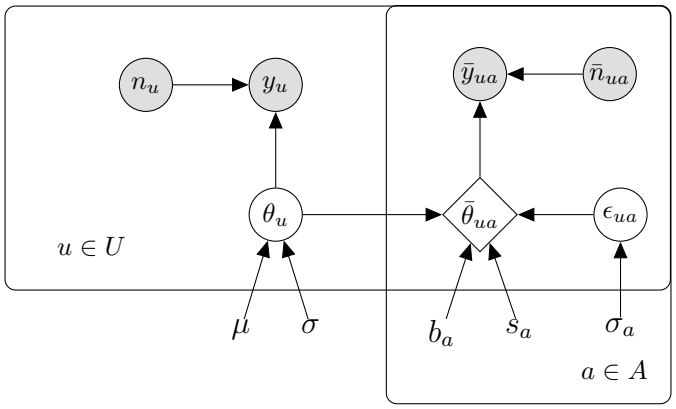

Figure 7: Plate diagram for statistical model.

We sampled 5000 users with at least 5 ratings for evaluation. For each user, we held out one rating as the test rating, generated a 100-item recommendation list, and measured the Mean Reciprocal Rank (MRR). We tuned each model's hyperparameters with scikit-optimize, optimizing MRR on a separate tuning set that was selected identically to the evaluation set.[6] We stopped tunning when the 5 best settings showed no more than 1% improvement in MRR. We exclude ALS on GR-E because it did not perform well after repeated tuning attempts. Implicit ALS worked well on GoodReads.

4.3 Statistical Analysis

Our statistical goal is to estimate the gender balance of user profiles, recommendation lists, and the propagation factor between them. There are several challenges that complicate doing this with commonly-used statistical techniques:

- Variance in user profile sizes makes it difficult to directly compare gender proportions between users (2 out of 5 and 20 out of 50 reflect very different levels of confidence).
- With many data sets and algorithm, we quickly run into large (and non-obvious) multiple comparison problems.
- We are interested in assessing distributions of bias, not just point estimates.

To address these difficulties, we model user rating behaviors using a hierarchical Bayesian model [Gelman et al., 2014] for the observed number of books by female authors out of the set of books with known authors. This model allows us to integrate information across users to estimate a user's tendency even when they have not rated very many books, and integrated Bayesian models enable us to robustly infer a number of parameters in a manner that clearly quantifies uncertainty and avoids many multiple-comparison problems [Gelman and Tuerlinckx, 2000]. We extend this

[6] To reduce the number of zeros, we tuned GoodReads using 1000-item lists instead of 100.

Table 5: Summary of key model parameters and variables.

Variable	Description
n_u	Number of known-gender books rated by user u
y_u	Number of female-authored books rated by u
θ_u	Probability of a known-author book rated by u being by a female author (smoothed user gender balance)
μ	Expected user gender balance, in log-odds ($\mathrm{E}[\mathrm{logit}(\theta_u)]$)
σ^2	Variance of user gender balance ($\mathrm{var}(\mathrm{logit}(\theta_u))$)
$\bar{n}_{ua}$	Number of known-gender books algorithm a recommended to user u
$\bar{y}_{ua}$	Number of female-authored books a recommended to u
$\bar{\theta}_{ua}$	Gender balance of algorithm a's recommendations for u
s_a	Regression slope of algorithm a (its responsiveness to user profile tendency)
b_a	Intercept of algorithm a (its baseline tendency)
σ_a^2	Residual variance of algorithm a (its variability unexplained by user tendencies)

to model recommendation list distributions as a linear function of user profile distributions plus random variance.

Figure 7 shows a plate diagram of this model, and Table 5 summarizes the key parameters; in the following sections we explain each of the components and parameters in more detail.

4.3.1 User Profiles

For each user, we observe n_u, the number of books they have rated with known author gender, and y_u, the number of female-authored books they have rated. From these observations, we estimate each user's author-gender tendency θ_u using a logit-normal model to address RQ2. The beta distribution is commonly used for modeling such tendencies, but the logit-normal has two key advantages: it is more parsimonious when extended with a regression, as we can compute regression coefficients in log-odds space, and it is substantially more computationally efficient to sample. In early versions of this experiment we also found that it fit our data slightly better.

We use the following joint probability as our likelihood model:

$$\begin{aligned} y_u &\sim \text{Binomial}(n_u, \theta_u) \\ \text{logit}(\theta_u) &\sim \text{Normal}(\mu, \sigma) \end{aligned}$$

$\text{logit}(\theta_u)$ is the log odds of a known-gender book rated by user u being written by a female author, and μ and σ are the mean and standard deviation of this user author-gender tendency. Negative values indicate a tendency towards male authors, and positive values a tendency towards female authors. θ_j is the corresponding probability or proportion in the range $[0, 1]$. When sampling from the fitted model, we produce a predicted θ', n', y', and observed ratio y'/n' for each sample in order to estimate the distribution of unseen user profiles.

We put vague priors on all parameters: $\sigma, \nu, \gamma \sim \text{Exponential}(0.1)$, as they are positive, and $\mu \sim \text{Normal}(0, 10)$. These priors provide diffuse density across a wide range of plausible and extreme values.[7]

4.3.2 Recommendation Lists

For RQ3 and RQ4, we model recommendation list gender distributions by extending our Bayesian model to predict recommendation distributions with a linear regression based on each user's smoothed proportion and per-algorithm slope, intercept, and variance. The regression is in log-odds (logit) space, and results in the following formula for estimating $\bar{\theta}_{ua}$:

[7] In early iterations of this work, we used broader priors; these vague priors are more in line with current STAN recommendations (see `https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations`), and do not affect inference conclusions.

$$\bar{y}_{ua} \sim \text{Binomial}(\bar{n}_{ua})$$
$$\text{logit}(\bar{\theta}_{ua}) = b_a + s_a \text{logit}(\theta_u) + \epsilon_{ua}$$
$$\epsilon_{ua} \sim \text{Normal}(0, \sigma_a)$$

The regression residual ϵ_{ua} captures variance in the relationship between users' and algorithms' recommendation proportions beyond that intrinsic in the use of a binomial distribution, and giving it per-algorithm variance allows for some algorithms being more consistent in their output than others. $\bar{n}_{ua}$ can differ between users and algorithms because the algorithms generate their recommendations without regard for author gender, and we remove unknown-gender books from the resulting lists for statistical analysis.

The result of our full model is that s_a captures how much an algorithm's output gender distribution varies *with* the input profile distribution, and σ_a^2 its variance *independent of* the input distribution. b_a expresses the algorithm's typical gender balance when the user's profile is evenly balanced (since the log of even odds is zero).

In the full model, the recommendation lists can affect the inferred parameters for user profiles, because the model is expressed as a factored joint probability distribution that includes all parameters. In practice, it is difficult to achieve separation, because we would need to either use point estimates for user profile tendencies in the recommendation list analysis (losing the rich information the first inference obtains about the *distribution* of profile bias, including the uncertainty in any particular user's tendency), or import the entire set of samples from the profile phase into the recommendation list phase (a process that is cost-prohibitive in current inference software).

4.3.3 Implementation

We fit and sample models with STAN [Carpenter et al., 2017], drawing 10,000 samples per model (4 NUTS chains each performing 1000 warmup and 2500 sampling iterations). We report results with the posterior predictive distributions of the parameters of interest, as estimated by the sampling process.

5 Profile and Propagation Results

In this section we present the results of our statistical analysis of user profiles and recommendations. We begin with characterizing the profiles of our sample users, and then proceed to analyze the resulting recommendations.

5.1 User Profile Characteristics

Under RQ2, we want to understand the distribution of users' author-gender tendencies, as represented by the proportion of known-gender books in each author's profile that are written by female

Table 6: Summary statistics for user profile gender distributions. μ is the posterior expected log odds of $P(\text{female}|\text{known})$; σ^2 is the posterior variance of that log odds; and θ' is the posterior expected proportion, or the mean $P(female|known)$ the model expects for new, unseen users.

	AZ	BX-E	BX-I	GR-E	GR-I
Obs. y/n	0.414	0.419	0.407	0.447	0.450
Std. Dev.	0.329	0.267	0.254	0.276	0.269
μ	-0.51	-0.40	-0.44	-0.27	-0.25
95%	(-0.58, -0.46)	(-0.44, -0.36)	(-0.48, -0.41)	(-0.32, -0.23)	(-0.30, -0.21)
σ	1.88	1.19	1.08	1.50	1.44
$\mathrm{E}[\theta']$ (post.)	0.42	0.42	0.41	0.45	0.45
Std. Dev.	0.30	0.23	0.21	0.27	0.26

authors. The histograms in Fig. 8 shows the distribution of observed author gender proportions, while Table 6 presents user profile summary statistics.

The Bayesian model from Section 4.3.1 provides more rigorous, smoothed estimates of this distribution. Table 6 describes the numerical results of this inference. The key parameters are μ, the average user's author-gender tendency in log-odds; σ, the standard deviation of user author-gender tendencies; and sampled θ values, the distribution of which describes the distribution of user author-gender tendencies expressed as expected proportions.

Figure 8 shows the densities of the author-gender tendency distribution, along with the densities of projected and actual observed proportions. The ripples in predicted and observed proportions are due to the commonality of 5-item user profiles, for which there are only 6 possible proportions; estimated tendency (θ) smooths them out. This smoothing, along with avoiding estimated extreme biases based on limited data, are why we find it useful to estimate tendency instead of directly computing statistics on observed proportions. The distribution of θ' — draws from the posterior distribution of a hypothetical new user — describes what the model has inferred about the distribution of user profile gender balances from the data it was provided. In the Amazon and BookCrossing data, we see high frequency of all-male and all-female profiles; as can be seen from the combination of smoothed tendency distribution and how it is reflected in predicted y/n distributions, this naturally arises from the right skew in the user tendency distribution combined with small profile sizes — an all-male profile is not just common in the data, but in the fitted model.

Comparing the observed and predicted y/n values in Fig. 8 provides a graphical assessment of model fit. The predicted values are samples of the observable gender balances that arise from θ' samples; under a well-fitting model, the distribution of these hypothetical users should be close to the distribution of observed users. To support direct comparison of the densities of observations and predictions, we resampled observed proportions with replacement to yield 10,000 observations. While there is some mild divergence in the observed and predicted distributions of high-female authors on the GoodReads data set, the models overall indicate good fit, and the means of smoothed, predicted, and observed proportions are all very close.

Table 7: Recommendation coverage and diversity statistics (implicit).

	AZ			BX			GR		
	Recs	Dist.	% Dist	Recs	Dist.	% Dist	Recs	Dist.	% Dist
ALS	250,000	17,757	7.1%	250,000	10,658	4.3%	250,000	16,382	6.6%
BPR	250,000	13,006	5.2%	250,000	42,161	16.9%	250,000	98,105	39.2%
II	249,949	120,791	48.3%	249,700	56,902	22.8%	250,000	25,506	10.2%
II	249,957	47,142	18.9%	248,439	17,978	7.2%	249,383	16,542	6.6%

Table 8: Recommendation coverage and diversity statistics (explicit).

	AZ			BX			GR		
	Recs	Dist.	% Dist	Recs	Dist.	% Dist	Recs	Dist.	% Dist
ALS	250,000	47,308	18.9%	250,000	65	0.0%	—	—	—
II	239,412	113,365	47.4%	248,316	18,588	7.5%	245,944	90,333	36.7%
UU	191,553	109,755	57.3%	219,082	43,475	19.8%	241,523	67,473	27.9%

Takeaway RQ2

We observe a population tendency to rate male authors more frequently than female authors in all data sets ($\mu < 0$), but to rate female authors more frequently than they would be rated were users drawing books uniformly at random from the available set (observed by comparing $\mathrm{E}[\theta']$ to each data set's fraction of female-authored books in Table 3). The average user author-gender tendency is slightly closer to an even balance than the set of rated books. We also found substantial variance between users about their estimated tendencies (s.d. of predicted θ exceeds 0.2; inferred $\sigma > 1$; both even-odds and book population proportions are within one s.d. of estimated means). This means that some users are estimated to strongly favor female authored books, even if these users are outnumbered by those that primarily read male-authored books.

5.2 Recommendation List Distributions

Our first step in understanding how collaborative filtering algorithms respond to this data bias is to examine the distribution of recommender list tendencies (RQ3). As described in 4.2, we produced 50 recommendations from each algorithm. Tables 7 and 8 show the basic coverage statistics of these algorithms. Users for which an algorithm could not produce recommendations are rare. We also computed the extent to which algorithms recommend different items to different users; "% Dist." is the percentage of all recommendations that were distinct items. Algorithms that repeatedly recommend the same items will be consistent in the gender distributions of their recommendations. ALS on BX-E did not personalize at all, so we omit it from analysis.

Table 9 provides the mean tendency for recommendation lists produced by each of our algorithms, plus the tendency of Most Popular and Highest Average Rating recommenders. These av-

Table 9: Mean / SD of rec. list female author proportions.

		AZ	BX	GR
Popular		0.472	0.405	0.424
Avg. Rating		0.291	0.205	0.125
Implicit	ALS	0.408 / 0.308	0.404 / 0.188	0.439 / 0.285
	BPR	0.407 / 0.284	0.424 / 0.274	0.440 / 0.316
	II	0.388 / 0.311	0.456 / 0.208	0.484 / 0.247
	UU	0.417 / 0.279	0.389 / 0.168	0.424 / 0.265
Explicit	ALS	0.405 / 0.150	0.301 / 0.012	—
	II	0.388 / 0.244	0.434 / 0.138	0.404 / 0.231
	UU	0.345 / 0.236	0.401 / 0.160	0.381 / 0.161

erages are in line with the user profile averages shown in Table 6.

Figures 9 and 10 show the density of recommendation list proportions, again showing the smoothed proportions with observed and predicted proportions for assessing model fit. The model fits quite well for explicit-feedback recommenders; some recommender and data set combinations on implicit-feedback, however, show significant effects that the model is not yet able to account for (as evidenced by the gaps between predicted and observed proportions). In particular, all algorithms on Amazon have curves not captured in the predicted distribution, and Item-Item on both BookCrossing and GoodReads exhibits a peak at about 0.35 that is not captured in the model. The result is that our model likely underestimates the extent to which these algorithms favor male-authored books. BPR on GoodReads favors both extreme-male and extreme-female distributions, as evidenced by the two peaks in its distribution. Identifying these effects and accounting for them in the model is left for future improvements of our experimental methodology; the quality of fit in these charts does affect our confidence in the inferences in the next section. The model predicts the implicit ALS algorithm's distribution relatively well, and the distribution shape is comparable to that of the input user profiles for each data set (compare with Fig. 8).

Explicit feedback algorithms in the majority of cases cases had highly concentrated distributions of smoothed balances, and low variance in observed balances. We discuss the differences between implicit and explicit response further in Sections 5.3 and 7.

Takeaway RQ3

Recommendation list average balances are comparable to user profile average balances, but otherwise there are notable differences in the *distribution* of balances. The Implicit ALS algorithm shows the most congruence between the distribution of recommendation list balances and user profile balances. BPR in particular has notable concentrations that decrease recommendation diversity with respect to user profile diversity, and reflect a pattern not yet captured in our model. Further research is needed to better understand what drives the distributions we observe and how to model the makeup of recommendation lists.

5.3 From Profiles to Recommendations

Our extended Bayesian model (Section 4.3.2) allows us to address RQ4: the extent to which our algorithms propagate individual users' tendencies into their recommendations (RQ4).

Figures 9–10 show the posterior predictive and observed densities of recommender author-gender tendencies, and Figures 11–12 show scatter plots of observed recommendation proportions against user profile proportions with regression curves (regression lines in log-odds space projected into probability space). Figure 13 shows the slope and intercept parameters with 95% credible intervals.

In implicit-feedback mode, most algorithms are quite responsive to user profile balances, with slopes greater than 0.5. The GoodReads data set seems to exhibit the best fit in Fig. 11, and shows the most direct reflection of user profiles into recommendation lists; it is also the densest, with users tending to have more ratings in their profiles, giving the recommender algorithms more to work with for producing accurate recommendations (see Fig. 6) and estimating users' profile tendencies. The ALS algorithm has regression parameters quite close to perfect propagation for all data sets, but especially GoodReads and Amazon (see Fig. 13). Explicit-feedback mode shows less responsiveness and stronger skews: all slopes are relatively small, and intercepts are negative (meaning a user with an evenly-balanced input profile will receive recommendations that have more men than women).

Takeaway RQ4

Implicit-feedback algorithms tend to reflect a user's profile gender balance in their recommendation lists. The strength and reliability of this propagation varies, but all data sets and implicit-feedback algorithms exhibit a clear linear trend. It is most pronounced in GoodReads, which has the most data for training; the implicit ALS algorithm is nearly a perfect line, and BPR amplifies user's tendencies towards female authors into their recommendation lists. Explicit-feedback algorithms are much less responsive to their users' input profiles, likely due to the fact that they rely on rating values, not the mere presence of a book.

6 Forced-Balance Recommendation

So far we have sought to measure, without intervention, the distribution of author genders of books recommended to users. This approach is quite reasonable given that neither past work, nor the analysis presented here, is sufficient to inform what recommendations *should* look like. Individual recommender systems professionals may, through other data, analysis, or philosophy, come to a conclusion about how they want their recommendation algorithms to behave.

In this section we address RQ5 with a suite of *forced-balance recommenders* that attempt to constrain the distribution in recommender output without substantially impacting recommendation quality. We consider very simple algorithms for understanding this tradeoff; the behavior of more sophisticated approaches such as calibration [Steck, 2018] or independence [Kamishima et al., 2018] are left for future work. As there is no general definition of "best tradeoff" between quality and gender distribution, nor clear consensus about exactly what to target in the first place, such

an analysis would be premature. Instead we seek to provide lower-limits to what can be expected from these type of tradeoffs with simple approaches. This analysis serves as a starting point for future explorations into recommender systems that deliberately pursue targeted changes in recommendation properties.

We consider three force-balance recommenders:

- single-pass force-balance (SingleEQ)
- multi-pass force-balance (GreedyEQ)
- multi-pass calibrate (GreedyReflect)

All three algorithms are implemented as a post-processor that can be applied to any base recommendation technique, much like Ziegler et al.'s topic diversification [Ziegler et al., 2005]. This means the primary input to these algorithms is an existing ranking of the item set. Often this input will be a list of items sorted by the prediction or ranking scores generated by a base algorithm. We operated the algorithms with a ranking over the entire item set as their input; for efficiency, truncated rankings could be used. All three algorithms start from the top of the input ranking and preserve it to varying degrees; they thus implicitly balance recommendation accuracy with gender representation by perturbing an accuracy-optimized ranking only insofar as adjustments are necessary to achieve their gender balance targets. Alongside the input ranking, all three force-balance algorithms also take the gender labels for each book and a target size for the list. The target size parameter allows for the common use of a recommendation algorithm in assembling a top-N list of fixed size.

The goal of the first two algorithms is to recommend approximately equal numbers of male- and female-authored books. In SingleEQ (Algorithm 1) this is accomplished in a single pass of the input recommendation list The algorithm is quite simple: for each item in the input base-algorithm ranking (in order) the algorithm either accepts the item (adding it to the output list of items) or reject it. Items are rejected if they would make the gender balance of the current output list further from our target. So while the current output list has more female authored books than male authored books it will reject female authored item recommendations[8]. Likewise, if the current output list has more male authored books, then it will reject additional male authored books. Note that books with unknown or unlinked gender are always recommended as they will have no effect on the known-gender gender balance of the generated recommendations. The algorithm proceeds in this manner, accepting and rejecting items from the base recommendation list, until the target recommendation size is reached.

Both GreedyEQ and GreedyReflect share the same general algorithm (Algorithm 2), and are structured more like a traditional greedy optimizer. The only difference is that GreedyEQ seeks a target balance of 0.5 while GreedyReflect targets the balance observed in each user's ratings.

The GreedyEQ algorithm proceeds iteratively, at each step selecting the next item to add to it's output list. Each step of the algorithm loops over the base recommendations selecting the top ranked item satisfying two constraints: 1) the item is not already in the output list, and 2) the

[8]This does not accommodate authors with non-binary gender identities. Our goal here is examine the behavior of simple mechanisms supported by available data.

Algorithm 1: Single-pass Equalize (SingleEQ)

Data: ranked list L, target length n, attribute $G : L \to m, f, \perp$
Result: ranked list L'
$L' \leftarrow$ empty list;
$n_f, n_m \leftarrow 0$;
for $i \in L$ **do**
  if $G(i) = \perp$ **then**
    add i to L';
  else if $G(i) = f \wedge n_f \leq n_m$ **then**
    add i to L';
    $n_f \leftarrow n_f + 1$;
  else if $G(i) = m \wedge n_m \leq n_f$ **then**
    add i to L';
    $n_m \leftarrow n_m + 1$;
  end
  if $|L'| \geq n$ **then**
    break;
  end
end

item would not lead to a worse gender imbalance. To determine if the item would lead to a worse imbalance we begin by estimating the current balance of the output list. If our current balance is more female heavy than our target balance we only add male-authored books. If our current balance is more male heavy than our target balance we only add female authored books. If our current balance is equal to our target balance we are willing to accept any book. As before, unkown and unlinked authors are recommended as they are reached by this algorithm.

This iterative process allows GreedyEQ to pick up items that were skipped in a past step, should the current gender balance of the output list allow them, leading to better recommendations. The cost for this improvement is taking many more passes over the item set, possibly increasing recommendation time, especially for large target recommendation sizes.

The third and final reranker, multi-pass calibrate (GreedyReflect), is based on Steck's concept of calibration [Steck, 2018]. Rather than targeting a gender balance of 0.5, it targets the balance observed in the user's ratings.

All three algorithms are designed to ensure that the output list will be at most one male- or female- authored book above (or below) the target gender balance, while being as close to the underlying ranking as possible. Due to the iterative nature of the algorithm, this will also hold true of every prefix of the output list, ensuring that the output list isn't separated into a clear "male half" and "female half" but instead has genders well-mixed throughout the list.

We repeated our evaluation from Section 4.2.1 with the reranking algorithms to measure their accuracy loss. Figure 14 shows the results of this experiment, and Table 10 shows the relative loss of balancing each algorithm for each data set. Most penalties are quite small at just a few percent;

Algorithm 2: Greedy Rebalance

Data: ranked list L, target length n, attribute $G : L \to m, f, \perp$, target balance p
Result: ranked list L'
$L' \leftarrow$ empty list;
$n_f, n_m \leftarrow 0$;
while $|L'| < n$ **do**
  $p' \leftarrow n_f/(n_f + n_m)$;
  for $i \in L \setminus L'$ **do**
    if $G(i) = \perp$ **then**
      add i to L';
      break;
    else if $G(i) = f \wedge p' \leq p$ **then**
      add i to L';
      $n_f \leftarrow n_f + 1$;
      break;
    else if $G(i) = m \wedge p' \geq p$ **then**
      add i to L';
      $n_m \leftarrow n_m + 1$;
      break;
    else // out of options, end early
      return L';
    end
  end
end

DataSet	Implicit	Algorithm	GreedyEQ	GreedyReflect	SingleEQ
AZ	False	ALS	3.23%	-0.57%	-5.60%
		II	3.65%	-0.01%	3.72%
		UU	-10.23%	0.85%	-10.23%
	True	ALS	8.11%	2.63%	13.09%
		BPR	6.18%	-0.98%	10.32%
		II	5.08%	1.34%	7.60%
		UU	4.65%	1.08%	8.69%
BX	False	ALS	0.82%	-2.24%	1.72%
		II	19.48%	-10.85%	35.42%
		UU	10.70%	-14.50%	-5.12%
	True	ALS	6.89%	3.59%	15.99%
		BPR	8.09%	3.08%	16.76%
		II	5.56%	2.46%	12.03%
		UU	4.24%	1.48%	9.66%
GR	False	II	7.40%	-0.50%	10.13%
		UU	25.01%	17.91%	33.56%
	True	ALS	4.75%	3.08%	11.36%
		BPR	7.08%	4.21%	13.52%
		II	3.23%	1.42%	6.65%
		UU	3.77%	2.17%	7.58%

Table 10: Accuracy loss for balancing genders.

the largest are (item-item on BX-E, user-user on GR-E) are on algorithms that do not perform well to begin with. In some cases the calibrated balancing even improves the recommender's accuracy slightly.

As expected, the multi-pass GreedyEQ algorithm generally outperforms SingleEQ. GreedyReflect, matching the user's profile balance instead of an arbitrary target of 0.5, usually performs the best.

Takeaway RQ5

We find, therefore, that it is possible to adjust the recommendation output balance with very simple approaches without substantial loss in accuracy. It also seems there is much room for more nuanced or refined adjustment. Again, we do not present these as particularly advanced approaches, but to establish an estimate of what should be possible. These results are consistent with those of Geyik and Kenthapadi [2018], where re-ranking techniques improved representation in job candidate search results without any harm to user engagement.

7 Discussion

We have observed the distribution of book author genders across the book recommendation pipeline (Fig. 1). Encouragingly for our societal goal of ensuring good representation in book authorship, representation of women seems to be higher in later stages of the pipeline: women write a greater share of rated books than cataloged books, and their books have more user interactions on average.

There is substantial variance between users in the gender balance of their historical book interactions, but on average, their profiles have better female author representation than the underlying book corpus does.

These author tendencies are then reflected into recommendations, particularly by implicit-feedback recommenders. Implicit-feedback recommendations were more reflective both of the overall distribution of user profile tendencies and each individual user's gender balance than explicit-feedback recommenders; this is likely because the explicit-feedback recommendations are primarily driven by the ratings that users give to books, rather than the presence of the book in the user's recommendation list. It is not surprising that the composition of a user's profile has a greater impact on algorithms that use the profile composition than it does on algorithms that use associated rating values, but it is useful to empirically document this difference in effect because it is difficult to predict *a priori* how algorithms will interact with particular socially-salient features of their input data that affect either its presence or its value. Perhaps in the future the social structure of recommendation data and consumption patterns will be sufficiently well-understood to make such predictions, but the current state of the art does not support them.

Recommender propagation of user profile balance seems to be both a blessing and a curse. On the one hand, it is encouraging that the algorithms are capturing and reflecting patterns in users' book consumption, whether those patterns are an actual gender preference or another preference that corresponds with author gender. Further, if a user wants to read books by underrepresented authors, and has found a number to put in their profile, a well-tuned collaborative filter may help them find more (although we need to empirically study recommender response to other axes of under-representation; we cannot assume that gender results will apply to e.g. ethnicity). On the other hand, if a user is reading predominantly majority authors, the collaborative filter will probably reinforce that tendency as well.

It is not yet clear what to *do* about this. The methods and results we have presented here are focused on describing what recommender system inputs and outputs look like, but we are also interested in how to deploy information access technologies to further social objectives. In addition to the roles Abebe et al. [2020] identify for computing in promoting social change, we think information access is a domain in which computing can be applied to directly catalyze positive social outcomes, particularly by promoting the work of content creators who have historically been overlooked. We have tested a few simple techniques for forcing particular representational goals, and found that they have little negative impact on recommendation accuracy, but whether and how to deploy such techniques is very much an open question.

In candidate sourcing and recruiting as a part of the hiring pipeline [Geyik and Kenthapadi, 2018], it seems clearly appropriate to deploy interventions to ensure representative search results. At least in the U.S. context, anti-discrimination law means that a recruiting platform's users are

already legally required to ensure some forms of representativeness. In other settings, however, it is less clear. Overriding the system's modeling of user preference to achieve the system designer's social goals may violate what agency users retain in their use of the recommender system [Ekstrand and Willemsen, 2016]. Leaving the system to propagate what patterns it will, however, may perpetuate inequities and deny content creators equal access to the creative marketplace (e.g. the goals outlined by Mehrotra et al. [2018]). The space of available interventions is not limited to either inaction or modifying the primary recommender, however; in some platforms, it may be feasible to deploy social nudges through additional recommendation experiences, such as adding a "New Authors You Might Love" feature that selects books for, among other things, author voices that are underrepresented in the corpus as a whole or in the individual user's historical activity.

Regardless of the appropriate solution, it is important to first understand what a system's data and behavior currently look like. We have presented results, reusable experimental methods, and a new composite data set for conducting such measurements of recommender systems. The next steps will be an ongoing discussion in the community of researchers and practitioners.

7.1 Limitations of Data and Methods

Our data and approach has a number of limitations that are important to note. First, book rating data is extremely sparse, and the BookCrossing data set is small, providing a limited picture of users' reading histories and reducing the performance of some algorithms. In particular, the high sparsity of the data set caused the MF algorithm to perform particularly poorly on offline accuracy metrics, so these findings may not be representative of its behavior in the wild; future work will need to test them across a range of recommender effectiveness levels and stages of system cold-start.

Second, our data and statistical methods only account for binary gender identities. While the MARC21 Authority Format supports flexible gender identity records (including multiple possibly-overlapping identities over the course of an author's life and nonbinary identities from an open vocabulary), VIAF does not seem to use this flexibility.

Third, we test a limited set of collaborative filtering algorithms. While we have chosen algorithms with an eye for diverse behaviors and global popularity, we must acknowledge that our selection of 4 algorithms is small in the face of algorithm diversity in the field. While our ultimate goal is to understand general trends, we acknowledge that our study does not evaluate enough algorithms to make claims about the entire field.

We consider it valuable to make forward progress in understanding the interaction of information systems with social concerns using the data we have available, even if that data has significant known weaknesses. We must, however, be reflective and forthright about the limitations of the data, methods, and resulting findings, and seek to improve them in order to develop a better understanding of the human impact of computing systems. Our experimental design can be readily extended to accommodate richer or higher-quality data sources and additional algorithms, and the code we provide for our experiments will facilitate such improvements. We have tested this reproducibility by re-running the experiments in the course of writing and revising this paper. Ultimately we see this as the first step in untangling a broader issue; we are actively exploring many extensions and improvements to this work.

7.2 Limitations of Current Results

Beyond the general limitations of our data and methods, there is much that our results here have left unexplored. We have only looked at uncontrolled distributions and correlations of author genders; we have not looked at any subdivisions, such as book genres. Author gender distributions may differ between genres or topics, and some of the effects we observe may be the result of user preferences for genres, topics, or other characteristics that happen to correlate with author gender for various reasons.

We believe observational, correlational studies such as the one we have presented have significant value in identifying the presence of potential effects. They are insufficient to establish causality, and they do not tell us *why* the effects are happening, but they provide insight into where to go looking to find the causal drivers of human and algorithmic behavior.

We hope that future work will uncover the factors that drive the relationships we have observed and yield deeper insight into both user behavior and the patterns that recommender systems can capture and reflect. We plan, of course, to carry out some of that work ourselves, but there is a great deal of space to explore. One particularly important next step is to adapt fairness constructs based on exposure [Diaz et al., 2020] or attention [Biega et al., 2018] to this problem setting; these account for rank position in addition to presence in a recommendation list, connect fair exposure to relevance, and are also more amenable to assessing fairness with respect to non-binary author attributes [Raj et al., 2020].

8 Conclusion and The Road Ahead

We have conducted an initial inquiry into the response of collaborative filtering book recommenders to gender distributions in the user preference data on which they are trained. Collaborative filtering algorithms trained on binary user-book interactions ("implicit feedback") tended to reflect the historical gender balance of users' reading patterns into their recommendations.

This paper is a first step in a much larger project to understand the ways in which recommendation algorithms interact with potentially discriminatory biases, and general behavior of recommendation technology with respect to various social issues. There are many future steps we see for advancing this agenda:

- Obtaining higher-quality data for measuring distributions of interest in recommender inputs and outputs. This includes obtaining data on non-binary gender identities and adopting statistical methods that can account for them.
- Examining other content creator features, such as ethnicity, in recommendation applications.
- Extending to additional algorithm families, such as content-based filters.
- Studying other domains and applications, such as movies, research literature, and social media.

- Develop more advanced algorithms that interact with various user or item characteristics of social concern; these could be developed to reflect organizational or societal goals or to help users further their individual goals [Ekstrand and Willemsen, 2016].
- Study the effect of existing refinements, such as diversification [Willemsen et al., 2016, Ziegler et al., 2005], on recommendation distributions.

We hope to see more work in the coming years to better understand ways in which recommender systems respond to and influence their sociotechnical contexts.

Acknowledgements

We thank Mucun Tian, Mohammed R. Imran Kazi, and Hoda Mehrpouyan for their contributions to the conference paper on which this work builds, and the People and Information Research Team (PIReT) for their support and feedback to help refine this research agenda. Computation performed on the R2 cluster [Boise State Research Computing Department, 2017].

References

R. Abebe, S. Barocas, J. Kleinberg, K. Levy, M. Raghavan, and D. G. Robinson. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 252–260, New York, NY, USA, Jan. 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372871. URL https://doi.org/10.1145/3351095.3372871.

G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005. ISSN 1041-4347. doi: 10.1109/TKDE.2005.99. URL http://dx.doi.org/10.1109/TKDE.2005.99.

M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, and A. Rieke. Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW):1–30, Nov. 2019. doi: 10.1145/3359301. URL https://doi.org/10.1145/3359301.

A. Bellogin, P. Castells, and I. Cantador. Precision-oriented evaluation of recommender systems: An algorithmic comparison. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, page 333–336, New York, NY, USA, 2011. ACM. ISBN 9781450306836. doi: 10.1145/2043932.2043996. URL http://doi.acm.org/10.1145/2043932.2043996.

A. Beutel, E. H. Chi, C. Goodrow, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, and L. Hong. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM

Press, 2019. ISBN 9781450362016. doi: 10.1145/3292500.3330745. URL http://dl.acm.org/citation.cfm?doid=3292500.3330745.

A. J. Biega, K. P. Gummadi, and G. Weikum. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 405–414. ACM, June 2018. ISBN 9781450356572. doi: 10.1145/3209978.3210063. URL https://dl.acm.org/citation.cfm?doid=3209978.3210063.

A. Billey, M. Haugen, J. Hostage, N. Sack, and A. L. Schiff. Report of the PCC ad hoc task group on gender in name authority records. Technical report, Program for Cooperative Cataloging, Oct. 2016. URL https://www.loc.gov/aba/pcc/documents/Gender_375%20field_RecommendationReport.pdf.

Boise State Research Computing Department. R2: Dell HPC intel e5v4 (high performance computing cluster), 2017. URL http://dx.doi.org/10.18122/B2S41H.

T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee and M. Sugiyama and U. V. Luxburg and I. Guyon and R. Garnett, editor, *Advances in Neural Information Processing Systems 29 (NIPS 2016)*. Curran Associates, Inc., July 2016. URL http://papers.nips.cc/paper/6227-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddi

D. Bucur. Gender homophily in online book networks. *Information sciences*, 481:229–243, May 2019. ISSN 0020-0255. doi: 10.1016/j.ins.2019.01.003. URL http://www.sciencedirect.com/science/article/pii/S0020025519300040.

J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, page 77–91. PMLR, 2018. URL http://proceedings.mlr.press/v81/buolamwini18a.html.

R. Burke. Multisided fairness for recommendation. *coRR*, July 2017. URL http://arxiv.org/abs/1707.00093.

R. Burke, N. Sonboli, and A. Ordonez-Gauger. Balanced neighborhoods for multi-sided fairness in recommendation. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 202–214, New York, NY, USA, 2018. PMLR. URL http://proceedings.mlr.press/v81/burke18a.html.

B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32, 2017. ISSN 1548-7660. doi: 10.18637/jss.v076.i01. URL https://www.jstatsoft.org/v076/i01.

O. Celma. *Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer, Berlin, Heidelberg, 2010. ISBN 9783642132865. doi: 10.1007/978-3-642-13287-2. URL https://link.springer.com/book/10.1007%2F978-3-642-13287-2.

S. Channamsetty and M. D. Ekstrand. Recommender response to diversity and popularity bias in user profiles. In *Proceedings of the 30th Florida Artificial Intelligence Research Society Conference*. AAAI Press, May 2017. URL https://aaai.org/ocs/index.php/FLAIRS/FLAIRS17/paper/view/15524/15019.

D. Cosley, D. Frankowski, L. Terveen, and J. Riedl. SuggestBot: Using intelligent task routing to help people find work in wikipedia. In *Proceedings of the 12th International Conference on Intelligent User Interfaces*, IUI '07, pages 32–41, New York, NY, USA, Jan. 2007. Association for Computing Machinery. ISBN 9781595934819. doi: 10.1145/1216295.1216309. URL https://doi.org/10.1145/1216295.1216309.

M. Deshpande and G. Karypis. Item-based Top-N recommendation algorithms. *ACM Transactions on Information Systems*, 22(1):143–177, Jan. 2004. ISSN 1094-9224. doi: 10.1145/963770.963776. URL https://doi.org/10.1145/963770.963776.

F. Diaz, B. Mitra, M. D. Ekstrand, A. J. Biega, and B. Carterette. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. ACM, Oct. 2020. doi: 10.1145/3340531.3411962. URL http://arxiv.org/abs/2004.13157.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. ACM. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL http://doi.acm.org/10.1145/2090236.2090255.

M. Ekstrand, J. Riedl, and J. A. Konstan. Collaborative filtering recommender systems. *Foundations and Trends® in Human-Computer Interaction*, 4(2):81–173, 2010. ISSN 1551-3955. doi: 10.1561/1100000009. URL http://dx.doi.org/10.1561/1100000009.

M. D. Ekstrand. LensKit for Python: Next-Generation software for recommender system experiments. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, 2020. doi: 10.1145/3340531.3412778. URL http://dx.doi.org/10.1145/3340531.3412778.

M. D. Ekstrand and J. A. Konstan. Recommender systems notation. Technical Report 177, Boise State University, 2019. URL https://scholarworks.boisestate.edu/cs_facpubs/177/.

M. D. Ekstrand and M. C. Willemsen. Behaviorism is not enough: Better recommendations through listening to users. In *Proceedings of the 10th ACM Conference on Recommender Systems*,

RecSys '16, page 221–224, New York, NY, USA, 2016. ACM. ISBN 9781450340359. doi: 10.1145/2959100.2959179. URL http://doi.acm.org/10.1145/2959100.2959179.

M. D. Ekstrand, M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, and M. S. Pera. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In S. A. Friedler and C. Wilson, editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency (PMLR)*, volume 81 of *Proceedings of Machine Learning Research*, pages 172–186, New York, NY, USA, Feb. 2018. PMLR. URL http://proceedings.mlr.press/v81/ekstrand18b.html.

D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian. Runaway feedback loops in predictive policing. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 160–171, New York, NY, USA, 2018. PMLR. URL http://proceedings.mlr.press/v81/ensign18a.html.

A. Epps-Darling, R. T. Bouyer, and H. Cramer. Artist gender representation in music streaming. In *Proceedings of the 21st International Society for Music Information Retrieval Conference*, page 248–254. ISMIR, Oct. 2020. URL https://program.ismir2020.net/poster_2-11.html.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, Aug. 2015. ISBN 9781450336642. doi: 10.1145/2783258.2783311. URL http://dl.acm.org/citation.cfm?doid=2783258.2783311.

S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im)possibility of fairness. *arXiv:1609.07236 [cs, stat]*, Sept. 2016. URL http://arxiv.org/abs/1609.07236.

B. Friedman and H. Nissenbaum. Bias in computer systems. *ACM Transactions on Information and System Security*, 14(3):330–347, July 1996. ISSN 1094-9224, 1046-8188. doi: 10.1145/230538.230561. URL http://doi.acm.org/10.1145/230538.230561.

A. Gelman and F. Tuerlinckx. Type S error rates for classical and bayesian single and multiple comparison procedures. *Computational Statistics*, 15(3):373–390, 2000. ISSN 0943-4062. doi: 10.1007/s001800000040. URL https://link.springer.com/article/10.1007/s001800000040.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. Hierarchical models. In *Bayesian Data Analysis*, page 101–138. CRC Press, 3rd edition, 2014. ISBN 9781439840955.

S. C. Geyik and K. Kenthapadi. Building representative talent search at LinkedIn. https://engineering.linkedin.com/blog/2018/10/building-representative-talent-search-at-linkedin, Oct. 2018. URL https://engineering.linkedin.com/blog/2018/10/building-representative-talent-search-at-linkedin. Accessed: 2018-10-25.

A. Gunawardana and G. Shani. Evaluating recommender systems. In *Recommender Systems Handbook*, pages 265–308. Springer, Boston, MA, 2015. ISBN 9781489976369, 9781489976376. doi: 10.1007/978-1-4899-7637-6_8. URL https://link.springer.com/chapter/10.1007/978-1-4899-7637-6_8.

F. Hamidi, M. K. Scheuerman, and S. M. Branham. Gender recognition or gender reductionism?: The social implications of embedded gender recognition systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 8. ACM, Apr. 2018. ISBN 9781450356206. doi: 10.1145/3173574.3173582. URL http://dl.acm.org/ft_gateway.cfm?id=3173582&type=pdf.

A. Hannak, C. Wagner, D. Garcia, M. Strohmaier, and C. Wilson. Bias in online freelance marketplaces: Evidence from TaskRabbit. In *Proceedings of the Workshop on Data and Algorithm Transparency*, 2016. URL http://datworkshop.org/papers/dat16-final22.pdf.

F. M. Harper and J. A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19:1–19:19, Dec. 2015. ISSN 2160-6455. doi: 10.1145/2827872. URL http://doi.acm.org/10.1145/2827872.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. Del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2649-2. URL http://dx.doi.org/10.1038/s41586-020-2649-2.

J. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 230–237. ACM, 1999. doi: 10.1145/312624.312682. URL http://portal.acm.org/citation.cfm?id=312682&dl=GUIDE&coll=GUIDE.

J. Herlocker, J. A. Konstan, L. Terveen, and J. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53, 2004. ISSN 1094-9224. doi: 10.1145/963770.963772. URL http://portal.acm.org/citation.cfm?id=963772.

A. L. Hoffmann. Data violence and how bad engineering choices can damage society. https://medium.com/s/story/data-violence-and-how-bad-engineering-choices-can-damage-society-39e44150e1d4, Apr. 2018. URL https://medium.com/s/story/data-violence-and-how-bad-engineering-choice Accessed: 2018-5-1.

K. Hosanagar, D. Fleder, D. Lee, and A. Buja. Will the global village fracture into tribes? recommender systems and their effects on consumer fragmentation. *Management Science*, 60(4):805–823, Nov. 2013. ISSN 0025-1909. doi: 10.1287/mnsc.2013.1808. URL https://doi.org/10.1287/mnsc.2013.1808.

J. C. Hu. The overwhelming gender bias in 'New York Times' book reviews. https://psmag.com/social-justice/gender-bias-in-book-reviews, Aug. 2017. URL https://psmag.com/social-justice/gender-bias-in-book-reviews. Accessed: 2020-5-12.

N. Hurley and M. Zhang. Novelty and diversity in Top-N recommendation – analysis and evaluation. *ACM Transactions on Internet Technology*, 10(4):14:1–14:30, Mar. 2011. ISSN 1533-5399. doi: 10.1145/1944339.1944341. URL http://doi.acm.org/10.1145/1944339.1944341.

J. Hutson, J. Taft, S. Barocas, and K. Levy. Debiasing desire: Addressing bias and discrimination on intimate platforms. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 18, Sept. 2018. doi: 10.1145/3274342. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3244459.

D. Jannach, L. Lerche, I. Kamehkhosh, and M. Jugovac. What recommenders recommend: An analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction*, 25(5):427–491, July 2015. ISSN 0924-1868, 1573-1391. doi: 10.1007/s11257-015-9165-3. URL http://link.springer.com/article/10.1007/s11257-015-9165-3.

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Recommendation independence. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 187–201, New York, NY, USA, 2018. PMLR. URL http://proceedings.mlr.press/v81/kamishima18a.html.

H. Kibirige, G. Lamp, J. Katins, A. O., gdowding, T. Funnell, matthias-k, J. Arnfred, F. Finkernagel, D. Blanchard, E. Chiang, S. Astanin, P. N. Kishimoto, stonebig, E. Sheehan, R. Gibboni, B. Willers, Pavel, Y. Halchenko, smutch, zachcp, J. Collins, R. K. Min, B. King, D. Brian, D. Arora, D. Brown, D. Becker, B. Koopman, and Anthony. has2k1/plotnine: v0.6.0, Aug. 2019. URL https://zenodo.org/record/3373970.

B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4):441–504, Oct. 2012. ISSN 0924-1868. doi: 10.1007/s11257-011-9118-4. URL https://doi.org/10.1007/s11257-011-9118-4.

R. Kuprieiev, D. Petrov, R. Valles, P. Redzyński, C. da Costa-Luis, A. Schepanovski, I. Shcheklein, S. Pachhai, J. Orpinel, F. Santos, A. Sharma, Zhanibek, D. Hodovic, Earl, A. Grigorev, N. Dash, G. Vyshnya, maykulkarni, Vera, M. Hora, xliiv, P. Rowlands, W. Baranowski, S. Mangal, and C. Wolff. DVC: Data version control - git for data & models, May 2020. URL https://zenodo.org/record/3813759.

N. Lathia, S. Hailes, L. Capra, and X. Amatriain. Temporal diversity in recommender systems. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 210–217. ACM, 2010. ISBN 9781450301534. doi: 10.1145/1835449.1835486. URL http://portal.acm.org/citation.cfm?id=1835486.

Library of Congress. MARC21 standards. Technical report, 1999. URL https://www.loc.gov/marc/.

K. Lum and W. Isaac. To predict and serve? *Significance*, 13(5):14–19, Oct. 2016. ISSN 1740-9713. doi: 10.1111/j.1740-9713.2016.00960.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2016.00960.x/abstract.

G. Magno, C. S. Araújo, W. Meira, Jr., and V. Almeida. Stereotypes in search engine results: Understanding the role of local and global factors. In *Proceedings of the Workshop on Data and Algorithm Transparency*, Sept. 2016. URL http://arxiv.org/abs/1609.05413.

J. McAuley, C. Targett, Q. Shi, and A. van den Hengel. Image-Based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 43–52, New York, NY, USA, 2015. ACM. ISBN 9781450336215. doi: 10.1145/2766462.2767755. URL http://doi.acm.org/10.1145/2766462.2767755.

W. McKinney and Others. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56, 2010. URL http://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf.

R. Mehrotra, J. McInerney, H. Bouchard, M. Lalmas, and F. Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pages 2243–2251. ACM, Oct. 2018. ISBN 9781450360142. doi: 10.1145/3269206.3272027. URL http://dl.acm.org/ft_gateway.cfm?id=3272027&type=pdf.

A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. Understanding the demographics of twitter users. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011. URL https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2816.

T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan. Exploring the filter bubble: The effect of using recommender systems on content diversity. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, page 677–686, New York, NY, USA, 2014. ACM. ISBN 9781450327442. doi: 10.1145/2566486.2568012. URL http://doi.acm.org/10.1145/2566486.2568012.

V. Pajović and K. Vyskocil. 2015 CWILA count methods and results. https://cwila.com/2015-cwila-count-methods-results/, Oct. 2016. URL https://cwila.com/2015-cwila-count-methods-results/. Accessed: 2018-5-7.

E. Pariser. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin, May 2011. ISBN 9781101515129.

I. Pilászy, D. Zibriczky, and D. Tikk. Fast ALS-based matrix factorization for explicit and implicit feedback datasets. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, page 71–78, New York, NY, USA, 2010. ACM. ISBN 9781605589060. doi: 10.1145/1864708.1864726. URL http://doi.acm.org/10.1145/1864708.1864726.

A. Raj, C. Wood, A. Montoly, and M. D. Ekstrand. Comparing fair ranking metrics. *coRR*, Sept. 2020. URL http://arxiv.org/abs/2009.01311.

J. Reback, W. McKinney, jbrockmendel, J. Van den Bossche, T. Augspurger, P. Cloud, gfyoung, Sinhrks, A. Klein, M. Roeschke, S. Hawkins, J. Tratner, C. She, W. Ayd, T. Petersen, M. Garcia, J. Schendel, A. Hayden, MomIsBestFriend, V. Jancauskas, P. Battiston, S. Seabold, chris-b, h-vetinari, S. Hoyer, W. Overmeire, alimcmaster, K. Dong, C. Whelan, and M. Mehyar. pandas-dev/pandas: Pandas 1.0.3, Mar. 2020. URL https://zenodo.org/record/3715232.

S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 452–461, Arlington, Virginia, United States, 2009. AUAI Press. ISBN 9780974903958. URL http://dl.acm.org/citation.cfm?id=1795114.1795167.

P. Resnick. Beyond bowling together: Sociotechnical capital. *HCI in the New Millennium*, 77:247–272, 2001. URL https://mccti.hugoramos.eu/Redes_Sociais_Online/TEXTOS_AULAS/TEXTO_AULA_07_Beyond%20Bowling%20Together%20SocioTechnical%20Capital_Resnick.pdf.

C. Riederer and A. Chaintreau. The price of fairness in location based advertising. *Fairness, Accountability and Transparency in Recommender Systems*, Aug. 2017. URL http://scholarworks.boisestate.edu/fatrec/2017/1/5.

A. Rosenblat and L. Stark. Algorithmic labor and information asymmetries: A case study of uber's drivers. *International Journal of Communication*, 10(0):27, July 2016. ISSN 1074-5351. URL http://ijoc.org/index.php/ijoc/article/view/4892/1739.

P. Sapiezynski, W. Zeng, R. E Robertson, A. Mislove, and C. Wilson. Quantifying the impact of user attentionon fair group representation in ranked lists. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pages 553–562, New York, NY, USA, May 2019. Association for Computing Machinery. ISBN 9781450366755. doi: 10.1145/3308560.3317595. URL https://doi.org/10.1145/3308560.3317595.

D. Shakespeare, L. Porcaro, E. Gómez, and C. Castillo. Exploring artist gender bias in music recommendation. *coRR*, Sept. 2020. URL http://arxiv.org/abs/2009.01715.

A. Singh and T. Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 2219–2228, New York, NY, USA, 2018. ACM. ISBN 9781450355520. doi: 10.1145/3219819.3220088. URL http://doi.acm.org/10.1145/3219819.3220088.

T. Spalding. Introducing thingISBN. https://blog.librarything.com/thingology/2006/06/introducing-thingisbn/, June 2006. URL https://blog.librarything.com/thingology/2006/06/introducing-thingisbn/.

A. Starke, M. Willemsen, and C. Snijders. Effective user interface designs to increase energy-efficient behavior in a rasch-based energy recommender system. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17, pages 65–73, New York, NY, USA, Aug. 2017. Association for Computing Machinery. ISBN 9781450346528. doi: 10.1145/3109859.3109902. URL https://doi.org/10.1145/3109859.3109902.

H. Steck. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 154–162. ACM, Sept. 2018. ISBN 9781450359016. doi: 10.1145/3240323.3240372. URL https://dl.acm.org/citation.cfm?doid=3240323.3240372.

J. Thebault-Spieker, B. Hecht, and L. Terveen. Geographic biases are 'born, not made': Exploring contributors' spatiotemporal behavior in OpenStreetMap. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork*, pages 71–82. ACM, Jan. 2018. ISBN 9781450355629. doi: 10.1145/3148330.3148350. URL https://dl.acm.org/citation.cfm?doid=3148330.3148350.

M. Thelwall. Reader and author gender and genre in GoodReads. *Journal of Librarianship and Information Science*, 51(2):403–430, June 2019. ISSN 0961-0006. doi: 10.1177/0961000617709061. URL https://doi.org/10.1177/0961000617709061.

M. van Alstyne and E. Brynjolfsson. Global village or Cyber-Balkans? modeling and measuring the integration of electronic communities. *Management Science*, 51(6):851–868, June 2005. ISSN 0025-1909. doi: 10.1287/mnsc.1050.0363. URL http://mansci.journal.informs.org/cgi/content/abstract/51/6/851.

S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, page 109–116, New York, NY, USA, 2011. ACM. ISBN 9781450306836. doi: 10.1145/2043932.2043955. URL http://doi.acm.org/10.1145/2043932.2043955.

VIDA. The 2016 VIDA count | VIDA: Women in literary arts. http://www.vidaweb.org/the-2016-vida-count/, Oct. 2017. URL http://www.vidaweb.org/the-2016-vida-count/. Accessed: 2018-5-7.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, Mar. 2020. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-019-0686-2. URL http://dx.doi.org/10.1038/s41592-019-0686-2.

M. Wan and J. McAuley. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 86–94. ACM, Sept. 2018. ISBN 9781450359016. doi: 10.1145/3240323.3240369. URL https://dl.acm.org/citation.cfm?doid=3240323.3240369.

M. C. Willemsen, M. P. Graus, and B. P. Knijnenburg. Understanding the role of latent feature diversification on choice difficulty and satisfaction. *User Modeling and User-Adapted Interaction*, 26 (4):347–389, Oct. 2016. ISSN 0924-1868, 1573-1391. doi: 10.1007/s11257-016-9178-6. URL https://link.springer.com/article/10.1007/s11257-016-9178-6.

K. Yang and J. Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, number Article 22 in SSDBM '17, pages 1–6, New York, NY, USA, June 2017. Association for Computing Machinery. ISBN 9781450352826. doi: 10.1145/3085504.3085526. URL https://doi.org/10.1145/3085504.3085526.

S. Yao and B. Huang. Beyond parity: Fairness objectives for collaborative filtering. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2925–2934. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6885-beyond-parity-fairness-objectives-for-collaborative-filtering.pdf.

M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. FA*IR: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 1569–1578. ACM, Nov. 2017. ISBN 9781450349185. doi: 10.1145/3132847.3132938. URL http://dl.acm.org/ft_gateway.cfm?id=3132938&type=pdf.

C.-N. Ziegler, S. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, pages 22–32, Chiba, Japan, 2005. ACM. ISBN 9781595930460. doi: 10.1145/1060745.1060754. URL http://portal.acm.org/citation.cfm?id=1060745.1060754.

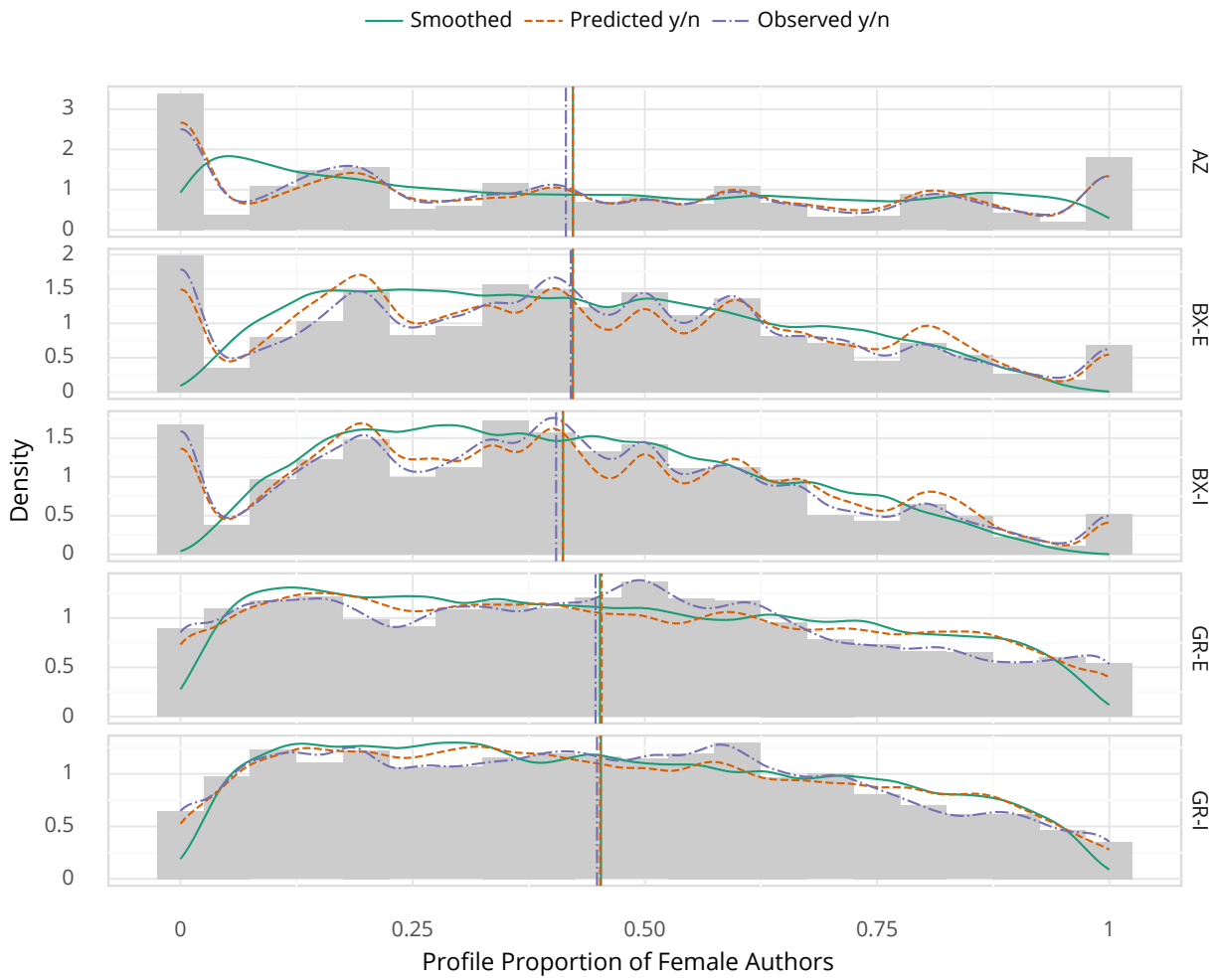

Figure 8: Distribution of user author-gender tendencies. Histogram shows observed proportions; lines show Gaussian kernel densities (bandwidth $1/2$ of Scott estimate) of smoothed tendencies (θ') along with observed and predicted proportions.

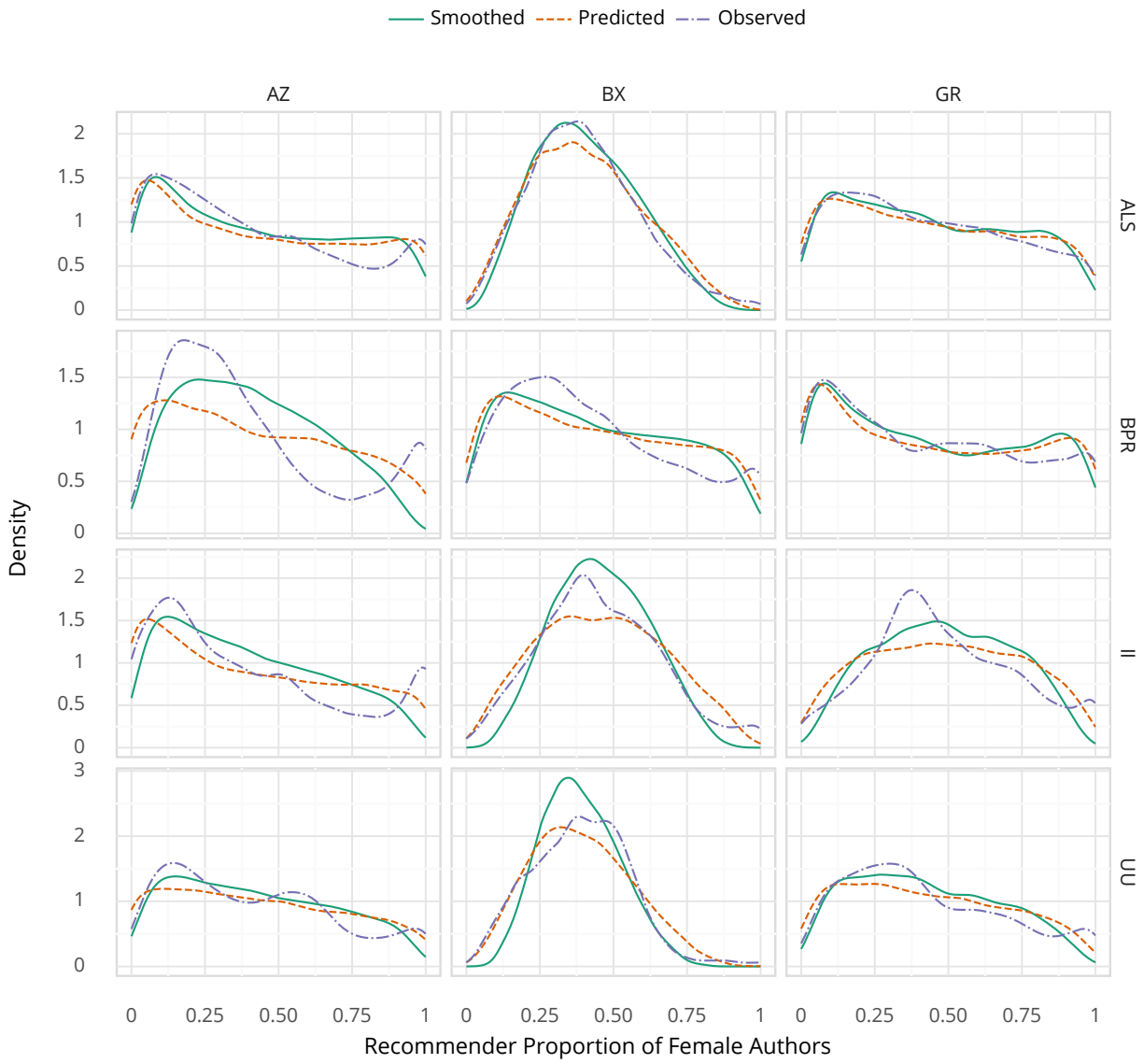

Figure 9: Posterior densities of recommender biases from integrated regression model (implicit feedback).

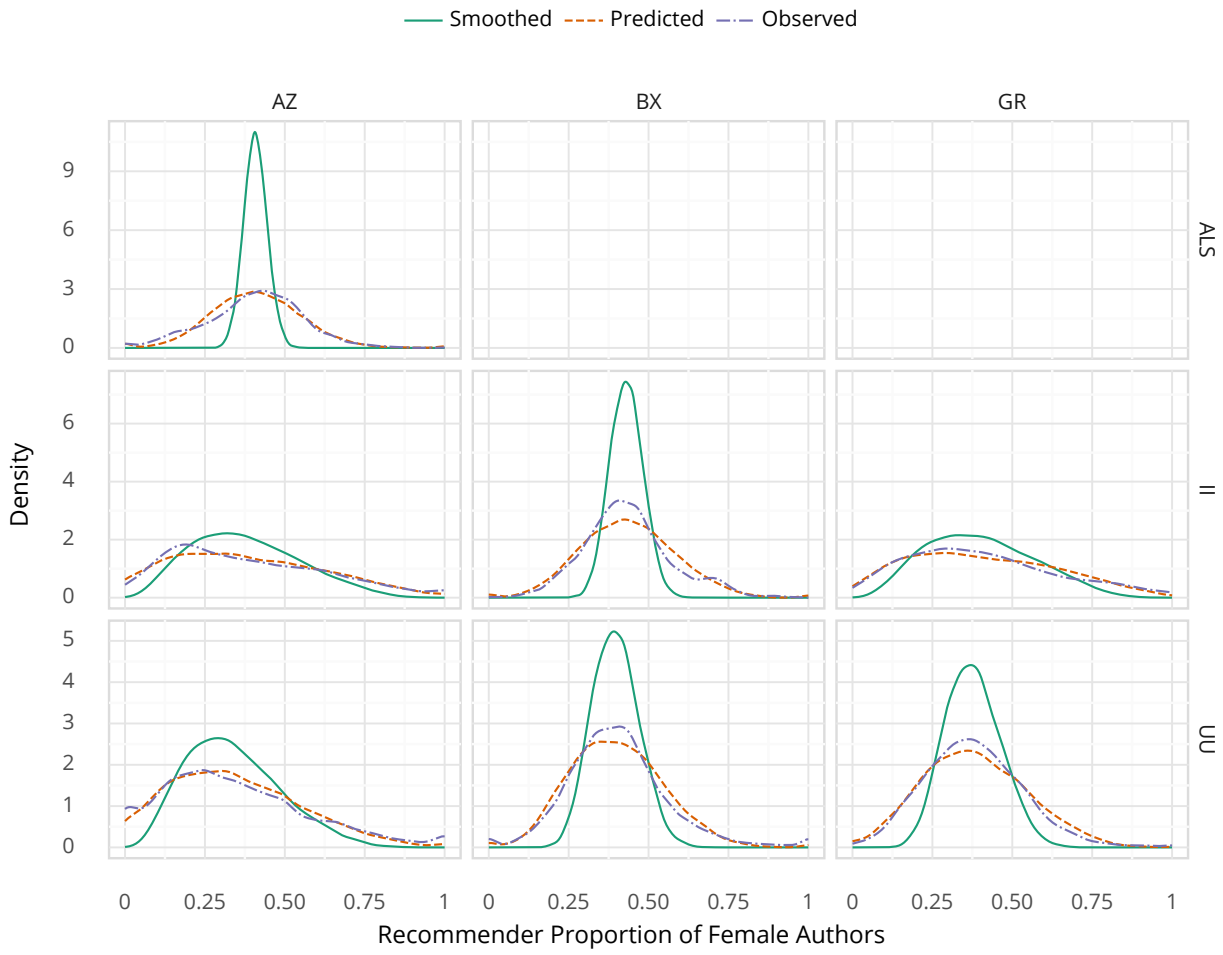

Figure 10: Posterior densities of recommender biases from integrated regression model (explicit feedback).

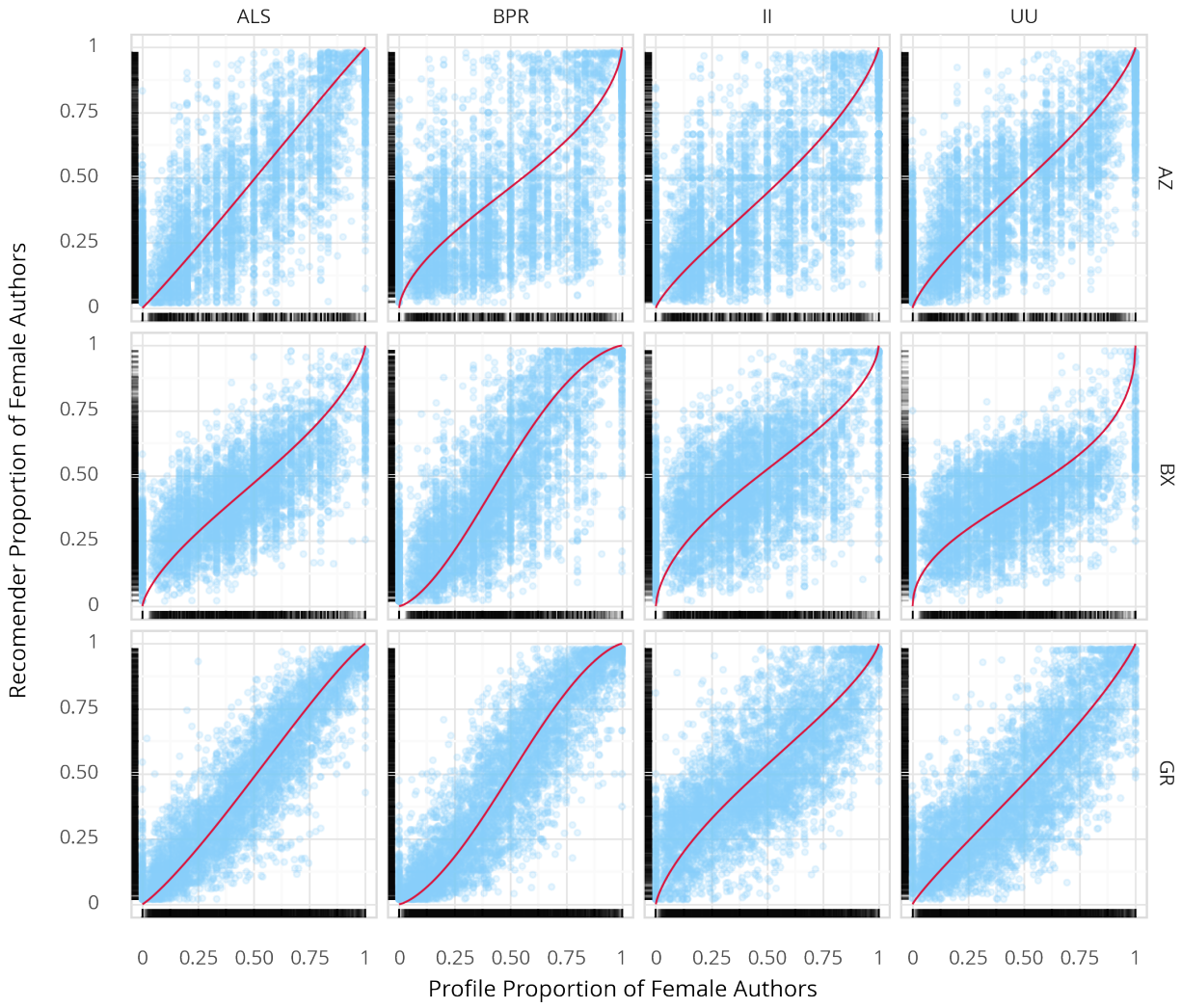

Figure 11: Scatter plots and regression curves for implicit feedback recommender response to individual users. Points are observed y/n proportions; curves are regression lines transformed from log-odds to proportions. Rug plots show marginal distributions.

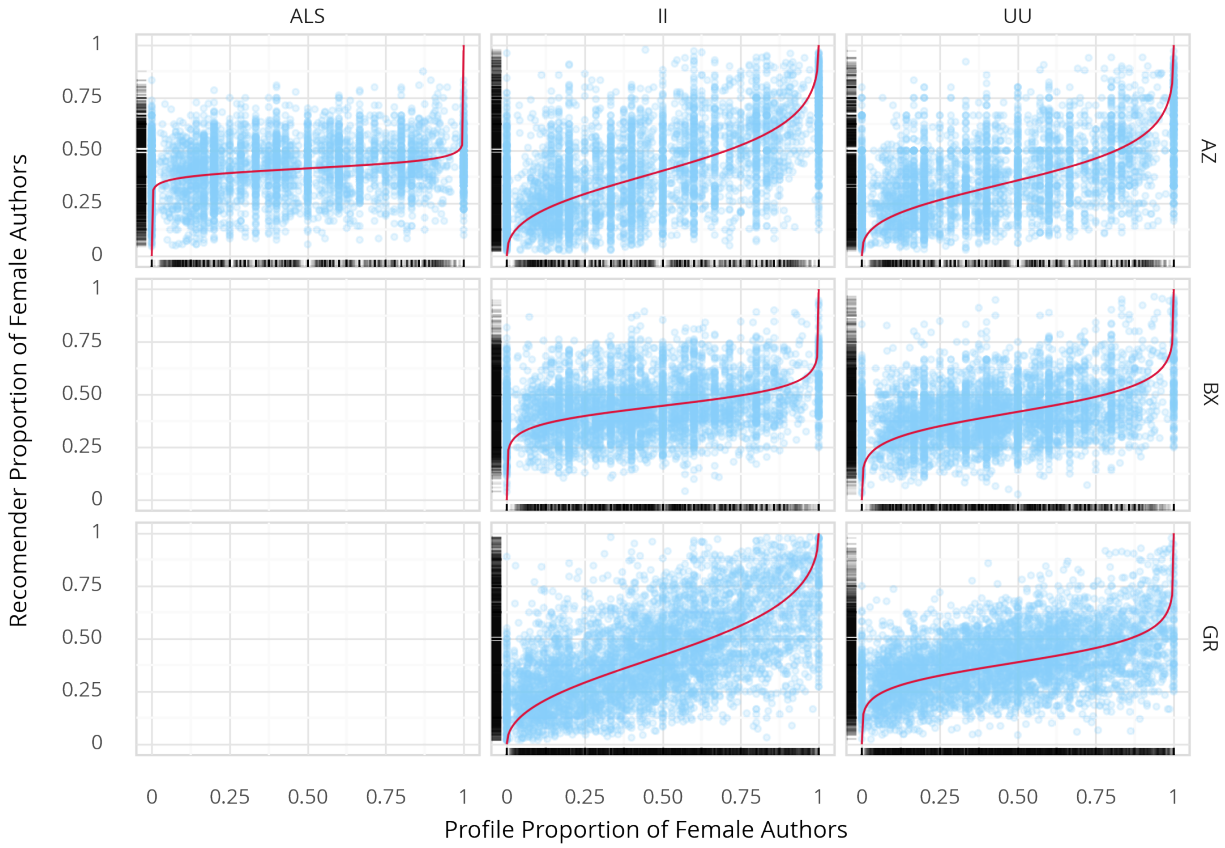

Figure 12: Scatter plots and regression curves for explicit feedback recommender response to individual users. Points are observed y/n proportions; curves are regression lines transformed from log-odds to proportions. Rug plots show marginal distributions.

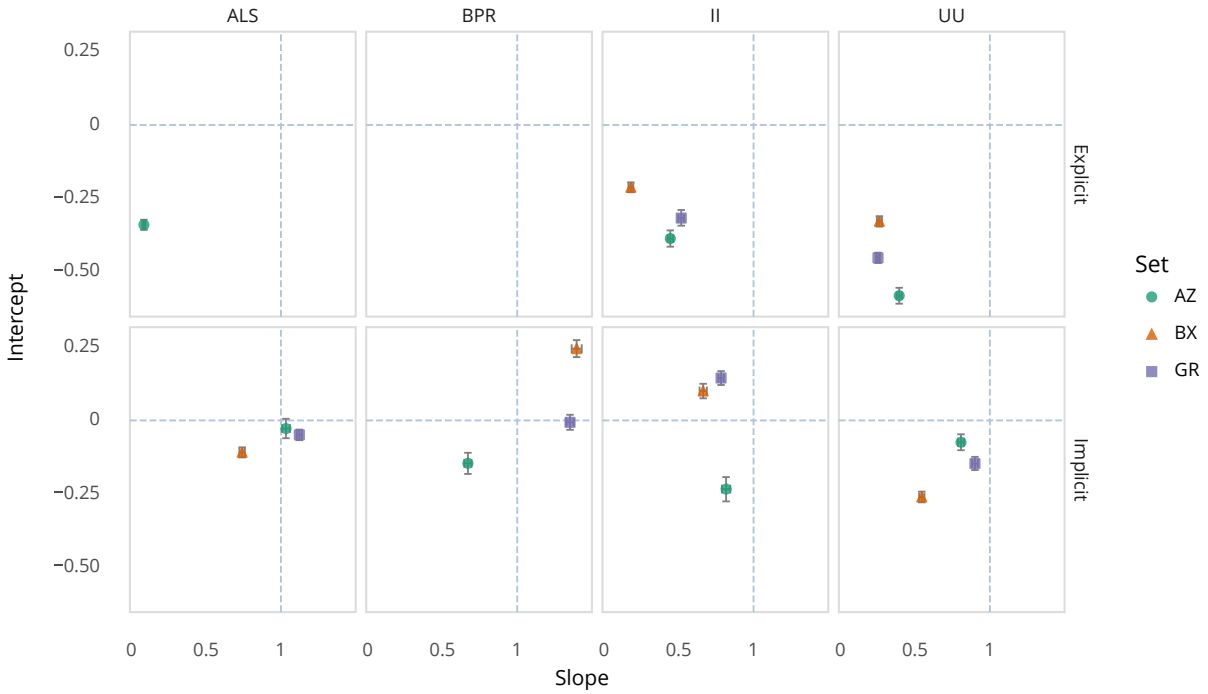

Figure 13: Slopes and intercepts for recommender models, with 95% intervals for the parameter estimate (slope intervals are narrower than the position dots). Dashed lines show perfect propagation (slope=1, intercept=0).

Figure 14: Top-N accuracy of natural recommenders and the Forced Balance strategies.