EXPLORING POTENTIALLY DISCRIMINATORY BIASES IN BOOK RECOMMENDATION

by

Mohammed Imran Rukmoddin Kazi

A thesis submitted to the Graduate College of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Science
with a Major in Computer Science
August 2016

Committee Members:

Michael D. Ekstrand, Chair

Byron Gao

Vangelis Metsis

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

DEDICATED

to

my MOTHER, FATHER, Wife

and

my FRIENDS

ACKNOWLEDGEMENTS

I would like to express my special appreciation and thanks to my advisor Dr. Michael D. Ekstrand, department of computer science at Texas State University, you have been a tremendous mentor for me. Your advice on both research as well as on my career has been invaluable. I could not have imagined having a better advisor and mentor for my work. He was never ceasing in his belief in me and always providing clear guidance.

I would like to thanks Dr. Byron Gao, department of computer science at Texas State University, to be part of my thesis committee and also providing motivating feedback for this work.

I would like to thanks, Dr. Vangelis Metsis, department of computer science at Texas State University, to be part of my thesis committee and also for providing constructive feedback.

I would like to thanks, Dr. Oleg Komogortsev, department of computer science at Texas State University, to be part of my thesis committee.

Finally, I must express my special thanks to my parents and to my wife for providing me with unfailing support and continuous encouragement for this work. This accomplishment would not have been possible without them.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

ABSTRACT

Recent issues which occurred in the field of artificial intelligence present disproportionality based on protected attributes such as sex, race, and ethnicity in their output had raised concerns. The algorithms used in AI may amplify or propagate biases which exist in the historical data and may reflect this in the output data. Computer world now does not consider this as an abstract fact and researchers are coming up with the new frameworks that modify the existing algorithms present in AI which aids these biases to be reduced to a reasonable level. Recommender System algorithms are well optimized with respect to accuracy and efficiency. But as recommender systems are built on top of Information Retrieval, Machine Learning, and Artificial Intelligence, these systems have high chances of producing a biased outcome. Our current research focus on building methodology for explores potentially discriminatory biases based on protected characteristics in Recommender System. Plus, the definition of discrimination in this work does not correlated with any particular definition which had been define in past. For this work we have taken Book Recommender as a basis for observation of the bias in both input and output of a recommender.

I. INTRODUCTION

Potentially Discriminatory Behavior in Artificial Intelligence

Artificial intelligence systems are all around us and these smart systems are used in different ways of our daily life. Further, these systems when integrated with different domains help people in decision making, i.e.:

- predicting purchases on online retail stores based on the user shopping history.
- evaluating an applicant's loan eligibility.
- suggesting recommendations about hotels.
- restaurants, to a mobile user as a personalized virtual assistant restaurant.
- assisting an educational institute in the processes of granting admits.
- assisting recruiter to filter out a job applicant.
- providing search recommendations in a search engine etc.

These are features of artificial intelligence gaining people trust.

As Artificial intelligence system have become more pervasive the concern has been raised as to whether they are fair or if they perpetuate or perhaps even create biases and potentially discriminatory outcomes i.e.

- using summon 2.0, an academic library discovery system, to search with keyword "stress in workplace" result in displaying documents related "women in the work force" (Reidsma 2016).
- tool like predictive policing, which is used to predicts the future crime's approximate location based on historical data about crime without specifying how

demographic information is excluded while computing the prediction (Madrigal 2016).

- AdFisher tool explore the behavior of ads setting page of google which render out high paid jobs ads for male compare to female job-seeker, (Datta, Tschantx, and Datta 2014).

Our current research work will focus developing a methodology for identifying and measuring potentially discriminatory output in recommender system based on input data.

Recommender System

Recommender Systems are built on top of Machine learning, Artificial Intelligence, and Information Retrieval algorithms (M. Ekstrand et al. 2011). These systems help users in discovering relevant items which they might had not experience before for example Movelens recommends list of movies based on user preferences, Amazon recommends books to user based on search and purchase history, and Yelp helps users to find good burger restaurants. The algorithms used in this system work with underlying data to predict or recommend list of items in application which come from different domains such as e-commerce, employment, and social media etc. Over a period, the algorithms used in this system are being significantly improved in terms of accuracy and computation time. But as these systems are exposed to data which is potentially biased, they are likely to replicate such biases into their outputs. Hence the application may narrow the outcome instead of expanding user options.

Current work

This work makes initial steps to form a methodology in understanding potentially discriminatory bias in recommender system. We do this by examine the genders of author of book rated by and recommended to user by finding answers for the below question:

RQ 1. How particular author's gender is distributed in user profile input data?

RQ 2. How particular author's gender is distributed in outcome of recommender system?

RQ 3. How much the variation in recommender output is depended on the user profile input data?

Overview of the Upcoming Chapters

The organization of our current work is as follows:

- In Chapter II we survey relevant studies on fairness and non discrimination aspect in machine learning application.
- In Chapter III, we describe about selection of dataset, processing the data and how the demography attribute value of author is derived and recommender experiment. Here, we also present a detailed discussion on the model design to observe bias and implementation details.
- Chapter IV, we present experimental results as well as analyses.
- Chapter V consists of the concluding remarks and future work.

II. BACKGROUND AND RELATED WORK

In the last decade or so, a strand of machine learning research has emerged that moves past the accuracy of the system and considers various other aspects of its outputs and applications, such as whether its output is fair (Custers et al. 2014). This is increasingly important as such systems are deployed more pervasively in society and used to make decisions with serious impact on peoples' lives such as credit and insurance decisions, incarceration, and job suitability(Sowell 2015).

When considering fairness and non-discrimination in machine learning applications, the problem is usually framed as learning from data that was produced by biased or discriminatory human processes, such as arrest records, but trying to learn a classifier whose output is non-discriminatory. Further, it is not sufficient to eliminate direct influence; under the theory of disparate impact, recently affirmed by the United States Supreme Court (Sowell 2015) discrimination arises if a protected group is disproportionately impacted by a decision process compared to an unprotected group, even if their protected status was not incorporated into the decision process. For example, a process that denies the right to vote to blacks more often than it denies it to whites is discriminatory, even if race is not used as a basis for the decisions. This is operationalized in machine learning by seeking to have the output be uncorrelated with the protected status while maintaining accuracy(Zafar et al. 2015).

Zliobiate and other researcher focusing on the concept of obey non-discrimination aspect in decision making and data mining machine learning algorithm(Zliobiate 2015; Jelveh and Luca 2015). For this they consider the example of the how financial institute use tools which are based on decision making and data retrieval algorithm evaluate outcome to approve a loan or credit card for a particular person. This research work's motivation was to figuring out the discrimination in the outcome of these algorithms and also to provide promising approach to increase the accuracy of it outcome. These researchers also propose constraints on classifier to avoid discriminatory outcomes.

Recommender systems are a common machine learning application which producing personalized suggestion to the users in a wide range of (M. D. Ekstrand, Riedl, and Konstan 2011; Konstan and Others 1997; Linden, Smith, and York 2003; Adomavicius and Tuzhilin 2005)(Ekstrand, Riedl, and Konstan 2011; Konstan and Others 1997; Linden, Smith, and York 2003; Adomavicius and Tuzhilin 2005). These algorithms can impact what people see of the world (Tufekci 2014). It is has been proposed, and is now generally agreed, that accuracy is not the only concern for a recommenders (McNee, Riedl, and Konstan 2006); our work will examine fairness as a new non-accuracy concern of recommendation.

III. METHODOLOGY

To achieve our research objectives, we have analyzed user ratings of books and recommendations produced from those ratings, using the author's gender as the protected characteristic. Using this model first we observed the distribution of female authors in the output of popular recommender algorithms. Then we have observed the distribution of the same author's gender in user profile data. We have constructed statistical model which compute probability distribution of user consumption rate based on author's gender. Then we have performed a comparison between user profile and recommender output data by plotting the distribution of female author.

Dataset

To study potentially discriminatory bias in recommender system we required a domain and data sets containing at least:

- Consumption or rating data to train a recommender.
- Information related to one or more protected characteristic i.e. gender, color and race etc.

Books rating are a natural choice, as authors have genders and ethnicities and author representation is a matter of raised concern. For this work, we combine two data sets:

- Book-Crossing provides ratings of books, giving as readable and preference data.
- Open Library provide book meta data including author.

We have joined these data sets by ISBN. We processed the data with Python scripts and a PostgreSQL database.

Author's Demographic Data

As we need to perform statistical computation base on the protected characteristics of author. Which open library has author name, it does not provide demographic data such as gender or ethnicity. As author's gender information was not available in the current data sets. Plus, we have considered only those relevant author for gather gender information whose book have rating information. To gather author's gender information, we tried two different approaches:

- Gender-Detect
- Wikidata

a. Gender-Detector

The author's gender data was derived from 'gender-detector' Python library (Vanetta 2014). Prediction is done base on Open Gender Tracker's Global Name Data from different countries (Bmerrill 2014) . The guess function of this library returns three possible outcome based on the given name: 'female', 'male' or 'unknown'. Unknown outcome occurs when internal computation value is below a preset threshold value. Plus, gender detector work with only the first name of person while detecting the gender.

Figure 1 shows the logic of our use of the gender-detect library to guess author's gender based on first name.

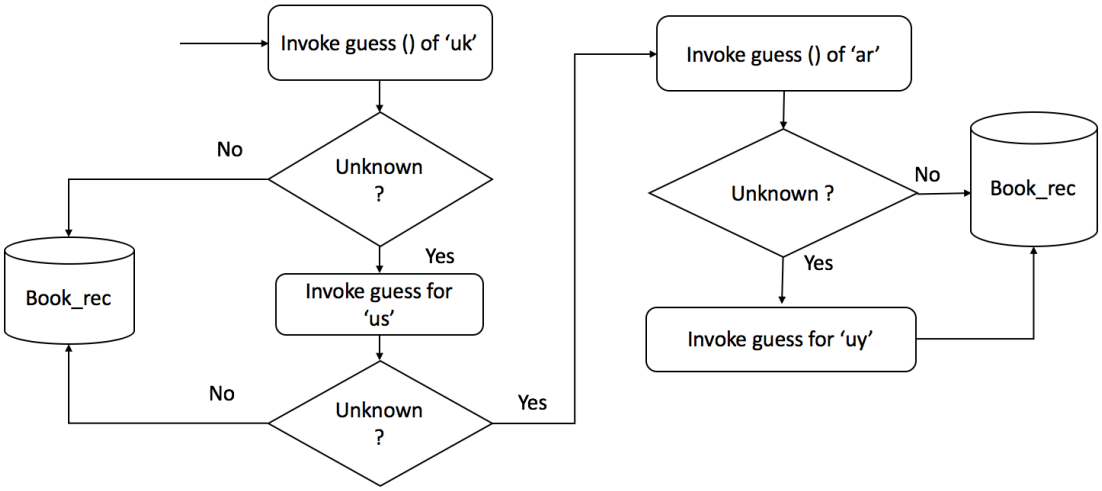

Figure 1 Gender Naming Repository

Table 1 denote associative country for each configured naming repository object

Table 1 Naming Repository

Objects	Country Name
detect_uk	United Kingdom
detect_us	United States
detect_ar	The government of the City of Buenos Aires
detect_uy	Uruguay

The reasons behind configuring all the available naming repository is that, these authors belong to different ethnicity or country and one naming repository might not contain enough data to predict gender for all of these authors.

Gender Detect logic:

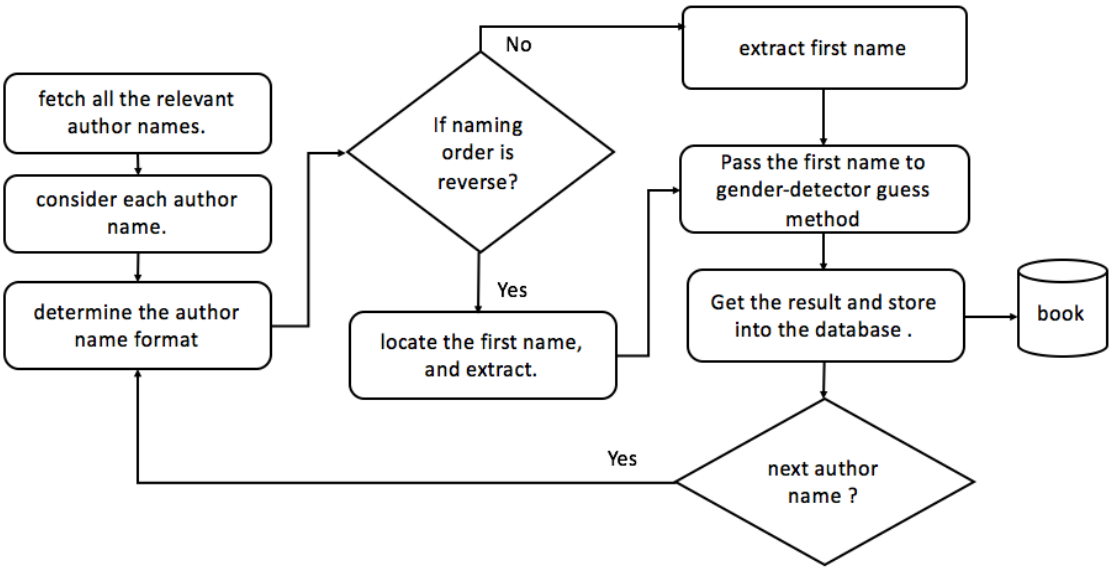

Figure 2 Gender Detector Flow Chart

Figure 2 shows the overall logic for our gender detection script. One of the particular problems that it needs to solve is that author names can be in two different formats in the Open Library data: 'First Last' (e.g. 'Eoin Colfer') or 'Last, First' (e.g. 'Jong, Enrica'). We handle this by checking if the name has a comma, and if it does, we reverse the name.

b. Wikidata

We tried getting author data from Wikidata (a sister project to Wikipedia), but it had insufficient coverage.

Sample Data

We analyzed a sample of 1000 users from the Book Crossing Dataset. Our sample data contains user rating for one thousand random users. We ignored books with unknown authors and restricted our user sample to user with at least 5 ratings count.

For each user in our sample, we computed the number of books read by female author and the total number of books.

Table 2 Sample Data Format

user	female author books read	total number books read
1331	1	5

Recommender Experiment

Recommender experiment for book data was performing with help of Lenskit utility, which is use to build recommender system for small and medium scale application (M. Ekstrand et al. 2011). For this experiment we used several popular recommender system algorithms, including non-personalized and collaborative filtering algorithms:

- UserUser collaborative filtering
- ItemItem collaborative filtering
- SVD collaborative filtering
- Item mean rating
- Popular Items

We had organized the book rating data into test data and train data so we could use Lenskit's train-test evaluator to run the experiment. The train data consist of all the book

rating data present in the database; the test data consisted of a single fake rating by each user in our sample, so that the evaluator would generate recommendations for them Recommended list for each user was set to size 100 items.

Lenskit Configuration

User-User Algorithm:

For this experiment the neighbor hood size was set to size 30 and cosine similarity over normalized vectors was used to compute user similarity.

Item-Item Algorithm:

For this experiment the neighbor hood size was set to size 20 and uservectornormalizer was bind with baselinesubtractinguservectornormalizer.

SVD Algorithm:

For this experiment the Feature Count was set to 40 and the Iteration Count was set to 125.

We analyze the following:

- Distribution and mean of the proportion of female-author books in user input profiles.
- Distribution and mean of the proportion of female-author books in recommendation lists.
- Relationship of recommended proportion to input proportion.

Bayesian Model

We have used a Bayesian model to infer a distribution of user and recommendation biases from observed ratings. We have used Bayesian approach to construct our model because in order to account for different users having different number of rating. We

have adapted our analysis from a similar problem in Chapter 5 of *Bayesian Data Analysis* (Gelman et al. 2013).

In our data set, for each user j, we have two observations:

- count of books by female authors read by user which is represented by y_j in the below diagram
- the total number of book read by user which is represented by n_j in the below diagram.

We model the data as being generated via a latent bias towards female authors, denoted by θ_j, for each user. We want to compute proportion distribution of given author's gender.

Our goal is to compute the probability distribution of these biases: $\boldsymbol{P(\theta|y)}$

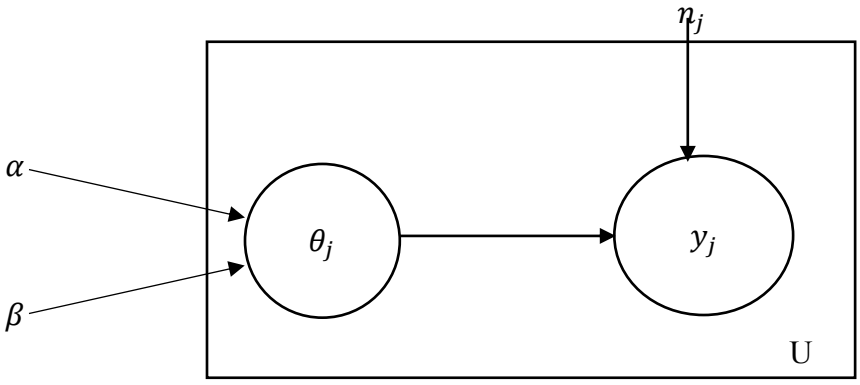

Figure 3 Plate Diagram

In more detail, we use the following model:

- $y_j \sim \text{Bin}(n_j, \theta_j)\ P(y|\theta)$
- $\theta_j \sim \text{Beta}\ (\alpha, \beta)\ \text{P}\ (y|\alpha, \beta)$

As per Bayesian Theorem:

- $P(\alpha, \beta|y, n) \propto P(y|\alpha, \beta, n)\ P(\alpha, \beta)$

Our likelihood model:

- $P(y_j|\alpha, \beta, n_j) = P(y_j|\theta_j, n_j)\ P(\theta_j|\alpha, \beta)$

Prior for this computation is:

- diffuse prior:

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-\frac{5}{2}}$$

Bayesian Procedure

We define the model:

y_j ~ Bin(n_j, θ_j) $P(y|\theta)$

θ_j ~ Beta (α, β) P ($y|\alpha, \beta$)

Then we derive the final equation, as per Eq. 5.8 in Gelman et al. is.

$$P(y_j|\alpha, \beta) = P(y_j|\theta_j, n_j)\, P(\theta_j|\alpha, \beta)$$

Compute the prior probabilities taking α, β

Finally , because for this experiment prior and likelihood functions operate over (α, β), but as the prior is over $\log\left(\frac{\alpha}{\beta}\right)$ and $\log(\alpha + \beta)$, need to compute the Jacobian for transformation; this is product of α, β $i.e.(\alpha \times \beta)$ and then compute the log of this product. Plus, transform the posterior on α, β to the transformed parameter space.

Now estimate the value of E[α]; by $\sum P(\alpha, \beta|y)$ to obtain expected value and repeat the same for β.Posterior expected value of θ can we computed with $\frac{\alpha}{\alpha+\beta}$

We computed posterior distribution in two ways:

Point wise, taking expected value of α, β and plug in to Beta to get $P\ (\theta|y, n)$

and the Integral, using distribution of α, β.

$$P(\theta|\mathbf{y}) = \iint P(\theta|\alpha,\beta)P(\alpha,\beta|\mathbf{y})\mathrm{d}\alpha\mathrm{d}\beta$$

Expected value of θ is computed by taking the weighted mean:

$$E[\theta] = \frac{\int_0^1 \theta\ P(\theta|\vec{y})d\theta}{\int_0^1 P(\theta|\vec{y}))\ d\theta}$$

We approximate made the integral by evaluating $P(\theta|\vec{y})$ at 1000 points on (0,1)

Limitations

For the current research work we have use one dataset which is book. We have considered only non-personalized and collaborative filtering algorithms. We have used python library which detect author's gender from person's first complete name hence there is possibility of producing 'unknown' as gender for particular name. Current approach of gender detect is for binary notion of gender. Have consider only one protected characteristic of author i.e. gender for this research work.

IV. RESULTS

Here we are presenting all the experimental results which will provide answers to all our 3 research question. Firstly, we have provided results related to the distribution of female author in user profile input data. Then we have provided results related to the distribution of female author are distribution in output of popular recommender algorithms. Later we have shown the results of a Bayesian statistical computation results which help in understanding how the gender bias is distributed in the both user profile input data and output of recommender algorithms. Finally we have shown the comparison of the results for female author in the user profile input data and output of various recommender algorithms.

Author Distribution in User Profile

In this experiment we have computed female author proportion for each user in user profile input data and overall female author proportion rate. This female distribution is from the sample data, which is selected from user profile input data. This computed proportion for female author was observing by plotting histogram.

Below formula was use to compute the overall proportion rate for female authors:

$$\frac{\sum(count\ female\ \ author\ books\ for\ all\ the\ user)}{\sum(\ total\ number\ of\ books\ \ for\ all\ the\ user\ )}$$

Below formula was use to compute the female author proportion for user rating and the same formula applied for recommender system output.

$$\frac{count\ of\ female\ author\ books}{total\ number\ of\ books}$$

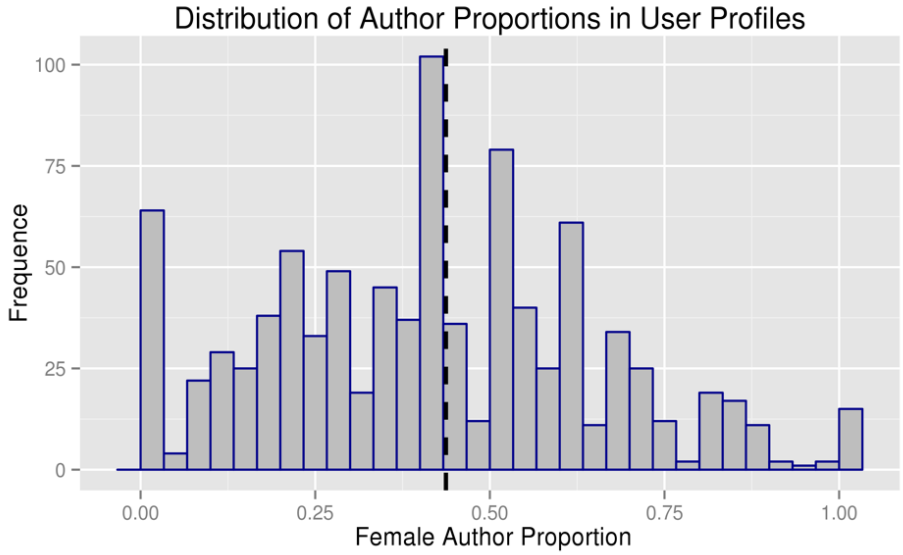

Figure 4 Female Authors in User Profile Input Data

dotted line: overall mean proportion for female author

Observation:

Overall female author proportion rate is 0.43. We can observe from this distribution in Figure 4 that users have a small but noticeable trend away from reading books by the female author.

Author Distribution in Non-Personalize Algorithm

We now examine the distribution of female author proportion for each user in Non-Personalize Recommender output. This computed proportion for female authors was observing using histogram. The below distribution of female author provide a partial answer to our second research question RQ 2. We have select those user detail which are are present in sample data. Below histogram represent distribution of female author proportion:

- User profile input data
- ItemMean Algorithm

- Popular Algorithm

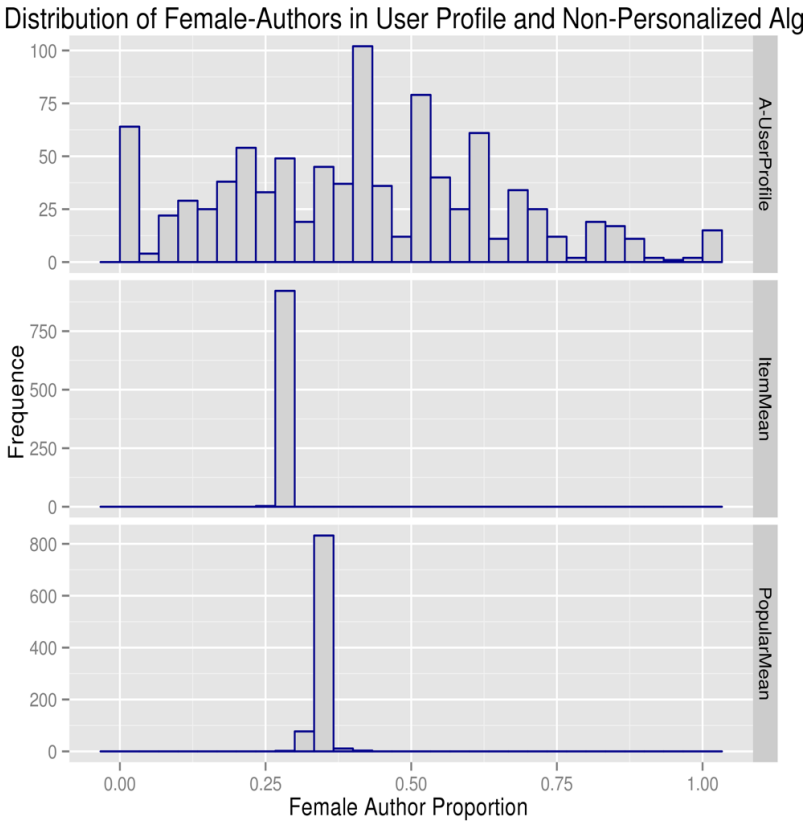

Figure 5 Female Author Distribution in Non-Personalize Algorithms

Observation:

The above histogram shows the distribution of female author in user profile input data and non personalized recommender algorithm's outcome:

- User Profile Input data, more number of user read less female author and over–all proportion rate for female author is 0.43.

- Item Mean Algorithms, more number of user read very less female author and over–all proportion rate for female author is 0.23 which is approximately half of the proportion value of female author in user profile input proportion.
- For Popular algorithm, more number of user read less female author and over–all proportion rate for female author is 0.30 which is considerably less than proportion value of female author in user profile input data.

Author Distribution in Personalize Algorithm

Over here we have computed distribution of female author proportion for each user in Personalize Recommender output. This computed proportion for female authors was observing using histogram. The below distribution of female author provide answer to our second research question RQ 2. Below histogram represent distribution of female author proportion in:

- User Profile Input Data
- SVD
- UserUser Collaborative Filtering

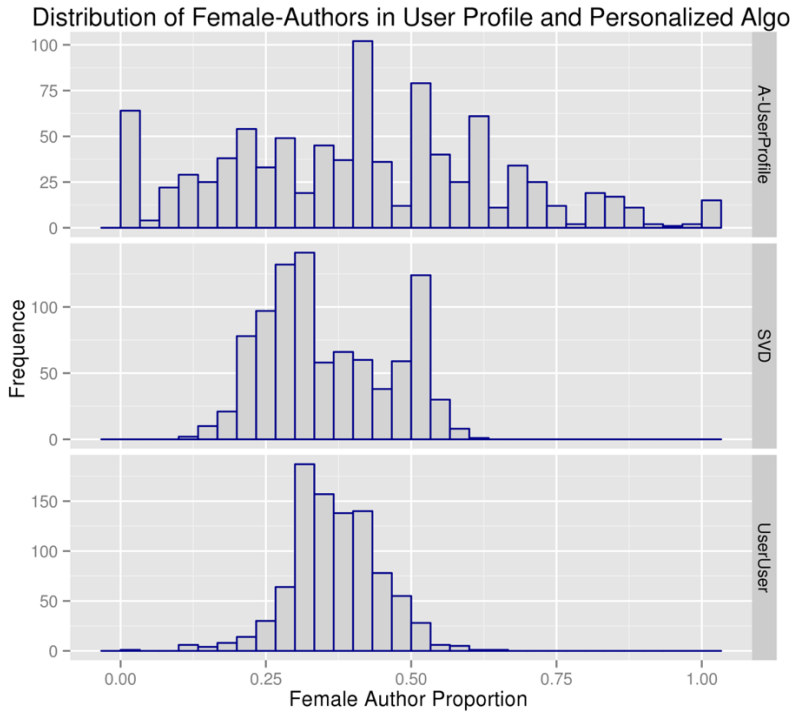

Figure 6 Female Author in Personalize Algorithms

Observation:

The above histogram shows the distribution of female author in user profile input data and output of personalized recommender algorithm:

- User Profile Input data, more number of user read less female author and over proportion rate for female author is 0.43.
- SVD Algorithm, more number of user read less female author and over–all proportion rate for female author is 0.34 which is considerably less than the proportion value of female author in user profile input proportion.
- For UserUser Algorithm, more number of user read less female author and over–all proportion rate for female author is 0.37 which is considerable is less than proportion value of female author in user profile input data.

Distribution of Gender Bias

With Bayesian model we are observing the computed probability of user consumption rate based on the bias in user profile and output of recommender algorithm.

a. Combine Posteriors Plots

As we have computed our posterior in two ways which is pointwise and integral form, we have combine the two posterior distributions which is reflected in the below figure 6.

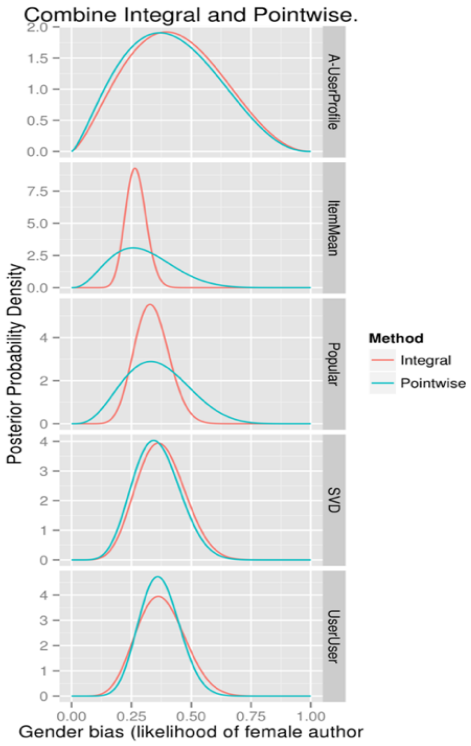

Figure 6 Posterior of Pointwise and Integral form

Observation:

- We had computed theta that minimizes $P(y_u|\theta, n_u) * P(\theta)$ where θ is from our initial inference , which is either from below mention points

- Blue line in the plot is Pointwise posterior which is computed by taking the expected value of α, β.
- Red line in the plot is integral posterior which is computed by taking the integral of the posterior distribution over α, β. In this the distribution of theta can serve as a prior for an additional inference step, here we had computed the mostly likely theta for a particular user.

- Plots represent combine posterior form of pointwise and integral:
  - User Profile Input Data.
  - Item Mean Algorithm.
  - Popular Algorithm.
  - SVD Algorithm.
  - UserUser Algorithm.

b. Credible Interval for Gender Bias

Creditable interval range for $P(\theta|y)$ for User Profile and Recommender System Algorithms. Credible interval range denote the that expected value of θ will fall in this range.

1. User Profile Input Data

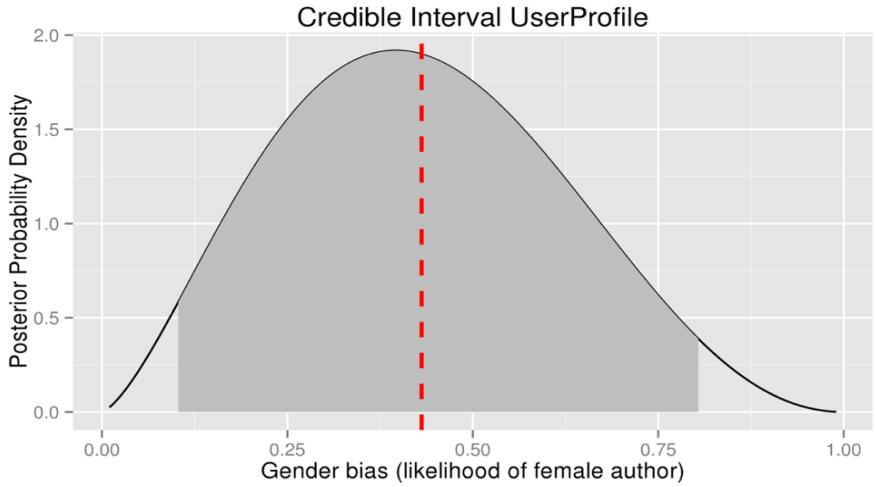

Figure 7 Posterior Distribution for User Profile (integral method), with expected value and 95% credible interval

dotted line: expected value of θ

Observation:

- Expected value of θ is 0.43
- Referring to figure 6 in which shaded part of the plot denote the credible interval range which is 0.10 to 0.80.
- For ItemMean algorithm observation in the figure 6, the expected value of θ is considerably less than threshold value of θ, thus we can determine that there is bias again the female author. The width of the curve for SVD algorithm much broader which denote more variance in the bias.
- User Profile excepted value of θ will serve as threshold value to recommender algorithms.

2. ItemMean Algorithm

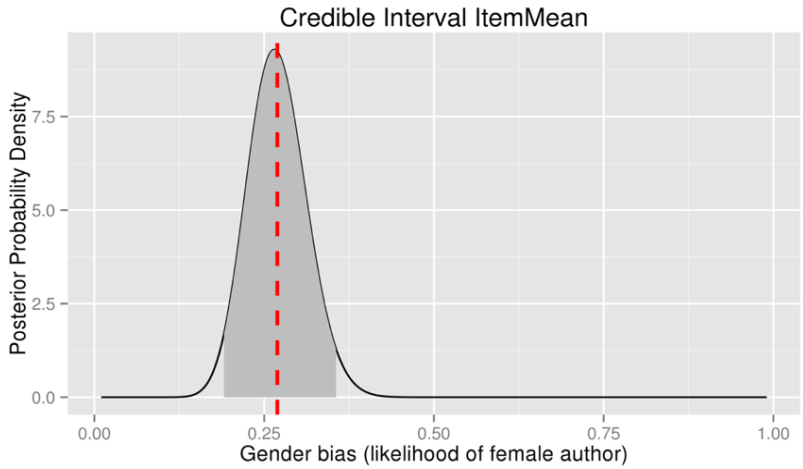

Figure 8 Posterior Distribution for ItemMean (integral method), with expected value and 95% credible interval

dotted line : weighted average of θ, where P⟨θ|y⟩is the weight.

Observation:

- Expected value of θ is 0.26.
- Referring to figure 7 in which shaded part of the plot denote the credible interval range which is 0.10 to 0.35.
- For ItemMean algorithm observation in the figure 6, the expected value of θ is considerably less than threshold value of θ, thus we can determine that there is bias again the female author. The width of the curve for SVD algorithm is not much broader which denote less variance in the bias.

3. Popular Algorithm

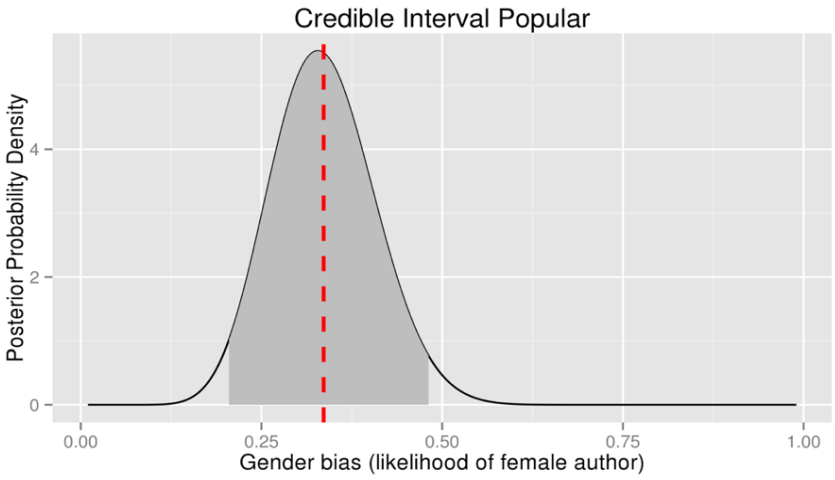

Figure 9 Posterior Distribution for Popular (integral method), with expected value and 95% credible interval

dotted line : weighted average of θ, where P⟨θ|y⟩is the weight.

Observation:

- Expected value of θ is 0.33
- Referring to figure 8 in which shaded part of the plot denote the credible interval range which is 0.20 to 0.48.
- For Popular algorithm observation in the figure 8, the expected value of θ is considerably less than threshold value of θ, thus we can determine that there is bias again the female author. The width of the curve for Popular algorithm is broad which denote variance in the bias.

4. SVD Algorithm

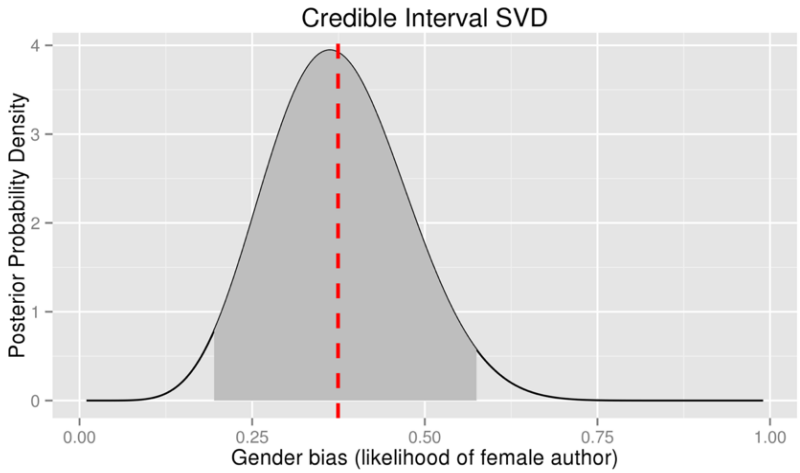

Figure 10 Posterior Distribution for SVD (integral method), with expected value and 95% credible interval

dotted line : weighted average of θ, where P⟨θ|y⟩is the weight.

Observation:

- Expected value of θ is 0.37.
- Referring to figure 9 in which shaded part of the plot denote the credible interval range which is 0.19 to 0.57.
- For SVD algorithm observation in the figure 6, the expected value of θ is considerably less than threshold value of θ, thus we can determine that there is bias again the female author. The width of the curve for SVD algorithm is broader which denote variance in the bias.

5. UserUser Algorithm

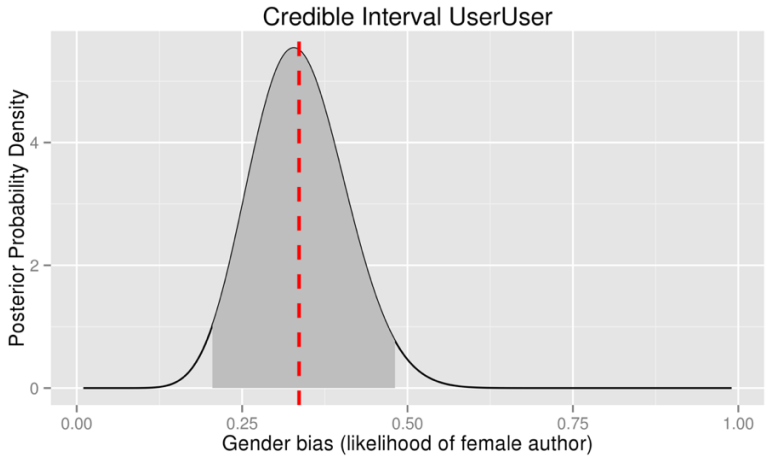

Figure 11 Posterior Distribution for UserUser (integral method), with expected value and 95% credible interval

dotted line : weighted average of θ, where P⟨θ|y⟩is the weight.

Observation:

- Expected value of θ is 0.34.
- Referring to figure 6 in which shaded part of the plot denote the credible interval range which is 0.20 to 0.48.
- For UserUser algorithm observation in the figure 6, the expected value of θ is considerably less than threshold value of θ, thus we can determine that there is bias again the female author. The width of the curve for UserUser algorithm is broad which denote variance in the bias.

Comparison of Author Distribution

Below plots represent the female author correlation between the user profile input data and the output of personalizes recommender algorithms, addressing RQ 3. For

comparison of female author distribution, we have taken only those user details from the recommender outputs which are present in sample data.

Scatter plots for Personalize Algorithm

Figure 13 represent distribution of female author proportion in User Profile data and in output of Recommender System algorithms.

a. SVD Algorithm

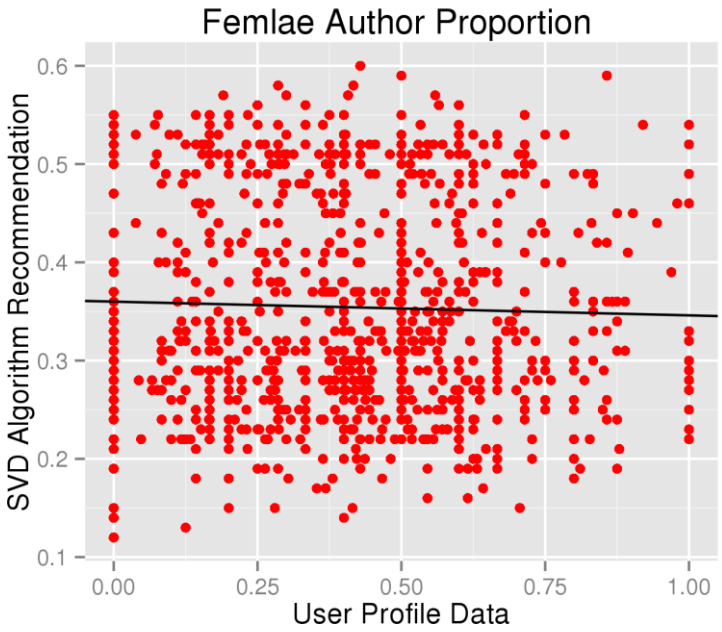

Figure 12 Scatter plot for User Profile vs SVD

x-axis: female author proportion value present in user profile data.

y-axis: female author proportion value present in outcome of SVD algorithms.

Observation:

slope of the mean absolute line value is negative and more of data points are not close to the line, thus user profile and recommender output data does not reflect strong correlation.

b. UserUser Algorithm

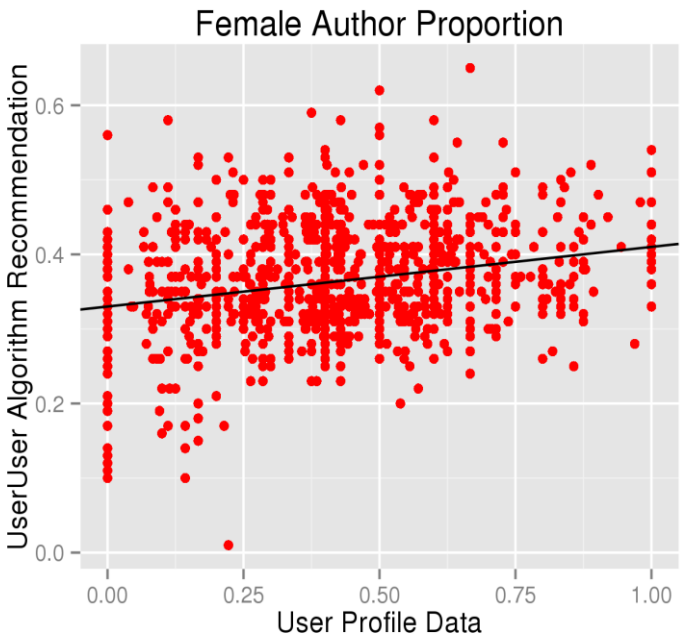

Figure 13 Scatter plots for User Profile vs UserUser

x-axis: female author proportion value present in user profile data.

y-axis: female author proportion present in outcome of UserUser algorithms.

Observation:

slope of mean absolute line value is positive and more of data points are close to the line, thus user profile and recommender output data does reflect correlation.

Predictive Linear Model

Table 2 represent the relationship between the input data and output data of recommender experiment. We have performed predictive linear test between the user profile input data again the outcome of the personalizing algorithm. We can observe that the p-value for UserUser is very low which determine significant in results. From the R^2 value for

useruser, we determine the presence of a relationship between user profile input data and output of this algorithm. Plus, from the coefficient value we say that output is getting affected by input and finally we have positive correlation for useruser. But for sad, as the p-value is on the higher end thus it makes the complete model insignificant.

Table 2 Predictive Model

Algorithm	Correlation	Coefficient	p-value	R^2
UserUser	**0.24**	**0.34**	**2.63e-14**	**0.07442**
SVD	-0.0308	0.36	0.3493	0.00094

V. CONCLUSION AND FUTURE WORK

Conclusion

In this work, we have successfully built a methodological model to explore the potentially discriminatory biases in outcomes of Book Recommender Systems. We have done this by taking the protected characteristic of author i.e. gender in Book Recommender System.

We have observed the distribution of female author gender in user profile input data, here the user tends to read less number female authors books. Then we have obverse the distribution female author output of non-personalize algorithms, here we found that many users read very small count of female author book, this count was less than that we observe in the user profile input data. Plus, we have performed the similar observation in the output of personalizing algorithms, here we found the similar level female author distribution was present as compared to user profile input data.

With the help of the statistical model, we have successfully computed the probability of user consumption rate based on the biases. This model also denoted that non personalize algorithms have strong potential biases against the female author books while the personalize algorithms maintain approximately similar biases rate as compared to biases in the user profile data. This shows that biases present in the input data is get replicated into the output of recommender system.

Finally, we have successfully computed linear predictive computation were we perform the comparison of female author distribution in user profile input data and output of personalize algorithms by this we have successfully observe the existence of relationship between input data and output of recommender system. Plus, this also denote particularly in UserUser algorithm that output is getting affected with input data.

Future Work

We are interested to tried out following things:

- We want to testing our current methodology with different type of recommender system e.g. Movies, Music, Restaurant etc.
- In the current work we have considered author gender as protected characteristics to observe potential discrimination. We would like to explore potentially discriminatory bias based on other demography attributes of author i.e. ethnicity, country, race of the author.
- We want to explore potential discrimination based on the user rating data.
- We want to consider content-based filtering algorithms for the recommender experiment.
- We are planning to implemented various potential definition of discrimination to observe different potentially discriminated biases and might help us to determine the level of fairness in outcome of recommender algorithm.

REFERENCES

Adomavicius, Gediminas, and Alexander Tuzhilin. 2005. "Toward the next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions." Ieeexplore.ieee.org. *Toward the next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions*. April 25. http://ieeexplore.ieee.org/xpl/abstractAuthors.jsp?arnumber=1423975.

Bmerrill, Jeremy. 2014. "Open Gender Tracker." *Beauvoir*. https://github.com/jeremybmerrill/beauvoir.

Custers, Bart, Toon Calders, Bart Schermer, and Zarsky. 2014. *Discrimination and Privacy in the Information Society*. Springer-Verlag Berlin Heidelberg.

Datta, Amit, Michael Carl Tschantx, and Anupam Datta. 2014. "A Tale of Opacity, Choice, and Discrimination." http://arxiv.org/pdf/1408.6491v2.pdf.

Ekstrand, Michael D, John T. Riedl, and Joseph A. Konstan. 2011. "Collaborative Filtering Recommender Systems." In *Collaborative Filtering Recommender Systems*, 82. Now Publishers.

Ekstrand, Michael, Michael Ludwing, Joseph A. Konstan, and John T. Riedl. 2011. "Rethinking the Recommender Research Ecosystem: Reproducibility, Openness, and Lenskit." In .

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. "Gelman's Chapter 5th." In *Bayesian Data Analysis*, 4.

Jelveh, Zubiin, and Michael Luca. 2015. "Towards Diagnosing Accuracy Loss in Discrmination-Aware Classification: An Application to Predictive Policing." In . http://www.fatml.org/papers/Jelveh_Luca.pdf.

Konstan, Joseph A., and Others. 1997. "Applying Collaborative Filtering to Usenet News." *GroupLens*. April 3. http://dx.doi.org/10.1145/245108.245126.

Linden, Greg, Brent Smith, and Jeremy York. 2003. "Amazon.com Recommendations: Item-to-Item Collaborative Filtering." *Ieeexplore.ieee.org*. January 22. http://dx.doi.org/10.1109/MIC.2003.1167344.

Madrigal, Alex. 2016. "Predictive Policing in Action." http://fairness.haverford.edu.

McNee, Sean M, John Riedl, and Joseph A. Konstan. 2006. "Being Accurate Is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems." In . ACM.

Reidsma, Matthew. 2016. *ALGORITHMIC BIAS IN LIBRARY DISCOVERY SYSTEMS*. http://fairness.haverford.edu/index.html.

Sowell, Thomas. 2015. "The 'Disparate Impact' Racket." *The "Disparate Impact" Racket*. April 14. http://www.realclearpolitics.com/articles/2015/03/10/the_disparate_impact_racket_125880.html.

Tufekci, Zeynep. 2014. "What Happens to Ferguson Affects Ferguson: Net Neutrality, Algorithmic Filtering and Ferguson." https://medium.com/message/ferguson-is-also-a-net-neutrality-issue-6d2f3db51eb0.

Vanetta, Marcos. 2014. "Gender-Detector." https://pypi.python.org/pypi/gender-detector.

Zafar, Muhammed Billal, Isabel Valere, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2015. "Fairness Constraints: A Mechanism for Fair Classification." In .

Zliobiate, Indrè. 2015. “On the Relation between Accuracy and Fairness in Binary Classification.” In *On the Relation between Accuracy and Fairness in Binary Classification*.