Exploring Author Gender in Book Rating and Recommendation

Michael D. Ekstrand
People & Information Research Team
Dept. of Computer Science
Boise State University
Boise, Idaho, USA
michaelekstrand@boisestate.edu

Mucun Tian
People & Information Research Team
Dept. of Computer Science
Boise State University
Boise, Idaho, USA
mucuntian@u.boisestate.edu

Mohammed R. Imran Kazi
Dept. of Computer Science
Texas State University
San Marcos, Texas, USA
md.kazi.1988@gmail.com

Hoda Mehrpouyan
Dept. of Computer Science
Boise State University
Boise, Idaho, USA
hodamehrpouyan@boisestate.edu

Daniel Kluver
Math, Statistics, & Computer Science
Macalester College
Minneapolis, Minnesota, USA
kluver@cs.umn.edu

ABSTRACT

Collaborative filtering algorithms find useful patterns in rating and consumption data and exploit these patterns to guide users to good items. Many of the patterns in rating datasets reflect important real-world differences between the various users and items in the data; other patterns may be irrelevant or possibly undesirable for social or ethical reasons, particularly if they reflect undesired discrimination, such as gender or ethnic discrimination in publishing. In this work, we examine the response of collaborative filtering recommender algorithms to the distribution of their input data with respect to a 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 differ in the gender distribution of their recommendation lists, and in the relationship of that output distribution to user profile distribution.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Social and professional topics** → *Gender*;

KEYWORDS

collaborative filtering; user impact; bias; discrimination

ACM Reference Format:
Michael D. Ekstrand, Mucun Tian, Mohammed R. Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring Author Gender in Book Rating and Recommendation. In *Twelfth ACM Conference on Recommender Systems (RecSys '18), October 2–7, 2018, Vancouver, BC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240323.3240373

RecSys '18, October 2–7, 2018, Vancouver, BC, Canada

1 INTRODUCTION

The evaluation of recommender systems has historically focused on the accuracy of recommendations [25, 42]. When it is concerned with other characteristics, such as diversity, novelty, and user satisfaction [28, 31, 49], it often continues to focus on 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 Uber [41], TaskRabbit [23], and search engines [35] interact with issues of bias, discrimination, and stereotyping.

Social impact is not a new concern in recommender systems. *Balkanization* [44] or *filter bubbles*, popularized by Pariser [39], are one example of this concern: do recommender systems enrich our lives and participation in society or isolate us in echo chambers? Recommender systems are intended to influence their users' behavior in some way; if they did not, there would be little reason to operate them. 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 that influence.

In this paper, we report observational results from our investigation into how recommender systems interact with author gender in book data and associated consumption and rating patterns. Our first step towards that end is to characterize the distribution of author genders in existing book data sets and the response of common collaborative filtering algorithms to that distribution. We address four research 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 the recommendations users receive from common collaborative

filtering algorithms? This measures the *overall* behavior of the 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.

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 that we seek to illuminate — recommender systems do not always propagate all input data patterns [7].

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 represents 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.

In the remainder of this paper, we describe our data set and experimental methodology, results of our experiment with existing algorithms, and some future directions for this line of research. The supplementary material archive accompanying this paper contains code to re-run our experiment and analyses.

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 identify relevant items amongst large sets of possibilities [1, 11]. 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.

While recommender evaluation and analysis often focuses on the accuracy of recommendations [25, 42], there has been significant work on non-accuracy dimensions of recommender behavior. Perhaps the best-known is diversity [49], sometimes considered along with novelty [28, 45]. Lathia et al. [32] examined the *temporal* diversity of recommender systems, studying whether they changed their recommendations over time. Other work has quantified recommendation bias with respect to classes of items [29].

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 [39, 44], mentioned earlier; recent work has sought to detect and quantify the extent to which recommender algorithms create or break down their users' information bubbles [37] and studied the effects of recommender feedback loops on users' interaction with items [27].

Other work seeks to use recommender technology to promote socially-desirable outcomes such as energy savings [43], better encyclopedia content [8], and new kinds of relationships [40].

2.3 Representation in the Book Industry

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 [38, 46]. We seek to understand how recommendation algorithms interact with such efforts: are they a help, a hindrance, or a neutral conduit?

2.4 Bias and Fairness in Algorithmic Systems

Questions of bias and fairness in computing systems are not new; Friedman and Nissenbaum [18] considered early on the ways in which computer systems can be (unintentionally) 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. [10] and Friedler et al. [17] have presented definitions of what it means for an algorithm to be *fair*. Feldman et al. [16] 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 [3] and predictive policing systems [15, 34], among others.

Burke [4] 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. [5] and Yao and Huang [48] have presented algorithms for C-fair collaborative filtering, and Ekstrand et al. [13] examine C-fairness in the accuracy of recommendation lists.

Our present study focuses on P-fairness. This dimension has not seen as much direct research, although it is related to historical concerns such as long-tail recommendation and item diversity [29]. Kamishima et al. [30] have presented algorithms for P-fair recommendation. In this paper, we present an offline empirical analysis of the P-fairness of several classical collaborative filtering algorithms and their underlying training data.

3 DATA AND METHODS

We address our questions through an experiment using publicly-available book data and common collaborative filtering techniques.

3.1 Data Sources and Integration

In order to analyze the demographic biases of user consumption patterns and resulting recommendations, we link multiple data sets to associate user ratings with book data. The accompanying code archive contains JavaScript, SQL, and R code for data import and integration, along with the LensKit recommender experiment code.

3.1.1 Book Consumption and Rating Data. We use two public sources of user consumption data: BookCrossing [49] and Amazon Book Reviews [36]. These data sets provide our historical user profiles and the training data for our collaborative filtering algorithms. Both are general reading data sets, consisting of user ratings for books across a wide range of genres and styles.

3.1.2 Book Metadata. We obtain book metadata, particularly author lists, by pooling records from Open Library[1] and the Library of Congress (LOC) MARC Open-Access Records[2].

3.1.3 Author Gender Data. We obtain author information from the Virtual Internet Authority File (VIAF)[3], a directory of author information compiled from authority records from the Library of Congress and other libraries around the world. Author gender identity is one of the available fields for many records.

The MARC21 data model [33] 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. Unfortunately, the VIAF does not use flexibility — all its assertions are "male", "female", or "unknown". This is a significant limitation that we discuss more fully in Section 5.1.

3.1.4 Linking Data Sets. We link book data with rating data by ISBN. To decrease sparsity, improve data linking coverage, and recommend at the level of creative works instead of individual editions, we link related ISBNs. We form a bipartite graph of ISBNs and records (LOC records, OpenLibrary "edition" records, and OpenLibrary "work" records when available) and consider each connected component to be a "book". Rarely (less than 1% of ratings) this causes a user to have multiple ratings for a book; we resolve multiple ratings by taking the median of their ratings.

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 (cleaning 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 cluster. 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 ensure good coverage while maintaining reasonable confidence in classification. Less than 5% of rated books have 'ambiguous' author genders.

3.1.5 Data Set Statistics. Tables 1–2 and Figure 1 summarize the results of integrating these data sets. In BookCrossing, approximately 75% of ratings are for books whose first author's gender we can identify; in Amazon, this drops to about 35% (27% of books). "% Female Books" and "% Female Ratings" indicate the fraction of known-gender books (or ratings) with female authors. While the data is somewhat 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.

[1]https://openlibrary.org/developers/dumps
[2]https://www.loc.gov/cds/products/marcDist.php
[3]http://viaf.org/viaf/data/

Table 1: Summary of book data

	LOC	All Known
ISBNs	6,522,453	20,485,242
'Books'	4,912,991	11,554,288
Matched Books	4,912,991	10,261,127
Known-Gender Books	2,567,646	4,251,559
Female-Author Books	580,719	1,009,290
Male-Author Books	1,986,927	3,242,269
% Female Books	29.2%	23.7%

Table 2: Summary of rating data

	BookCrossing	Amazon
Ratings	1,149,780	22,507,155
Users	105,283	8,026,324
Rated ISBNs/ASINs	340,554	2,330,066
Rated 'Books'	295,935	2,286,656
Matched Books	240,255	1,083,066
Known-Gender Books	166,928	616,317
Female-Author Books	66,524	181,850
Male-Author Books	100,404	434,467
% Female Books	39.9%	29.5%
% Female Ratings	45.3%	36.2%

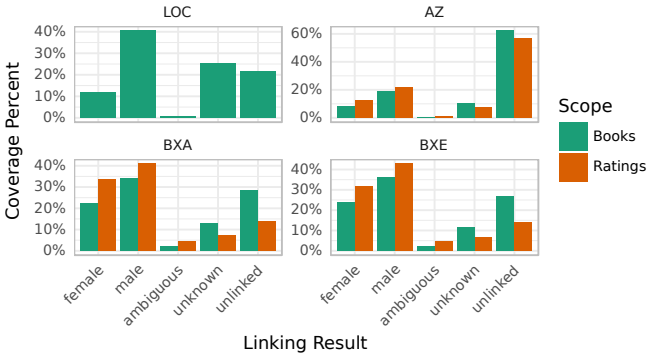

Figure 1: Results of data linking and gender resolution. LOC is the set of books with Library of Congress records; other panes are the results of linking rating data.

3.2 Experimental Design

Our main experiment has several steps:

(1) Import and integrate data, detect author genders.
(2) Sample 1000 users, each of whom has rated at least 5 books with known author gender, for analysis. This sampling keeps the analysis computationally tractable while considering enough users to draw statistically valid conclusions.
(3) Quantify gender distribution in sample user profiles (RQ2).

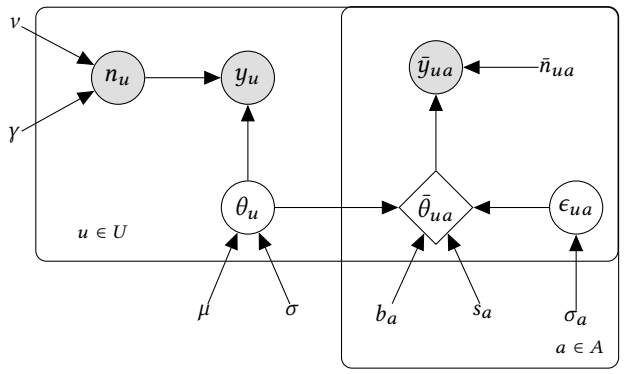

Figure 2: Plate digram for statistical model.

(4) Produce 50 recommendations for each of the sample users, using the entire data set data set for training.
(5) Compute recommendation list gender distribution (RQ3) and compare with user profile distribution (RQ4).

Gender could correlate with book rating behavior in two ways: the *selection* of books to read and/or rate, and the *rating values* given to those books the user has chosen to read. To account for these distinct effects, we run two versions of this experiment: one uses the rating values with explicit-feedback algorithm configurations; the other ignores rating values, treating each implicit or explicit as a "read" event, and operates the collaborative filters in implicit feedback mode. For BookCrossing, we perform these variants with separate user samples: explicit data sampled from the BookCrossing data set excluding implicit-only ratings (BXE), and implicit data sampled from the whole data set discarding rating values (BXA). For Amazon (AZ), we use a single user sample but compute recommendations both with and without rating values.

3.3 Recommending Books

We used the LensKit toolkit [12] to produce 50 recommendations for each of our 1000 sample users using the following algorithms:

- UserUser, a user-based collaborative filter [24].
- ItemItem, an item-based collaborative filter [9].
- MF, the FunkSVD matrix factorization algorithm [19].
- PF, hierarchical Poisson factorization [22].

We tuned each algorithm by optimizing nDCG on a train-test split of the consumption data. Full configurations are in the accompanying code; as we examine the behavior of algorithms, not test their effectiveness, precise details are omitted for space.

3.4 Statistical Analysis

We model user rating behaviors using a hierarchical Bayesian model [20] for the observed number of books by female authors out of the set of books with known authors. We extend this to model recommendation list distributions as a linear function of user profile distributions plus random variance. We select this strategy for three reasons: the variance in user profile sizes makes it difficult to directly compare gender proportions between users, a hierarchical Bayesian 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 of the multiple-comparison problems that often plague this kind of analysis [21].

Table 3: 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 author-gender balance)
μ	Expected user gender balance, in log-odds (E[logit(θ_u)])
σ^2	Variance of user gender balance
$\bar{n}_{ua}$	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
σ_a^2	Residual variance of algorithm a (its variability unexplained by user tendencies)

Figure 2 shows the plate diagram for our model, and Table 3 summarizes the key variables.

3.4.1 User Profiles. For each user, we observe n_u, the number of books they have rated with known author genders, 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[4] to address RQ2. We also model n_u as a random variable with a negative binomial distribution to produce more realistic predicted observations for unseen users to test model fit. We use the following joint probability as our likelihood model:

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

$\text{logit}(\theta_u)$ 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.

All parameters have vague priors: $\sigma, \nu, \gamma \sim \text{Exponential}(0.001)$, as they are positive, and $\mu \sim \text{Normal}(0, 100)$. These priors provide diffuse density across a wide range of plausible and extreme values.

[4] We also tested the more traditional beta model, but the logit-normal is more computationally efficient, better fits the data (using ELPD_{loo} as estimated by the R `loo` package), and produces a more internally consistent model when we extend it to handle recommendation lists via regression.

3.4.2 Recommendation Lists. For RQ3 and RQ4, we model recommendation list gender distributions by extending our Bayesian model to incorporate observed recommendation distributions via a linear regression based on a user's smoothed proportion and per-algorithm slope, intercept, and variance. This results in the following formula for estimating $\bar{\theta}_{ua}$:

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

For recommendation lists, we omit the binomial distribution used for modeling user profiles, and instead directly compute $\bar{\theta}_{ua} = (\bar{y}_{ua} + 1)/(\bar{n}_{ua} + 2)$. This change is because recommendations are not independent between users, and the highly consistent recommendations produced by some algorithms cause a binomial model to fit poorly (observed proportions are severely underdispersed). The regression residual ϵ_{ua} captures variance in the relationship between users' and algorithms' recommendation proportions, and giving it per-algorithm variance allows for some algorithms being more consistent in their output than others. The result 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.

3.4.3 Implementation. We fit and sample models with STAN 2.17.3 [6], drawing 10,000 samples per model (4 NUTS chains each performing 2500 warmup and 2500 sampling iterations). We report results with the posterior predictive distributions of the parameters of interest, as estimated by the sampling process.

4 RESULTS

In this section we present our experimental results with existing algorithms on our three data sets. We begin with characterizing the profiles of our sample users, and then proceed to analyze the resulting recommendations.

4.1 Baseline Distribution

The statistics of our underlying data sets, as presented in Tables 1–2, address RQ1. If we consider LOC as a representative sample of books-in-print, 29.2% of books with known author genders are female-authored; when considering all books appearing in our catalog data set, the proportion drops to 23.7%. Rating data has a more balanced distribution: 39.9% of known-gender books in BookCrossing are written by women, and 45.3% of ratings of known-gender books are for female-authored books. In Amazon the proportions are closer to baseline but are still higher (29.5% of books and 36.2% of ratings). If women are underrepresented among published authors, these results suggest that such underrepresentation is reduced in ratings from the systems we consider.

4.2 User Profile Characteristics

Our investigation into RQ2 focuses on 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 authors. Figure 3 shows the distribution of user profile sizes,

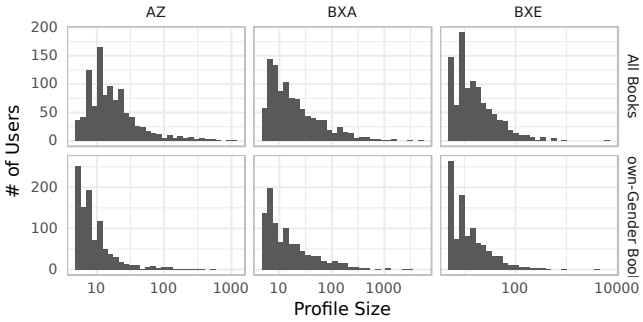

Figure 3: Distribution of user profile sizes.

Table 4: Summary statistics for user profiles gender distributions (log odds of P(female|known)).

	BXA	BXE	AZ
Mean Obs. Proportion	0.410	0.408	0.366
Std. Dev.	0.252	0.255	0.325
μ (est. mean log odds)	-0.42	-0.45	-0.83
95% Interval for μ	(-0.50, -0.34)	(-0.53, -0.37)	(-0.97, -0.70)
σ (est. sd log odds)	1.03	1.11	1.77
Posterior Mean θ	0.42	0.40	0.37
Std. Dev.	0.23	0.23	0.28

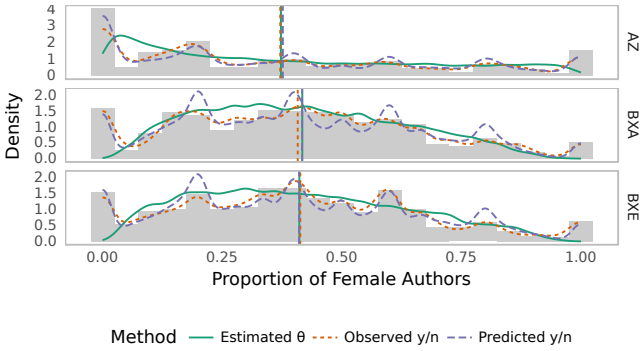

Figure 4: Distribution of user author-gender tendencies. Histogram shows observed proportions; lines show kernel densities of estimated tendencies (θ') along with observed and predicted proportions.

and Figure 4 shows the distribution of observed author gender proportions. Table 4 presents user profile summary statistics.

The Bayesian model from Section 3.4.1 provides more rigorous, smoothed estimates of this distribution. Table 4 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 4 shows the densities of the author-gender tendency distribution, along with the densities of projected and actual observed

Table 5: Recommendation coverage and diversity statistics.

	BXA			BXE			AZ (Implicit)			AZ (Explicit)		
	Users	Dist. Items	% Dist.	Users	Dist. Items	% Dist.	Users	Dist. Items	% Dist.	Users	Dist. Items	% Dist.
Profile	*1,000*	*35,187*	*66.5*	*1,000*	*24,913*	*73.6*	*1,000*	*27,525*	88.2	*1,000*	*27,525*	88.2
UserUser	1,000	6,007	12.0	988	6,235	12.7	1,000	15,343	30.7	939	25,853	55.1
ItemItem	1,000	21,282	42.6	997	10,174	20.4	999	33,363	67.7	999	22,360	45.6
MF	1,000	140	0.3	1,000	264	0.5	1,000	164	0.3	1,000	651	1.3
PF	1,000	1,506	3.0	1,000	4,105	8.2	1,000	2,746	5.4	1,000	3,538	7.0

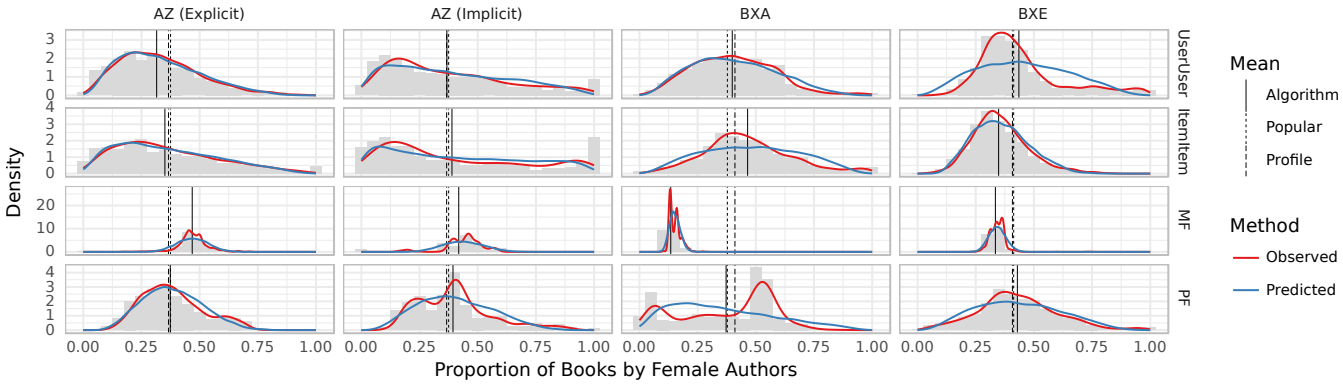

Figure 5: Posterior densities of recommender biases from integrated regression model.

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. To support direct comparison of the densities of observations and predictions, we resampled observed proportions with replacement to yield 10,000 observations.

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. The average user author-gender tendency is slightly closer to an even balance than the set of rated books. We also found a large diversity amongst 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. The Amazon data set has the strongest tendency ($\mu = -0.82$, mean(θ) = 0.37, sd(θ) = 0.28), with a particular spike in highly-male profiles.

4.3 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 3.3, we produced 50 recommendations from each algorithm. Table 5 shows the basic coverage statistics of these algorithms along with corresponding user profile statistics. 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.

Table 6: Mean / SD of rec. list female author proportions.

	BXA	BXE	AZ (Implicit)	AZ (Explicit)
Popular	0.458	0.500	0.364	0.364
Rating	—	0.383	—	0.222
UserUser	0.399 / 0.180	0.435 / 0.190	0.315 / 0.186	0.367 / 0.278
ItemItem	0.465 / 0.200	0.348 / 0.124	0.351 / 0.245	0.389 / 0.336
MF	0.134 / 0.027	0.334 / 0.039	0.468 / 0.079	0.418 / 0.124
PF	0.372 / 0.208	0.429 / 0.177	0.374 / 0.144	0.394 / 0.177

Table 6 provides the mean tendency for recommendation lists produced by each of our algorithms, plus the tendency of Most Popular and Highest Average Rating recommenders. Figure 5 shows the density of observed recommendation list proportions.

All recommenders were more consistent in their tendencies than the underlying user profiles. UserUser, Item Item, and PF exhibit significant variance in the items they recommend and the gender distributions of their output lists, though all are more concentrated than the user profile distribution. Their mean tendencies are comparable to input profile tendencies, as well as the popular-item tendency; the exact relationship varies from data set to data set. The MF algorithm barely personalizes at all, likely due to the high sparsity of the data; as a result, its observed tendency is much

Table 7: Regression parameters for algorithms with 95% credible intervals.

	UserUser			ItemItem			MF			PF		
	b_a	s_a	σ_a	b_a	s_a	σ_a	b_a	s_a	σ_a	b_a	s_a	σ_a
BXA	-0.151	0.656	0.574	0.179	0.662	0.742	-1.718	0.013	0.180	-0.468	0.542	1.020
	(-0.20,-0.10)	(0.60,0.71)	(0.54,0.61)	(0.12,0.24)	(0.60,0.72)	(0.70,0.78)	(-1.73,-1.71)	(0.00,0.03)	(0.17,0.19)	(-0.54,-0.40)	(0.46,0.62)	(0.97,1.07)
BXE	-0.139	0.162	0.906	-0.573	0.129	0.531	-0.652	0.002	0.161	-0.166	0.298	0.772
	(-0.20,-0.08)	(0.10,0.22)	(0.87,0.95)	(-0.61,-0.54)	(0.09,0.16)	(0.51,0.56)	(-0.66,-0.64)	(-0.01,0.01)	(0.15,0.17)	(-0.22,-0.11)	(0.25,0.35)	(0.74,0.81)
AZ (Implicit)	-0.127	0.688	0.715	0.094	0.863	0.895	-0.244	0.011	0.364	-0.224	0.287	0.537
	(-0.19,-0.06)	(0.65,0.73)	(0.68,0.76)	(0.02,0.17)	(0.81,0.92)	(0.84,0.95)	(-0.27,-0.22)	(-0.00,0.02)	(0.35,0.38)	(-0.26,-0.18)	(0.26,0.31)	(0.51,0.56)
AZ (Explicit)	-0.580	0.322	0.681	-0.380	0.438	0.852	-0.117	0.006	0.273	-0.403	0.141	0.525
	(-0.63,-0.53)	(0.29,0.35)	(0.65,0.71)	(-0.44,-0.32)	(0.40,0.48)	(0.81,0.89)	(-0.14,-0.10)	(-0.00,0.02)	(0.26,0.29)	(-0.44,-0.37)	(0.12,0.16)	(0.50,0.55)

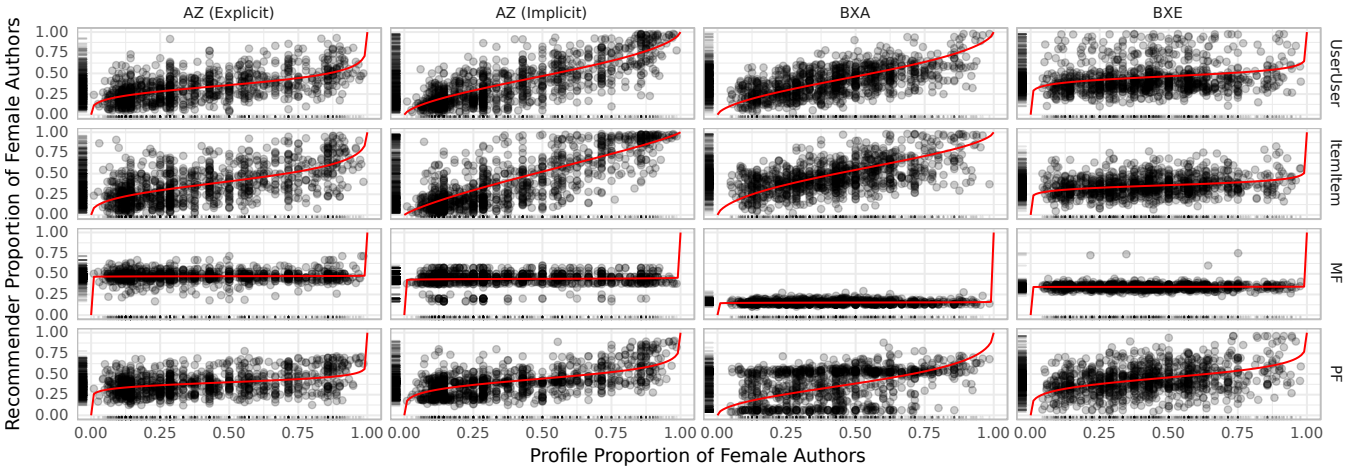

Figure 6: Scatter plots and regression curves for recommender response to individual users.

more concentrated. In the BookCrossing data, it tends to favor male authors more than the underlying data would support; in implicit feedback mode, it is highly biased towards male authors with respect even to the baseline distributions.

4.4 From Profiles to Recommendations

Our extended Bayesian model (Section 3.4.2) allows us to address RQ4: the extent to which our algorithms propagate individual users' tendencies into their recommendations (RQ4).

Figure 5 shows the posterior predictive and observed densities of recommender author-gender tendencies, and Figure 6 shows scatter plots of observed recommendation proportions against user profile proportions with regression curves (regression lines in log-odds space projected into probability space). Table 7 contains the parameters inferred for each regression line.

The k-NN algorithms are responsive to individual users' historical tendencies, as indicated by the positive slopes (s_a) with credible intervals excluding zero. MF is almost entirely unresponsive; PF responds some but not nearly so much as the k-NN algorithms and exhibits higher independent variance (σ_a^2).

The posterior model also does not fit as well for PF, because of its combination of responsiveness and global consistency. Our model can fit generally unresponsive curves such as MF, and generally responsive curves such as the k-NN models; PF sits in an awkward place. Visual inspection of the scatter plot suggests that there is a strong component with consistent tendencies, but the regression may accurately model the remaining users. Future work will use a model that can better account for some global consistency.

4.5 Summary

RQ1 — Baseline Gender Distribution Known books are significantly more likely to be written by men than by women; representation among rated books is more balanced.

RQ2 — User Input Gender Distributions User are diffuse in their rating tendencies, with an overall trend favoring male authors but less strongly than the baseline distribution.

RQ3 — Recommender Output Distributions Different CF techniques produce recommendations with quite different distributions. Matrix factorization on BookCrossing produced reliably male-biased recommendations, while nearest-neighbor and PF techniques were closer to the user profile tendency while being less diffuse than their inputs. Some algorithm and data set combinations resulted in recommendations that were more balanced than their inputs.

RQ4 — Distribution Propagation Most algorithms reflected some of each user's profile tendency in their recommendations; this effect was substantially stronger for implicit-feedback recommendations than explicit-feedback. Classical

matrix factorization did not exhibit significant personalization of any kind, likely due to data sparsity.

5 DISCUSSION

We found that users in the BookCrossing data set exhibit mild, diffuse tendency towards books written by men; users in the Amazon data set exhibit a somewhat stronger but still highly diffuse tendency. Both tendencies are more evenly balanced than the set of available of books. Collaborative filtering algorithms trained on this data exhibit remarkably different behavior; some learn substantially stronger and more consistent tendencies, in some cases producing lists more imbalanced than the item universe. Others propagate users' tendencies into their recommendation lists.

Nearest-neighbor recommenders in implicit feedback mode also propagated much of each user's profile tendencies into their recommendations. One interpretation of this is that they are partially picking up on a user's preference for books by male or female authors and reflecting this preference in the recommendations, which is what we would expect from a personalized recommender algorithm. The matrix factorization technique we tested consistently exhibited a much stronger bias towards male authors than was present in the input data, and was largely oblivious to individual users' preferences or biases (or, indeed, their book preferences). It also did not produce very accurate recommendations in our parameter tuning compared to the other algorithms.

The answer to the question "how do recommenders interact with gender distributions?" is therefore not simple. It has good company with other questions of the social impact of recommendations; for example, contrary to the filter bubble hypothesis, recommender algorithms had a *diversifying* effect on users' viewing portfolios in one movie recommendation service [37]. Exact answers likely depend on algorithm, application, and a number of other variables.

5.1 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 non-binary identities from an open vocabulary), VIAF does not seem to use this flexibility. The result is that gender minorities are not represented, or are misgendered, in the available data; we agree with Hoffmann [26] that this is a significant problem.

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 5 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.

6 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. The algorithms differed in their response to these distributions.

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 extending our statistical methods to account for them.
- Examining other content creator features, such as ethnicity, in recommendation applications.
- Studying other domains and applications, such as movies, research literature, and social media.
- Considering additional recommendation techniques.
- Studying change in distribution over time of both user consumption patterns and recommender outputs.
- 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 [14].
- Study the effect of existing refinements, such as diversification [47, 49], 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.

ACKNOWLEDGMENTS

We thank anonymous reviewers for helpful feedback on earlier versions of this paper. Experiments made use of the R2 cluster [2] operated by the Boise State Research Computing Department and hardware donated by Micron Technology.

REFERENCES

[1] G Adomavicius and A Tuzhilin. 2005. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* 17, 6 (2005), 734–749.

[2] Boise State Research Computing Department. 2017. R2: Dell HPC Intel E5v4 (High Performance Computing Cluster). (2017). https://doi.org/10.18122/B2S41H

[3] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*. http://tiny.cc/r330vy

[4] Robin Burke. 2017. Multisided Fairness for Recommendation. (July 2017). arXiv:cs.CY/1707.00093 http://arxiv.org/abs/1707.00093

[5] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced Neighborhoods for Multi-sided Fairness in Recommendation. In *Proc. FAT* 2018 (Proceedings of Machine Learning Research)*, Vol. 81. PMLR, New York, NY, USA, 202–214. http://proceedings.mlr.press/v81/burke18a.html

[6] Bob Carpenter, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A Probabilistic Programming Language. *Journal of Statistical Software* 76, 1 (2017), 1–32. https://doi.org/10.18637/jss.v076.i01

[7] Sushma Channamsetty and Michael D Ekstrand. 2017. Recommender Response to Diversity and Popularity Bias in User Profiles. In *Proc. FLAIRS 30*. AAAI Press. https://aaai.org/ocs/index.php/FLAIRS/FLAIRS17/paper/view/15524/15019

[8] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. 2007. SuggestBot: Using Intelligent Task Routing to Help People Find Work in Wikipedia. In *Proc. ACM IUI '07*. ACM, 32–41. https://doi.org/10.1145/1216295.1216309

[9] Mukund Deshpande and George Karypis. 2004. Item-based Top-N Recommendation Algorithms. *ACM Transactions on Information Systems* 22, 1 (2004), 143–177.

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proc. ITCS '12*. ACM, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[11] Michael Ekstrand, John Riedl, and Joseph A Konstan. 2010. Collaborative Filtering Recommender Systems. *Foundations and Trends in Human-Computer Interaction* 4, 2 (2010), 81–173.

[12] Michael D Ekstrand, Michael Ludwig, Joseph A Konstan, and John T Riedl. 2011. Rethinking the Recommender Research Ecosystem: Reproducibility, Openness, and LensKit. In *Proc. ACM RecSys '11*. ACM, New York, NY, USA, 133–140. https://doi.org/10.1145/2043932.2043958

[13] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Proc. FAT* 2018 (Proceedings of Machine Learning Research)*, Vol. 81. PMLR, New York, NY, USA, 172–186. http://proceedings.mlr.press/v81/ekstrand18b.html

[14] Michael D Ekstrand and Martijn C Willemsen. 2016. Behaviorism is Not Enough: Better Recommendations Through Listening to Users. In *Proc. ACM RecSys 2016*. ACM, New York, NY, USA, 221–224. https://doi.org/10.1145/2959100.2959179

[15] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2017. Runaway Feedback Loops in Predictive Policing. (June 2017). arXiv:cs.CY/1706.09847 http://arxiv.org/abs/1706.09847

[16] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proc. ACM KDD 2015*. ACM, 259–268. https://doi.org/10.1145/2783258.2783311

[17] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. *arXiv:1609.07236 [cs, stat]* (23 Sept. 2016). http://arxiv.org/abs/1609.07236

[18] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Transactions on Information Systems* 14, 3 (July 1996), 330–347. https://doi.org/10.1145/230538.230561

[19] Simon Funk. 2006. Netflix Update: Try This at Home. http://sifter.org/~simon/journal/20061211.html. (11 Dec. 2006). Accessed: 2010-4-8.

[20] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2014. Hierarchical Models. In *Bayesian Data Analysis* (3rd ed.). CRC Press, 101–138.

[21] Andrew Gelman and Francis Tuerlinckx. 2000. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* 15, 3 (2000), 373–390. https://doi.org/10.1007/s001800000040

[22] Prem Gopalan, Jake M Hofman, and David M Blei. 2013. Scalable Recommendation with Poisson Factorization. *arXiv:1311.1704 [cs, stat]* (Nov. 2013). http://arxiv.org/abs/1311.1704

[23] Aniko Hannak, Claudia Wagner, David Garcia, Markus Strohmaier, and Christo Wilson. 2016. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit. In *Proc. DAT Workshop*. http://datworkshop.org/papers/dat16-final22.pdf

[24] Jonathan Herlocker, Joseph A Konstan, Al Borchers, and John Riedl. 1999. An Algorithmic Framework for Performing Collaborative Filtering. In *Proc. ACM SIGIR 1999*. ACM, 230–237. https://doi.org/10.1145/312624.312682

[25] Jonathan Herlocker, Joseph A Konstan, Loren Terveen, and John Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems* 22, 1 (2004), 5–53.

[26] Anna Lauren Hoffmann. 2018. Data Violence and How Bad Engineering Choices Can Damage Society. (April 2018). https://medium.com/s/story/39e44150e1d4 Accessed: 2018-5-1.

[27] Kartik Hosanagar, Daniel Fleder, Dokyun Lee, and Andreas Buja. 2013. Will the Global Village Fracture Into Tribes? Recommender Systems and Their Effects on Consumer Fragmentation. *Management Science* 60, 4 (Nov. 2013), 805–823. https://doi.org/10.1287/mnsc.2013.1808

[28] Neil Hurley and Mi Zhang. 2011. Novelty and Diversity in Top-N Recommendation – Analysis and Evaluation. *ACM Transactions on Internet Technology* 10, 4 (March 2011), 14:1–14:30. https://doi.org/10.1145/1944339.1944341

[29] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction* 25, 5 (25 July 2015), 427–491. https://doi.org/10.1007/s11257-015-9165-3

[30] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2018. Recommendation Independence. In *Proc. FAT* 2018 (Proceedings of Machine Learning Research)*, Vol. 81. PMLR, New York, NY, USA, 187–201. http://proceedings.mlr.press/v81/kamishima18a.html

[31] Bart Knijnenburg, Martijn Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (1 Oct. 2012), 441–504. https://doi.org/10.1007/s11257-011-9118-4

[32] Neal Lathia, Stephen Hailes, Licia Capra, and Xavier Amatriain. 2010. Temporal diversity in recommender systems. In *Proc. ACM SIGIR 2010*. ACM, 210–217. https://doi.org/10.1145/1835449.1835486

[33] Library of Congress. 1999. *MARC21 Standards*. Technical Report. https://www.loc.gov/marc/

[34] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (Oct. 2016), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

[35] Gabriel Magno, Camila Souza Araújo, and Wagner Meira, Jr. 2016. Stereotypes in Search Engine Results: Understanding The Role of Local and Global Factors. In *Proc. DAT Workshop*. http://datworkshop.org/papers/dat16-final35.pdf

[36] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proc. ACM SIGIR 2017 (SIGIR '15)*. ACM, 43–52. https://doi.org/10.1145/2766462.2767755

[37] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2014. Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity. In *Proc. WWW 2014*. ACM, New York, NY, USA, 677–686. https://doi.org/10.1145/2566486.2568012

[38] Vesna Pajović and Kristina Vyskocil. 2016. 2015 CWILA Count Methods and Results. (Oct. 2016). https://cwila.com/2015-cwila-count-methods-results/ Accessed: 2018-5-7.

[39] Eli Pariser. 2011. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin.

[40] Paul Resnick. 2001. Beyond Bowling Together: Sociotechnical Capital. *HCI in the New Millennium* 77 (2001), 247–272.

[41] A Rosenblat and L Stark. 2016. Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers. *Inter. Journal of Communication* 10 (2016), 27.

[42] Guy Shani and Asela Gunawardana. 2010. Evaluating Recommendation Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B Kantor (Eds.). Springer, 257–297.

[43] Alain Starke, Martijn Willemsen, and Chris Snijders. 2017. Effective User Interface Designs to Increase Energy-efficient Behavior in a Rasch-based Energy Recommender System. In *Proc. ACM RecSys 2017*. ACM, New York, NY, USA, 65–73. https://doi.org/10.1145/3109859.3109902

[44] Marshall van Alstyne and Erik Brynjolfsson. 2005. Global Village or Cyber-Balkans? Modeling and Measuring the Integration of Electronic Communities. *Management Science* 51, 6 (June 2005), 851–868. https://doi.org/10.1287/mnsc.1050.0363

[45] Saúl Vargas and Pablo Castells. 2011. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *Proc. ACM RecSys 2011*. ACM, New York, NY, USA, 109–116. https://doi.org/10.1145/2043932.2043955

[46] VIDA. 2017. The 2016 VIDA Count | VIDA: Women in Literary Arts. (Oct. 2017). http://www.vidaweb.org/the-2016-vida-count/ Accessed: 2018-5-7.

[47] Martijn C Willemsen, Mark P Graus, and Bart P Knijnenburg. 2016. Understanding the role of latent feature diversification on choice difficulty and satisfaction. *User Modeling and User-Adapted Interaction* 26, 4 (Oct. 2016), 347–389. https://doi.org/10.1007/s11257-016-9178-6

[48] Sirui Yao and Bert Huang. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. In *Advances in Neural Information Processing Systems 30*. 2925–2934. http://tiny.cc/a330vy

[49] Cai-Nicolas Ziegler, Sean McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving Recommendation Lists through Topic Diversification. In *Proc. WWW 2005*. ACM, 22–32. https://doi.org/10.1145/1060745.1060754